The molecular roots of evolution

Philip Ball

This is an extended version of my article "Celebrate the Unknowns" published in *Nature* **496**, 419-420 (2013). It acknowledges subsequent comments about precisely what Francis Crick said of the Central Dogma, namely that he stated it much more precisely than did Jim Watson, Marshall Nirenberg, or most textbooks.

Attempts to understand the complexity of the genome are complicating the picture of how DNA sequence relates to physiological function and heredity. Where does this leave the gene-centred view of evolution?

Next year's diamond jubilee of the discovery of DNA's molecular structure by Crick, Watson and their collaborators will rightly celebrate how, by showing that hereditary information can be encoded in the double helix, their work launched the genomic age. Yet the conventional narrative – in which those 1953 papers lead inexorably to the Human Genome Project and the dawn of personalized medicine tailored to our genes – is as misleading as the popular narrative of gene function itself, in which a DNA sequence is translated into a protein and ultimately into a phenotype. Both stories owe their tenacity to the alluring elegance of linear causality. But it's not that simple.

It's often said today that the more closely we look at the genome, the less we understand it. We should put it more baldly: we do not truly know what genes are or how they work. We don't know what most of our DNA is for, nor how (or to what extent) it governs organismal traits. At the very least, the 'blueprint' rhetoric that surrounded the Human Genome Project now looks seriously misleading. It's tempting to call DNA not a blueprint but a crib-sheet, comprehensible only with extensive, implicit contextualization – except that even this remains too close to the old list-of-parts picture that seems now profoundly undermined.

What often goes unremarked among the revisionism that current research is prompting is that DNA was supposed to be the missing part of the puzzle in evolutionary theory: the repository of Mendelian hereditary factors. If we don't actually understand DNA after all, then how can we expect to know how, at the molecular level, evolution works?

Here are just a few of the questions that seem increasingly to complicate the popular 'standard model' of genetics:

- only about 5% of the mammalian genome seems to be under selection, while a significantly higher proportion now appears to be functional in some sense
- genes operate in complex networks of interaction in which cause and effect are often hard to discern and certainly nonlinear

- the complexity of the human body arises from the complexity of the network and perhaps the functions of non-coding DNA, not from a surfeit of genes
- there is a serious shortfall between the known heritability of particular traits (such as disease susceptibility) and the genes that seem associated with it
- the forces directing the evolution of the mammalian genome act non-uniformly across the genome for reasons, and with consequences, that are not well understood[1]
- projects such as ENCODE[2] that interrogate the transcriptional status of the genome challenge the distinctions of coding and non-coding regions, and propose that the very definition of a gene is not clear
- mammalian genomes are replete with 'pseudogenes', which are inherited and transcribed but do not code for proteins, generally have no identifiable function and seem free from selection
- genes are regulated not just directly by other genes but by higher-order spatial organization of the chromosomes, which can provide a memory and computational system complementary to that encoded in the primary genomic sequence.[3]
- non-genetic information can be inherited, for example by epigenetic mechanisms that pass on regulatory chemical markers on gene loci, causing functional effects without any change in the primary nucleotide sequence.[4] An extreme example might be prions, which allow inheritance of environmentally acquired traits via the self-propagation of protein folds.[5]

All this is well known to researchers in the field, and indeed it is precisely because there is no consensus, and often heated argument, about what it all entails for the way cells and organisms work that the study of the molecular mechanisms of genetics and evolution is now so rich and exciting. The lacunae in this field should be emerging as a major narrative in biology, comparable in some ways to the disruptive discovery of 'dark energy' (or something like it) in cosmology. But so far, it is not. One could easily imagine, from popular discussions of evolution, that today's age of genomics is doing no more than filling in the gaps of the classical Neodarwinian synthesis and its post-Crick/Watson molecular embodiment. "The post-Watson-Crick synthesis gave a very neat picture, but one that was too strong, with a limited view of the potentiality of the genome", says Mark Gerstein of Yale University, one of the team leaders of ENCODE.

What do these developments mean for evolutionary theory? It remains beyond serious doubt that Darwinian natural selection, acting at the level of the phenotype, drives much and possibly most evolutionary change. Moreover, Mendel's principle of inheritable, discrete 'particulate factors', later called genes, can explain how this works. Further, the mathematical population genetics devised in the early to mid-twentieth century to account in broad terms for how genes spread remains largely secure. None of this depends on a particular microscopic notion of what a gene is.

What is not clear is how natural selection is effected at the molecular level: to put it crudely, 'where selection happens'. This fact is fudged in accounts that still present the gene as though it were a little autonomous stretch of DNA intent on copying itself. Take this recent description from a prominent popularizer:

> At stake is the level at which Darwinian selection acts… biologists with non-analytical minds warm to multi-level selection: let a thousand flowers bloom and let Darwinian selection choose among all levels in the hierarchy of life. But it doesn't stand up to serious scrutiny. Darwinian selection is a very particular process, which demands rigorous understanding… the gene doesn't belong in the hierarchy. It is on its own as a "replicator," with its own unique status as a unit of Darwinian selection.[6]

There is no hint here that the very notion of 'gene' has become problematic[7]. Yet as evolutionary biologist Michael Lynch of Indiana University says, "the gene has always been a much more fuzzy concept than we like to tell undergraduate classes". Some molecular biologists barely use the word at all these days. Perhaps more profoundly, the statement above neglects, with false confidence, the fact that there is no comprehensive theory of how phenotype – *the only level at which natural selection can plausibly act* – is related to genotype. "It's striking how, for some biologists, a 70 to 80-year-old theory is still seen as the 'latest and greatest'", says Andreas Wagner, an evolutionary biologist at the University of Zurich. "That theory, which assumes that genes interact simply and additively, is outdated, but it was successful and to some degree very effective."

A part of the problem is a common one in science: an addiction to old, almost colloquial words (like 'life' and 'force') that retain heuristic value while being scientifically imprecise. It is hard to avoid speaking of the genotype and phenotype as though these are well established phenomena that cohere at a particular point in the complexity of an organism. But there are many phenotypes, from protein to limb – and some argue, to social group – each with its own mapping to genotype. And as Wagner says, "what ENCODE shows is that, even if you look just at the molecular phenotype, the complexity is already staggering".

Life beyond genes

The ENCODE project set out systematically to map out which parts of the human chromosome are transcribed, how transcription is regulated, and how this is affected by chromatin structure and histone modification.[2] It has revealed that there is much more to genome function than is encompassed in the roughly 1 percent that contains our 20,000 or so genes. It concludes that at least 80 percent of the genome is transcribed, much of which lies outside protein-coding regions. It is not clear how much of this transcription relates to actual function – some geneticists and evolutionary biologists point out that just because genomic regions are transcribed doesn't guarantee that the RNA transcripts serve any biological role.[8] But the ENCODE team argues that at least some of those transcripts could provide a reservoir of molecules that might find regulatory functions: in other words, a pool of potentially useful variation. Such findings lead them to propose that

> the transcript be considered as the basic atomic unit of inheritance… the term gene would then denote a higher-order concept intended to capture all those transcripts (eventually divorced from their genomic locations) that contribute to a given phenotypic trait.[9]

This rather incendiary proposal has elicited mixed responses. "The statement that the transcript is the unit of heredity is clearly ludicrous because transcripts are not inherited", says evolutionary biologist Patrick Phillips of the University of Oregon. "This controversy may have more to do with the dual roles of DNA as both the units of inheritance and as the starting point for most molecular processes. It retains both of these roles because of its capacity to store information. The two roles do need to be separated for understanding both functional molecular biology and its implications for evolutionary change." That is perhaps the crux of the matter, since it is precisely those dual roles – DNA as genotype and as the wellspring of phenotype – that have made DNA the focus of the traditional view of genetics.

The fact that some non-protein-coding regions of the genome encode RNAs with regulatory functions has been known for decades. But the full complexity and diversity of non-translated RNAs has only emerged much more recently, and the ENCODE results seem to take it to a new level. Certainly it seems clear now that the tiny proportion that codes for proteins and is thought to be under selection is only a small part of what is actually transcribed. One emerging view is that much of the additionally transcribed material is also under selection, but very weakly so. But that remains disputed – some feel that the transcription of non-coding material may be noise, irrelevant to function and evolution.

One particular component of this 'dark matter' is pseudo-genes: sequences that may be inherited but are non-functional, non-coding and of uncertain selective status. They are often recognizable by their genetic disablements, such as stop codons in the midst of the sequence. We don't know what pseudogenes are; it's likely that they are not a single class of objects at all. Some might be merely 'dying genes' which have lost their function, or transcripts inadvertently reverse-transcribed back into the genome at some random position but stillborn for lack of promoter elements to activate them: old-style 'junk DNA', in other words. Others do seem to have regulatory roles, probably via RNA interference – which means strictly speaking that they are not really pseudogenes at all.[10]

According to Gerstein, the ENCODE results suggest that, in contrast to the conventional Watson/Crick view, there is a whole level of adaptation and selection that operates beyond the level of proteins. "The complexity that is constantly being revealed by projects such as ENCODE are telling us that we really don't understand the genotype-phenotype map at a functional level", Phillips says. "This in turn means that we don't really understand how the specifics of how evolutionary forces will influence any given genome, even if we have a very good handle on how forces act in general."

The ENCODE findings are by no means the only results that are unsettling old assumptions about genes and evolution. They add to a growing perception that the regulatory network of RNA molecules might provide a pool of potential

variation and a locus for adaptation to the environment that does not require the riskier gambit of altering protein-coding sequences in DNA.[11] According to John Mattick of the Garvan Institute of Medical Research in Sydney, Australia,

> Adaptation by selecting from random mutations of DNA is a very slow process for complex organisms like us with low numbers of progeny and long generation times. So it may become advantageous to evolve strategies for directing mutations to more productive (regulatory) areas, and reduce the risk of damaging consequences. It may have also been advantageous to evolve an RNA-based 'adaptation network' that can adapt to the prevailing environment, without such adaptations needing to be inscribed in the DNA system that encodes it. In that case, there is reduced selective pressure on the DNA itself; environmental adaptation can be accommodated in the plasticity of an RNA regulatory network, with the possibility that this may be inherited and ultimately re-written to the DNA.

Regulation of genes can also be effected via epigenetic molecular alterations to DNA, such as the addition of a methyl group or modification of histones in chromatin. Many of these regulatory chemical markers are inherited, including some that govern susceptibility to diabetes and cardiovascular disease[4]. Genes may also be regulated by the spatial organization of the chromosomes, which is in turn affected by epigenetic markers[3]. Although such effects have long been known, it has been claimed that their prevalence and persistence may be much greater than was previously thought, and so it is no longer unreasonable to suggest that they might have adaptive and evolutionary significance.[12] Again, the significance of epigenetics in the function of cells and evolution of organisms is disputed; some say that histone-modifying enzymes cannot be involved in gene regulation, or at least not selectively, because they lack sequence specificity [Ptashne].

Some researchers even dispute that, once one reaches the molecular level, natural selection is still the dominant mechanism of evolutionary change. Lynch asserts that "at the molecular level, we definitely do not agree that evolution happens largely by natural selection." He argues that random genetic drift can play a major role in the evolution of genomic features such as non-coding introns, especially in small populations where such effects (due to the imperfect operation of natural selection) are stronger – and that includes humans. Lynch has shown that natural selection does not in any case necessarily lead to lasting enhancement of fitness. On the contrary, it can generate cellular complexity – a redundant accumulation of molecular pathways – that is at best evolutionarily and potentially burdensome and non-adaptive.[13] He points out that "ingrained belief in the extraordinary power of selection" tends to promote the view that biological imperfections at the molecular level are the consequence of compromises that selection must make when two traits share a genetic basis (the phenomenon of pleiotropy), rather than their perhaps being an inevitable result of the molecular basis of evolution. Such work underscores Phillips' claim that "understanding the short-term response to selection does not tell us how and/or why complex systems accumulate as they do over time."

Hidden in the net

Pleiotropy and the more general notion of epistasis (interactions between genes) could in fact be regarded as traditionalist inklings of the complications caused by the networked character of the genome. But these no longer look like special cases; they are probably the norm, and they impose constraints on genetic evolution that are poorly understood and hard to discern. Genes that are deeply embedded in networks, for example, cannot easily be mutated without highly nonlinear and perhaps catastrophic consequences. "Our current formulations of evolutionary models really can't deal very effectively with complex regulatory networks", says Phillips. "Evolutionary theory has been completely derelict in making predications regarding what kind of genetic networks we should be expecting to see emerging from these functional genomic studies."

Nevertheless, developmental biologists have long debated to what extent the complexity and apparently modular operation of the genome – for example, in the patterning mechanisms of *Hox* genes – impose constraints on the kinds of changes that natural selection can induce. And it's not just about constraint: complexity can also generate opportunity, new possibilities for change that could not be achieved via a linear genotype-to-phenotype progression. Some argue that mutations may be accumulated in cryptic form, free from selective 'weeding', thanks to the robustness of networks in maintaining a particular phenotype: in other words, the complexity of the genome provides a degree of 'evolutionary capacitance'. This hidden variation might be unmasked by some new environmental stress that brings selective pressure to bear, and the genetic 'preadaptation' could then unleash adaptations for which the multiple mutations could not have happened gradually under selective pressure, enabling the organism to leap across evolutionary peaks. Some of these ideas are venerable, but they are constantly acquiring new wrinkles. It has been proposed, for example, that the response of the chaperone Hsp90 to environmental change might alter the strength of the coupling of genotype to phenotype and thus release cryptic variation.[14] The potentially major implications of how such genetic preadaptation influences evolution lack any consensus.

Underlying all these matters is the question of whether there are likely to be any general principles of how evolutionary change happens at the molecular level – something, that is, to replace the old notion that 'DNA mutates, and 'good' mutations spread'. Of course, both of those statements remain true in themselves, and sometimes their connection is transparent and closely documented. But in general the link between a miscopied nucleotide during meiosis and an organism that is more or less 'fit' is extremely murky and ineluctably complicated: evolution negotiates many paths, making it up as it goes.

That needn't be as haphazard as it sounds. One thing evolution must provide is robustness: a capability to survive in a changing environment, whether that entails temperature variations, lack of nutrients or water, or unknown predators. This isn't something that a gene mutation confers: that alone might, on rare occasions, offer improved prospects to one particular challenge, but with the loss of the previous function. So it seems likely that evolutionary innovation doesn't happen this way. Rather, a particular phenotype, arising from a complex network of gene interactions, is likely to have a great many genotypic 'solutions', and so a

new phenotype might appear without the loss of the existing one.[15] In this way, organisms may acquire robustness to a wide range of environmental changes, as well as sequentially specializing for particular circumstances. That's not a property of any DNA sequence per se, but of the higher levels of gene organisation. To put it another way, mutations on DNA are simply means to the end that allows regulatory, metabolic or other networks to produce new phenotypes in a robust, redundant manner. This is still, in a sense, natural selection from among random mutations, but the DNA sequence is then the 'unit of evolution' only in much the same way as notes and chords are the units of Beethoven's Kreutzer sonata – 'meaning' resides only in their combinations. To think, then, that evolution might be comprehended by laying homologous sequences side by side is misguided.

One consequence of this more complex picture of the genotype–phenotype relationship is that it imposes some clear constraints on the forms that mutations can viably produce. Such constraints are evident in reality – for example, mutagenesis experiments in *Drosophila* in the 1980s elicited only a small number of variants in embryonic segmentation patterns. In particular, evolutionary stasis, which is well attested the fossil record, can result from the phenotypic uniformity of a large number of closely related genotypes – it might take a long time for mutations to 'find' the genotypic network of a superior phenotype, triggering a burst of innovation.[16] Here natural selection is, in effect, temporarily neutered by an inability for random mutation to generate any diversity from which to select.

Deal with it

In short, the current picture of how and where evolution operates, and how this shapes genomes, is something of a mess, with many voices saying many things, some overlapping, some contradictory. That's not a criticism, but a vote of confidence in the healthy state of molecular and evolutionary biology after what seems like several decades of the somewhat complacent view that its all over bar the sequencing. Just as there has never been a more exciting time to be a cosmologist, so that should be true in genetics. Yet barely a whisper of this vibrant debate reaches the public, who have been fed assurances that, on the one hand, DNA is as solipsistic a blueprint as ever, and on the other, genomics is steadily answering questions rather than raising them. Even the ENCODE findings, although they enjoyed wide media coverage, were widely presented as "there's a thing", with little hint the implications for how evolution and genomes work.

Why is this? I can offer only speculations. One is sentimentality. Biology is so complicated and plagued by exceptions that it is deeply painful to relinquish the apparent promise, after Watson and Crick, of simplicity (and double-helical elegance) at its core.

Besides, the change in narrative is much more easily grasped for dark energy – compelled by a single astonishing revelation – than for molecular evolution, where old never-quite-settled arguments about such issues as where and how widely adaptation occurs are now colliding with new questions about non-

coding RNA, epigenetics and genomic network theory. It's not clear where to look, or which story to tell.

And the rhetoric surrounding the Human Genome Project has raised the stakes too high. By appearing to promise knowledge of "the instructions to make a human" rather than stressing its immense practical value, the project yoked itself to a notion of genetic function and evolution that now seems questionable. It is one thing to have to revise our ideas about the cosmos, another to admit we don't understand ourselves as well as we'd thought.

There may also be anxiety that admitting any uncertainty about the mechanisms of evolution will be exploited by those who seek to undermine it. Certainly, popular accounts of epigenetics and the ENCODE results have been much more coy about the evolutionary implications than the developmental ones. But we are grown up enough to be told about the doubts, debates and discussions that are leaving the putative Age of the Genome with more questions than answers. Tidying up the story risks not only bowdlerizing the science but creating straw men for its detractors. As Lynch has said,[17] simplistic portrayals of evolution can simply encourage equally simplistic demolition.

When the structure of DNA was first deduced, it seemed to supply the final part of a beautiful puzzle whose solution began with Darwin and Mendel. The beauty and simplicity of that picture has proved too alluring. For its jubilee, we should do DNA a favour and lift some of the awesome responsibility for evolution's complexity from its shoulders.

References

1. R. H. Waterston *et al*., *Nature* **420**, 520-562 (2002).
2. ENCODE special issue, *Nature* **489**, 45-113 (2012).
3. Prohaska, S. J. *et al*., *J. Theor. Biol.* **265**, 27 (2010).
4. Jablonka, E. & Raz, G., *Q. Rev. Biol.* **84**, 131–76 (2009).
5. Halfman, R. & Lindquist, S. *Science* **330**, 629-632 (2010).
6. Dawkins, R. *Prospect* June (2012).
7. Stadler, P. F., Prohaska, S. J., Forst, C. V. & Krakauer, D. C. *Theory Biosci.* **128**, 165-170 (2009).
8. Doolittle, W. F. *Proc. Natl Acad. Sci. USA* **110**, 5294–5300 (2013).
9. Djebali, S. *et al.*, *Nature* **489**, 101-108 (2012).
10. Sasidharan, R. & Gerstein, M., *Nature* **453**, 729-731 (2008).
11. Mattick, J. S. *Proc. Natl Acad. Sci. USA* **109**, 16400–16401 (2012).
12. Mattick, J. S. *FEBS Lett*. **585**, 1600–1616 (2011).
13. Lynch, M., *Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.1216130109 (2012).
14. Jarosz, D. F. & Lindquist, S. *Science* **330**, 1820 (2010).
15. Wagner, A. *Trends in Genetics* **27**, 397-410 (2011).
16. Wagner, A. *Trends Ecol. Evol.* **26**, 577-584 (2011).
17. Lynch, M. *Nature* **435**, 276 (2005).