

The g-value paradox

Matthew W. Hahn* and Gregory A. Wray

Department of Biology, Duke University, Durham, North Carolina 27708, USA

*Correspondence (email: mwh3@duke.edu)

INTRODUCTION

The February 15th, 2001 announcement of the draft human genome sequence was the culmination of a momentous undertaking. The analyses of this sequence (International Human Genome Sequencing Consortium 2001; Venter et al. 2001) predicted a surprisingly modest 31,000 genes for *Homo sapiens* (although this number has yet to be finalized), as compared with estimates as high as 140,000 genes just a few years ago (Fields et al. 1994). Given this, we narrowly top the list for the eukaryotic genomes that have been completely sequenced (Table 1). Even though sequencing the human genome may be merely a first pass at a deeper understanding of our biology, one fact stands out as demanding an immediate explanation: Why do humans have so few genes?

The assumptions and chauvinism implicit in this question—that humans are vastly more complex than the other fully sequenced eukaryotes and should therefore have a commensurately larger suite of genes—are difficult to argue clearly and may be even more difficult to justify biologically (McShea 1996). Still, it is hard to deny our intuitive perception that the number of genes in a genome should be roughly correlated with complexity and that organismal complexity can be ranked as yeast < nematodes < flies < humans (we reserve judgment on the relative position of the “green fly,” *Arabidopsis*). However, the number of genes in the genomes of these organisms does not match our naive expectation.

This disjunction between the number of genes and organismal complexity, what we call the “G-value paradox,” parallels the finding during the 1950s that the physical size of genomes does not correlate with organismal complexity, a relationship known as the C-value paradox (Cavalier-Smith 1985; Appendix and Table 1). The finding that much of the genome contains noncoding repeats and “junk” DNA seemed to resolve the C-value paradox. Implicitly, this resolution rested on the assumption that once noncoding DNA was taken into account, the total number of genes would then correlate with organismal complexity (Cavalier-Smith 1985). However, the published G values of the completely sequenced eukaryotes make it clear that we have not yet resolved the C-value paradox—it has merely given way to the G-value paradox.

Just as the discovery of noncoding DNA seemed to resolve the C-value paradox, so a few simple observations may in time resolve the G-value paradox. These observations all attempt to give more value to each of our genes and thus to give us a more accurate genomic predictor of organismal complexity by identifying the true measure of information encoded by a genome, the “I-value” (Appendix; the concept of “information” with respect to genes is itself highly debated, see Oyama 1985 and Sarkar 1998, but a philosophical resolution to this issue is beyond the scope of this article). Some of the observations we discuss here have been offered as *the* answer to explaining our modest number of genes (Davidson 2001; Hanke et al. 1999; Szathmary et al. 2001), whereas some have been invoked in combination (International Human Genome Sequencing Consortium 2001; Petsko 2001). These observations indicate that the evolution of organismal complexity will typically involve changes in the genome that are subtler than simply adding genes. The C-value paradox was resolved by a plea to the G value; a resolution of the G-value paradox may be offered by a plea to the I value.

However, what if no measure of genomic information content, no matter how precise, correctly predicts organismal complexity? Our last observation will attempt to undermine the basic assumption that organismal complexity somehow corresponds to even a refined measure of genomic complexity.

cis REGULATION: DIVERSIFYING FUNCTION THROUGH TRANSCRIPTIONAL CONTROL

Many genes are expressed at several times and places during development. In principle, increasing the complexity of a protein’s expression profile allows it to carry out a greater diversity of functions (Davidson 2001). Approximately 50% of the human genome is made up of repeated elements, and only 5% of the remainder is actually transcribed (International Human Genome Sequencing Consortium 2001); this leaves an enormous fraction of the genome with the potential to contain cis-regulatory elements. Although not all of the noncoding regions may be used or needed for cis regulation, it does not appear that many eukaryotes are space limited (Table 1).

Table 1. Comparisons of C value to G value

Species	C value*	G value†
<i>Saccharomyces cerevisiae</i>	12	6000
<i>Drosophila melanogaster</i>	120	14000
<i>Caenorhabditis elegans</i>	97	19000
<i>Arabidopsis thaliana</i>	125	26000
<i>Homo sapiens</i>	2900	31000

*In millions of base pairs.

†Number of estimated genes.

Sources: Goffeau et al. 1996, Adams et al. 2000, The *C. elegans* Sequencing Consortium 1998, The Arabidopsis Genome Initiative 2000, International Human Genome Consortium 2001, Venter et al. 2001.

INTERGENIC COMBINATORICS: AN ACCELERATING RATE OF RETURN

As the number of genes in an organism increases, the combinations of encoded protein that can operate together to perform complex functions increase much faster. This is true of both metabolic (Fell and Wagner 2000; Jeong et al. 2000) and signaling (Szathmáry et al. 2001) protein networks. If each gene in the *C. elegans* genome produced just one protein, 162 million pairwise combinations of these proteins would be possible; in comparison, 480 million combinations are possible with the complement of human genes (Sokal and Rohlf 1995). Adding just 100 genes to our genome opens up 3.1 million additional pairwise combinations. For a rather modest increase in G value, an organism can in principle gain a disproportionate amount of complexity.

MORE FUNCTIONS PER GENE: “SWISS-ARMY KNIFE” PROTEINS

The proportion of multifunctional proteins encoded in our genome appears to be higher than that in flies and nematodes (International Human Genome Sequencing Consortium 2001); that is, each of our proteins has, on average, a greater number of distinct biochemical functions than do those of either of *D. melanogaster* or *C. elegans*. This has been described as the “Swiss-army knife” explanation for our smaller than expected G value.

ALTERNATIVE SPLICING: FROM GENOME TO TRANSCRIPTOME

By the best estimates currently available, 59% of our genes are alternatively spliced during transcription (Hanke et al. 1999; International Human Genome Sequencing Consortium 2001). If one only considers splice variants that affect protein-coding regions, one obtains an estimate of approximately 69,000 distinct protein sequences encoded by our genome. This is a greater than 300% increase on the number of genes. In com-

parison, the nematode genome contains a smaller proportion of alternatively spliced genes, producing at most 25,000 proteins (International Human Genome Sequencing Consortium 2001; The *C. elegans* Sequencing Consortium 1998). This represents only a 33% increase over a simple count of genes.

POSTTRANSLATIONAL MODIFICATION: FROM TRANSCRIPTOME TO PROTEOME

Following translation, a variety of modifications can further increase the number of functionally distinct proteins encoded by a single gene. Common modifications include glycosylation, proteolytic cleavage, and phosphorylation (Alberts et al. 1989). A comparison of the human proteome (the complement of proteins in a cell) to the transcriptome (the complement of transcripts in a cell) will give us an estimate for how prevalent this mechanism is in our genome.

GENE REDUNDANCY: G-VALUE INFLATION

Sometimes the G value will overestimate the information encoded in our genome. Duplications of single genes, whole chromosomes, or even whole genomes are common evolutionary events (Cavalier-Smith 1985). Fully 40% of the loci in the *C. elegans* genome are the result of tandem duplications (The *C. elegans* Sequencing Consortium 1998), which may account for much of its unexpectedly large G value compared with *Drosophila*. Single knockouts of duplicated genes in mice often have no effect (Hurst and Smith 1999), suggesting a significant degree of informational redundancy among duplicated loci within mammalian genomes.

THERE IS NO G-VALUE PARADOX: RUBE GOLDBERG GENOMES

Even given the six points just discussed, genomic complexity may not always be correlated with organismal complexity. Like a Rube Goldberg Contraption, our genome may be overly complex: There may be a simpler way to encode our bodies and behaviors than actually exists in our genome. Indeed, the G-value paradox (and the C-value paradox) is only a paradox under the assumption that organismal complexity somehow corresponds to the number of components in a genome (even a more refined number than a count of genes). Complexity theory addresses the emergent properties of a set of interactors and emphasizes the lack of a simple relationship between the complexity of instructions (genes) and the complexity of the product (organism) (Kauffman 1993). For example, biosynthesis of the amino acid proline requires three different enzymes in enterobacteria, two en-

zymes in fungi and higher plants, but only a single, unrelated enzyme in the gram-positive bacteria, *Clostridia* (Herrmann and Somerville 1983). In fact, variation in the number of genes necessary to synthesize a given amino acid is common across the tree of life (Herrmann and Somerville 1983). This is not functional redundancy but simply different recipes for the same product. The historical necessity of working with existing materials also means that the number of biological instructions often does not correlate with the complexity of biological products. During the development of *C. elegans*, 131 cells are produced at various times and places that subsequently undergo programmed cell death (this is approximately one eighth of the total cells produced) (Sulston and Horvitz 1977). Like the human appendix, these cells may have evolved in ancestors in whom they had a function.

CONCLUSIONS

Our previous discussion emphasizes that a count of the number of genes in a genome may underestimate or overestimate the information encoded by those genes and may be no better than genome size at predicting organismal complexity. Instead, data from the sequenced eukaryotic genomes implicate more subtle genetic changes as being responsible for changes in morphological and behavioral complexity. However, a refined estimate of genomic information content, no matter how sophisticated, may not correlate simply with organismal complexity either. Within the next few years, some lowly organism (our guess is that it will likely be a plant) will be found to have a G-value in excess of our own. When that day arrives and we lose our exalted status as G-value champions, we can console ourselves with the knowledge that (genome) size is not everything.

Acknowledgments

We thank members of the Rausher and Wray Laboratories, T. Oakley, C. Cunningham, F. Nijhout, and M. Rausher for reading and discussing this article. We give special thanks to L. Moyle and D. McShea for thorough editing and comments.

REFERENCES

- Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., and Amanatides, P. G., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185–2195.
- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. 1989. *Molecular Biology of the Cell*. 2nd Ed. Garland Publishing, New York.
- Cavalier-Smith, T. 1985. *The Evolution of Genome Size*. John Wiley & Sons, New York.
- Davidson, E. H. 2001. *Genomic Regulatory Systems: Development and Evolution*. Academic Press, San Diego.
- Fell, D. A., and Wagner, A. 2000. The small world of metabolism. *Nat. Biotechnol.* 18: 1121–1122.
- Fields, C., Adams, M. D., White, O., and Venter, J. C. 1994. How many genes in the human genome. *Nat. Genet.* 7: 345–346.
- Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., and Feldmann, H., et al. 1996. Life with 6000 genes. *Science* 274: 546–567.
- Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., and Lehmann, G., et al. 1999. Alternative splicing of human genes: more the rule than the exception? *Trends Genet.* 15: 389–390.
- Herrmann, K. M., and Somerville, R. L. (eds). 1983. *Amino Acids: Biosynthesis and Genetic Regulation*. Addison-Wesley Publishing Company, Reading, MA.
- Hurst, L. D., and Smith, N. G. C. 1999. Do essential genes evolve slowly? *Curr. Biol.* 9:747–750.
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
- Jeong, H., Tombor, B., Albert, R., Oltval, Z. N., and Barabási, A.-L. 2000. The large-scale organization of metabolic networks. *Nature* 407: 651–654.
- Kauffman, S. A. 1993. *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University Press, Oxford.
- McShea, D. W. 1996. Metazoan complexity and evolution: Is there a trend? *Evolution* 50: 477–492.
- Oyama, S. 1985. *The Ontogeny of Information*. Cambridge University Press, Cambridge.
- Petsko, G. A. 2001. Size doesn't matter. *Genome Biol.* 2: 1003.1–1003.2.
- Sarkar, S. 1998. *Genetics and Reductionism*. Cambridge University Press, Cambridge.
- Sokal, R. R., and Rohlf, F. J. 1995. *Biometry*. W.H. Freeman and Company, New York.
- Sulston, J. E., and Horvitz, H.R. 1977. Post-embryonic cell lineages of the nematode *Caenorhabditis elegans*. *Dev. Biol.* 56: 110–156.
- Szathmáry, E., Jordan, F., and Pal, C. 2001. Can genes explain biological complexity? *Science* 292: 1315–1316.
- The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.
- The *C. elegans* Sequencing Consortium. 1998. Sequence and analysis of the genome of *C. elegans*. *Science* 282: 2012–2018.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., and Sutton, G. G., et al. 2001. The sequence of the human genome. *Science* 291: 1304–1351.

APPENDIX. GLOSSARY

C value

The amount of DNA found in a haploid genome, usually measured in millions of base pairs (Mb) or picograms (pg). The “C” is used to denote the constancy of this number in a species.

G value

The number of genes found in a haploid genome. These numbers are usually a combination of known genes and predicted open reading frames. Posttranscriptional processes such as alternative mRNA splicing and phosphorylation are not taken into account in calculating this number.

I value

The amount of information encoded by a genome. This estimate attempts to incorporate both the effective number of genes (by including alternative splicing, posttranslational modification, multidomain proteins, and gene redundancy) and the complexity added by gene expression and interacting genes.