

The ENCODE Project: missteps overshadowing a success

Sean R. Eddy

Two clichés of science journalism have now played out around the ENCODE (Encyclopedia of DNA Elements) project. ENCODE’s publicity first presented a misleading “all the textbooks are wrong” narrative about noncoding human DNA. Now several critiques of ENCODE’s narrative have been published, and one was so vitriolic that it fueled “undignified academic squabble” stories that focused on tone more than substance. Neither story line does justice to our actual understanding of genomes, to ENCODE’s results, or to the role of big science in biology.

ENCODE aimed to catalog all functional elements in the human genome, where “all” and “functional element” should both be in scare quotes. In genomics, all is a term of art, meaning reasonably comprehensive. Even complete genome sequences only asymptotically approach actual completeness, despite being well defined goals of known and finite size. ENCODE’s goal was much more nebulous, because there are three senses in which “functional element” is ill defined and had to be operationalized.

First, ENCODE had to choose a specific set of genome-wide biochemical assays as proxies for biological function. ENCODE assayed the locations of RNA transcription, of several DNA-binding proteins, of several specific histone modifications, and of DNase I hypersensitivity sites.

Second, different functional elements are active in different conditions: different cell types, developmental stages, environmental conditions, *ad infinitum*. ENCODE had to limit the scope of different cell types and conditions it examined. Its assays were run primarily on a small number of cultured cell lines: three cell types in particular (two immortal cancer cell lines and an embryonic stem cell line), and to lesser extent on 144 other cell types.

Third, biological phenomena are generally a continuum – for example, a DNA binding protein binds all possible DNA sequences, albeit with different affinities. The continuous quantitative terrain of each ENCODE assay’s results was converted into discrete “functional elements” by choosing statistical thresholds based on reproducibility across replicates.

The results, published simultaneously in 30 papers [1], consist of a matrix of assays versus

cell types, totalling 1640 datasets. Each dataset can be laid as a track of discrete boxes along the human genome, showing the genomic locations of all the reproducible biochemical activities that have been operationally defined as discrete functional elements.

ENCODE's feature maps illustrate the human genome. They are unquestionably useful, despite the incomplete and ill defined nature of having to operationalize something as complex as biological function. Many labs will use ENCODE datasets without needing (or being able) to generate such systematic high quality data themselves, a *sine qua non* of useful big science investment in biology.

All the textbooks are wrong?

The ENCODE summary paper [2] added up how much of the genome was covered by the mapped functional elements, arriving at a figure of 80.4% (dominated by RNA transcription, which alone covered 62%). Couched in precise language about the proxy that was being counted – regions of reproducible biochemical activity – reporting these coverage statistics was innocuous, even *pro forma* for a genomics project. The summary paper also correctly noted that with respect to two of the ways functional element had been operationalized – because only a few cell types were studied, and because thresholds had to be drawn – the coverage numbers were conservative underestimates.

In attempting to popularize the result and the project, in press releases and interviews, ENCODE leaders fit the results to a seductive “all the textbooks are wrong” narrative. The noncoding 99% of the human genome was described as an enigmatic sea of dark matter that had been dismissed as junk out of ignorance and arrogance. ENCODE claimed to have revolutionized our view of the genome by showing that more than 80% of the genome was functional, thus disproving junk DNA. The New York Times' coverage is representative of the received story:

The human genome is packed with at least four million gene switches that reside in bits of DNA that once were dismissed as “junk” but that turn out to play critical roles in controlling how cells, organs and other tissues behave.... At least 80 percent of this DNA is active and needed. [3]

Thus all reproducible biochemical events were claimed to be “critical” and “needed”. But

ENCODE had not shown what fraction of these activities play any substantive role in human gene regulation, nor was the project designed to show that. There are other well-studied explanations for reproducible biochemical activities besides crucial human gene regulation, including residual activities (pseudogenes), functions in the molecular creatures that infest eukaryotic genomes (transposons, viruses, and other mobile elements), and noise. Far from disproving junk DNA, ENCODE's operationalized definition of function *included* junk DNA.

An undignified academic squabble?

Several critiques making these points (and more) have appeared [4-7]. Of these, Ford Doolittle's piece [4] is perhaps the deepest examination of where ENCODE's narrative failed. Besides regulatory DNA (promoters, enhancers, and such), a large fraction of mammalian genomes (at least 50% in humans) is composed of mobile DNA elements, most of which are defective and decaying. Moreover, eukaryotic genome sizes are too variable and too large to be accounted for by natural selection for host functions. Thus genomes contain a lot of DNA that's not critically important for host functions, much of which arises from mobile element replication: a concept irreverently called junk DNA. The fact that mobile elements autonomously replicate and spread explains why a host genome has difficulty getting rid of this DNA.

ENCODE's reproducible biochemical activities do not bear on the concept of junk DNA because they not distinguish between activities important to human biology, versus activities expected to be found in junk DNA, including host suppression of mobile element replication, and mobile elements regulating and expressing themselves as they try to replicate and evade host surveillance. ENCODE's data illuminate the battlefield, but ENCODE should have acknowledged the known combatants.

One critique, by Graur and colleagues [6] – angry, dogmatic, scattershot, sometimes inaccurate – garnered widespread media coverage because its tone played into the “undignified academic squabble” cliché of science journalism. Attention focused on the squabbling more than the substance, and probably led some to wonder whether the arguments were just quibbling over the semantics of the word function.

In trying to conceptualize the forces that act on genome evolution, it's not just semantics. We can envision the human genome as a perfectly honed machine, or we can think of it as a wild landscape littered and layered with successions of decomposing molecular replicators, like dead

weeds decaying into fertile soil. How much DNA does it take to *design* a human? How much DNA does it take to *evolve* a human? They aren't the same question, and the gap between them is where we seek an understanding of genome evolution.

The Random Genome Project: the missing negative control.

The Doolittle and Graur critiques touch on a different sense in which ENCODE's interpretation needs to be questioned. It's one thing to distinguish human gene regulatory functions from other biological functions, such as mobile DNA element functions. Another question is whether we ought to accept the oft unstated assumption, prevalent in genomics at present, that any specific and reproducible biochemical event *must* correspond to a meaningful biological function. In this view, all biochemically active sequences are nonrandom and honed by selection, and natural selection does not tolerate any wasted activities.

Another view is that biology is noisy. Population genetic theory says natural selection has limited power [8] so we should expect that every biochemical machine is only just good enough for its job, tolerating some level of background biochemical activity [9]. Those in the first camp have typically argued against noise by arguing that biochemical events are not just a nonspecific background haze over the genome, but instead occur at highly specific, reproducible locations in specific cell types. An example of this argument is that a cell type-specific expression pattern is often taken to be sufficient evidence for functionality of a noncoding RNA transcript. But this misunderstands what we should expect noise to look like in a genome.

To clarify what noise means, I propose the Random Genome Project. Suppose we put a few million bases of entirely random synthetic DNA into a human cell, and do an ENCODE project on it. Will it be reproducibly transcribed into mRNA-like transcripts, reproducibly bound by DNA-binding proteins, and reproducibly wrapped around histones marked by specific chromatin modifications? I think yes.

A striking feature of genetic regulation is that regulatory factors (proteins or RNAs) generally recognize and bind to small sites, small enough that any given factor will find specific binding sites even in random DNA. Promoters, enhancers, splice sites, poly-A addition sites, and other functional features in the genome all have substantial random occurrence frequencies. These sites are not nonspecific in a random genome. They are specific sequences, albeit randomly occurring and not under selection for any function.

Would biochemical activities in the random genome be regulated under different conditions? For example, would they be cell type-specific? Surely yes, because the regulatory factors themselves (such as transcription factors) are regulated and expressed in specific cell types and conditions.

Suppose we identified a cell type-specific transcription unit in the random genome, and we knock it down or out. Would we see a measurable molecular phenotype (for example, by measuring genome-wide gene expression from the rest of the genome)? I think yes. Any biochemical activity perturbs the system to some degree, if only by soaking up regulatory factors and changing their free concentrations. When does a measurable phenotype stop being a mere perturbation and start becoming evidence of an important function? I am not convinced that there's a qualitative distinction.

Even as a thought experiment, the Random Genome Project states a null hypothesis that has been largely absent from these discussions in genomics. It emphasizes that it is reasonable to expect reproducible biochemical activities – even measurable knockout phenotypes – in random unselected DNA. Could the Random Genome Project be done for real? Certainly we could use genome synthesis methods to synthesize a specific random sequence, and I can think of cheaper means since we only need *any* random sequence. To a small extent the experiment has already been done, where people have already measured specific biochemical activity (such as enhancers that drive specific expression patterns) in exogenous or random DNA, either in their negative controls or in *in vitro* selection experiments in random sequence libraries. It would be interesting to collate such examples. It would also be interesting to study some natural examples of large exogenous DNA insertions, such as organellar genomes that integrate into nuclear DNA, or bacterial genome integrations into eukaryotic genomes (such as the *Wolbachia* genome that integrated into the *Drosophila ananassae* genome) although there is some danger in studying sequences that may have co-evolved with the eukaryotic host's biochemical machinery.

Why big science fails even as it succeeds.

Clearly the arguments over ENCODE are also fueled by unease over big science and small science approaches in biology. It is bewildering to see ENCODE take an eminently reasonable data generation project and spin it so inexpertly as a hypothesis test with supposedly revolutionary conclusions. Does it suggest that big science is out of control? I suggest the deeper

problem is that we aren't sufficiently comfortable with defending what big science in biology is *for*.

There are three categories of big science: the big experiment, the map, and the leading wedge.

A big experiment is driven by a single question or hypothesis test, but requires a large scale community investment. Like any experiment, it would generally include an experimental design, positive and negative controls, and validation experiments that test conclusions from multiple angles. A failure mode in a big experiment is the difficulty in planning experiments properly in a committee process.

A map is a data resource – comprehensive, complete, closed ended – to be used by multiple groups, over a long time, for multiple purposes. The decision to build a map is a cost/benefit calculation, weighed against individual labs who are already making piecemeal maps in an ill coordinated fashion, especially when small groups lack technical expertise to make the map well. A failure mode with a map is to miscalculate the cost/benefit analysis and make a map that too few individual labs will use.

A leading wedge is a massed technology development effort, in an area where we need radically better methods. There is a driving goal that is visionary, but perhaps arbitrary. The actual long term goal is to develop and democratize a breakthrough technology, making it cheap and effective for routine use in individual labs, for individual well designed experiments. A failure mode in a leading wedge is the inertia of the big science phase, failing to transition as quickly as it should to democratization to small labs.

We have been too shy to defend maps and leading wedges in biology. Maps and leading wedges are about enabling small science; they are not the science itself. Coordinated investment in infrastructure is somewhat new to our culture of hypothesis driven, question driven science. We feel a strong temptation to spin all big science projects in biology as big experiments, whereas the complexity of biology – its lack of unifying theory, its wealth of fascinating and crucial detail – is such that big experiments are nearly nonexistent in our field. Our important insights are born and tested in small labs.

For example, the Human Genome Project (HGP) was sometimes spun as a revolution in *understanding* the human genome, and this was wrong. But the project's critics sometimes fell into a trap themselves, of not recognizing the value of the HGP as both a map and a leading

wedge. The value of the human genome sequence as a resource is indisputable. A massive revolution in sequencing technology also resulted, and that technology is now in the hands of individual labs.

ENCODE and some of its critics have fallen into similar traps. In trying to make the result sound important, ENCODE's publicity spun it retrospectively as a hypothesis test, but ENCODE was not designed to test anything. ENCODE is a map. It should have been published and defended as such. And while its critics argue over an interpretation that wasn't in ENCODE's mission to begin with, ENCODE's planners should also recognize that as ENCODE now moves into a new funding phase, it may be headed for a failure mode in its actual mission. The cost/benefit calculation is rapidly changing. ENCODE's technologies (all based on high throughput sequencing) are now widely and inexpensively available in individual labs.

The next big thing.

The United States government is now considering the Brain Activity Map (BAM) project, a big science project with the aim of recording every neuron in a brain. In my taxonomy, the Brain Activity Map is *not* a map, nor an experiment; it is all leading wedge. Recording from every neuron in a large brain is an arbitrary aspirational goal, like sending a person to the Moon. Neural activity is entirely open ended, different in every question we ask and every condition we examine. But neuroscience is in part technology limited, asking questions about neural circuits and neural systems using current technologies that record from relatively few neurons at once. Reaching BAM's goal of recording every neuron in a brain will do little in itself to advance our understanding of brains, and it will not produce much in the way of a stable and reusable data resource, but it would be a landmark achievement in enabling individual neuroscientists to record thousands to millions of neurons simultaneously, in model neural systems, for a multitude of individual experiments. Recognizing that BAM is all leading wedge, not a map and not an experiment, is crucial to its planning, its support, and its success.

References

1. ENCODE Project Consortium. (2012) www.encodeproject.org.
2. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

3. Kolata, G. (2012) Bits of mystery DNA, far from 'junk', play crucial role. New York Times, 5 September 2012.
4. Doolittle, W.F. (2013) Is junk DNA bunk? A critique of ENCODE. Proc. Natl. Acad. Sci. USA, in press.
5. Eddy, S.R. (2012) The C-value paradox, junk DNA and ENCODE. Curr. Biol. 22, R898–899.
6. Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., and Elhaik, E. (2013) On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. Genome Biol. Evol., in press; doi:10.1093/gbe/evt028.
7. Niu, D. K., and Jiang, L. (2013) Can ENCODE tell us how much junk DNA we carry in our genome? Biochem. Biophys. Res. Commun. 430, 1340–1343.
8. Lynch, M. (2007) *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.
9. Struhl, K. (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat. Struct. Mol. Biol. 14, 103–105.

HHMI Janelia Farm Research Campus,

Ashburn VA 20147, USA.

E-mail: eddys@janelia.hhmi.org