Genet. Res., Camb. (1967), **9**, *pp.* 141–147
Printed in Great Britain

An approach to the problem of whether clustering of functionally related genes occurs in higher organisms*

BY R. C. ELSTON AND EDWARD GLASSMAN

Departments of Biostatistics and Biochemistry, and the Genetics Curriculum,
University of North Carolina at Chapel Hill, U.S.A.

(*Received* 12 *April* 1966)

The clarification of the function of gene clusters in operons in bacteria has drawn attention to the possibility of similar clusters existing in higher organisms. Except for some highly questionable cases there is no evidence that this is the case. It may well be, however, that genes which control morphological structure, and which thus are not directly related biochemically, are clustered on chromosomes so that coordinate effects can be exerted on the same organ or tissue. Clustering may provide a mechanism for common control of such genes at critical times in development. Until such time as the genetic maps of higher organisms approach the density of phage (even the genetic map of *Drosophila melanogaster* is sparse compared with T4), this concept cannot be tested directly. However, statistical approaches to this problem do exist; one is presented here.

This paper is an attempt to answer the following question: is there in the genome of *D. melanogaster* any organization in the locations of functionally related genes, or are functionally related genes situated at random? This question can be divided into two parts, each of which will be discussed separately. The first part is concerned with whether or not functionally related genes have any tendency to be found on particular chromosomes, and the second part is concerned with whether, within a given chromosome, functionally related genes are situated at random along the length of that chromosome.

To answer both parts of the question, preliminary data on the mutants of *D. melanogaster* were collected in such a manner as to minimize bias. Braver (1956) classified mutants 'according to body parts affected' and according to chromosome, but without listing map locations. Map locations were taken from Bridges & Brehme (1944), on which Braver's work was originally based. In the analyses discussed below only those mutants are included for which a definite location is given by Bridges and Brehme. No attempt was made to check further on the location or on the phenotype; such a procedure introduces no bias in the results obtained from these data.

Since the statistical procedures that are used require that the numbers involved are not too small, some modifications of the classification as given by Braver were

* Supported in part by research grant GM 08202 and Career Development Awards (1-K3-GM-31, 732-01 and GM-K3-14, 911) from the Public Health Service.

https://doi.org/10.1017/S0016672300010442 Published online by Cambridge University Press

made. Firstly, the following three classes have been obtained by pooling the appropriate sections, care being taken to include each distinct mutant only once: (i) change in size or number of bristles and hairs; (ii) bristles and hairs of the head; (iii) bristles and hairs of the thorax. Secondly, certain other classes listed by Braver, in which a total of less than ten mutants exist (e.g. antenna, arista, etc.) are excluded. Thirdly, all mutants on the fourth chromosome were excluded. Furthermore, the four following classes were also excluded, since the genes involved in each can hardly be considered as a single functionally related group: larvae, lethals, pupae, and temperature sensitive. The remaining mutants are then found

Table 1. *Numbers of mutants classified by chromosomes and body parts affected*

	Chromosome		
	I	II	III
Abdomen	10	11	7
Body			
Color	11	5	7
Shape	6	6	7
Size	15	13	8
Bristles and hairs			
Change in size or number	20	13	15
Shape	5	5	3
Head	10	2	5
Thorax	14	21	22
Eyes			
Color	9	17	22
Size and shape	28	28	23
Texture	22	23	14
Legs	3	9	4
Sterility			
Male	7	11	5
Female	14	15	4
Thorax	10	8	10
Wings			
Color	11	5	6
Curvature	12	27	13
Length	19	34	14
Margin effects	13	10	10
Position held to body	12	22	15
Veins	19	31	12
Width	7	24	6
Total	277	340	232
Percentage of all mutants	32·6	40·0	27·3
Percentage of total map length	23·6	38·6	37·9

χ^2 (42 d.f.) = 65·4, $P \sim 0·01$.

https://doi.org/10.1017/S0016672300010442 Published online by Cambridge University Press

to occur as indicated in Table 1, where the numbers of mutants occurring in each chromosome are given for each of twenty-two different classes.

To answer the first part of our question, i.e., whether functionally related genes have any tendency to be found on particular chromosomes, the data in Table 1 can be considered as a simple 22×3 contingency table and a homogeneity χ^2 statistic with 42 degrees of freedom calculated. The result, as indicated at the foot of the table, is significant at approximately the 1% level. We therefore conclude that the relative probabilities that genes should occur in chromosomes I, II or III are not the same for all functional groups. This conclusion assumes firstly that the classes considered are related to functional groups, and secondly that the mutations of the different types were ascertained independently of their location. The test discounts the fact that, in the sample of genes obtained, the proportions found in the three chromosomes are different from the proportions that the three chromosomes bear, in map units, to the total genome. (See last two lines of Table 1.)

To answer the second part of our question, i.e., whether within a given chromosome there is any tendency among functionally related genes to cluster, those classes in Table 1 in which there are a larger number of mutants were further analyzed; in particular, the three classes pertaining to eyes and the five larger classes of the seven pertaining to wings were subjected to further analysis as indicated in Tables 2 and 3.

The method of analysis is based upon the following considerations. If a set of genes are distributed at random along a chromosome, then, if the whole length of the chromosome is divided into a number of intervals of equal length, the number of genes occurring in any one interval is expected to follow a Poisson distribution. Since no 'true' scale on which to measure chromosome distances is known, we arbitrarily use map distances. The choice of the length of interval to take is arbitrary, as far as the null hypothesis of a Poisson distribution is concerned, but can greatly affect whether or not clustering, if present, will be detected. Intuitively it would appear that clustering, in the broadest possible sense, will probably be best detected if the length of the interval is so chosen that the mean number of genes per interval approximates unity; clustering of two or more genes together will then be detected as an excess of intervals in which more than one or no genes occur, and a deficiency of intervals in which just one gene occurs, as compared to what is expected on the basis of a Poisson distribution. (The choice of interval length that makes the mean number approximately unity can even more clearly be seen to be appropriate if the opposite alternative hypothesis is considered, i.e. that there is some tendency towards regular spacing of the genes. For if the mean is unity and the genes are strictly equidistant along the chromosome, then all the intervals will contain one gene and none will contain more than one or no genes.) For convenience, we have used the same length of interval throughout, namely three map units; the twenty-four distributions given in Table 2 then have means that range approximately from 0·3 to 1·3. The successive intervals were defined as from 0 up to and including 3 units, from more than 3 up

https://doi.org/10.1017/S0016672300010442 Published online by Cambridge University Press

to and including 6 units, etc. Twenty-two intervals were taken for Chromosome I, thirty-six for Chromosome II and thirty-five for Chromosome III.

For each of the twenty-four distributions a simple χ^2 statistic with one degree of freedom can be calculated testing the goodness of fit of a Poisson distribution by pooling the number of intervals in which more than one and no mutants occur. The square root of this, with a positive sign if the departure from a Poisson distribution is in the direction of clustering, and a negative sign if the departure is in the direction of regular spacing, is a standardized normal deviate under the null hypothesis. The calculated deviates are given in Table 3, together with the significance levels appropriate for a one-sided test. It should be noted that one

Table 2. *Numbers of 3-unit intervals containing indicated numbers of eye and wing mutants*

	Chromosome																
	I					II							III				
Number of mutants in 3-unit interval	0	1	2	3	4	0	1	2	3	4	5	7	0	1	2	3	4
Eyes																	
Color	14	7	1	0	0	24	8	3	1	0	0	0	21	10	2	0	2
Size and shape	8	6	3	4	1	18	12	4	1	0	1	0	20	9	4	2	0
Texture	10	7	2	1	2	24	7	3	1	0	0	1	25	7	2	1	0
Wings																	
Curvature	13	7	1	1	0	17	13	4	2	0	0	0	27	5	1	2	0
Length	13	4	1	3	1	15	14	3	3	0	1	0	25	6	4	0	0
Margin effects	10	11	1	0	0	30	5	0	0	0	1	0	29	3	2	1	0
Position to body	14	4	4	0	0	20	11	4	1	0	0	0	24	9	1	0	1
Veins	12	3	5	2	0	17	12	3	3	1	0	0	26	6	3	0	0

Table 3. *Normal deviates for data in Table* 2

	Chromosome			
	I	II	III	Total
Eyes				
Color	−0·49	0·95	0·62	1·08
Size and shape	0·82	0·30	1·04	2·16
Texture	0·48	1·81*	0·91	3·20*
Wings				
Curvature	−0·02	−0·09	1·54*	1·43
Length	1·78*	−0·27	1·29*	2·80*
Margin effects	−1·73	1·05	1·86*	1·18
Position to body	1·35*	0·33	0·29	1·97
Veins	2·22*	0·38	0·99	3·59*
Total	4·41*	4·46*	8·54**	17·41***

* $0·01 < P < 0·1$. ** $0·001 < P < 0·01$. *** $P < 0·001$.

of these deviates (that for margin effects on chromosome I) would be considered significant at the 10% level in the direction of regular spacing if a two-sided test were performed. However, of the twenty-four deviates in the body of the table only five are negative, as contrasted with an expected twelve if the genes were all randomly distributed within the chromosome. Furthermore these normal deviates can be summed, and, on the assumption of independent Poisson distributions, the sum of n such deviates will be a normal random variable with mean 0 and variance n. It is clear from the table that if sufficiently many are summed there is a highly significant result. The twenty-four distributions are not really independent, as a certain amount of pleiotropy occurs in the data. This, however, is far too little in amount to account for the significance found; and in any case if there is any tendency for pleiotropic genes to fall into clusters with respect to more than one of the functions involved, this in itself is of great interest.

This tendency towards clustering of genes which affect the same structure might have physiological significance, or it could merely reflect the known fact that the genes of *D. melanogaster*, when considered altogether in terms of map distances, are not located at random along the chromosomes. For example, the region of the genetic map on the third chromosome between 5 and 25 has only four known loci, whereas the region between 40 and 60 on the same chromosome has over fifty loci. To distinguish between these two possibilities the analysis has been repeated, but in such a way as to discount the known clustering of all loci; this is achieved by transforming the scale on which the chromosome distances are measured before dividing each chromosome up into new intervals (n-locus intervals). The results are given in Tables 4 and 5, which are analogous to Tables 2 and 3.

The practical procedure used to determine these intervals is, for each chromosome separately, as follows: all the loci given by Bridges & Brehme (1944) are listed in order and every nth locus in that list marks the end of an interval. This has the effect of measuring chromosome distances on a scale on which each interval contains the same number (n) of known loci. For convenience, we have used only one value of n for each chromosome, namely $n = 8$ for chromosomes I and II, and $n = 10$ for chromosome III. These values were chosen so as to make the overall mean number of mutants per interval very close to unity for each of the three sets of eight distributions in Table 4. The result is that there are seventeen such intervals for chromosome I, twenty-two for chromosome II, and fifteen for chromosome III; in each case there is a partial interval at the end of the chromosome that is ignored, and this accounts for the fact that the total number of mutants in each of the twenty-four distributions is not always the same in Tables 2 and 4.

The normal deviates for the twenty-four distributions are given in Table 5; there are more negative entries in the body of the table than there are positive entries, and so the data would suggest there is more evidence for regular spacing than there is for clustering. For this reason the significance level used is that appropriate for a two-sided test. It is quite clear, both from the non-significance of the grand total and the fact that barely more than 10% of the statistics are

https://doi.org/10.1017/S0016672300010442 Published online by Cambridge University Press

Table 4. *Numbers of* n-*locus** *intervals containing indicated numbers of eye and wing mutants*

	Chromosomes														
	I					II					III				
Number of mutants in n–locus interval	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Eyes															
Color	10	5	2	0	0	10	7	4	1	0	4	4	4	3	0
Size and shape	2	9	2	3	1	4	12	4	1	1	4	2	7	1	1
Texture	3	10	2	1	1	7	9	4	2	0	7	4	2	2	0
Wings															
Curvature	8	6	3	0	0	7	5	8	2	0	6	6	2	1	0
Length	6	7	2	1	1	6	6	6	3	1	5	6	4	0	0
Margin effects	7	7	3	0	0	13	8	1	0	0	10	2	1	2	0
Position to body	9	6	0	2	0	6	12	3	1	0	4	8	2	1	0
Veins	6	5	5	1	0	4	11	5	1	1	5	8	2	0	0

* $n = 8$ for chromosomes I and II, 10 for chromosome III.

Table 5. *Normal deviates for data in Table* 4

	Chromosome			
	I	II	III	Total
Eyes				
Color	0·16	0·42	0·64	1·22
Size and shape	−1·73*	−1·82*	1·63	−1·92
Texture	−1·97*	−0·40	0·81	−1·56
Wings				
Curvature	−0·04	1·29	−0·29	0·96
Length	−0·38	0·71	−0·26	0·07
Margin effects	−0·48	−0·78	1·71*	0·45
Position to body	−0·04	−1·73*	−1·33	−3·10*
Veins	0·63	−1·41	−1·40	−2·18
Total	−3·85	−3·72	1·51	−6·06

* $0{\cdot}01 < P < 0{\cdot}1$.

significant at the 10% level, that on this new scale of measurement functionally related genes can be regarded as occurring randomly along the chromosome.

We see therefore that the clustering of functionally related genes within a chromosome appears, on the basis of the data analyzed here, to be completely accounted for by the general clustering of all genes. The tendency for functionally related genes to occur on a particular chromosome, however, cannot be accounted for by the different proportions of the genes borne by each of the chromosomes. It should be noted that the test performed here for clustering within chromosomes cannot be expected to be very powerful in detecting particular types of clustering, since it is designed only to detect clustering of two *or* more genes. We may also expect some loss of power from not using the best interval length, i.e., that interval length that leads to a mean number of mutants per interval of unity, for each of the twenty-four distributions. Nevertheless the method clearly detected the expected clustering on a scale of map units, and yet suggested, for the transformed scale, that any departure from randomness is in the direction of regular spacing.

We may note in conclusion that there are ways in which the data could be improved and lead to a better analysis. Bridges and Brehme's list was compiled in 1944 and many more loci and their mutant phenotypes have been reported since then. When a new edition is available, these can be added to this analysis. Pseudoalleles are listed by Braver as a single gene, and these might well be a source of clustering. We have simply ignored the fact that there is any pleiotropism, since we do not know the true primary function of the proteins coded by pleiotropic genes: if this were known a clearer analysis would perhaps be possible. Finally, a denser genetic map would be of great help. This might be achieved by increasing the number of known genetic loci of the X-chromosome of *D. melanogaster* by accumulating temperature-sensitive sex-linked lethals. Since these lethals will survive at some temperature, the functions of the genes involved might be more easily determined. Extensive data of this type might shed new light on gene clustering in higher organisms. The present analysis, however, does not detect any functional significance for the arrangement within chromosomes of the known genes in *D. melanogaster*.

SUMMARY

A statistical analysis of the genome of *D. melanogaster* indicates that functionally related genes tend to be found on a particular chromosome and, when their locations within a given chromosome are measured in terms of map units, show a tendency to cluster; this clustering within chromosomes, however, is completely accounted for by the known clustering of all genes within chromosomes. Thus the analysis does not reveal any obvious functional significance for the arrangement of the genes within the chromosomes of this organism.

REFERENCES

Braver, Norma B. (1956). *The mutants of* Drosophila melanogaster *Classified According to Body Parts Affected.* Carnegie Institution of Washington Publication 552A, Washington, D.C.

Bridges, Calvin B. & Brehme, Katherine S. (1944). *The Mutants of* Drosophila melanogaster. Carnegie Institution of Washington Publication 552, Washington, D.C.

https://doi.org/10.1017/S0016672300010442 Published online by Cambridge University Press