FROM SINGLE GENE TO PHENOTYPE: QUESTIONING A DISCRETE GENE IN EXPLAINING PHENOTYPE DIVERSITY

Karl Smith Byrne (ks677@cam.ac.uk). University of Cambridge

Supervisado por: José Miguel García Ramírez. Departamento de Psicología Social. Universidad de Granada

Fecha de recepción: 24 de enero de 2014.
Fecha de revisión: 19 de marzo de 2014.
Fecha de aceptación: 11 de abril de 2014.

Summary

The great diversity of phenotypes across organisms raises the question of how it emerged from the digital DNA sequence. Often the question is summarized as 'how many genes do we really need?' The benefit of answering this is readily apparent; particularly since sequencing the genome, research has sought the origin of normative and pathological phenotypes in our genes. However, in response, the literature will retort that neither the number of genes nor the size of the genome make robust predictions about phenotype complexity or diversity. For example, a common sea flea, *Daphnia pulex*, has ~31,000 genes, compared with our ~23,000. Given the gene-centric state of current biology, the questions these comparisons advance about the power of the gene are disconcerting. The remit of what follows is to address the value of quantifying genes to explain the phenotype. The heterogeneous nature of gene definitions in the literature necessitates brief discussion of gene ontology. Following this, I will discuss how function emerges from the transcript, and the resultant translated functional product. However, questioning the power of the gene should be taken in tandem with questioning its scope; this discussion will end on a brief survey of the proportion of the phenotype that should readily be attributed to non-DNA inheritance information, which highlights the need or systems-based approaches to phenotype variability.

Key words: Transcript Function, Protein Interaction, non-DNA Inheritance, Systems Approaches

Introduction

Initially, in the early 19th century the gene was defined loosely as a basic unit of heredity (Morgan et al. 1915). However, during the subsequent century research additionally defined it as a discrete locus, that produced a single polypeptide, and that was a transcribed code in the DNA molecule (Watson & Crick, 1953; Beadle & Tatum, 1941). These studies cumulatively produce an oft cited definition: a unit of heredity, which is a discrete segment of DNA that codes for a polypeptide chain. However, as sequencing techniques advance, this definition becomes increasingly unable to cope with emerging data; for brevity, select examples are highlighted. Primarily, not all genes were found to code for proteins: ~50% of transcriptional units in mice code for functional RNA molecules. Those coding for proteins do not conform to a single protein model; a single gene locus may encode a variety of proteins via alternate splicing. By

extracting exons to be combined in multiple permutations, alternate splicing may generate numerous mRNA codes that translate into unrelated proteins. Further, such single gene loci are not found linearly discrete. Discontinuous sense and anti-sense organisation allows for a single gene to be subsumed within the intron of another gene (Gerstein et al., 2007). Lastly, recent literature suggests the coding/regulatory distinction may be obsolete. Transcription factors have been found to bind within ~14% of exon nucleotides, which demonstrates that genes may simultaneously code for proteins and cell regulation (Stergachis et al., 2014; Pesole, 2008).

Although these studies represent great advances in the field of genetics, they also necessitate an updated definition of the gene that allows the maintenance of its integral relationship with the phenotype. The working definition used henceforth is: *The gene is a union of functional genomic sequences encoding a coherent set of potentially overlapping functional products* (Gerstein et al., 2007).

To evaluate well the contribution to phenotypic diversity, it is therefore necessary to clarify a functional transcript, and its relation to the gene, and to expound the role of resultant products in the phenotype.

A typical definition of a functional transcript is a unit of RNA or DNA which, when transcribed, reliably translates into a functional product. However, improved sequence technologies reveal a number of exceptions that appear functional, yet challenge primacy of a functional dichotomy. Consider, the non-sense mediated decay pathway (NMD), which normally removes transcripts with premature stop-codons. While not directly influencing the phenotype, coding genes may alternate transcription between coding regions and distal NMD transcripts as a rate limiting factor in protein production (Huang et al., 2010). Clearly not functional as a coding gene would be described, NMD transcripts temporarily gain function, in the context of gene regulatory processes. Given such findings, it is of benefit to consider the inverse: What is a non-functional transcript? A non-functional transcript is one created by biological mechanism, as opposed to imperfect stochastic artifacts of transcription, which does not contribute directly to phenotypic diversity. Often, transposable elements (TE), such as Alu repeats, will be given as example of a non-functional product. However, recent literature finds exception; a subset of TE has indeed been found functional, and demonstrates splice signals as a substrate for exon creation. Further, TE are inserted, without detrimental effect, into functional transcripts – TE overlap with ~62% of long non-coding RNA and UTR (Sorek 2007; Shen et al. 2011). In light of these findings, it is difficult to parse the functional from the non-functional transcript, and thus, precludes the ability to quantify gene effect from the functionality of the entire genome - clouded by the vastness of genomic data.

Notably, the model producing such difficulties derives the genome as static; that transcript function is fixed for context temporally. However, next-generation technologies are finding TE more numerous. These additional data have discovered that many TE are recent evolutionary developments, and may be considered as sources of genome innovation. Evolution is a trial and error process, and it is possible that ‘non-functional’ transcripts are intermediaries in the evolutionary testing process. Although such dynamism may blur the lines between genetic noise and novelty, it is clear that the genome should benefit from the flexible functionality of ‘in progress’ transcripts. Such as with NMD and TE, specific contexts may highlight their minor functionality as beneficial to fitness, and facilitate development (Mudge, Frankish, & Harrow, 2013). Conceptually, then, it appears more beneficial to re-evaluate the transcript as being functionally transcribed. That although the gene may be to a certain degree a fixed sequence, the context and need for novelty may supplement with a

fluidity of function in additional transcripts. However, a practical challenge remains; such a model of transcript function requires a flexible structural framework to allow for the quantification of the genetic effect. Many exceptions to traditional functionality are compounded by the inability of static genome structure to accommodate the simultaneous activation of mutually distal elements.

Such difficulty may arise from the conceptual simplification that the genome is linear and two dimensional, when in reality it is a dynamic, three-dimensional structure. To facilitate gene expression, local genome folding regularly occurs between remote regulatory elements and genes, and between genes and transcription factors. This amendment may reconcile the flexible functionality and preserve gene function effects. For example, in expression of erythroid genes, mutually distal genes are found to relocate to an active transcription site. The necessary distal elements are trafficked to a central scaffold of pre-assembled transcriptional complexes, for context-specific expression (Mercer & Mattick, 2013). Note that movement implied is restricted. Although ligand-induced changes may expand genomic regions, it does not allow major reorganization of chromosomes. Such dynamism may help facilitate the conceptual advances in understanding of transcript functionality. However, in doing so, the genome topology is required to adopt a more regulatory role to preserve the measure of the transcriptional gene unit. Further, that this new framework accommodate existing genetic phenomena. For example, the ability of the genome to appropriate a distinct structure by context reveals a modular genetic design. Specifically, those coding and regulatory elements may be incorporated alternatively along a single sequence dependent on necessity. The benefit is readily apparent: a three-dimensional conception allows for the overlap of genes, and facilitates the role of epigenetic effects. Histone modifications and methyl groups act in symbiosis with extant topological constraints, and constrict possible intron and exon boundaries to create the central scaffold for transcription compartments (Luco et al., 2011; Kornblihtt et al, 2013).

Further mechanisms for such dynamism may be found in the specificity offered by transcription factor use. Single gene loci have ~3 promoters on average, these may drive multiple sequences simultaneously by local folding to fulfill necessity. This facilitates a complex array of function specific activation in genes, in a complex of transcripts; for example the developmentally specific Hox gene activation. Crucial for body plan development, Hox genes display collinear activation across overlapping domains. The activation occurs according the current developmental conditions, the body axis, of the embryo. The process involves the successive relocation of genes to an active transcription compartment; however, inactive Hox genes are isolated in a single structure delimited from the active flanking regions (Noordermeer et al., 2011). By integrating this structure with the concept of a functional transcript, it becomes possible to delineate the gene as a unit in measure of the phenotype. The functional transcript may thus be revised as a group of functionally related DNA or RNA sequences, which when transcribed robustly contribute to phenotypic variability. Importantly, these additions do not require a large redefinition, as the gene may be seen as largely constant. However, the three-dimensional model makes explicit that transcript function occurs in complex hubs of activation; many genes are concurrently trafficked to direct a specific phenotype. Multi-gene activation may, thus, question the value of the single gene unit; a question that is best answered by their translated products, and the phenotype.

Typically, a functional product is considered the result of translation that produces a measureable effect on phenotypic diversity, with a primary example being amino acids and resultant proteins. The premise behind the contribution to the phenotype is well explained through gene-wide association analysis designs; that a protein will produce a

difference in the phenotype, based on the fixed sequence variants, and thus, that a reliable functional product is based at the single gene loci. However, many such studies report disturbingly low genetic contributions to behavior. One recent study of obesity aetiology, found that 1% of total expected variation was due to genetic factors. A similar study of breast cancer risk factors found no benefit of introducing common genetic variants. As such, emerging proteomics research is questioning the benefit of relating the functional locus of phenotype impact to a single gene product, such as a protein. For example, the functional locus may be removed from a fixed single location, and be dependent on the context, as in the protozoan T. Brucei. It infects mammalian blood streams and evades host immune function by regularly altering its glycoprotein surface. Importantly, this phenotype is controlled by ~1000 VSG genes, which are transcribed one at a time – multiple fixed gene loci and products may specify a single phenotype, which questions the value of affixing product function to the single gene (Ruiz-narváez, 2012).

The prior example is, however, a single protein-based trait, and may not be a representative evaluation of the power of single protein-gene contribution to the phenotype. A more complex example can be taken from the well-annotated *lac* operon involved in lactose digestion. An operon is a segment of gene comprising regulatory transcripts and multiple code that produce multiple proteins that act as a polycistronic mRNA with related metabolic functions. The *lac* operon is an example of how a complex trait is well described by considering the interaction of many products (Beckwith et al., 1970; Bassford et al., 2003). Such a function would be missed if single gene protein behavior were assumed, and is in accord with recent literature that states only rarely can discrete biological function be attributed to individual products. Rather, most biological constituents for the phenotype arise from a complex network of pairwise interactions.

A more expansive example can be drawn from the seminal protein-protein interaction yeast study on *Saccharomyces cerevisiae*. Jeong and collegues studied the impact on the phenotype of a protein deletion from the removal of a single gene. Firstly, 93% of all proteins had at least five pairwise interactions. Secondly, results found that the phenotypic variability attributable to the gene deletion was largely described by the topological position of its resultant protein, in the complex network of molecular interactions. Notably, not all protein deletions were found equally essential to function. In contrast, only 21% when deleted proved lethal to the organism (Jeong et al., 2001). That the importance for phenotype maintenance was proportionate to the degree of connectivity emphasizes the relevance of protein-protein interaction for describing behavior. Additionally, when proteins that are less integrated were deleted, the phenotype was more likely to buffer against deleterious effects. Thus, despite the importance of discrete biological function, the phenotype is better described by the integration of gene products, than by a singular protein.

These studies describe the functional contribution of the genes to the phenotype as being somewhat removed from the single gene locus. Rather, the single gene may serve as a subroutine or heuristic template, to be supplemented with necessary additional material. Further, that the power and robusticity of our genes is due to their flexibility to produce a cumulatively measured phenotypic effect. These studies, however, omit any quantification of the total effect; to what extent is genetic explanation the whole explanation?

The question over number of necessary genes for behavior readily implies that given a complete understanding of DNA, one should be able to return a human being; that non-DNA differences do not contribute to the phenotype and are not inherited. However, it

is important to acknowledge that DNA does not need to code for the entire phenotype. Emergent properties from physical attributes, such as the properties of water that contribute crucially to the formation of the cell lining, are not directly coded for in DNA. Evolution is described as being parsimonious in creating genetic information. Given that cells are widely believed to have evolved before DNA, there is little reason that DNA would code for properties already present. For clarity, the point is that the structures of a flower, for example, may emerge in part from nature taking a 'free ride' from the properties that arise from the physico-chemical universe (non-DNA), and from DNA (Nottale & Auffray, 2008).

Additionally, it is clearly not true that DNA *does* indeed encode all biological systems. Firstly, an organism will also inherit much non-DNA information: a fully fertilized egg and maternal antibodies; and RNA, the centriole, and other non-DNA components from the sperm. Such information serves as a robust source of trans-generational phenotype plasticity, and as a trigger for development, without which life would not begin. However, difficulty may arise in comparison of these different forms of inheritance information. The eukaryotic cell is a very complex structure, and difficult to represent in multidimensional computational space. Yet, even when simplified to exclude redundancies, such as binning similar structural data into categories of organelles, comparable effective information content to that of the genome is estimated from 1% of structural information. Given the structural information is largely what constrains aforementioned protein interactions, it appears foolhardy to omit these data (Noble, 2011).

Studies that have explicitly researched non-DNA inheritance information have produced informative data on the development, particularly, on skeletal morphology. When a fertilized mouse egg is carried to term in a mouse of a different strain, the number of vertebrae of the born mouse is dependent on the mouse that bore it, not its genetic mother (McLaren & Michie, 1958). Similarly, more recent studies of cross-species cloning with fish of differing genera find that the skeletal morphology is a mix of both DNA and structural inheritance information. Goldfish eggs inserted into carp nuclei develop with vertebrae number congruous with carp, which is attributed to the egg cytoplasm regulatory effects on gene expression in early development (Sun et al., 2005). Lastly, surgical modifications made to alter the orientation of cilia patterns in paramecium, are robustly inherited by daughter cell across multiple generations (Beisson & Sonneborn, 1965). These data clearly identify patterns of non-DNA inheritance information on the phenotype. Although it is not readily comparable, such information makes a major contribution to the development and fitness of the organism. It is thus of benefit to consider that the extent of gene contribution to the phenotype is tempered by extant data from other sources.

Conclusion

Ultimately, little benefit appears from quantifying single gene effect to explain the phenotype. A brief survey of the literature highlights the wonderful flexibility of the genome in producing our phenotype; yet it also reveals that its power is greatly rooted in its complex interconnectivity. Admittedly, there are difficulties inherent in these integral structures for delimiting single gene contributions. However, that we can begin to appreciate the reach of this dynamic structure is comforting; identifying non-DNA inheritance information, at least, delimits the contribution of genetic factors, if also clouding the role of the gene. Future research should redress the gene as a functional template that is a valuable, but partial truth in a wider system of interactions. Irrespective, the benefit of our confusion is apparent as it represents both the dearth of genetic information yet to be explained, and also the Socratic precursor for progress.

References

Bassford, P., Beckwith, J., Berman, M., Brickman, E., Casadaban, M., Guarente, L., ... & Silhavy, T. (1980). Genetic fusions of the lac operon: a new approach to the study of biological processes. *Cold Spring Harbor Monograph Archive*, *7*, 245-261.

Beadle, G.W. and Tatum, E.L. (1941). Genetic control of biochemical reactions in Neurospora. Proc. Natl. Acad. Sci., *27*, 499–506.

Beckwith, J. R., & Zipser, D. (Eds.). (1970). *The lactose operon* (Vol. 1). Cold Spring Harbor Lab.

Beisson, J. & Sonneborn, T. M. (1965) .Cytoplasmic inheritance of the organization of the cell cortex in paramecium Aurelia. Proc. Natl Acad. Sci. USA, *53*, 275– 82.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., … Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome research*, *17*, 669–81.

Huang, X.-Q., Lui, S., Deng, W., Chan, R. C. K., Wu, Q.-Z., Jiang, L.-J., … Gong, Q.-Y. (2010). Localization of cerebral functional deficits in treatment-naive, first-episode schizophrenia using resting-state fMRI. *NeuroImage*, *49*, 2901–6.

Jeong, H., Mason, S. P., Barabási, a L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, *411*, 41–2.

Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. (2011). Epigenetics in alternative pre-mRNA splicing. Cell, *144*, 16–26

Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ. 2013. Alternative splicing:Apivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol 14: 153–165

McLaren, A. & Michie, D. (1958) An effect of uterine environment upon skeletal morphology of the mouse. Nature, *181*, 1147–1148.

Mercer, T. R., & Mattick, J. S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome research*, *23*, 1081–8.

Morgan, T.H., Sturtevant, A.H., Muller, H.J., and Bridges, C.B. (1915). The mechanism of Mendelian heredity. Holt Rinehart & Winston, New York.

Mudge, J. M., Frankish, A., & Harrow, J. (2013). Functional transcriptomics in the post-ENCODE era. *Genome research*, *23*, 1961–73.

Noble, D. (2011). Differential and integral views of genetics in computational systems biology. *Interface focus*, *1*, 7–15.

Noordermeer, D., LeleuM, Splinter E, Rougemont J, De LaatW, Duboule D. (2011).Thedynamic architectureofHoxgene clusters. Science, *334*, 222–225.

Nottale, L. & Auffray, C. (2008) . Scale relativity and integrative systems biology 2. Macroscopic quantum-type mechanics. Progr. Biophys. Mol. Biol., *97*, 115–157.

Pesole, G. (2008). What is a gene? An updated operational definition. *Gene*, *417*, 1–4.

Ruiz-Narvaez, E. A. (2011) What is a functional locus? Understanding the genetic basis of complex phenotypic traits. Med Hypotheses, *76*, 638-42

Shen S, Lin L, Cai JJ, Jiang P, Kenkel EJ, Stroik MR, Sato S, Davidson BL, Xing Y. (2011).Widespread establishment and regulatory impact of Alu exons in human genes. Proc Natl Acad Sci, *108*, 2837–2842

Sorek R. (2007). The birth of new exons: Mechanisms and evolutionary consequences. RNA, *13*, 1603–1608.

Stergachis, A. B. et al. 2013. Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution. *Science*, *342*, 1367-1372.

Sun, Y. H., Chen, S. P.,Wang, Y. P., Hu,W. & Zhu, Z. Y. (2005) . Cytoplasmic impact on cross-genus cloned fish derived from transgenic common carp (Cyprinus carpio) nuclei and goldfish (Carassius auratus) enucleated eggs. Biol. Reprod., *72*, 510–515.

Watson, J.D. and Crick, F.H.C. 1953. A structure of deoxyribonucleic acid. Nature, *171*, 964–967.