just one of a number of opportunistic tumors that may affect these immunosuppressed individuals. He and his colleagues have now found three individuals, young gay males with characteristics similar to those of the AIDS patients, who have a squamous carcinoma of the tongue. This cancer is rarely seen in young nonsmokers like these patients. One of the men was a lover of a patient with Kaposi's sarcoma. In addition, the San Francisco group has identified four AIDS patients who have a Burkitt's-like lymphoma.

Ziegler says, "The question arises—are these individuals susceptible to the cancers because of the activation of DNA viruses that are passed between the individuals?" All three of the cancers have been linked with members of the DNA-containing herpes virus family. Oral cancers, such as the squamous cell carcinoma, have been associated with HSV, Burkitt's lymphoma with Epstein-Barr virus (EBV), and, as already noted, Kaposi's sarcoma with CMV.

In fact, the Ziegler group has identified CMV DNA and proteins in sarcoma cells from AIDS patients, but not in normal cells from adjacent tissue. And they have found EBV in tumor cells from two of the lymphoma patients. Ziegler suggests that the immunosuppression of the patients may have allowed activation of the viruses, thus leading to the cancers. If his hypothesis is borne out, there would be another link for the chain of evidence supporting a causative role for the herpes viruses in cancer.

In addition to a possible role of the viruses in the etiologies of these cancers, there appears to be a genetic element influencing who gets the cancer, at least for Kaposi's sarcoma. According to Friedman-Kien, Laubenstein, and Rubinstein, there is an association between the sarcoma and a particular HLA antigen, the one designated DR5. The DR antigens, which are encoded by genes in the major histocompatibility complex, are thought to be involved in the regulation of immune responses. "Between 50 and 60 percent of the patients have HLA-DR5 in the classic as well as the homosexual variety of Kaposi's," says Friedman-Kien. "This indicates a genetic predisposition." Exactly how the DR5 antigen predisposes to Kaposi's sarcoma is not understood. Safai and Marilyn Pollack of Memorial Sloan-Kettering Cancer Center have similar findings.

In general, AIDS is providing researchers with a wealth of clues for investigating how the immune system works normally and how its malfunction can result in disease, including cancer. Meanwhile, a major effort is under way at CDC and elsewhere to pinpoint the agent or agents that cause the disease. "Identification of the cause and then prevention are the major goals," Gottlieb says.—JEAN L. MARX

Repeated DNA Still in Search of a Function

As new families of DNA sequences are discovered the picture becomes more complicated, more fascinating, and more mysterious

A large and varied zoo of repetitive DNA sequences from eukaryotic organisms was on view at a workshop* held in mid-July at the National Institutes of Health (NIH), Bethesda. "It is truly astonishing that even in 1982 people can still report the discovery of a family of 100,000 members of a repeated sequence that makes up 3 percent of a mammalian genome." Thus comments Giorgio Bernadi, of the University of Paris, who organized the meeting jointly with Maxine Singer of the National Cancer Institute. "It's clear there are many more families to be found," adds Singer, "and the discovery process is moving at a tremendous pace."

Interest in repetitive DNA sequences goes back many years but, as with many aspects of molecular biology, the advent of recombinant DNA technology and DNA sequencing now permits previously unmatched scrutiny of the structures of interest. It was therefore not surprising that the NIH meeting showed a heavy emphasis on new structural, rather than functional, information. For example, new repeat families were described, sequences of known ones were clarified, relationships between families were explored, and so on. Singer expressed a common frustration when she said, "We all go on grinding out the data on structure without thinking enough about what it means."

The truth is, however, that the functions of the large and motley collection of repeated DNA families are proving particularly resistant to elucidation. Putative functions are many, including, variously, involvement in chromosome pairing, control of gene expression, processing of messenger RNA precursors, and participation in DNA replication. So far none has been established, save for the single exception of a small family that gives rise to 7*S* RNA, a molecule that recently was serendipitously discovered to be an essential component of a particle that mediates the secretion of proteins from cells.

For Eric Davidson, of the California Institute of Technology, a key message of the NIH workshop was the inference that many families of repeated sequences dispersed throughout the genome have been mobile through evolutionary time. "The evidence for mobility is indirect, but compelling," he suggests. If mobility is a reality, and most agree that it probably is, then it seems likely that at least some members of repeat families will have important effects in the genome, even if they have no formal function. Enhancing recombination and altering rates of gene expression are obvious possibilities, while the initiation of new species is a more recondite proposal.

Some repetitive DNA will undoubtedly be shown to have a function, in the formal sense; some will likely be shown to exert important effects; and the remainder may well have no function or effect at all and can therefore be called selfish DNA. Repetitive DNA constitutes a substantial proportion of the genome (up to 90 percent in some cases), and there is considerable speculation on how it will eventually be divided between these three groups. Current bets would put a small fraction in the function category, with distribution of the rest rising steeply through the effect and selfish categories.

Eukaryotic DNA can conveniently be divided into three classes. First is single copy sequence DNA, which contains but is not exclusively composed of, protein coding sequences. Second is moderately

*International Workshop on Highly Repeated DNA Sequences in Eukaryotes, sponsored by the National Cancer Institute and the Fogerty International Center, 12 to 14 July.

0036-8075/82/0813-0621$01.00/0

repetitive DNA, some families of which are dispersed throughout the genome while others are clustered in tandem repeats at the centers (centromere) and ends (telomere) of the chromosomes. In addition, members of functional genes can formally be placed in this group, such as those for ribosomal RNA, transfer RNA, histones, actin, β-globin, and immunoglobulins. But for the most part the class is composed of sequences of unknown function repeated up to 10^5 times and dispersed around the genome.

Third is the highly repetitive DNA, sequences repeated a million or more times, some of which are dispersed while most are clustered at centromeres and telomeres. This group of clustered sequences is often known as satellite DNA and is thought by some to be influential in the pairing of chromosomes both in cell division and in the fusion of male and female gametes.

Compared with the middle repetitive sequences, satellite DNA is generally rather simple. Although the amount of satellite may vary dramatically in quantity between different species, its complexity as measured by the number of different families of repeat sequences it contains is usually, but not always, low. For the most part satellite is not transcribed into RNA, and those instances where it is might very well be the result simply of readthrough from the transcription of legitimate genes, such as histones, embedded in the repeated sequences.

The long tandem arrays of satellite DNA are not, incidentally, immune to invasion by the apparently mobile dispersed middle repetitive sequences. Singer noted the presence of members of three such repeat families—one 300 base pairs long, the second 875 base pairs, and the last close to 6400—within the predominant satellite DNA of African green monkey.

As might be expected, there are echoes of structural similarity between the satellite sequences of related species. More striking, however, is the clear distinction between the satellite of one species and that of another, no matter how closely related. Does this picture merely reflect a high rate of change of satellite sequences, which obviously must be faster than the rate of speciation? Or, given the putative role of satellite in chromosome pairing, could a shift in the identity of satellite composition in a geographically isolated population of organisms initiate incipient speciation through effective reproductive isolation?

The uniformity of sequence within families of satellite DNA, and within families of dispersed repeats too, in any one species is striking. There must be mechanisms that homogenize the sequences against the constant tendency to diverge. As a consequence, the clustered and dispersed classes of repeated sequence display a remarkable constancy of sequence against a background of steady modification in much of the single copy sequence DNA.

The sharp distinction in sequence between satellites of related species, and to a somewhat lesser extent in dispersed sequences too, indicates that when a change does occur in repeat families, it runs through very rapidly. Not only does the sequence of repeat families appear to shift abruptly, but so too does the number of members in the family, again as indicated by differences between species. "Some of these repeat families seem to explode in number when new species form," observes Davidson.

"Some of these repeat families seem to explode in number when new species form."

The long tandem repeats of satellite DNA allow explanations in direct mechanical terms of some of the properties of the group. For instance, when homologous chromosomes pair there is only a small chance of perfect matching of the tandem arrays, one against the other. The extensive repeats allow for imperfect pairing within the arrays and subsequent unequal crossing-over. The result of unequal crossing-over is reciprocal amplification and deletion of sequences. In addition to fluctuation in the number of repeat units, unequal crossing-over also generates sequence homogeneity in tandem arrays. The same process can apparently occur between nonhomologous chromosomes too, thus spreading homogeneity in the repeat sequences wherever they are located in the genome.

A second process of sequence homogenization is known as gene conversion, and this is not a consequence of tandem repeats. In some as yet to be determined fashion, similar sequences on the same chromosome, or on homologous or nonhomologous chromosomes, are compared against each other and any differences between them may be corrected. If the repair were a stochastic process, sometimes one sequence being corrected, sometimes the other, then the process of homogenization or the spread of a new sequence through a family would be a question of drift.

If, however, as Gabriel Dover of the University of Cambridge, England, argues, correction is sometimes biased in one direction because of favorable chemical or mechanical equilibria, then the process would constitute what he calls "molecular drive." Dover argues that molecular drive might sometimes be important in initiating speciation through bringing about a concerted shift in structure in a family of repeated sequences within a population of individuals.

Satellite DNA unquestionably is a puzzle. What determines the number of copies in a repeat family? And how does the genome tolerate so much of it? Perhaps, as Singer has recently promulgated, just a small fraction of the satellite sequences is essential to some genomic function while the remainder is harmless surplus. This, she indicates, is a comfortable middle ground between the extreme selfish DNA position, which sees no function in all this "junk DNA," and the adaptationist position, which looks for functions in every structure. The same questions and speculations can be applied to dispersed repetitive DNA.

A number of patterns are emerging from a comparison of dispersed repeated sequences in higher and lower eukaryotic organisms. For instance, the sea urchin, about which most is known through the work of Davidson, Roy Britten, and their colleagues, has several thousand repeat families, some of which have as few as 100 members. Mammals, by contrast, appear to have relatively few families, each of which contains many thousands, sometimes hundreds of thousands, of members. One family, the Alu family, which in humans has between 3×10^5 and 5×10^5 repeats of a 300-base-pair sequence, has been the prototype of the large dispersed repeat family in mammals. It is found in primates and in rodents (where it measures 135 base pairs in length).

Important as Alu is, one effect of the NIH meeting was to set it in better perspective, with the description of several new very abundant repeat families in mammals. For instance, Hans Zachau of the University of Munich reported a 475-base-pair repeat in mouse, with an estimated family number of 10^5.

Singer discussed some preliminary data on two new repeat families in primates. One of them, designated Ret, measures 875 base pairs and may be repeated about 2×10^4 times in the Afri-

can green monkey genome. Another, as yet of undetermined length, may be substantially more numerous. Nicholas Hastie, of Roswell Park Memorial Institute, Buffalo, reported three relatively short repeat families in mouse to genome, each measuring between 200 and 500 base pairs, and each is repeated between 7×10^4 and 1.8×10^5 times. Many participants talked privately about early data on still other families.

"I was impressed by all these reports," says Carl Schmid of the University of California at Davis. "I hadn't expected to see this kind of information come up so quickly," Even with the hegemony of Alu broken by the appearance on the scene of new important repeat families, the picture in mammalian genomes remains distinct from that in sea urchins. Davidson warns, however, that beneath the impressive presence of a few large families might lurk a large number of small, functionally more important, families. Schmid points out that just three families, the Alu, and Kpn (a 6400-base-pair family), and a diverse family designated U, constitute half of the repeat families in the human genome. He admits, however, that these could be "a smokescreen hiding a considerable number of small families, like those in the sea urchin."

Even if the comparative pattern between the sea urchin and mammals is not as distinct as it appears to be on current data, it is apparent that the lower organism does not possess the very large families so characteristic of mammalian genomes.

One distinction that can be made when looking at dispersed sequences as a whole is the length of the repeat unit. Singer recently suggested the notation SINES, for short interspersed repeated segments, and LINES, for long interspersed repeated segments. SINES are typically less than 500 base pairs long and are repeated some 10^5 times. LINES, which until 3 years ago had been unknown in mammals, measure more than 5000 base pairs in length and are repeated perhaps 10^4 times in the genome. "This notation at least has the benefit of getting away from Alu-like and non-Alu as descriptions of other families," says Singer. Inevitably, though, as more and more sequences are found some begin to fall in the gap between the two groups.

Another distinction between repeated units that, as Schmid notes, might make a useful comparison is the nature of the flanking sequences. In *Drosophila* and yeast, for example, there are small families of long repeat units that are flanked first by long terminal repeats and then by short direct repeats that seem to indicate the insertion into a genomic target site. There is now no doubt that these families are transposable elements that carry their own molecular apparatus for excision from and insertion into the genome. Moreover, these elements have significant effects on the activity of both single genes and large multigene loci.

So far no repeat family in mammals, amphibians, or sea urchins has been shown to possess these characteristic features of transposability. And yet there is a widely held belief that the dispersed sequences are mobile. One piece of indirect evidence is the fact that they are all flanked by short direct repeats, putative target sites in the genome that double up on insertion of the repeat element. Earlier this year Singer and Giovanna Grimaldi presented persuasive evidence of the duplication of a target site that now flanks an Alu sequence embedded in the α-satellite of the African green monkey genome.

Other indirect evidence for mobility includes the presence of a short repetitive sequence in an intron of the rat growth hormone gene and its absence in the equivalent place in the human gene and the different distribution of SINES in the clusters of globin genes.

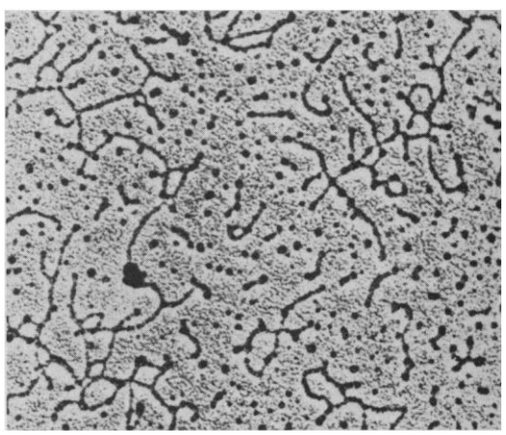
Davidson has for some time been interested in the notion of dispersed repeated sequences "diffusing around the genome in evolutionary time," taking with them elements that might control gene expression. He noted the recent discovery of a similar 21-base-pair element some 400 nucleotides upstream from three different genes, each of which is under glucocorticoid control. Does this small sequence function as a receptor in the control of gene expression? And is it the kind of element that might diffuse around the genome, its insertion at a propitious site bringing other genes under glucocorticoid control?

Gerald Fink, of Cornell University, described an 8-base-pair element upstream from yeast genes involved in amino acid synthesis. In this case the element is demonstrated to be important for gene expression. The same speculations can be applied here.

"The puzzle is," says Davidson, "that the important controlling elements we know about are small, much smaller than the typical member of a repeat family." Perhaps, he suggests, the repeat units are vehicles for controlling elements. In which case it might be rewarding to look for signs of transposition in areas that flank known controlling elements.

One observation that might be taken as evidence of function in repeated sequences is the frequency of transcription into RNA. A significant proportion of nuclear RNA contains transcripts of repeated sequences, although 90 percent of this is lost in RNA processing and exit to the cytoplasm. Davidson and his colleagues have shown that in sea urchin the spectrum of repeat families that are transcribed changes during development, an appealing argument for some regulatory function. Most intriguing, however, is the discovery that only a small proportion of any repeat family is ever transcribed. "Most members appear to be quiescent, which must make you cautious when isolating samples in search of their function."

There is a persistent theme here that just a small fraction of repeat families is functional, the balance being present as a side consequence of amplification processes that give a product that has no function but can be tolerated. Alan Weiner of Yale University described results from work on the U repeat family which indicate the production of incomplete copies of putative functional genes which are then inserted back into the genome as inactive members of the family. The process involves reverse transcription, from RNA to DNA, as does his suggestion for the spread of the Alu family. "I find this particularly exciting," says Schmid. "It may be that we shall be able to describe many of the repeat families as multiple pseudogenes."



Interspersed repeat sequences in maternal messenger RNA of sea urchin eggs can be detected by renaturation experiments.

It is clear that, from their abundance, their unusual structure, and their frequent transcription, dispersed repetitive DNA families cannot be ignored. But it is equally clear that for the most part they, like their tandemly repeated relatives, remain a phenomenon in search of a function.—ROGER LEWIN