

The C-value paradox, junk DNA, and ENCODE

Sean R. Eddy

What's the C-value paradox?

You might expect more complex organisms to have progressively larger genomes, but eukaryotic genome size fails to correlate well with apparent complexity, and instead varies wildly over more than a 100,000-fold range. Single-celled amoebae have some of the largest genomes, up to 100-fold larger than the human genome. This variation suggested that genomes can contain a substantial fraction of DNA other than for genes and their regulatory sequences. Thomas dubbed it the “C-value paradox” in 1971.

The C-value paradox is related to another puzzling observation, called “mutational load”: the human genome seems too large, given the observed human mutation rate. If the entire human genome were functional (in the sense of being under selective pressure), we would have too many deleterious mutations per generation. By 1970, rough calculations had suggested to several authors that maybe only 1-20% of the human genome could be genic, with the rest evolving neutrally or nearly so.

So why not call it the “genome size paradox”? What's a “C-value”?

“C-value” means the “constant” (or “characteristic”) value of haploid DNA content per nucleus, typically measured in picograms (1 picogram is roughly 1 gigabase). Around 1950, the observation that different cell types in the same organism generally have the same C-value was part of the support for thinking that DNA was responsible for heredity.

I don't see why it's a paradox. Maybe you just don't understand how to measure complexity. More advanced machines can have fewer parts. Or maybe amoebae are more complex than humans.

For sure, we don't understand how to meaningfully measure an organism's complexity, and we don't have any theoretical basis for predicting how many genes or regulatory regions one needs. But the C-value paradox isn't just an observation that different species have different genome sizes, it's the observation that even *similar* species can have quite different genome sizes. For example, there are many examples of related species in the same genus that have

haploid genome sizes that differ by 3- to 8-fold, especially in plants, as seen in species of rice (*Oryza*), *Sorghum*, or onions (*Allium*). The maize (*Zea mays*) genome expanded by about 50% in just 140,000 years since its divergence from *Zea luxurians* (and not merely by polyploidization). Unlike what we expect of genes and regulatory sequences, which generally evolve slowly and conservatively, for some reason genome size can change rapidly on evolutionary timescales.

OK, cool. I've already come up with some hypotheses. Maybe the extra DNA has a structural role in the nucleus...

Remember, the C-value paradox is old. Many hypotheses have been proposed and carefully weighed in the literature. At first, people looked for explanations in terms of some functional significance of the extra DNA – an adaptive function that would maintain nongenic, nonregulatory DNA by natural selection. But to explain mutational load (and more modern observations of comparative genomics, that only a small fraction of most eukaryotic genome sequence is conserved and under selective pressure), you have to posit an adaptive role where only the bulk amount of the DNA matters, not its specific sequence. To explain the C-value paradox, you have to explain why this bulk amount would vary quite a bit even between similar species. Although some such adaptive explanations have been speculated, a rather different line of thinking, starting with Ohno and others in the early 1970's, ultimately led to a reasonably well-accepted explanation of the C-value paradox.

What's the explanation for the C-value paradox?

Genomes carry some fraction of DNA that has little or no adaptive advantage for the organism at all. Some genomes carry more than others, and some genomes carry quite a lot of it. Ohno, who believed that strongly polarizing statements clarify scientific debate, called this “junk DNA”.

So the idea is that all noncoding DNA is junk DNA?

No. Of course we've also known since the earliest days of molecular biology (including the Jacob/Monod lac operon paradigm) that genes are regulated by sequences that often occur in noncoding DNA. Rather the idea is, there's a fraction of DNA that is useful and functional for the organism (genes and regulatory regions) which does more or less scale with organismal

complexity, and a “junk” fraction which varies widely in amount, creating the C-value paradox.

I’m having a hard time with your derogatory term “junk”.

Ohno’s zest for polarizing provocation went too far. Far from clarifying, his term tends to incense people, and the science behind the idea gets muddled. If you like, call it “nonfunctional” DNA instead – and by nonfunctional, we mean “having little or no selective advantage for the organism”. These words, especially “for the organism”, will become important.

How much nonfunctional DNA an organism would harbor will be a tradeoff between how deleterious it is to carry versus how easy it is to get rid of. It’s actually not obvious that extra DNA would be all that deleterious; DNA replication is a relatively small part of the energy budget of most organisms. Still, DNA deletions are common enough mutations. If there were even a small selective disadvantage to having a junky genome, especially in species with large population sizes (where small selection coefficients have more effect) and fast growth rates (where an obese genome might especially be a hindrance), it would be surprising to see a lot of nonfunctional DNA.

That’s what I mean. Natural selection wouldn’t tolerate junk. If you can’t explain how this extra DNA got there and why it’s maintained, “junk DNA” is an argument from ignorance. You can’t just assume it’s junk.

Ohno was mostly focused on pseudogenes, which do occur, but not nearly in large enough numbers to explain the C-value paradox. So indeed, what Ohno’s idea lacked to make it convincing was an observable mechanism that creates large amounts of junk DNA rapidly, faster than natural selection deletes it. In 1980, two landmark papers, by Orgel and Crick and by Doolittle and Sapienza, established a strong case for such a mechanism. They proposed that “selfish DNA” elements, such as transposons, essentially act as molecular parasites, replicating and increasing their numbers at the (usually slight) expense of a host genome. Selfish DNA elements function for themselves, rather than having an adaptive function for their host.

The massive prevalence of transposable elements in eukaryotic genomes was only just becoming appreciated at the time. One transposable element in humans, called Alu, occurs in about a million copies and accounts for about 10% of our genome. Almost all copies of transposons in genomes are partial or defective elements that were inserted in the evolutionary

past and are now decaying away, largely by neutral mutational drift. Active DNA transposons (one kind of “selfish DNA”) generate a mass of decaying dead transposons (one source of “junk DNA”).

We can affirmatively identify transposon relics by computational genome sequence analysis methods. These studies show that transposable elements invade in waves over evolutionary time, sweeping into a genome in large numbers, then dying and decaying away. 45% of the human genome is detectably derived from transposable elements. The true fraction of transposon-derived DNA in our genome must be greater, because neutrally evolving sequences decay so rapidly that after only a hundred million years or so, they eventually become too degraded to recognize. The C-value paradox is mostly (though not entirely) explained by different loads of decaying husks of transposable elements. Larger genomes have a larger fraction of transposon relics.

But transposons are functional. Their DNA is biochemically active and they can encode proteins. Organisms have layers of DNA regulatory systems for suppressing transposon activity. There are many examples of organisms co-opting (“domesticating” or “exapting”) transposon functions. Transposons are interesting!

All true. To me, “junk DNA” is a colloquial term of endearment, and a reminder of the history of ideas in the field. You can forget the polarizing term so long as you remember the data it stands for: astonishing genome size variation, mutational load, a small fraction of conserved DNA, and the large fraction of eukaryotic genomes that is composed of neutrally decaying transposon relics. These data support a view that eukaryotic genomes contain a substantial fraction of DNA that serves little useful purpose for the organism, much of which has originated from the replication of transposable (selfish) elements.

Evolution is sure to repurpose some fraction of this vast quantity of DNA in interesting ways, sometimes reshaping something to play a new organismal role. (As Sydney Brenner put it, garbage is stuff you throw out, but junk is the interesting stuff you keep around that might be useful someday.) Mutational load arguments and comparative genomics suggest that co-option is the exception, not the rule. Doolittle and Sapienza made an important point: “[w]hen a given DNA... can be shown to have evolved a strategy (such as transposition) which insures its genomic survival, then no other explanation for its existence is necessary.”

If the C-value paradox was more or less resolved long ago, why bring it up again?

Recently, the ENCODE project has concluded that 80% of the human genome is reproducibly transcribed, bound to proteins, or has its chromatin specifically modified. In widespread publicity around the project, some ENCODE leaders claim that this biochemical activity disproves junk DNA. If there is an alternative hypothesis, it must provide an alternative explanation for the data: for the C-value paradox, for mutational load, and for how a large fraction of eukaryotic genomes is composed of neutrally drifting transposon-derived sequence. ENCODE hasn't done this, and most of ENCODE's data don't bear directly on the question. Transposon-derived sequence is generally expected to be biochemically active by ENCODE's definitions – lots of transposon sequences are inserted into transcribed genic regions, mobile transposons are transcribed and regulated, and genomic suppression of transposon activity requires DNA-binding and chromatin modification.

The question that the “junk DNA” concept addresses is not whether these sequences are biochemically “active”, but whether they're there primarily because they're useful for the organism. Sequence conservation analyses, including ENCODE's, consistently indicate that only around 5-20% of the human genome is under detectable selective pressure. Some additional fraction of sequences has probably evolved new human-specific regulatory functions that are not conserved with other closely related species, but ENCODE's publicized interpretation would require that such nonconserved regulatory sequences account for 80-95% of the genome, far outnumbering evolutionary conserved regulatory sequences. Given the C-value paradox, mutational load, and the massive impact of transposons, the data remain consistent with the view that the nonconserved 80-95% of the human genome is mostly composed of nonfunctional decaying transposons: “junk”.

Where can I find out more?

Cavalier-Smith, T., ed. (1985) *The Evolution of Genome Size*. (John Wiley & Sons.)

Doolittle, W.F. and Sapienza C. (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601-603.

Gregory, T.R., ed. (2005) *The Evolution of the Genome*. (Elsevier Academic Press.)

Kidwell, M.G. (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115, 49–63.

Lynch, M. (2007) *The Origins of Genome Architecture*. (Sinauer Associates.)

Ohno, S. (1972) So much “junk” DNA in our genome. *Brookhaven Symp. Biol.* 23, 366–370.

Orgel, L.E. and Crick, F.H.C. (1980) Selfish DNA: the ultimate parasite. *Nature* 284, 604–607.

Thomas, C.A. Jr. (1971) The genetic organization of chromosomes. *Annu. Rev. Genet.* 5, 237-256.

HHMI Janelia Farm Research Campus, Ashburn VA 20147, USA. E-mail:
eddys@janelia.hhmi.org