The EMBO Journal Vol.17 No.19 pp.5521–5524, 1998

EMBO WORKSHOP REPORT

Trinucleotide Expansion Diseases in the Context of Micro- and Minisatellite Evolution Hammersmith Hospital, April 1–3, 1998

John M.Hancock and Mauro F.Santibáñez-Koref

Comparative Sequence Analysis Group, MRC Clinical Sciences Centre, Hammersmith Hospital, London W12 0NN, UK

Introduction

In 1991, La Spada *et al.* and Kremer *et al.* showed that spinal and bulbar muscular atrophy and Fragile X syndrome are caused by expansions of CAG and CGG repeats, respectively. Since then, numerous other diseases have been found to result from a similar mechanism of expansion. Expansions have been found affecting repeats in exons, 5′ and 3′ UTRs and introns. A considerable amount of effort has been put into understanding the process of this expansion, its relationship to disease and the behaviour of these repeats within human populations—all essential prerequisites for providing effective genetic counselling, and, in the longer run, developing effective therapies.

The revelation that triplet repeats can make a significant pathological impact was a surprise—since the earliest discovery of simple sequences (now generally known as microsatellites) they have been regarded as essentially neutral structures, useful as genetic markers but of no evolutionary significance beyond that. Despite (or, in part, because of) this assumed neutrality, evolutionary biologists and population geneticists had focused a considerable amount of attention on the pattern and process of repetitive sequence evolution. Much of this knowledge remains at the fringes of studies of the diseases themselves, while evolutionary biologists working on repetitive sequences are largely ignorant of the biological processes that contribute to instability in triplet repeats. The aim of this meeting was to improve the cross-fertilization between these two, not necessarily mutually exclusive, groups of researchers.

Three main strands could be discerned from the presentations. We will describe these loosely as: the genomic and evolutionary context of repeats, mechanisms of mutation at repeats and the molecular genetic basis of repeat instability.

Genomic and evolutionary context of repeats

Surveys of the frequencies and distributions of repeats in eukaryotic genomes, based on database searches, show that certain repeats are more common than others, and that those involved in triplet expansion diseases are relatively rare. At the DNA level, searches that ignore the coding/non-coding potential of sequences show that poly(A/T) repeats are the most common mononucleotide repeats, poly(CA/GT) the most common dinucleotides, poly(AAT) is the most common trinucleotide, while amongst tetranucleotides, G/A rich repeats are the most common (Jörg Epplen, Bochum; Hanah Margalit, Jerusalem). Interestingly, zinc finger proteins that can bind to many of these repeats have been identified (Epplen).

Within proteins, repetition has also been surveyed at the level of amino acids. Here polyglutamine repeats, those coded for by poly(CAG) [and potentially poly(CAA)] sequences, are the the most common, and may provide the raw material for the evolution of new proteins or protein domains (Howard Green, Harvard). Many long poly-amino-acid repeats are encoded by interrupted codon repeats (Green; Margalit), and such patterns are also seen to accumulate during the evolution of disease-related repeats at the Fragile X (Evan Eichler, Cleveland) and SCA1 (John Hancock, London) loci. Is there some specific mechanism that introduces interruptions into repeats (Green), or is this simply an effect of selection against the phenotypic consequences of repeat instability? Another possible constraint on the types of repeat that accumulate within coding regions is the requirement to avoid erroneous RNA–RNA interactions in the cytoplasm (Donald Forsdyke, Kingston).

A number of arguments can be made for the neutrality (non-functionality) of repeats within genomes. These include the observation that, even when they are within exons, they vary widely in length between species, and may even include frameshifts (Green; Diethard Tautz, München). In the red flour beetle *Tribolium castaneum*, repeats are very rare and are absent from 15 proteins whose homologues in *Drosophila* contain them, which suggests that they are dispensable. Further, the genetic load of mutations within microsatellites would be lethal unless the mutations were effectively neutral (Tautz). Dmitry Gordenin (Research Triangle Park) pointed out that during the replication of a genome there may be 100 times more mutations at microsatellites than there are point mutations.

Despite these arguments, the existence of diseases caused by repeat expansion provides a strong basis for selection to act on these repeats. Evolutionary biologists are rarely able to focus on the proximal causes of phenotypes acted on by selection, but the triplet expansion diseases provide us with detailed analyses of a variety of different kinds of processes. These vary from the shutting down of a specific gene due to methylation of the region around the CGG repeat in Fragile X syndrome to what appear to be more complex interactions between a number of closely packed genes at the myotonic dystrophy locus (Keith Johnson, Glasgow). Field effects resulting from the formation of altered chromatin conformations at the myotonic dystrophy repeat may affect the expression of nearby genes (Johnson), and similar effects may be seen in transgenic mice in which the targeted constructs contain CTG tandem repeats (Richard Festenstein, London). The phenotypic consequences of glutamine expansions remain poorly understood. Philippe Djian (Meudon) presented evidence that expanded polyglutamine repeats act as substrates for transglutaminase, giving rise to protein–protein cross-linking in aggregates, while Jean-Marc Gallo (London) showed that Androgen Receptor molecules with expanded polyglutamine repeats are targets of enhanced proteolysis.

Evolutionary analysis also provides evidence that selection acts on repeats, at least within coding regions. One argument derives from their long evolutionary persistence, which might not be expected if they were wholly neutral (Eichler; Hancock), and from the length limits that are evident on amino acid repeats (Green). Gabby Dover (Leicester) suggested that repeats could participate in coevolution between regulatory sequences in genes and the protein factors binding them. Repeats might also participate in coevolution between interacting protein molecules (Hancock). Perhaps the clearest example of selective advantage conferred by repeats is seen in pathogenic bacteria. Here, slippage events may either vary the phase of the start codon with respect to the rest of the coding region, as seen in the SINQ glycosyl transferase of the Gram-negative bacterium *Haemophilus influenzae*, or alter the organization of gene promoters, as seen in the *H.influenzae pilin* gene. Both of these phenomena provide means for the bacteria to avoid host immune responses (Richard Moxon, Oxford).

Repeats might also confer more indirect advantages. Analyses of the divergence of sequences near repeats suggested that, outside a so-called transitional zone which was probably derived from the microsatellite itself, neutral divergence decreased with increasing array size both for CA microsatellites (Mauro Santibáñez-Koref, London) and CAG repeats in both disease-associated and non-associated loci (Hancock). This might provide a selective advantage to genes containing repeats because of a decreased rate of deleterious mutation (Hancock).

Population analyses of anonymous (repeats whose location with respect to genes is uncharacterized), gene-associated and disease-associated triplet repeats (Ranjan Deka, Cincinnati) reveal a striking difference between gene-associated (and anonymous) repeats and disease-associated repeats. Disease-associated repeats show more alleles and higher heterozygosity, allele size variance and mutation rates than other repeats, and show a tendency towards expansion while other repeats tend to contract.

An intriguing question remains of how repeats originate within genomes. Hanah Margalit described analyses which showed a strong association of many classes of repeat, particularly A/G rich ones, with transposons, suggesting that they arise during the integration process or as poly(A) tails from reverse-transcribed RNAs (Alec Jeffreys, Leicester). An alternative suggestion by Miroslav Radman (Paris) was that microsatellites could originate as 'ancestral glue'—self-complementary repetitive sequences are ideal for repairing double-strand breaks if other sequence information is absent. Distributions of repeats within genomes might also be influenced by their location with respect to replication origins (Hancock) (repeats are exclusively found associated with replication origins in mitochondria; Rus Hoelzel, Durham) and local levels of mutability (Santibáñez-Koref; Hancock).

Mechanisms of mutation at repeats

Micro- and minisatellites are generally distinguished on the basis both of their basic motif length and their mutational mechanisms, microsatellites being thought to mutate primarily by strand slippage during replication and minisatellites by recombinational processes. The well-studied minisatellite locus MS32 mutates almost entirely by recombinational processes (Jeffreys) but others, such as MSY1 on the human Y chromosome (Mark Jobling, Paul Taylor and Matthew Hurles, Leicester) mutate by a more slippage-like process. Craig Primmer (Helsinki) provided examples of sequences that appear to be crossing the boundary between the two classes, as well as showing differing modes of mutation. David Leach (Edinburgh) showed that recombination can be stimulated by the formation of unusual DNA structures during replication of CXG microsatellites.

Alec Jeffreys described studies on MS32 using a novel PCR method to carry out allele-specific, single-cell analyses of recombination. Recombination at MS32 is asymmetric, with different alleles showing different frequencies. He showed that a recombination hotspot lies at the 5′ end of the minisatellite array. A single point mutation in this hotspot abolishes the ability of the mutable allele to act as an acceptor of recombination. This emphasized how elusive the influences of genomic context on the mutational dynamics of such sequences can be.

It is intriguing in this context that Fragile X CGG repeat arrays show polarity of mutation (James Macpherson, Salisbury; Eichler) in a manner reminiscent of minisatellites. Mark Hirst (Oxford) showed that, in a *Saccharomyces cerevisiae* model system, instability in CGG repeats was orientation dependent. However, most of the expansions observed appeared to arise during the integration process, perhaps reflecting the repair of single-stranded intermediates. Deletions, on the other hand, appeared to be due to recombination with motifs in flanking sequences. In human populations, the Fragile X array shows a wide variety of mutational patterns, apparently depending on the chromosomal haplotype on which the array resides (Macpherson). So complex is this pattern that single-nucleotide polymorphisms (SNPs) may provide a better basis than the characteristics of the repeat itself for predicting the likelihood of mutational change at a given allele. Some rare Fragile X haplotypes may be particularly prone to rare types of mutational event, so-called concatenated mutation (Anna Murray, Salisbury). This kind of complexity also seems consistent with complex interactions between genomic context and mutational process. Adding to this complexity, Miroslav Radman pointed out that the polymorphism at microsatellites could have the effect of inhibiting homologous recombination.

The question of the relative importance of intra- and interallelic processes in microsatellite evolution is also reflected in one of today's most controversial topics in microsatellite biology—do the skews observed in the length distributions of microsatellites, and the apparent tendencies of these repeats to expand during evolution in some lineages (David Rubinsztein, Cambridge; Bill Amos, Cambridge; Eichler; Hancock), reflect an effect of heterozygosity in increasing microsatellite mutation rate (which could be mediated by way of increased mutation rate in individuals with large interallelic differences in length) (Amos), or is a dependence of mutation rate on array length sufficient to explain these effects? Christian Schlötterer (Wien) pointed out that individuals with long microsatellite alleles are also likely to have large length differences between alleles and David Goldstein (Oxford) mentioned that microsatellite mutations show a threshold effect characteristic of a slippage-like process. Rosalind Harding (Oxford) suggested that the observed skews in microsatellite allele length could be explained by length dependence of mutation rates. The observation that the average CAG repeat is longer in rodents, and especially rats, than in humans provides a cautionary note that general conclusions should not always be drawn from studies of subsets of data (Hancock). Another intriguing suggestion is that microsatellite length distributions seem to show an upper size limit (Amos). This may be due to a length threshold above which arrays are prone to deletion (Amos), or it might reflect the interplay between different length-dependencies of insertion and deletion rates (Richard Sibly, Reading).

Despite these uncertainties, techniques of phylogenetic reconstruction using microsatellite lengths can be remarkably successful. David Goldstein outlined an analysis of a Y-specific microsatellite in Jews who claimed to be members of the Cohen priestly caste, whose status is inherited down the male line. These individuals were enriched for a particular compound haplotype whose coalescence time was estimated at 3000 years ago, a date corresponding closely to the construction of the first Temple in Jerusalem.

Long, disease-causing repeats would be expected to be lost from populations through selection and genetic drift. Ranajit Chakraborty's presentation, given by Marek Kimmel (Houston), suggested that segregation distortion at the myotonic dystrophy locus is sufficiently strong to maintain the frequency of pre-mutation-length alleles in populations. This effect was estimated to be significant in female but not in male transmissions. Small-pool PCR analysis of sperm showed no evidence of segregation distortion in male carriers of expanded alleles at the same locus (Darren Monckton, Glasgow).

Molecular genetic basis of repeat instability

Long, disease-causing repeats show dynamic mutation, that is extreme meiotic instability. An explanation for this requires mechanisms other than standard strand slippage during DNA replication because it is characterized by large jumps in repeat length which are uncharacteristic of slippage. The driving force in most analyses of this process is the idea that tandem repeats of CAG and CGG can form stem–loop structures that give rise to replication errors. For example, formation of non-standard structures during replication might give rise to recombinational repair at double-strand breaks, which in turn might give rise to dynamic mutation (Leach). Alternatively, dynamic mutation might result from an inability of the appropriate nucleases to process structures formed within the 5′ flaps of Okazaki fragments (Gordenin).

Christopher Pearson (Houston) provided the most direct evidence for the formation of such structures (S-DNA, for slipped-strand DNA) in the form of electron micrographs of irregular structures formed by repeat sequences. These structures were bound by hMSH2, a component of the human mismatch repair system. He also showed that interruptions in the repeat structure reduced the formation of stem–loops, consistent with the stabilizing effect of repeat interruptions at disease loci. Mismatch repair has been shown to be important for the repair of replication errors at repeats in both *Escherichia coli* (Kristina Schmidt, Edinburgh) and *S.cerevisiae* (Gordenin), but this is more relevant to regular strand-slippage processes.

There have been a number of indications of the importance of the transcriptional status of repeats to their stability in model systems. The formation of stem–loop structures in *E.coli* has different effects on plaque growth depending on the type of repeat, its length (and in particular the length of the terminal loop of the structure) and, for CGG repeats, the reading frame in which it is incorporated (Leach). In *E.coli*, transcription into a repeat substantially increased its instability (Richard Bowater, London) while James McClellan (Portsmouth) showed that unusual RNA structures also induce pausing during transcription. Formation of stem–loop structures also affects the processivity of methyltransferase (Steven Smith, Duarte).

DNA repeats can form more complex structures than simple stem–loops, such as triplexes or quadruplexes, which might contribute to their instability (Karen Usdin, Bethesda). However, when sequences capable of forming such structures are put into transgenic mice, they do not show dynamic mutation (Usdin; Monckton), suggesting that effects other than DNA structure, such as repair efficiency and/or genomic location, may influence the instability of repeats. The longer mean length of CAG repeats in rodent genomes might also indicate that these species are more tolerant of long repeats than are humans (Hancock). Despite this, transgenic models can have notable successes in replicating the properties of the human diseases. Darren Monckton described a transgenic mouse line which showed tissue-specific, age-dependent and expansion-biased somatic instability in a manner reminiscent of that seen in humans. This effect was seen in one line out of three, which again may reflect an importance of genomic location to repeat instability.

Conclusions

The existence of trinucleotide expansion diseases has stimulated a massive amount of research which is providing many new genetic insights, particularly at the molecular and population levels. This has shown us that the processes underlying the mutation and evolution of simple DNA sequences are more complex than was previously thought. A detailed understanding of the basis of the trinucleotide expansion diseases provides a previously unavailable insight into the kinds of selective factors that might affect simple sequence evolution. On the other hand, evolutionary ideas about the origins and potential functions of simple sequences can inform our understanding of the origins, persistence and potential functional significance of trinucleotide repeats.

Acknowledgements

We thank Philippe Djian for pre-publication access to his report on this meeting and Keith Johnson for his comments.

References

Kremer,E.J., Pritchard,M., Lynch,M., Yu,S., Holman,K., Baker,E., Warren,S.T., Schlessinger,D., Sutherland,G.R. and Richards,R.I. (1991). Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science*, **252**, 1711–1714.

La Spada,A.R., Wilson,E.M., Lubahn,D.B., Harding,A.E. and Fischbeck,K.H. (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, **352**, 77–79.