Chapter 10
Genetic Recombination

Mary Sara McPeek

Genetic recombination and genetic linkage are dual phenomena that arise in connection with observations on the joint pattern of inheritance of two or more traits or genetic markers. For example, consider two traits of the sweet pea, *Lathyrus odoratus*, an organism studied in depth by Mendel [9]: flower color, with purple (dominant) and red (recessive) phenotypes, and form of pollen, with long (dominant) and round (recessive) phenotypes. Under the Mendelian model for flower color (recast in more current terminology), each plant carries two alleles for flower color, one inherited from each parent, where each allele can be one of two types, denoted P and p. The pair of alleles carried by a plant is known as its genotype. Plants with genotype PP or Pp have purple flowers, while plants with genotype pp have red flowers. Mendel's First Law can be interpreted as specifying that a parent plant passes on a copy of one of its two alleles to each offspring, with each parental allele having an equal chance of being copied, and with this occurring independently across offspring and across parents. Similarly, each plant carries two alleles for form of pollen, where each of these can be L or l. Plants with genotype LL or Ll have long pollen, while plants with genotype ll have round pollen. Suppose one crossed a true-breeding parental line having purple flowers and long pollen (all individuals having genotype $PPLL$) with a true-breeding parental line having red flowers and round pollen (all individuals having genotype $ppll$). Then the offspring of that cross, known as the F_1 generation, would all have genotype $PpLl$, resulting in purple flowers and long pollen. Suppose a backcross were performed, in which F_1 individuals were crossed with individuals from the $ppll$ parental line. In this example, genetic **linkage** would refer to a tendency for pairs of alleles inherited from the same parent, such as the pair PL or the pair pl, to be transmitted together during meiosis, while genetic **recombination** would refer to the event that an individual transmits a pair of alleles that were inherited from different parents, such as the pair Pl or pL. If we let $0 \leq \theta \leq .5$ denote the **recombination fraction**, which is the probability of a recombination between

M.S. McPeek
Departments of Statistics and Human Genetics, University of Chicago
e-mail: mcpeek@uchicago.edu

S. Dudoit (ed.), *Selected Works of Terry Speed*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4614-1347-9_10, © Springer Science+Business Media, LLC 2012

the genes for these two traits in a single meiosis, then in the backcross offspring, we expect individuals with genotypes $PpLl$, $ppll$, $Ppll$ and $ppLl$ to occur with relative frequencies $(1-\theta)/2$, $(1-\theta)/2$, $\theta/2$ and $\theta/2$, respectively.

A long-standing, important application of the ideas of linkage and recombination is to construction of genetic maps [15] and to subsequent localization of genes (or other genetic variants of interest) on those maps. The key observation is that the recombination fraction between a pair of genetic markers tends to increase with the chromosomal distance between them, with markers on different chromosomes having recombination fraction .5. Thus, by merely observing patterns of joint inheritance of traits, one can make inference about which trait genes lie on the same chromosome, chromosome, and make estimates of distances between them. The basic ideas of and mathematics behind linkage and recombination were developed early in the 20th century [10, 15, 5]. Notably, these problems attracted the interest of R. A. Fisher [3].

Starting in the early 1980's, there was a resurgence of interest in the problem of genetic map construction, spurred by the development of recombinant DNA technology which resulted in the ability to collect genotype data on large numbers of neutral genetic markers throughout the human genome [1] as well as genomes of model organisms. It was not long after these technological breakthroughs occurred that Terry shifted much of his energy and interest into the field of statistical genetics, near the beginning of the explosion of new data and resulting need for new statistical models and methods. In human data, the map construction problem called for more sophisticated statistical analysis than that typically required in experimental organisms. In model organisms, experimental crosses can often be planned in such a way that it is feasible to simply observe the relative frequency of recombinants in any given interval and convert it to a distance using a "map function", an analysis method that we will call the "two-point analysis." However, in humans, crosses cannot be planned, and so any given human meiosis would typically be uninformative for some of the markers of interest. (For example, in the sweet pea example above, all meioses from an individual with genotype $Ppll$ would be uninformative for recombination between these two genes, because the recombinant and non-recombinant allele pairs are indistinguishable.) When many genetic markers are considered simultaneously in each meiosis, and many meioses from different individuals (with different patterns of informativeness) are analyzed together, substantial additional information, beyond that available from a two-point analysis, can typically be obtained by a joint analysis using a suitable statistical model for joint recombination events among a collection of genetic markers.

Thus, the statistical challenges of genetic mapping in humans naturally led to consideration of probability models for the crossover process that causes the observation of recombination. In humans and other diploid eukaryotes, crossing over takes place during a phase of meiosis in which the two parental versions of a given chromosome have each been duplicated, and all four resulting strands or chromatids are lined up together, forming a tight bundle. located along this bundle, with each crossover involving exactly two of the four chromatids. It is assumed that the two chromatids involved in any particular crossover are nonsister chromatids, that is,

the two chromatids cannot be the two identical copies of one of the parent's versions of the chromosome. After crossing over has occurred, the four resulting chromatids are each mosaics of the original parental chromosomes. Keeping in mind this framework, one can consider two key aspects of the model: (1) the distribution of crossover points along the bundle of four chromatids and (2) the choice of nonsister pair of chromatids involved in each crossover. Perhaps the simplest useful model is the no-interference model of Haldane [5], which models aspect (1) by assuming that the crossover points form a homogeneous Poisson process and models aspect (2) by assuming that each nonsister pair is equally likely to be chosen for each crossover, independently across crossovers. **Interference** refers to deviation from Haldane's model. Interference, in the form of local inhibition of crossover points on a resulting single chromatid, was readily apparent in early *Drosophila* data [16, 11]. It is convenient to refer to failure of assumption (1) of Haldane's model as **crossover interference** and failure of assumption (2) of Haldane's model as **chromatid interference**.

Under the assumption of no chromatid interference (NCI), Speed et al. [14] derive a set of constraints, on the multilocus recombination probabilities, that are necessary and sufficient to ensure the existence of a counting process model for the distribution of crossover points along the bundle of four chromatids. They apply these constraints to prove a consistency result for the maximum likelihood estimate of the map order of a finite number of genetic markers along a chromosome. Specifically, they show that, under the assumption of NCI, in the case of complete data, i.e. when all meioses are informative for all markers, if maximum likelihood estimation is performed assuming the Haldane model, then the MLE will converge almost surely to the true map order, even when the true crossover point process is not Poisson (it can be any counting process).

The idea that the assumption of NCI imposes constraints on multilocus recombination probabilities is developed further in Zhao et al. [18], in which the main goal is assessment of the empirical evidence for chromatid interference. This paper extends the constraints from single spore data (such as that from humans and *Drosophila*) to tetrad data (from organisms such *Neurospora crassa*, *Saccharomyces cerevisiae* and *Aspergillus nidulans*) in which data on all 4 chromatid strands are available for each meiosis, providing much more information about strand choice and, hence, allowing a more powerful test of the NCI assumption. An efficient iterative algorithm for maximum likelihood estimation under the constraints is developed, and a likelihood ratio test is proposed to assess whether there is evidence that the constraints are not satisfied by the multinomial model assumed to generate the data. An empirical bootstrap approach is used to assess significance. Some of the experiments did provide evidence for chromatid interference, but overall there was no consistent pattern. The extent and type of chromatid interference seemed to vary across organisms and across experiments. Because the loci considered in these experiments are functional genes, as opposed to neutral markers, it is possible that differential viability may play a role in the results as well. In single-spore data, in particular, the constraints imposed by NCI are rather weak, and the available data do not provide

much power to contradict them. Therefore, it seemed reasonable to assume NCI and focus attention on models for the crossover process.

Because the Haldane no-interference model was so clearly contradicted by most of the available, relevant data, Terry was somewhat concerned about relying on it for map inference. If a more flexible, yet still parsimonious and tractable, model could be developed and shown to fit the data better, Terry reasoned, it could be useful for a range of applications in genetic inference. This problem is addressed by McPeek and Speed [8], in which a range of point process models, involving one or two additional parameters, are fit to *Drosophila* data by maximum likelihood. Goodness of fit of the models is assessed, and the pattern of interference generated by each model is compared to that in data. The most promising model that emerges from this study, the **gamma model**, is a stationary gamma renewal process on four strands, combined with the assumption of NCI to generate a thinned process. In addition to fitting the data better and providing a pattern of interference that mimics that in data, the gamma model is also parsimonious and, when an integer shape parameter is used, results in efficient computational methods. This promising model is further developed in Zhao et al. [19], in which the gamma model with integer shape parameter is referred to as the **chi-square model** because it results in a stationary renewal process having chi-square interarrivals (with even degrees of freedom) for the process on a single strand. The model is fit to datasets from a number of different organisms, with different datasets from the same organisms having similar estimated shape parameter. The results of the analyses suggest that it may be reasonable to use an organism-specific shape parameter to model interference.

In a closely-related line of research, Terry and colleagues sought to connect probability modeling of the crossover process with the initially mysterious-seeming map functions commonly used in two-point analysis. A **map function** is used to convert probability of recombination across an interval to genetic distance of the interval, where **genetic distance** is defined as the expected number of crossovers per strand per meiosis. A difficulty in application of map functions to multilocus analyses is that when there are more than three markers, the multilocus recombination probabilities cannot be uniquely determined from the map function [3]. Earlier work [4, 13, 7] had proposed to solve this identifiability problem by constraining the probability of an odd number of crossovers across a union of disjoint intervals to depend only on the total length of these intervals. However, this is not a biologically plausible assumption, and, as shown by Evans et al. [2], assuming NCI, the class of count-location models [6, 12] is the only class of models having map functions that satisfy this constraint. Zhao and Speed [17] remove this biologically implausible constraint, and instead solve the general problem of developing stationary renewal process models that can generate specific map functions. They show that in most cases of previously-proposed map functions, one can construct a stationary renewal process that generates the map function. Furthermore, they show that this stationary renewal process can typically be approximated quite well by the gamma or chi-square model. The useful practical consequence of this is that two-point analyses using a particular map function can easily be extended to more informative multipoint analyses, an approach that is particularly valuable in the presence of missing data.

References

[1] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.*, 32:314–331, 1980.

[2] S. N. Evans, M. S. McPeek, and T. P. Speed. A characterisation of crossover models that possess map functions. *Theor. Popul. Biol.*, 43:80–90, 1993.

[3] R. A. Fisher. The theory of linkage in polysomic inheritance. *Phil. Trans. Roy. Soc. B*, 233:55–87, 1947.

[4] H. Geiringer. On genetic map functions. *Ann. Math. Statist.*, 142:1369–1377, 1944.

[5] J. B. S. Haldane. The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.*, 8:299–309, 1919.

[6] S. Karlin and U. Liberman. A natural class of multilocus recombination processes and related measures of crossover interference. *Adv. Appl. Prob.*, 11: 479–501, 1979.

[7] U. Liberman and S. Karlin. Theoretical models of genetic map functions. *Theor. Popul. Biol.*, 25:331–346, 1984.

[8] M. S. McPeek and T. P. Speed. Modeling interference in genetic recombination. *Genetics*, 139:1031–1044, 1995.

[9] G. Mendel. Versuche über pflanzenhybriden. *Verh. Naturf. Ver. Brünn*, 4:3–44, 1866.

[10] T. H. Morgan. Random segregation versus coupling in Mendelian inheritance. *Science*, 34:384, 1911.

[11] H. J. Muller. The mechanism of crossing-over. *Am. Nat.*, 50:193–221, 284–305,350–366,421–434, 1916.

[12] N. Risch and K. Lange. An alternative model of recombination and interference. *Ann. Hum. Genet.*, 43:61–70, 1979.

[13] F. W. Schnell. Some general formulations of linkage effects in inbreeding. *Genetics*, 46:947–957, 1961.

[14] T. P. Speed, M. S. McPeek, and S. N. Evans. Robustness of the no-interference model for ordering genetic markers. *Proc. Natl. Acad. Sci. USA*, 89:3103–3106, 1992.

[15] A. H. Sturtevant. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J. Exp. Zool.*, 14:43–59, 1913.

[16] A. H. Sturtevant. The behavior of the chromosomes as studied through linkage. *Zeit. f. ind. Abst. u. Vererb.*, 13:234–287, 1915.

[17] H. Zhao and T. P. Speed. On genetic map functions. *Genetics*, 142:1369–1377, 1996.

[18] H. Zhao, M. S. McPeek, and T. P. Speed. Statistical analysis of chromatid interference. *Genetics*, 139:1057–1065, 1995.

[19] H. Zhao, T. P. Speed, and M. S. McPeek. Statistical analysis of crossover interference using the chi-square model. *Genetics*, 139:1045–1056, 1995.