Investigative Genetics

EDITORIAL Open Access

ENCODE and its first impractical application

Bruce Budowle

The C value paradox, as initially coined, was encountered in early eukaryotic genomic studies with the oddity that genome size was not necessarily correlated with organism complexity [1]. With the discovery of non-coding DNA in the 1970s, it became apparent that the size of the eukaryotic genome was not related to the number of genes contained within it. Indeed, only a small portion (approximately 2%) of the human genome carries coding genes [2-4], the rest being the so-called "junk DNA" [5]. The human genome project further elucidated the number of genes in our genomes - counting a paltry 20,000 to 25,000 genes [2-4]. With so few genes one might ask "how could such a complex organism as *Homo sapiens* pass on the necessary genetic blueprint to the next generation?" An equally enticing question could be "how could nature be so wasteful and commit so much junk DNA to the human genome?" The Encyclopedia of DNA Elements (ENCODE) project has shed some light on these two questions. There is not one paper to cite but greater than 30 studies [6] that were coordinated and published in concert describing the results of a multi-year consortium effort to catalogue the functional elements of human DNA. Hundreds of authors reported on analyses of thousands of data sets. A good summary of the work is captured in the ENCODE Project Consortium's September 2012 publication titled "An integrated encyclopedia of DNA elements in the human genome" [7]. The ENCODE project identified a large number of functional elements, defined as sites that encoded a product or exhibited a biochemical signature in the human genome. The power of current DNA sequencing technologies made the Consortium project possible. The depth of analysis is impressive. In this one paper more than 1,600 data sets were analyzed for a multitude of elements including human protein-coding and non-coding RNAs, pseudogenes, RNA from different cell lines, binding locations of a number of DNA-binding proteins and RNA polymerase components, DNase I hypersensitive sites, locations for histone modifications, and DNA methylation. The most exciting finding and one that may begin to address the two questions posed above was that 80.4% of the genome has a biochemical function, that is, it is covered by or near at least one ENCODE-identified element. More precisely, a large portion of the human genome contains a regulatory event. The authors state that "95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction. . ., and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE." The outcome is that the noncoding junk DNA is far from being useless genome filler. Instead, seemingly inert DNA can influence functional genes. The nature of genetic and epigenetic control is quite complex and exquisite and today all that more appreciated. ENCODE is a public resource that will contribute substantially to the understanding of gene expression and mechanisms of disease and, hopefully, cures.

Surprisingly though, what might be the first application of ENCODE data is not directed toward improving human health through molecular biology. Bolstered by the newly found functional nature of a greater portion of the junk DNA, an appeal in a U.S. Court has been brought forward in part on a basis that information derived from typing the short tandem repeat (STR) markers used in forensic human identification worldwide violates an individual's privacy [8]. The argument exploits ENCODE data to suggest that there is some noticeable predictive power hidden with the forensic STRs related to the health status of an individual. After all the Consortium publication suggests that "Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation." Such ENCODE information should be understood and the limitations should be appreciated; we are far from extracting predictive power and unlikely to do so with the forensically-relevant STRs. Even without knowing any causal relationship, as now might be intimated by some with the ENCODE project data, association studies between the forensic STR loci and disease genes generally have come up empty, providing little if

Correspondence: Bruce.Budowle@unthsc.edu
Department of Forensic and Investigative Genetics, Institute of Applied Genetics, University of North Texas Health Science Center at Fort Worth, 3500 Camp Bowie Blvd, Fort Worth, TX 76107, USA

any predictive power. One could argue that one STR, the TH01 locus, has been shown to have some effect on expression at the gene *Tyrosine Hydroxylase* [9] well before the onset of the ENCODE project. Still armed with such information, there is little predictive power regarding an individual's health status by knowing the allelic repeat state of the TH01 locus. Such limited power is not surprising; next generation sequencing and genome wide association studies overwhelmingly find that most diseases are genetically complex and one marker provides little value in determining risk of disease or outcome from therapeutics.

It would be a shame that the phenomenal effort that brought forth ENCODE might be misused to attempt to breach the foundations of forensic DNA typing. ENCODE's value is in laying a foundation of the intricate functionality of the human genome that someday may help improve the human condition. Certainly, claims of privacy violations via human identification by STR typing are unfounded and criticizing this powerful forensic tool, based on ENCODE data, does not improve the human condition.

Received: 3 January 2013 Accepted: 3 January 2013
Published: 17 January 2013

References

1. Thomas CA Jr: **The genetic organization of chromosomes.** *Annu Rev Genet* 1971, **5:**237–256.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, *et al*: **The sequence of the human genome.** *Science* 2001, **291:**1304–1351.
3. International Human Genome Sequencing Consortium, Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860–921.
4. e! Ensembl: *Human Assembly and Gene Annotation, Coding Genes*. http://useast.ensembl.org/Homo_sapiens/Info/Annotation/#genebuild.
5. Ohno S: **So much "junk" DNA in our genome.** *Brookhaven Symp Biol* 1972, **23:**366–370.
6. The ENCODE Project: *ENCyclopedia Of DNA Elements*. 2012. http://www.genome.gov/10005107.
7. The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57–74.
8. Krakauer H: *Appeal against DNA fingerprinting cites ENCODE project*. 2012. http://www.newscientist.com/article/dn22331-appeal-against-dna-fingerprinting-cites-encode-project.html.
9. Albanèse V, Biguet NF, Kiefer H, Bayard E, Mallet J, Meloni R: **Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite.** *Hum Mol Genet* 2001, **10:**1785–1792.

doi:10.1186/2041-2223-4-4
Cite this article as: Budowle: ENCODE and its first impractical application. *Investigative Genetics* 2013 **4**:4.