AN ANALYSIS OF LOCAL VARIABILITY OF FLOWER COLOR IN LINANTHUS PARRYAE*

SEWALL WRIGHT
The University of Chicago[1]

Received November 9, 1942

THE DISTRIBUTION OF LINANTHUS PARRYAE

EPLING and DOBZHANSKY (1942) have recently published a very detailed account of the distribution of two alternative characteristics, blue and white flowers, of a diminutive annual plant, *Linanthus Parryae*, in a portion of its range in the Mojave desert. It seemed of interest to make an analysis of their data from the standpoint of the theory of isolation by distance discussed in the preceding paper (WRIGHT 1943).

The region in question is about 80 miles long and averages about 10.5 miles wide. It stretches in an east-west direction along the piedmont north of the San Gabriel and San Bernardino Mountains and is largely isolated from other populations of the species. Data were obtained at stations every half mile along the principal roads, including two or three parallel roads at most places. At each station if the plants were present, counts were made of four samples of 100 plants each, spaced at intervals of approximately 250 feet at right angles to the road.

The vegetation in this piedmont belt is stated to be homogeneous. A number of species of shrubs are listed as characteristic. "The spacing of these shrubs is wide, and it is doubtful if they cover as much as 60 percent of the ground. Around the base of most of the bushes, the soil has accumulated so as to form a slight mound. *Linanthus Parryae* occupies the ground between the mounds, forming a widespread reticulum which is interrupted only by the stream beds or depositions alluded to above." The authors find nothing to suggest any selective differential.

Some of the conclusions reached by the authors are given as follows. "The apparent complexity of the distribution pattern of white and blue flower color in *Linanthus Parryae* can be reduced to a relatively simple scheme. The blue was found principally in three or four 'variable areas.' Outside these areas the blue was encountered sporadically, as would be expected if it were introduced there only on rare occasions through mutation or through occasional transport of 'blue' pollen or seed. Within the variable areas, the white and blue occurred side by side, and the population was differentiated into an extremely fine mosaic of microgeographic races. Pure white and pure blue colonies occurred at distances as small as 500 feet. Nevertheless, populations found one mile or less apart, resemble each other more than do populations taken at random in

* A portion of the cost of composing the mathematical formulae is borne by the Galton and Mendel Memorial Fund.

[1] Acknowledgment is made to the DR. WALLACE C. and CLARA A. ABBOTT MEMORIAL FUND of the UNIVERSITY OF CHICAGO for assistance in connection with the calculations.

the variable areas." This is followed by comparison with the U-shaped distribution of gene frequencies which the present author has deduced as characteristic in effectively small populations.

THE HIERARCHY OF SUBDIVISIONS

For more detailed mathematical analysis, it is convenient to define a hierarchy of subdivisions. Six compact, approximately equal, primary subdivisions are recognized along the length of the range. Each of these includes five secondary subdivisions. Each of these in turn includes four tertiary subdivisions. The tertiary subdivisions were chosen so as to include three stations as far as possible (two stations in six cases, four stations in two cases). The stations, as noted above, typically include four samples, but many of them contain less.

SMALLER SUBDIVISIONS	PRIMARY SUBDIVISIONS (EAST TO WEST) I	II	III	IV	V	VI	TOTAL
Secondary	5	5	5	5	5	5	30
Tertiary	20	20	20	20	20	20	120
Stations	57	59	60	60	61	59	356
Samples	198	211	214	214	218	203	1258

The population density in 1941 was found to vary from 1 to 26 per square foot (average 9.7) in the variable areas and from 1 to 48 per square foot (average 7.4) in the predominantly white areas. The area occupied by an average sample of 100 of the plants may thus be taken as about 12 square feet. It is stated, however, that in unfavorable years the species may be found only in sparse concentration or abundant only locally. Since the effective size of a population depends much more on the number of productive individuals in unfavorable than in favorable years, it is probable that the typical density is very much less than 100 per 12 square feet.

The average distance between samples at a typical station (four samples spaced at 250 foot intervals in a line) is 417 feet. Thus a station may be considered as representative of a circle of about this radius and hence of an area of about 5.4×10^5 square feet ($417^2\pi$) or 0.020 square mile. A station would contain about 45,000 sample areas if these were closely packed. However, since it is stated that about 60 percent of the ground is occupied by other vegetation, this estimate is to be reduced accordingly. Using round numbers, this indicates that there were about 2×10^4 sample areas in the population represented by a station in 1941. But if unfavorable years are taken into account, this number would probably have to be reduced enormously.

The average distance between stations in a group of three at half mile intervals is two-thirds of a mile. A tertiary subdivision may thus be considered to be representative of an area of about 1.4 square miles ($=(\frac{2}{3})^2\pi$) and thus would contain about 70 station equivalents. No doubt there should be some

reduction to allow for interruptions. We shall use 50 as a round number for the station equivalents in a tertiary subdivision.

The secondary subdivisions typically include 12 stations spaced at half mile intervals, often along a straight road, but more often involving roads at right angles to each other or two close parallel roads. If along a straight line, the average distance between included stations is $2\frac{1}{6}$ miles. We shall consider a secondary subdivision to represent an area of about 14 square miles.

The area occupied by the entire population is about 840 square miles. The average area of one of the six primary subdivisions is thus about 140 square miles. If each of these were completely filled by its five recorded secondary subdivisions, the average area represented by each of the latter would be 28 square miles, just twice that estimated above. There were, however, large territories far from the roads which were not sampled. The smaller estimate accordingly seems preferable. These estimates are summarized below on the basis of the numbers in 1941 and on an arbitrary hypothesis.

TABLE 1

The hierarchy of subdivisions.

	AREA	ESTIMATED PLANTS IN 1941	NUMBER OF UNITS (A) 1941	NUMBER OF UNITS (B) ARBITRARY
Total population	840 sq. mi.	6×10^{10}	6×10^{8}	6×10^{6}
Primary subdivisions	140 sq. mi.	10^{10}	10^{8}	10^{6}
Secondary subdivisions	14 sq. mi.	10^{9}	10^{7}	10^{5}
Tertiary subdivisions	1.4 sq. mi.	10^{8}	10^{6}	10^{4}
Stations	0.02 sq. mi.	2×10^{6}	2×10^{4}	200
Samples	12 sq. ft. (A) 1200 sq. ft. (B)	100	1	1

It would appear that the total number of plants of *Linanthus Parryae* in this region was between 10^{10} and 10^{11} in 1941. The average effective size of breeding population over a period of years is probably much less. For the sake of comparison, two widely different hypotheses will be used for the area from which the parents of individuals are drawn: (A) the area occupied by 100 plants in 1941 (12 square feet on the average), (B) an area 100 times as large to allow for years in which the population is sparse. Estimate (A) would be practically the minimum possible even if 1941 were a typical year. Estimate (B) is quite arbitrary.

THE SIGNIFICANCE OF DIFFERENTIATION WITH STATIONS

The first statistical question that requires consideration is whether or not there are greater differences among samples from the same station than are expected from the accidents of sampling. Let p be the actual but unknown frequency of blue in a homogeneous local population. Let p_0 be the observed frequency in a random sample of N individuals. Let $\delta p = p_0 - p$. Then $\sigma_{\delta p}^2 = p(1-p)/N$. This may be estimated from the observed frequency by apply-

ing the usual Gaussian correction after substituting p_0 for p. Thus for L samples from the same homogeneous population

$$\frac{1}{L}\sum_{1}^{L} p_0(1-p_0) = \frac{1}{L}\sum_{1}^{L}(p+\delta p)(1-p-\delta p) \tag{1}$$
$$= p(1-p) - \sigma^2_{\delta p} \quad \text{if} \quad L \doteq \infty$$
$$= (N-1)\sigma^2_{\delta p}$$

$$\sigma^2_{\delta p} = p_0(1-p_0)/(N-1) \tag{2}$$

taking $p_0(1-p_0)$ as the best estimate of the theoretic mean, obtainable from a single sample.

In a group of K samples of N each, not necessarily drawn from a homogeneous population, and with observed variance $\sigma_{p_0}{}^2 = \sum_1^K (p_0-\bar{p}_0)^2/K$

$$\overline{\sigma^2_{\delta p}} = \frac{1}{K}\sum_{1}^{K}[p_0(1-p_0)/(N-1)] \tag{3}$$

$$\overline{\sigma^2_{\delta p}} = [\bar{p}_0(1-\bar{p}_0) - \sigma^2_{p_0}]/(N-1). \tag{4}$$

The deviation of the frequencies, p_0, of samples about the unknown actual frequency, $\bar{p}$, of the whole heterogeneous population may be analyzed into two independent components, the deviation from the observed mean frequency $\bar{p}_0$ of the K samples and the deviation of this from $\bar{p}$

$$(p_0-\bar{p}) = (p_0-\bar{p}_0) + (\bar{p}_0-\bar{p}) \tag{5}$$

$$\sigma^2_{(p_0-\bar{p})} = \frac{1}{L}\sum_{1}^{L}\sum_{1}^{K}(p_0-\bar{p}_0)^2/K + \sigma^2_{(\bar{p}_0-\bar{p})}/K \tag{6}$$

$$\frac{K-1}{K}\sigma^2_{(p_0-\bar{p})} = \frac{1}{L}\sum_{1}^{L}\sigma^2_{p_0}. \tag{7}$$

If the observed variance $\sigma_{p_0}{}^2$ be treated as representative of the theoretical average $\sum\sigma_{p_0}{}^2/L$, we obtain the formula with ordinary Gaussian correction for the uncertainty of the mean. This, however, will not do in this case, because $\sigma_{p_0}{}^2$ is not independent of $\bar{p}_0$. If, for example, the frequencies of blue in four samples of 100 plants are 100, 100, 0, 0, respectively, the observed variance, $\sigma_{p_0}{}^2 = \frac{1}{4}$, is the theoretical maximum, and application of the Gaussian correction gives the impossible estimate $\sigma^2_{(p_0-\bar{p})} = \frac{1}{3}$. Since $\sigma_{p_0}{}^2$ necessarily approaches 0 as $\bar{p}_0$ approaches either 0 or 1, assume as a first approximation that $\sigma_{p_0}{}^2 = C\bar{p}_0(1-\bar{p}_0)$ where C is a constant for a given group of samples. Then

$$\frac{K-1}{K}\sigma^2_{(p_0-\bar{p})} = \frac{C}{L}\sum_{1}^{L}\bar{p}_0(1-\bar{p}_0) \tag{8}$$

$$= C\bar{p} - \frac{C}{L}\sum_{1}^{L}\bar{p}_0{}^2 \tag{9}$$

(10) But $$\sigma^2_{(\bar{p}_0-\bar{p})} = \frac{1}{L}\sum_{1}^{L}\bar{p}_0^2 - \bar{p}^2 \quad \text{and also} \quad \sigma^2_{(p_0-\bar{p})}/K.$$

(11) Thus $$\frac{K-1}{K}\sigma^2_{(p_0-\bar{p})} = C\bar{p}(1-\bar{p}) - \frac{C}{K}\sigma^2_{(p_0-\bar{p})}$$

(12) $$\sigma^2_{(p_0-\bar{p})} = KC\bar{p}(1-\bar{p})/(K+C-1).$$

The best estimate of $\bar{p}(1-\bar{p})$ is $\bar{p}_0(1-\bar{p}_0)$. An approximation for the variance of sample frequencies, corrected for uncertainty of the mean is thus

(13) $$\sigma^2_{(p_0-\bar{p})} = K\sigma^2_{p_0}/(K+C-1) \quad \text{where} \quad C = \sigma^2_{p_0}/\bar{p}_0(1-\bar{p}_0).$$

It may be noted that if C is small there is an approach to the ordinary Gaussian correction, but if $C=1$ (as in the extreme case cited above) there is no Gaussian correction at all, and impossible estimates are avoided.

This variance includes the sampling errors in the determination of the local frequencies as well as the variance due to real differentiation of local populations. To obtain an estimate of the latter, the mean sampling variance, $\overline{\sigma^2_{\delta p}}$ as given by (4) must be subtracted.

(14) $$\sigma_p^2 = \frac{K\sigma_{p_0}^2}{(K+C-1)} - \frac{\bar{p}_0(1-\bar{p}_0) - \sigma_{p_0}^2}{(N-1)}.$$

While this seems to be as good an estimate as it is practicable to obtain from a single group of samples, it has serious limitations if $\bar{p}_0$ is close to 0 or 1. There is only one type of distribution in a group of four samples of 100 plants each for which $\bar{p}_0=.0025$—namely, samples with the frequencies 0, 0, 0, 1. Formula (14) gives very nearly $\sigma_p^2=0$ (exactly if C is treated as 0), but obviously no information is given (or can be given) on differentiation among samples from stations for which $\bar{p}$ (as opposed to $\bar{p}_0$) is .0025. Again, in such a station as 103 with frequencies 0, 0, 0, 5 the estimate of σ_p^2 is the maximum possible from a group of four samples with mean $\bar{p}_0=.0125$ but is much less than might occur among stations for which $\bar{p}=.0125$. The estimate from $\bar{p}_0$ and $\sigma_{p_0}^2$ is satisfactory for values of $\bar{p}_0$ less than .25 or greater than .75 (if four samples) only if the extreme type of distribution (0, 0, 0, n) is rare. In the present data, distributions of this extreme type are abundant below $\bar{p}_0=.05$ and above $\bar{p}_0=.95$ (24 in 40 stations, excluding those in which all plants or all but one were alike). There were 61 stations in which there were from 5 to 95 percent blues. Among these, only two showed the most extreme possible differentiation for their average (namely, stations 137 with 0, 0, 30, 0 blues and station 371 with 0, 100, 100 blues). The following estimates are based on these 61 stations.

NO. OF SAMPLES PER STATION	NO. OF STATIONS	ESTIMATED σ_p^2	ESTIMATED $\sigma^2_{\delta p}$
2	8	.0394	.0012
3	16	.0704	.0013
4	37	.0269	.0017
Total	61	.0400	.0015

The estimates of σ_p^2 varied enormously. In one case (station 292) the estimated variance was less (by an insignificant amount) than that expected from the accidents of sampling. At the opposite extreme were stations 371 (0, 100, 100) and 27 (0, 99, 100), which were largely responsible for the high mean estimate of σ_p^2 in stations with three samples. On the average, the total estimated variance is about 28 times that expected from accidents of sampling. There is thus no doubt of the reality of the differentiation among samples from the same station.

SIZE OF THE PARENTAL POPULATION ON FOUR HYPOTHESES

Of greater interest for our present purpose is the variability of gene frequencies. Unfortunately this depends on the answers to a number of questions which could be obtained only by experiments which have not yet been made.

It is conceivable that the difference between blue and white is not genetic at all, but this is highly improbable. Assuming that blue and white differ genetically, it makes a difference whether there is exclusive cross pollination, exclusive self fertilization, or some intermediate condition. While self pollination appears improbable, we shall consider this possibility as well as that of cross pollination. Under exclusive self pollination, the mode of inheritance makes no difference in the distribution of the blue and white clones. If, however, there is cross pollination, the estimates of gene frequencies from observed phenotypic frequencies depend on the mode of inheritance. We shall consider three extreme hypotheses—namely, that blue is recessive, that it is dominant, and that it depends on multiple factors and a threshold.

The primary purpose of the analysis will be to find the effective size of the population from which parents (of adjacent individuals in the case of exclusive self fertilization) must be drawn to account for the observed distribution as a cumulative consequence of sampling, according to the theory of isolation by distance recently presented (WRIGHT 1943). The possibility that the distribution may be affected by differential selection must also be examined. We shall consider first the differentiation demonstrated above to occur within stations and after this the differentiation among larger populations.

If there is exclusive self fertilization, we are concerned with the quantity $E = \sigma_p^2/\bar{p}(1-\bar{p})$, which measures the correlation between adjacent plants relative to the population of the station (WRIGHT 1943). In this formula, $\bar{p}$ is the mean, σ_p^2 the variance of the frequency of blue among the unit populations from which adjacent plants are drawn. However, since these populations are unknown, we can only calculate E from the samples of 100 plants. This has been done separately for each of the 61 stations in which $\bar{p}_0$ was between .05 and .95, using formula (14) to estimate σ_p^2 in each case. The values of E ranged from 0.00 to 1.00 in a very asymmetrical distribution. The distribution of $\sqrt{E}$ was more nearly normal.

If the hypothesis of exclusive self fertilization is correct, the variation of $\sqrt{E}$ should be independent of $\bar{p}_0$. This was tested by calculating the correlation and regression coefficients. The statistical constants are given in table 2. It

turns out that the regression coefficient of $\sqrt{\bar{E}}$ on $\bar{p}_0$, .22±.10, is slightly more than twice its standard error. This is not in good agreement with the hypothesis, although it cannot be considered to eliminate it.

The effective size of the parental population may be estimated from the average value of E. In these 61 cases the average value was .210. (It may be noted that this is considerably larger than the square of the average value of $\sqrt{\bar{E}}$ given in table 2 because of the great variability among stations. In fact $\bar{E}=\overline{(\sqrt{E})^2}+\sigma^2_{\sqrt{E}}=.155+.055=.210$.)

TABLE 2

Statistical constants under four different genetic hypotheses. $\bar{p}_0$,$\bar{q}$ and $\bar{m}$ are the stations means, and $\bar{\bar{p}}_0$, $\bar{\bar{q}}$ and $\bar{\bar{m}}$ are the means of these means for the group of stations considered. In the first column, b is the regression of $\sqrt{E}$ on $\bar{p}_0$ and r is the correlation between these variables. Similarly b and r refer to the corresponding variables in the other columns.

(1) SELF FERTILIZATION		CROSS FERTILIZATION SINGLE GENE DIFFERENCE			MULTIPLE ADDITIVE FACTORS AND THRESHOLD	
			(2) BLUE DOMINANT	(3) BLUE RECESSIVE		
RANGE ($\bar{p}_0$)	.05 to .95	RANGE (q)	.10 to .90	.10 to .90	RANGE ($\bar{m}$)	−1.85 to +1.85
NO.	61	NO.	47	63	NO.	57
$\bar{\bar{p}}_0$	.440	$\bar{\bar{q}}$	.380	.517	$\bar{\bar{m}}$	−.212
$\sigma_{\bar{p}}$	.254	$\sigma_{\bar{q}}$	.202	.245	$\sigma_{\bar{m}}$	.932
$\sqrt{\bar{E}}$	.394	$\sqrt{\bar{F}}$	.380	.373	$\sqrt{\bar{F}}$	.399
$\sigma_{\sqrt{E}}$	.234	$\sigma_{\sqrt{F}}$	.256	.229	$\sigma_{\sqrt{F}}$	.232
r	+.24±.10	*r*	+.56±.10	−.21±.12	*r*	+.16±.13
b	+.22±.10	*b*	+.70+.16	−.20+.12	*b*	+.040+.033
$\bar{E}$	.210	$\bar{F}$	.210	.192	$\bar{F}$	.213
N (A)	45	N (A)	25	27	N (A)	25
N (B)	25	N (B)	14	15	N (B)	14

If for the moment we assume that the samples correspond in size to the groups from which the parents of any individual are drawn (hypothesis (A)) an estimate can be made of the effective population number of these groups from the theory of area continuity (WRIGHT 1943, fig. 7). Under this theory it requires a parental population of about 45 to give a value of E of .210 in a total (station) including 2×10^4 of these unit populations. This can agree with the actual number of plants in a sample (100) on taking account of the fact that these varied enormously in productivity. However, as noted, 1941 was a year of exceptional abundance. On the average it might require a considerably larger area than indicated in this year to provide an effective population number of 45. But in this case there would be less than 2×10^4 such groups in the area represented by a "station," which in turn would require a smaller estimate of the population number of the parental group. If, for example, a station includes only 200 parental groups (hypothesis (B)) instead of 20,000,

the effective population number of these under the theory would be about 25.

Consider next estimates under the hypothesis of prevailing cross pollination with blue dependent on a single differential gene. If blue is dominant, its gene frequency in a sample is given by $q = 1 - \sqrt{1 - p_0}$. A variation, δp in phenotypic frequency, implies the variation $\delta q = \delta p / 2\sqrt{1-p}$ in gene frequency. Thus the sampling variance of gene frequencies is given by the formula

$$(15) \qquad \begin{aligned} \overline{\sigma_{\delta q}^2} &= \frac{1}{K}\sum_1^K \left[\frac{\sigma_{\delta p}^2}{4(1-p_0)}\right] \\ &= \frac{1}{K}\sum_1^K \left[\frac{p_0(1-p_0)}{4(N-1)(1-p_0)}\right] = \frac{\bar{p}_0}{4(N-1)}. \end{aligned}$$

The mean gene frequency, $\bar{q}$, and the variance, σ_q^2, were estimated for each station by the following formulae in which N was 100 and K usually 4.

$$(16) \qquad \bar{q} = \frac{1}{K}\sum_1^K q$$

$$(17) \qquad \sigma_q^2 = \frac{K\sigma_{q_0}^2}{(K + C - 1)} - \frac{\bar{p}_0}{4(N-1)} \quad \text{where} \quad \sigma_{q_0}^2 = \frac{1}{K}\sum_1^K q^2 - \bar{q}\sum q$$

$$\text{and} \qquad C = \sigma_{q_0}^2/\bar{q}(1-\bar{q}).$$

The quantity $F = \sigma_q^2/\bar{q}(1-\bar{q})$ (slightly different from the provisional quantity C) should be independent of $\bar{q}$ on a valid hypothesis. This was tested by calculating the regression of $\sqrt{F}$ on $\bar{q}$ for all cases in which $\bar{q}$ was between .10 and .90 (47 in number). There turns out to be a highly significant positive regression $+.70 \pm .16$. This means that there is a great deal too much variability at stations in which blue is common as compared with that in stations in which it is rare to be compatible with this hypothesis.

The average value of F on this hypothesis is .210 which would imply a parental population of about 25 if it is assumed that there are 2×10^4 or more of these in the area represented by a station (hypothesis (A)). However, if there are fewer per station, the estimate must be decreased. If, for example, there are only 200 parental populations per station area (hypothesis (B)) the estimate of effective size is about 14.

If blue is assumed to be recessive, its gene frequency in a sample is given by $q = \sqrt{p_0}$. Phenotypic variation δp implies $\delta q = \delta p / 2\sqrt{p}$ and mean sampling variance $\sigma_{\delta q}^2 = (1-\bar{p}_0)/4(N-1)$. There were 63 stations for which $\bar{q}$ was between .10 and .90 under this hypothesis. The values of F, $\sqrt{F}$, and σ_q^2 in each of these were estimated by formulae analogous to those used under the hypothesis of dominance.

It turned out that the regression of $\sqrt{F}$ on $\bar{q}$ is negative, $-.20 \pm .12$. There is not enough variability within stations in which blue is common to satisfy this hypothesis in contrast with the situation if blue is assumed dominant. The regression is less than twice its standard error, however, so that the data cannot be considered to be incompatible with recessiveness of blue.

The average value of F on the hypothesis of recessiveness is .192 and thus only slightly different from that on the hypothesis of dominance. The estimates of effective size of parental group are accordingly practically the same (27 under hypothesis (A), 15 under hypothesis (B)).

The third hypothesis with respect to the mode of inheritance, assuming cross fertilization, is that the alternative characters blue and white depend on whether the cumulative effects of multiple factors (with no dominance or interaction) exceed a certain threshold. The common type of polydactyly in guinea pigs is an example of a character in which there is a superficial simulation of simple Mendelian heredity but in which data from F_3 and later generations indicate the above mechanism (WRIGHT 1934). The methods used in the analysis of polydactyly may be applied except that the correction for the sampling error was given incorrectly. (Fortunately the correction was so small that this had no appreciable effect on the results.)

Assume that the multiple factors determine a substantially normal distribution on a primary scale but that the phenotype depends on whether the value on this scale is above or below a threshold. It is convenient to take the threshold as the zero point and take as the unit of measurement the standard deviation (on this primary scale) characteristic of a sample in the station under consideration. The proportion above the threshold in a given sample with mean at m on the scale (and thus threshold at $-m$ relative to the mean) is as follows in terms of probability integral,[2]

$$\text{pri } x_1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_1} e^{-x^2/2} dx$$

$$p = 1 - \text{pri } (-m) = \text{pri } m \tag{18}$$

$$m = \text{pri}^{-1} p. \tag{19}$$

A phenotypic variation δp implies $\delta m = \delta p/y$, where y is the ordinate of the unit normal curve at the threshold. Thus the sampling variance for m is $\sigma_{\delta m}^2 = p_0(1-p_0)/y^2(N-1)$. This was found for each sample and averaged to find the correction to be applied in calculating σ_m^2.

$$\sigma_m^2 = \left[\sum_1^K m^2 - \bar{m} \sum_1^K m \right] / (K-1) - \sum_1^K [p_0(1-p_0)/y^2]/K(N-1). \tag{20}$$

To estimate the inbreeding coefficient F from σ_m^2 it may be noted first that with multiple factors and no dominance or factor interaction, F measures the proportional *decrease* in the variance of characters within random breeding

[2] In a number of papers (WRIGHT 1926, 1934, etc.) prf x_1 has been used for the area of the probability curve between the mean and the deviation x_1, as a multiple of the standard deviation. This has the disadvantage that the inverse function is two-valued unless areas below the mean are treated as negative. The form used above avoids this difficulty. The recognized form erf $x_1 = (2/\sqrt{\pi}) \int_0^{x_1} e^{-x^2} dx$ is inconvenient for several reasons.

subgroups (WRIGHT 1921). The contribution of a pair of alleles A, a to the variance of a subgroup under the above conditions is $2q(1-q)(A)^2$ where (A) is the effect of replacing a by A. The average contribution in an array of such subgroups is $2[\bar{q}(1-\bar{q})-\sigma_q^2](A)^2$. But $\sigma_q^2=\bar{q}(1-\bar{q})F$, where F is the inbreeding coefficient of individuals relative to the total. Thus the average contribution of A, a to intragroup variance may be written $2(1-F)\bar{q}(1-\bar{q})(A)^2$. The total intragroup variance is merely the sum of such contributions under the conditions.

$$\sigma_g^2 = 2(1 - F)\sum [\bar{q}(1 - \bar{q})(A)^2]. \tag{21}$$

Next it may be noted that F measures the proportional *increase* in the variance of the *total* population resulting from subdivision into inbred lines. This total variance (σ_t^2) is compounded of the intragroup variance (σ_g^2) given above and the intergroup variance (σ_m^2). The contribution of A, a to the mean of each subgroup is of the form $2q(A)$. The variance of these is $4\sigma_q^2(A)^2 = 4F\bar{q}(1-\bar{q})(A)^2$.

$$\sigma_m^2 = 4F\sum [\bar{q}(1 - \bar{q})(A)^2] \tag{22}$$

$$\sigma_t^2 = \sigma_g^2 + \sigma_m^2 = 2(1 + F)\sum [\bar{q}(1 - \bar{q})(A)^2]. \tag{23}$$

(24) Thus $$F = \sigma_m^2/(2\sigma_g^2 + \sigma_m^2).$$

In the present case, $\sigma_g^2=1$ by hypothesis.

$$F = \sigma_m^2/(2 + \sigma_m^2). \tag{25}$$

Estimates of F have been made separately by this formula for each of 57 stations in which $\bar{m}$ was between -1.85 and $+1.85$. Again $\sqrt{F}$ should be independent of $\bar{m}$ if the hypothesis is valid. The regression of $\sqrt{F}$ on $\bar{m}$ turned out to be positive but not significant ($+.040\pm.033$). From this standpoint this hypothesis is satisfactory, more satisfactory in fact than any of the others considered.

The mean value of F was .213, substantially the same as under the hypothesis that blue depends on a single dominant (F=.210) or a single recessive (F=.192). The estimate of the effective size of the parental group is accordingly approximately the same under all of these hypotheses: 25 to 27 if there are 2×10^4 or more such groups in the area represented by a station but less if the number of random breeding units in a station is less than 2×10^4 (about 14 or 15 if the number of groups is as small as 200).

The effects of partial determination of the character by environmental factors are obvious under the last hypothesis. If there are environmental effects on individuals, not related to their location (for example, change in color with aging), the genetic component of intragroup variance would be less than 1, but σ_m^2 would not be affected. In this case, F is larger than estimated above, requiring N (size of random breeding unit) to be smaller to account for the observed differentiation within stations. In the case of *Linanthus Parryae*, the observations of EPLING and DOBZHANSKY tend to rule this out.

If, on the other hand, there are environmental influences (such as character

of soil) that make a difference between samples but rarely between individuals of the same sample, it is σ_m^2 that must be reduced if it is to represent genetic differentiation and the estimate of F becomes smaller than calculated above. In this case, a larger value of N is implied. If intersample variance is almost entirely environmental, even though genetic segregation occurs within samples, F is almost zero, and it is implied that N is so large (for example, over 1000) that there is virtual panmixia throughout the station. There is little likelihood, however, that flower color is due to such environmental effects in *Linanthus Parryae.*

ANALYSIS OF VARIABILITY WITHIN THE RANGE

It is next of interest to analyze the variability in the region as a whole. The mean gene frequencies in the primary and secondary subdivisions were as follows, assuming blue to be recessive.

TABLE 3

Gene frequencies in the primary and secondary subdivisions assuming blue to be recessive.

PRIMARY SUBDIVISION	SECONDARY SUBDIVISION A	B	C	D	E	TOTAL
I	.573	.504	.717	.657	.302	.551
II	.339	.032	.007	.005	.008	.078
III	.009	.000	.000	.000	.000	.002
IV	.010	.005	.000	.000	.068	.017
V	.126	.004	.002	.000	.000	.027
VI	.000	.106	.411	.224	.014	.151
Total						.137

The distribution on the map is shown in figure 1. This brings out the three separate centers of high frequency, one in I overlapping II, a second in VI and a third in the southern parts of IV and V, to which EPLING and DOBZHANSKY called attention.

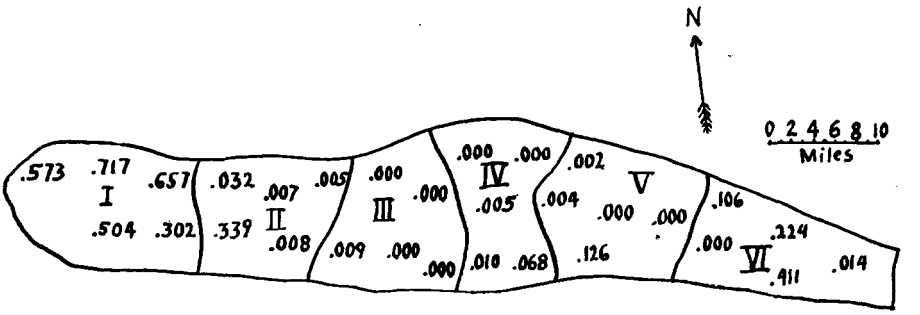

FIGURE 1.—The geographical distribution of the frequency of the gene for blue flowers (if recessive) in the portion of the range of *Linanthus Parryae* investigated by EPLING and DOBZHANSKY. The six primary subdivisions are indicated by Roman numerals.

It will be convenient to symbolize the various levels in the hierarchy by subscripts: 1 for a primary subdivision, 2 for a secondary, 3 for a tertiary, 4 for a

station, and 5 for a sample. Thus q_4 represents the mean gene frequency in a station, K_4, the number of recorded stations in a tertiary population, etc., $\sigma_{4.3(G)}^2$ is used for the gross variance of q_4 (stations) within the next higher category (with the modified Gaussian correction for uncertainty of the mean, but without correction for the sampling variance of the next lower level); $\sigma_{4.3}^2$ is the final estimate including this last correction. Finally, $\sigma_{4.t}^2(=\sigma_{4.3}^2+\sigma_{3.2}^2+\sigma_{2.1}^2+\sigma_{1.t}^2)$ is the variance of q_4 within the total population.

The variance of samples within stations was obtained for all of the 356 stations.

$$\sigma^2_{5\cdot4(G)} = \frac{\sum_1^{356}[\sum q_5^2 - q_4\sum q_5]}{\sum_1^{356}[K_5 - 1 + \bar{C}_4]} = \frac{7.2523}{959.3} = .0076$$

$$\overset{2}{\sigma}_{5\cdot4} = .957\overset{2}{\sigma}_{5\cdot4(G)} = .0073.$$

The summations in the bracket refer to the K_5 (usually four) samples of a station. $\bar{C}_4=.161$ is the average value of $C_4(=\sigma_{q_0}^2/\bar{q}(1-\bar{q}))$ for the 63 stations in which q_4 was between .10 and .90. The mean sampling variance of q_5 in these 63 stations was such that $\sigma_{5.4}^2/\sigma_{5.4(G)}^2$ was .957. This ratio is carried over to all stations.

The variance of stations within the 120 tertiary subdivisions was as follows.

$$\overset{2}{\sigma}_{4\cdot3(G)} = \frac{\sum_1^{120}[\sum q_4^2 - q_3\sum q_4]}{\sum_1^{120}[K_4 - 1 + \bar{C}_3]} = \frac{4.7112}{267.9} = .0176$$

$$\overset{2}{\sigma}_{4.3} = \overset{2}{\sigma}_{4\cdot3(G)} - 356\overset{2}{\sigma}_{5\cdot4(G)}/1258 = .0176 - .0021 = .0154.$$

The number of stations within tertiary subdivisions (K_4) was usually 3. There were 30 tertiary subdivisions in which q_3 was between .10 and .90.

$$\bar{C}_3 = \frac{1}{30}\sum_1^{30}\left[\frac{\sum q_4^2 - q_3\sum q_4}{K_4q_3(1-q_3)}\right] = .266.$$

Similarly the variance of tertiary subdivisions within the 30 secondary subdivisions was as follows.

$$\overset{2}{\sigma}_{3\cdot2(G)} = \frac{\sum_1^{30}[\sum q_3^2 - q_2\sum q_3]}{\sum_1^{30}[K_3 - 1 + \bar{C}_2]} = \frac{2.1386}{97.1} = .0220$$

$$\overset{2}{\sigma}_{3\cdot2} = \overset{2}{\sigma}_{3\cdot2(G)} - 120\overset{2}{\sigma}_{4\cdot3(G)}/356 = .0220 - .0059 = .0161.$$

Here K_3 was regularly 4. The average value of $\bar{C}_2$, calculated from ten secondary subdivisions for which q_2 was between .10 and .90, was .237.

The variance of secondary subdivisions within the six primary subdivisions was as follows.

$$\overset{2}{\sigma}_{2\cdot1(G)} = \frac{\sum_{1}^{6}\left[\sum q_2^2 - q_1 \sum q_2\right]}{\sum_{1}^{6}\left[K_2 - 1 + \overline{C_1}\right]} = \frac{.3218}{25.0} = .0129$$

$$\overset{2}{\sigma}_{2\cdot1} = \overset{2}{\sigma}_{2\cdot1(G)} - \overset{2}{\sigma}_{3\cdot2(G)}/4 = .0129 - .0055 = .0074.$$

Here K_2 was regularly 5. There were three primary subdivisions with q_1 considered large enough to warrant estimation of C_1—namely, I, II, and VI with values of C_1 of .084, .237 and .182 and $\overline{C_1} = .168$.

Finally the variance of the six primary subdivisions within the total was as follows.

$$\overset{2}{\sigma}_{1\cdot t(G)} = \frac{\sum_{1}^{6} q_1^2 - q_t \sum_{1}^{6} q_1}{(K_1 - 1 + C_t)} = \frac{.2197}{5.3} = .0414$$

$$\overset{2}{\sigma}_{1\cdot t} = \overset{2}{\sigma}_{1\cdot t(G)} - \overset{2}{\sigma}_{2\cdot1(G)}/5 = .0414 - .0026 = .0388.$$

Here

$$C_t = \frac{\sum q_1^2 - q_t \sum q_1}{6q_t(1-q_t)} = .309.$$

There is little doubt of the reality of differentiation at all of these levels except for the case of secondary within primary subdivisions in which $\sigma_{2.1(G)}^2$ is only 2.3 times the variance expected from accidents of sampling at the next lower level. The extreme departures from normality in the distribution of q, however, must be kept in mind. They make the applicability of FISHER'S (1938) z test somewhat dubious.

The variance of the groups at each level within the total may be obtained by adding the variances down to the level in question. Thus $\sigma_{1\cdot t}^2 = .0388$, $\sigma_{2\cdot t}^2 = .0462$, $\sigma_{3\cdot t}^2 = .0623$, $\sigma_{4\cdot t}^2 = .0777$ and $\sigma_{5\cdot t}^2 = .0850$. By dividing these expressions by $q_t(1-q_t)$ where $q_t = .137$ we obtain values of $\sigma_{1\cdot t}^2/q_t(1-q_t)$ etc. This expression has been shown to be equal to $(F_t - F_i)/(1-F_i)$ under area continuity, in the absence of disturbing factors (WRIGHT 1943). It is .717 for samples, .656 for stations, .525 for tertiary groups, .390 for secondary groups, and .327 for primary groups. In figure 2 the square roots of these figures are plotted against number of random breeding units and compared with the theoretical curves for $N = 10$ and $N = 20$ deduced in the paper referred to above. The number of random breeding units is as given in table 1 using the arbitrary assumption (B) that there are 200 of these to a station.

From inspection of this figure it appears that there is considerably more variability of the higher categories than expected from that within stations. It would require a random breeding unit of considerably less than ten to account for this variability in contrast with about 15 indicated by the variability within stations on the assumption made above. If the samples are considered to be the random breeding units (2×10^4 per station equivalent) the

discrepancy is greater, since the variability within stations indicates a random breeding unit of about 27 on this hypothesis, while the figures for the higher categories are only slightly modified.

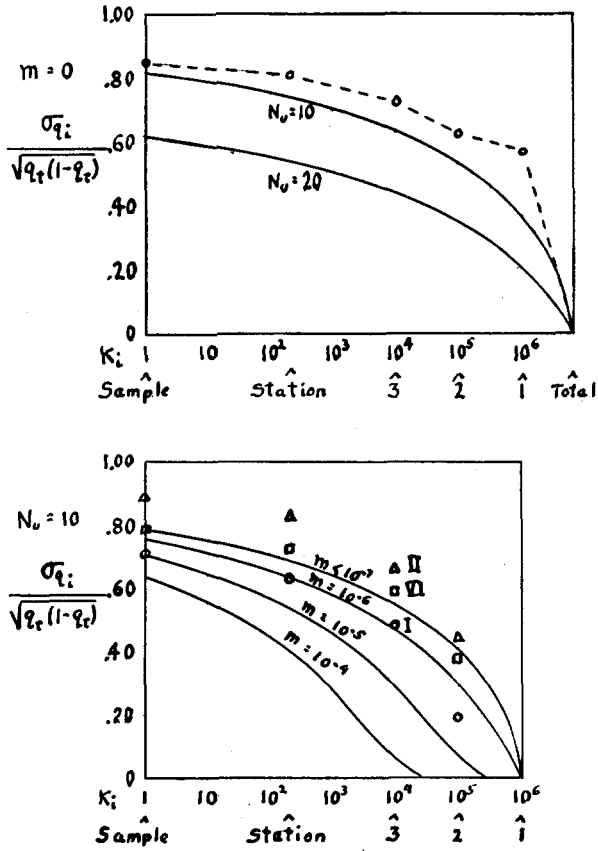

FIGURE 2.—The small circles connected by broken lines indicate the calculated variability of the frequency of the gene for blue (if recessive) among subgroups (samples, stations, tertiary, secondary and primary subdivisions of the range studied). Comparisons are made with the theoretical variability on the hypotheses that the effective number of individuals in a random breeding unit is 10 or 20, and that long range dispersal or mutation is negligible.

FIGURE 3.—The triangles indicate the calculated variability of the gene frequency of blue (if recessive) among subgroups of the primary subdivision II. The squares and circles do the same for primary subdivisions VI and I, respectively. These are compared with the theoretical amounts of variability on the hypothesis that N = 10 and that there is long range dispersal or mutation at rates up to $m = 10^{-4}$.

The most serious discrepancy is in the great variability of the primary subdivisions. In this case, however, the comparison with theory is hardly a fair one. The elongated range along the piedmont would favor a greater amount of differentiation than indicated by the theory of area continuity. There is some approach to the conditions of linear continuity.

TABLE 4

Analysis of variability in six primary subdivisions (I to VI). The numbers in column 1 designate the level in the hierarchy. Column 2 gives the number of groups at each level included in the primary subdivision. Column 3 gives the sum of squared deviations of q from the mean of the next higher category. This is divided by the entries in column 4, $\sum(K-1+C)$ to give the entries in column 5, which are gross intragroup variances not corrected for the accidents of sampling. The sampling variances are given in column 6. Subtraction of these from the entries in column 5 give the net intragroup variances of column 7. The running sum of these (column 8) gives the variance of the groups at each level within the primary subdivisions. In column 9 these are divided by $q_1(1-q_1)$ where q_1 is the mean gene frequency for the primary subdivision (table 3) to give the estimate of the quantity $(F_t-F_1)/(1-F_1)$. The square roots of these quantities are the ordinates in figure 3. Column 10 gives the ratio of the gross variance (column 5) to that expected from sampling (column 6).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10
I	2	5	.1037	4.2	.0249	.0172	.0077	.0077	.031	1.4
	3	20	1.1140	16.2	.0688	.0161	.0527	.0604	.244	4.3
	4	57	1.9389	42.3	.0458	.0072	.0386	.0990	.400	6.3
	5	198	3.7814	150.2	.0252	.0009	.0243	.1232	.498	—
II	2	5	.0854	4.2	.0205	.0062	.0143	.0143	.199	3.3
	3	20	.3990	16.2	.0246	.0069	.0178	.0321	.446	3.6
	4	59	.8993	44.3	.0203	.0017	.0186	.0507	.704	11.9
	5	211	.9825	161.5	.0061	.0002	.0059	.0565	.786	—
III	2	5	.0001	4.2	.0000	.0000	.0000	.0000	.006	—
	3	20	.0003	16.2	.0000	.0000	.0000	.0000	0	—
	4	60	.0037	45.3	.0001	.0001	.0000	.0000	0	—
	5	214	.0675	163.7	.0004	.0000	.0004	.0004	.213	—
IV	2	5	.0034	4.2	.0008	.0006	.0002	.0002	.014	1.3
	3	20	.0382	16.2	.0024	.0025	(−.0001)	.0001	.004	0.9
	4	60	.3440	45.3	.0076	.0007	.0069	.0070	.428	10.8
	5	214	.4094	163.7	.0025	.0001	.0024	.0094	.576	—
V	2	5	.0125	4.2	.0030	.0005	.0025	.0025	.096	5.8
	3	20	.0339	16.2	.0021	.0025	(−.0004)	.0021	.081	0.8
	4	61	.3484	46.3	.0075	.0011	.0065	.0085	.331	7.0
	5	218	.6391	166.8	.0038	.0001	.0037	.0122	.473	—
VI	2	5	.1167	4.2	.0280	.0085	.0195	.0195	.152	3.3
	3	20	.5533	16.2	.0342	.0090	.0252	.0446	.348	3.8
	4	59	1.1768	44.3	.0266	.0026	.0240	.0686	.535	10.2
	5	203	1.3724	153.5	.0089	.0003	.0086	.0772	.603	—

ANALYSIS OF VARIABILITY WITHIN THE PRIMARY SUBDIVISIONS

Because of the above consideration it is desirable to analyze the variability within the primary subdivisions. This has been done, with the results presented in table 4. Judging from the ratio of the gross variance to that expected from sampling (column 10), there appears to be significant differentiation at all levels within the primary subdivisions in which blue was not rare—namely, I($\bar{q}$=.551), II($\bar{q}$=.078), and VI ($\bar{q}$=.151) with one exception. The exception

is the absence of significant differentiation of secondary subdivisions in I. In addition, there was significant differentiation among secondary subdivisions in V and of stations in both IV and V in spite of the rarity of blues in these areas ($\bar{q}=.027$ in V, $\bar{q}=.017$ in IV). The data from III, which was almost uniformly white flowered ($\bar{q}=.002$), are wholly inadequate for any estimates. The significance of differentiation among samples within stations has not been determined separately for the primary subdivisions. As noted, the variance in stations in which $\bar{q}$ was between .10 and .90 was about 24 times that expected from the accidents of sampling and undoubtedly significant.

In figure 3 the values of $\sigma_{x,t}/\sqrt{q_t(1-q_t)}$ as estimates of $\sqrt{(F_t-F_i)/(1-F_i)}$ under the theory of area continuity are plotted against the estimates of the number of random breeding units included in the category in question, within primary subdivisions I, II and VI. Again it is assumed arbitrarily that there are 200 random breeding units in a station.

The curves for the three primary subdivisions agree with each other as well as can be expected. They are somewhat more nearly parallel to the theoretical curve for $N=10$ than when differentiation was considered relative to the entire range. There is still, however, more variability of the higher categories at least within II and VI than expected from the theory of area continuity. Again the discrepancy would be much more serious if the random breeding units are identified with the 1941 samples, 2×10^4 to a station.

There are other factors, however, that must be considered. Even if there are only 200 random breeding units per station, the number in one of the primary subdivisions is of the order 10^6 (and in the whole range considered here, 6×10^6). This would be the typical number of generations to common ancestors of individuals that are far apart if the theory of area continuity applies strictly. In this case it would require something like a million years for a local colony to spread over this area, which is certainly highly improbable. However, means of dispersal to great distances so rare that only a minute fraction of the population of any occupied region has such an origin would enable the species to spread over a large suitable range in a few years. On the other hand, the effects of even a minute amount of replacement by a random sample of the species are not negligible.

It was shown in the preceding paper that such replacement in the proportion m (whether due to long range dispersal or mutation or, as accurately as possible, of uniform selection) removes nearly all random differentiation of populations more than $1/m$ times the random breeding unit and considerably reduces such variability in populations one tenth of this size. Thus reversible mutation between blue and white at rates of the order of 10^{-6} per generation should practically eliminate random differentiation of primary subdivisions and somewhat reduce that of secondary subdivisions, even under hypothesis B (in which these are 10^6 and 10^5 times the random breeding unit respectively). Admixture of a random sample of the species into all populations at the rate 10^{-4} per generation would practically eliminate all random variability of tertiary and larger subdivisions under the same hypothesis. Under hypothesis A this would eliminate random variability even among stations. The varia-

bility of samples within stations, however, would not be affected under hypothesis B and not very much under hypothesis A. The expected variability at each level in the hierarchy is shown in figure 3 for $N = 10$ and $m = 10^{-4}$, 10^{-5}, 10^{-6} and $m < 10^{-7}$.

The large amount of differentiation actually found at all levels up to the highest indicates that some other factor than mere accumulation of sampling differences has been at work.

One possibility is that there are irregularities in the distribution, including differences in density. This would cause a greater amount of random differentiation of the larger categories than expected under a uniform distribution.

The most obvious possible factor that could counterbalance the effects of long range dispersal and mutation, however, is differential selection. Mr. W. HOVANITZ, in a personal communication, suggests that the climatic conditions may differ sufficiently near the ends of the region studied (I and VI, in which blue was relatively common) from those in the middle (where blue was rare) to make such an interpretation plausible. It is less plausible for the differences among secondary and tertiary subdivisions of the same primary subdivision.

There is a possibility, however, of selective differentiation even in the absence of any environmental differentiation. As noted in previous papers, the random differences in gene frequency, occurring in all series of alleles up to a certain level in the hierarchy, create a unique genetic system in each locality. Slightly different adaptive systems may be arrived at in different localities. If the gene or genes which distinguish blue and white play a role in any such systems, this would give a basis for locally different selection pressures.

The distribution of blue and white can be accounted for most easily by supposing that most of the differentiation of the smaller categories is random in character and due to the accumulation of sampling accidents in random breeding groups of one or two dozen productive individuals per year but that at the higher levels, processes which tend to pull down random differentiation such as mutation and especially occasional long range dispersal are counterbalanced by selective differentials between local genetic systems.

SUMMARY

The detailed account of the distribution of blue and white flowers of the annual plant *Linanthus Parryae* in a region of the Mojave desert by EPLING and DOBZHANSKY provides interesting material for comparison with the theoretical amount of random differentiation in a population that is continuous but in which dispersal is severely restricted.

For this purpose the 840 square miles studied is broken up into a hierarchy of subdivisions. There proves to be highly significant differentiation of samples (of 100 plants) within stations (representative of about 0.02 square miles). There is significant differentiation of stations within tertiary subdivisions (about 1.4 square miles), of these within secondary subdivisions (about 14 square miles), of these within primary subdivisions (about 140 square miles), and very marked differentiation of the primary subdivisions along the somewhat narrow piedmont zone.

Assuming that the difference is a genetic one, four hypotheses are considered in connection with the variability of samples within stations. It is improbable that reproduction is by self fertilization, but if it is, it would require that the population from which adjacent individuals are derived be about 45 to account for the observed variability, accepting the density of population found in 1941 as typical. This, however, was an unusually favorable year. Arbitrarily assuming 200 units per station instead of 20,000 as indicated in 1941, an effective population number of about 25 per unit is indicated.

If there is predominant cross fertilization, it makes no appreciable difference whether blue depends on a single dominant or a single recessive gene or on multiple factors and a threshold. Under any of these hypotheses, the effective population number of the random breeding unit comes out 25 to 27 if the 1941 estimate of numbers is accepted and 14 or 15 if the area occupied by a parental unit is assumed to be 100 times as large.

The amount of differentiation of the higher categories is somewhat greater than expected as a random consequence of that of the lower categories, under continuity of area and with negligible rates of long range dispersal and mutation (rates less than 10^{-7} per generation) and no differential or other selection. This is especially true of the primary subdivisions, but here the theory is unsatisfactory because of the elongated character of the range, which favors excessive differentiation.

However, there is somewhat too much variability of the higher categories even within the compact primary subdivisions to be accounted for as wholly random under the assumption above. As it is highly probable that the theoretical values should be substantially reduced because of long range dispersal at rates greater than 10^{-6} per generation, there is probably a counterbalancing influence of differential selection. This would not necessarily depend on differential environmental conditions. It could be a by product of the development of different genetic systems by the process of random differentiation.

LITERATURE CITED

EPLING, C., and TH. DOBZHANSKY, 1942 Genetics of natural populations. VI. Microgeographical races in *Linanthus Parryae*. Genetics **27**: 317–332.

FISHER, R. A., 1938 Statistical methods for research workers. 7th ed. 356 pp. Edinburgh: Oliver and Boyd.

WRIGHT, S., 1921 Systems of mating. Genetics **6**: 111–178.

1926 A frequency curve adapted to variation in percentage occurrence. J. Amer. Statist. Ass. **21**: 162–178.

1934 An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics **19**: 506–536.

1943 Isolation by distance. Genetics **28**: 114–138.