JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, 2012
doi: 10.1111/j.1556-4029.2012.02252.x
Available online at: onlinelibrary.wiley.com

Letter to the Editor—Out with the "Junk DNA" Phrase

Sir,

What started as a clever talk title by Susumu Ohno (1) to describe non-protein-coding DNA (ncDNA) quickly became a ubiquitous phrase ("junk DNA") causing substantial confusion and distraction from a more sophisticated and accurate appreciation of the majority of the human genome that does not encode for proteins. While much of the scientific community rejected the vernacular—with some prominent scholars calling the characterization "ambiguous and even derogatory" (2)—the term persisted widely, in concept and informally, in academic literature and popular media. The scientific intricacies of the many humanomes (i.e., the genome, exome, transcriptome, and proteome) are relatively poorly understood by those outside the relevant disciplines, though those within the relevant disciplines cannot deny the importance of better understanding the molecular and cellular roles of ncDNA. While popular media are picking up on the general scientific sentiment that ncDNA has some importance (see, e.g., [3]), there is considerable confusion about (i) what the "junk" vernacular originally referred, (ii) what the implications of the subsequent scientific rejections of that vernacular are, and (iii) what the current characterizations of CODIS markers non-protein-coding DNA are. This confusion is apparent in important court opinions (e.g., [4–6]). Misconceptions of "junk DNA" are shaping the judiciary's perception of the loci used to in the standard CODIS profile and, subsequently, the judiciary's perception of the privacy implications of a CODIS profile and the appropriateness of the "fingerprint analogy" (e.g., [7,8]).

There was never a consensus among scientists that ncDNA was deserving of the "provocative term" coined by Ohno (9). While the diversity of non-protein-coding regions remained poorly understood for decades, at least four hypotheses explained the maintenance of these seemingly nonfunctional regions of the genome. The "selectionist hypothesis" posited that these regions regulate gene expression (10). The "neutralist hypothesis" posited these regions have no function but are transmitted passively as relics of evolutionary processes (10). The "intragenomic selectionist hypothesis" posited that non-protein-coding regions actively promote their own transmission and accumulate because of their elevated reproduction rate relative to protein-coding regions (10).[1] The "nucleotypic hypothesis" posited that these regions act to maintain structural integrity of the genome (10). When Ohno himself first used the term "junk DNA" to refer to all ncDNA, he had explicitly stated, "Certain untranscribable and/or untranslatable DNA base sequences appear to be useful…" (1, p. 367) Sydney Brennar, a molecular biologist, had distinguished "junk" from "garbage," explaining that, while garbage is worthless, used up, and thrown away, junk is of potential value and stored for unspecified future use (e.g., [9]). The characterization of ncDNA as "junk DNA" ultimately had the effect of "repel[ling] mainstream researchers" from studying it (11, p. 1246). It was not until the 1990s that scientists gave increasing attention to "junk DNA"—along with increased attention to every aspect of the genome spurred by the Human Genome Project—and began to appreciate the diversity of ncDNA "not [as] a single midden

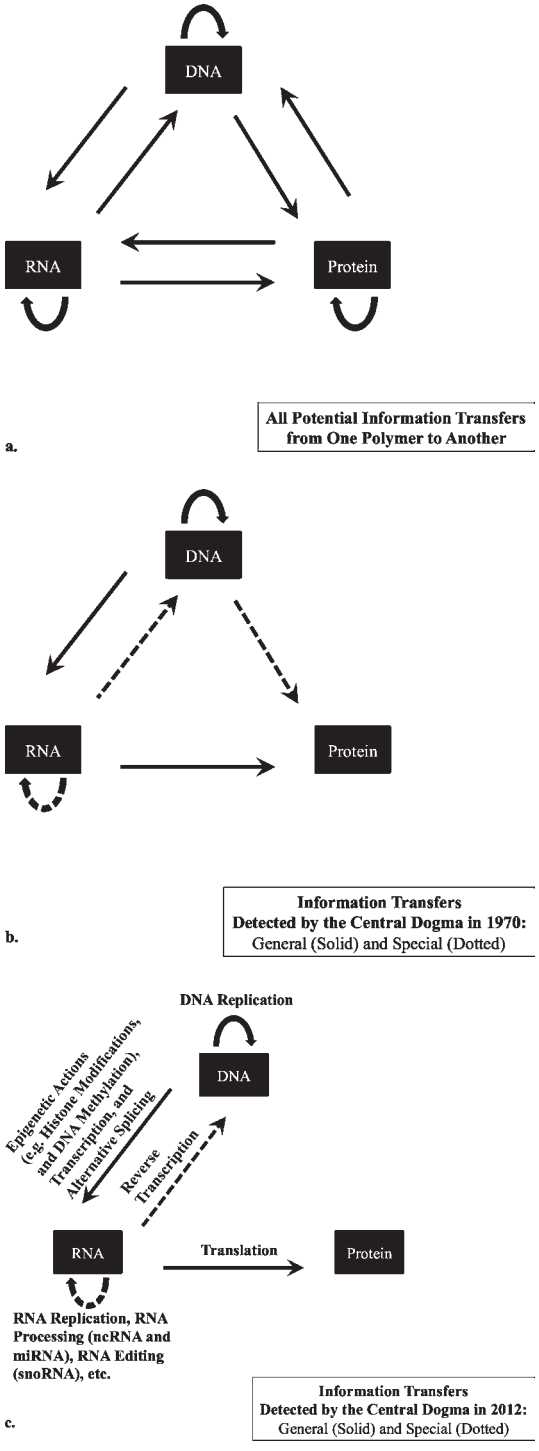

FIG. 1—*Information transfers and the central dogma. Panel (a) shows all possible transfers (adapted from Crick's figure 1 [16]). Panel (b) shows detected transfers in 1970 (adapted from Crick's figure 3 [16]). Panel (c) shows detected transfers in 2011 with notations summarizing some of the complexities of gene expression (adapted from Mattick's figure 1 (20) and Slack's figure 1 [21]). Correction made after online publication 7 Sept 2012: References in Fig 1 legend updated to reflect correct information.*

[1]This hypothesis was accompanied by the coinage of another unfortunate phrase, "selfish DNA."

TABLE 1—*Summary of non-protein-coding genomic elements.*

Non-Protein-Coding Genomic Element			Brief Description
Transcription regulatory elements			Molecular elements considered typical of gene structure, such as promoters, enhancers, and intronic splicing signals (21)
Introns			Segments of DNA located within genes that interrupt or separate exons from one another
5′ and 3′ untranslated regions		UTRs	Transcribed DNA sequences preceding (5′ UTR) and following (3′ UTR) coding sequences containing regulatory elements, such as binding sites for microRNAs (miRNAs), and polyadenylation signals (22)
RNA-specifying genes	MicroRNAs	miRNAs	Destabilize or inhibit the translation of targeted mRNAs; 19–25 nucleotides in length (23)
	Transfer RNAs	tRNAs	Facilitate translation by transporting specific amino acids to the ribosome; *c.* 80 nucleotides in length (23)
	Ribosomal RNAs	rRNAs	Facilitate the movement of tRNAs along the mRNA during translation; 4 types (18S, 28S, 5.8S, and 5S) (23)
	Spliceosomal RNAs	snRNAs	Facilitate the processing of pre-mRNAs (i.e., help splice introns that are not self-splicing); 5 types (U1, U2, U4, U5, and U6) (23)
	Small Nucleolar RNAs	snoRNAs	Facilitate posttranscription modifications of rRNAs, tRNAs, and snRNAs; 2 types (H/ACA box and C/D box) (23)
	Piwi-Interacting RNAs	piRNAs	Protect the integrity of the genome in germline cells during spermatogenesis; 26–34 nucleotides in length (23)
	RNAse P/MRP genes		Process tRNA and rRNA precursors (23)
	Long noncoding RNAs	lncRNAs	*c.* 200+ nucleotides in length, such as *XIST*, which silences an X chromosome during X-inactivation (23)
Repeat elements	Satellite DNA		DNA sequences often near centromeres and telomeres α-satellite or alphoid DNA, a 171-bp sequence that is repeated in tandem and clustered at the centromeres of all chromosomes. Repeat size of satellite DNA may be between 2 and 2000 bp and the size of the repeat array may be greater than 1000 bp (10,21)
	Minisatellites or Variable Number Tandem Repeats	VNTRs	Repeat units of 10–200 bp clustered into repeat arrays of 10–100 units Found near the telomeres (the terminal ends of chromosomes), but are also distributed across the chromosomes (10,21)
	Microsatellites or Short Tandem Repeats	STRs	Repeat units of 2–5 bp arranged in arrays of 10–100 units (10,21)
	Short Interspersed Nucleotide Elements	SINEs	*c.* 1,500,000 copies of SINEs present in the genome account for more than 10% of the genome (10,24)
	Long Interspersed Nucleotide Elements	LINEs	*c.* 850,000 copies of LINEs present in the genome, account for roughly one-fifth of the genome (10,24)
	Retrovirus-like Elements		*c.* 450,000 copies present in the genome (24)
	Transposons		*c.* 300,000 copies present in the genome (24)
Pseudogenes			Exhibit similarity to genes but lack introns and promoters and contain poly-A tails. Most pseudogenes have lost the ability to be transcribed (10,21,25)

heap…but [as] a complex mix of different types of DNA, many of which are vital…" (12, p. 608). Table 1 provides a summary of non-protein-coding elements of the genome.

The "-omic revolutions" that are dramatically and rapidly changing our understanding of the genome have not called into question the central dogma per se (as shown in Fig. 1), although they have certainly nuanced it by stressing the importance of noncoding function and have also challenged the conceptualization and definition of a "gene" (e.g., [13]). The components and physical boundaries of genes are no longer clear and discrete. Genes are more than just exons stitched together during transcription and subsequently translated into proteins. For example, in different contexts different combinations of exons may be used rather than all of them. Accordingly, the definition of a gene has been broadened to encompass not only the exonic sequence but also introns and intronic splicing sites, as well as promoters, enhancers, and other cis- and trans-regulatory elements (i.e., the factors located close to and far from the exons, respectively) that contribute to known phenotypes or functions. With the term "gene" increasingly being used to specify not only DNA sequences that encode proteins but also DNA sequences that do not encode protein but do specify RNA transcripts with known function, the term may be increasingly confusing to nonscientists and may be of diminishing operational value to scientists (see, e.g., [13]). With this in mind, we can leave the "junk" vernacular behind and refocus our attention to the current understanding of the human genome's structure and function and, specifically, how the standard and recommended CODIS markers (14,15) are characterized within this context.

Armed with the scientific and technological advances of the last 40–50 years, scientists in 2012 are able to better appreciate the complexities of the informational transfers articulated in 1958 as "the central dogma." (Coincidentally, despite over-generalizations and an array of distinct ideas attributed to it, "the central dogma"—as clarified by Francis Crick in 1970—did *not* stipulate that information transfer was only and always transferred from DNA to RNA to protein, did *not* stipulate that RNA lacked function aside from encoding proteins and "sa[id] nothing about control mechanisms" or gene expression [16, p. 562]). The diverse origins, characteristics, functions, and evolution of non-protein-coding regions of the genome are given increasing attention as scientists move beyond a simple Mendelian (one gene-one trait or disease) model and seek a more holistic understanding of human inheritance. This increased appreciation for

non-protein-coding regions of the genome does not, however, inherently give rise to increased significance of the diverse array of particular types of non-protein-coding regions.

Recent court opinions have asserted the markers in the standard CODIS profile were characterized as "junk DNA, because 'they are thought not to reveal anything about trait coding'" (e.g., [17, p. 5]). However, the 13 standard CODIS loci were attributed (indeed, burdened) with the label "junk DNA" because they are all microsatellites, and hence non-*protein*-coding. Indeed, the phrase "trait coding" itself reflects a dearth of genetic literacy among the legal profession. That those 13 specific loci—as well as the recent recommendation of 11 additional loci—were chosen for inclusion in a panel designed for identification purposes with an emphasis placed on a lack of association with known phenotypes (14,15) is an entirely separate issue from the loci being non-protein-coding elements. Accordingly, it is appropriate to encourage the discontinued characterization that CODIS loci are "junk DNA" (see also [18]). It is also appropriate to warn nonscientists that to imply the CODIS loci are each or collectively involved in gene expression and are now important for a wide array of traits and conditions of biomedical relevance is unfounded (19).

Selection of loci used for identification purposes is not a permanent, unalterable decision. Rather, it is possible for the forensic science community to revisit such decisions periodically and substitute markers in the event statistical associations, causal relationships, or predictive value for biomedically relevant phenotypes become known. Selection of markers for identity should be directed by the inherent usefulness of each marker to discriminate individuals and the experimental ease of amplification, rather than the negative qualitative value of the marker in detecting phenotype. Moreover, the arbitrariness of marker selection must be kept in mind—which phenotypes are considered "sensitive" or "medically relevant" are themselves subjective determinations and not universally agreed. Normative arguments surrounding the use of genetic information for molecular photofitting or phenotyping or the storage and unrestricted analysis of DNA samples can and must be kept separate from questioning whether it is scientifically possible to select a set of markers that are of value restricted to identification purposes. The scientific community, should it choose to do so, can relegate the "junk DNA" phraseology to the history books and forge ahead to a more nuanced understanding of genomics and the central dogma.

Acknowledgments

This work was funded by Grant No. P50HG004487-05 and Grant No. K99HG006446 from the National Human Genome Research Institute (NHGRI). The content is solely the author's and does not necessarily represent the official views of the NHGRI or the University of Pennsylvania.

References

1. Ohno S. So much 'junk' DNA in our genome. Brookhaven Symp Biol 1972;23:366–70.
2. Brosius J, Gould SJ. "On Genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci U S A 1992;89(22):10706–10.
3. Gibbs WW. The unseen genome: gems among the junk. Sci Am 2003;289(5):46–53.
4. *People v. Buza*, San Francisco Co. Super. Ct. SCN 207818 (First App. Dist, Ct. of App. CA, Aug. 4, 2011) at 22 (citing Gibbs' "Gems among the Junk" Sci Amer 2003 and explaining that the "quantity and nature" of information decipherable from the CODIS profile "will undoubtedly increase").
5. *United States v. Mitchell*, 652 F.3d 387 (3rd Cir (Pa) 2011), Dissenting opinion at 19 (stating "…it is little comfort that only so-called 'junk DNA' is used to compile a suspect's DNA profile" and noting "'with advances in technology, junk DNA may reveal far more extensive genetic information.' United States v. Kriesel, 379 F.3d 941, 947 (9th Cir. 2007).")
6. *State v. Abernathy*, No. 3599-9-11 (Vt. Super. Ct. June 1, 2012).
7. *United States v. Shavlovsky*, 2012 WL 652672 (E.D. Cal. 2012).
8. *United States v. Kriesel*, 508 F.3d 941, 947-948 (9th Cir. 2007).
9. Rabinow P. The anthropology of reason. Anthropol Today 1992;8(5):7–10, 7–8, paraphrasing Syndey Brenner.
10. Graur D, WH L. Fundamentals of molecular evolution, 2nd edn. Sunderland, MA: Sinauer, 2000;14, 274–5, 386–7, 392-4.
11. Makalowski W. Genomics. Not junk after all. Science 2003;300 (5623):1246–7.
12. Nowak R. Mining treasures from 'Junk DNA'. Science 1994;263 (5147):608–10.
13. Gingeras TR. Origin of phenotypes: genes and transcripts. Genome Res 2007;17:682–90.
14. Hares DR. Expanding the CODIS core loci in the United States. Forensic Sci Int Genet 2012;61(1):e52–e54.
15. Ge J, Eisenberg A, Budowle B. Developing criteria and data to determine best options for expanding the core CODIS loci. Investig Genet 2012;3:1. Doi: 10.1186/2041-2223-3-1.
16. Crick F. Central dogma of molecular biology. Nature 1970;2278 (5258):561–3.
17. *People v. Buza*, San Francisco Co. Super. Ct. SCN 207818 (First App. Dist, Ct. of App. CA, Aug. 4, 2011) at 5 (quoting Haskell v. Brown, 677 F.Supp.2d 1187, 1190 (N.D.Cal. 2009)).
18. Kaye DH. Please let's bury the junk: the CODIS loci and the revelation of private information. NW Univ Law Rev 2007;102:70–81.
19. Katsanis SH, Wagner JK. Characterization of the standard and recommended CODIS markers. J Forensic Sci 2012; doi: 10.1111/j.1556-4029.2012.02253.x [Epub ahead of print].
20. Mattick JS. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. BioEssays 2003;25:930–9.
21. Slack FJ. Regulatory RNAs and the demise of 'junk' DNA. Genome Biol 2006;7(9):328. Doi: 10.1186/gb-2006-7-9-328.
22. Maroni G. Molecular and genetic analysis of human traits. Malden, MA: Blackwell Science, 2001;57, 58–61, 61–65, 68, 82.
23. Neilson JR, Sandberg R. Heterogeneity in mammalian RNA 3′ end formation. Exp Cell Res 2010;316(8):1357–64.
24. Wright MW, Bruford EA. Naming 'junk': human non-protein coding RNA (ncRNA) gene nomenclature. Hum Genomics 2011;5(2):90–8.
25. Jobling MA, Hurles ME, Tyler-Smith C. Human evolutionary genetics: origins, peoples & disease. New York, NY: Garland Publishing, 2004;31.
26. Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DRF. Pseudogenes: pseudo-functional or key regulators in health and disease? RNA 2011;17:792–8.

Jennifer K. Wagner,[1] J.D., Ph.D.
[1]Center for the Integration of Genetic Healthcare Technologies, University of Pennsylvania, Philadelphia, PA, 19104.
E-mail: jennifer.wagner@uphs.upenn.edu