Review Article

The ENCODE Project: Deciphering the Human Genome's Rosetta Stone

Aftab J. Ahmed, PhD
Chief Executive Officer
Bio-Aging, Inc.
Peoria, Illinois

Abstract

The successful completion of the Encyclopedia of DNA Elements (ENCODE) Project in 2007 continues the pursuit of a systematic approach to catalogue and annotate genetic perturbation and its relation with disease susceptibility. The ENCODE Project is a logical step in augmenting the genomic sequence data produced by the Human Genome Project (HGP) with functional information. The various initiatives to isolate the causative gene associations with common diseases have met with understandable enthusiasm. These are, however, initial steps in better understanding human biology. To fully realize the potential of these powerful approaches, an appraisal of their ramifications will increasingly be in order. This article presents both the promise and challenges of genetic medicine in a broad perspective.

Key words: Human genome, functional genomics, drug development, personalized medicine, single nucleotide polymorphisms (SNPs), genetics, healthcare policy.

It seems an ever-greater emphasis is being placed on harnessing the power of genome mining. The year 2007 might well end up being thought of as a turning point for research on the human genome, as the imbrications of the genome's code began to yield inklings of links between DNA sequences and human health and disease, the underlying objective of these efforts. For example, the sequencing of personal genomes of such luminaries as James Watson and Craig Venter grabbed headlines.[1,2] Similarly, genome-wide associations between "gene variants" and some of the more common diseases generated quite a bit of excitement (*vide infra*). To top it all off, came the results from the pilot phase of the ENCODE (Encyclopedia of DNA Elements) Project.

These forays simultaneously shed light on the complexity of the genome as much as they dispensed a sobering dose of caution in the interpretation of emerging results. A common thread ran through these multifarious approaches: to increase the odds in favor of personalized medicine. The Human Genome Project (HGP) provided the impetus for these attempts, in which the sequencing of the entire human DNA resulted in annotated sequences for potential genes and other genomic characteristics being contextualized.[3,4] As such, the HGP laid the foundation for functional genomics. While significant, HGP was only a first step in "whole-genome" approaches to foster better understanding of human biology. Genomes of several species have been subsequently sequenced and have sharpened the definition of key genomic features under evolutionary constraints, including presumed functional elements such as protein-coding sequences (or exons) and

Correspondence should be directed to

Aftab J. Ahmed, PhD
aftabahmed@msn.com

other regulatory regions.

To amplify the results of HGP, the International Haplotype Mapping Project (HapMap) has catalogued the localization and distribution of common genetic variants within and among human populations.[5] These gene variants reflect differences — and, hence, heterogeneity — in specific regions of DNA. In turn, this has ascribed additional dimensionality to the sequence information generated. The human genome comprises roughly 5-6 billion individual code letters, or nucleotides (bases). While more than 99.5% of the nucleotide sequences are identical between any two people, still millions of differences exist in base sequences, or genetic variants, in the DNA, called single nucleotide polymorphisms (SNPs). It is these variations that account for genetically determined differences among individuals. These differences are inherited in large blocks of haplotypes, which are closely linked genetic markers on a chromosome that are inherited together. Because a haplotype is identified by the pattern of SNPs, interindividual differences are estimated to number around 15 million or so in the human genome. As of mid-2007, the HapMap Project had charted more than 3 million SNPs in different haplotypes.[6] Thus, whereas the HGP provides linear DNA sequence and a gene list with possible (computational) prediction of function, the HapMap Project furnishes information on alleles and haplotypes without correlation to the phenotype.

The ENCODE Project has set for itself a formidable goal in bridging the gaps. It seeks to map the arrays of DNA sequence elements, which not only include genes per se but also regulatory sequences controlling them, including, but not exclusive to, their promoters and enhancers, repressors and silencers, exons, and RNA transcripts.[7] A research consortium undertook the gargantuan task of investigating the diversity of DNA sequences, albeit in only 1% (comprising some 30 million nucleotides, or 30 megabases) of the human genome, which was represented by 44 regions with 400 known genes selected by certain criteria, encompassing the well-studied genes and availability of comparative sequence data from other species. Of course, the intent is to extend the pilot project to the entire genome in the future. It is anticipated that the insights garnered from the ENCODE Project should help answer functional and mechanistic questions pertaining to health and disease.

One of the most salient themes to emerge from this exercise in functional genomics is that nearly the entire genome may be expressed as primary RNA transcripts that extensively overlap and, importantly, include noncoding sequences.[6] For many years, most of the human genome has been thought to be "junk." In that sense, the DNA may be considered a molecular equivalent of "bloatware," which are trial programs in new personal computers that take up hard drive space and slow down the computer. Such was the picture of human genome that emerged as early as the 1970s, which is in stark contrast to the expectation that the DNA would be likely pared down to its bare essentials. Instead, it was found that the bulk of DNA was nonfunctional with only about 1.2% of its sequences encoding proteins.

Because DNA is transcribed in its near entirety , it muddles the central dogma of genetics, which could hardly be simpler: DNA is transcribed into RNA that is then translated into protein(s), the workhorses of the cell. The preponderance of RNA transcripts bespeaks not only of a literal glut of regulatory switches, but it is also considered, tantalizingly, to confer a potential adaptational advantage. When introns (stretches of intervening sequences interspersed within the coding regions of genes) were discovered, it was immediately assumed that these sequences were nonfunctional, even though they were transcribed into RNA. By consensus, they were considered leftovers of the genome's early evolutionary history. Given that roughly 45% of the mammalian genome is derived from transposons, which are mainly parasitic hitchhikers, the concept of "selfish DNA" germinated, which reinforced the view of eukaryotic genomes, including the human, as largely comprising evolutionary detritus. With its findings that roughly 93% of the DNA is transcribed into RNA in different cell types, the ENCODE Project has now allowed a few big proverbial elephants to enter the drawing room. The profusion of RNA transcripts suggests that considerable information — regulatory in the first instance — must lie outside of the exonic boundaries of DNA sequences specifying proteins.

This is evidenced by the rather complex pattern of transcription of DNA sequences into nonprotein-coding RNAs (ncRNAs). This somewhat extravagant expression of the genetic information appears to be developmentally regulated and hints at a function that has yet to be fully appreciated.[9] Simply, the function of ncRNAs is increasingly apparent in the regulation of diverse cellular processes and may be significant in human health. Accordingly, changes in

ncRNAs have been implicated in heart attacks and cancer.[10] Equally, some ncRNAs are expressed in the brain, and at least one is involved in the behavioral response. Thus, the untranslated RNA transcript BC1 is normally expressed in the mouse brain. Knockout mice strains without this RNA display no gross physical impairment but were shown to have reduced exploratory behavior and, consequently, a higher mortality rate in field experiments.[11]

Taken together, these considerations upend conventional thinking that the genome largely comprises junk DNA. In other words, the possibility is quite real that most of the human genome may be functional after all. If so, this would lead to the conclusion that our understanding of genetic programming in complex organisms has been fundamentally misunderstood over the past 50 years or so. The source of this misunderstanding may be attributed to the prevalent model that genetic information is expressed as, and transacted by, proteins. This clearly emanated from the early work on bacteria in which most genes, in fact, code for proteins. The ENCODE Project suggests that RNA-based networks would likely coordinate the developmental and homeostatic expression of the sum total of proteins in humans and, of course, other eukaryotic organisms. The notion that network(s) of RNA transcripts play a regulatory function is not entirely novel, though, and has been proposed previously.[12-3] The ENCODE Project data put this contention on a more solid footing.[14]

The ENCODE Project embodies the race to identify SNP variations in the DNA that are more frequent in patients with common, multifactorial diseases and, therefore, serve as markers for susceptibility. There appears to be a gold rush of sorts for new initiatives to compile databases to tackle various diseases. Hence, the National Human Genome Research Institute (NHGRI) announced the pilot phase of The Cancer Genome Atlas (TCGA) project to map the "cancer genome" with specific reference to human lung, brain, and ovarian cancers.[15] Likewise, a map of the structural variations in the human genome (defined as genomic changes, beyond the single base-pair substitutions in SNPs, involving chunks of DNA sequences leading to insertions, deletions, inversions, duplications, etc.) is being contemplated to better understand the genetic basis of disease.[16] Also, a self-styled connectivity map is in progress to deploy genetic signatures and connect the dots of gene(s) and disease with small-molecule therapeutics.[17] Well ahead of such initiatives, though, are the genome-wide associations (GWA) between gene variants and the seven more common diseases, as recently reported by the Wellcome Trust Case Control Consortium (WTCCC)[18] surveying diabetes, hypertension and, among other conditions, rheumatoid arthritis. The WTCCC study confirms the association of some genes for which disease associations have been previously established. It also identifies putative novel genes that may affect susceptibility to these diseases. While associations between biological traits and diseases have an impressive genealogy,[19] the WTCCC study was greatly facilitated by the information compiled in the HGP and HapMap databases to examine genetic variations at 500,000 different loci in the genomes of roughly 17,000 unrelated individuals. An important conclusion of the WTCCC study is that variations responsible for diseases afflicting broader swaths of populations are manifold. Some of these are in the exons, others are found in noncoding sequences, and yet others are found within the confines of "gene deserts," chromosomal regions entirely devoid of genes. What this unremittingly demonstrates is that the challenge to understand the biological function of genomic regions associated with disease risk would be daunting, indeed.

To assess the impact of GWA studies, the current predicate for identification of a gene for a specific disease must be revisited. The working model is derived from investigations of relatively rare diseases and has definitively afforded association with mutation(s) in a single gene. Because such mutations are predicted to disrupt the function of the encoded protein, the affected gene is consequently considered causative. The genetic architecture, so to speak, of multifactorial diseases (for example, diabetes or asthma) is unlikely to be based on simple devastating mutations. Such diseases arise from the combined risk induced by an unknown number of genetic variations, some of which may not code for a protein(s) and may be inherently difficult to identify. It is so that GWA studies strive to survey genetic variations, including noncoding regions.

Empirically, locations in the genome that are separated by a short DNA stretch (or a relatively small number of nucleotide pairs) are oftentimes in linkage disequilibrium (LD). In plain language, it means that there is a statistically robust association between variations of any two genetic loci under scrutiny, whether or not they are on the same chromosome.

Practically, this allows survey of variations throughout the genome rather precisely, simply by genotyping a subset of polymorphic loci. Thus, GWA studies rely on the presumption that LD enables one SNP to function as a marker for association to other sequences in the genome. It is based on this strategy that numerous reports on GWA for several diseases was published in 2007.[20-1]

Inasmuch as the prized pot at the end of this genomic rainbow began as a trickle a little over 2 years ago (starting in 2005), it portends to potentially become a tsunami of data and inundate researchers and clinicians alike. Primarily, however, the emergent complex pattern of associations begs the question of what exactly would this information bode in clinical management of diseases. Specifically, what does a disease risk of two to three times the general population in individuals with one gene copy mean? This consideration is pertinent, as most of the genes identified thus far appear to enhance the risk incrementally. Whereas a 50% increase in risk sounds alarming, it is quite modest in absolute terms, actually. For instance, if a woman with a 3% age-related risk of developing breast cancer also carries two copies of the most aggressive breast cancer variants in her DNA, she would be 1.6 times more likely to develop breast cancer. Her overall risk, however, will have increased to a mere 4.8%.

By the same token, an offshoot of genetic medicine is surmised to advise predisposed individuals to minimize their risk by taking preventive measures and modifying lifestyle. It is not entirely clear whether a layperson would be able to meaningfully utilize this information to lower their disease risk by lifestyle practices such as regular exercise or dietary changes. These intricacies are difficult to effectively communicate to patients, let alone those at risk or the "worried well" because statistical significance does not necessarily translate to clinical relevance. In contrast, however, a more common association might significantly affect disease prevalence and profile in a population at large. That is, even if the risk may not ring alarm bells for an individual, a variant's broader dispersion in the population may account for a far larger number of cases and may have profound repercussions for public health policy.

It is quite conceivable, however, that genome mining may stimulate research on physiological pathways not traditionally targeted for drug development. A poignant example of this is the concurrent susceptibility to both diabetes and heart disease. A recently discovered variant for heart diseases falls in the same region of chromosome 9 as a new diabetes variant. That diabetes and coronary artery disease more often than not are co-present should not come as a surprise to the practicing physician, as data corroborate this correlation. The question is what does fine-mapping of genetic variations contribute to the understanding of disease process(es). It is telling, therefore, that when the complete DNA sequence of James Watson and Craig Venter was in the news, even professional publications were at a loss to offer meaningful interpretations of the sequence data. One point was repeatedly noted that Venter's DNA showed a variant that predisposed him to heart disease. This was an odd point of emphasis, since his family history must have alerted Venter to that risk already. While a variation in Venter's DNA may represent an increased risk, it should be pointed out, that at least some of such associations, in the end, may have nothing to do with disease causation. Neither bioinformatics tools nor conceptual mechanisms are available to screen out spurious correlations, since their frequency is expected to increase with greater sequence heterogeneity in the genome.

All these attempts to map the human genome promise to geneticize medicine overtly, and the process to that effect is in full swing. The hope, of course, is to usher in personalized medicine. One of the concerns with this indefatigable faith in genomic medicine is that racially and ethnically organized groups may begin to hover around the powerful symbol of "iconic" genes, as is apparently the case with Tay-Sachs disease, sickle cell anemia, and cystic fibrosis.[22] To an extent, this is foreshadowed by the approval of BiDil by the U.S. Food and Drug Administration (FDA), an antihypertensive drug labeled for use exclusively in patients of African-American heritage. Consequently, genomic medicine — however imminent, or not, its impact on human health and disease ultimately may prove to be — must also address policy development issues. These should include how heavily should the weight of profit motive be in shaping the future of medicine and society, who should bear those risks, and what limitations might necessarily be placed on individu-

als to shape their genetic make-up.

Such issues are still on the horizon, as the age of personalized medicine still remains illusory as of yet, but their contours are taking palpable shape. Molecular hieroglyphics in the genome's Rosetta stone are gradually but certainly being deciphered. It remains to be seen whether its decipherment lends itself to a coherent text, a meaningful narrative pregnant with implications. It is not clear whether carvings in this euphemistic stone might merely provide general rules of thumb rather than specific and precise insights. Hence, expectations vested in this mammoth task should remain anchored in a realistic framework, especially because DNA sequencing is increasingly becoming cost-effective, and bioinformatics tools are being refined apace. On the other hand, the fact that common drug prescriptions do not take into consideration a patient's weight or comorbid conditions should give proponents of geneticized medicine a pause as to how many therapies of the future could realistically be based on encyclopedic genetic information alone. Additionally, how cytochrome P450 isozymes determine drug metabolism may be a controlling factor in personalized medicine as well.

With the intense activity in genome mining, it is easy to put in mind Mr. Woodhouse, that comical hypochondriac in Jane Austen's *Emma* who blames the multiplicity of his ailments on the rain, the cold, and the piece of wedding cake he devoured. It is not difficult to imagine his consternation on discovering that even his minor health challenges may well be due to his genetic luck of the draw. Nonetheless, the door to genomic medicine has creaked open ever so slightly to give a glimpse of what the clinical practice of the future may hold. It is certain, however, that translation of genetic susceptibility into sound medical practice would necessitate much larger populations in GWA than have been conscripted in studies heretofore.

References

1. Levy S, Sutton G, Ng P, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007;5(10)e254.
2. James Watson's Personal Geonome Sequence. http://jimwatsonsequence.cshl.edu/cgi-perl/gbrowse/jwsequence. Accessed June 28, 2008.
3. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
4. The International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):915-6.
5. Kent WJ, Sugent CW, Furey TS, et al. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437:1299.
6. http://www.genome.gov/10001688. Intenational HapMap Project. Accessed June 17, 2008.
7. http://www.genome.gov/10005107. The ENCODE Project: encyclopedia of DNA elements. Accessed June 17, 2008.
8. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447:799.
9. Mattick J, Makunin I. Non-coding RNA. Human Mol Genet. 2006;15:R17.
10. Mimouni N. NcRNAs: lost in translation. IBScientific J of Sci. 2007;2:27.
11. Lewejohann L, Skryabin B, Sachser N, et al. Role of a neuronal small non-messenger RNA; behavioural alterations in BC1-deleted mice. Behav Brain Res. 2004;154(1):273-89.
12. Cheng J, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 2005;308(5725): 1149-54.
13. Carnici P, Sandelin A, Lenhard B, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006;38:626.
14. Denoeud F, Kapranov P, Ucla C, et al. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res. 2007;17(6)746-59.
15. Collins F, Barker A. Mapping the cancer genome. Sci Am. 2007;296(3):50-7.
16. The Human Genome Structural Variation Working Group. Completing the map of human genetic variation. Nature. 2007;38:447:161.
17. Lamb J, Crawford E, Peck D, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929-35.
18. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. Nature. 2007;447(7145):661-78.

19. Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet. 2006;38(6):659-62.
20. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. 2007;445(7130):881-5.
21. Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglycerdide levels (Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institute of BioMedical Research). Science. 2007;316(5829):1331-6.
22. Wailoo K, Pemberton S. The troubled dream of genetic medicine: Ethnicity and innovation in Tay-Sachs, cystic fibrosis, and sickle cell disease, The Johns Hopkins University Press, Baltimore, Maryland: 2006.