Biochemical and Biophysical Research Communications 430 (2013) 1340–1343

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Mini Review

Can ENCODE tell us how much junk DNA we carry in our genome?

Deng-Ke Niu *, Li Jiang

MOE Key Laboratory for Biodiversity Science and Ecological Engineering and Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Article history:
Received 25 November 2012
Available online 22 December 2012

Keywords:
ENCODE
Biochemical activity
Repressive transcription
Conservation
Knockout

ABSTRACT

One of the large, unsolved problems in human genetics is the proportion of functional sequences in genomes. Recently, the encyclopedia of DNA elements consortium revealed that the majority of the genome is biochemically active, which were described as biochemical functions. This has been used as evidence to pronounce the death of the junk DNA concept. In evolutionary biology, junk DNAs are sequences whose gain or loss does not seriously affect fitness of the host organism. In the human genome, a large amount of biochemical activity should be to repress the sequences so as to avoid their harmful expression. The biochemical activity is very different from functionality in the light of evolution. The single nucleotide polymorphism sites associated with disease and other phenotypes may be functional, but their abundance in the active genome regions is not reliable evidence of functionality. Because of sequence-independent functions, the proportion of functional regions would be underestimated when sequence constraints are used alone. Knockout may be the most effective means of distinguishing functional sequences from junk DNA.

1. Introduction

The sequencing of the human genome has revealed that coding sequences cover only about 1.2% of the euchromatic genome [1]. A big unknown of this is the estimate of non-coding sequences that are functional. A systematic approach to identification of the functional sequences was initiated in 2003 for this reason. Now, an encyclopedia of DNA elements (ENCODE) has been published for the human genome [2]. The vast majority (80.4%) of the genome was found to have some sort of biochemical activity (e.g. RNA transcription, transcription-factor-binding, chromatin structure, and histone modification) in at least one cell type. Given that the project assayed only 147 of the hundreds of human cell types, the total amount of active sequence is surely underestimated. One likely estimation is that almost all of the genome is biochemically active [3]. The ENCODE consortium has maintained that the majority of the human genome may affect cellular and large-scale phenotypes and thus should be as described as having biochemical functions. This conclusion was further interpreted as the death knell for the junk DNA concept [3–5]. The human genome now seems to be perfectly designed, as advocated by creationists. In this paper, we first review the inaccuracy of this interpretation of biochemical activity as evidence against the junk DNA concept and then discuss alternative strategies for the estimation of the proportion of DNA with biological functions.

2. Differences between biochemical activities and biological functions and relevance to the junk DNA concept

Using extensive biochemical analyses, the ENCODE consortium found that the vast majority of the human genome was not inert but rather involved in various biochemical activities [2]. In addition to the 39.54% of the genome that is made up of protein-coding genes, they found that intergenic sequences were widely transcribed. In total, transcriptionally active sequences made up 62% of the genome. The addition of transcription factor binding sites, DNase I hypersensitive sites, methylated DNA, and other regions that showed evidence of regulation and interaction brought this figure up to 80.4% [2]. It is very likely to approach 100% if all cell types and all DNA-binding proteins are assayed [3]. In fact, there is one biochemical activity that is already known to cover the entire human genome, DNA replication. DNA sequences are replicated simply because of their existence in the nucleus, without any implications regarding their functionality. In the other extreme, a small proportion of the genome, coding for proteins, rRNAs, and tRNAs, is not only biochemically active but also has unquestionable biological function. The debate is whether the transcription of unknown non-coding RNAs, the binding of tran-

Abbreviations: ENCODE, encyclopedia of DNA elements; GWAS, genome-wide association study; SNP, single nucleotide polymorphism.

* Corresponding author. Fax: +86 10 58802064.
E-mail addresses: dkniu@bnu.edu.cn, dengkeniu@hotmail.com (D.-K. Niu).

0006-291X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2012.12.074

scription factors, methylation, and other biochemical activities are really convincing evidence of functionality.

First, transcriptional activity is very different from biological functionality. Not all transcription has a functional purpose. Transcription may be repressive transcription [6]. In many eukaryotes, the transcription of non-coding RNAs from heterochromatic regions and repetitive elements is required for the formation of repressive chromatin structures within these loci [7,8]. The repressive transcription of these loci indicates that active expression of them is probably harmful, at least in the cell types in which the repressive transcription was observed. Further evidence is required to determine whether these loci are transcribed in other types of cells for functional purposes. Although we do not know the exact rate at which repressive transcription takes place in the human transcriptome, at least, it is inaccurate to assign functionality using transcriptional processes alone.

In addition, many eukaryotic transcripts are recognized as aberrant mRNA and thus degraded by mRNA surveillance mechanisms [9,10]. These aberrant transcripts may have been caused by either expression errors of functional genes or the transcription of aberrant genes or non-genetic sequences. Although the transcription and degradation of these aberrant RNA transcripts are surely costly to the cells, they are not necessarily eliminated by natural selection. In organisms with small effective population sizes like humans and mice, the accumulation of neutral and slightly deleterious sequences is inevitable. For example, the transcription of introns exerts an energetic burden upon the organism. However, quantitative estimation of the energetic burden of a long intron in a highly expressed human gene indicate that the energetic cost is still too negligible to be an effective force to reduce intron size [11].

DNA methylations and histone modifications are signs of repression in transposable elements [12–15]. Because transposable elements occupy almost half of the human genome, it is natural that repressive activities would be detected in numerous parts of the human genome. Like repressive transcription, DNA methylation and histone modifications do not indicate anything about the biological functions of the loci.

DNase I hypersensitivity is a good marker of chromatin accessibility, which facilitates the recruitment of transcriptional machines and regulatory proteins. The binding of transcription factors to a given region indicates that transcription is being regulated in that region. These biochemical activities are evidence of the transcriptional potential of the loci and nearby sequences. If the transcription of numerous loci is as minimally regarded as biological function, these activities, which are hallmarks of transcriptional potential, may not be useful indicators of biological function.

On the other side, we have definite evidence for the function of some non-coding sequences. For example, the RNAs transcribed from mouse B2 and human Alu SINEs were found to bind RNA polymerase II and switch global gene expression upon cell stress [16]. And the more simple repeats, microsatellites, have direct role in chromatin organization [17]. Undoubtedly, many regions of the parts of the genome that ENCODE found to have biochemical activity are actually functional. However, biochemical activity itself is not reliable evidence of the proportion of functional sequences in the human genome [18]. The ENCODE results cannot demonstrate or even provide evidence that the human genome is perfectly designed in either the creationist or evolutionary sense. As a concept in evolutionary biology, the junk DNA survives after ENCODE.

3. Association with human disease or other phenotype is not definite evidence of biological function

In recent years, genome-wide association studies (GWAS) have linked thousands of DNA variations to hundreds of human traits ranging from normal phenotypes to diseases. Most of these variations are found in non-coding regions. Using a detailed map of the biochemical activities of non-coding regions in the human genome, the ENCODE consortium makes a large step from the identification of associations to the interpretations of underlying causality [5]. For example, they found that 12% of GWAS single nucleotide polymorphism (SNP) sites are in transcription-factor-occupied regions and 34% in DNase I hypersensitive regions [2]. The enrichment of the GWAS SNPs in the active regions of the genome is much stronger evidence of functionality for the non-coding regions than the biochemical activity.

However, cautions must still be taken before claiming that these regions are functional. First, the 4492 GWAS SNP sites make for a very low coverage on the active non-coding regions observed by the ENCODE consortium. For example, the transcription-factor-binding regions that cover 231 Mb contain only 539 GWAS SNP sites. One SNP site corresponds to 429 kb of transcription-factor-binding region, on average. As transcription-factor-binding sites are typically 10 nt long [19], we might deduce the functionality of a non-coding region with 10 bases (or even 100 bases) from an SNP site. However, it is too bold to assign functionality to a 429 kb segment solely by the presence of a trait-associated SNP.

Second, not all diseases or phenotypic traits are selectively visible. In organisms, like humans, the force of natural selection declines with age, especially after reproductive stage. Alleles that are neutral at young ages but deleterious at old ages could accumulate in population. A collection of the diseases and phenotypic traits associated with these alleles is senescence. This is the evolutionary mechanism of aging proposed as the mutation accumulation model [20]. Evidence for this model is accumulating [21,22]. In this way, the existence of the SNP sites in association with these traits and the regulatory elements involved in the expression of these traits are not maintained by natural selection. Their presence in and absence from the human genome are regulated by chance. They are not necessary for the survival of humans as a species. Therefore, the enrichment of the GWAS SNP sites in the active regions of the genome is not solid evidence for functionality.

In addition, an association between any SNP site and any given disease does not necessarily indicate that the DNA region is functional with respect to normal phenotypic features. A loss-of-function mutation can convert a normal phenotypic trait into a disease trait. The existence of this type of SNP has implications for the functionality of the locus. In contrast, gain-of-function mutations create disease from nothing. The existence of this type of SNP shows the mutational hazard of the loci but it does not have any implications for functionality.

4. Functional sequences include but are not limited to sequences under purifying selection at the nucleotide level

As discussed above, it is not appropriate to define the proportion of functional sequences as the proportion of sequences with biochemical activity. More reliable ways of distinguishing functional sequences from junk DNA are discussed below.

If a DNA element is functional, its loss, gain, or change will affect the fitness of the host organism. Purifying selection takes place to preserve the element in the genome. That is, the element is under evolutionary constraints. In the light of natural selection, sequence conservation has usually served as a proxy for functionality [23,24]. Recent surveys of sequences under purifying selection showed that 5% of the human genome is conserved across all mammals and 4% is subject to lineage-specific constraints [25,26]. The coding sequences make 1.22% of the human genome [1,2]. When the 9% constraint sequences are considered as functional regions in the human genome, the non-coding part area is

about 6.4 times the size of the coding area. That is, 6.4 kb of non-coding sequences carry out the expression and regulation of 1 kb of coding sequences. The gene expression and regulatory machine is not unimaginably simple.

Many highly conserved elements had not been found to be biochemically active in previous studies [27,28]. This was used as a reason to abandon sequence constraint as a proxy for potential function [28]. Now, the reason is not sound because the ENCODE results indicate that all the human genome are biochemically active [2,3].

Although the sequence constraint could be used as a proxy for potential functionality, we believe that this method underestimates the proportion of functional sequences if it is limited to counting constraints at the nucleotide level. Many non-coding sequences may be retained in the genome because of their sequence-independent effects on fitness. A number of hypotheses have been proposed regarding possible sequence-independent functions of non-coding sequences. Here are some examples. The nucleoskeletal theory of the evolution of genome size proposes that a large amount of non-coding sequences exist because of their role in determining nuclear volume [29]. It was also proposed that introns may be retained in eukaryotic genes because of their ability to reduce transcription-associated damages to nearby exons [30,31]. A more reliable hypothesis is that the transcription and splicing of introns may be one mechanism by which the temporal expression of various genes is regulated [32–35]. Recently, the introns in the mouse gene *Hes7* were found to cause a 19 min delay in its expression [36]. Removal of the introns abolished the oscillating expression of the gene, leading to severe segmentation defects in mouse development [36]. In addition, the exon–exon junction complexes deposited on mRNA transcripts during splicing are required for the recognition of aberrant mRNA molecules with premature translation-termination codons in mammals [10]. The specific nucleotide sequences of the introns are not required, but the introns are retained in the genome because their splicing products provide markers for the mRNA surveillance process. Introns and their splicing have been found to enhance gene expression [37,38]. However, whether expression changes in most genes are under natural selection is still in debate [39,40]. It is premature to regard the enhancing effect as an adaptive role of introns.

In addition, there are many non-coding RNAs (like the U12-type snRNAs [41]) that are conserved in secondary and higher structures, but not in the nucleotide sequences.

5. Knockout may be the most effective strategy for determining functional sequences

Compared with the strategies discussed above, knockout may be more effective to understand the function of a gene or a DNA segment. By comparing the knockout organism to a wild-type organism, definite conclusion on the functionality of the knockout sequence could be approached. However, the interpretation of knockout results must be made cautiously if only a limited number of experimental conditions have been tested. In yeast (*Saccharomyces cerevisiae*) grown in rich medium at 30 °C, only 15% of the protein-coding genes were found to be essential for growth [42]. However, assays of the knockout strains under 1144 different sets of chemical conditions indicated that nearly all yeast genes are essential to the survival of the yeast organism in nature [43]. Similar patterns have been observed for the spliceosomal introns of *S. cerevisiae*. The majority of introns can be removed without serious effects on growth under laboratory conditions, but these same introns may be required for growth under stress conditions [44,45]. In humans, many experiments are not practical. However, exhaustive knockout experiments can be carried out in mammals such as monkeys. Then the proportion of functional sequences in the human genome can be approximately estimated.

Acknowledgments

We would like to thank Yu-Fei Yang for his helpful suggestions on this paper. This work was supported by the National Natural Science Foundation of China (grant numbers 31121003 and 31071112).

References

[1] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931.
[2] The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74.
[3] E. Yong, ENCODE: the rough guide to the human genome, Discover Magazine (2012).
[4] E. Pennisi, ENCODE project writes eulogy for junk DNA, Science 337 (2012) 1159–1161.
[5] J.R. Ecker, W.A. Bickmore, I. Barroso, J.K. Pritchard, Y. Gilad, E. Segal, Genomics: ENCODE explained, Nature 489 (2012) 52–55.
[6] M.G. Guenther, R.A. Young, Repressive transcription, Science 329 (2010) 150–151.
[7] M. Buhler, D. Moazed, Transcription and RNAi in heterochromatic gene silencing, Nat. Struct. Mol. Biol. 14 (2007) 1041–1048.
[8] S.I. Grewal, S.C. Elgin, Transcription and RNA interference in the formation of heterochromatin, Nature 447 (2007) 399–406.
[9] C.J. Shoemaker, R. Green, Translation drives mRNA quality control, Nat. Struct. Mol. Biol. 19 (2012) 594–601.
[10] O. Isken, L.E. Maquat, Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function, Gene. Dev. 21 (2007) 1833–1856.
[11] Y.-F. Huang, D.-K. Niu, Evidence against the energetic cost hypothesis for the short introns in highly expressed genes, BMC Evol. Biol. 8 (2008) 154.
[12] F.K. Teixeira, V. Colot, Repeat elements and the *Arabidopsis* DNA methylation landscape, Heredity 105 (2010) 14–23.
[13] J.A. Yoder, C.P. Walsh, T.H. Bestor, Cytosine methylation and the ecology of intragenomic parasites, Trends Genet. 13 (1997) 335–340.
[14] J. Zempleni, Y.C. Chew, B. Bao, V. Pestinger, S.S. Wijeratne, Repression of transposable elements by histone biotinylation, J. Nutr. 139 (2009) 2389–2392.
[15] C. Liu, F. Lu, X. Cui, X. Cao, Histone methylation in higher plants, Annu. Rev. Plant Biol. 61 (2010) 395–420.
[16] S.L. Ponicsan, J.F. Kugel, J.A. Goodrich, Genomic gems: SINE RNAs regulate mRNA production, Curr. Opin. Genetics Dev. 20 (2010) 149–155.
[17] Y.-C. Li, A.B. Korol, T. Fahima, A. Beiles, E. Nevo, Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review, Mol. Ecol. 11 (2002) 2453–2465.
[18] S.R. Eddy, The C-value paradox, junk DNA and ENCODE, Curr. Biol. 22 (2012) R898–R899.
[19] A.J. Stewart, J.B. Plotkin, Why transcription factor binding sites are ten nucleotides long, Genetics 192 (2012) 973–985.
[20] D. Fabian, T. Flatt, The evolution of aging, Nat. Educ. Knowl. 3 (2011) 9.
[21] K.A. Hughes, J.A. Alipaz, J.M. Drnevich, R.M. Reynolds, A test of evolutionary theories of aging, Proc. Natl. Acad. Sci. USA 99 (2002) 14286–14291.
[22] J.L. Graves, Gene expression in late-life, Front. Genet. 3 (2012) 156.
[23] C.I. Castillo-Davis, The evolution of noncoding DNA: how much junk, how much func?, Trends Genet 21 (2005) 533–536.
[24] C.P. Ponting, R.C. Hardison, What fraction of the human genome is functional?, Genome Res 21 (2011) 1769–1776.
[25] L.D. Ward, M. Kellis, Evidence of abundant purifying selection in humans for recently acquired regulatory functions, Science 337 (2012) 1675–1678.
[26] K. Lindblad-Toh, M. Garber, O. Zuk, M.F. Lin, B.J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, et al., A high-resolution map of human evolutionary constraint using 29 mammals, Nature 478 (2011) 476–482.
[27] The ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 447 (2007) 799–816.
[28] J.A. Stamatoyannopoulos, What does our genome encode?, Genome Res 22 (2012) 1602–1611.
[29] T. Cavalier-Smith, Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox, J. Cell Sci. 34 (1978) 247–278.
[30] D.-K. Niu, Y.-F. Yang, Why eukaryotic cells use introns to enhance gene expression: Splicing reduces transcription-associated mutagenesis by inhibiting topoisomerase I cutting activity, Biol. Direct 6 (2011) 24.
[31] D.-K. Niu, Protecting exons from deleterious R-loops: a potential advantage of having introns, Biol. Direct 2 (2007) 11.
[32] D. Gubb, Intron-delay and the precision of expression of homoeotic gene products in *Drosophila*, Dev. Genet. 7 (1986) 119–131.

[33] I.A. Swinburne, P.A. Silver, Intron delays and transcriptional timing during development, Dev. Cell 14 (2008) 324–330.
[34] I.A. Swinburne, D.G. Miguez, D. Landgraf, P.A. Silver, Intron length increases oscillatory periods of gene expression in animal cells, Genes Dev. 22 (2008) 2342–2346.
[35] C.S. Thummel, Mechanisms of transcriptional timing in *Drosophila*, Science 255 (1992) 39–40.
[36] Y. Takashima, T. Ohtsuka, A. Gonzalez, H. Miyachi, R. Kageyama, Intronic delay is essential for oscillatory expression in the segmentation clock, Proc. Natl. Acad. Sci. USA 108 (2011) 3300–3305.
[37] H. Le Hir, A. Nott, M.J. Moore, How introns influence and enhance eukaryotic gene expression, Trends Biochem. Sci. 28 (2003) 215–220.
[38] A.B. Rose, S. Emami, K. Bradnam, I. Korf, Evidence for a DNA-based mechanism of intron-mediated enhancement, Front. Plant Sci. 2 (2011) 98.
[39] F. Staubach, M. Teschke, C.R. Voolstra, J.B. Wolf, D. Tautz, A test of the neutral model of expression change in natural populations of house mouse subspecies, Evolution 64 (2010) 549–560.
[40] Y. Gilad, A. Oshlack, S.A. Rifkin, Natural selection on gene expression, Trends Genet. 22 (2006) 456–461.
[41] J.J. Turunen, E.H. Niemelä, B. Verma, M.J. Frilander, The significant other: splicing by the minor spliceosome,, WIREs RNA 4 (2013) 61–76.
[42] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, et al., Functional profiling of the *Saccharomyces cerevisiae* genome, Nature 418 (2002) 387–391.
[43] M.E. Hillenmeyer, E. Fung, J. Wildenhain, S.E. Pierce, S. Hoon, W. Lee, M. Proctor, R.P. St Onge, M. Tyers, D. Koller, et al., The chemical genomic portrait of yeast: Uncovering a phenotype for all genes, Science 320 (2008) 362–365.
[44] J. Parenteau, M. Durand, S. Veronneau, A.-A. Lacombe, G. Morin, V. Guerin, B. Cecez, J. Gervais-Bird, C.-S. Koh, D. Brunelle, et al., Deletion of many yeast introns reveals a minority of genes that require splicing for function, Mol. Biol. Cell 19 (2008) 1932–1941.
[45] J. Parenteau, M. Durand, G. Morin, J. Gagnon, J.-F. Lucier, R.J. Wellinger, B. Chabot, S. Abou Elela, Introns within ribosomal protein genes regulate the production and function of yeast ribosomes, Cell 147 (2011) 320–331.