Kosambi and the Genetic Mapping Function

K K Vinod

K K Vinod is a Senior Scientist at the Rice Breeding and Genetics Research Centre, Indian Agricultural Research Institute, Aduthurai, Tamil Nadu. His research interests are on genetics and breeding of rice especially on tolerance to abiotic stresses like salinity and low nutrients. He works on QTL mapping and marker assisted selection for low nitrogen and phosphorus response, drought tolerance and grain quality in rice.

One of the activities that geneticists pursue is the generation of genetic maps of organisms that chart the locations of the different genes on the chromosomes. Damodar Dharmanada Kosambi was not a geneticist by training and profession, but a mathematician. Yet, in 1944, Kosambi wrote a classic paper about mapping function in genetics in the journal *Annals of Eugenics*. His lone paper in genetics was tilted 'The estimation of map distance from recombination values'. The mapping function that Kosambi derived is still widely used by geneticists the world over for the mapping of genomes.

Keywords
Chromosome-mapping, recombination, genetic interference.

1. Introduction

A major feature of sexual reproduction is the phenomenon of recombination that occurs during the formation of gametes by the process of cell division known as meiosis. During meiosis, the homologous chromosomes (maternal and paternal) pair up and exchange genetic material by a process known as crossing over or recombination. The net result is that the genetic alleles carried on the chromosomes are shuffled to generate variability. The frequency at which recombination occurs between two loci on the same chromosome can be measured by setting up crosses between genetically defined individuals and measuring the frequencies of the progeny carrying combinations of traits that are different from that seen in the parents. The larger the genetic distance between the loci, the higher will be the recombination frequency. Mapping functions are nothing but mathematical adjustments in the estimated genetic distance between two loci located on a chromosome, calculated based on recombination values. Adjustments are necessary to accommodate uncertain events of crossovers and crossover suppression that happen during meiotic division. This process of estimating the genetic

distance (not the physical distance) between two loci on a chromosome is called linkage mapping.

1.1 *Genetic Linkage and Mapping – A Brief History*

Gregor Johann Mendel discovered that segregation of simple traits in pea plants occurred by way of discrete factors that are today known as genes. He showed that the factors responsible for different traits are inherited independent of each other, which he formulated as one of the laws of heredity – the law of independent assortment. The dawn of the twentieth century witnessed dramatic development in the science of genetics, following the rediscovery of Mendel's findings in 1900. One of the major discoveries that followed immediately was the identification that the chromosomes are the carriers of Mendel's factors. Since the number of chromosomes in any organism is smaller than the number of traits, each chromosome has to carry multiple factors. In that case, it is to be expected that factors located on the same chromosomes will be inherited together – in other words they will be linked to each other.

Gregor Johann Mendel discovered that segregation of simple traits in pea plants occurred by way of discrete factors that are today known as genes.

Box 1. The Chromosome Theory and Linkage

1. Genes are arranged on the chromosome in a linear fashion.
2. Genes located in the same chromosome tend to stay together during every generation or inheritance.
3. The distance between two genes decides whether they will be inherited together and if so at what frequency.

Around the year 1905, William Bateson, Edith Saunders and Reginald Punnett explored the ways by which traits were assorted by studying individuals with different contrasting traits. They found that not all traits perfectly assorted among themselves and followed Mendelian ratio. Some traits (flower colour and pollen shape in sweet pea plants) showed a sort of 'coupled-with-each-other' effect resulting in deviation from the expected ratio [2]. Conspicuously, the deviations did not follow any expected ratio and the progenies were more conservative towards parental types and less towards recombinants. Predominance of parental phenotypes led Bateson and his associates to hypothesize that there was *some 'coupling' between pollen shape and flower colour*, and that this coupling or association resulted in the observed deviation from independent assortment.

Answer to this puzzle came from Thomas Hunt Morgan's famous 'fly room' of Columbia University in 1910[1]. While experiment-

[1] See *Resonance*, Vol.8, No.11, 2003.

Morgan and his colleagues found that genes responsible for certain traits (for example, eye colour in *Drosophila*) reside on the X chromosome based on the inheritance pattern of the trait.

ing with the fruit fly (*Drosophila melanogaster*) that he introduced to genetics as a model organism, Morgan and his colleagues found that genes responsible for certain traits (for example, eye colour in *Drosophila*) reside on the X chromosome based on the inheritance pattern of the trait. This was the first hypothesis on the presence of a gene on a particular chromosome, physical evidence of which appeared shortly after, in 1914, from one of Morgan's students, Calvin Blackman Bridges. By then, Morgan and his students identified many genes linked to the X chromosome. By consolidating the information on the genes that show a 'coupled-with-each-other' effect from his own and other labs, Morgan wrote in a 1911 issue of *Science* that "*Instead of random segregation in Mendel's sense we find "association of factors" that are located near together in the chromosomes. Cytology furnishes mechanism that the experimental evidence demands*", clearly hypothesising on linkage and recombination through crossing over. In linkage, more parental types occurred due to the absence of crossing over, while recombinants were results of crossing over. So the first assumption was that those traits showing parental-type predominance resulted from genes that are close together on a chromosome, and those assorted well were either far apart on a chromosome or were on different chromosomes. Linkage was an essential requirement to ascertain the proximity of genes.

Taking this idea further from his professor, Alfred Henry Sturtevant, a nineteen-year undergraduate student of Morgan, went on to develop the first genetic map of the X chromosome of *Drosophila* with six linked factors in 1913. Sturtevant's map was accurate with regard to placement of genes and was logically strong. He described a genetic map as a linear arrangement of genes on a chromosome and wrote that "*the proportion of cross-over could be used as an index of the distance between any two factors*". Sturtevant described that 'map distance' is not physical distance but rather was some joint function of length and strength over a region of chromosome.

Sturtevant described that 'map distance' is not physical distance but rather was some joint function of length and strength over a region of chromosome.

Map to Mapping Functions

In 1916, Morgan and Bridges consolidated the findings so far and published a book *Sex-Linked Inheritance in Drosophila*, in which the first elaborate map of the X chromosome appeared (*Figure* 1). They described that the factors (the word 'gene' was not used in this book) are strictly linearly arranged on chromosomes and their distances are additive of their recombination fraction [3]. The year 1919 witnessed a war of arguments between William Earnest Castle and Morgan on the hypothesis of linear arrangement of factors. Castle argued that the linear arrangement of genes was 'doubtful', and Morgan and his students' assumptions were 'absurd', citing many of Morgan's and Sturtevant's data. Castle's arguments centred on two points, there was no perfect additivity of the map distances, and genetic distances of more than 50 was improbable. This made Castle to present a three-dimensional model for the arrangement of genes, mainly to support the additivity problem. Morgan and his students (Bridges and Sturtevant) countered Castle by arguing that his approach of combinatory data analyses from independent experiments was wrong and unacceptable. Further, they demonstrated that occurrence of double crossovers between two loci which are wider apart gets unaccounted, resulting in underestimation of recombination frequencies. They also argued that the map unit of more than 50, as shown in their map, was actually cumulative distances of adjacent loci and not the observed crossover fractions.

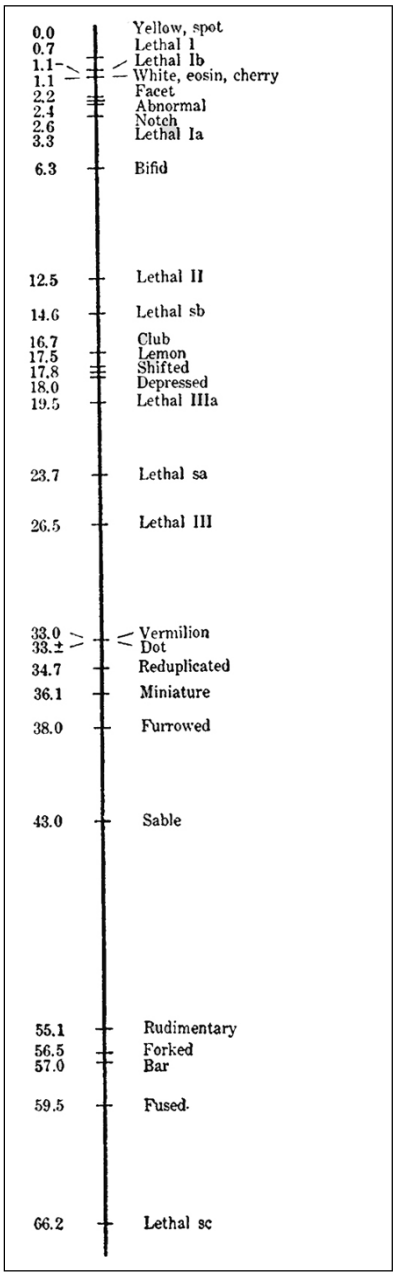


Figure 1. The first linkage map of X chromosome of Drosophila with map order of 29 loci and distances (From Morgan and Bridges 1916).

2. Haldane's Mapping Function

Settling the arguments of the Castle–Morgan, in the same year, John Burdon Sanderson Haldane[2], a British geneticist, wrote a classic paper in the *Journal of Genetics* describing "*a more accurate theory of the relations* inter se *of low cross-over values, and of their connexion with the distances apart of the loci of factors in a chromosome*". One can see that Haldane went way

[2] See *Resonance*, Vol.3, No.12, 1998.

[3] See *Resonance*, Vol.2, No.9, 1997.

[4] See *Resonance*, Vol.4, No.12, 1999.

ahead in thinking and described that the Sturtevant's theory of proportionality gives consistent result only when crossover values are small, and is not accurate for larger values, because multiple crossovers over large distances often go unnoticed. So Haldane presented a more accurate way of calculating map distance, using a mathematical function which later came to be known as a 'mapping function'. Haldane later became one of the architects of population and quantitative genetics along with Ronald Fisher[3] and Sewall Wright[4].

Haldane in his paper [4], quantified the genetic linkage and described that 1% of recombination between two loci can be taken as a unit of map distance, and named it as centimorgan (cM). Haldane found that relationship between recombination frequency and map distances deviate from linearity when the frequencies near 10% and beyond. This is because when the loci are farther apart on a chromosome, the chance of the occurrence of multiple crossovers increases proportionately to the distance. In the event of multiple crossovers, the even number of crossovers (2,4,…) often go unaccounted, resulting in underestimation of map distance. Haldane assumed that crossovers occur randomly and independently over the entire chromosome. If p_k is the probability of k crossovers between two loci, then the recombination fraction r, determined by genetic crosses is the sum of probability of odd number of crossovers,

$$r = p_1 + p_3 + p_5 + p_7 + \cdots \quad (1)$$

Haldane quantified the genetic linkage and described that 1% of recombination between two loci can be taken as a unit of map distance, and named it as centimorgan (cM).

while the map length d is,

$$d = p_1 + 2p_2 + 3p_3 + 4p_4 + \cdots \quad . \quad (2)$$

Haldane showed that p_k follows a Poisson distribution, $p_k = e^{-d}d^k/k!$ and then the map length remains to be d, while the recombination fraction is

$$\begin{aligned} r &= e^{-d}\,(d/1! + d^3/3! + d^5/5! + d^7/7! + \cdots) \\ &= e^{-d} \sinh d \\ &= \tfrac{1}{2}\,(1 - e^{-2d}) \quad , \end{aligned} \quad (3)$$

$$d = -\tfrac{1}{2} \ln (1 - 2r) . \qquad (4)$$

This has become the famous Haldane's mapping function. Haldane's mapping function therefore, adjusts the observed proportion of recombinant gametes for unobserved multiple crossovers so that map distances are additive. When $r < 10\%$, loci are located close together, the amount of double crossovers within such a small interval is negligible, and therefore $r = d$ and additive, an equivalent to Morgan's map distance. Further, when $r = 50\%$, the function estimates the map distance to be infinite indicating that two loci are inherited independently and when two loci if separated by 50 cM distance, there may be an estimated proportion of 32% recombination ($r = 0.32$) between the loci. Haldane concludes his paper stating that "*the theory developed... fits all the observed data in plants*".

Notwithstanding the arguments of Castle–Morgan debate, and the assumptions of Haldane, the additivity of recombination fraction still remained elusive.

3. Interference

Notwithstanding the arguments of Castle–Morgan debate, and the assumptions of Haldane, the additivity of recombination fraction still remained elusive. In their augments, in a way, both Morgan and Castle erred. Morgan erred on the argument of 'complete' additivity, while Castle could not substantiate his claims because he depended mainly on Morgan's data. While presenting the first genetic map Sturtevant wrote that, "*the evidence, so far as it goes, indicated that the occurrence of one*

Box 2. Interference

The existence of the phenomenon of interference is known for almost a century now, but how it is exerted still remains a mystery. In most eukaryotes, during prophase of the meiotic division, recombination events are induced on homologous chromosomes. The phenomenon of interference was found to occur among most of these recombination events – but not all, resulting in widely spaced crossing overs. It has been postulated that some sort of chiasma discouraging signals arise from the region on crossing over which spread in a gradient fashion along the chromosome arm on both directions from the chiasma, but without any conclusive proof. Interference has been found to exert its effects on the entire chromosome. In eukaryotes it can act over mega basepair length of DNA. However, studies on the varying chromosomal lengths in different organisms have shown that interference is not a property of DNA itself.

crossover makes another one less likely to occur in the same gamete", a phenomenon now known as interference.

For the three loci P, Q and R occurring in that order on a chromosome closely placed or linked, the pairwise recombination fraction r, between the loci r_{PQ}, r_{QR} and r_{PR}, expected from two point crosses, should follow the relation,

$$r_{PR} = r_{PQ} + r_{QR} - 2r_{PQ}r_{QR}, \qquad (5)$$

where $r_{PQ}\, r_{QR}$ is the expected double crossover frequency that occurs simultaneously between P and Q and between Q and R. However, deviations from this expected double crossover frequency have been found to occur commonly. This deviation, as suggested by Sturtevant, is a result of crossover suppression between adjacent loci. To incorporate this departure, the equation (5) is to be modified by incorporating a correction term C, as $r_{PR} = r_{PQ} + r_{QR} - 2Cr_{PQ}r_{QR}$, where C has been defined as the coefficient of coincidence, and $(1 - C)$ as interference. Therefore, C is an expectation, equivalent to,

$$C = \frac{r_{12}}{2r_{PQ}\, r_{QR}},$$

where r_{12} is the actual or true double crossover frequency, against which the observed double crossover frequency may not be equal. If both are equal, that is $r_{12} = 2r_{PQ}\, r_{QR}$, then $C = 1$, and $(1 - C) = 0$, there is no interference.

Although, Sturtevant's data suggested that interference differs from locus to locus, Morgan believed that in very close loci, interference was complete.

Although, Sturtevant's data suggested that interference differs from locus to locus, Morgan believed that in very close loci, interference was complete, that is $C = 0$. Thus the term $2Cr_{PQ}r_{QR}$ becomes 0, and $r_{PR} = r_{PQ} + r_{QR}$, the complete additivity. In practice, however, the existence of interference may cause over or underestimation of linkage between adjacent loci over a long chromosome, irrespective of their physical closeness. Further, chromosome themselves have regions varying in recombination events (*Figure* 2). Surprisingly, Haldane was silent on interference; or rather he might have felt it of negligible influence.

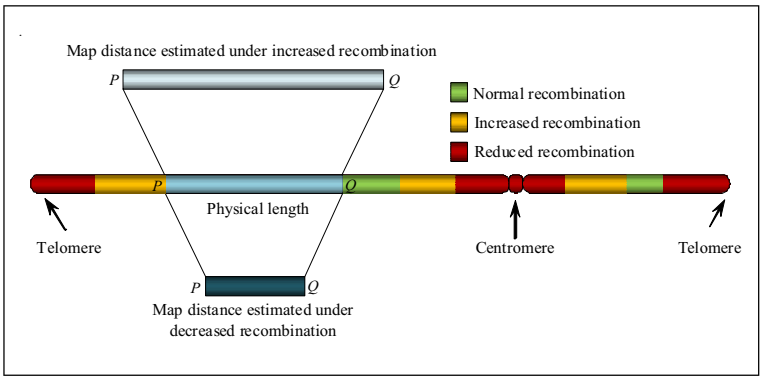
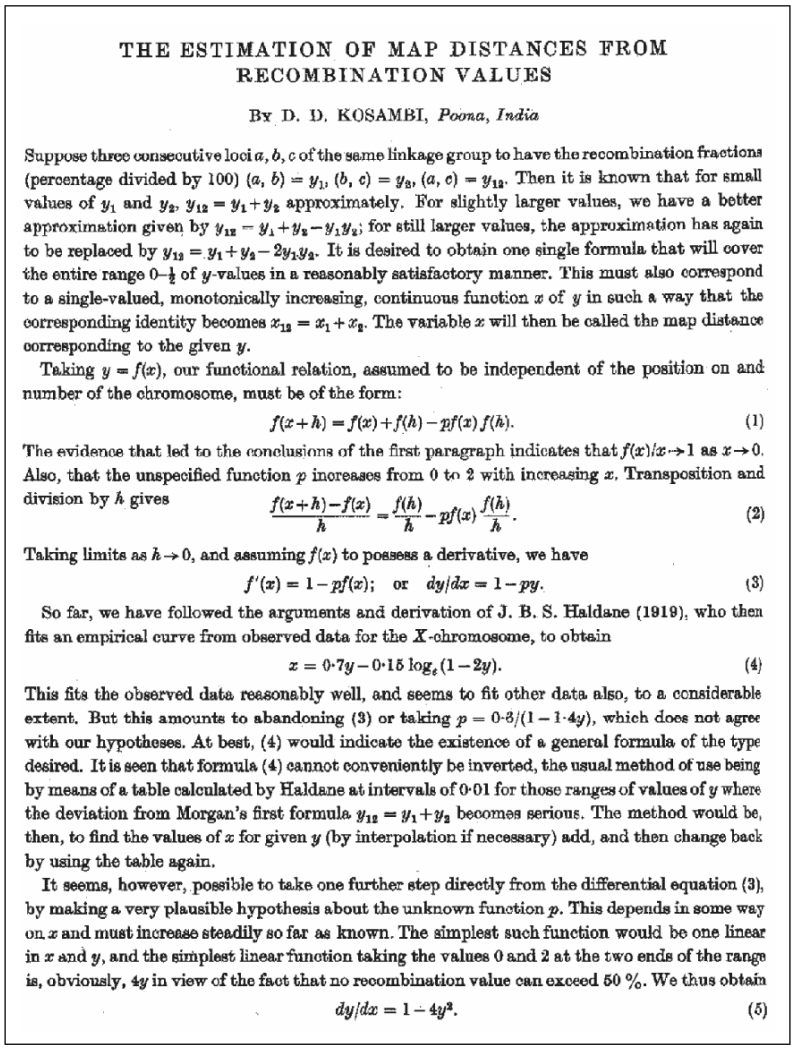


Figure 2. Recombination events vary on different regions of a chromosome. They are typically suppressed near telomeres and centromeres. Estimation of map distances therefore varies depending on the region of chromosome being mapped.

4. Kosambi's Mapping Function

Haldane's mapping function was purely mathematical, and was rather unequivocally accepted; it had found innumerable applications in linkage map development. By then Kosambi, himself a mathematical genius perhaps was keenly observing the developments in quantitative genetics. The period between 1920 and 1944 was flooded with mathematical and statistical applications in biology especially in quantitative genetics by Fisher, Wright and Haldane. The fact that Kosambi followed the works of Haldane and others was clear from what he wrote in his paper published in the third issue of Volume 12 of the journal *Annals of Eugenics* in 1944 (*Figure* 3): "*A comprehensive recasting of available data on map distances is not possible at present, because I have no access to the necessary bibliographic material, and also because a good deal of the data seems to have been estimated by statistically unsatisfactory methods*". Kosambi was addressing again the additivity problem, where he found that interference had caused so much erroneous estimations that *certain loci refer to whole sections of the chromosome*. Kosambi said that the use of Haldane's mapping function under such circumstances was limited. Obviously, his thinking had a lot of influence from Fisher.

Kosambi described that, since the map distance is a function of expected number of crossovers, the following derivative can be obtained from equation (5),



THE ESTIMATION OF MAP DISTANCES FROM RECOMBINATION VALUES

By D. D. KOSAMBI, *Poona, India*

Suppose three consecutive loci a, b, c of the same linkage group to have the recombination fractions (percentage divided by 100) $(a, b) = y_1$, $(b, c) = y_2$, $(a, c) = y_{12}$. Then it is known that for small values of y_1 and y_2, $y_{12} = y_1 + y_2$ approximately. For slightly larger values, we have a better approximation given by $y_{12} = y_1 + y_2 - y_1 y_2$; for still larger values, the approximation has again to be replaced by $y_{12} = y_1 + y_2 - 2y_1 y_2$. It is desired to obtain one single formula that will cover the entire range 0–½ of y-values in a reasonably satisfactory manner. This must also correspond to a single-valued, monotonically increasing, continuous function x of y in such a way that the corresponding identity becomes $x_{12} = x_1 + x_2$. The variable x will then be called the map distance corresponding to the given y.

Taking $y = f(x)$, our functional relation, assumed to be independent of the position on and number of the chromosome, must be of the form:

$$f(x+h) = f(x) + f(h) - pf(x)f(h). \quad (1)$$

The evidence that led to the conclusions of the first paragraph indicates that $f(x)/x \to 1$ as $x \to 0$. Also, that the unspecified function p increases from 0 to 2 with increasing x. Transposition and division by h gives

$$\frac{f(x+h)-f(x)}{h} = \frac{f(h)}{h} - pf(x)\frac{f(h)}{h}. \quad (2)$$

Taking limits as $h \to 0$, and assuming $f(x)$ to possess a derivative, we have

$$f'(x) = 1 - pf(x); \quad \text{or} \quad dy/dx = 1 - py. \quad (3)$$

So far, we have followed the arguments and derivation of J. B. S. Haldane (1919), who then fits an empirical curve from observed data for the X-chromosome, to obtain

$$x = 0{\cdot}7y - 0{\cdot}15 \log_e(1-2y). \quad (4)$$

This fits the observed data reasonably well, and seems to fit other data also, to a considerable extent. But this amounts to abandoning (3) or taking $p = 0{\cdot}3/(1 - 1{\cdot}4y)$, which does not agree with our hypotheses. At best, (4) would indicate the existence of a general formula of the type desired. It is seen that formula (4) cannot conveniently be inverted, the usual method of use being by means of a table calculated by Haldane at intervals of 0·01 for those ranges of values of y where the deviation from Morgan's first formula $y_{12} = y_1 + y_2$ becomes serious. The method would be, then, to find the values of x for given y (by interpolation if necessary) add, and then change back by using the table again.

It seems, however, possible to take one further step directly from the differential equation (3), by making a very plausible hypothesis about the unknown function p. This depends in some way on x and must increase steadily so far as known. The simplest such function would be one linear in x and y, and the simplest linear function taking the values 0 and 2 at the two ends of the range is, obviously, $4y$ in view of the fact that no recombination value can exceed 50 %. We thus obtain

$$dy/dx = 1 - 4y^2. \quad (5)$$

Figure 3. The first page of DD Kosambi's solo paper in genetics published in the journal Annals of Eugenics in 1944.

$$\frac{dy}{dx} = 1 - 2Cr. \quad (6)$$

Kosambi's mapping function adjusts the map distance based on interference which changes the proportion of double crossovers.

Kosambi took the value of $C = 2r$ to obtain

$$\frac{dy}{dx} = 1 - 4r^2, \quad (7)$$

which was integrated to obtain the equation

$$2r = \tanh 2d; \; d = \frac{1}{4}\ln\frac{1+2r}{1-2r}, \quad (8)$$

$$r = \frac{(e^{4d} - 1)}{2(e^{-4d} + 1)}. \qquad (9)$$

Kosambi's mapping function adjusts the map distance based on interference which changes the proportion of double crossovers. What makes Kosambi's mapping function more unique than Haldane's is the rationale behind it. It says that the crossover interference depends on the size of the chromosome segment, and such interferences are absent when the segment size is sufficiently large. Interference increases when the segment size decreases. Kosambi wrote [1] that "*the simplest .... function would be one linear in* x (map distance) *and* y (recombination fraction), *and the simplest linear function taking the values* 0 *and* 2 *at two ends of the range is obviously* 4y (equivalent to 2*C* in the equations above) *in view of the fact that no recombinant value can exceed* 50%."

Kosambi, while explaining his concept surprisingly used the term "markers" although modern genetic markers were unknown during his period. He actually was referring to an unknown or imaginary loci linked to a known gene, the same function modern molecular markers are known for!

5. Mapping Functions – How and When

The mapping functions are therefore mere correction factors of the estimated map distance between two loci. Real recombinations are biological phenomena that are under the influence of biological factors. Therefore in the true sense, applying a universal mapping function to a complex biological phenomenon is less meaningful. This leaves us with three choices of mapping functions, with respect to three levels of interference, viz., complete interference, incomplete interference, and no interference. Complete interference does not allow double crossovers, so we have Morgan's mapping function of true additivity; incomplete interference allow some double crossovers, hence we have Kosambi's mapping function; and for no interference, we have Haldane's mapping function.

It can now be seen that, all the three mapping functions converges to similar estimates of map distance when the two loci in question are close enough.

It can now be seen that, all the three mapping functions converges to similar estimates of map distance when the two loci in question are close enough, so that not more than one crossing over is possible between them. So larger the distance and if independence of crossing over is assumed, one can use Haldane's mapping function, and if interference is to be accounted use of Kosambi's mapping function is advocated. To explain it in another way, if the expected double crossover equals the observed double crossover, use Haldane's; and if any deviation is seen between the expected and observed values of double crossovers, Kosambi's function would be expected to give the best results. After Haldane and Kosambi, many mapping functions have been developed [5], but none has become as popular as the original ones.

From the practical point of view, the use of Kosambi's function is widely practiced, because of its advantage over Haldane's function for interference, even though Kosambi's map distances do not provide exact additivity as that of Haldane's. Nevertheless, many recombination data, gathered over almost a century in a wide range of organisms, roughly exhibit a level of interference nearly corresponding to Kosambi's estimates. Therefore, many of the published genetic maps are based on Kosambi distances. Furthermore, recently it has been demonstrated that unbiased estimates of variation of Kosambi's map distances are lower than that of Haldane's in a multi-point analysis of a real data, suggesting that the Kosambi's distances are more accurate than Haldane's [6]. We must be humble before the genius of a multi-visionary, a great mind the science of genetics was fortunate to have, Damodar Dharmananda Kosambi!

Suggested Reading

[1] D D Kosambi, The estimation of map distance from recombination values, *Annals of Eugenics*, Vol.12, pp.172–175, 1944.

[2] W Bateson, E R Saun-ders and R C Punnet, Experimental studies in the physiology of heredity, *Reports to the Evolution Committee of the Royal Society*, Vol.2, pp.80–99, 1905.

[3] T H Morgan and C B Bridges, *Sex linked inheritance in Drosophila*, Carnegie Institution of Washington. p.88, 1916.

[4] J B S Haldane, The combination of linkage values, and the calculation of distance between linked factors, *Journal of Genetics*, Vol.8, pp.299–309, 1919.

[5] B H Liu, *Statistical Genomics: Linkage, Mapping, and QTL Analysis*, CRC Press, p.611, 1997.

[6] M Huehn, Random variability of map distances based on Kosambi's and Haldane's mapping functions, *J. Appl. Genet.*, Vol.51, pp.27–31, 2010.

Address for Correspondence
K K Vinod
IARI
Rice Breeding and Genetics
Research Centre
Aduthurai 612 101
Tanjavur District, TN, India
Email: kkv_gen@iari.res.in;
kkvinodh@gmail.com