PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Transcriptional Noise and the Evolution of Gene Number

Adrian Bird and Susan Tweedie

Phil. Trans. R. Soc. Lond. B 1995 **349**, 249-253
doi: 10.1098/rstb.1995.0109

Transcriptional noise and the evolution of gene number

ADRIAN BIRD AND SUSAN TWEEDIE

Institute of Cell and Molecular Biology, University of Edinburgh, Kings Buildings, Edinburgh EH9 3JR, U.K.

SUMMARY

Several proposals are made to explain the apparent increase in complexity of certain lineages during evolution. The proposals (not made in this order) are: (1) that gene number is a valid measure of biological complexity; (2) that gene number has not increased continuously during evolution, but has risen in discrete steps; (3) that two of the biggest steps occurred at the transition from prokaryotes to eukaryotes and the transition from invertebrates to vertebrates; (4) that these steps were made possible by 'systemic' changes in the way that genetic information is managed in the genome; (5) that the ability to silence inappropriate promoters is the primary limitation on gene number; (6) that the invention of nucleosomes (and perhaps the nuclear membrane) facilitated the evolution of eukaryotes from prokaryotic ancestors; (7) that the spread of low density methylation throughout the genome facilitated the evolution of vertebrates from invertebrate ancestors.

DNA METHYLATION AND TRANSCRIPTIONAL NOISE

To understand organisms at the molecular level, it is not enough to list the component molecules and their interactions with one another. We also need a historical explanation of how things got to be that way. Indeed, for some biological problems an evolutionary approach, involving comparative analysis of homologous systems in related organisms, is an essential starting point for understanding. DNA methylation is a case in point, as the study of vertebrate genomes alone gives us an unrepresentative view of its biological role. It has been known for some time that the extensive genomic DNA methylation seen in vertebrates is exceptional (Bird & Taggart 1980). Methylation of invertebrate genomes is confined to a small fraction of the genome, and in some cases (e.g. *D. melanogaster* and *C. elegans*) may be absent altogether. The data on what sequences are methylated in the invertebrates are incomplete, but there is reason to believe that the main targets are transposable elements and other potentially damaging DNA sequences that have been silenced by methylation (discussed in Bird 1993). Genes do not appear to be methylated in invertebrates. Thus the primary function of DNA methylation in these organisms is apparently protection of the genome by neutralizing of potentially disruptive elements (Bird 1993). In vertebrates, by contrast, the genome as a whole is heavily methylated, and most genes are methylated to some extent.

A surprising aspect of the spread of DNA methylation through the vertebrate genome is the resulting close proximity between genes and methyl-CpG. This brings with it a significantly increased mutational load, as 5-methylcytosine is a potent mutagen (Bird 1980; Jones *et al*. 1992). Direct evidence for the adverse consequences of methylation is provided by an analysis of mutations that give rise to human genetic diseases. Over one-third of point mutations involve the conversion of CpG to TpG (Cooper & Youssouffian 1988). As a consequence of its instability, the density of CpG in the genome as a whole is greatly reduced, being stabilized at approximately one-quarter of the frequency that would be expected on a random basis (Josse *et al*. 1961; Sved & Bird 1990). This means that the dinucleotide CpG occurs about once every 100 bases throughout the genome. Since most CpGs are methylated, this is also the approximate density of methylation in the genome.

Could low density methylation of this kind exert a potentially beneficial effect that may counterbalance the deleterious consequences of mutation? It has been known for some time that CpG methylation causes transcriptional repression (reviewed in Bird 1993). Experiments in our laboratory have implicated proteins that bind specifically to methylated DNA as transcriptional repressors (Boyes & Bird 1991, 1992). One such protein, MeCP1, interacts with DNA with an affinity that is proportional to the density of methyl-CpGs, and appears to compete with transcription factors for access. The outcome of the competition depends upon the density of methylation (the higher the density, the more likely MeCP1 is to win and repress) and the strength of the promoter (the stronger the promoter, the more likely transcription is to occur). The winner takes all, as methylated genes are either transcribed at the full rate or are highly repressed. Intriguingly, the density of methyl-CpG that prevails in the genome (one per 100 base pairs) is sufficient to repress weak promoters without affecting strong ones. For example, the γ-globin gene in HeLa cells (where its promoter is weak) can be totally repressed by methylation of the few CpGs that it has, but this same level of methylation is compatible with maximal transcription if the promoter is strengthened (Boyes & Bird 1992). Thus low density methylation has the characteristics of a noise reduction system that eliminates background levels of transcription but leaves authentic transcription unaffected. The suggestion here is that this characteristic provides the primary evolutionary advantage of genome-wide methylation in the vertebrates (see also Bird 1995).

GENE NUMBER AT THE INVERTEBRATE–VERTEBRATE TRANSITION

Another apparent difference between vertebrates and invertebrates concerns gene number. It can be seen from table 1 that estimate for eukaryotes generally fall into two categories (Bird 1995). Vertebrate estimates are between 50000 and 100000 genes per genome, whereas no other eukaryotic organism has been found to exceed 25000 genes. Table 1 should not be taken at face value, as some estimates of gene number are not accurate. Nevertheless, it is striking that vertebrate gene numbers are so much higher than the others. We shall consider the idea that the high gene number in vertebrates is not part of a continuum among eukaryotic organisms, but represents a quantal upward shift at the invertebrate/vertebrate boundary. The capacity for more genes in the vertebrates is proposed to be a direct consequence of the cooption of DNA methylation as a novel mechanism for reducing transcriptional noise.

Why should there be a relation between noise reduction and gene number? A hypothetical answer can be deduced from analysis of gene expression patterns in vertebrate and non-vertebrate cells. Table 2 shows estimated numbers of transcripts in two mammalian cell types (HeLa and L cells) and a *Drosophila* cultured cell line, based on RNA reassociation analysis (Bishop *et al.* 1974). Within the limits of the technique, it is apparent that the approximate numbers of transcripts needed by a *Drosophila* cell and a mammalian cell are similar. Comparative studies of messenger RNA complexity in other cell and tissue types corroborate the general impression that the diversity of transcripts (and therefore proteins) needed by these disparate cell types is roughly the same.

Table 2. *Similar complexity and number of messenger RNAs in tissue culture cells of an invertebrate* (Drosophila) *and two vertebrates* (*human and mouse*)

(The data are taken from Levy & McCarthy (1975), Ryffel & McCarthy (1975) and Williams & Penman (1975). It is assumed that the average size of mRNAs is 2000 nucleotides.)

species/cell line	RNA complexity	mRNA number
Drosophila melanogaster Schneider cells	13400	6700
Mus musculus L cells	12500	6300
Homo sapiens HeLa cells	18800	9400

Early studies of mRNA populations also gave important information about gene expression in different cells of an organism (Hastie & Bishop 1976). Most messenger RNAs are common to all cell types. In other words, there is a common set of mRNAs encoding housekeeping proteins that is found in all kinds of cells. The differences between one cell type and another is due to proteins encoded by a relatively small number of mRNAs that are cell type-specific. Combination of this information with the data on gene number permits a diagrammatic representation of the transcription patterns in vertebrate and invertebrate organisms (figure 1). Five cell types of each organism are represented. As discussed above, the number of genes that are active in each cell type is approximately similar. Most of the active genes are common to all cell types of that organism, whereas a minority represent 'tissue-specific' transcripts that define the properties of the particular cell (nerve, muscle, etc.). It is apparent from the figure that only a small fraction of the total number of genes is expressed in any vertebrate cell type, whereas in the invertebrate example most of the genes are actively expressed in each cell. Thus a crucial

Table 1. *Approximate estimates of gene number in free living organisms*

(The origin of the values is given in Bird (1995).)

species	type	gene number	method
prokaryotes			
E. coli	bacterium	4000	ORF
eukaryotes (except vertebrates)			
S. lemnae	ciliated protozoan	12000–15000	reassociation
O. similis	ciliated protozoan	12000	reassociation
S. cerevisiae	fungus	7000	ORF
D. discoidium	slime mould	12000	reassociation
D. melanogaster	arthropod	12000–16000	reassoc/ORF
C. elegans	nematode	17800	ORF
S. purpuratus	echinoderm	< 25000	reassociation
vertebrates			
Fugu	fish	50000–100000	sequence homol.
M. musculus	mammal	80000	CpG islands
H. sapiens	mammal	80000	CpG islands
H. sapiens	mammal	60000–70000	cDNA tagging

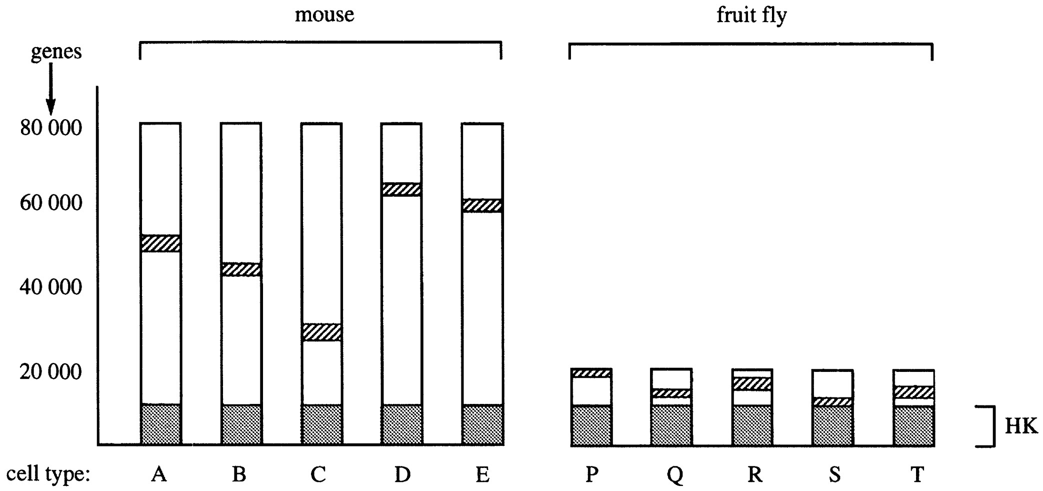

Figure 1. Proportions of active and silent genes in vertebrate and invertebrate cell types. Five cell types of mouse (A–E) and fruit fly (P–T) are represented. The height of each bar represents the total number of genes in that cell, which is the same for all tissues of the organism. Active genes are shaded. Dark shading at the bottom of each bar shows the number of genes that are commonly expressed in all tissues (HK for housekeeping genes). Diagonal shading shows the minority of tissue specific genes in each cell type whose expression specifies its identity. The number of active genes in each mouse and fly cell type is approximately the same (see table 2), but the number of genes in total is very different (see table 1). Hence there is a large difference in the number of silent genes (unshaded regions) between invertebrate and vertebrate cell types. All values are approximate.

difference between the vertebrates and the invertebrates resides in the number of genes that are kept silent. Taking literally the approximate values represented in figure 1, it appears that a vertebrate cell has about 60000 silent genes whereas an invertebrate cell has about 6000 silent genes, ten times fewer.

Figure 1 also makes explicit the advantage in terms of organismal complexity of having more genes. The number of different cell types that *Drosophila* can programme is less than the number that a mouse can programme. This is because the number of separate cell identities that are possible depends on the number of available permutations of the tissue-specifying genes. Obviously, the more genes there are, the more identities can be specified.

How silent are the unused genes? Polymerase chain reaction technology allows us to estimate that for several 'silent' genes there is one transcript in approximately 10^4 mammalian cells (S. Tweedie & A. Bird, unpublished observations). This seems to be a remarkably high level of repression until one considers the number of genes to which it applies. A cell with 60000 'silent' genes (figure 1) would contain six inappropriate transcripts from these repressed genes, each cell having a different selection. At first sight, a few 'rogue' transcripts may not appear to be capable of causing adverse biological effects. Recall, however, that most messenger RNAs are present at five to ten copies per cell (Hastie & Bishop 1976). This means that a single mRNA per cell is only a few-fold less abundant than most of that cell's appropriate messengers. This emphasizes the importance of keeping the selection of spurious transcripts to a minimum.

How is noise reduction achieved? One component of the system is likely to be chromatin, which is known to repress basal transcription without affecting the expression of activated promoters (Laybourn & Kadonaga 1991). It is hypothesized here that chromatin by itself is not enough. Help is required from genome-wide low density methylation, a noise reduction system that is peculiar to the vertebrates. Together, DNA methylation and chromatin prevent transcription from the very large number of 'silent' genes and in so doing they allow a cell to express only the genes that programme its identity. Invertebrates have fewer silent genes in each cell type (figure 1) and therefore need less stringent noise reduction mechanisms.

In summary, it is proposed that gene number is limited by the efficiency of mechanisms that restrain transcription from inappropriate genes (see Bird 1995). By co-opting DNA methylation vertebrates have improved noise reduction and as a result have raised the ceiling on gene number. New genes to fill the void arose by duplication and divergence of old ones (Ohno 1970); hence the plethora of gene families in present day vertebrates. Gene duplication itself is an extremely frequent event. What is rare in evolution is significant improvement in the capacity to retain and control all the novel genetic functions that duplication and divergence can generate.

With more genes available, vertebrates built bodies that were more complex than those of their invertebrate ancestors. Whether vertebrates are really more complex than invertebrates can be debated at length, because the meaning of complexity in a biological context is difficult to pin down. This problem is circumvented here by equating complexity with the number of genes an organism has.

GENE NUMBER AT THE PROKARYOTE–EUKARYOTE TRANSITION

Is it possible that the transition from prokaryotes to eukaryotes also depended upon changes in genome management analogous to those at the vertebrate–

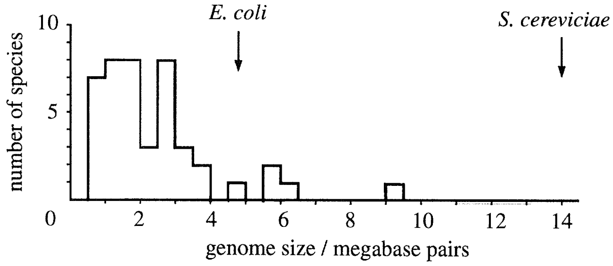

Figure 2. Genome size in bacteria (Krawiec & Riley 1990). By assuming that the density of genes in all species is the same as that in *E. coli* (one gene per 1000 nucleotide pairs on averages (Burland *et al.* 1993)), it is possible to deduce the number of genes in each species by multiplying the genome size in megabases by 1000. Thus the mean number of genes in this selection is about 2500 per genome.

invertebrate boundary? One striking difference between the groups concerns genome packaging. Bacterial genomes are largely exposed to diffusable components in the cell. Gene silencing depends upon repressor molecules that are dedicated to the operon concerned, such as the *lac* and *trp* repressors (Hoopes & McClure 1987). Eukaryote genomes, on the other hand, are sequestered from cellular factors by the nuclear membrane and by packaging into nucleosomes. Nucleosomes in particular appear to be adapted for reduction of transcriptional noise, as they have been shown to greatly reduce spurious ('basal') transcription, without affecting output from authentic promoters (Laybourn & Kadonaga 1987). These changes in genome accessibility correlate with an apparent increase in the number of genes per genome in the eukaryotes compared with prokaryotes (table 1). The parallel between this transition and the vertebrate–invertebrate transition discussed above is evident (see Bird 1995).

Gene number in prokaryotes can be inferred reliably from genome size, as sequencing projects indicate that genes are maximally packed in the genome at a density of about one gene per 1000 base pairs. A spread of genome sizes from prokaryotic organisms indicates an average of about 2500 genes in a typical prokaryotic genome (figure 2). Gene number in eukaryotes cannot be inferred from genome size, as the density of genes is extremely variable. Nevertheless, as shown in table 1, estimates (of variable accuracy) have been made, and eukaryotes clearly have several-fold more genes than most prokaryotes. It is proposed that the global repression mechanism associated with nucleosomes (and perhaps the nuclear membrane) facilitated this growth in gene number. The implication of this hypothesis is that the fundamental difference between a eukaryote and a prokaryote is the increased number of genes in the latter.

Fossil remains indicate that prokaryotes were the only living forms on Earth for at least a billion years. Eukaryotes appeared at the end of this period, between 1.5 and 2 billion years ago. The colossal delay may be due to the unlikelihood that a novel mechanism of this kind could be introduced without severely compromising the organism's fitness. Nucleosomes, for example, could not have suddenly appeared in the genome of a prokaryote without killing it. More feasible is that these structures were advantageous for some other reason in the first instance, but were by chance preadapted for the repression role. The analogy with DNA methylation may ultimately be helpful. The methylation system appears to have evolved initially as part of a system for detecting and neutralizing selfish elements in the genome. DNA methylation was therefore preadapted as a repressor, and it is not difficult to envisage a scenario that could have lead to its spread through the genome as a dampener of illegitimate transcription. Nucleosomes could have been similarly preadapted, though we have little idea of their ancestral function. The study of bacteria that diverged more recently from the eukaryote lineage may shed light on this question.

This work was funded by the Wellcome Trust, The Imperial Cancer Research Fund and the Howard Hughes Medical Institute.

REFERENCES

Bird, A. P. 1980 DNA methylation and the frequency of CpG in animal DNA. *Nucl. Acids Res.* **8**, 1499–1594.

Bird, A. P. 1993 Functions for DNA methylation in vertebrates. *Cold Spring Harb. Symp. quant. Biol.* **58**, 281–285.

Bird, A. P. 1995 Gene number, noise reduction and biological complexity. *Trends Genet.* **11**, 94–100.

Bird, A. P. & Taggart, M. H. 1980 Variable patterns of total DNA and rDNA methylation in animals. *Nucl. Acids Res.* **8**, 1485–1497.

Bishop, J. O., Morton, J. C., Rosbach, M. & Richardson, M. 1974 Three abundance classes in Hela cell mRNA. *Nature, Lond.* **250**, 199–203.

Boyes, J. & Bird, A. 1991 DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell* **64**, 1123–1134.

Boyes, J. & Bird, A. 1992 Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *Embo Jl.* **11**, 327–333.

Cooper, D. N. & Youssoufian, H. 1988 The CpG dinucleotide and human genetic disease. *Hum. Genet.* **78**, 151–155.

Hastie, N. D. & Bishop, J. O. 1976 The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9**, 761–774.

Hoopes, B. C. & McClure, W. R. 1987 Strategies in regulation of transcription initiation. In *Escherichia coli and Salmonella typhimurium: cellular and molecular biology* (ed. F. Neidhardt *et al.*), pp. 1231–1240. Washington D.C.: American Society for Microbiology.

Jones, P. A., Rideout, W. M., Shen, J.-C., Spruck, C. H. & Tsai, Y. C. 1992 Methylation, mutation and cancer. *BioEssays* **14**, 33–36.

Josse, J., Kaiser, A. A. & Kornberg, A. 1961 Enzymatic synthesis of deoxyribonucleic acid. VII Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J. biol. Chem.* **236**, 864–875.

Krawiec, S. & Riley, M. 1990 Organization of the bacterial chromosome. *Microbiol. Rev.* **54**, 502–539.

Laybourn, P. J. & Kadonaga, J. T. 1991 Role of nucleosomal cores and histone H1 in regulation of transcription by RNA polymerase II. *Science, Wash.* **254**, 238–245.

Levy, W. & McCarthy, B. 1976 Messenger RNA complexity in *D. Melanogaster. Biochemistry, Wash.* **14**, 2440–2446.

Ohno, S. 1970 Evolution by gene duplication. Berlin: Springer-Verlag.

Ryffel, G. & McCarthy, B. 1975 Complexity of cytoplasmic RNA in different mouse tissues measured by hybridisation of polyadenylated RNA to complementary cDNA. *Biochemistry, Wash.* **14**, 1379–1385.

Sved, J. & Bird, A. 1990 The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natn. Acad. Sci. U.S.A.* **87**, 4692–4696.

Williams, J. & Penman, S. 1975 The messenger RNA sequences in growing and resting mouse fibroblasts. *Cell* **6**, 197–206.