Proc. Nat. Acad. Sci. USA
Vol. 70, No. 2, pp. 535–537, February 1973

A Proposal for the Structure of the *Drosophila* Genome

(repetitive DNA/unique DNA/chromomere/circle formation/chromosome structure)

JAMES BONNER AND JUNG-RUNG WU*

Division of Biology, California Institute of Technology, Pasadena, Calif. 91109

Contributed by James Bonner, December 11, 1972

ABSTRACT **We propose a structure for the genome of *Drosophila melanogaster* in which each chromatid of each chromomere (band) consists on the average of about 30–35 different sequences of single-copy (unique) DNA, each on the average about 750 base pairs in length. These are separated from one another by stretches of the middle repetitive (reiterated) DNA, which in *D. melanogaster* makes up about 15% of the genome. These stretches are about 100–150 base pairs in length and are all of the same sequence or family in each individual chromomere and of a different family (sequence) in each different chromomere. Our proposed structure of the *Drosophila* genome is in accord with all of the known facts concerning the physical chemistry and molecular biology of *Drosophila* DNA.**

It is known that the haploid genome of *Drosophila melanogaster* contains 0.12 pg of DNA or about 1.2×10^8 base pairs. Of this amount, about 6% is serially repetitive (1)†, with a repetition number of several thousand, and is centromeric (2, 3). We do not here concern ourselves further with this portion of the genome. Of the remainder, about 79% (of the total genome) consists of single-copy or unique DNA, presumably the structural genes, while 15% consists of sequences that are repetitive. The latter portion, sometimes referred to as the middle repetitive DNA, consists of sequences that are repeated an average of about 30–35 times per haploid genome, although this average includes family sizes from about 10 to about 100 (1, 2, 4).

Further facts

We have previously described the results of electron microscopic studies of the size and distribution of the middle repetitive sequences of *Drosophila* DNA (1)†. The middle repetitive sequences are short, about 100–150 base pairs in length, and are dispersed throughout the single-copy DNA of the genome, each repetitive sequence being separated from the next by, on the average, 750 base pairs of single-copy DNA. Since we know the amount of DNA contained in the middle repetitive sequences ($0.15 \times 1.2 \times 10^8 = 1.8 \times 10^7$ base pairs per genome) and the average length of these sequences (125 base pairs), we can calculate their number ($1.8 \times 10^7/125$) to be 1.44×10^5. Since each family of middle repetitive sequences consists of 30–35 similar or identical members, we calculate that the number of families of different sequence is ($1.44 \times 10^5/32.5$), about 4500.

A great deal is also known about the *Drosophila* genome from both genetics and cytology. Thus, the giant polytene chromosomes of the *Drosophila* salivary gland are organized into discrete dense bands or chromomeres, separated by less dense interbands. The bands consist of densely packed DNA, each of the some 2000 chromatids that it contains being packed to a density about 70-times that of the contour length of the DNA included in that chromatid (5). The interbands, on the contrary, contain DNA of a density consistent with the concept that each chromatid consists of an extended single DNA double helix (6).

The number of chromomeres, or bands, in the *D. melanogaster* genome is said to be of the order of 3500–5000 (6). Let us assume that the real number is an average of these two figures, namely 4250. Although the chromomeres or bands vary in size, they contain on the average ($1.2 \times 10^8/4.25 \times 10^3$) 28,000 base pairs of DNA or sufficient to code for about 30–35 enzyme molecules of average molecular weight.

Our model

We now suggest the proposition: each chromomere (and we now define chromomere as band and accompanying interband) contains on the average about 30–35 different single-copy DNA sequences, each accompanied by a middle repetitive segment 100–150 base pairs in length; the latter are all of the same sequence family. This proposal is suggested, in the first place, by the coincidence between the number of families of middle repetitive DNA sequences in *D. melanogaster* (4500) and the number of chromomeres in the *Drosophila* genome (about 4300). Our proposal is further supported in the most dramatic and remarkable manner by the work of Thomas *et al.* (7). These investigators have shown that if native eukaryotic DNA is sheared, the ends of the fragments are then caused to become single-stranded by treatment with an appropriate exonuclease (each end is resected optimally by about 450 bases); if the fragments are then briefly reannealed (to a criterion such that only repetitive DNA can reanneal), then circles are formed with a considerable frequency. This shows, as pointed out by Thomas *et al.* (7), that in eukaryotic DNA some segments of the genome are serially repetitious.

In contrast to the model of Thomas *et al.* (7), which proposes that the structural genes are serially repetitious (which is not compatible with the results of studies on the kinetics of reannealing of eukaryotic DNA), we propose that merely the interspersed repetitive DNA sequences are serially repetitious,

* Present address: Institute of Molecular Biology, Florida State University, Tallahassee, Fla. 32306.

† In this paper we report the observed length of the double-stranded middle repetitive segments, together with their accompanying bushes of collapsed single-stranded DNA, as equivalent to 150–200 base pairs. We consider our best estimate of the length of the repetitive segment itself to be 100–150 base pairs.

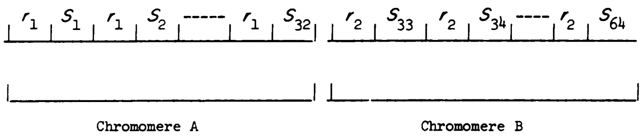

FIG. 1. Proposed model of the chromosome of *D. melanogaster*. Repetitious sequences are designated r_1, r_2, etc., each 125 base pairs in length, and single-copy sequences as S_1, S_2, etc., each on the average 750 base pairs in length.

and that this is so over a finite length of the genome, namely over the length of one band of the *D. melanogaster* genome.

Analysis

Our proposal is in agreement with the findings of Thomas *et al.* (7) concerning the size to which *Drosophila* DNA must be sheared in order to maximize circle formation (by the further treatments noted above). Let us suppose that each band consists on the average of 35 stretches of single-copy DNA, each 750 base pairs in length, and each accompanied by a repetitious sequence 125 base pairs in length and present in the band as 35 identical copies. The total contour length of the DNA in each chromatid of an average band would then be, therefore, 3.4 Å × 875 × 35 = 10.8 μm. On each side of our chromomere is another, similar in structure but with a different family of repetitious sequences. Let us then shear the DNA randomly and ask to what size must we shear to maximize the probability that two identical repetitious sequences will be contained in each fragment.

Clearly the minimum length must be equal to that of one of the average stretches of single-copy DNA plus two adjacent repetitive sequences. This length is (750 + 2 × 125) × 3.4 Å = 0.34 μm. As the length is increased, probability of ultimate circle formation should increase until the length of the fragment becomes such that, on the average, it includes DNA of more than one average band. We analyze the problem as follows: let us contemplate the *Drosophila* chromosome as modeled in Fig. 1. We will suppose that the chromomere (band) is made up of the number, c, of identical repetitious sequences of length, r, each attended by a single-copy (unique) sequence of length, u. The total length of the chromomere is then

$$\text{length of chromomere} = c(u + r). \qquad [1]$$

What is the probability, P_1, if we shear somewhere to the right of the beginning of chromomere A, a fragment of length F will include two r regions?

$$\text{If } F < u + 2r, P_1 = 0$$

$$\text{If } F = u + 2r, P_1 = 1/(u + r)$$

$$\text{If } F = u + 2r + 1, P_1 = 2/(u + r)$$

In general, $P_1 = (F - u - 2r + 1)/(u + r)$, or about $(F - u - 2r)/(u + r)$ for

$$2u + 3r > F > u + 2r. \qquad [2]$$

Next, let us imagine that an exonuclease resects our fragments of length F. Let e = resection length − r. What is the probability that the left end of our fragment starts within the distance e of an r? (We neglect effects at the ends of the chromomere in this approximation.) This probability will be

$$P_{2(\text{left})} = e/(u + r) \qquad [3]$$

There are periodic effects that depend on F as a function of $c(u + r)$ that we average over the chromomere. Therefore,

$$P_{2(\text{right})} = e/(u + r) \qquad [4]$$

And in the sum, for both ends of a fragment of length F the probability, P_2, that the exonuclease will expose two repetitious segments r is[‡]:

$$P_2 = [e/(u + r)]^2 \qquad [5]$$

We next ask, what is the probability, P_3, that the right end of a fragment will protrude beyond chromomere A and into chromomere B? We might imagine that circle-productive fragments cease at $F = c(u + r) + e$, or perhaps at $F = c(u + r)$ because we have already dealt with resection effects as P_2. Let us merely conclude that there are $c(u + r)$ starting (shearing) points of which F are excluded:

$$P_3 = \frac{c(u + r) - F}{c(u + r)} = 1 - \frac{F}{c(u + r)} \qquad [6]$$

The overall probability of circle formation, P_{CF}, from resected fragments of *Drosophila* DNA is therefore:

$$P_{\text{CF}} = P_1 \cdot P_2 \cdot P_3 \quad \text{or}$$

$$P_{\text{CF}} = P_1 \left(\frac{e}{u + r}\right)^2 \left[1 - \frac{F}{c(u + r)}\right] \qquad [7]$$

Let us remember that:

$$P_1 = 0 \text{ if } F < u + 2r, \text{ and that } P_1 = 1 \text{ if } F > 2u + 3r.$$

We plot P_{CF} as a function of F in Fig. 2, together with the relevant experimental data on circle formation as a function of F from Lee and Thomas (7). The predictions from our model correspond to the findings of Lee and Thomas (7) for *Drosophilia* DNA, with the exception that the maximum P_{CF} that they find is about 16.5% rather than the 14% predicted by our simple model. Suppose, however, that only 100 complementary base pairs are required for circle formation rather than the 125 that we have stipulated. In this case, our calculated maximum P_{CF} would rise to 16%.

The melting temperature of the rings formed by *Drosophila* DNA by the methods of Thomas *et al.* is lower than that of DNA of infinite length by about 4–5°C[§]. This value is in agreement with our suggestion that the reannealed ends are short, about 100–150 base pairs in length, rather than longer, as would be the case if the reannealed structures were structural genes.

[‡] The formulations of Eqs. **3**, **4**, and **5** are based on the assumption that r base pairs are required for stable ring formation. If the number required is in fact less than r, then in these three equations r should be replaced by the required number, for example $r/2$, etc.

[§] Thomas *et al.* conclude that the T_m of circle opening is less than that of infinitely long native DNA by only about 1°. We have replotted their data, placing more weight on the values $<T_m$ than they have done, and conclude that the true lowering is more nearly 5°.

DISCUSSION

The structure of the *Drosophila* genome suggested above is in accord with all of the facts concerning *Drosophila* DNA gathered by such diverse physical techniques as rate of reannealing of denatured DNA ("Cot curve") of Laird and McCarthy (4), Wu *et al.* (1), and others; electron microscopic analysis of repetitive sequence length and distribution of Wu *et al.* (1); and circle formation by reannealed, previously sheared, and resected DNA fragments of Thomas *et al.* and Lee and Thomas (7). We can ask next, however, is our model in accord with the genetic analysis of the *Drosophila* genome? Although most or all genetic evidence indicates that each chromomere operates as a single functional unit (8), it is nonetheless also true that a single chromomere may contain more than one distinguishable genetic locus. Thus the white locus, which is contained in one chromomere, has been subdivided into five subloci or pseudoalleles (9). These are distinguished from one another on the basis of the fact that each sublocus may crossover with the others. Mutants in each sublocus do not complement those within the same sublocus or those of other subloci. It is possible, therefore, that each chromomere may contain several or many such subloci that have not as yet been recognized. It may be particularly difficult to recognize them since, as pointed out above, chromomeres do operate in general as single functional units. In this connection it is of interest to note that transcription of the DNA of a band is apparently initiated simultaneously in all chromatids at a single point in that band, and proceeds from there (10). It is said for *Chironymus*, although this is not known to be true for *D. melanogaster*, that the RNA transcribed from a single chromatid of a single band (puffing band) may be as long as the contour length of the DNA of the chromatid of that band, that is, that the whole chromatid is transcribed as a single unit (11).

We now turn to the question, why is it that the middle repetitive segments of the *Drosophila* genome are interspersed among the unique segments in the way that we have found? There are of course many possibilities. Among these we point to the following:

(*i*) The middle repetitive segments may have to do with the way in which DNA is packed, densely, into the band structure; that is, that segment of DNA that contains a single family of repetitive DNA segments is that segment that packs into a single chromomere. How repetitive segments might lend themselves to packing we do not know.

(*ii*) The middle repetitive segments may have to do with the subsequence processing of the transcribed giant RNA into individual messenger fragments.

(*iii*) The middle repetitive segments may be secondary control elements of transcription, either modulating transcription within the chromomere or all responsible to a primary, but external (to the chromatid as here visualized), signal. A subset of this view would be that one, perhaps the first in the series of middle repetitive segments of a single chromomere, generates the signal that controls those further along the length of the chromomere.

(*iv*) The middle repetitive segments may be the sites (or one class of site) of crossingover. It is of interest in this connection that not only does crossingover occur in the white locus between pseudoalleles or subloci, but also that unequal crossingover can take place between pseudoalleles (12). According to the view proposed here of the structure of the genome, a middle repetitive segment of DNA must be present between each sublocus of the white locus. If crossingover occurs only within such segments, then unequal crossingover would be an occasional, but expected, result.

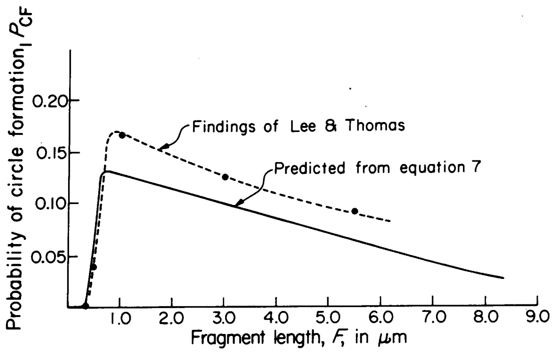

FIG. 2. Probability of circle formation (P_{CF}) from resected (by 450 bases at each end) fragments of *D. melanogaster* DNA of fragment length *F*. Calculated according to Eq. **7** in text. $(u + r)$ are assumed equal to 875 base pairs, *e* to equal 450 − 125 = 325 bases, and *c* to equal 35. The experimental findings of Lee and Thomas (7) for the same DNA are included.

We hope that our proposed structure may lead to a deeper understanding of the genetics of *Drosophila*.

We thank Profs. Eric Davidson, Norman Davidson, and Edward Lewis for their suggestions, advice and counsel, and Prof. Norman Davidson for his tutorials in statistical calculations. This work was supported in part by U.S. Public Health Service Research Grant GM-13762 and in part by a postdoctoral fellowship of the U.S. Public Health Service (J.-R. W.).

1. Wu, J.-R., Hurn, J. & Bonner, J. (1972) *J. Mol. Biol.* **64,** 211–219.
2. Botcham, M., Kram, R., Schmid, C. W. & Hearst, J. E. (1971) *Proc. Nat. Acad. Sci. USA* **68,** 1125–1129.
3. Rae, P. (1970) *Proc. Nat. Acad. Sci. USA* **67,** 1018–1025.
4. Laird, C. D. & McCarthy, B. J. (1969) *Genetics* **63,** 865–882.
5. DuPraw, E. H. (1970) *DNA and Chromosomes* (Holt, Rinehart & Winston, New York), p. 243.
6. Swift, H. (1962) in *The Molecular Control of Cellular Activity,* ed. Allen, J. M. (McGraw-Hill, New York), p. 73; Bridges, C. B. (1935) *J. Heredity* **26,** 60–64; Bridges, C. B. (1938) *J. Heredity* **29,** 11–16.
7. Thomas, C. A., Hamkalo, B. A., Misra, D. N. & Lee, C. S. (1970) *J. Mol. Biol.* **51,** 621–632; Thomas, C. A., Lee, C. S., Pyritz, R. E. & Bick, M. D. (1972) "Closing the rings," in *J. Gen. Physiol. Symp.*, in press; Lee, C. S. & Thomas, C. A., Jr. (1972) "Formation of rings from *Drosophila* DNA fragments," in press.
8. Judd, B. H., Shen, M. W. & Kaufman, T. C. (1972) *Genetics* **71,** 139–156.
9. Judd, B. H. (1959) *Genetics* **44,** 34–42.
10. Berendes, H. D. (1971) in *Control Mechanisms of Growth and Differentiation.* Society for Experimental Biology no. 25 (Academic Press, London), pp. 145-161.
11. Daneholt, B. J., Edström, E., Egyhazi, E., Lambert, B. & Ringborg, U. (1969) *Chromosoma* **28,** 399–417; *Chromosoma* (1969) **28,** 418–429; Daneholt, B. (1970) *J. Mol. Biol.* **49,** 381–391.
12. Judd, B. H. (1961) *Proc. Nat. Acad. Sci. USA* **47,** 545–550.