Heredity (1977), **38** (1), 37-45

THE PATTERN OF LOCI ON *DROSOPHILA* CHROMOSOMES

P. R. McCARTNEY, J. H. RENWICK and MADELEINE R. MUNDAY*
London School of Hygiene and Tropical Medicine, London WC1E 7HT

Received 14.vii.76

Summary

In this study, we have attempted to detect and describe patterns in the arrangement of gene loci within chromosomes of a metazoon. Known loci (842 of them) on the first three chromosomes of *Drosophila melanogaster* have been characterised according to a list of 45 properties and the distributions of these properties have been examined systematically.

For all these properties, there is little, if any, evidence of clustering between known loci that are not close enough to belong to the same 10-locus group. Our analysis has, however, revealed evidence confirming that some properties show a tendency to cluster *within* a 10-locus group. Even this tendency is not strong, except for some of the morphological properties. However, it is apparently at variance with the findings of Elston and Glassman (1967), based on fewer data.

1. Introduction

From our knowledge, however limited, of evolutionary mechanisms, we can be sure of the existence of some residual pattern in the arrangement of gene loci on the chromosomes of any species (see Renwick, 1972). Aspects of the subject are also discussed by Fox and Abächerli (1971); Hood *et al.* (1975) and Tartof (1975).

The null hypothesis is that there is no pattern among syntenic loci (*i.e.* loci on the same chromosome pair), relating their ordinal positions to their properties; or, in other terms, that those loci that possess a particular property (or set of properties) are randomly positioned in the whole sequence of syntenic loci. The aim of Elston and Glassman (1967) was to show whether deviation from such randomness within a chromosome could be detected. The fact that they were unable to detect such a deviation led to assertions that pattern is non-existent or at least not detectable.

The power of any specific test to demonstrate a deviation will, of course, vary with the nature of the pattern that is actually present. It is well known that, for testing the randomness of artificially generated pseudorandom numbers (for example), no single test of randomness is by itself adequate to exclude all patterns. As it happens, some of our results, though not conclusive by themselves, do now conform with the *a priori* and general expectation of non-randomness.

The description of the residual pattern and the assessment of its consistency or otherwise from chromosome to chromosome and from species to species are the ultimate problems to which our studies have been directed. One approach is to generate an index of similarity between two loci from their properties and to study its relationship with map distance. However,

* Present address: Department of Gynaecology, St Thomas's Hospital Medical School, London SE1 7EH

early in this work it became obvious that, where clustering did exist, it was usually of a weak nature and, even when detectable, was often in the form of more than one cluster on the same chromosome (as it is in the examples given in tables 1 and 2). In such instances, one locus, though it might be close to other members of a cluster of loci similar to itself, might concurrently be remote from each member of a second cluster of the same type of loci. Thus a simple, monotonic relationship between similarity and distance will not hold.

These observations forced us to revert to two distributional tests, using the data on individual properties, namely the χ^2 dispersion test and an adaptation of the runs test. Following Elston and Glassman (1967), we allowed for the uneven distribution of all detected loci along the chromosome map by ignoring distances for these tests and by considering only the linear order of loci. Minor differences in statistical technique and the availability of additional data have led to results differing in part from those of Elston and Glassman.

2. The drosophila data

From the valuable compilation of *D. melanogaster* data by Lindsley and Grell (1967), each locus was scored for the presence or absence of each of 45 properties. Strictly, the properties belong to one or more mutants at a locus rather than to the locus itself. In choosing the 45 properties to be scored, we made no attempt to obtain independence of one property from another. About two-thirds of the properties are of a morphological nature and the remainder could be crudely described as concerned with function (see table 3). Some loci are assigned only one property but most loci are assigned more, the maximum being in fact 10 properties, the mean 3·74. Presence, particularly of secondary properties, is undoubtedly underestimated, but the degree of this is probably independent of locus position.

To avoid more relevant scoring errors (of a systematic kind), the loci on a chromosome were scored in alphabetical order rather than in the order of the loci on the chromosome. On the first, second and third chromosomes, 394, 253, 195 loci, respectively, were scored. We have accepted Lindsley and Grell's decisions about which closely-linked mutants represent separate loci. We recognise, however, that there must inevitably be errors in those decisions, if only because the results of complementation tests on all pairs of closely linked mutant " alleles " were not and still are not available—far from it. Further, even where complementation between two mutants has been demonstrated, the mutants could still theoretically belong to different complementation groups within the same locus. We return to this problem in the general discussion.

Unlike Elston and Glassman, we have not eliminated properties such as lethality at various stages of development, since the argument that that property embraces a variety of mechanisms, though valid, applies in some degree to all the other properties also.

(i) *Distribution of occurrences of a property within groups of loci*

The linear sequence of loci on a chromosome was fragmented into as many non-overlapping groups of 10 adjacent loci as possible, the remaining

TABLE 1

Numbers of loci concerned with homoeosis in consecutive groups of 10 loci on each of the main drosophila chromosomes

Chromo- some	Numbers of homoeotic loci in g consecutive groups of 10 loci	Dispn. χ^2	d.f. (g − 1)	P (disp. test)	No. of groups below mean n	No. of groups above mean m	No. of runs u'	Max. no. runs that would give $P < 0{\cdot}05$ $u_{0{\cdot}05}$	P (runs test)	P (combined)
1	000001000000000000001000000000000000000	(37·39)	38	n.s.	37	2	5	2	n.s.	n.s.
2	0000001000100100100000000	(21·34)	24	n.s.	21	4	9	3	n.s.	n.s.
3	0001000123110510000	42·85	18	$<0{\cdot}001$ (0·00084)	11	8	7	6	$P<0{\cdot}1$ (0·088)	$P<0{\cdot}001$ (0·00078)

The dispersion χ^2 was calculated as $(g-1)$ (observed variance)/(expected variance). It was highly significant for chromosome 3. The dispersion χ^2s are given in parentheses for the other chromosomes, whose loci possess the homoeotic property too rarely to allow a meaningful test. The runs test (Swed and Eisenhart, 1943) by itself was not significant for this homoeotic property on any of the chromosomes, but, for chromosome 3, the combined significance level—by Fisher's technique using exact significance probabilities—was still highly significant.

TABLE 2

Numbers of loci concerned with size change in consecutive groups of 10 loci on each of the main drosophila chromosomes

Chromosome	Numbers of size-change loci in consecutive groups of 10 loci	Mean no. per group	Dispn. χ^2	d.f. (g−1)	P	No. of groups below mean n	No. of groups above mean m	No. of runs u'	Max. no. runs that would give $P<0{\cdot}05$ $u_{0\cdot05}$
1	331542424300401410534144022113561131615	2·59	66·41	38	<0·005	19	20	23	14
2	5531301101101112123123140	1·72	35·84	24	<0·1	15	10	12	8
3	4103042122101531132	1·89	24·61	18	n.s.	9	10	11	6
All		2·17	126·86	80	<0·002				

The dispersion χ^2 was calculated as $(g-1)$ (observed variance)/(expected variance). It was highly significant overall and on chromosome 1 individually. The runs test was not significant for this size property on any of the chromosomes.

Table 3

Outcome of dispersion tests for each of 31 morphological and 14 other properties on each chromosome

Chromosome 1		Chromosome 2		Chromosome 3		All chromosomes		Property	χ^2_2 between
dir.	N	dir.	N	dir.	N	dir.	ΣN	(i) Morphological	chromosomes
–	6	–	3	.	0	.	9	Whole fly	4·9
+	83	+	12	+	20	+	115	Body	**39·9**
–	15	–	3	–	8	–	26	Head	5·2
+	26	+	14	+ +	20	+ +	60	*Thorax*	3·9
+	28	+	8	+	4	+	40	Abdomen	**9·7**
–	4	+	5	–	3	. .	12	Ocelli	1·0
+	102	–	56	+	53	. .	211	Eyes	1·8
–	3	.	1	–	3	. .	7	Aristae	1·7
+	7	.	0	.	1	.	8	Antennae	**7·6**
+	144	+	85	–	51	. .	280	Wings	**6·2**
+ +	26	+	30	+	17	+ +	73	*Wing venation*	5·3
+	13	+	8	+	8	+	29	Legs	0·3
+	49	–	22	–	14	. .	85	Females	4·7
–	63	+	22	+	12	. .	97	Males	**15·4**
+	97	+	45	–	31	. .	173	Genitalia	**7·6**
+	15	+ +	7	+	9	+§	31	*Hairs*	1·1
+	91	+	51	+	43	+	185	Bristles	0·8
+ +	104	+	44	+	36	+ +	184	*Size change*	**8·8**
+	12	+	8	+ +	5	+	25	*Size increase*	0·2
+ +	92	+	36	–	31	. .	159	*Size decrease*	**9·7**
–	11	–	7	+	10	. .	28	Absence of part	2·4
+	12	+	21	–	15	. .	48	Number change	**10·4**
–	5	–	15	+	12	. .	32	Number increase	**11·0**
–	5	+ +	6	–	3	. .	14	*Number decrease*	1·1
–	211	+	125	+	76	. .	412	Morphology	**10·7**
+	41	+	25	+	15	+	81	Texture	1·1
–	2	–	4	+ +	15	. .	21	*Homoeosis*	**24·7**
–	2	.	0	.	0	.	2	Asymmetry	.
–	4	+	7	–	5	. .	16	Internal organs	3·3
–	2	–	4	+ +	5	. .	11	*Halteres*	4·6
.	0	–	4	.	1	.	5	Other parts	**7·6**
								(ii) Functional	
+	76	+	42	+	47	+	165	Colour	4·0
+	96	+	41	–	26	. .	163	Function	**12·3**
+	41	–	23	–	33	. .	97	Posture	**7·2**
.	1	.	0	.	0	.	1	Behaviour	.
+	43	–	55	–	32	. .	130	Lethal at any stage	**14·0**
+	15	+	8	.	1	.	24	as pupa	**7·0**
–	3	+	6	.	1	.	10	as larva	4·0
–	5	.	1	.	0	.	6	as embryo	4·6
–	5	–	4	–	4	–	13	Tumour	0·5
–	4	+	6	+	8	. .	18	Enzyme, etc.	5·0
–	9	–	14	+	15	. .	38	Hatching date	**10·0**
.	1	.	0	.	1	.	2	Next generation	.
.	0	.	1	.	1	.	2	Resistance	.
.	0	–	2	.	1	.	3	Meiosis	.

Key

N Indicates the number of occurrences of a property on a particular chromosome.
Direction of outcome (*dir.*).

+ Indicates that the test statistic (not printed) fell *above* its expected median value in the direction of clustering (but not significantly so).

+ + Indicates that it fell *significantly above* it (P < 0·05 on a two-tailed test).

– Indicates that it fell *below* it.

. Indicates numbers inadequate for testing (0 or 1) on at least one chromosome.

. . Indicates that the statistic showed apparent inconsistency—it deviated from the median in different directions on different chromosomes.

Significance was reached for the nine morphological properties [in italic type] and for no function-related properties. The last column gives the goodness-of-fit χ^2 (2 d.f.) to test whether or not, for each property, the numbers of occurrences, N, as shown on the three chromosomes, are proportional to the numbers of loci on them (394, 253, 195 respectively). For 18 properties, the χ^2 (in **bold face**) exceeded 6·0 ($P < 0·05$), often by a wide margin.

§ Significant (P < 0·01) when dispersion and runs tests are jointly pooled over three chromosomes.

few loci at the right end of the chromosome (and all those on the short chromosome 4) being discarded. For each property, the observed number of loci possessing that property was found for each group. We computed the observed numbers of groups with 0 loci possessing the property; those with 1 locus possessing it; with 2 loci possessing it; . . . up to all the group's loci possessing it.

We chose a group-size of 10 loci for pragmatic reasons. This round number had been used before; a larger number would have generated too few groups for a sensitive test, a smaller one would have made the test sensitive to only tight forms of clustering. No attempt was made to define formally the group-size that optimised some compound function of the various desired characteristics.

(ii) *Test of dispersion*

The observed values were used, for each property separately, to assess their observed variance. This variance was compared with the variance to be expected on the null hypothesis that the ordering of loci is random in respect of this property. The expected variance is npq, where n is the number of loci per group (*i.e.* 10), p and q are the probabilities, respectively for absence and presence of the property as regards a randomly-chosen locus on this chromosome. An estimate ($\hat{q}$) of q is given by the observed frequency of the property among loci on this particular chromosome, and $\hat{p}$ is $(1-\hat{q})$.

In their analysis of n-locus intervals, Elston and Glassman retained the Poisson approximation to the binomial distribution. This, though appropriate elsewhere in their paper, was unnecessarily inaccurate here, particularly as their group size was only 8 or 10 and as the occurrence rate, q, was substantial. The criteria of Feller (1957) were therefore not satisfied. Indeed, their use of the Poisson approximation introduces a slight negative bias in the normal deviates of their table 5, and thus hinders their chances of finding significant clustering.

Subject to certain conditions about the underlying distribution, the ratio of observed to expected variances, when multiplied by $(g-1)$, is distributed as a chi-squared with $(g-1)$ degrees of freedom (see, for example, Armitage, 1971), where g is here the number of groups.

A simplified form of this χ^2 test was also used by Elston and Glassman (1967). The data at the time of their analysis were sufficiently limited in bulk that only dichotomisation seemed justified, the two groups being (a) those with exactly one locus possessing the property and (b) those with some other number of loci, including zero. The size of groups had been chosen as 8 (but 10 for chromosome 3) to ensure a mean occurrence-rate of approximately one locus possessing the property per group.

3. Results

On chromosome 1, three properties, denoted in abbreviated form by the terms, *wing-venation*, *size-change* and *size-decrease*, are significantly non-random at the 5 per cent level or beyond. On chromosome 2, *hairs* and *a decrease in numbers of a part* are significantly non-random. On chromosome 3, the properties, *thorax*, *size-increase*, *halteres*, *homoeosis* (*i.e.* the substitution of one body part, say a leg, by another body part, say an antenna), are significantly non-random, the first and the last beyond the 0·005 level. In each case, a two-tailed test has been used.* The short chromosome 4 was ignored.

A total of 112 tests on individual chromosomes could be performed usefully—*i.e.* the number of occurrences available for them is adequate. (For a further 23, the number available is not adequate—0 or 1.) Although the non-independence of many of the 45 properties makes hazardous any summation over them, it is pertinent to note that all nine significant results are in the direction of clustering, as are the majority of non-significant results. Not one significant result in favour of a systematic distribution of loci was found.

As we assume it to be independent across chromosomes, the statistic and also its degrees of freedom could each be usefully summated over all three chromosomes, for each property (see table 3). There are 35 properties that could have been treated in this way, but to avoid the combining of effects of opposite sign, we carried out summation for only those 13 properties for which, additionally, the statistic for each of the chromosomes was consistently on the same side of its expected median value. For only two properties was the χ^2, consistently for all 3-chromosomes, **less** than its median value, whereas, for 11 properties, it was consistently in **excess** of that value. For three of these 11 properties—*thorax*, *wing-venation* and *size-change* (the last being set out in table 2)—this overall excess, indicative of clustering, was significant ($P < 0{\cdot}05$), if we assume homogeneity among chromosomes.

(i) *Discussion of dispersion test*

The dispersion test we used would have registered a bigger deviation from randomness if, instead of taking the number of loci on each chromosome as given, we had treated the whole genome as a unit. That bigger deviation we would have observed would have reflected two component phenomena —clustering within the chromosome and uneven allocation between the chromosomes. The latter effect is already well recognised. It is discussed by Elston and Glassman and it is confirmed here in the statistical significance of the goodness-of-fit χ^2 tests for 18 of the 40 properties that could be tested (see last column of table 3). Some of the deviations from simple proportionality, *e.g.* those for loci affecting the *whole body*, are very large.

In interpreting the results within a chromosome, we have borne in mind that, from a multiplicity of significance tests, even if the properties were independent, a predictable number of the tests are expected to be significant in either direction even on the null hypothesis. But all the properties that manifest significant non-randomness do so in the direction of clustering and all are from the class concerned with morphology. Indeed nine of the 31 properties in the morphological class are significantly in favour of clustering either on one of the chromosomes individually, or on the whole set. Not one is significant from the class of 14 functional properties. Admittedly, this apparent disproportion is, at least in part, a reflection of the fact that

* A result in the left tail could point to a systematic arrangement, with a tendency for like loci to be equally spread. Elston and Glassman's analysis gave some such indication but this largely disappears if the analysis is repeated without the Poisson approximation.

numbers (N in table 3) are more adequate for detecting significance in the one class than in the other, so it is far from certain that there is, in addition, a true difference in clustering tendency between the two classes.

(ii) *Tight and loose clustering*

The test of dispersion that has been employed is expected to detect close clustering more efficiently than loose clustering, because it takes no account of any similarity of scores in contiguous groups (of 10 loci each). However, of the two levels of clustering—loose and tight—the former is certain to exist and is therefore more interesting to pursue.

(iii) *Test of runs*

With this in mind, a runs test was carried out on each chromosome, for each property (Swed and Eisenhart, 1943). We examined the sequence of groups, after classifying them by whether or not the frequency of occurrence of the property among loci of the group exceeded the average frequency over all groups. For example, as shown in table 2, on chromosome 1, among 39 consecutive groups of 10 loci each, the mean number of *size-change* loci is 2·59. Twenty groups exceed this mean and these occur (in 12 unbroken runs) in an overall sequence of 23 runs (u). But this overall sequence would have had to consist of as few as 14 runs ($u_{\cdot 05}$) to have indicated a clustering tendency of this *size-change* property at a significance level of 0·05.

The sensitivity of the test and the nature of the patterns being sought both depend to an unfortunate extent upon the arbitrary choice of group size. The 10-locus group corresponds, on average, to about two map units on chromosome 1 and about five on the other chromosomes. The choice of a smaller group size would have led to a more sensitive test but any result that was significant would have referred mainly to a tight form of clustering.

In the event, the occasions when the number of runs was significantly less than the expected number (tabulated in part by Swed and Eisenhart, 1943), were no more frequent than would have been predicted in a multiplicity of significance tests. In other words, loose clustering, if it exists, was not detected.

4. General discussion

As already mentioned, the properties are related to each other, sometimes very obviously (*e.g. size change*, *size-increase*, *size-decrease*) so no summation over properties is undertaken.

It is reassuring that the dispersion test is sensitive enough to have detected, in particular, that element of obvious non-randomness manifested by *homoeotic* loci (see table 1). This was not detected by similarity-index methods presumably because of the diluting effect of the other 44 properties. This diluting could be particularly important when a non-monotonic relationship between similarity and distance exists for one or more of these other properties, as discussed in the introduction.

In selecting a suitable test, our preference has naturally been for one that was at least able to demonstrate obvious clusters such as the *homoeotic* ones. However, this type of selectivity merges imperceptibly into one that gives the results we fancy. To escape from this situation, we intended to choose a method by its appropriate behaviour with one set of data and then to use it as the preferred method for all other data. We can still eventually do this, but, for the moment, the methods chosen—the dispersion test and the runs test—are both unsuitable for the more scanty data so far available on organisms other than Drosophila.

In *D. melanogaster* for all 45 properties, the runs test gives little, if any, evidence of clustering between loci that are not close enough to belong to the same 10-locus group. The dispersion test has, however, revealed evidence confirming that some properties, particularly morphological ones, show tendencies to cluster *within* a 10-locus group. With notable exceptions, even these tendencies (towards *close* clustering) are not strong.

Some caution must also be exercised because of the insecurity of Lindsley and Grell's decisions about which mutants, in a closely-linked series, belong to separate loci. We expect one systematic type of error from interpreting two close loci as one (*e.g.* on the basis of no recombination) and a countervailing type from interpreting one locus as two (when independently-arising mutants could not be tested against each other). The relative frequency of occurrence of these two types of error is difficult to assess.

For any constant size of group, the runs test, on the sequence of groups, and the dispersion test, which is invariant to rearrangement of groups, are independent. The right-tail probabilities from the two tests (each for three chromosomes) were therefore combined for each property, by the $(-2\Sigma lnP)$ method of Fisher (1946). The results (converted to two-tailed significance levels) neither strengthen nor weaken more than marginally the overall indications for clustering obtained from the dispersion test alone. The details are therefore not presented.

5. References

ARMITAGE, P. 1971. *Statistical methods in medical research.* Blackwell Scientific Publications, Oxford.

ELSTON, R. C., AND GLASSMAN, E. 1967. An approach to the problem of whether clustering of functionally related genes occurs in higher organisms. *Genet. Res.*, *9*, 141-147.

FELLER, W. 1957. *An Introduction to Probability Theory and its Applications*, Vol. I, pp. 98 and 143. Wiley, New York.

FISHER, R. A. 1946. *Statistical methods for research workers*, 10th Edn., p. 100. Oliver and Boyd, London.

FOX, D. J., AND ABÄCHERLI, E. 1971. Drosophila enzyme-genetics: a table. *Experientia*, *27*, 218-220.

HOOD, L., CAMPBELL, J. H., AND ELGIN, S. C. R. 1975. The organization, expression and evolution of antibody genes and other multigene families. *Annl. Rev. Genet.*, *9*, 305-353.

LINDSLEY, D. L., AND GRELL, E. H. 1967. *Genetic variations of drosophila melanogaster.* Carnegie Institution, Washington.

RENWICK, J. H. 1972. Comments on the chromosome maps of man and other vertebrates. *Bull. Europ. Soc. Hum. Genet.*, *5*, 108-119.

SWED, F. S., AND EISENHART, C. 1943. Tables for testing randomness of grouping in a sequence of alternatives. *Ann. Math. Stat.*, *14*, 66-87.

TARTOF, K. D. 1975. Redundant genes. *Annl. Rev. Genet.*, *9*, 355-385.