Journal of Biology
J. Biol.
BioMed Central

Opinion

Evolutionary genomics and the reach of selection

Laurence D Hurst

Address: Department of Biology and Biochemistry, University of Bath, Bath, Somerset BA2 7AY, UK. Email: l.d.hurst@bath.ac.uk

Published: 23 February 2009

Journal of Biology 2009, **8**:12 (doi:10.1186/jbiol113)

The electronic version of this article is the complete one and can be
found online at http://jbiol.com/content/8/2/12

Abstract

Unexpected findings in evolutionary genomics both question the role of selection in genome evolution and clarify how genomes work.

Why is studying the way that genes and genomes evolve interesting? There are many generally accepted answers. Looking for places in a genome that are highly conserved is an efficient means to locate functionally important sequences, usually genes or gene regulatory domains. Conversely, unusually fast-evolving sequences can suggest where Darwinian selection might have acted to cause important differences between species. We can discover which gene families can be easily expanded or lost, which species are related to which others, and where genes have been transferred horizontally between species rather than being transmitted by descent. But if you ask me what I think is especially interesting about evolutionary genomics then let me give a bit of history.

In the 1970s and 80s there was a large school of evolutionary biology, much of it focused on understanding animal behavior, that to a first approximation assumed that whatever trait was being looked at was the product of selection. Richard Dawkins is probably the most widely known advocate for this school of thought, John Maynard Smith and Bill (WD) Hamilton its main proponents. The game played in this field was one in which ever more ingenious selectionist hypotheses would be put forward and tested. The possibility that selection might not be the answer was given short shrift.

By contrast, during the same period non-selectionist theories were gaining ground as the explanatory principle for details seen at the molecular level. According to these models, chance plays an important part in determining the fate of a new mutation - whether it is lost or spreads through a population. Just as a neutrally buoyant particle of gas has an equal probability of diffusing up or down, so too in Motoo Kimura's neutral theory of molecular evolution an allele with no selective consequences can go up or down in frequency, and sometimes replace all other versions in the population (that is, it reaches fixation). An important extension of the neutral theory (the nearly-neutral theory) considers alleles that can be weakly deleterious or weakly advantageous. The important difference between the two theories is that in a very large population a very weakly deleterious allele is unlikely to reach fixation, as selection is given enough opportunity to weed out alleles of very small deleterious effects. By contrast, in a very small population a few chance events increasing the frequency of an allele can be enough for fixation. More generally then, in large populations the odds are stacked against weakly deleterious mutations and so selection should be more efficient in large populations.

In this framework, mutations in protein-coding genes that are synonymous - that is, that replace one codon with another specifying the same amino acid and, therefore, do not affect the protein - or mutations in the DNA between genes (intergene spacers) are assumed to be unaffected by selection. Until recently, a neutralist position has dominated thinking at the genomic/molecular level. This is indeed reflected in the use of the term 'junk DNA' to describe intergene spacer DNA.

These two schools of thought then could not be more antithetical. And this is where genome evolution comes in. The big question for me is just what is the reach of selection. There is little argument about selection as the best explanation for gross features of organismic anatomy. But what about more subtle changes in genomes? Population genetics theory can tell you that, in principle, selection will be limited when the population comprises few individuals and when the strength of selection against a deleterious mutation is small. But none of this actually tells you what the reach of selection is, as *a priori* we do not know what the likely selective impact of any given mutation will be, not least because we cannot always know the consequences of apparently innocuous changes. The issue then becomes empirical, and genome evolution provides a plethora of possible test cases. In examining these cases we can hope to uncover not just what mutations selection is interested in, but also to discover why, and in turn to understand how genomes work. Central to the issue is whether our genome is an exquisite adaption or a noisy error-prone mess.

The contest between function and noise

Consider, for example, the problem of transcription. Although maybe only 5% of the human genome comprises genes encoding proteins, the great majority of the DNA in our genome is transcribed into RNA [1]. In this the human genome is not unusual. But is all this transcription functionally important? The selectionist model would propose that the transcription is physiologically relevant. Maybe the transcripts specify previously unrecognized proteins. If not, perhaps the transcripts are involved in RNA-level regulation of other genes. Or the process of transcription may be important in keeping the DNA in a configuration that enables or suppresses transcription from closely linked sites.

The alternative model suggests that all this excess transcription is unavoidable noise resulting from promiscuity of transcription-factor binding. A solid defense can be given for this. If you take 100 random base pairs of DNA and ask what proportion of the sequence matches some transcription factor binding site in the human genome, you find that upwards of 50% of the random sequence is potentially bound by transcription factors and that there are, on average, 15 such binding sites per 100 nucleotides. This may just reflect our poor understanding of transcription factor binding sites, but it could also mean that our genome is mostly transcription factor binding site. If so, transcription everywhere in the genome is just so much noise that the genome must cope with.

The problem of alternative transcripts is very similar. In the original view of the gene, one gene made one transcript, which made one protein. For many organisms (such as bacteria and yeast) this model is still pretty good. For us it isn't. Latest estimates suggest that the vast majority of human protein-coding genes can make many different (alternative) messenger RNA molecules from a single transcript. In no small part this is achieved by the cleavage and splicing of one transcript in many different ways, each producing a different set of protein-coding sections (exons), the non-coding sections (introns) being removed (Figure 1). But why this richness? Again, a good case can be made for both the selectionist and the noise view.

A selectionist model would suppose that each transcript has a role and is made when and where it is needed. In *Drosophila*, different splicing of transcripts from genes in the sex-determination pathway in males and females is central to the establishment of sex differences in development, suggesting that in this case exact coordination is critical. Similarly, many human genes are differently spliced in neurons, so producing ion channels with different sequences and different biophysical or regulatory properties. Alternatively, splicing may be inherently error-prone and many of these alternatives may be just so much rubbish. Again a defense can be given. The human genome is unusual in having many and large introns. Finding small exons in the sea of non-protein-coding material must be a formidable computational task for our cells and hence is potentially error-prone.

Recently, some evidence has been presented to support the noisy splice model. The single-celled protist *Paramecium* has short introns, some of which contain stop codons. Interestingly, introns that are a multiple of three nucleotides long are much more likely to contain a stop than those that are not [2]. Why might this be? *Paramecium*, like other eukaryotes, has a system called nonsense-mediated decay that eliminates mRNAs that contain a stop codon where they should not have one - it is, in effect, a quality-control mechanism. As codons are three nucleotides long, if an intron that is not a multiple of three long is not removed, it will induce a change in the reading frame (a frameshift) and is likely to make an mRNA with an out-of-place stop codon; this mRNA will be removed by the quality-control system (Figure 1a). One that is a multiple of three, however, will not induce a frameshift (Figure 1b). To remove these transcripts would

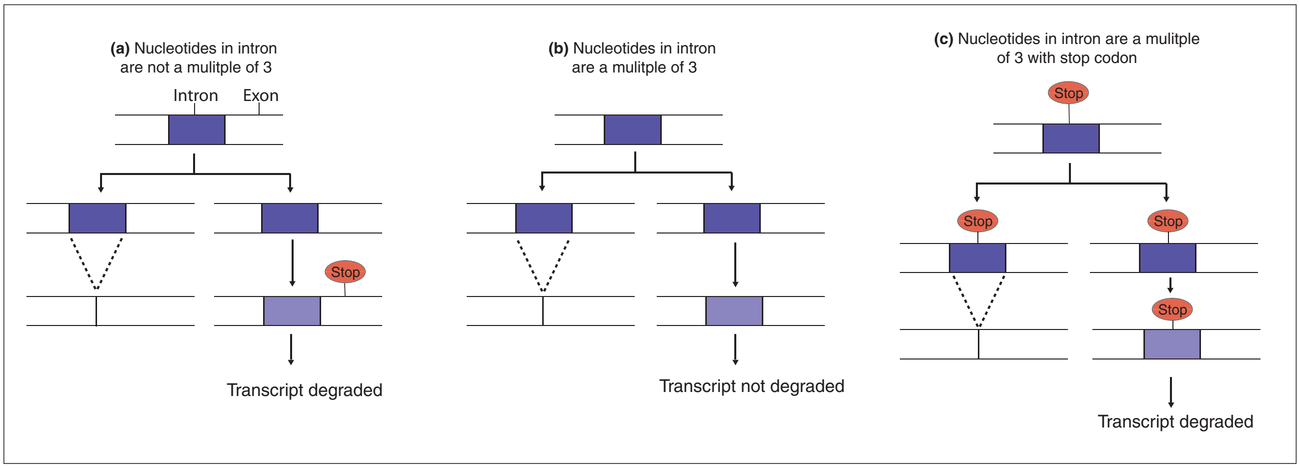
(a) Nucleotides in intron are not a mulitple of 3
Intron
Exon
Stop
Transcript degraded
(b) Nucleotides in intron are a mulitple of 3
Transcript not degraded
(c) Nucleotides in intron are a mulitple of 3 with stop codon
Stop
Stop
Stop
Stop
Transcript degraded

Figure 1

The protist Paramecium has short introns in which some contain stop codons. Introns in multiples of threes are more likely to contain a stop codon as a fail-safe measure for correct splicing. **(a)** The failure in removal of an intron that is not a multiple of three will cause a frameshift and this will most likely introduce an out of place stop codon in the resulting mRNA. This transcript can then be degraded by nonsense mediated decay (NMD). **(b)** When an intron that is a multiple of three long is not removed, it will not cause a frameshift and therefore the mis-spliced transcript will not be degraded. **(c)** This can be overcome by having stop codons in introns of multiple of three. Therefore when the intron is not removed, NMD can act on the incorrectly placed stop codon and remove the transcript.

require a stop codon in the intron as a fail-safe measure (Figure 1c). The excess of stop codons in introns that are multiples of three is hence parsimoniously explained if we suppose splicing to be inherently error-prone.

A very direct measure of noise is variation in the expression level of the same gene in many otherwise identical cells. For some genes, there is a lot of variation, given the mean abundance, for others much less. In yeast, for example, 'essential' proteins (those whose absence is lethal) tend to have low-noise expression [3]. Other proteins, notably those for the import of metabolites from the environment into the cell, tend to be very noisy. Is this between-gene variation in noise itself adaptive? *A priori*, essential genes would be expected to be tightly regulated and to have low variation in expression: if levels of the protein accidentally slip too low, the cell might die. Does the noise of highly noisy genes exist to enable a response to a variable environment - or because selection doesn't care?

Two favorite examples from my laboratory bear on issues on which confident neutralist statements were commonplace: that selection will not affect synonymous mutations in mammals and that the location of genes in the genome is irrelevant. We found that in regard to synonymous mutations the strict-neutralist position is hard to substantiate in mammals, but for previously unrecognized reasons. Mammalian genes are unusual in having a very low ratio of coding sequence to intronic DNA. This presents our cellular machinery with an unusual problem, namely correctly identifying the ends of numerous small exons. The solution mammals appear to have reached is to allow a specific class of proteins (SR proteins) to bind in immature RNA to exonic splice enhancer (ESE) motifs, these being located predominantly at the ends of exons [4]. The need to specify these motifs, however, ensures that many synonymous mutations are under probably strong selection, as failure of splicing is potentially highly deleterious [5]. Indeed, upwards of 40 diseases are associated with synonymous mutations that disrupt splicing [5]. Both the choice of which codon to use and rates of evolution of synonymous sites [6] are affected by the need to specify ESEs.

The issue of gene location gets to the heart of the relationship between genome organization and the control of gene expression. The simplest model supposes that a gene with its relevant upstream control elements is enough to dictate expression of that gene. Those working on transgenes (genes inserted by researchers into a genome) know from experience that this is a limited model, as most new transgene inserts will not be expressed appropriately, if at all. There is now abundant evidence that within a genome, genes with similar expression patterns cluster together [7] - that is, they are syntenic. Whether this reflects selection or noise remains the key issue. A simple model supposes that most DNA in any given cell type is packaged in such a way as to be largely unavailable for transcription. The unpacking of the DNA to enable expression from one

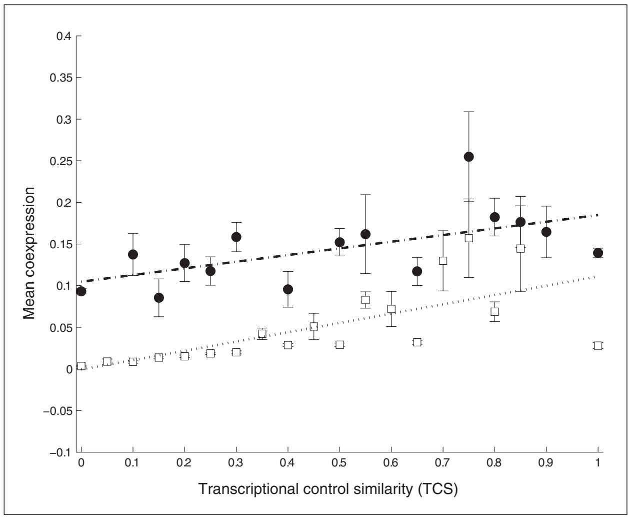

Figure 2
Influence of genomic co-localization on gene coexpression as a function of similarity in transcriptional control. Transcriptional control similarity (TCS) is a measure of the similarity in the suite of transcription factors that regulate a pair of genes. A score of zero means no similarity, a score of one means the very same transcription factors regulate the two genes. Mean levels of coexpression of neighboring genes are shown as black circles and of non-neighbors as white squares; error bars represent standard error of the mean. Note that neighboring genes with no transcription factor control similarity (TCS = 0) have, on average, the same level of coexpression as two unlinked genes with TCS = 1. Adapted from [9].

gene can potentially influence, by accident, the expression of neighboring genes. Indeed, the transcription rate of a transgene corresponds to that of the genes adjacent to the position in which it is inserted [8].

The alternative possibility is that genes expressed together are in close vicinity because selection has favored specific patterns of coordinated expression. Comparative genomics can help resolve this issue. Do coexpressed genes tend to be preserved in synteny more than expected, as predicted by a selectionist model? To a limited degree this can be the case [7]. However, we also find that neighboring genes have more coordinated expression patterns (when one gene is upregulated the neighbor is as well; when downregulated they tend to be downregulated in concert) than expected simply because of being next to each other on the chromosome [9]. This can also be shown experimentally: two transgenes are coexpressed when inserted adjacent to one another but not when inserted in genomically different locations [10]. The quantitative extent to which this is true is striking. On average, genes that are regulated by completely different sets of transcription factors have, if the genes are neighbors, the same degree of coexpression as a pair of unlinked genes that have exactly the same set of transcription factors regulating them (Figure 2) [9].

Genomic noise abatement: a new view of gene and genome evolution?

What I find so tantalizing about these issues is three-fold. First, the facts so often conflict with our prior assumptions: the very fact of widespread transcription conflicts with the previous assumption that DNA that was not protein-coding must be silent junk. Second, both the 'perfectly-formed-genome' model and the 'noisy-rubbish' model look reasonable given what we know about the mechanism of gene expression. For example, that RNA can function as a regulatory molecule is not in question. The issue is whether this explains the vast amounts of transcription. Finally, no matter which answer is right, we will have learned something profound and new about how genomes function. They may be vastly more organized than often supposed, or they may be error-prone machines with a potential problem of unwanted transcripts.

This last issue opens up an important new avenue and way of thinking about genomes. The selection operating in genomes may not be so much to optimize gene function as to minimize the consequences of its inherently error-prone nature. Put differently, if genomes are subject to error-prone transcription, splicing and translation, then this would create the conditions for the evolution of quality-control and noise-abatement measures. I have already mentioned nonsense-mediated decay as a suggested quality-control mechanism. The richness of ESEs in genomes with small exons and large introns is parsimoniously explained as a result of selection for splice noise reduction. Recently, I and my colleagues speculated that as expression noise is likely to be lower in genomic domains that always have DNA accessible for transcription, this could explain why essential genes cluster together in the genome [11]. This is consistent with the finding that in yeast the chromosome ends, which are domains of very high expression noise, are home to an order of magnitude fewer essential genes than elsewhere on the chromosomes.

Where next: why evolutionary genomics should go extinct

Not so long ago molecular genetics and evolutionary genetics were typically considered two distinct disciplines largely not talking to each other. Now the two need each other more than ever and this trend can only continue. To really understand how genomes evolve, we need more than the statistical tests for selection provided by the past three decades of population genetics research. We need to understand the mechanisms of gene transcription, of splicing, of translation, regulation, repair and recombination, these details being provided by molecular biology. Indeed, convincing demonstration of selection on synonymous mutations required specification of the mechanism of accurate splicing, the standard statistical tests being indecisive. Conversely, for molecular geneticists the tool-kit of evolutionary genomics is ever expanding: multisequence alignment, phylogenetic reconstruction, tests for selection, DNA footprinting and so on. The ultimate success of evolutionary genomics will be its demise, not because its tools and techniques are not needed, but rather because they are so integral that they are simply part of one field, a sort of post-post-genomics.

References

1. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization.** *Nat Rev Genet* 2007, **8:**413-423.
2. Jaillon O, Bouhouche K, Gout JF, Aury JM, Noel B, Saudemont B, Nowacki M, Serrano V, Porcel BM, Ségurens B, Le Mouël A, Lepère G, Schächter V, Bétermier M, Cohen J, Wincker P, Sperling L, Duret L, Meyer E: **Translational control of intron splicing in eukaryotes.** *Nature* 2008, **451:**359-362.
3. Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS: **Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise.** *Nature* 2006, **441:**840-846.
4. Fairbrother WG, Holste D, Burge CB, Sharp PA: **Single nucleotide polymorphism-based validation of exonic splicing enhancers.** *PLoS Biol* 2004, **2:**E268.
5. Chamary J-V, Parmley JL, Hurst LD: **Hearing silence: non-neutral evolution at synonymous sites in mammals.** *Nat Rev Genet* 2006, **7:**98-108.
6. Parmley JL, Chamary JV, Hurst LD: **Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers.** *Mol Biol Evol* 2006, **23:**301-309.
7. Hurst LD, Pal C, Lercher MJ: **The evolutionary dynamics of eukaryotic gene order.** *Nat Rev Genet* 2004, **5:**299-310.
8. Gierman HJ, Indemans MHG, Koster J, Goetze S, Seppen J, Geerts D, van Driel R, Versteeg R: **Domain-wide regulation of gene expression in the human genome.** *Genome Res* 2007, **17:**1286-1295.
9. Batada NN, Urrutia AO, Hurst LD: **Chromatin remodelling is a major source of coexpression of linked genes in yeast.** *Trends Genet* 2007, **23:**480-484.
10. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S: **Stochastic mRNA synthesis in mammalian cells.** *PLoS Biol* 2006, **4:**e309.
11. Batada NN, Hurst LD: **Evolution of chromosome organization driven by selection for reduced gene expression noise.** *Nat Genet* 2007, **39:**945-949.