Letter to the Editor

Linkage Disequilibrium and Molecular Drive in the rDNA Gene Family

There is currently much debate on the degree of linkage disequilibrium in multigene families for it is an indication of the differences between individuals (on which selection may act) in the extent of homogenization of a mutation by the genomic mechanisms of turnover that underpin molecular drive (SEPARACK, SLATKIN and ARNHEIM 1988; NAGYLAKI 1988). Much of the debate has been theoretical leading to different conclusions, depending on adjustments of the starting assumptions. Recently, however, SEPARACK, SLATKIN and ARNHEIM (1988) provide interesting data on linkage disequilibrium in the human rDNA family, from which they suggest "that sister chromatid exchanges are much more important than homologous or nonhomologous recombination, and that molecular drive may not apply to the evolution of the rDNA family." Is this suggestion justified? And what are the factors that influence the interaction between molecular drive and natural selection?

Two of the seminal publications on the evolutionary dynamics of the rDNA multigene family by Norman Arnheim and his colleagues (ARNHEIM *et al.* 1980; KRYSTAL *et al.* 1981) showed that: (1) species diagnostic mutations have spread through the rDNA of several ape species (a pattern of distribution known as concerted evolution—for reviews see DOVER 1982; ARNHEIM 1983); (2) such homogenized mutations could be found in all five tandem arrays located on five pairs of nonhomologous chromosomes in human individuals; (3) fluctuations in copy number of whole rDNA units, and of an array of subrepeats within the IGS (intergenic spacer) of each unit (spacer length variants), reflect the activities of unequal crossing over; (4) several, but not all, partially homogenized mutations (observed as rDNA units with and without a given mutation) were not restricted to one array.

KRYSTAL *et al.* (1981) concluded from these studies that "analysis of the distribution of three human rDNA polymorphisms among individual nucleolus-organizer-containing chromosomes is also consistent [along with the homogeneity patterns of species-specific mutations] with the model of genetic exchanges among rDNA sequences on nonhomologous chromosomes. Nucleolus organizers on nonhomologous chromosomes generally share the polymorphic forms found in the population as a whole." Without actually calling it such, ARNHEIM and colleagues were describing the process which came to be known as molecular drive. This is the process for spreading a variant repeat both through a family (homogenization) and, concomitantly, through a sexual population (fixation) as a consequence of continual fluctuations in the copy number of repeats within individuals produced by a variety of biased and unbiased mechanisms of genetic turnover (one of which is unequal crossing over) operating *within* and *between* homologous (and where necessary nonhomologous) chromosomes.

The rDNA of other species, in addition to many other repetitive families, also have both completely homogenized and polymorphic mutations shared by homologous and nonhomologous chromosomes (STRACHAN, WEBB and DOVER 1985). Additionally, in *Drosophila melanogaster* all *X* and *Y* chromosome rDNA arrays in the species share fully homogenized mutations which coexist with a specific mutation restricted to the *Y* chromosome arrays (COEN, STRACHAN and DOVER 1982).

How do we reconcile such seemingly paradoxical observations of fully and partially homogenized mutations shared between chromosomes coexisting with occasional mutations limited to a single chromosomal array (observed as linkage disequilibrium) in one and the same gene family?

In my view the answer is probably simple: interpretations of differences in mutant gene distribution should consider both the ages and locations of mutations *vis-à-vis* exchange breakpoints, and not just differences in rates of unequal exchange around the karyotype (for discussion see COEN and DOVER 1983). A chromosome-specific mutation (by which I mean a mutation shared ideally by all chromosomes of a given type) can be considered as a relatively recent event which has only had time to be homogenized and fixed by a combination of unequal exchanges between-sister chromatids and between-homologous chromosomes. A mutation shared around the karyotype would be older and would have existed long enough to have been transferred from array to array by rarer nonhomologous chromosome exchanges. A mutation limited only to chromosome haplotypes could be considered either as the youngest of all or as having occurred in a position that is unlikely to switch to a nonsister chromatid, no matter how many such exchanges might be occurring (see below). Calculations of expected distributions of mutations at different periods of time need to consider their times and places of origins as well as differentials in rates, units, biases and locations of unequal exchange around the karyotype, the size of the rDNA family, and the size and breeding structure of the populations (DOVER 1982, 1986; OHTA and DOVER 1983, 1984). A blanket homogeneity is neither expected nor observed under molecular drive, because the process is not instantaneous.

What is known about variant rDNA gene distribution in humans, and what caution needs to be exercised in the interpretation of this knowledge? From the data presented in SEPARACK, SLATKIN and ARNHEIM (1988) it is clear that, in contrast to the patterns of sharing of previously described mutations, there is a mutation within a restriction enzyme cleavage site (*Bgl*II) approximately 2 kb from the end of a small array of subrepeats (each of 700 bp) within the IGS and that this mutation is in linkage disequilibrium with some spacer length variants that are themselves due to fluctuations in copy number of the 700-bp repeats (caused by unequal exchanges at the periodicity of the subrepeats). This linkage disequilibrium generates between-individual differences in the frequency of the *Bgl*II mutation, which SEPARACK, SLATKIN and ARNHEIM interpret as being inconsistent with molecular drive from which they claim "one would not expect between-individual differences in the frequency of a variant gene." It is difficult to see on what basis they derived such an expectation to which their observations do not fit. Expectation of variant gene distribution depends on the parameters listed in the previous paragraph, for which only one relevant figure is provided: the estimated time of origin of the *Bgl*II mutation being no younger than 50,000–500,000 yr before the present.

If the two markers used by SEPARACK, SLATKIN and ARNHEIM (1988) are to segregate independently and not show linkage disequilibrium then there is a requirement for frequent interchromosome recombination between the markers, assuming that the *Bgl*II mutation arose as a point mutation in one rDNA gene containing a particular length variant. This recombination event would need to occur during unequal crossingover at the longer periodicity of the whole rDNA unit. Hence, the frequency of this event would depend on the rate at which this level of unequal alignment occurred and on the probability that the break point took place within the 2-kb interval in an rDNA unit of approximately 47 kb. Rates of unequal crossing over should not be calculated from the linkage disequilibria data, for that would lead us into a circular argument. The rates have to be calculated independently. In Drosophila and yeast there are genetic ways for doing this (COEN and DOVER 1983; SZOSTAK and WU 1980). With regard to the probabilities of occurrence of breakpoints along the 47-kb rDNA unit, it is highly unlikely that they are equal given that the spacer is known to be composed of an heterogeneous mixture of sequences containing Alu repeats, simple-sequence DNA, single-copy DNA etc., and given that there is a breakpoint hotspot within a 55-bp sequence in each of the 700-bp subrepeats (ERICKSON and SCHMICKEL 1985). The location of these breakpoints *within* one of the genetic markers used for the analysis of linkage disequilibrium means that great care needs to be taken over the use of the methodology for measuring linkage disequilibria based on more traditional analysis of polymorphisms of single-copy genes. This is particularly problematical when one of the two markers being used is variation in the copy number of spacer subrepeats. This variation is due to the two levels of unequal crossing over (at the periodicities of the subrepeats and of the whole rDNA unit) which cause wide fluctuations in copy number of rDNA units and their internal repeats both between individuals (with respect to a given chromosome) and between all rDNA carrying chromosomes. In humans these differences range from tens to hundreds of repetitive units (KRYSTAL *et al.* 1981). Hence, many of the length variants could have arisen repeatedly and independently at each rDNA. Accordingly, it is not possible to evaluate estimates of D and D' *as if* the two marked morphs arose on one original chromosome as single unique events. Because of this it is difficult to understand the following statement by SEPARACK, SLATKIN and ARNHEIM: "That the value of D' in each population is different from 1.0 indicates that both polymorphisms have been present for sufficiently long in humans that there has been time for recombination to partially disassociate the length variants from the *Bgl*II polymorphism." Not only is this statement unquantifiable in the absence of knowledge of unequal exchange rates around the karyotype and of the probability of breakpoint events within the requisite 2-kb unit under scrutiny, as discussed above, but it is also unrealisitc because the generation of length variants has a totally separate dynamics from the unique point-mutational event that gave rise to the original single rDNA unit with the *Bgl*II mutation. The multiple occurrences of individual length variants also affect the large fluctuations between chromosomes and between individuals in the copy-number of rDNA units per array, which in turn will affect the statistics of expectations of linkage disequilibria (*i.e.*, expectations of individuals missing one rDNA "haplotype") against which observations are compared.

In order to understand the potential interaction between molecular drive and natural selection we need to consider the population dynamics of the former, in so far as this can be generalized. The population dynamics of molecular drive are inevitably complex, but at the very least they will be affected by the large differences in magnitude between the mutation rate ($\sim 10^{-6}$ per kilobase per generation), the rate of turnover (~10^{-2}–10^{-4} per kilobase per generation) and the fast rate at which the sexual process randomizes chromosomes at each generation. On this basis it was proposed (DOVER 1982; OHTA and DOVER 1984) that at any given generation the differences between individuals in the ratio of "old" to "new" variant copies of a gene family will be small *relative* to the large difference between the extreme beginning and end states of all genes being "old" or all genes being "new." The size of the variance of the ratio of "old" to "new" at any given moment depends critically on the parameters discussed above. SEPARACK, SLATKIN and ARNHEIM (1988) distort this broad working principle by writing that "the theory of molecular drive suggests that transposition, gene conversion and unequal crossing over are possibly more important than natural selection for the evolution of multigene families because soon after a variant arises molecular drive tends to limit the amount of variation among individuals on which natural selection can act." It has been made explicit from the beginning that the extent to which natural selection hinders, promotes or ignores a molecularly driven mutation and its phenotypic consequences depends on the ecology of the population and the ontogeny of the individuals. If the carriers of the new mutant genes cannot be tolerated then selection would be able to discriminate between individuals carrying different numbers of mutant genes no matter how small these differences might be (DOVER 1982). Second, it is clear that neither selection nor drift can be directly responsible for homogenization because they cannot transfer genetic information from one locus to another. For this we need to exploit the variety of turnover mechanisms that have been shown to be operating in all examined eukaryotic species. Hence, it can never be the case of genomic turnover being *more important* than selection *per se*, because they are not formally comparable. They are two distinct processes operating at different levels, whose contributions to the evolution of a multigene family are affected by different parameters, the details of which need to be explored case by case (DOVER 1987). Unfortunately, whereas the contributions of turnover are directly observable, the contributions of selection are not, except possibly in cases of "molecular coevolution" between a changing gene family and other functionally interacting gene products (DOVER and FLAVELL 1984; ARNHEIM 1983; HANCOCK and DOVER 1988; HANCOCK, TAUTZ and DOVER 1988).

Seemingly paradoxical differences in the distributions and frequencies of variant genes in one and the same multigene family are a common observation. They should not be interpreted as signifying the absence or presence of molecular drive in a crude polarization of interpretation; rather they indicate that there are a number of parameters that affected the within- and between-individual frequencies of different variant genes in a gene family from their inception through to their possible final homogenization and fixation. The extent to which selection plays with these different molecularly driven distributions depends on the complex parameters affecting a slowly changing population in its natural environment. The magnitude of individual differences in the degree of homogenization cannot be taken as evidence for or against either the internal genomic forces responsible for molecular drive or the external ecological forces responsible for natural selection. The expectation of genetic cohesion of a population throughout a period of change by molecular drive pertains to idealized conditions, analogous to the idealized conditions at the basis of the Hardy-Weinberg equilibrium of Mendelian systems (DOVER 1982). Observed deviations from expectations in both cases tell us something interesting about real biological systems; they are not a basis for negating the reality of the stochastic processes that went into the formulations of the null hypotheses.

GABRIEL A. DOVER
Department of Genetics
University of Cambridge
Cambridge CB2 3EH
England

LITERATURE CITED

ARNHEIM N., 1983 Concerted evolution of multigene families, pp 38–61 in *Evolution of Genes and Proteins*, edited by M. NEI and R. K. KOEHN. Sinauer, Sunderland, Mass.

ARNHEIM N., M. KRYSTAL, R. SCHMICKEL, G. WILSON, O. RYDER and E. ZIMMER, 1980 Molecular evidence for genetic exchanges among ribosomal genes on non-homologous chromosomes in man and apes. Proc. Natl. Acad. Sci. USA **77:** 7323–7327.

COEN, E. S., T. STRACHAN and G. A. DOVER, 1982 The dynamics of concerted evolution of rDNA and histone gene families in the *melanogaster* species subgroup of *Drosophila*. J. Mol. Biol. **153:** 841–870.

COEN, E. S., and G. A. DOVER, 1983 Unequal crossing over and the coevolution of X and Y rDNA arrays in *D. melanogaster*. Cell **33:** 849–855.

DOVER, G. A., 1982 Molecular drive: a cohesive mode of species evolution. Nature **299:** 111–117.

DOVER, G. A., 1986 Molecular drive in multigene families; how biological novelties arise, spread and are assimilated. Trends Genet. **2:** 159–165

DOVER, G. A., 1987 DNA turnover and the molecular clock, J. Mol. Evol. **26:** 47–58

DOVER, G. A., and R. B. FLAVELL, 1984 Molecular coevolution; DNA divergence and the maintenance of function. Cell **38:** 623–624.

ERICKSON, J. M., and R. D. SCHMICKEL, 1985 A molecular basis for discrete size variation in human ribosomal DNA. Am. J. Hum. Genet. **37:** 311–325.

HANCOCK, J. M., and G. A. DOVER, 1988 Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. Mol. Biol. Evol. **5:** 377–391

HANCOCK, J. M., D. TAUTZ and G. A. DOVER, 1988 Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster.* Mol. Biol. Evol. **5:** 393–414

KRYSTAL, M., P. D'EUSTACHIO, F. H. RUDDLE and N. ARNHEIM, 1981 Human nucleolus organizers on non-homologous chromosomes can share the same ribosomal gene variants. Proc. Natl. Acad. Sci. USA **78:** 5744–5748.

NAGYLAKI, T., 1988 Gene conversion, linkage and the evolution of multigene families. Genetics **120:** 291–301.

OHTA, T., and G. A. DOVER, 1983 Population genetics of multigene families that are dispersed into two or more chromosomes. Proc. Natl. Acad. Sci. USA **80:** 4079–4083.

OHTA, T., and G. A. DOVER, 1984 The cohesive population genetics of molecular drive. Genetics **108:** 501–521.

SEPARACK, P., M. SLATKIN and N. ARNHEIM, 1988 Linkage disequilibrium in human ribosomal genes: implications for multigene family evolution. Genetics **119:** 943–949.

STRACHAN, T., D. A. WEBB and G. A. DOVER, 1985 Transition stages during molecular drive in multiple copy DNA families in Drosophila. EMBO J. **4:** 1701–1708.

SZOSTAK, J. W., and R. WU, 1980 Unequal crossingover in the ribosomal DNA of *Saccharomyces cerevisiae.* Nature **284:** 426–430.