Translating the ENCyclopedia Of DNA Elements Project findings to the clinic:
ENCODE's implications for eye disease.

Paul G Sanfilippo PhD,^{1,2} Alex W Hewitt PhD,^{1,2}

1. Centre for Eye Research Australia, University of Melbourne, Royal Victorian Eye and Ear Hospital, Melbourne, Australia.
2. Lions Eye Institute, University of Western Australia, Centre for Ophthalmology and Visual Science, Perth, Australia.

Address for correspondence:

Dr Alex Hewitt.
Centre for Eye Research Australia
Royal Victorian Eye and Ear Hospital
32 Gisborne Street,
East Melbourne, Victoria, Australia 3002
Telephone: +61 3 9929 8713
Fax: +61 3 9929 8711
Email: hewitt.alex@gmail.com

Declaration: There are no competing or conflicts of interest issues arising from this work.

Acknowledgments and Financial Support:

This work was supported by the Ophthalmic Research Institute of Australia.
AWH is supported by a NHMRC Early Career Fellowship and an Alcon Young Investigator grant.

ABSTRACT:

Approximately ten years after the Human Genome Project unraveled the sequence of our DNA, the ENCyclopedia Of DNA Elements (ENCODE) Project sought to interpret it. Data from the recently completed project has shed new light on the proportion of biologically active human DNA, assigning a biochemical role to much of the sequence previously considered to be 'junk'. Many of these newly catalogued functional elements represent epigenetic mechanisms involved in regulation of gene expression. Analogous to an Ishihara plate, a gene-coding region of DNA (target dots) only comes into context when the non-coding DNA (surrounding dots) is appreciated. In this review we provide an overview of the ENCODE project, discussing the significance of these data for ophthalmic research and eye disease. The major shortfall in the application of ENCODE data for ocular research is in the relative paucity of eye tissue used for analysis. Nonetheless, the novel insights afforded by the ENCODE project into the pathophysiology of disease, will in time allow for the development of new therapeutic strategies in the management of common blinding disorders.

KEY WORDS: epigenetic, methylation, macular degeneration, glaucoma, myopia.

INTRODUCTION

April 2012 commemorated the centennial anniversary of the sinking of the Titanic. During the one hundred years since the maritime tragedy, advances in the field of molecular biology have seen the discovery of the structure of DNA and sequencing of the human genome.¹⁻³ While the unveiling of our genetic blueprint was celebrated as a fundamental milestone in biology, it was acknowledged that it represents but the tip of the iceberg in our understanding of the molecular mechanisms of disease. In essence, publication of the human genome laid bare our biological building blocks, though biologists were yet to decipher the instruction manual for assembly.

To address this gap, the ENCyclopedia Of DNA Elements (ENCODE) project was established in 2003 to uncover new information about the genome beyond the DNA sequence.⁴ Comprising the collaborative efforts of an international consortium of over 400 scientists examining approximately 150 different cell types (including human cell lines from: conjunctival fibroblast; iris pigment epithelial, retinal pigment epithelia and retinoblastoma), the main goal of the project was to identify functional elements involved in gene regulation and expression.⁴ Thus, in addition to the ~1.5% of DNA, which is known to 'encode' the ~20,800 annotated proteins, the ENCODE project brought into sharp relief the lion's share of the human genome - the intractable intergenic deserts colloquially termed 'junk' and thought to have a limited biological role. ENCODE researchers surveyed the complete genomic landscape for a multitude of functional elements including protein-coding and non-coding RNA's, gene regulators (e.g. promoters, enhancers, silencers), and elements linked to 'epigenetic' mechanisms.⁴ Epigenetics can be defined as somatically heritable variation in gene expression that do not involve changes in nucleotide base sequence,⁵ and include such processes as DNA methylation, histone modification and nucleosome remodeling (Figure 1). Essentially, these are biochemical features of the genome that superimpose (epi, 'upon' or 'above') the DNA sequence, yet are still active in modulating gene expression.⁵

In the human genome, DNA methylation occurs primarily on cytosines in 'CpG' dinucleotides that are found at a low density but may also be in regions enriched for

CG dinucleotides, 'CpG islands,' which are present at ~60% of gene promoters.⁶ Isolated CG dinucleotides are frequently methylated, with dense methylation in promoter regions generally associated with transcriptional silencing.⁶ There are a variety of known histone modifications which directly influence gene expression. Histones can be considered as the spools for which DNA wraps around. Specific histone modifications can have dramatic differences in gene transcription. For example mono-methylation of the histone H3 lysine 4 (H3K4Me1) marks are often found near regulatory elements, whilst tri-methylation of the histone H3 lysine 4 (H3K4Me3) is frequently found near gene promoter regions.^{7,8} Acetylation at histone H3 lysine 27 (H3K27ac) are commonly identified near active regulatory elements.⁹

Better knowledge of the location, amount and functional role of epigenetic elements will enable a greater understanding of how the information contained within DNA is utilised in the differentiation and homeostasis of various cells and tissue. Data from the ENCODE project will also have important implications for the study and treatment of disease. The application of genome-wide association study (GWAS) design to the investigation of complex disease has unearthed a multitude of loci.¹⁰ Nonetheless, the precise pathogenic or functional variant at each statistically implicated site has generally been difficult to identify. Disease causing variants in Mendelian conditions have typically been easily characterized due to their clear disruption of protein coding DNA sequence.¹¹ Conversely in complex diseases, such as age-related macular degeneration (AMD) and glaucoma,¹²⁻¹⁵ single nucleotide polymorphisms (SNPs) that are unequivocally associated with a phenotype or disease have frequently been found to be intergenic, lying outside known protein coding domains. Data generated through the ENCODE project has helped to assign function to these poorly characterized regions.

MAJOR FINDINGS FROM ENCODE

In September 2012, initial results from the ENCODE project were released as a companion set of 30 papers published across three journals. The over-arching finding from the project that created much excitement (and some controversy) in the scientific community, was confirmation that the proportion of biologically active

human DNA had been generally under appreciated.^{4,16} In their overview paper, the consortium reported that at least 80% of the genome demonstrated biochemical function and was thus capable of being transcribed.⁴ While much of this could be assigned a regulatory role in moderating the expression levels of coding DNA, it nonetheless challenges the gene-centric paradigm that biology has been founded on. Interestingly, Greally used the metaphor of the Ishihara test in this context and results from the ENCODE project provide a cogent foundation for his argument.¹⁷ From an holistic viewpoint, a functional understanding of our genome (i.e. visualising the pattern on an Ishihara plate) requires an appreciation not only of the genes which compose it (target dots) but also the bulk of 'non-gene' DNA sequences (surrounding dots)(Figure 2).

At a fundamental level, the significance of the ENCODE data has engendered discussion amongst scientists as to whether the current concept of the gene as a minimum unit of heredity bears re-evaluation. Djebali and colleagues describe a new landscape of DNA transcription in human cells, with the discovery of thousands of novel RNAs, shrinkage of the intergenic space and overlapping of gene boundaries formerly assumed to be distinct loci.⁸ Accordingly, a compelling case can be made that RNA rather than DNA, and thus the transcript rather than the gene should be viewed as the 'atomic' component of inheritance.⁸ Under this model of genomic organisation, the term 'gene' would signify a higher-order element representing a collection of transcripts governed by regulatory elements to produce variation in a particular phenotypic trait. There is a primordial connotation associated with such a model given the long-standing debate surrounding the chicken-and-egg dilemma that symbolises the evolutionary roles of RNA versus DNA. It is interesting to note that the ratio of 'non-coding' to total genomic DNA scales with complexity of an organism.¹⁸ Certainly, what has become known as the RNA world hypothesis, remains best among alternative explanations for the early development of life.¹⁹ In this, it is postulated that the first stages of biological evolution involved RNA, with the molecules original role expanded beyond simply existing as a facilitator of DNA transcription.

The ENCODE project has produced a comprehensive catalogue of previously unannotated functional elements that act in concert to modulate spatial, temporal and cell-specific characteristics of gene expression. One of the approaches utilised to identify these regulatory regions included filtering their chromatin signatures,²⁰ the characteristic chromatin (protein-DNA complex that forms the cell nucleus) structures associated with various gene regulators and epigenetic histone modifications. Thurman and colleagues used a DNase 1 hypersensitivity assay to map active sites.²¹ Short segments of 'unraveled' chromatin are highly sensitive to cleavage by various nucleases such as DNase 1 or formaldehyde.²² They identified approximately 3 million such sites including almost all those previously validated experimentally, in addition to many novel sites, most of which exhibit highly cell-specific regulation. Neph *et al.* similarly used DNase 1 footprinting to scan the genome for recognition motifs, the short DNA sequences that facilitate protein binding.²³ Approximately 8.4 million distinct short sequence elements were identified, roughly doubling the number of known recognition sequences. Taken together, these new data will help to cultivate novel hypotheses regarding the functional variants at loci, genetically associated or implicated in a particular disease. For example, Maurano *et. al.* showed that numerous GWAS SNPs were co-located at DNase hypersensitivity sites that were associated with promoters of distant genes, thus expanding the range of plausible causal variants.²⁴ Consequently, future work will need to thoroughly interrogate the regulatory elements of a gene when assigning causative risk.

ENCODE and eye disease.

It is clear that despite many inroads being made into understanding the genetic mechanisms of many ocular diseases, we are only beginning to uncover similarly important epigenetic factors. A limited number of epigenetic factors have now been implicated in the pathogenesis of AMD and cataract. Though other diseases such as glaucoma and diabetic retinopathy have to-date been poorly studied on the epigenetic front.

Hunter *et al.* used an array-based DNA bisulfite sequencing method to investigate the DNA methylation of retina pigment epithelium and choroid between patients with AMD and age-matched controls.²⁵ Interestingly they found that *glutathione S-transferase isoform mu1 (GSTM1)* and *mu5 (GSTM5)* promoter methylation, together with mRNA levels, significantly differed between AMD versus normal retinas.²⁵ Recently, Wei and colleagues harnessed the power of studying monozygotic twins who were discordant for AMD to identify a distinct methylation signature for disease.²⁶ They revealed a significantly decreased level of methylation on the *interleukin 17 receptor C (IL17RC)* promoter in peripheral blood samples of AMD cases compared to controls, and then went on to show that hypomethylation mirrored an elevated gene expression in the macular of patients with choroidal neovascularisation or geographic atrophy.²⁶

The utility of incorporating the ENCODE data in disease discovery was showcased by Schaub and colleagues.²⁷ They reported a strongly supported functional SNP in linkage disequilibrium with an associated lead variant (rs380390) in the *complement factor H* gene, associated with AMD.²⁷ Interestingly, despite the number of other ocular GWAS, no other functional variants for ocular traits were highlighted.²⁷

The major shortfall in the application of ENCODE data for ocular research is in the relative paucity of eye tissue used for analysis (Figure 3). Given the high cost associated with analysis, the ENCODE investigators prioritized research into cell lines which were widely used, as such no in-depth expression or histone analysis were performed on ocular specific cell lines. With the evolution of biotechnology it certainly would seem timely for a reinvigoration of the NEIbank resource.²⁸ Nonetheless, despite the issues of tissue specificity of epigenetic marks, the ENCODE resource can provide general insight into global gene regulation and could still be used to explore loci of interest. For example Verhoeven *et al.* recently mined ENCODE's integrated regulation data in loci associated with myopia and found many associated loci contained these regulatory elements supporting a potential pathogenic mechanism.²⁹

As well as providing insight into pathogenic mechanisms, a greater understanding of epigenetics will allow for novel therapeutic avenues to be explored. Drugs altering methylation profiles are currently being used to treat specific diseases,³⁰ and over time the repertoire of functional epigenetic agents is likely to increase. Furthermore, genetic and epigenetic discovery allows for drug repositioning,³¹ and in time decoding the molecular mechanisms of blinding disease will uncover novel, therapeutic targets.

CONCLUSION:

With the publication of the most recent installment of findings from the ENCODE consortium, major improvements in our understandings of the “logic of biology” have been made. It is interesting to note that whilst the number of protein coding genes in the human genome had been greatly overestimated, the functional and epigenetic signatures of the genome, which are now being uncovered, were generally under appreciated.

The enormous wealth of data generated by the ENCODE project has gone a long way in answering the intriguing question of ‘what is the purpose?’ for the vast majority of the human genetic code, given that <2% is known to be translated into proteins.^{1,2} The discovery that much of the genome is biochemically active has fundamentally changed our perceptions of its organisation and complexity. Rather than a gene-centric view of the genome, it is generally accepted that a wider scope is required to fully appreciate the complex nature of our molecular milieu. The millions of new functional elements that the ENCODE project has identified are one of these dimensions; genetic and epigenetic mechanisms that function akin to components of a massive, complicated switchboard, managing a network of cellular events required for cellular function and tissue integration. By modulating the processes of replication, transcription and translation, these regulatory switches can turn specific genes on or off, thus affecting the timing and rate of protein production.

Undoubtedly, the most important implication of the ENCODE findings will lie in their translation to clinical relevance. Over the years, GWAS data have revealed a rich

trove of genetic variants associated with an increased susceptibility for a range of common blinding diseases. It has become evident that many loci play a role in gene expression, but until now knowledge of their biological underpinnings have generally remained elusive. The improved understanding in the biochemistry of gene regulation brought to bear on genomics research through the ENCODE project has helped identify regulatory sites in these loci and will continue to shed light on their pathophysiological basis for disease. The quest to discover and annotate disease-causing genetic variants should no longer result in molecular dead-ends, when those variants occur within previously poorly annotated regions. Nonetheless, additional epigenetic study of ocular specific tissue is required. In this way, the data generated through ENCODE has charted the way to new therapeutic strategies in the treatment of disease and a horizon free of genomic icebergs.

REFERENCES:

1. Venter JC, Adams MD, Myers EW *et al.* The sequence of the human genome. *Science* 2001; **291**: 1304-51.
2. Lander ES, Linton LM, Birren B *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860-921.
3. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 1953; **171**: 737-8.
4. Consortium EP, Dunham I, Kundaje A *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57-74.
5. American Association for Cancer Research Human Epigenome Task F, European Union NoESAB. Moving AHEAD with an international human epigenome project. *Nature* 2008; **454**: 711-5.
6. van Vliet J, Oates NA, Whitelaw E. Epigenetic mechanisms in the context of complex diseases. *Cell Mol Life Sci* 2007; **64**: 1531-8.
7. Benevolenskaya EV. Histone H3K4 demethylases are essential in development and differentiation. *Biochem Cell Biol* 2007; **85**: 435-43.
8. Djebali S, Davis CA, Merkel A *et al.* Landscape of transcription in human cells. *Nature* 2012; **489**: 101-8.
9. Creyghton MP, Cheng AW, Welstead GG *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* 2010; **107**: 21931-6.
10. Hindorff LA, Sethupathy P, Junkins HA *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009; **106**: 9362-7.
11. Stone EM. Finding and interpreting genetic variations that are important to ophthalmologists. *Trans Am Ophthalmol Soc* 2003; **101**: 437-84.
12. Consortium DG, Fritsche LG, Chen W *et al.* Seven new loci associated with age-related macular degeneration. *Nat Genet* 2013; **45**: 433-9.
13. Burdon KP, Macgregor S, Hewitt AW *et al.* Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. *Nat Genet* 2011; **43**: 574-8.

14. Thorleifsson G, Walters GB, Hewitt AW *et al.* Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma. *Nat Genet* 2010; **42**: 906-9.
15. Lu Y, Vitart V, Burdon KP *et al.* Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat Genet* 2013; **45**: 155-63.
16. Eddy SR. The ENCODE project: Missteps overshadowing a success. *Curr Biol* 2013; **23**: R259-61.
17. Greally JM. Genomics: Encyclopaedia of humble DNA. *Nature* 2007; **447**: 782-3.
18. Mattick JS. A new paradigm for developmental biology. *J Exp Biol* 2007; **210**: 1526-47.
19. Bernhardt HS. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)(a). *Biol Direct* 2012; **7**: 23.
20. Wang Z, Zang C, Rosenfeld JA *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 2008; **40**: 897-903.
21. Thurman RE, Rynes E, Humbert R *et al.* The accessible chromatin landscape of the human genome. *Nature* 2012; **489**: 75-82.
22. Song L, Zhang Z, Grasfeder LL *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* 2011; **21**: 1757-67.
23. Neph S, Vierstra J, Stergachis AB *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 2012; **489**: 83-90.
24. Maurano MT, Humbert R, Rynes E *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012; **337**: 1190-5.
25. Hunter A, Spechler PA, Cwanger A *et al.* DNA methylation is associated with altered gene expression in AMD. *Invest Ophthalmol Vis Sci* 2012; **53**: 2089-105.

26. Wei L, Liu B, Tuo J *et al.* Hypomethylation of the IL17RC promoter associates with age-related macular degeneration. *Cell Rep* 2012; **2**: 1151-8.
27. Schaub MA, Boyle AP, Kundaje A *et al.* Linking disease associations with regulatory information in the human genome. *Genome Res* 2012; **22**: 1748-59.
28. Wistow G, Peterson K, Gao J *et al.* NEIBank: genomics and bioinformatics resources for vision research. *Mol Vis* 2008; **14**: 1327-37.
29. Verhoeven VJ, Hysi PG, Wojciechowski R *et al.* Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet* 2013; **45**: 314-8.
30. Krecmerova M, Otmar M. 5-azacytosine compounds in medicinal chemistry: current stage and future perspectives. *Future Med Chem* 2012; **4**: 991-1005.
31. Hurle MR, Yang L, Xie Q *et al.* Computational drug repositioning: from data to therapeutics. *Clin Pharmacol Ther* 2013; **93**: 335-41.

Figure 1.

Zooming into the epigenetic landscape. Gene expression (as represented by RNA translation) is directly influenced not only by the DNA sequence, but also through DNA methylation, histone modifications and nucleosome remodeling.

Figure 2.

The role of non-coding DNA in defining cellular function. Analogous to an Ishihara plate, a gene-coding region of DNA (target dots) only comes into context when the non-coding DNA (surrounding dots) is appreciated.

Figure 3.

UCSC genome browser tracks of integrated regulation from the ENCODE project for the chromosome 10q26 region, which has been strongly associated with age-related macular degeneration. Panel A displays all available ENCODE data, whilst Panel B displays only information obtained from ocular-related cell lines (retinal pigment epithelia – HRPEpiC; conjunctival fibroblast - HconF; iris pigment epithelial – HIPEpiC; retinoblastoma - WERI-Rb-1). Note the lack of tissue specific information relating to histone modifications (H3K27Ac; H3K4Me1; H3K4Me3) and transcription.

A

B