Proc. Indian natn Sci Acad. **B69** No. 4 pp 447-452 (2003)

Gene Birth, Death, Modification, Poaching, Crippling, Dimorphism and Culling: The Challenge for Genomics

LINDSAY G COWELL[1], N AVRION MITCHISON[2*], BRIGITTE MULLER[3], LAURIE G SMITH[4] and NADIA M TERRAZZINI[5]

[1]*Department of Immunology, Campus Box 3010, Duke University Medical Center, Durham, NC 7701, UK*
[2]*Department of Immunology, Windeyer Institute of Medical Science, 46 Cleveland Street, London 1T 4JF, UK*
[3]*Deutsches Rheuma Zentrum Berlin, Schumannstrasse 21/22, 10117 Berlin, Germany*
[4]*Section of Cell and Developmental Biology,University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093-0116, USA*
[5]*Department of Life Sciences, University of East London, Romford Road, E15 4LZ, London, UK*

(Received on 20 June 2002; Accepted after revision on 18 September 2002)

This brief survey covers the main events in the evolution of eukaryotic genes in broad brush style. It concentrates on regulatory DNA, an area that has been relatively neglected, and where we believe that the present case-by-case analysis is likely to be supplemented by more general, genomics-based methods. It is biased towards immunology, in part because the immune system relies heavily on polymorphism of regulatory DNA to provide flexibility and in part because of our own field of interest. It gives a central place to recent work that has shown how analysis of electronic genomes can be used to trace gene duplication and its consequences. It mentions cellular systems that offer models for the study of evolution of regulatory DNA on a small scale. It alludes to the unanswered question of how genomes adjust their overall size.

Key Words: Regulatory DNA, Gene duplication, Gene promoters in cancer, Herpes viruses, Major histocompatibility complex (MHC)

The subject of this review can be put in the following nutshell. A century and a half has passed since we learned to interpret form in terms of function and evolution. Several decades have passed since we learned to do the same for coding genes and protein structure. Yet for regulatory DNA we have hardly begun. This is reflected in the fact that EMBOSS (the European Molecular Biology Open Software Suite) contains 43 programs for handling proteins, and just one for regulatory DNA. The gap is remarkable, as it is now widely accepted that a full understanding of form and function in biology depends ultimately on knowing when, where and how much individual genes are expressed (Carroll 2000, Davidson 2001).

The current edition of *The Molecular Biology of the Cell* demonstrates where we now stand (Alberts et al.2002). It illustrates transcription factors assembling around RNA polymerase2 at the start of transcription site, but the figures are purely diagrammatic and make no attempt to show the way the DNA folds or to present a three dimensional structure. At present we know only that the full structures will be enormously difficult to solve – at least an order of magnitude more difficult than solving the structure of any single protein. And an understanding will be even harder to obtain of how such a structure operates and could have evolved.

This is an area hitherto dominated by biochemistry, using such tools as EMSA (electrophoretic mobility shift assay), foot printing, reporter constructs and so on. Nevertheless we are convinced that the genomics based evolutionary approach has much to offer. Close examination of how promoters have evolved, particularly in micro-evolution, will surely contribute to the final solution. This review therefore deals with the following three questions:

E-mails: [1] *lgcowell@pengo4.mc.duke.edu;* [3]Mitte ueller@drfz.de; [4]lsmith@biomail.ucsd.edu; [5]nadia@whstaff1.uel.ac.uk;
* Corresponding author : [2]n.mitchison@ucl.ac.uk; Tel. 44 (0) 20 76799354; Fax 44 (0) 20 76799357

1. How far has the genomic approach got in identifying regulatory DNA?
2. What can we learn from the pattern of variation in the DNA thus identified?
3. Can we in this way build a picture of the evolution of regulatory DNA?

To clarify, most regulatory DNA is contained within the promoter region upstream of coding sequences, but it would be a mistake simply to equate regulatory function with that location. Important regulatory activity has been detected far upstream of the proximal promoter region, within introns, and downstream of the coding regions. Rather, the question is whether functionally active sequences can be identified directly from their sequence properties, rather than operationally. Recent mainstream approaches to this problem have been reviewed elsewhere (Stormo 2000, Fujibuchi et al. 2001, Coleman et al. 2002). Questions one and two are closely linked, for functional activity can up to a point be identified from the pattern of variation (conservation of transcription factor binding sites, but with functionally active variation grouped around them). Question three is also closely linked, as an important test of functional activity is whether it is conserved during evolution.

Recent surveys emphasize the importance of regulatory gene sequences in evolution, but without taking us far into detail (Mitchison 1997, Carroll 2000, Enard et al. 2002). Often cited is the striking study of domestication in maize from its teosinte ancestor.A major area of research is the multifactorial diseases of man and their animal models. This has yielded conspicuous associations of disease susceptibility with polymorphisms in regulatory DNA (Mitchison 2001), notably in autoimmune (Becker et al. 1998) and cardiovascular disease (Pailard et al. 1999). We exclude from this generalization the special case of polymorphism in the coding sequences of the “extrovert” genes, which encode proteins such as hemoglobin and the MHC (major histocompatibility complex) that bind directly to parasite and other foreign molecules (Mitchison 1997). These are where the “red queen” evolutionary strategy enlarges structural diversity (Sasaki et al. 2002). We also exclude from consideration microbial evolution, which has its own rules. So far these instances of domestication and disease association have been studied largely on a case-by-case basis; the question now is whether they can be brought together by the application of genomics.

We note that India plans to invest heavily in genomics research (Mudur 2001), and hope that the approaches discussed here will proved relevant.

Birth and Subsequent Modification of Genes

Gene duplication has generally been regarded as a necessary source of evolutionary novelty.A recent study of the origin and subsequent fate of duplicated genes is based on the genomes of man, mouse, chicken, nematode, fly, rice, *Arabidopsis*, and yeast (Lynch & Conery 2000, see also further discussion in the same journal). Each known open reading frame (after appropriate filtration) was compared with all others from the same species so as to identify gene duplicates, and diversity between the duplicates was then analyzed for nucleotide substitution at replacement (R) versus silent (S) codon sites. The R/S ratio is a useful parameter, since silent substitutions provide a “biological clock”, while replacements measure the pressure of natural selection. The study yielded important information about the survival, modification and silencing (death) of duplicated genes. For our present purpose it demonstrates two important points. One is that plants are relatively rich in duplications, a point that has a long history tracing back to their high level of chromosomal rearrangements that provide plasticity in adapting to varied environments (Barnes 2002, Comai 2000). Plants occupy a high position on the genomics agenda. The second point is that duplicated genes that have diverged in their coding sequences can routinely be identified in the published electronic data bases by genomics. The promoters of these genes could potentially be tracked, and should provide extensive material for further genomic analysis.

Limited gene sets are already being subjected to this type of analysis. For instance the alpha interferon genes have undergone extensive duplication, with as many as fifteen serially duplicated genes running along human chromosome 9. Their coding sequences show only minor variation, and it seems likely that it is the need to make different responses in different cell types that maintains this extensive duplication. One promoter variant in these genes has already been shown to alter the response of reporter constructs (Raj et al. 1991). It is likely that the same kind of analysis could also be applied to chemokines (Yoshie et al. 2001).

These are approaches based on intraspecies analysis. Clearly comparison between species can yield further information of the same sort. For both types of analysis the interval since the start of diversification is critical. If the interval is too short the extent of diversification may be too small, and if too long the informative diversification may get swamped by accumulation of random polymorphism (junk DNA). With long intervals, for instance in trans-family comparisons within the MHC, the only feature of a promoter that is conserved may be the canonical transcription factor binding sites (Benoist & Mathis 1990). The variation around these sites may be what is most informative, but may no longer be detectable because of the surrounding junk. In this kind of analysis what matters is the variation that is maintained by selection. For man the most informative comparison may be with the apes, although it is too early to be sure (Enard et al. 2002). Indeed drawing the limits of informative comparison will be an important task for the future genomics, with different types of gene no doubt having different limits.

Poaching

In this and the following section we mention approaches that have not yet yielded subject matter for genomics, but which we suspect will eventually do so. Viruses occasionally pick up host genes and incorporate them into their own genomes. Because viruses evolve so fast this poaching provide valuable information about the evolution of regulation. Herpes viruses are conspicuous examples; where 10 – 15% of the ~100 genes in their genome have been acquired in this way (Alba et al. 2001a,b). They comprise a large family of viruses, within which some 4,000 genes have so far been sequenced and subjected to genomic analysis. They offer rich material for study of the evolution of regulatory DNA sequences. It seems that these viruses must often evolve their own regulatory DNA, since the viral genes of host origin usually lack introns and must therefore have been picked up from host mRNA. Nevertheless their regulatory DNA cannot be entirely independent of the host, since it depends on the host cell for much of its machinery of expression.

Crippling

Sometimes genes come under strong selective pressure. Cancer is an outstanding example, where proteins not required for cell survival or growth tend to be dispensed with, and those that mediate drug resistance get over expressed. Two groups of studies have examined these effects on regulatory DNA sequences. One is in the Reed-Sternberg cell, believed to be the central culprit in Hodgkin's Lymphoma. This cell is related to the B cell, as it has rearranged immunoglobulin genes although these are transcriptionally inactive. Reed-Sternberg cells have been found with an immunoglobulin promoter that is "crippled" by mutation in the immunoglobulin octamer, although this form of inactivation seems to be unusual (Theil et al. 2001). Crippling mutations have also been found in the Fas promoter, again consonant with the needs of a cancer cell (Muschen et al. 2000). Crippled MHC gene promoters have been found in melanoma cell lines, consonant with the dispensed-function hypothesis but consonant also with the hypothesis of immune surveillance (Lee et al. 2000). In the UK, we understand from personal information that of the Cancer Genome Project of the Wellcome Trust Genome Campus at Hinxton has for the time being chosen not to investigate mutation in regulatory DNA, although that may change.

Another approach would be to put promoters under pressure in an experimental system. Here again there is some background information from experiments with the MHC. Many years ago George Snell, the co-discoverer of the MHC, bred congenic mouse strains in which the MHC from one strain was introduced into another by back-crossing, which at the time represented an important step forward in understanding the MHC. It thus became possible to breed mice hemizygous for a congenic MHC, induce cancer, and then immuno-select the cancer cells (by transplantation) for loss of one of the two "allelic" MHC's. Two groups thought – at least for a while – that they had succeeded in doing so (Mitchison 1956, Dalianis et al. 1981). However, the work was never pursued to the point where immunoselection was applied to the product of a single gene (the full classical MHC of the mouse is now known to have five linked loci). With modern knowledge of the MHC this is something that could easily be done, and would seem to offer a fair chance of picking up crippling mutations in the MHC promoters. Furthermore it would be of interest to perform the immune selection *in vitro* as well as *in vivo*. Monoclonal antibodies could be used to select against any of a wide range of single gene products expressed on the cell surface. Indeed selection for or against expression of any gene of interest, by one means or another, could be of interest.

Promoter Variation Reflects Variation in Coding Sequences

It has long been realized that variation in one set of coding sequences can select for variation in other linked sequences, a process referred to as hitch-hiking (Smith & Haigh 1974). The MHC is a good place to look for such an effect, because of the high level of polymorphism in coding sequences maintained by balancing selection (heterozygote advantage). Such an effect has indeed been found, over distances far longer than the classical promoter (Beck & Trowsdale 2000). Yet it seems likely that MHC promoters would also be subject to direct natural selection. To try to sort out these two effects, we sequenced a series of murine MHC promoters mainly derived from wild mice (Mitchison & Roes 2002). The upshot was that both effects seem to operate. Linkage disequilibrium with polymorphic coding sequences – hitch hiking – seems necessary to maintain extensive promoter polymorphism. The most telling support for this view comes from study of CIITA (Class II trans-activator). This gene is a major modulator of MHC II expression, yet being non-polymorphic in its coding sequences it is also non-polymorphic in its promoter. On the other hand the hitch hiking effect is clearly not sufficient to account for all the features of MHC promoter polymorphism. Although these MHC II promoters are extraordinarily diverse, a set of MHC class I promoters from the same genomic DNA that we also sequenced showed little diversity. As MHC I coding sequences are as diverse as those of MHC II, we attributed the difference to the greater range of selective pressures operating on MHC II gene expression. Furthermore, the polymorphic residues in the MHC II promoters were predominantly located near transcription factor binding sites.

We also examined human MHCII promoters (Held et al. submitted). We employed a biophysical approach to detect the ability of varying promoter sequences to bind transcription factors. First we obtained synthetic samples of naturally occurring oligonucleotides that varied slightly in sequence. Each oligonucleotide was then tested for its ability to bind nuclear extracts, by means of plasmon resonance. The only variation that affected the level of binding occurred in sequences adjacent to the TATA box. Furthermore this effect could be verified by showing that the oligonucleotides detected in this way also yielded different values by EMSA (the electrophoresis mobility shift assay mentioned above).

The general message is that in any systematic scan for promoter variation it would make sense to pay special attention to genes with polymorphic coding sequences. They seem to be where a meaningful distribution of variable residues is most likely to occur. The known mouse MHC II promoter variants provide a model of what can be hoped for and their further exploration has become an urgent matter. We need to know whether reporter constructs can be used to explore their functional consequences. It would be particularly helpful to learn whether, for instance, substitutions near the CRE sequence (the cyclic-AMP response element) alter the response to agents that regulate the cAMP level. Ultimately, we need to explore gene-reshuffling, where different promoters would be tested *in vivo* for their effect on the same coding sequence.

Gene Dimorphism

An unexpected feature of the polymorphism emerged during the above study of mouse MHC II promoters. To our surprise, each of the three promoters under study turned out to have two basic polymorphic type, each made up of a linked series of alternative nucleotide substitutions. The separation into two haplotypes was not complete, as limited recombination between different sites within each promoter evidently takes place. The haplotypes were not in linkage disequilibrium between the three loci. Thus we were confronted with three instances of promoter dimorphism, for which we had no explanation. Enlarging on this limited data, we speculated that dimorphism (in either regulatory or coding sequences) may play a part in the birth and death of genes. Conceivably, we proposed, it may assist newly duplicated genes to diversify at the early time when one of the pair is at greatest risk of extinction. A dimorphic gene might duplicate and subsequently recombine with the old, non-duplicated gene of the opposite dimorphic type, thus generating a new chromosome in which the two dimorphic types are both present in tandem.

A second possibility, we proposed, is that dimorphism could assist the loss of function of a duplicated gene. One member of a pair of duplicated genes (that have undergone differentiation) might develop a dimorphism, in which one of the haplotypes mimics the function of the other duplicated gene. This would make the other gene redundant, and leaves it prone to extinction. We proposed that this could explain why the H2E gene so often looses function, since its function is mimicked (up to a point) by one of the dimorphic haplotypes at H2A.

These two possibilities are opposed to one another, in a way that might be expected of genomes that seem to vary habitually in size. This is all highly speculative, and hardly worth mentioning except that promoter dimorphism does not seem to have been encountered previously. It would be worth keeping an eye out for in the future.

Culling

Granted that genes undergo unending cycles of birth and death by loss of function, how come that the genome does not fill up with pseudogenes? Interpreting their work on the compact genome, Venkatesh et al. (2000) ask "Is the Fugu gene density a primitive trait retained from the ancestral duplicate invertebrate genome? This seems unparsimonious, as the pufferfish belongs to a highly derived teleost lineage, and the majority of teleosts that diverged earlier from the ancestral lineage have larger genomes. It seems more likely that a strong bias towards deletions in the Fugu has compacted the genome.". About the nature of that bias the Brenner group has nothing to say, and nor do we. Machinery for culling does seem to exist, and in general presumably serves to keep the genome within bounds of some sort. We are left with an important but unanswered question.

From Informatics to Experiment, and Back Again

The scope outlined here for future genomics calls for the gathering of further data, particularly in the areas of (i) population genetics, (ii) biochemistry with reporter constructs, EMSA and plasmon resonance and (iii) reversed genetics with promoter reshuffling. In population genetics we expect increasing use will be made of DNA banks, such as that mentioned here of mice offered by the Jackson Laboratory. These data will in turn demand further genomics. The trade will be two-way, and will require scientists able to handle both approaches.

References

Alba M M, Das R, Orengo C A & Kellam P 2001a Genomewide function conservation and phylogeny in the Herpesviridae; *Genome Res.* **11** 43-54

______, Lee D, Pearl F M, Shepherd A J, Martin N, Orengo C A and Kellam P 2001b VIDA: a virus database system for the organization of animal virus genome open reading frames; *Nucleic Acids Res.* **29** 133-136

Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P 2002 *Molecular Biology of the Cell*, 4th edition (New York: Garland Science)

Barnes S 2002 Comparing Arabidopsis to other flowering plants; *Curr. Opin. Plant Biol.* **5** 128-134

Beck S and Trowsdale J 2000 The human major histocompatability complex: lessons from the DNA sequence; *Annu. Rev. Genomics Hum. Genet.* **1** 117-137

Becker K G, Simon R M, Bailey-Wilson J E, Freidlin B, Biddison W E, McFarland H F and Trent J M 1998 Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases; *Proc. Natl. Acad. Sci. USA* **95** 9979-9984

Benoist C and Mathis D 1990 Regulation of major histocompatibility complex class-II genes: X, Y and other letters of the alphabet; *Annu. Rev. Immunol.* **8** 681-715

Carroll S B 2000 Endless forms: the evolution of gene regulation and morphological diversity; *Cell* **101** 577-580

Coleman S L, Buckland P R, Hoogendoorn B, Guy C, Smith K and O'Donovan M C 2002 Experimental analysis of the annotation of promoters in the public database; *Hum. Mol. Genet.* **11** 1817-1821

Comai L 2000 Genetic and epigenetic interactions in allopolyploid plants; *Plant Mol. Biol.* **43** 387-399

Dalianis T, Ahrlund-Richter L, Merino F, Klein E and Klein G 1981 Reduced humoral and cellular cytotoxic sensitivity in histocompatibility variants of the YAC (Moloney) lymphoma; *Immunogenetics* **12** 371-380

Davidson E H 2001 *Genomic Regulatory Systems: Development and Evolution* (New York: Academic Press)

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis G M, Bontrop R E and Paabo S 2002 Intra- and interspecific variation in primate gene expression patterns; *Science* **296** 340-343

Fujibuchi W, Anderson J S and Landsman D 2001 PROSPECT improves cis-acting regulatory element prediction by integrating expression profile data with consensus pattern searches; *Nucleic Acids Res.* **29** 3988-3996

Heldt C, Listing J, Sözeri O and Müller B Differential expression of HLA II genes associated with disease progression in rheumatoid arthritis (to be submitted)

Lee T J, Kim S J and Park J H 2000 Influence of the sequence variations of the HLA-DR promoters derived from human melanoma cell lines on nuclear protein binding and promoter activity; *Yonsei Med. J.* **41** 593-599

Lynch M and Conery J S 2000 The evolutionary fate and consequences of duplicate genes; *Science* **290** 1151-1155

Mitchison A 1997 Partitioning of genetic variation between regulatory and coding gene segments: the predominance of software variation in genes encoding introvert proteins; *Immunogenetics* **46** 46-52

Mitchison N A 1956 Antigens of heterozygous tumours as material for the study of cell heredity; *Proc. Royal Physical Soc.* **25** 45-48

______ 2001 Polymorphism in regulatory gene sequences; *Genome Biol.* **2** 1-6

______ and Roes J 2002 Patterned variation in murine MHC promoters; *Proc. Natl. Acad. Sci. U S A* **19** 10561-10566

Mudur G 2001 India invests heavily in genomics research; *BMJ* **322** 576

Muschen M, Re D, Brauninger A, Wolf J, Hansmann M L, Diehl V, Kuppers R and Rajewsky K 2000 Somatic mutations of the CD95 gene in Hodgkin and Reed-Sternberg cells; *Cancer Res.* **60** 5640-5643

Paillard F, Chansel D, Brand E, Benetos A, Thomas F, Czekalski S, Ardaillou R and Soubrier F 1999 Genotype-phenotype relationships for the renin-angiotensin-aldosterone system in a normal population; *Hypertension* **34** 423-942

Raj N B, Au W C and Pitha P M 1991 Identification of a novel virus-responsive sequence in the promoter of murine interferon-alpha genes; *J Biol. Chem.* **266** 11360-11365

Sasaki A, Hamilton W D and Ubeda F 2002 Clone mixtures and a pacemaker: new facets of Red-Queen theory and ecology; *Proc. R. Soc. Lond. B Biol. Sci.* **269** 761-772

Smith J M and Haigh J 1974 The hitch-hiking effect of a favourable gene; *Genet. Res.* **23** 23-35

Stormo G G 2000 DNA binding sites: representation and discovery; *Bioinformatics* **16**

Theil J, Laumen H, Marafioti T, Hummel M, Lenz G, W irth T and Stein H 2001 Defective octamer-dependent transcription is responsible for silenced immunoglobulin transcription in Reed-Sternberg cells; *Blood* **97** 3191-3196

Venkatesh B, Gilligan P and Brenner S 2000 Fugu: a compact vertebrate reference genome; *FEBS Lett.* **476** 3-7

Yoshie O, Imai T and Nomiyama H 2001 Chemokines in immunity; *Adv. Immunol.* **78** 57-110