Conceptual and empirical challenges of ascribing functions to transposable elements

by

Tyler A. Elliott[1], Stefan Linquist[2], and T. Ryan Gregory[1,*]

[1] Department of Integrative Biology, University of Guelph, Guelph, Ontario N1G 2W1 Canada

[2] Department of Philosophy, University of Guelph, Guelph, Ontario N1G 2W1 Canada

telliott@uoguelph.ca

linquist@uoguelph.ca

* Corresponding author

E-mail: rgregory@uoguelph.ca
Phone: 1-519-824-4120, x58053
Fax: 1-519-767-1656

Publication Reference: Elliott, T.A. Linquist, S. Gregory, T.R. (2014). Conceptual and empirical challenges of ascribing functions to transposable elements. *The American Naturalist*, 184(1): p. 14-24

Keywords: causal role, ENCODE, genome size, junk DNA, selected effects, transposons

Short Title: Function and transposable elements

Manuscript Type: Article

Abstract

The media attention and subsequent scientific backlash engendered by the claim, announced by spokespeople for the Encyclopedia of DNA Elements project (ENCODE), that 80% of the human genome has a "biochemical function" highlights the need for a clearer understanding of function concepts in biology. This article provides an overview of two major function concepts that have been developed in the philosophy of science – the "causal role" concept and the "selected effects" concept – and their relevance to ENCODE. Unlike some previous critiques, the ENCODE project is not considered problematic because it employed a causal role definition of function (which is relatively common in genetics), but because of how this concept was misused. In addition, several unique challenges that arise when dealing with transposable elements (TEs), but which were ignored by ENCODE, are highlighted. These include issues surrounding TE-level versus organism-level selection, the origins versus the persistence of elements, and accidental versus functional organism-level benefits. Finally, some key questions are presented that should be addressed in any studies aiming to ascribe functions to major portions of large eukaryotic genomes, the majority of which is made up of transposable elements.

Introduction

Though their quantity, diversity, activity level, and composition vary considerably among species, it is becoming increasingly clear that transposable elements (TEs) represent the dominant type of DNA sequence within most eukaryotic genomes (Gregory 2005). Notably, transposable elements and inactive remnants thereof make up two thirds of the total DNA content of the human genome (de Koning et al. 2011). Whereas the human genome contains approximately 20,000 protein-coding genes, it is home to more than 3 million recognizable transposable element copies (International Human Genome Sequencing Consortium 2001). Moreover, animal genome sizes are known to vary more than 7,000-fold, with most of this diversity thought to be the result of differential TE abundance.

Observations such as these raise important questions about the effects of TEs on host organisms and the conditions that contribute to their spread and persistence over evolutionary time. Perhaps the most deeply entrenched view holds that transposable elements are in some sense "functional," meaning that they confer some benefit to the genome/organism in which they are found. Going back as far as McClintock (1950), TEs were thought to play an essential role in gene regulation. Other authors have suggested that TEs play a crucial role in generating genetic variation through their mutagenic effects (e.g., McClintock 1984; Biémont and Vieira 2006), while still others posit that the presence of large swaths of non-coding DNA buffering the protein-coding genes against mutations (e.g., Yunis and Yasmuneh 1971; Patrushev and Minkevich 2008). In fact, organism-level functions have been proposed for every new type of non-coding DNA sequence upon its discovery, and TEs are no exception. A prevailing assumption for many decades has been that any genetic element that is so widespread must be functional, or else it would have been eliminated by natural selection (see Appendix 1).

It is simply not true that non-coding DNA has long been dismissed as worthless junk and that functional hypotheses have only recently been proposed – despite the oft-repeated cliché in media reports and the introductions of far too many scientific papers. Indeed, it was specifically in reaction to the persistent assumption that most or all of the genome *is* functional that the classic "selfish DNA" papers were written in 1980 (Doolittle and Sapienza 1980; Orgel and Crick 1980). However, even then, the possibility was left open that at least some transposable elements would prove to be functional at the organism level:

> It would be surprising if the host genome did not occasionally find some use for particular selfish DNA sequences, especially if there were many different sequences widely distributed over the chromosomes. One obvious use … would be for control purposes at one level or another. (Orgel and Crick 1980).

As Orgel et al. (1980) noted, it is an empirical question as to the proportion of transposable elements that have taken on organism-level functions. Many individual examples are now known of transposable elements that have been co-opted into organism-level functional roles (see Sinzelle et al. 2011). However, these still represent a tiny minority of TEs, and the conditions that generate and sustain organism-level functions remain the subject of considerable debate.

An unfortunate obstacle to progress in this debate has been a lack of clarity regarding definitions of "function." Recently, these conceptual issues reached center stage with the rise of the Encyclopedia of DNA Elements project (ENCODE). This project aimed "to delineate all functional elements encoded in the human genome" by cataloguing "regions of transcription, transcription factor association, chromatin structure and histone modification" (The ENCODE Project Consortium 2012). The ENCODE project involved more than 400 scientists, cost around $200 million (Maher 2012), and culminated in the simultaneous publication of 30 papers in

September 2012. It created a large data resource enabling future analysis of the human genome, but nearly all of the extensive media coverage of the project focused on a single result:

> These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions (The ENCODE Project Consortium 2012).

This claim, that 80% of the human genome exhibits "biochemical function," was widely (mis)interpreted as indicating that most (perhaps even all) non-coding DNA is biologically functional at the organism level, and thus that the ENCODE results overturned the concept of "junk DNA" (Appendix 1) This was clearly the interpretation promoted by ENCODE leadership and by summaries published in *Nature* and *Science* (Ecker et al. 2012; Pennisi 2012; Appendix 1). Almost immediately, these claims were subject to staunch criticism primarily on the grounds that the ENCODE definition of "function" was not only extremely liberal, it also involved equivocating between so-called "biochemical function" (i.e., a positive result in at least one of the chosen assays in at least one cell type) and more commonly understood meanings of the word (e.g., Doolittle 2013; Eddy 2013; Graur et al. 2013; Niu and Jiang 2013). Much of this rather heated debate hinges on the definition of "function" as applied to non-coding DNA sequences.

This paper has two objectives. The first is to review the concepts of function that have been developed in the philosophy of science and their application to non-coding DNA elements, including most recently by ENCODE. In the process, we note that some critiques of ENCODE's use of "function" have been somewhat oversimplified – it is not the concept of function that was used that is problematic per se, but how it was used by ENCODE. The second goal is to explore some specific complications to functional interpretations that arise as a result of the unique biological features of transposable elements (TEs). These complications include the need to consider multiple levels of selection, the distinction between the origin and the persistence of a

particular DNA element, and the distinction between functions and accidental benefits for the host.

Two conceptions of function in the philosophy of biology

The ENCODE claim that >80% of human DNA sequences exhibit a ”biochemical function” has prompted many molecular biologists to examine the meaning of this term. To this end, some critics of ENCODE (Graur et al. 2013; Doolittle 2013) have appealed to work on concepts of function in the philosophy of science, in particular the distinction between “selected effect” (SE) functions and “causal role” (CR) functions.

A selected effect function is any capacity that has been shaped or maintained by natural selection in the past. Hence, to ascribe an SE function is to make a historical claim. By contrast, CR functions are ascribed to the parts of a system that contribute to any system-level capacity of interest (Cummins 1975). Unlike SE functions, CR functions do not hinge on facts about evolutionary history (Amundson and Lauder 1994). Instead, CR functional ascriptions identify the relations between the capacity of a system and the activities of its component parts. It is sometimes said that SE functions explain why a trait exists, while CR functions explain how it works.

Although some critics of ENCODE appeal explicitly to this philosophical distinction between CR and SE functions, they have done so in a way that many philosophers would find somewhat problematic. For example, Doolittle claimed that, “most philosophers of biology, and likely, most practicing biologists when pressed, would endorse some form of the selected effect (SE) definition of function”(Doolittle 2013; see also Graur et al. 2013).Thus, the criticism of ENCODE is simply that they have used a CR definition of function, which some authors

consider inherently flawed. However, philosophers have identified numerous biological applications of the CR concept (Brandon 2011) for example in the field of functional morphology(Amundson and Lauder 1994). It is argued that these applications could not be served by an SE concept alone, and that the CR concept is in fact more basic than the SE concept (Griffiths 2006; but see Rosenberg and Neander 2007, Linquist forthcoming). Although these issues are unresolved, many philosophers allow that both SE and CR concepts play appropriate, but distinct epistemic roles.

Understanding how philosophers have reached this pluralist near-consensus requires recapping some key moves in this debate. Unlike the physical sciences, where purpose-oriented concepts of function were abandoned after the death of Aristotelian physics, function-talk persists in the biological sciences. Philosophers of science recognize that modern biologists could not be using "function" in the classical, teleological sense. So what else might they mean by this term? Initially, the only plausible answer to this question seemed to be that all biological functions are SE functions (Milikan 1984; Rosenberg, 1985; Neander 1991). This interpretation has several advantages. It allows one to draw a clear distinction between adaptive functions and accidental benefits. It also establishes objective criteria for ascribing functions to biological traits.

However, the prevalence of the SE definition was soon called into question. . Some philosophers argued that the SE-function concept is too restrictive and hence cannot account for all legitimate functional ascriptions in biology (Amundson and Lauder 1994; Griffiths 1994). They argued that within entire branches of biology the use of "function" carries no commitments about selective history. For example, an oncologist might be interested in the conditions that promote metastasis in some cell lineage. A particular genetic mutation might contribute to this

process, in which case, one might say that the mutation *functions* in this capacity. Advocates of CR functions in biology sought to make room for these non-SE, but nonetheless standard uses of "function" in the life sciences.

Today, perhaps the closest thing to a consensus among philosophers of biology is that each function concept is associated with a distinct type of explanatory goal. On this view, the SE function concept is appropriate for developing evolutionary or "ultimate" explanations; while the CR concept is appropriate for explaining "proximate" mechanisms. As philosopher Paul Griffiths has recently stated, "unless anatomy, physiology, molecular biology, developmental biology, and so forth turn their attention specifically to evolutionary questions, they investigate functions in the causal [role] sense." (Griffiths, p. 3. 2006). Importantly, philosophical debates on this issue have taken their lead from biologists' use of language. If a large and productive community of researchers appeal to functions in the causal role sense, this is taken by philosophers as good evidence that the concept is doing good epistemic work. However, the ENCODE controversy calls this very methodology into question. ENCODE's use of "function" would likely be very surprising to many philosophers who take scientific practice at face value. Philosophers have long recognized an in-principle weakness in CR functions, but have assumed that this shortcoming is more of a conceptual than a practical problem. The ENCODE controversy reveals that philosophers' worst fears can become a reality.

A problem with CR functions, long recognized by philosophers but often dismissed as trivial, stems from their interest-relativity. According to the CR mode of functional analysis, the elements of a system that are identified as functional depend entirely on the system-level capacity that a researcher chooses to investigate. In principle, a researcher could select any system-level capacity for functional analysis. Taken to the extreme, this can result in some rather

odd-sounding functions. For example, one might claim that the function of evaporation and condensation is to produce rain, or that the function of tectonic movement is to cause earthquakes. As absurd as they sound, strictly speaking these are legitimate applications of the CR function concept. Nevertheless, this "permissiveness problem" is often dismissed as a "philosopher's problem," arising in principle but never in practice because scientists do not actually employ the concept so loosely. Real working biologists, it is often assumed, are too careful to use the CR function concept so liberally or, even worse, in ways that might be intentionally self-serving or misleading.

The legitimacy of this assumption is called into question by ENCODE, in which affiliated researchers have adopted an extremely liberal criterion for ascribing causal role functions to genetic elements. According to ENCODE, a sufficient condition for qualifying as "biochemically functional" is that a sequence of DNA exhibits at least one of the following properties, at least once, in at least one of the 147 cell types analyzed: (1) it is transcribed into RNA (but not necessarily translated into a protein), or (2) it contains or is adjacent to a transcription binding factor, or (3) it is a methylated CpG dinucleotide, or (4) it is located in an area of open chromatin, or (5) it is found organized in nucleosomes containing certain histone modifications (The ENCODE Project Consortium 2012). Obviously, such permissive criteria will identify a great many genetic elements regardless of whether they have been under selective pressure or contribute to any meaningful organism-level capacities. And, in fact, some of these analyses are likely to give high rates of positive results simply by chance (Graur et al. 2013; White et al. 2013). However, the problem is not simply that ENCODE has adopted the CR concept; rather, the issue is that their assays detect "function" only in the sense of giving a positive result on their assays without any demonstration of actual biological significance. It is entirely possible that an

otherwise *biologically inert* strand of DNA will bind to a protein, and thereby qualify as "functional" according to the ENCODE criteria.

An analogy may be useful here. Imagine an individual who wishes to use a metal detector to find valuables on a beach. First, he visits a jewelry store in order to establish the machine's ability to detect gold and silver Satisfied, he begins scanning the beach. Occasionally he hits upon an old nail or a c bottle cap, which causes the metal detector to light up and emit a sound. Technically, triggering the metal detector could be considered a CR function of these pieces of discarded metal. However, this is not the same thing as locating treasures, and it would be false to assume that every hit with the detector was identifying something useful just because this was the case in the jewelry store. Yet, this appears to be what ENCODE has done by employing assays that are normally used to find unambiguously functional elements (e.g., genes) and then considering any positive result elsewhere in the genome to be an indication of "biochemical function". Some of the hits identified by ENCODE may indeed be gold, but most could be bottle caps.

To summarize, critics of ENCODE who rely on philosophical insights in their critique of ENCODE have been somewhat too focused on the use of CR function concepts *per se*. They are right about the importance of the distinction between CR and SE functions but it is not the case that CR functions are widely recognized among philosophers as inherently faulty. To the contrary, most philosophers of biology, and arguably most biologists (including geneticists) when pressed, would recognize that CR functional claims play a valid role in the context of proximate explanations so long as the concept is not misused. Of course, there remains an open question as to whether CR functions *should* be so used in place of SE functions in the context of developing proximate explanations (see Linquist, forthcoming). There is also an evident

disadvantage in having multiple concepts of function in use, namely that this may invite confusion or even outright equivocation – this is especially true if only the word "function" is used, without reference to the actual concept being employed (Doolittle et al., in review).

In any case, the issue with ENCODE is not simply one of semantics. Although its use of a CR concept of function is not necessarily problematic in itself, this has been implemented in an extraordinarily loose manner. Moreover, the fact that most of the human genome is made up of transposable elements greatly complicates the application of ENCODE's criteria. As discussed below, there are important distinctions regarding the functions of transposable elements (TEs) which can only be drawn by using the SE rather than the CR concept of function. These distinctions are of more than just evolutionary interest – they are also relevant when investigating TEs on a proximate, mechanistic level in terms of their effects on organismal phenotypes.

Common criteria for identifying functions in non-mobile genetic elements.

A variety of criteria are commonly used to identify functions in non-mobile genetic elements including protein-coding genes, regulatory domains, and other such sequences. Several such criteria are reviewed briefly below before considering how their application becomes complicated when applied to the majority of the genome made up of transposable elements. It is particularly relevant to note that some of these criteria generate evidence for CR functions rather than SE functions, indicating that CR functions are not considered inherently flawed within genetic analyses.

1) *Context-specific transcription*

One criterion used to suggest function is context specificity of an RNA transcript. This includes tissue specificity, developmental stage specificity, or stimulus specific activation, such

as in response to stress (eg., Özgür et al. 2012; Belmonte et al. 2013). In such cases, context specificity is regarded as a form of adaptive specialization: a less adapted genetic element would not become active under such specific circumstances, or, so the thinking goes. This assumption can of course be challenged on the grounds that some context specific activations are non-functional (see section 4). But the identification of a context specific transcript is at least suggestive of functionality and worthy of more detailed investigation. In terms of the aforementioned distinction among function concepts, the operative sense of "function" in these cases appears to be the SE variety. Specialization and adaptation are the grounds on which functions are being ascribed.

2) Positional information

A second criterion for ascribing function to a DNA sequence occurs in context of genome-wide bioinformatics studies. These studies identify associations between particular DNA sequences and a wide range of genomic properties that might indicate a functional role. For example, DNA sequences located upstream of a known coding region are often good candidates for the assignment of functional roles. Here the underlying assumption is that the proximity of these sequences to structurally significant sites implies a regulatory role. In this case, the operative sense of function is a CR notion. A genetic sequences is being identified as functional (or not) according to its role in development. It is perhaps taken for granted that selection will often act on these sequences. But such historical questions about evolutionary origin and maintenance are of secondary importance. As with the previous criterion, evidence of active location is imperfect evidence of function. It is possible that these sequences are not actually playing a role in development. It is therefore important to supplement positional information with additional data.

3) *Sequence conservation*

A third criterion for identifying function is sequence conservation among species. Conservation can be measured by percent identity the of coding sequence, by shared base pairs outside the coding sequence, or by conserved position in the genome. The underlying assumption is that such instances of conservation are evidence of purifying selection (Lindblad-Toh et al., 2011). Hence, this criterion clearly assumes an SE-function concept.

4) *Experimental manipulation*

The last method for identifying functional relevance of a DNA sequence is through experimental manipulation. This can involve reporter-gene assays for transcription and protein production, such as placing a putative regulatory sequence upstream of a reporter gene and assaying for the production of mRNA and/or protein (Xiong et al., 2012). Sequences that are found in coding regions can also be mutated and their phenotypic effects observed (Kim et al., 2010). Or, their transcripts can be targeted for silencing by short interfering RNAs (siRNAs) and the effects on the phenotype of the organism or tissue in question can be observed (Kleimhammer et al., 2010). In all such cases, the search for function involves manipulating a sequence and looking for direct effects on the phenotype. Importantly, this criterion does not directly investigate questions of adaptive significance. Although phenotypic effects are often under selection pressure, this is by no means guaranteed. At the same time, using manipulation as a criterion for identifying functions fits squarely within the tradition of investigating CR functions. Hence, a CR function is the prevailing conception of function in these cases.

Unique Challenges for Ascribing Function to TEs

As the brief overview above reveals, both CR and SE functional concepts are operative within the standard criteria used to ascribe functions to non-TE sequences. Indeed, some combination of these may provide the most compelling evidence that a particular DNA sequence is functional in a biologically meaningful sense. Unfortunately, the ENCODE claim of 80% function in the human genome was based strictly on the CR function concept. "Biochemical function", as they used the term, simply referred to a positive result in at least one of their chosen assays, making this part of ENCODE's analysis entirely "closed" and self-referential (Table 1). That is to say, it lacked any connection to information external to the system, such as experimental evidence of phenotypic impacts at the organismal level or comparisons among species to demonstrate sequence conservation.

As others have noted, ENCODE could simply have chosen a slightly more liberal criterion – say, that the sequence is replicated, or that it contains a suitable binding site for DNA polymerase – and they would have been guaranteed to identify "function" in 100% of the genome (Graur et al. 2013). Presumably, the ENCODE authors would consider such a result trivial and uninformative, but this raises the question as to why they chose the assays that they did. It seems reasonable to conclude that they did so because similar assays have been successful in detecting sequences with biological functions at the organism level (e.g., protein-coding genes, regulatory regions, etc.). However, there is a crucial distinction between ENCODE and the previous work that used such criteria: ENCODE examined the entire genome, most of which is not protein-coding genes or obvious regulatory domains but is primarily made up of transposable elements and their remnants.

Unlike protein-coding genes and other non-mobile regions, TEs possess unique biological properties that must be taken into account when interpreting the results of analyses like those outlined above or employed in the ENCODE project. As noted, transposable elements and their non-autonomous derivatives are by far the most common sequences in the human genome (de Koning et al. 2011), which means that a large portion of the sequences identified as "functional" by ENCODE must fall within TE-derived sequences. Moreover, some of these TEs will still be active or retain some of their former biological activity (eg. the ability to recruit transcription factors). However, ENCODE applied its criteria unilaterally across all components of the genome and neglected to consider TEs within their proper biological and evolutionary context. In particular, the mobile nature of TEs creates at least three major sources of complexity: 1) by introducing a second level at which evolutionary processes can operate (Doolittle 1989), such that SE functions may relate to the TE level rather than the standard organism level, 2) by potentially shifting from a parasitic element to one with an organism-level function ("exaptation"; Gould and Vrba 1982), such that the origin of a sequence and the reasons for its persistence are decoupled, and 3) by adding the possibility that a given TE merely has beneficial side-effects for the organism but that these represent, at most, CR functions.

1) Organism-level versus TE-level evolution

Some transposable element-derived sequences are known to be important for gene regulation, as part of normal developmental processes, in the vertebrate immune system, and in various other ways. On the other hand, many are also implicated as disease causing mutagens. TEs exhibit many characteristics in common with viruses, and as such they are most often characterized as parasites of the host genome. Like viruses, active TEs harness the host's replication machinery, but they are able to move about and become duplicated independently of

the rest of the genome. Moreover, TEs exhibit heritable variation in their ability modify copy rate, as well as in their capacity to avoid deletion. In other words, they display the set of properties that are sufficient for evolution through natural selection (Lewontin, 1970).

The fact that TEs may undergo evolution at their own intragenomic level greatly complicates efforts to assign them SE functions at the organism level because it introduces alternative explanations that must be ruled out. For example, a discovery that TEs exhibit widespread sequence conservation may be evidence of organism-level selection because of the TEs phenotypic effects (especially if only certain TE insertions are conserved), but it could also be the result of intragenomic selection on transposition ability (e.g., if all active TE copies are conserved). The question, even when evidence of SE functions is found, is *cui bono* – who benefits, the organism, the TE, or both?

As a notable example, it has often been reported that TEs become active when the organism/cell encounters stress. In many cases, this correlation has been interpreted to indicate that the TEs play an adaptive role in the cellular stress response (e.g., McClintock 1984,Shapiro 2011, Chénais et al. 2012) and thus that this represents an SE function of the TEs involved. However, there are several possible TE-level explanations that would need to be ruled out in order for this hypothesis to hold. For example, it is possible that stress causes a breakdown in the repression mechanisms that normally keep TEs in check, thereby allowing them to become active. It is also possible that both TEs and stress-response genes are normally inactivated by being methylated or packaged into chromatin, and that activating the stress-response genes also activates nearby TEs as a side effect. Another possibility is that TEs respond to stress in the host cell and become active in preparation to “abandon ship” and facilitate transfer to a new host.

These alternative hypotheses are testable, but so far they have generally been missing from discussions of the correlation between TE activity and stress.

2) Origin versus persistence

One of the primary points raised in the original "selfish DNA" papers (Doolittle and Sapienza 1980; Orgel and Crick 1980) was that organism-level functions are not necessary to explain the existence of large quantities of transposable element DNA. The fact that they are capable of autonomous or semi-autonomous replication means that TEs can exist simply because they are good at existing. Similarly, one need not find an organism-level function to explain the presence of every virus or bacterium in the human body.

That said, it is clear that some TEs have been co-opted to serve important roles in normal genome function. So, even though these sequences may begin as parasites, their continued persistence may in some cases relate to their effects on host phenotypes. In other words, there may be a shift in the level of selection that accounts for the types and quantities of certain TEs within a genome. As noted by Gould and Vrba (1982), this process of "exaptation" means that the explanation for a trait's origin and that for its persistence (e.g., current function) may be very different. In the case of TEs, it is critical to distinguish between questions of origin versus current persistence, because there are multiple possible explanations available, especially for persistence.

One concrete example of exaptation is that of the *RAG1* protein of vertebrates. This protein helps to carry out a process known as V(D)J recombination where different exons are shuffled and ligated together to form the first step in the production of an antigen-specific binding protein known as an antibody (Oettinger et al. 1990). *RAG* proteins mediate these

precise rearrangements of DNA by binding to recognition signal sequences (RSS) between target exons and excising the intervening sequences into a circular piece of DNA for degradation (Jones and Gellert 2004). Cutting and ligation of DNA strands to accomplish this task is performed by a fusion protein derived from the transposases of elements from the *Transib* and *CMC* superfamilies (Kaptinov and Jurka 2005; Panchin and Moroz 2008). These TEs are no longer capable of independent replication and no active members of either superfamily are found in the human genome. Hence these TEs appear to be exapted for their host-level functions.

Many cases are less clear cut, however. An interesting case is the telomeric non-long terminal repeat (non-LTR) retrotransposons in drosophilids. A telomerase gene has not been found in *D. melanogaster* and their telomeres are primarily composed of three different non-LTR retrotransposons named *HeT-A*, *TART* and *TAHRE* (Pardue et al. 1996; Abad et al. 2004).These elements, which are no longer capable of independent replication, rely on one another's promoter capacity, reverse transcriptase, and *Gag* proteins for proper replication and targeting to telomeres (Danilevskaya et al. 1997, Rashkova et al. 2002; Rashkova et al. 2003, Shpiz et al. 2007). This interdependency among TEs, along with the absence of host-generated teleomeres, suggest coevolution with each other and with the host. But one might argue that these elements have not been completely exapted, as their protein coding regions alone have not been incorporated into the host genome like in *RAG*1. Instead, these telomeric elements might represent a point along a symbiotic continuum reflecting a mutualistic relationship with the host, as opposed to one that is purely parasitic or purely serving only the host (Kidwell and Lisch 2001, Durand and Michod, 2010).

3) Biological functions versus beneficial side-effects

The attribution of organism-level functions to TEs becomes most complex in cases where specific properties or behaviors of TEs appear to confer benefits to both TEs and the host organism. In such instances, the origin of the TE and the evolution of its current effects are potentially explained exclusively by selection at the TE level, with the positive effect on the host being viewed as merely a beneficial side-effect. Alternatively, the TE may have begun as a parasite but then became modified by selection on organismal phenotypes as part of its domestication into a host-level functional role. Associated changes in the TE sequence and/or the interaction between TE and host may then become beneficial to the TE in terms of allowing it to persist within the genome without being deleted. Thus, not only can there be evolution at multiple levels, but there can be synergistic or countervailing interactions between those levels depending on the changes engendered by evolutionary processes at each level.

A relevant example is provided by the process of double stranded break (DSB) repair. Retrotransposon DNA has been found at the repair sites of DSB in yeast. This association pattern is sometimes taken to suggest that TEs have a host-level function in DNA repair (Moore and Haber; Teng et al. 1996). In fact, there is probably a better explanation for this pattern. Transposable elements in general have an association with breaks in DNA and the host proteins mediating the repair of those breaks. Both DNA transposons and retrotransposons must cut DNA at locations where they insert, with DNA transposons cutting when they excise. These breaks must then be repaired by cellular repair proteins to ensure not only host survival, but also element survival. Thus, proteins involved with recognizing and repairing DNA breaks tend to associate with elements and the proteins they encode (Beall et al. 1994; Downs and Jackson 1999; Gasior et al. 2006). The intrinsic association between TEs and repair proteins suggest the

possibility of a co-evolutionary, antagonistic relationship. On this view, TEs require the enzymatic abilities of repair proteins to repair breaks in DNA, but repair proteins are selected to limit the spread of TEs and their damage to the genome (Sawyer and Malik 2006). Even when the ability of retrotransposons to cut target DNA is impaired, they have been shown to insert back into the genome through so-called endonuclease independent mechanisms, at the sites of DSBs (Morrish et al. 2002; Sen et al. 2007; Ichinaga and Okada 2008; Srikanta et al. 2009). Futher support comes from studies in yeast and mouse fibroblasts showing that mitochondrial and contaminant *E. coli* DNA are used as substrates for repair (Yu and Gabriel 1999; Lin and Waldman 2001). These findings suggest that cells will use any abundant extra-nuclear DNA (not just TEs) to repair breaks in the genome. Taken together, these data suggest a viable alternative to the hypothesis that TEs have been selected for a role in host DNA repair (Eickbush 2002). Although DSB repair may be an accidental benefit of TE replication, this is not necessarily a host-level function.

Questions to ask when attempting to identify functions for TEs The attribution of functions to the majority of the genome that is made up of transposable elements is complex for a number of reasons, both conceptual and empirical. As yet, there is no clear-cut set of procedures to reliably and unambiguously resolve these issues. However, much confusion is avoided if the purpose of an investigation is made explicit by addressing a number of key questions. A list of such questions is provided below, and although it is not exhaustive, it provides a means of alleviating some of the problems that have plagued the ENCODE project.

1. *Is the objective of the study to identify CR functions, and how are these distinguished from false positives?*

In some cases, the goal of a study may simply be to catalogue the positive results of some particular set of assays (i.e., closed-system CR functions which may or may not have SE functions). There are ways in which this can be useful, for example by identifying sequences of potential biological importance that can then become the target of further study to determine what, if any, their biological functions may be. Misrepresentations about the genome being "80% functional" aside, a catalogue of this sort was, in fact, the objective and most significant outcome of the ENCODE project.

However, it must be noted that positive results in any particular assay are not sufficient – these must be compared against a null hypothesis that provides an indication of how many false positives are expected due to chance. Notably, a recent study by White et al. (2013) highlighted this issue by showing that even randomly-generated sequences will tend to provide reproducible results in assays of regulatory capability. Likewise, de Souza et al. (2013) have argued that very few of the proposed cases of TEs taking on functions as regulatory binding sites are supported by sufficient evidence. Thus, without a null model that quantifies the expected rate of false positives and clear standards of evidence for demonstrating biological significance, even assessments of closed-system CR functions may be greatly exaggerated.

2. What other evidence will be included in order to assess potential biological functions?

The objective of a study may be to identify not just closed-system CR functions, but sequences of biological significance at the organism level, possibly including SE functions. As this would no longer be a closed-system analysis (i.e., with "function" defined only in terms of a positive result on a chosen assay), it must make reference to external information to validate claims of biological significance. For example, it may include comparisons across taxa to search for evidence of phylogenetic conservation or correlations between genetic properties and

phenotypic traits across species. It could also make reference to comparisons among individuals, if there are differences in genomic properties and quantifiable traits within a species. Or, a study could make use of experimental manipulations (knock outs, deletions) to examine effects on organismal phenotypes. Without such additional information, any conclusions regarding actual biological impact (let alone SE functions) are speculative at best.

3. *If the goal is to identify organism-level (SE) functions of TEs, how will alternative explanations be ruled out?*

Once again, it is important to bear in mind that the unique properties of TEs introduce important complexities into assessments of function at the organism level such that identifying a correlation between TE presence/activity and a particular trait does not, by itself, provide evidence that the TE has an SE function in this regard. For example, transcription alone does not automatically indicate an organism-level function for a retrotransposon because passage through an RNA intermediate is part of their own replication mechanism. As noted above, becoming active during stress or being incorporated into double-strand break repairs could also be explained from a TE-level perspective without these being organism-level SE functions of the elements involved. To reiterate, there are important distinctions that must be made between the origins of a DNA sequence (e.g., as a parasitic TE), the reasons for its persistence and abundance (e.g., it continues as an active parasite or it has been coopted by the host genome), and explanations for correlations with organismal traits (e.g., it serves a role at the organism level and has been under selection for this reason, or it merely exerts beneficial side effects for the host).

Concluding Remarks

The possibility that the majority of non-coding DNA plays an important functional role at the organism level has been actively discussed for many decades. While it is not true that most of the genome was simply dismissed as useless junk, there have long been legitimate debates regarding the percentage of DNA that is biologically important in large eukaryotic genomes. This is a question that will require both empirical data and conceptual clarification to resolve.

For example, the recent claims by the ENCODE project leadership that 80% of the human genome can be assigned a "biochemical function" are highly misleading because of the way in which the concept of "function" was employed. The issue is not simply that ENCODE made use of a causal role definition of function rather than a selected effects definition, as the CR definition is relatively common in genetics Rather, it is because ENCODE misapplied this definition of function by using criteria that were far too broad. Equivocation between this loose concept of causal role function and phenotypically relevant biological functions exacerbated the confusion surrounding the ENCODE results.

As described in this review, ascribing functions to specific components of the genome is uniquely challenging when the sequences involved are transposable elements. Their capacity form autonomous replication creates several major complications that confound the use of functional assessments that are typically implemented in studies of genes or regulatory regions. These unique challenges were ignored by ENCODE because the entire human genome was treated in the same way, despite the fact that it is made up primarily of TEs. Future work that aims to provide an estimate of the percentage of DNA in the human genome with a biologically meaningful function at the organism level will therefore require a much more sophisticated approach that takes these issues into account.

Acknowledgements

This work was supported by an Ontario Graduate Scholarship (OGS) to T.A.E. and a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to T.R.G. We thank Ford Doolittle, Susan Kalisz, Kimberley Hughes, and an anonymous reviewer for comments that improved the manuscript.

Literature Cited

Abad, J. P., B. de Pablos, K. Osoegawa, P. J. de Jong, A. Martín-Gallardo, and A. Villasante. 2004. *TAHRE,* a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. Molecular Biology and Evolution 21:1620-1624.

Amundson, R. and G.V. Lauder. 1994. Function without purpose: the uses of causal role function in evolutionary biology. Biology and Philosophy 9: 443-469.

Beall, E. L., A. Admon, and D. C. Rio. 1994. A *Drosophila* protein homologous to the human p70 Ku autoimmune antigen interacts with the P transposable element inverted repeats. Proceedings of the National Academy of Sciences of the United States of America 91:12681-12685.

Belmonte, M. F., R. C. Kirkbride, S. L. Stone, J. M. Pelletier, A. Q. Bui, E. C. Yeung, M. Hashimoto, J. Fei, C. M. Harada, M. D. Munoz, B. H. Le, G. N. Drews, S. M. Brady, R. B. Goldberg, and J. J. Harada. 2013. Comprehensive developmental profiles of gene activity in

regions and subregions of the *Arabidopsis* seed. Proceedings of the National Academy of Sciences of the United States of America 110:E435-444.

Biémont, C. and C. Vieira. 2006. Genetics: Junk DNA as an evolutionary force. Nature 443: 521-524.

Brandon, R.N. 2011. A general case for functional pluralism. in *Function: Selection and Mechanisms* (ed. P. Huneman) 97-104 (Springer)

Cummins, R. 1975. Functional analysis. The Journal of Philosophy 72: 741-765.

Chénais, B., A. Caruso, S. Hiard, and N. Casse. 2012. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. Gene 509:7-15.

Danilevskaya, O. N., I. R. Arkhipova, K. L. Traverse, and M.-L. Pardue. 1997. Promoting in tandem: the promoter for telomere transposon *HeT-A* and implications for the evolution of retroviral LTRs. Cell 88:647-655.

de Koning, A. P. J., W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. 2011. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genetics 7:e1002384.

de Souza, F.S.J., L.F. Franchini, and M. Rubinstein. 2013. Exaptation of transposable elements into novel *cis*-regulatory elements: is the evidence always strong? Molecular Biology and Evolution 30: 1239-1251.

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

Doolittle, W. F. 1989. Hierarchical approaches to genome evolution. Canadian Journal of Philosophy Supp 14:101-133.

Doolittle, W. F. 2013. Is junk DNA bunk? A critique of ENCODE. Proceedings of the National Academy of Sciences of the United States of America 10.1073/pnas.1221376110.

Downs, J. A., and S. P. Jackson. 1999. Involvement of DNA end-binding protein Ku in Ty element retrotransposition. Molecular and Cellular Biology 19:6260-6268.

Durand, P. M., and R. E. Michod. 2010. Genomics in the light of evolutionary transitions. Evolution 64:1533-1540.

Ecker, J.R., W.A. Bickmore, I. Barroso, J.K. Pritchard, Y. Gilad, and E. Segal. 2012. Genomics: ENCODE explained. Nature 489: 52-55.

Eddy, S. 2012. The C-value paradox, junk DNA and ENCODE. Current Biology 22: R898-R899

Eickbush, T. H. 2002. Repair by retrotransposition. Nature Genetics 31:126-127.

Gasior, S. L., T. P. Wakeman, B. Xu, and P. L. Deininger. 2006. The human LINE-1 retrotransposons creates DNA double-strand breaks. Journal of Molecular Biology 357:1383-1393.

Graur, D., Y. Zheng, N. Price, R. B. R. Azevedo, R. A. Zufall, and E. Elhaik. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biology and Evolution 5:578-590.

Griffiths, P. 1994. Cladistic classification and functional explanation. Philosophy of Science 61: 216-227.

Griffiths, P.E. 2006. Function, homology and character individuation. Philosophy of Science 73:1-25.

Gregory, T. R. 2005. Synergy between sequence and size in large-scale genomics. Nature Reviews Genetics 6:699-708.

Gould, S. J., and E. S. Vrba. 1982. Exaptation-a missing term in the science of form. Paleobiology 8:4-15.

Hangauer, M.J., I.W. Vaughn, and M.T. McManus. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genetics 9: e1003569.

Ichiyanagi, K., and N. Okada. 2008. Mobility pathways for vertebrate L1, L2, CR1, and RTE clade retrotransposons. Molecular Biology and Evolution 25:1148-1157.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Jones, J. M., and M. Gellert. 2004. The taming of a transposon: V(D)J recombination and the immune system. Immunological Reviews 200:233-248.

Kapitonov, V. V., and J. Jurka. 2005. RAG1 core and V(D)J recombination signal sequences were derived from *Transib* transposons. PLoS Biology 3:e181.

Kidwell, M. G., and D. R. Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution 55:1-24.

Kim, D.-U., J. Hayles, D. Kim, V. Wood, H.-O. Park, M. Won, H.-S. Yoo, T. Duhig, M. Nam, G. Palmer, S. Han, L. Jeffery, S.-T. Baek, H. Lee, Y. S. Shim, M. Lee, L. Kim, K.-S. Heo, E. J. Noh, A.-R. Lee, Y.-J. Jang, K.-S. Chung, S.-J. Choi, J.-Y. Park, Y. Park, H. M. Kim, H.-M. Park,

K. Kim, K. Song, K. B. Song, P. Nurse, and K.-L. Hoe. 2012. Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. Nature Biotechnology 28:617-623.

Kleinhammer, A., W. Wurst, and R. Kühn. 2010. Gene knockdown in the mouse through RNAi. Methods in Enzymology 477:387-414.

Lewontin, R. 1970. The Units of Selection. *Annual Review of Ecology and Systematics* 1:1-18.

Lin, Y., and A. S. Waldman. 2001. Capture of DNA sequences at double-strand breaks in mammalian chromosomes. Genetics 158:1665-1674.

Lindblad-Toh, K., M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 78:476-482.

Maher, B. Fighting about ENCODE and junk. Nature News Blog, September 6th 2012

McClintock, B. 1950. The origin and behaviour of mutable loci in maize. Proceedings of the National Academy of Sciences of the United States of America 36:344-355.

McClintock, B. 1984. The significance of responses of the genome to challenge. Science 226:792-801.

Milikan, R. 1984. Language, Thought, and Other Biological Categories. Cambridge, M.A.: MIT Press.

Moore, J. K., and J. E. Haber. 1996. Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. Nature 383:644-646.

Morrish, T. A., N. Gilbert, J. S. Myers, B. J. Vincent, T. D. Stamato, G. E. Taccioli, M. A. Batzer, and J. V. Moran. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nature Genetics 31:159-165.

Neander, K. 1991. Functions as selected effects: The conceptual analysts defense. Philosophy of Science 58: 168-184.

Niu, D.-K., and L. Jiang. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? Biochemical and Biophysical Research Communications 430: 1340-1343

Oettinger, M. A., D. G. Schatz, C. Gorka, and D. Baltimore. 1990. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science 248:1517-1523.

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

Orgel, L.E., F.H.C. Crick, and C. Sapienza. 1980. Selfish DNA. Nature 288: 645-646.

Özgür, E., U. Mert, M. Isin, M. Okutan, N. Dalay, and U. Gezer. 2013. Differential expression of long non-coding RNAs during genotoxic stress-induced apoptosis in HeLa and MCF-7 cells. Clinical and Experimental Medicine 13:119-126.

Pedersen, E. S. Lander, and M. Kellis. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 78:476-482.

Pennisi, E. 2012. ENCODE Project Writes Eulogy for Junk DNA. Science 337: 1159-1161.

Panchin, Y., and L. L. Moroz. 2008. Molluscan mobile elements similar to the vertebrate Recombination-Activating Genes. Biochemical and Biophysical Research Communications 369:818-823.

Pardue, M.-L., and P. G. DeBaryshe. 2011. Retrotransposons that maintain chromosome ends. Proceedings of the National Academy of Sciences of the United States of America 108:20317-20324.

Patrushev, L.I. and I.G. Minkevich. 2008. The problem of eukaryotic genome size. Biochemistry 73: 1519-1552.

Rosenberg, A. 1985. The Structure of Biological Science. Cambridge, MA, Cambridge University Press.

Sawyer, S. L., and H. S. Malik. 2006. Positive selection of yeast nonhomologous end-joining genes and a retrotransposon conflict hypothesis. Proceedings of the National Academy of Sciences of the United States of America 103:17614-17619.

Sen, S. K., C. T. Huang, K. Han, and M. A. Batzer. 2007. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. Nucleic Acids Research 35:3741-3751.

Shapiro, J.A. 2011. *Evolution: A View from the 21st Century*. FT Press Science, Upper Saddle River, NJ.

Shpiz, S., D. Kwon, A. Uneva, M. Kim, M. Klenov, Y. Rozovsky, P. Georgiev, M. Savitsky, and A. Kalmykova. 2007. Characterization of *Drosophila* telomeric retroelement *TAHRE*: transcription, transpositions, and RNAi-based regulation of expression. Molecular Biology and Evolution 24:2535-2545.

Sinzelle, L., Z. Izsvák, and Z. Ivics. 2009. Molecular domestication of transposable elements: From detrimental parasites to useful host genes. Cellular and Molecular Life Sciences 66:1073-1093.

Srikanta, D., S. K. Sen, E. M. Conlin, and M. A. Batzer. 2009. Internal priming: an opportunistic pathway for L1 and *Alu* retrotransposition in hominins. Gene 448:233-241.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.

Teng, S., B. Kim, and A. Gabriel. 1996. Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. Nature 383:641-644.

White, M.A., C.A. Myers, J.C. Corbo, and B.A. Cohen. 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proceedings of the National Academy of Sciences of the United States of America 110: 11952-11957.

Rashkova, S., S. E. Karam, R. Kellum, and M.-L. Pardue. 2002. Gag proteins of the two *Drosophila* telomeric retrotransposons are targeted to chromosome ends. The Journal of Cell Biology 159:397-402.

Rashkova, S., A. Athanasiadis, and M.-L. Pardue. 2003. Intracellular targeting of Gag proteins of the *Drosophila* telomeric retrotransposons. Journal of Virology 77:6376-6384.

Xiong, A.-S., R.-H. Peng, J. Zhuang, J. Davies, J. Zhang , and Q.-H. Yao. 2013. Advances in directed molecular evolution of reporter genes. Critical Reviews in Biotechnology 32:133-142.

Yu, X., and A. Gabriel. 1999. Patching broken chromosomes with extranuclear cellular DNA. Molecular Cell 4:873-881.

Yunis, J.J. and W.G. Yasmineh. 1971. Heterochromatin, satellite DNA, and cell function. Science 174: 1200-1209.

Table 1. Summary of the causal role (CR) and selected effects (SE) concepts of function, with additional distinctions and objections relevant to their use in discussions of transposable elements

Causal Role (CR) Functions		Selected Effect (SE) Functions	
Basic definition:	Main objection:	Basic definition:	Main objection:
The capacity of some lower level component to make a functional contribution to a system level capacity that is selected by an investigator.	Functions are investigator relative and insufficiently constrained.	Any capacity of a system for which that system was under natural selection in the past.	Too limited to capture all meaningful senses of "function" in biology.
Further distinctions:	Proposal:	Further distinctions:	Proposal:
Closed systems are entirely self-referential, with causal role functions determined only in terms of the assays used and not connected to broader systems outside the study. **Open systems** are not exclusively self-referential, but include connections to information about broader systems, such that functions identified may not simply be assay-specific CR functions but may also be shown to be SE functions with biological significance for organismal phenotypes.	The problem with ENCODE is that it adopts a closed system approach to the identification of CR functions.	**Host level** SE functions are capacities of TEs that were selected to benefit hosts. **TE level** SE functions are capacities that were selected to benefit TEs. **Origin functions** are selected effects (at either level) for which an element became established in a population. **Maintenance functions** are selected effects (at either level) for which an element is maintained in a population **Accidental benefits** are capacities that might increase an element's frequency, but for which there has been no selection.	For genetic elements that fall within mobile elements it is necessary to observe these distinctions.