

# A FULLER THEORY OF "JUNCTIONS" IN INBREEDING

Sir RONALD A. FISHER  
Department of Genetics, Cambridge

Received 15.xii.53

## I. THE PLACE OF THE THEORY OF JUNCTIONS

THE theory of junctions was first put forward in the third chapter of my book *The Theory of Inbreeding*, published in 1949. It is set out in Section 14, pages 49-61. Since that time further examination of the problem and new calculations allow the subject to be exhibited in a more complete form, and an inconsequential error to be corrected. I agree with the reviewer (C.A.B.S.) of the first edition, that the subject should have been given a separate chapter. In this paper I hope to give a more complete account of the present state of knowledge.

Each locus of the germ plasm will be represented in the foundation material on a finite number of strands; for example, on four strands for an inbred line propagated by sib-mating, or by parent offspring mating, but on eight strands for double cousin mating, where two pairs are to be used, or for self-incompatible tetrasomic plants. After  $n$  generations more than one strand may be present at some loci, but at an increasing number all but one will have disappeared.

With sib-mating, for example, the probability of not yet being homogenic has been shown to be exactly

$$\frac{1}{10}(5+16\epsilon)\epsilon^n - 2^{-n} + \frac{1}{10}(13-16\epsilon)\left(\frac{1}{2}-\epsilon\right)^n + \frac{1}{5} \cdot 2^{-2n}$$

As, in this expression the second, third and fourth terms are, when  $n$  exceeds 10, somewhat trifling in magnitude, and decreasing rather rapidly, it has been convenient to consider the later stages of the inbreeding process in the light of the asymptotic theory, in which the first term only is retained. Out of a total length  $L$ , then, the length of strand, on any metric whatever, expected to be still heterogenic will be expressed as

$$l = \zeta \epsilon^n L,$$

where  $\zeta$  stands for

$$\frac{1}{10}(5+16\epsilon) = 1.7944272$$

approximately, and  $\epsilon$  for

$$\frac{1}{4}(\sqrt{5}+1) = 0.80901699.$$

So expressed, the theory supplies the average value of the length of strand still heterogenic, but gives no idea of the extent to which this length may be expected to vary from case to case. This will depend greatly on the number of heterogenic tracts in which the portion still heterogenic is broken up. It is a function of the theory of junctions to supply means of calculating the expected number of such tracts. Indeed, the asymptotic theory may be completed on this basis, since in the limit both the number of tracts and their average length are found to diminish. With tracts widely separated, so that the survival of each is independent of the existence of others, the distribution of tract number must tend to the Poisson limit, for which the frequency of there being  $n$  tracts is

$$e^{-m} \frac{m^n}{n!},$$

where  $m$  is the number expected. Similarly, as the tracts grow shorter by random attrition, the distribution of length among existing tracts must tend to the limit

$$df = \frac{1}{a} e^{-x/a} dx.$$

If  $X$  stands for the sum of  $n$  independent values of  $x$ , the distribution of  $X$  must be

$$df = \frac{1}{a^n (n-1)!} \cdot X^{n-1} e^{-X/a} dX$$

for any given number  $n$ , and for all  $n$  possible with the frequencies of the Poisson series the distribution is

$$df = e^{-\frac{x}{a}-m} \sum_{n=1}^{\infty} \frac{1}{n!(n-1)!} \left(\frac{mX}{a}\right)^n \frac{dX}{X}$$

together with the finite frequency,  $e^{-m}$ , at the limit,  $X = 0$ . The mean value of  $X$  in the expression above is easily seen to be  $ma$ . Consequently, no more is needed for the complete asymptotic distribution than to supplement the expected length heterogenic with the expected number of heterogenic tracts.

The continuous part of the distribution may be expressed in terms of a Bessel Function, *i.e.* as

$$\frac{m}{\sqrt{Xl}} e^{-m(1+\frac{X}{l})} I_1(2m\sqrt{Xl}) \cdot dX,$$

where  $l$  is the expected total length heterogenic, and

$$I_1(u) = \frac{1}{i} \mathcal{J}_1(iu) = \frac{1}{2} u + \frac{1}{2^2 \cdot 4} u^3 + \frac{1}{2^2 \cdot 4^2 \cdot 6} u^5 + \dots$$

The expected value of  $e^{tX}$  for the distribution

$$\frac{1}{a^n (n-1)!} X^{n-1} e^{-x/a} dX$$

is  $(1-at)^{-n}$ ;

and  $\sum e^{-m} \frac{m^n}{n!} \cdot (1-at)^{-n}$

is equal to  $e^{m/(1-at)-m} = e^{mat/(1-at)}$ ,

so the Cumulative Function is

$$mat/(1-at) = lt/(1-\frac{lt}{m}),$$

whence, as has been seen, the mean is

$$l = am,$$

the variance

$$\frac{2l^2}{m} = 2ma^2,$$

and the other cumulants  $\begin{cases} K_3 = 6ma^3, \\ K_4 = 24ma^4, \\ K_5 = 120ma^5, \end{cases}$

and so on.

In these expressions it should be noted that the condensation at zero is included.

Dr J. H. Bennett has also shown how the variance may be obtained by considering the survival of the heterogenic condition at two loci simultaneously. By this path also he has verified the variance deduced above, both for the case of sib-mating, and for some other cases.

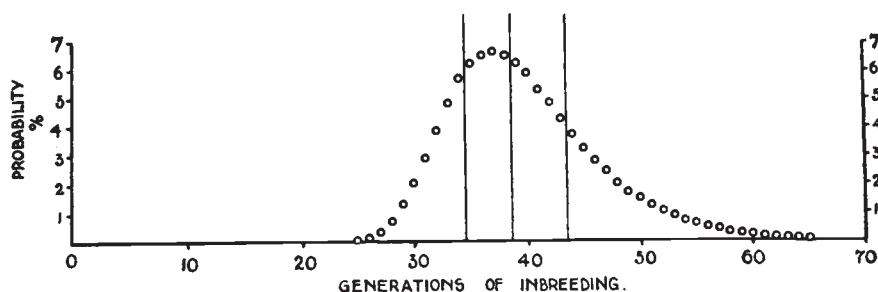


FIG. 1. Probability of elimination of the last remaining Heterogenic Tract at each generation of sib-mating.

## 2. THE PRODUCTION AND SURVIVAL OF JUNCTIONS

A new junction is formed whenever a crossover occurs in a region for which the parent organism is heterozygous. Since "map length" is a metric defined by the frequency of crossing over, junction-formation will be uniform on this metric. Knowing the frequency of all mating types at each locus, the rate of production of junctions is easily calculated. After production the junction will be inherited like a point-mutation. It may be lost; in the course of time it may come to prevail in all four strands; after a given number of generations,

however, it has a calculable chance of being neither lost nor fixed. If it were fixed the strands on both sides of the junction would be homogenic. If they are heterogenic on one side only, the junction must bound a heterogenic tract. If they are heterogenic on both sides, the junction must lie within such a tract. The number of tracts is calculable from the number of external junctions. It is most convenient to calculate first the expected number of all junctions, and then by a more complex argument to calculate the smaller number of internal junctions, the subtraction of which will give the number of external junctions.

The total number of junctions, external and internal, at any stage may be calculated by the method of Section 14. The types of mating capable of giving rise to a new junction at any locus are :

Type	Frequency
$aa = ab$	$u$
$ab = ab$	$v$
$ab = ac$	$x$
$aa = bc$	$y$
$ab = cd$	$z$

where  $a, b, c, d$ , represent local portions of the kinds of strand available.

Since in types  $u$  and  $y$  only one parent is heterozygous, only two of the gametes used will come from a heterozygous parent, so that in the four gametes used the rate of occurrence of new junctions per unit map length is

$$2u + 4v + 4x + 2y + 4z.$$

This frequency is calculable after any number  $r$  of generations of inbreeding, and is found in fact to be

$$\frac{1}{5}(4 + 24\epsilon)^r + \frac{1}{5}(16 - 24\epsilon)\left(\frac{1}{2} - \epsilon\right)^r.$$

The first mating involving a junction formed in gametogenesis of the  $r^{\text{th}}$  generation is of the type

$$je = ee$$

in the  $(r+1)^{\text{th}}$  generation, where  $j$  stands for the strand with a junction, and  $e$  for any strand of type  $a, b, c, d$  unbroken at this point. The survival of the junction is analogous to that of a point mutation, and the probability of it being neither fixed nor eliminated after  $n-r-1$  further generations is

$$\frac{1}{20}(5 + 16\epsilon)^{n-r-1} + \frac{1}{20}(13 - 16\epsilon)\left(\frac{1}{2} - \epsilon\right)^{n-r-1} + \frac{1}{10}\left(\frac{1}{4}\right)^{n-r-1}$$

The expected number of junctions, external and internal at the  $n^{\text{th}}$  generation is found by multiplying these expressions together and adding for all values of  $r$  from 0 to  $n-1$ . In the summation we use the relations.

$$\sum_{r=0}^{n-1} \alpha^r \beta^{n-r-1} = \frac{1}{\alpha - \beta} (\alpha^n - \beta^n)$$

whenever  $a \neq \beta$ ,  
 while the sum is  $na^{n-1}$   
 when  $a = \beta$ .

The total density of junctions external and internal, evaluated in this way comes to

$$\frac{1}{5} \left\{ (8+8\epsilon)n + (428-528\epsilon) \right\} \zeta \epsilon^n$$

where the factor  $\zeta \epsilon^n$  represents the fraction of the total map length heterogenic.

### 3. INTERNAL JUNCTIONS

The frequency of junctions internal to heterogenic tracts remains to be subtracted.

So far as concerns the leading term, that involving  $\epsilon^n$ , these are derived in two ways. (i) At a point with chromosomes of only two kinds, of mating type  $u$  or  $v$ , a crossover may produce a new type of chromosome resembling one kind on the left and the other on the right, which will be an internal junction so long as both the original kinds are still present, but will become external if either of these be lost. Using the analogy of a junction with a mutation, we may think of such mating types as trigenic, of types  $x$  or  $y$ .

In mating type  $u$  (as shown in fig. 2) in which two strands are produced by a heterozygote, one half of the next generation appear

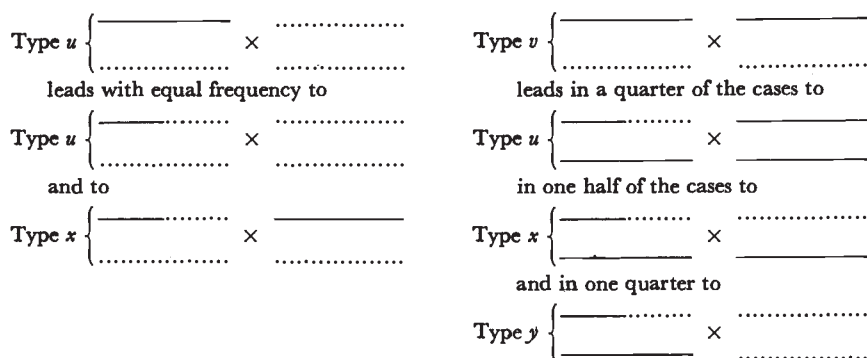


FIG. 2.—Modes of generating trigenic internal junctions from digenic types of mating.

as type  $x$ , while from mating type  $v$  producing four such strands, one half are of type  $x$  and one quarter of type  $y$ . The rate of production, per unit map length is therefore

$$u + 3v,$$

and substituting

$$u = (20 - 24\epsilon) \zeta \epsilon^n$$

$$v = (40\epsilon - 32) \zeta \epsilon^n,$$

the rate of production of such trigenic types is seen to be

$$(96\epsilon - 76)Z\epsilon^r.$$

As to their survival, the coefficient both of  $x_0$  and of  $y_0$  in the expectation of  $x+y+z$  taken  $n-r-1$  generations later is

$$\left(\frac{1}{2}\right)^{n-r-1}.$$

But

$$\begin{aligned} & \sum_{r=0}^{n-1} (96\epsilon - 76)Z\epsilon^r \left(\frac{1}{2}\right)^{n-r-1} \\ &= (96\epsilon - 76)Z \frac{1}{\epsilon - \frac{1}{2}} \left\{ \epsilon^n - \left(\frac{1}{2}\right)^n \right\} \end{aligned}$$

So the coefficient of  $Z\epsilon^n$  in the expected number of internal junctions of this type, after  $n$  generations is

$$\begin{aligned} & (96\epsilon - 76)4\epsilon \\ &= (96 - 112\epsilon) \\ &= 5.39010 \end{aligned}$$

(ii) There is also a second type of internal junction with frequency of the same order, which at first I overlooked; for starting with a trigenic or tetragenic mating type, a junction may be formed between two strands, and the junction together with a strand not involved in the junction may both survive as a digenic type in some later generation. Since such a digenic type will thereafter become rarer only as a power of  $\epsilon$ , the fact that a rapidly decaying type  $x$ ,  $y$  or  $z$  is needed to initiate the system, will not prevent internal junctions of this second kind, formed for the most part early in the inbreeding process, from contributing a term in  $\epsilon^n$ . This process is illustrated in fig. 2.1.

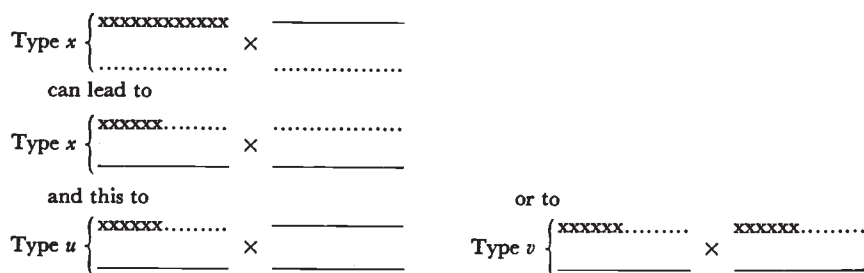


FIG. 2.1.—An illustrative mode of generating digenic internal junctions from trigenic types of mating.

If we let  $j$  stand for a junction supposedly between strands  $a$ ,  $b$ , and let  $c$ ,  $d$  stand for strands of different origin from these, the number

of mating types which may be distinguished is 15. Of the three ultimate digenic types, the absolute complexities are already known as follows :—

Mating types	Absolute complexity
$cj = cc, jj$ $u$	$5 + 16\epsilon \div 20$
$cj = cj$ $v$	$6 + 20\epsilon \div 20$
$cc = jj$ $w$	$8 + 24\epsilon \div 20$

A class of mating types which lead only to these apart from other types with the junction lost or external, may thus be evaluated by the frequency with which the types of known complexity are generated. So for four types involving  $a, c, j$  the complexities are as follows :—

Mating types	Absolute complexity
$ac = aj$	$2 + 6\epsilon \div 20$
$ac = cj$ or $aj = cj$	$3 + 10\epsilon \div 20$
$aa = cj$	$1 + 4\epsilon \div 20$
$aj = cc$ or $ac = jj$	$4 + 12\epsilon \div 20$

The four mating types in  $c, d, j$  can also be expressed in terms of the first three.

$cd = cj$	$5 + 16\epsilon \div 20$
$cc = dj$	$5 + 16\epsilon \div 20$
$cj = dj$	$6 + 20\epsilon \div 20$
$cd = jj$	$8 + 24\epsilon \div 20$

Two types involving  $a, b, c, j$  can be expressed in terms of the first seven.

$ab = cj$	$1 + 4\epsilon \div 20$
$ac = bj$	$2 + 6\epsilon \div 20$

Finally, the two remaining types involving  $a, c, d, j$  can be expressed in terms of the first eleven.

$cd = aj$	$4 + 12\epsilon \div 20$
$ad = cj$	$3 + 10\epsilon \div 20$

With these fifteen evaluations it is easy to find the contribution of type  $x(ab = ac)$  which out of 4 gametes produced by heterozygotes has

$ac = aj$	$\frac{1}{2}$	worth	$1 + 3\epsilon \div 20$
$bc = cj$	1		$3 + 10\epsilon \div 20$
$cj = aa$	$\frac{1}{2}$		$0.5 + 2\epsilon \div 20$
$ab = cj$	$\frac{1}{2}$		$0.5 + 2\epsilon \div 20$
$ac = bj$	$\frac{1}{2}$		$1 + 3\epsilon \div 20$

with total  $6 + 20\epsilon \div 20$

Equally, from type  $y$  with two gametes produced by the heterozygote, all will be of type  $ac = cj$ , giving again,

$$6 + 20\epsilon \div 20$$

Finally, from type  $z$  of four gametes from heterozygotes, there will be

$ac = cj$	2	$6 + 20\epsilon \div 20$
$ad = cj$	2	$6 + 20\epsilon \div 20$

Hence the value from all sources is

$$\begin{aligned} & (6+20\epsilon)(x+y+2z) \div 20 \\ &= (6+20\epsilon)L_r \div 20 \\ &= (6+20\epsilon)\left(\frac{1}{2}\right)^r L_0 \div 20 \\ &= (12+40\epsilon)\left(\frac{1}{2}\right)^r \div 20 \end{aligned}$$

since if  $z_0 = 1$ ,  $L_0 = 2$ .

As before, then, we may evaluate the series

$$\begin{aligned} & \sum_{r=1}^{n-1} (12+40\epsilon)\left(\frac{1}{2}\right)^r \epsilon^{n-r-1} \div 20 \\ &= (12+40\epsilon) \frac{1}{\epsilon - \frac{1}{2}} \left\{ \epsilon^n - \left(\frac{1}{2}\right)^n \right\} \div 20, \end{aligned}$$

of which the term in  $\epsilon^n$  is

$$\begin{aligned} & (12+40\epsilon)(4\epsilon) \div 20 \\ &= (40+128\epsilon) \div 20 \\ &= 4\mathcal{Z}. \end{aligned}$$

To the total of

$$(96-112\epsilon)\mathcal{Z}\epsilon^n$$

internal junctions of the first kind, we have then to add

$$4 \quad \mathcal{Z}\epsilon^n$$

of the second, making

$$(100-112\epsilon)\mathcal{Z}\epsilon^n$$

in all.

Now the total density of junctions was found to be

$$\frac{1}{5} \left\{ (8+8\epsilon)n + (428-528\epsilon) \right\},$$

and subtracting the internal junctions, there must be

$$\begin{aligned} & \frac{1}{5} \left\{ (8+8\epsilon)n - (72-32\epsilon) \right\} \\ &= \frac{1}{5} (8+8\epsilon) \left\{ n - \frac{1}{5} (58-52\epsilon) \right\} \\ &= 2.8944(n-3.1862) \end{aligned}$$

external junctions expected.

The number of separate heterogeneous tracts will, apart from tracts ending on the terminus of a chromosome, be half of this, or

$$1.4472(n-3.1862),$$

and the typical length of such a tract will be

$$\frac{69}{n-3.2} \text{ centimorgans}$$

when  $n$  is as great as 10 or so.

Expressions retaining only the leading term are, of course, approximations, and may not be very accurate in the first 10 generations.



For these table 11.0 gives the total number of junctions produced, and the numbers expected at any stage to have been lost, or fixed in tracts already homogeneous, or external, namely at the boundary of a homogeneous with a heterogeneous tract, or finally, internal, within a tract still heterogeneous. These numbers relate to each 100 cM map length.

TABLE 11.0  
*Fate of junctions produced in the first 10 generations*

Gen.	Lost	Fixed	External	Internal	Total produced
1	0.	0.	0.	4.	4.
2	1.	0.	0.75	6.25	8.
3	2.5625	0.0625	1.96875	6.40625	11.
4	4.23438	0.23438	3.09961	5.93164	13.5
5	5.88672	0.51172	3.94434	5.15723	15.5
6	7.43262	0.87012	4.46219	4.36008	17.125
7	8.84253	1.28003	4.69404	3.62090	18.4375
8	10.10516	1.71454	4.70267	2.97763	19.5
9	11.22258	2.15227	4.55113	2.43340	20.35938
10	12.20262	2.57762	4.29324	1.98121	21.05469
	$18 - \alpha\epsilon^n(n + \beta)$	$6 - \alpha\epsilon^n(n + \gamma)$	$2\alpha\epsilon^n(n - 5)$	$\eta\epsilon^n$	$24 - \delta\epsilon^n$
	$\alpha = (18 + 58\epsilon)/25 = 2.5969194$		$\beta = (58 - 52\epsilon)/5 = 3.186223$		
	$\beta = (187 - 178\epsilon)/5 = 8.298995$		$\gamma = (26 + 72\epsilon)/5 = 16.849845$		
	$\gamma = (-13 + 22\epsilon)/5 = 0.959675$		$\delta = (32 + 112\epsilon)/5 = 24.521981$		

The rate at which junctions are lost is greatest in the fourth generation, while the rate at which they become fixed reaches its maximum in the ninth. Indeed, for all junctions fixed the average stage of inbreeding is so high as 13 generations. External junctions are highest after 8 generations, and this is the stage at which the number of heterogenic tracts reaches its culmination. Internal junctions are highest after only 3 generations. The table is supplemented by the formulæ in  $\epsilon^n$  appropriate to later stages.

A distinction may be drawn between end-tracts, terminated by the end of a chromosome, and enclosed tracts. If the total map length is  $L$ , and the haploid number of chromosomes is  $\nu$  we may expect :

$$\begin{array}{ll}
 \text{End tracts} & 2\nu\zeta\epsilon^n \\
 \text{Enclosed tracts} & \{1.45(n-3.19)L - \nu\}\zeta\epsilon^n \\
 \text{Total} & \{1.45(n-3.19)L + \nu\}\zeta\epsilon^n.
 \end{array}$$

Since the average length still heterogenic is

$$L\zeta\epsilon^n$$

the standard length of a heterogenic tract may be taken to be

$$1 \div \{1.45(n-3.19) + \nu/L\}$$

In mice there are 20 pairs of chromosomes, with a total map length roughly estimated to be about 2500 centimorgans. For various

numbers of generations of sib-mating the following table (11.1) shows the expected numbers of tracts of these two kinds, and the expected total map length occupied by them.

TABLE 11.1  
*Numbers and expected length of heterogeneous tracts*

Genera- tions	Expected number			Total length cM.	Expected length attached to each dominant cM.
	End tracts	Enclosed tracts	Total number		
5	23.633	47.488	71.121	1477.0	...
10	8.582	49.375	57.957	536.4	18.0
15	2.988	30.432	33.420	186.77	12.25
20	1.035	15.228	16.263	64.71	8.68
25	0.359	6.900	7.259	22.43	6.43
30	0.124	2.954	3.078	7.772	5.29
35	0.043	1.218	1.261	2.693	4.50
40	0.015	0.490	0.505	0.933	3.91

TABLE 11.2  
*Probability of complete homogeneity at different stages  
of inbreeding by sib-mating*

Generation	Probability	Generation	Probability
25	0.00070	45	0.82005
26	0.00216	46	0.84850
27	0.00565	47	0.87286
28	0.01278	48	0.89360
29	0.02557	49	0.91118
30	0.04603	50	0.92601
31	0.07565	51	0.93847
32	0.11508	52	0.94892
33	0.16386	53	0.95765
34	0.22056	54	0.96492
35	0.28313	55	0.97098
36	0.34910	56	0.97599
37	0.41606	57	0.98019
38	0.48187	58	0.98364
39	0.54479	59	0.98651
40	0.60358	60	0.98888
41	0.65745	61	0.99084
42	0.70601	62	0.99246
43	0.74916	63	0.99379
44	0.78708	64	0.99489
45	0.82005	65	0.99580

From the expected total number of tracts

$$m = \{1.4472(n-3.1862)L + v\} \epsilon^n$$

the probability of complete homogeneity is calculated, when  $m$  is sufficiently small, as  $e^{-m}$ .

With  $v = 20$ ,  $L = 25$ , table 11.2 shows the probability of complete homogeneity from 25 to 65 generations. In fig. 1, p. 189, is shown the distribution of the generation number at which complete homogeneity first occurs. It is probable, however, that the mutation rates at the different loci might supply a new mutation in one of the four strands in every 20 generations or so. Consequently, for long-continued inbreeding we should approach not final homogeneity but a statistical equilibrium in which one or more mutations would occasionally be segregating. The distribution is, of course, a discontinuous one, though bearing some resemblance to the continuous distribution

$$df = d(e^{-x}).$$

Dr Bennett has made a comparative study of these distributions for parent-offspring matings, selfed tetrasomics, and other important types of inbreeding.

#### 4. REFERENCES

- FISHER, R. A. 1949. *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh.  
BENNETT, J. H. 1953. Junctions in inbreeding. *Genetica*, 26, 392-406.