CALCULATING LINKAGE INTENSITIES FROM F_3 DATA*

F. R. IMMER¹

*Office of Sugar Plant Investigations, Bureau of Plant Industry,
U. S. Department of Agriculture, University Farm, St. Paul, Minnesota*

Received April 10, 1933

It has been shown by FISHER (1928) that no statistic for estimating linkage intensities can have a smaller sampling variance, in the theory of large samples, than that given by the solution of the equation of maximum likelihood. The method of maximum likelihood may be used, then, as a standard by which the efficiency of other formulae for calculating linkage intensities may be determined. FISHER (1930) determined the efficiency of several other methods for calculating linkage intensities from F_2 data. IMMER (1931) presented a study of the efficiency of the four-fold correlation method. FISHER and BALMUKAND (1928) provided a table to facilitate the calculation of linkages, from F_2 data, by the product ratio method. HUTCHINSON (1929) applied the method of maximum likelihood to problems of linkage between two genes when the characters were determined by duplicate of complementary factors, and IMMER (1930) provided tables which could be used to reduce to a minimum the labor necessary in calculating linkage intensities for various factorial combinations when the product ratio method is used.

It remained for FISHER (1930) to point the way to the solution of linkage problems when F_3 data are utilized in order to determine the F_2 genotypes. This can be accomplished most efficiently by an easy extension of the method of maximum likelihood. The present paper is a continuation of this study.

If we designate p (expressed as a decimal fraction) as the crossover percentage in repulsion crosses, with $1-p$ as the crossover percentage in coupling crosses, we may calculate easily the expected frequency of occurrence of the four expected phenotypes in F_2 . The expected ratio for the AB , Ab , aB and ab phenotypes in F_2 will be $1/4 (2+p^2, 1-p^2, 1-p^2, p^2)$. If we designate the observed frequency of occurrence of the AB , Ab , aB and ab phenotypes as a , b , c and d we may calculate the linkage intensity satisfying the equation of maximum likelihood from

$$\frac{a}{2+p^2} - \frac{b+c}{1-p^2} + \frac{d}{p^2} = 0, \quad (I)$$

* Part of the cost of the mathematical composition is paid by the GALTON AND MENDEL MEMORIAL FUND.

¹ The writer wishes to express his appreciation to Dr. R. A. FISHER for his criticisms and suggestions during the course of this study, begun while the writer was a Fellow of the NATIONAL RESEARCH COUNCIL at the ROTHAMSTED EXPERIMENTAL STATION, Harpenden, Herts., England.

or, expressed more conveniently for calculation, $n p^4 - (a - 2b - 2c - d) p^2 - 2d = 0$, where p or $1 - p$ is the crossover percentage, as defined above, and n is the total frequency. The amount of information supplied by the data will be

$$\frac{2n(1+2p^2)}{(1-p^2)(2+p^2)} \quad (\text{II})$$

and the standard error of p will be given by the square root of the reciprocal of the amount of information or $\sqrt{\frac{(1-p^2)(2+p^2)}{2n(1+2p^2)}}$.

A very common problem confronting the plant geneticist is one in which one of the two factor pairs being studied acts as a zygotic lethal when homozygous recessive, rendering the plant unclassifiable for the other factor pair in the lethal phenotypes. This results in the typical 9:3:4 ratio when the two factor pairs are inherited independently. Assuming that the aB and ab phenotypes die, or are indistinguishable, the expected frequency ratio in F_2 will be $1/4 (2+p^2, 1-p^2, 1)$. The expectation in the $aB+ab$ class, being independent of p , will supply no information on linkage. Designating the observed numbers in the AB and Ab classes as a and b , as before, we may find the value of p which will satisfy the conditions of maximum likelihood. Multiplying the observed numbers in each class by the logarithm of the expected, and summing, we obtain, $a \log (2+p^2) + b \log (1-p^2) + (c+d) \log 1$. Maximizing this equation for p we obtain

$$\frac{2ap}{2+p^2} - \frac{2bp}{1-p^2} = 0, \quad (\text{III})$$

and the value of p can be calculated readily from $p = \sqrt{\frac{a-2b}{a+b}}$. This

formula was first given by COLLINS (1924). The formula for the standard

error of p can be calculated from the expression $\frac{2ap}{2+p^2} - \frac{2bp}{1-p^2}$. Dif-

ferentiating again with respect to p , substituting the expected for the observed frequencies and changing signs of the terms we obtain the amount of information contained in the data in the form

$$\frac{3np^2}{(2+p^2)(1-p^2)} \quad (\text{IV})$$

The reciprocal of this, or $\frac{(2+p^2)(1-p^2)}{3np^2}$, will be the sampling vari-

ance of p and the square root of this quantity will be the standard error of p . By dividing the variance obtained when all classes are viable in F_2 by the above variance, we may find the efficiency of linkage determinations made when two classes are lethal compared with classifications in which four classes are viable. This is given by

$$\frac{(1-p^2)(2+p^2)}{2n(1+2p^2)} \div \frac{(1-p^2)(2+p^2)}{3np^2} = \frac{3p^2}{2(1+2p^2)}$$

This tends toward zero as p tends toward zero, becomes .25 when $p = .50$ and rises to .50 when $p = 1.00$. It is obvious, therefore, that the error of linkage calculations based on repulsion crosses in which one character is lethal in the homozygous recessive condition is very high compared with determinations in which all classes are viable. Even in coupling crosses the relative accuracy does not exceed one-half that obtained when all classes are viable for any value of p . For close linkage in repulsion crosses the amount of information supplied by classification for only the two distinguishable phenotypes is extremely small for ordinary populations studied.

ROBERTSON, DEMING and KOONCE (1932) studied the linkage relations between the factor pairs *Ycyc* and *Fcfc* determining green vs. virescent seedlings and green vs. chlorina seedlings, respectively. Plants of the *ycyc* were lethal in the seedling stage rendering it impossible to classify the plants for the *Fcfc* factor pair when the *Ycyc* factor pair was homozygous recessive. The cross was one of repulsion. They observed in F_2

Phenotype	Symbol	Number of plants
<i>Yc Fc</i>	a	4165
<i>Yc fc</i>	b	1823
<i>Yc Fc + yc fc</i>	c+d	1853
<hr/>		<hr/>
Total	n	7841

The linkage intensity would then be given by

$$p = \sqrt{\frac{4165 - 2(1823)}{4165 + 1823}} = .2944 \text{ or } 29.44 \text{ percent crossing over.}$$

The amount of information contained in these data would be given by equation IV, or

$$\frac{3(7841)(.2944)^2}{\{1 - (.2944)^2\} \{2 + (.2944)^2\}} = 1069.74.$$

The standard error of p would be

$$\sqrt{\frac{1}{1069.74}} = .0306 \text{ or } 3.06 \text{ percent.}$$

A condition frequently encountered in F_2 classification is one in which the genetic constitution of the material is such that the AB and Ab phenotypes can not be separated, resulting in an expected 12:3:1 ratio when $p = .50$. For such cases the value of p satisfying the conditions of maximum likelihood will be given by

$$p = \sqrt{\frac{d}{c+d}}$$

with a standard error of p given by $\sqrt{\frac{1-p^2}{n}}$.

Compared with classifications in which all classes are distinguishable the above classification is 100 percent efficient when $p = 0$, falls to 75 percent efficiency when $p = .50$ and tends toward 50 percent as p tends toward 1.00.

FIGURE 1.—Graph showing amount of information per plant, in percent of a backcross, in F_2 data when; (A) all classes are fully viable and distinguishable, (B) the AB and Ab phenotypes can not be separated, and (C) the aB and ab phenotypes can not be separated.

The relative accuracy of F_2 classifications when (1) all phenotypes are distinguishable, (2) the AB and Ab phenotypes are indistinguishable or (3) the aB and ab phenotypes are indistinguishable is shown graphically in figure 1. The information per plant grown and classified from these

three types of separations is given in percent of the information which would be supplied by a single plant from a backcross.

The accuracy of F_2 data in the coupling phase when all classes are viable, approaches the accuracy of a backcross as the recombination approaches zero. When the aB and ab phenotypes are indistinguishable for one of the factor pairs the amount of information supplied for close linkage by the F_2 repulsion phase is very small.

The estimation of the linkage from F_2 classification must be based on the ratio of the observed numbers of individuals in the different phenotypes. With the exception of the double recessive class the F_2 phenotypes are made up of two or more genotypes. By growing and classifying progenies of selfed F_2 plants in F_3 it is possible, with suitable numbers in the F_3 lines, to determine the genotype of each F_2 plant. The linkage may be estimated, then, from the frequency of occurrence of the various genotypes. The manner in which F_3 data may be used to supplement the F_2 will be discussed in some detail.

ESTIMATING LINKAGE INTENSITIES FROM F_3 DATA

In table 1 are given the expected frequencies, in terms of p , of the different genotypes which may be distinguished in progenies of selfed F_2 plants. Designating the probability of occurrence of the AB , Ab , aB and ab gametes as p , $1-p$, $1-p$ and p , respectively, and assuming that crossing over occurs equally frequently in both male and female gametogenesis the probability of occurrence of the different genotypes is found quickly by multiplying the probability of the male by the female gametes.

TABLE 1

Frequency of occurrence of the nine genotypes expected from progenies of selfed F_2 plants.

	AA		Aa		aa
BB	p^2		$2p(1-p)$		$(1-p)^2$
Bb	$2p(1-p)$	$2p^2$	$2(1-p)^2$		$2p(1-p)$
bb	$(1-p)^2$		$2p(1-p)$		p^2

The genetic expectation in the $AaBb$ genotype will depend on the genetic recombination which produced this genotype. If the recombination is one of coupling the expectation will be $2p^2$ and if it is repulsion it will be $2(1-p)^2$. The total expectation for the $AaBb$ genotype will be $2p^2 + 2(1-p)^2$ or $2(1 - 2p + 2p^2)$. In table 2 are given the symbols to be used to designate the observed frequency of the individuals of each genotype actually observed in an experiment.

We will consider first the separation of the "doubly dominant" F_2 progenies into genotypes by growing progenies of selfed F_2 plants in F_3 .

"Doubly dominant" is used to denote the AB phenotypes; that is, plants in which at least one gene of each of the two factor pairs is dominant. We will assume also that the $AaBb$ genotypes are not to be separated into two components depending on their origin from coupling or repulsion recombination. The expected frequency of the two kinds of $AaBb$ genotypes combined will be, therefore, $2(1-2p+2p^2)$. The total expected frequency ratio in the 4 genotypes of the AB phenotype, it will be remembered,

$$\text{is } \frac{2+p^2}{4}.$$

TABLE 2

Symbols used to designate the observed frequency of occurrence of individuals of various genotypes.

	AA		Aa		aa
BB	e		f		l
Bb	g	h		i	m
bb	j		k		n

The value of p which will satisfy the conditions of maximum likelihood will be found by multiplying the observed frequencies in the different genotypes by the logarithm of the expected, summing and making this quantity a maximum, note being taken of the basis on which these are expected. Thus,

$$e \log \frac{p^2}{2+p^2} + (f+g) \log \frac{2p(1-p)}{2+p^2} + (h+i) \log \frac{2(1-2p+2p^2)}{2+p^2}$$

when maximized for p will give the equation of maximum likelihood in the form

$$\frac{2e+f+g}{p} - \frac{f+g}{1-p} - \frac{2(h+i)(1-2p)}{1-2p+2p^2} - \frac{(e+f+g+h+i)2p}{2+p^2} = 0. \quad (V)$$

This formula, on substituting the actual numbers, found in an experiment, for e, f, g, h and i will yield the best possible value of p .

ROBERTSON et al (1930), in the linkage study referred to previously, grew and classified 746 F_3 progenies from as many F_2 plants of the $YcFc$ phenotype. The genotype constitution of these plants was as follows:

<i>Genotype</i>	<i>Symbol</i>	<i>Number of F_3 lines</i>
$Yc Yc Fc Fc$	e	31
$Yc yc Fc Fc$	f	157
$Yc Yc Fc fc$	g	145
$Yc yc Fc fc$	h+i	413

Substituting the observed frequencies in equation V we obtain

$$\frac{364}{p} - \frac{302}{1-p} - \frac{826(1-2p)}{1-2p+2p^2} - \frac{1492p}{2+p^2} = 0.$$

The exact value of p can be approached most rapidly by substituting likely values for p and interpolating. Thus, we may estimate the value of p near .30 and substitute in the formula given above. The result is illustrated below.

	$p=.30$	$p=.29$	$p=.2994$
$\frac{364}{p}$	1213.33	1255.17	1215.765
$-\frac{302}{1-p}$	-431.43	-425.35	-431.059
$-\frac{826(1-2p)}{1-2p+2p^2}$	-569.66	-589.80	-570.891
$-\frac{1492p}{2+p^2}$	-214.16	-207.61	-213.771
Total	-1.92	32.41	.044

The result of substituting $p=.30$ being -1.92 we assume that p is but slightly less than .30 and substitute $p=.29$; obtaining 32.41 as a total. The true value must give zero as a total so we interpolate by calculating $.29+.01$ ($32.41 \div 34.33$) and find $p=.2994$. Substituting $p=.2994$ we find the total deviation from zero to be .044. This serves as a check on the previous calculation as well as means of obtaining a more accurate value if one is required.

The amount of information contained in the data from this type of analysis in F_3 may now be calculated. To do so we differentiate equation V, that is,

$$\frac{2e+f+g}{p} - \frac{f+g}{1-p} - \frac{2(h+i)(1-2p)}{1-2p+2p^2} - \frac{(e+f+g+h+i)2p}{2+p^2}$$

with respect to p and obtain

$$-\left\{ \frac{2e+f+g}{p^2} + \frac{f+g}{(1-p)^2} - \frac{2(h+i)4p(1-p)}{(1-2p+2p^2)^2} + \frac{2(e+f+g+h+i)(2-p^2)}{(2+p^2)^2} \right\}.$$

Substituting the expected for the observed frequencies we have

$$-\frac{2}{2+p^2} \left\{ \frac{p(2-p)}{p^2} + \frac{2p(1-p)}{(1-p)^2} - \frac{2(1-2p+2p^2)4p(1-p)}{(1-2p+2p^2)^2} + \frac{(2-p^2)}{(2+p^2)} \right\}$$

which reduces to

$$-\frac{4(2-6p+3p^2+4p^3)}{p(1-p)(2+p^2)^2(1-2p+2p^2)}.$$

Changing signs we find the amount of information per F_2 plant, classified into genotypes in F_3 , to be

$$\frac{4(2-6p+3p^2+4p^3)}{p(1-p)(2+p^2)^2(1-2p+2p^2)}. \quad (\text{VI})$$

The information supplied by n F_3 lines will be n times this quantity and the standard error of p will be the square root of the reciprocal of the amount of information or

$$\sqrt{\frac{p(1-p)(2+p^2)^2(1-2p+2p^2)}{4n(2-6p+3p^2+4p^3)}}.$$

Turning to the data of ROBERTSON, DEMING and KOONCE we find the numerical amount of information by substituting $p = .2994$ in equation VI,

giving 3254.47. The standard error will be $\sqrt{\frac{1}{3254.47}} = .0175$ or 1.75 percent.

Since the F_2 classification is concerned only with a separation of the phenotypes and the F_3 classification involves the separation of the doubly dominant (AB) phenotypes into their component genotypes the information contained in the latter will be supplementary to that contained in the former. In the data of ROBERTSON et al the best value of p for both F_2 and F_3 data will be found by combining the formulae given by the maximum likelihood solution for F_2 and F_3 classification; that is, equations III and V. Thus,

$$2p \left(\frac{4165}{2+p^2} - \frac{1823}{1-p^2} \right) + \frac{364}{p} - \frac{302}{1-p} - \frac{826(1-2p)}{1-2p+2p^2} - \frac{1492p}{2+p^2} = 0.$$

Solving this by the method of approximation illustrated before we find the value of p which will satisfy best both the F_2 and F_3 data to be .2981 or 29.81 percent crossing over. The standard error to be attached to this linkage can be calculated readily from the total amount of information as

calculated from equations IV and VI. Since the F_2 supplied 1069.74 units of information and the F_3 supplied 3254.47, the total will be 4324.21. The standard error of p calculated for both F_2 and F_3 data will then be

$$\sqrt{\frac{1}{4324.21}} = .0152 \text{ or } 1.52 \text{ percent. The crossover percentage for both}$$

F_2 and F_3 data is, therefore, 29.81 percent with a standard error of 1.52 percent.

One other source of information remains in the F_3 data obtained from selfed plants of the doubly dominant F_2 phenotypes. This is the separation of the two types of "double heterozygotes," depending on their origin from coupling or repulsion recombination; where "double heterozygote" is used to denote plants heterozygous for both factor pairs. Since the expected frequency for coupling and repulsion, in the double heterozygous genotypes is $2p^2$ and $2(1-p)^2$, respectively, the value of p which will satisfy best the conditions of maximum likelihood will be given by

$$h \log \frac{2p^2}{2(1-2p+2p^2)} + i \log \frac{2(1-p)^2}{2(1-2p+2p^2)}$$

which will be a maximum when

$$\frac{2h}{p} - \frac{2i}{1-p} + \frac{2(h+i)(1-2p)}{1-2p+2p^2} = 0. \quad (\text{VII})$$

The amount of information, per F_3 line classified, in this separation is found to be given by

$$\frac{4}{(1-2p+2p^2)^2}. \quad (\text{VIII})$$

This amount of information is supplementary to both the F_2 and that given in equation VI for the simple separation of the doubly dominant F_2 phenotypes into homozygous and single and double heterozygotes and can be added to them if this separation is made.

The number of plants per F_3 line needed in order to separate the two types of double heterozygotes will depend on the linkage intensity. The linkage may be calculated for each doubly heterozygous progeny and the separation into coupling or repulsion recombination made on this basis. Due to the error of random sampling the calculated values of p will vary around the true value as origin. A point may be calculated, however, from which the deviation of the true p and q (where $q = 1 - p$) will be twice their standard error. Letting such a midpoint be p' it will be possible to classify each F_3 doubly heterozygous progeny as arising through coupling or repulsion recombination depending on whether the calculated value of p

is greater or less than this midpoint, p' . When all classes are viable the number of individuals needed, per F_3 family, in order that the difference between the true value of p or q and this midpoint shall be twice the standard error of p or q will be given by

$$\sqrt{n} = 2 \frac{\left\{ \sqrt{\frac{(1-p^2)(2+p^2)}{2(1+2p^2)}} + \sqrt{\frac{(1+q^2)(2+q^2)}{2(1+2q^2)}} \right\}}{q-p}.$$

From this we find that 25 plants per F_3 line are sufficient for this separation when the crossover percentage is less than 22 percent. As the crossover percentage increases to 25 percent 33 plants would be required, for 30 percent crossing over 54 plants per F_3 line are needed, at 35 percent 97 are required, at 40 percent 223 are needed and at 45 percent the number rises to 898. It will be only for linkages which do not greatly exceed 22 percent that this separation is worthwhile.

The midpoint or p' from which the separation is to be made of the coupling or recombination double heterozygotes, with all classes viable, will be given by

$$p' = p + \frac{2}{\sqrt{n}} \sqrt{\frac{(1-p^2)(2+p^2)}{2(1+2p^2)}}.$$

The value of p' for varying percentages of crossing over is shown graphically in figure 2.

FIGURE 2.—Graph showing value of midpoint (p') to be used in separating double heterozygotes, by F_3 classification, as arising through coupling or repulsion recombination.

This midpoint, p' , is .9008 for 1 percent crossing over, falls to .7845 for

5 percent crossing over, to .7046 for 10 percent crossing over and to .6057, .5453, .5112 and .5000 as the crossover percentage increases to 20, 30, 40 and 50 percent respectively. Due to the fact that the error of the linkage intensities increases as p decreases, p' will exceed .5 for all values of the crossover percentage less than 50 percent.

We may consider next the classification of the "singly dominant" selfed progenies of F_2 plants into homozygotes and heterozygotes as a result of F_3 classification. "Singly dominant" is used to denote F_2 plants homozygous recessive for one factor pair and having at least one gene of the other factor pair in the dominant condition. We will consider only the Ab phenotype since the expectation for the aB phenotype is the same. The equation of maximum likelihood would take the form

$$j \log \frac{(1-p)^2}{1-p^2} + k \log \frac{2p(1-p)}{1-p^2}$$

which, when maximized for p would give

$$\frac{k}{p} - \frac{2j+k}{1-p} + \frac{(j+k)2p}{1-p^2} = 0 \quad (\text{IX})$$

as the formula for determining the value of p from such a separation. Differentiating again with respect to p , substituting the expected for the observed frequencies and changing signs we have the amount of information per F_2 plant classified by F_3 separation in the form

$$\frac{2}{p(1+p)(1-p^2)} \quad (\text{X})$$

The standard error of p , based on n such F_3 lines would be

$$\sqrt{\frac{p(1+p)(1-p^2)}{2n}}$$

The value of p which will satisfy the conditions of maximum likelihood when both F_2 and F_3 data are used can be obtained readily by simply combining in an additive manner the terms on the left hand side of the equality sign applicable to the particular separations involved and setting the entire quantity equal to zero. This was illustrated in the simple case of the data of ROBERTSON et al. The standard error to be attached to the value of p found from a combination of data from different classifications is obtained readily from the total amount of information; the latter being the sum of the information obtained from the different sources individually. In table 3 is illustrated the additive manner in which such information combines.

TABLE 3

Illustration of the manner of combining information from different types of classification when all phenotypes are viable and can be separated.

TYPE OF SEPARATION	INFORMATION PER PLANT	NUMBER OF PLANTS	TOTAL INFORMATION
F_2	$\frac{2(1+2p^2)}{(1-p^2)(2+p^2)}$	n	$\frac{2n(1+2p^2)}{(1-p^2)(2+p^2)}$
F_3 separation of AB phenotypes ignoring origin of double heterozygotes	$\frac{4(2-6p+3p^2+4p^3)}{p(1-p)(2+p^2)^2(1-2p+2p^2)}$	$\frac{n(2+p^2)}{4}$	$\frac{n(2-6p+3p^2+4p^3)}{p(1-p)(2+p^2)(1-2p+2p^2)}$
F_3 separation of double heterozygotes	$\frac{4}{(1-2p+2p^2)^2}$	$\frac{2n(1-2p+2p^2)}{4}$	$\frac{2n}{1-2p+2p^2}$
F_3 separation of Ab phenotypes	$\frac{2}{p(1+p)(1-p^2)}$	$\frac{n(1-p^2)}{4}$	$\frac{n}{2p(1+p)}$
F_3 separation of aB phenotypes	$\frac{2}{p(1+p)(1-p^2)}$	$\frac{n(1-p^2)}{4}$	$\frac{n}{2p(1+p)}$
Total	$\frac{2}{p(1-p)}$	n	$\frac{2n}{p(1-p)}$

Multiplying the information per plant by the number of plants expected for each type of separation we obtain the total amount of information; as given in the last column in the table. Adding the different amounts of information obtainable from F_2 classification and the separation of the doubly and singly dominant F_2 progenies into genotypes by F_3 classifica-

tion we obtain $\frac{2n}{p(1-p)}$ as the total amount of information or $\frac{2}{p(1-p)}$

as the amount of information per F_2 plant completely classified. This is twice the amount of information given by a backcross, as expected, since the number of crossovers in both male and female gametogenesis are known. In actual experiments the numerical amount of information, obtained from the various sources is added and the standard error to be attached to the crossover percentage calculated from a combination of several types of separation determined from the total amount of information.

COMPARISON OF F_2 AND F_3 DATA WHEN ALL
GENOTYPES ARE FULLY VIABLE

From doubly dominant F_3 progenies the information is of three types, (1) from small families of about 25 plants it will always be possible to separate the homozygotes from the single and double heterozygotes. The ratio of the amount of information from this source to that from a single F_2 plant is

$$\frac{4(2-6p+3p^2+4p^3)}{p(1-p)(2+p^2)^2(1-2p+2p^2)} \div \frac{2(1+2p^2)}{(1-p^2)(2+p^2)} = \frac{2(2-6p+3p^2+4p^3)(1+p)}{p(1+2p^2)(2+p^2)(1-2p+2p^2)}$$

When p is less than .08 this value exceeds 25, which shows the advantage of growing F_3 progenies from double dominants when p is less than .08 for this first purpose. In other words, F_3 classification of the above type will give more information per plant grown than will the F_2 when there is less than 8 percent crossing over as determined from repulsion crosses.

The second source of information (2) expected from doubly dominant F_3 progenies when the families are sufficiently large or the linkage sufficiently close for the two types of double heterozygotes to be separated, since the proportion of the two genotypes depends on the linkage, will add

to the precision of our estimate by an amount $\frac{4}{(1-2p+2p^2)^2}$ for each double heterozygote discriminated or

$$\frac{4}{(1-2p+2p^2)^2} \times \frac{2(1-2p+2p^2)}{2+p^2} = \frac{8}{(2+p^2)(1-2p+2p^2)}$$

for each double dominant examined.

Once the two types of double heterozygotes have been separated the third source of information (3) obtained will be ordinary F_2 data obtained from the segregation of the double heterozygotes in F_3 .

Adding the information available from the three sources from complete classification of the doubly dominant F_3 progenies such classification will give more information per plant grown than will the F_2 when p is .26 or less. For higher values of p F_2 classification will give a greater amount of information per plant grown and classified than would F_3 classification.

Twenty-five plants would be sufficient to separate the singly dominant (Ab or aB phenotypes) into genotypes by F_3 classification. The ratio of information from singly dominant F_3 progenies to that given by F_2 will be

$$\frac{2}{p(1+p)(1-p^2)} \div \frac{2(1+2p^2)}{(1-p^2)(2+p^2)} = \frac{2+p^2}{p(1+p)(1+2p^2)}$$

This exceeds 25 when p is less than .075 so that for closer repulsion than this, F_3 progenies of this kind would supply more information per plant grown than would an F_2 of equal size, considering 25 plants per F_3 line sufficient.

Since 25 plants per line would be sufficient to separate the double dominants into homozygotes and single or double heterozygotes and to separate the single dominants into homozygotes or heterozygotes it would be of interest to determine which would give the greater amount of information.

Dividing $\frac{4(2-6p+3p^2+4p^3)}{p(1-p)(2+p^2)(1-2p+2p^2)}$ by $\frac{2}{p(1+p)(1-p^2)}$ we find that

this is less than unity when p lies between .18 and .69. Therefore it would be more profitable to grow and classify, in F_3 , singly dominant progenies than to classify doubly dominant progenies into homozygotes and single and double heterozygotes when p is greater than .18 and less than .69, that is, when the crossover percentage is greater than 18 percent in repulsion or 31 percent in coupling. For closer linkage equation VI would supply more information than would IX.

In figure 3 is given a comparison of the relative accuracy of linkage calculations made from F_3 data when (A) the separation of the doubly dominant progenies is made completely, including the separation of the double heterozygotes as to origin in coupling or repulsion, (B) the genotypic constitution of the single heterozygotes is determined and (C) separation of the double dominants into homozygotes and single and double heterozygotes only, ignoring the separation of the double heterozygotes as to origin in coupling or repulsion.

The lines for B and C in the above graph cross at the points $p = .18$ and $p = .69$. The value of separating the two types of double heterozygotes is shown by the difference between curves A and C.

COMPARISON OF F_2 AND F_3 CLASSIFICATION WHEN THE HOMOZYGOTES FOR ONE OF THE CHARACTERS DIE

When two of the four F_2 phenotypes are inviable when homozygous recessive for one of the factor pairs and do not permit of classification for the other character in these phenotypes, complete classification would be possible only in the remaining two classes, giving a theoretical frequency ratio of $1/4\{2+p^2, 1-p^2, 1\}$. The two classes which die would supply no information on linkage. Since the amount of information available from F_2 separation is very small for close linkage in repulsion it would be expected that F_3 classification would here be extremely valuable.

With two classes of F_2 phenotypes indistinguishable and only small progenies grown in F_3 , failure to recover the aB phenotypes in F_3 from selfed double heterozygotes due to chance would render it impossible to distinguish the $AaBb$ F_2 genotype from the $AABb$ by F_3 classification. If the requirement of five expected plants in the aB phenotype (resulting from selfing the F_2 $AaBb$ genotype) be set up as a minimum to guard

FIGURE 3.—Graph showing amount of information, in percent of a backcross, per F_2 plant whose genotype is determined by growing selfed progenies in F_3 when; (A) the double dominants are completely classified, (B) the single dominants are separated into homozygotes and heterozygotes and (C) the double dominants are separated into homozygotes and single and double heterozygotes only, ignoring the separation of the double heterozygotes into two components depending on origin.

against chance loss of all plants of this phenotype, this number would be expected from 25 individuals per F_3 line only when p is less than the number given by $\frac{25(1-p^2)}{4} = 5$ or when p is less than .447. For higher

values of p a larger F_3 progeny would need to be grown in order to separate the double heterozygote from the $AABb$ genotype, the number required from any given value of p being given by substituting x for 25 in the above equation and solving for x .

The number of plants per F_3 line necessary to separate the double heterozygotes into the two genotypes as to origin in coupling or repulsion in order that the deviation of p and q ($= 1+p$) from the midpoint should be twice the standard error of p or q , respectively, would be given by

$$\sqrt{n} = 2 \frac{\left\{ \sqrt{\frac{(1-p^2)(2+p^2)}{3p^2}} + \sqrt{\frac{(1-q^2)(2+q^2)}{3q^2}} \right\}}{q-p}$$

The number of plants required (n) would be 250 per F_3 line for 20 percent crossing over and rises as the crossover percentage increases or decreases from this point. Apparently, then, it would be useless to contemplate separation of the double heterozygotes into the two types. When the doubly dominant F_3 progenies are tested, therefore, the separation could be made only for the homozygous dominants and single and double heterozygotes, without attempting to separate the latter as to origin. This could be accomplished with 25 plants per F_3 line when p is less than .447, while a greater number of plants would be required when p exceeds this point.

Such being the case the doubly dominant progenies would be more profitable when p is less than .18 and the single dominants would be more profitable when p exceeds .18. With all classes viable the double dominants again became more profitable than the single dominants when p exceeded .69, that is, less than 31 percent crossing over in coupling. With two of the four classes of phenotypes dying, or indistinguishable, the number of plants per F_3 line for the separation of the double dominants when p exceeds .69 would exceed 25 by an amount which would render the single dominants the most profitable on the basis of plants grown, for all values of p in excess of .18.

Comparing the amount of information in the separation of the double dominants into homozygotes and single and double heterozygotes by F_3 classification with that available from F_2 separation we have

$$\frac{4(2-6p+3p^2+4p^3)}{p(1-p)(2+p^2)(1-2p+2p^2)} \div \frac{3p^2}{(1-p^2)(2+p^2)} = \frac{4(2-6p+3p^2+4p^3)(1+p)}{3p^3(2+p^2)(1-2p+2p^2)}$$

This exceeds 25 when p is less than 31.5 percent so that for closer linkage than this F_3 classification of this type would be more profitable than an F_2 population of the same number of plants, assuming 25 plants grown per F_3 line.

Comparing the singly dominant F_3 progenies with the F_2 we obtain

$$\frac{2}{p(1+p)(1-p^2)} \div \frac{3p^2}{(1-p^2)(2+p^2)} = \frac{2(2+p^2)}{3p^3(1+p)}$$

This exceeds 25 when p is less than .347 which shows the F_3 classification of this type is more profitable, per plant grown, than the F_2 for smaller values of p . It is apparent that when two phenotypes are indistinguishable in F_2 the error is so much greater than when all phenotypes may be discriminated that F_3 classification is more profitable over a wider range of p .

SUMMARY

1. Formulae are presented for estimating the linkage intensity from F_2 data when two of the four classes are indistinguishable. The relative accuracy of such estimations of linkage compared with estimations made when all four classes are distinguishable is given.

2. The manner in which data from F_3 progenies, of selfed F_2 plants, may be used in calculating linkages is illustrated.

3. Doubly dominant F_3 progenies will supply more information per plant grown when separated into homozygotes and single and double heterozygotes alone, than will F_2 classification when p is less than 8 percent in repulsion. When the double heterozygotes are separated as well and the additional F_2 information from the F_3 lines is utilized such double dominant progenies, classified into genotypes in F_3 , supply more information per plant grown than the F_2 when p is less than 26 percent in repulsion.

4. Singly dominant F_3 progenies are more valuable than an F_2 of equal members when p is less than 7.5 percent in repulsion.

5. When all genotypes are viable the doubly dominant F_3 progenies separated into homozygotes and single and double heterozygotes supply more information than single dominant F_3 progenies when p is less than .18 or greater than .69.

6. When the homozygous recessives, for one of the characters, die or are indistinguishable F_3 separation of the double dominants into homo-

zygotes and single and double heterozygotes proves much more profitable than growing an F_2 of equal size when p is less than .315. Such doubly dominant F_3 progenies are more profitable than single dominant progenies when p is less than .18. For looser linkage in repulsion, or in coupling crosses, the single dominants would be preferable.

7. The manner of estimating the linkage intensity from F_2 data and F_3 data of various kinds is illustrated. The standard error to be attached to this crossover percentage is calculated from the total amount of information.

LITERATURE CITED

- COLLINS, G. N., 1924 Measurement of linkage values. *J. Agric. Res.* **27**: 881-891.
- FISHER, R. A., 1928 Statistical methods for research workers. Ed. 2. Edinburgh: Oliver & Boyd.
- 1930 Statistical methods for research workers. Ed. 3. Edinburgh: Oliver & Boyd.
- FISHER, R. A. and BALMUKAND, BHAI, 1928 The estimation of linkage from the offspring of self heterozygotes. *J. Genet.* **20**: 79-92.
- HUTCHINSON, J. B., 1929 The application of the "method of maximum likelihood" to the estimation of linkage. *Genetics* **14**: 519-537.
- IMMER, F. R., 1930 Formulae and tables for calculating linkage intensities. *Genetics* **15**: 81-98.
- 1931 The efficiency of the correlation coefficient for estimating linkage intensities. *Amer. Nat.* **65**: 567-572.
- ROBERTSON, D. W., DEMING, G. W. and KOONCE, DWIGHT, 1932 Inheritance in barley. *J. Agric. Res.* **44**: 445-466.