The HUGO Journal
a SpringerOpen Journal

REVIEW Open Access

The extent of functionality in the human genome

John S Mattick[1,2,3*] and Marcel E Dinger[1,2]

Abstract

Recently articles have been published disputing the main finding of the ENCODE project that the majority of the human genome exhibits biochemical indices of function, based primarily on low sequence conservation and the existence of larger genomes in some ostensibly simpler organisms (the C-value enigma), indicating the likely presence of significant amounts of junk. Here we challenge these arguments, showing that conservation is a relative measure based on circular assumptions of the non-functionality of transposon-derived sequences and uncertain comparison sets, and that regulatory sequence evolution is subject to different and much more plastic structure-function constraints than protein-coding sequences, as well as positive selection for adaptive radiation. We also show that polyploidy accounts for the higher than expected genome sizes in some eukaryotes, compounded by variable levels of repetitive sequences of unknown significance. We argue that the extent of precise dynamic and differential cell- and tissue-specific transcription and splicing observed from the majority of the human genome is a more reliable indicator of genetic function than conservation, although the unexpectedly large amount of regulatory RNA presents a conceptual challenge to the traditional protein-centric view of human genetic programming. Finally, we suggest that resistance to these findings is further motivated in some quarters by the use of the dubious concept of junk DNA as evidence against intelligent design.

Introduction

Recently there has been renewed discussion and controversy surrounding the extent and density of biochemical and biological functionality in the human genome (Graur et al. 2013; Doolittle 2013; Niu and Jiang 2013), prompted by the conclusion of the recent ENCODE studies (Dunham et al. 2012), following earlier analyses (Pheasant and Mattick 2007), that much if not most of the human genome may be functional. In particular the paper by Graur et al. 2013 has attracted particular attention because, unusually for a scientific paper, it employs not only logical argument to dispute but also sarcasm to ridicule the ENCODE conclusions.

Review

Putting polemic and ideology (see below) aside for the moment, the substantive scientific argument of Graur et al. is based primarily on the apparent lack of sequence conservation of the vast majority (~90%) of the human genome, suggesting that this indicates lack of selective constraint (and therefore function). The fundamental flaw, however, in this argument is that conservation is relative, and its estimation in the human genome is largely based on the questionable proposition that transposable elements, which provide the major source of evolutionary plasticity and novelty (Brosius 1999), are largely non-functional. This argument also overlooks a number of other assumptions and considerations that are tacitly embedded in conservation comparisons and their interpretation (Pheasant and Mattick 2007):

(i) relative conservation imputes function, but lack of (discernable) conservation imputes nothing (Pang et al. 2006), especially when there may be high turnover (Smith et al. 2004; Frith et al. 2006), different evolutionary rate classes in different types of functional elements (Taylor et al. 2006; Oldmeadow et al. 2010), and/or extended evolutionary distances involved (think '*frere*' and '*brother*' for a linguistic analogy);

(ii) like words, regulatory sequences have more relaxed structure-function constraints than protein-coding sequences, which encode analog devices with strict chemical requirements. Indeed this is well supported by comparative analysis of gene promoters, which

* Correspondence: j.mattick@garvan.org.au
[1]Garvan Institute of Medical Research, Darlinghurst NSW 2010, Australia
[2]St Vincent's Clinical School, University of New South Wales, Kensington NSW 2052, Australia
Full list of author information is available at the end of the article

nobody disputes are functional, but where orthologous function can be retained over large evolutionary distances in the absence of any recognizable primary sequence conservation (see e.g. Fisher et al. 2006);

(iii) regulatory sequences are the main genetic substrates for the exploration of phenotypic diversity in animals, by orchestrating the differential expression of a relatively stable and largely orthologous set of protein-coding genes (Pheasant and Mattick 2007; Taft et al. 2007; Carroll 2008), which diverge under positive selection for lineage-specific adaptive radiation;

(iv) the conclusion of lack of conservation of most of the human genome is largely based on a circular comparison with the rate of evolution of pan-mammalian ancient 'repeats', a slightly pejorative term referring to transposon-derived sequences (many with RNA origins), which are assumed to be largely non-functional and therefore evolving 'neutrally'. That is, one assumes that a subset of the genome is evolving neutrally and is therefore indicative of the rate of unconstrained divergence, then finds that most of the rest of the genome is behaving similarly, which is therefore concluded to also be non-functional. If the first assumption is incorrect, and increasing evidence suggests that it may be (Oldmeadow et al. 2010; Faulkner et al. 2009; Baillie et al. 2011) (although this is disputed in Graur et al. 2013), the derived conclusion of non-functionality of the rest of the genome is also incorrect (Pheasant and Mattick 2007).

The fact is we simply do not know - most elements in the human genome have not been subject to functional analysis, which itself is fraught with ascertainment difficulties (see e.g. Lewejohann et al. (2004) for a retrotransposon-derived example). While others have provided superficially independent evidence that ancient repeats are neutrally evolving, based on indel distribution rather than primary sequence comparison (Lunter et al. 2006), this is subtly subject to similar circular logic and lack of acknowledgement that protein-coding (and some miRNA) sequences may have structure-function constraints and therefore mutational patterns different from those in cis-regulatory sequences and other classes of trans-acting regulatory RNAs that emanate from the genome (Pang et al. 2006; Dinger et al. 2009);

(v) even if ancient repeats are neutrally evolving (which we think unlikely), the extant comparison set is restricted to those whose orthology is recognizable, some barely so, and therefore represents the more conserved end of a starting population whose full original distribution is unknown, thereby underestimating to an unknown extent the true 'neutral' evolution rate and therefore the extent of conservation of the remainder of the genome (Pheasant and Mattick 2007).

The other substantive argument that bears on the issue, alluded to in the quotes that preface the Graur et al. article, and more explicitly discussed by Doolittle (Doolittle 2013), is the so-called 'C-value enigma', which refers to the fact that some organisms (like some amoebae, onions, some arthropods, and amphibians) have much more DNA per cell than humans, but cannot possibly be more developmentally or cognitively complex, implying that eukaryotic genomes can and do carry varying amounts of unnecessary baggage. That may be so, but the extent of such baggage in humans is unknown. However, where data is available, these upward exceptions appear to be due to polyploidy and/or varying transposon loads (of uncertain biological relevance), rather than an absolute increase in genetic complexity (Taft et al. 2007). Moreover, there is a broadly consistent rise in the amount of non-protein-coding intergenic and intronic DNA with developmental complexity, a relationship that proves nothing but which suggests an association that can only be falsified by downward exceptions, of which there are none known (Taft et al. 2007; Liu et al., 2013).

In contrast to these uncertain indices, estimations and interpretations, the major fact to emerge from the ENCODE studies (Birney et al. 2007; Dunham et al. 2012) and their predecessors (Cheng et al. 2005; Carninci et al. 2005) is that the vast majority of the mammalian genome is differentially transcribed in precise cell-specific patterns (Mercer et al. 2008) to produce large numbers of intergenic, interlacing, antisense and intronic non-protein-coding RNAs, which show dynamic regulation in embryonal development (Dinger et al. 2008; Guttman et al. 2011; Ng et al. 2012), tissue differentiation (Sunwoo et al. 2009; Pang et al. 2009; Mercer et al. 2010; Askarian-Amiri et al. 2011) and disease (Gupta et al. 2010; Khaitan et al. 2011), with even regions superficially described as 'gene deserts' expressing specific transcripts in particular cells (Mercer et al. 2012; Roberts and Pachter 2011). Moreover, there is increasing evidence of their functional relevance (Mattick 2009b) and that a major function of these noncoding RNAs is to guide chromatin-modifying complexes to their sites of action, to supervise the epigenetic trajectories of development (Mattick and Gagen 2001; Dinger et al. 2008; Nagano et al. 2008; Pandey et al. 2008; Khalil et al. 2009; Mattick et al. 2009; Koziol and Rinn 2010; Spitale et al. 2011) - which appears to comprise a far greater fraction of human genetic programming than expected (Mattick 2004) in order to specify the architecture of the organism at a level of detail well beyond mere cell-type specification (Mattick et al. 2010).

Given these observations, we would submit that differential expression (including extensive alternative splicing) of RNAs is a far more accurate guide to the functional content of the human genome than logically circular assessments of sequence conservation, or lack thereof. Assertions that the observed transcription represents random noise (tacitly or explicitly justified by reference to stochastic ('noisy') firing of known, legitimate promoters in bacteria and yeast), is more opinion than fact and difficult to reconcile with the exquisite precision of differential cell- and tissue-specific transcription in human cells (for a recent debate see van Bakel et al. 2010; Clark et al. 2011). Moreover, where tested, these noncoding RNAs usually show evidence of biological function in different developmental and disease contexts, with, by our estimate, hundreds of validated cases already published and many more en route, which is a big enough subset to draw broader conclusions about the likely functionality of the rest. It is also consistent with the specific and dynamic epigenetic modifications across most of the genome, and concurs with the ENCODE conclusion that 80% of the genome shows biochemical indices of function (Dunham et al. 2012). Of course, if this is true, the long-standing protein-centric zeitgeist of gene structure and regulation in human development will have to be reassessed (Mattick 2004, 2007, 2011), which may be tacitly motivating the resistance in some quarters.

There may also be another factor motivating the Graur et al. and related articles (van Bakel et al. 2010; Scanlan 2012), which is suggested by the sources and selection of quotations used at the beginning of the article, as well as in the use of the phrase "evolution-free gospel" in its title (Graur et al. 2013): the argument of a largely nonfunctional genome is invoked by some evolutionary theorists in the debate against the proposition of intelligent design of life on earth, particularly with respect to the origin of humanity. In essence, the argument posits that the presence of non-protein-coding or so-called 'junk DNA' that comprises >90% of the human genome is evidence for the accumulation of evolutionary debris by blind Darwinian evolution, and argues against intelligent design, as an intelligent designer would presumably not fill the human genetic instruction set with meaningless information (Dawkins 1986; Collins 2006). This argument is threatened in the face of growing functional indices of noncoding regions of the genome, with the latter reciprocally used in support of the notion of intelligent design and to challenge the conception that natural selection accounts for the existence of complex organisms (Behe 2003; Wells 2011).

Conclusions

It is our position that these arguments are misguided. Indeed, we have refuted the specific claims that most of the observed transcription across the human genome is random (Clark et al. 2011; Mercer et al. 2012) and put forward the case over many years that the appearance of a vast layer of RNA-based epigenetic regulation was a necessary prerequisite to the emergence of developmentally and cognitively advanced organisms (Mattick 1994; Mattick and Gagen 2001; Mattick 2004; Amaral et al. 2008; Mattick 2009a, 2011). This case is, moreover, entirely consistent with the broad tenets of evolution by natural selection, although it may not be easily reconcilable with current population theory and current ideas of evolutionary neutrality. In any case, that our understanding of the remarkably complex processes underlying the molecular evolution of life, including the likely evolution of evolvability (Mattick 2009c), is incomplete should not be surprising. With the emergence of transformative technologies, such as massively parallel sequencing, which provide tools to view the inner molecular workings of the genome that were inconceivable less than a decade ago, it is as important as ever that we as scientists remain open to observations that challenge even the most fundamental paradigms that exist within biology today.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JSM and MED wrote the manuscript. Both authors read and approved the final manuscript.

Author details
[1]Garvan Institute of Medical Research, Darlinghurst NSW 2010, Australia. [2]St Vincent's Clinical School, University of New South Wales, Kensington NSW 2052, Australia. [3]School of Biotechnology & Biomolecular Sciences, University of New South Wales, Kensington NSW 2052, Australia.

Received: 28 April 2013 Accepted: 2 July 2013
Published: 15 July 2013

References

Amaral PP, Dinger ME, Mercer TR, Mattick JS (2008) The eukaryotic genome as an RNA machine. Science 319:1787–1789

Askarian-Amiri ME, Crawford J, French JD, Smart CE, Smith MA, Clark MB, Mattick JS (2011) SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer. RNA 17:878–891

Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Faulkner GJ (2011) Somatic retrotransposition alters the genetic landscape of the human brain. Nature 479:534–537

Behe MJ (2003) A functional pseudogene: an open letter to Nature. http://www.discovery.org/a/1448

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, de Jong PJ (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816

Brosius J (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115–134

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Hayashizaki Y (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563

Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25–36

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Gingeras TR (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149–1154

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Mattick JS (2011) The reality of pervasive transcription. PLoS Biol 9:e1000625

Collins FS (2006) The Language of God: a Scientist Presents Evidence for Belief. Free Press, New York

Dawkins R (1986) The Blind Watchmaker: Why the evidence of Evolution Reveals a Universe without Design. W. W, Norton & Company, Inc, New York

Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Mattick JS (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res 18:1433–1445

Dinger ME, Amaral PP, Mercer TR, Mattick JS (2009) Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Brief Funct Genomics Proteomics 8:407–423

Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci USA 110:5294–5300

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Birney E (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Carninci P (2009) The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41:563–571

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS (2006) Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312:276–279

Frith MC, Ponjavic J, Fredman D, Kai C, Kawai J, Carninci P, Sandelin A (2006) Evolutionary turnover of mammalian transcription start sites. Genome Res 16:713–722

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, 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 5:578–590

Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Chang HY (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464:1071–1076

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Lander ES (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477:295–300

Khaitan D, Dinger ME, Mazar J, Crawford J, Smith MA, Mattick JS, Perera RJ (2011) The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion. Cancer Res 71:3852–3862

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Rinn JL (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA 106:11667–11672

Koziol MJ, Rinn JL (2010) RNA traffic control of chromatin complexes. Curr Opin Genet Dev 20:142–148

Lewejohann L, Skryabin BV, Sachser N, Prehn C, Heiduschka P, Thanos S, Prior H (2004) Role of a neuronal small non-messenger RNA: behavioural alterations in BC1 RNA-deleted mice. Behav Brain Res 154:273–289

Liu G, Mattick JS, Taft RJ (2013) A meta-analysis of the genomic and transcriptomic composition of complex life. Cell Cycle 12:127–138

Lunter G, Ponting CP, Hein J (2006) Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol 2:e5

Mattick JS (1994) Introns: evolution and function. Curr Opin Genet Dev 4:823–831

Mattick JS (2004) RNA regulation: a new genetics? Nat Rev Genet 5:316–323

Mattick JS (2007) A new paradigm for developmental biology. J Exp Biol 210:1526–1547

Mattick JS (2009a) Deconstructing the dogma: a new view of the evolution and genetic programming of complex organisms. Ann N Y Acad Sci 1178:29–46

Mattick JS (2009b) The genetic signatures of noncoding RNAs. PLoS Genet 5:e1000459

Mattick JS (2009c) Has evolution learnt how to learn? EMBO Rep 10:665

Mattick JS (2011) The central role of RNA in human development and cognition. FEBS Lett 585:1600–1616

Mattick JS, Gagen MJ (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol 18:1611–1630

Mattick JS, Amaral PP, Dinger ME, Mercer TR, Mehler MF (2009) RNA regulation of epigenetic processes. Bioessays 31:51–59

Mattick JS, Taft RJ, Faulkner GJ (2010) A global view of genomic information–moving beyond the gene and the master regulator. Trends Genet 26:21–28

Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS (2008) Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA 105:716–721

Mercer TR, Qureshi IA, Gokhan S, Dinger ME, Li G, Mattick JS, Mehler MF (2010) Long noncoding RNAs in neuronal-glial fate specification and oligodendrocyte lineage maturation. BMC Neurosci 11:14

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Rinn JL (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotech 30:99–104

Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, Feil R, Fraser P (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322:1717–1720

Ng SY, Johnson R, Stanton LW (2012) Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. Embo J 31:522–533

Niu DK, Jiang L (2013) Can ENCODE tell us how much junk DNA we carry in our genome? Biochem Biophys Res Commun 430:1340–1343

Oldmeadow C, Mengersen K, Mattick JS, Keith JM (2010) Multiple evolutionary rate classes in animal genome evolution. Mol Biol Evol 27:942–953

Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Kanduri C (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell 32:232–246

Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22:1–5

Pang KC, Dinger ME, Mercer TR, Malquori L, Grimmond SM, Chen W, Mattick JS (2009) Genome-wide identification of long noncoding RNAs in CD8+ T cells. J Immunol 182:7738–7748

Pheasant M, Mattick JS (2007) Raising the estimate of functional human sequences. Genome Res 17:1245–1253

Roberts A, Pachter L (2011) RNA-Seq and find: entering the RNA deep field. Genome Med 3:74

Scanlan J (2012) The Designer's Detritus: ENCODE, Junk DNA, and Intelligent Design Creationists. Scitable (Nature Education), http://www.nature.com/scitable/blog/student-voices/the_designers_detritus_encode_junk

Smith NG, Brandstrom M, Ellegren H (2004) Evidence for turnover of functional noncoding DNA in mammalian genome evolution. Genomics 84:806–813

Spitale RC, Tsai MC, Chang HY (2011) RNA templating the epigenome: Long noncoding RNAs as molecular scaffolds. Epigenetics 6:539–43

Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL (2009) MEN ε/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. Genome Res 19:347–359

Taft RJ, Pheasant M, Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 29:288–299

Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CA (2006) Heterotachy in mammalian promoter evolution. PLoS Genet 2:e30

van Bakel H, Nislow C, Blencowe BJ, Hughes TR (2010) Most "dark matter" transcripts are associated with known genes. PLoS Biol 8:e1000371

Wells J (2011) The Myth of Junk DNA. Discovery Institute Press, Seattle

doi:10.1186/1877-6566-7-2
Cite this article as: Mattick and Dinger: **The extent of functionality in the human genome.** *The HUGO Journal* 2013 **7**:2.