Studying Recombination with High-Throughput Sequencing: An Educational Primer for Use with "Fine-Scale Heterogeneity in Crossover Rate in the *garnet-scalloped* Region of the *Drosophila melanogaster* X chromosome"

Caiti S. Smukowski Heil and Mohamed A. F. Noor[1]
Biology Department, Duke University, Durham, North Carolina 27708

SUMMARY An article by Singh and colleagues in this issue of *GENETICS* quantifies variation in recombination rate across a small region of the *Drosophila melanogaster* genome, providing an opportunity for instructors of genetics to introduce or reinforce important concepts such as recombination and recombination rate variation, genome sequencing, and sequence features of the genome. Additional background information, a detailed explanation of the methods used in this study, and discussion questions are provided.

Related article in *GENETICS*: Singh, N. D., E. A. Stone, C. F. Aquadro, and A. G. Clark, 2013 Fine-scale heterogeneity in crossover rate in the *garnet-scalloped* region of the *Drosophila melanogaster* X chromosome. Genetics 194: 375–387.

Background

RECOMBINATION (or crossing over) is the equal exchange of genetic material between homologous chromosomes during prophase of meiosis I. The physical link that is established from this exchange helps create the tension necessary for chromosomes to be pulled to opposite poles during anaphase, ensuring that each egg or sperm ends up with the correct number of chromosomes. Failure of this process can result in an egg or sperm having too many or too few chromosomes, which is fatal or severely developmentally debilitating.

Singh *et al.* (2013) explore differences in the frequency of recombination events across part of the *Drosophila melanogaster* genome. As they discuss, researchers have found that recombination events do not happen randomly across the genome in most organisms studied. Instead, some areas of the genome are far more likely to be the sites of recombination events (crossovers) than other parts of the same genome. Humans have narrow (2000 bp) "hotspots" of recombination punctuated by long stretches of zero or near-zero recombination, and there has been a lot of recent excitement about a protein (PRDM9) whose amino acid sequence determines where in the human genome these hotspots occur (Baudat *et al.* 2010). However, these patterns have not been examined with the same precision in most other taxa, and there are many unanswered questions. Are there narrow hotspots of recombination in other taxa—in other words, can one observe very narrow regions (1000, 5000, or 10,000 bp) of the genome bearing much higher rates of recombination than their neighbors? What causes any observed recombination rate variation? In the study we review, the authors show that the model species *D. melanogaster* indeed exhibits recombination rate variation at very fine scales. They also look at associations between recombination rate and sequence features, aimed at understanding what causes recombination rate variation in *Drosophila*. We begin by explaining several of the methods they employ to accomplish this.

Classical Genetics

One of the coolest things about the study by Singh *et al.* (2013) is that the authors use a classical *Drosophila* genetics technique dating back over 100 years. In the early 1900s, Thomas Hunt Morgan, a fly geneticist at Columbia University,

doi: 10.1534/genetics.113.150771
[1]Corresponding author: Biology Department, Duke University, Box 90338, Durham, NC 27708. E-mail: noor@duke.edu

discovered sex chromosome inheritance using several mutations on the X chromosome (Morgan 1910). Like humans, *Drosophila* have two sex chromosomes, where XX is female and XY is male. A recessive mutation on the X chromosome will have a different pattern of inheritance than a mutation on an autosome, because males carrying only one copy of the X chromosome will have a visible phenotype, while females, carrying two copies of the X, will not. While experimenting with transmission of these X chromosome mutations, Morgan crossed a normal-winged, white-eyed fly (+ w) to a short-winged red-eyed fly (s +); in the F_2 he found short-winged white-eyed flies (s w) (Morgan 1911; Sturtevant 1913). This combination is possible only if an "interchange of materials between homologous chromosomes occurs," or what we know today as recombination/crossing over (Sturtevant 1913).

If Morgan's experiment looks familiar to you, that is good! Singh and colleagues did something very similar. They too used two recessive mutations in genes on the X chromosome, in this case *garnet* (dark red eye color) and *scalloped* (scalloped/fringed wing). In the first cross, they mate a wild-type female (+ +, normal eyes and wings) to a *garnet-scalloped* male (*g sd*, dark red eyes and scalloped wings) to get F_1 females that are heterozygous for these mutations (phenotype: normal eyes and wings). The researchers are interested in the meiosis occurring in these females (*Drosophila* males do not have recombination), and they can see the products of meiosis by analyzing their offspring. Remember, males must contribute the Y chromosome to their sons and their X chromosome to their daughters, so sons will always receive their only X chromosome from their mother. If no recombination occurs between her homologous X chromosomes, the son will receive either his mother's *garnet-scalloped* chromosome (*g sd*) or her wild-type chromosome (+ +). If recombination does occur, the sons will receive either *garnet* wild type (*g* +) or wild-type *scalloped* (+ *sd*). Because the authors are interested only in cases where recombination does occur, the researchers screened flies for dark red eyes and normal wings (*g* +) or normal eyes and scalloped wings (+ *sd*) phenotypes.

In the end, they manually screened 92,105 males and recovered 6716 recombinant males (2483 + *sd* and 4233 *g* +). That is a lot of fly sorting! However, they observed something unexpected in the proportions of recombinants. As the authors point out, instead of finding equal proportions of each recombinant class as predicted from normal Mendelian inheritance, they found many more *g* + recombinants than + *sd* recombinants. While there could have been some trouble with manually identifying these mutations, the best explanation seems to be that there is a viability affect with the flies bearing the *scalloped* mutation. For unknown reasons, fewer flies carrying the *scalloped* genotype reach adulthood, skewing the distribution toward *g* + recombinants. Regardless of the proportions, the researchers can use the number of recombinants to measure the amount of recombination going on in this interval by dividing the total number of recombinants by the total number of individuals surveyed (6716/92105 = 0.073). This gives us the percentage recombination 7.3%, or 7.3 cM (centiMorgans, named after Thomas Hunt Morgan).

Pinpointing the Crossovers

The authors inferred that one single crossover happened between the *g* and *sd* loci to produce each of the 6716 recombinant males. But the authors were not (as) interested in the overall recombination fraction between these two mutations, which are >2 million bases apart. Instead, they wanted to examine *where* in the 2 million bases the crossovers occurred. They ask: Do the recombination events happen in very localized hotspots as they do in mammals? And is there variation in recombination rate when one looks across this 2 million-base span at only 5000 bases at a time, such that some spans of 5000 bp are more likely to have crossovers than others?

To answer these questions and pinpoint where the crossovers occurred, the authors needed a lot more markers. In between *g* and *sd*, they developed 451 markers to get an average distance of <5000 bp between markers. The most commonly used molecular genetic markers today are single nucleotide polymorphisms (SNPs)—nucleotide positions in the genome at which individuals in a population vary (*e.g.*, one individual may have a "T" while another has a "G"). With 6716 recombinant flies to study, and 451 markers, the assays would require >3 million genotypes if each fly was studied for each marker separately! While technologies exist to genotype many markers in single individuals (*e.g.*, SNP genotyping chips used by direct-to-consumer companies such as 23andMe), it is more technically challenging to genotype a moderate number of markers in a very large number of individuals.

The authors employed a combination of creative approaches to alleviate this problem. The first approach leveraged shotgun sequencing. Shotgun sequencing became popular in the mid-1990s with the combination of lower-priced automated DNA sequencing and better computational tools. This approach reenergized the original human genome project and led to its timely completion in 2001 (Lander *et al.* 2001; Venter *et al.* 2001). It is now considered the standard for such procedures.

Shotgun sequencing involves randomly breaking the genome into small chunks and sequencing these multiple, overlapping pieces many times. For our fictional example (see Figure 1A), eight sequencing "reads" (individual outputs from the automated sequencer), where each read is 6 nucleotides (nt) long, would provide 48 nt of sequence: bigger than the total sequence in the figure (its total "genome"). However, the coverage is random, so rather than generating 8 reads, your first step is to generate 80 reads, hence providing >10 times as much sequence as needed (called "10× coverage" of this mini-genome), but increasing the probability that the whole genome is covered by at least one read. The second step is to align the reads computationally

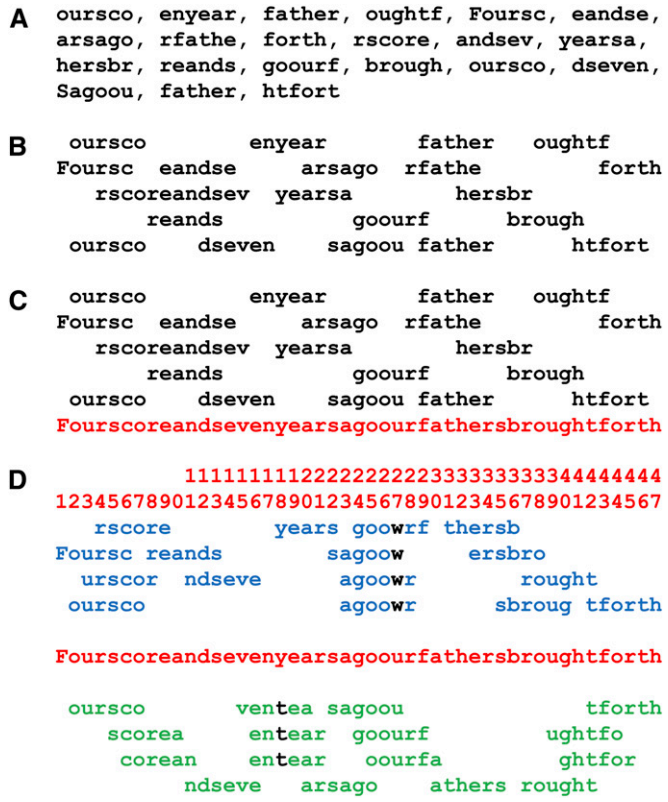

Figure 1 Illustration of principle of shotgun sequencing, assembly, and alignment. In this fictional example, individual sequence reads are drawn at random and are six letters in length. (A) The raw output of individual sequence reads. (B) How one computationally seeks overlapping motifs (stretches of sequence) among the individual reads, hoping to eventually cover the full message (genome). (C) How the aligned reads in B can be put together to produce the full message (a new assembled "reference genome"). (D) A step further: the full "reference" genome sequence is available in red, and positions in it are numbered at the top. The blue sequence reads are from a second individual or set of clones and are aligned to the existing full genome sequence. The green sequence reads are from a third individual or set of clones and are also aligned to the existing full genome sequence. We see that the blue reads differ from the red reference genome (and green) by a SNP in position 27. The green sequence reads differ from the red reference genome (and blue) by a SNP in position 18. Thus, the green and blue sequences differ by two SNPs, and these SNPs can be used as genetic markers.

to look for overlap in a subset of the nucleotides (Figure 1B). Finally, with that alignment, one assembles a "consensus" sequence, which presumably matches the complete original DNA sequence that was broken up (Figure 1C).

The project by Singh *et al.* (2013) has one facet that is simpler than the example above and two facets that are more challenging. The simpler facet is that they already have an assembled *D. melanogaster* sequence with which to work (Adams *et al.* 2000). When they obtain their reads, they merely have to match them to the existing sequence and not worry about assembly (Figure 1D). One difficult facet is that the reads may not match perfectly—there is variation in DNA sequence among individual fly isolates (strains), and the strains they are sequencing are not perfect matches at all nucleotide sites to each other or to the originally published *D. melanogaster* genome sequence (sites 18 and 27 in Figure 1D). However, the differences should only be at 1–2% of nucleotide sites, so this does not provide a great challenge—indeed, it is precisely those sites that differ between the two strains in their cross that are used as the SNP markers. The other challenge is that they are working with a specific 2.1 million-base region and do not want to generate sequence for the full 150 million bases of the *Drosophila* genome. To this end, they used a "hybridization array" that made it so that almost the only DNA sequences they examined were from the relevant 2.1 million-base region of the X chromosome, rather than everywhere in the genome. Basically, using probes based on the published sequence from the X chromosome region of interest, they selectively capture the small pieces of DNA that are most similar to this region and discard the rest, thus saving money and time on unnecessary sequencing and alignment. This process is often called a "target-enrichment" strategy.

Thus far, the authors have generated sequences from their two lines, and they have identified nucleotide differences (SNP markers) to be used for mapping. However, they still must identify the genotype for all 6716 recombinant progeny at the 451 markers. To accomplish this, they use a modification of the target-enriched shotgun-sequencing approach on pools of DNA from 100 different test-cross males each bearing the same recombinant genotype (either + *sd* or *g* +). They literally mixed DNA samples from 100 different flies each having the same recombinant genotype into a common tube (the "pool"), repeated to make 25 pools in all, and then used genotypes for SNP markers from the 25 pools to estimate recombination rate.

How does this work? Let us look at the + *sd* group of the recombinant flies. The *garnet* gene is on the X chromosome at position 13,630,000. The *scalloped* gene is on the X chromosome at position 15,700,000. If a male offspring from a test-cross has the wild-type allele at the *garnet* gene, it must have come from the wild-type grandparent (called UgX54A, blue in Figure 2). All (100%) of male testcross + *sd* flies will have the wild-type allele from that grandparent at the *garnet* gene (vertical bar labeled 1), since they clearly have the wild-type eye-color phenotype. In contrast, none of the + *sd* flies will have the wild-type allele from the UgX54A grandparent at the *scalloped* gene (vertical bar labeled 4), since they clearly exhibit the mutant wing-shape phenotype. For markers in between these two genes (vertical bars 2 and 3), a subset of flies may have the UgX54A allele. Imagine that exactly the same number of flies had the UgX54A allele at markers 2 and 3—that would imply that there was never a crossover event between markers 2 and 3. In contrast, a big difference between these two markers in fractions bearing the *g*–*sd* grandparent's allele (Figure 2 has one-fifth red for marker 2 *vs.* four-fifths red for marker 3) suggests that a lot of recombination occurs between these two markers.

The authors obtained sequences from the pools of 100 flies and examined how often SNP alleles from the UgX54A

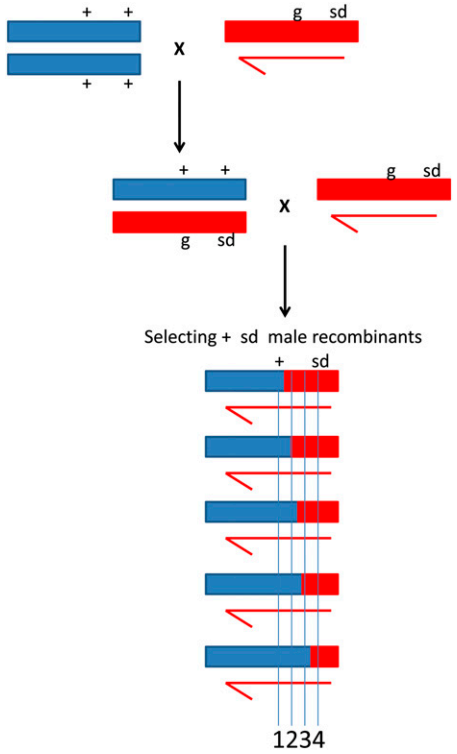

Figure 2 Revised version of Singh *et al.*'s Figure 1 to show the cross they performed. The particular example shows just the + *sd* recombinant males that they selected. The four vertical lines indicate positions of genetic markers on the X chromosome. The reader should note that all recombinant individuals have the allele derived from the wild-type (blue) parent at marker 1 (near the *garnet* gene), while all recombinant individuals have the allele derived from the mutant (red) parent at marker 4 (near *sd*).

grandparent *vs.* the *g sd* grandparent were observed and used the "allele frequency" approach described above to infer how much recombination occurred between each pair of adjacent markers. Moving from left to right across Figure 3b in Singh *et al.* (2013) shows that more of the offspring had the allele derived from the *g–sd* grandparent's allele, analogous to the transition from 1/5 at marker 2 to 4/5 at marker 3 in our fictitious example discussed above and in Figure 2. This procedure was done separately for the + *sd* recombinants and the *g* + recombinants, as depicted in Singh *et al.*'s Figure 3b *vs.* Figure 3a (Singh *et al.* 2013), respectively, and the patterns are opposite of each other, as expected. Fortunately, estimates from these two separate sets were highly correlated in how much recombination (in units of percentage recombination per million bases, cM/Mb) occurred (Singh *et al.* 2013, Figure 3c, blue lines *vs.* red lines). This is a good indication that the authors correctly inferred recombination events in all of their flies.

Investigating Variation in Crossover Frequency

One of the overarching questions for researchers studying recombination is determining how recombination events are distributed. Previous work in *Drosophila* has shown that some genomic regions experience much more recombination than others. However, until now, no one has looked at very small regions to determine whether a similar pattern of recombination variation exists. Is recombination still heterogeneous at very fine scales? Singh *et al.* (2013) answer with a definitive yes! Indeed, it appears that recombination is significantly heterogeneous at the 5-kb scale: rates vary over 90-fold (for example, in some 5-kb intervals, the recombination rate might be as low as 0.3 cM/Mb, while in other intervals it may be as high as 27 cM/Mb). The authors detect variation at every scale they investigate (10 kb, 72×; 20 kb, 41×; 50 kb, 8×; 100 kb, 3x), revealing an unprecedented view of fine-scale recombination variation in *Drosophila*.

Understanding fine-scale variation informs us how recombination works in *Drosophila*, how it is similar to and different than other organisms, and potentially how recombination sites are selected. By looking at the correlation of DNA sequence features with recombination rate variation at different scales, researchers can interpret how recombination influences the distribution of sequence features and in turn, how sequence features influence the distribution of recombination. To this end, Singh *et al.* (2013) analyze associations between recombination rate and sequence motifs, codon bias, gene density, repeat content, GC content, diversity, and divergence.

Conclusions

Singh and colleagues have leveraged classical genetic techniques combined with modern sequencing technology to reveal fine-scale recombination rate variation, previously unknown in *Drosophila*. Both the methods used in this study and its results are based on concepts important to a genetics student. At the most basic level, an understanding of the process of meiosis and recombination is a foundation of most genetics classes. Identifying parental and recombinant classes is a technique many students will be familiar with through the application of three-point mapping to identify the order of genes on a chromosome. Familiarity with shotgun sequencing and such technologies is necessary in this genomics age, and the combination of recombination (or linkage disequilibrium more specifically) and genotyping/sequencing for use in mapping studies like genome-wide association studies is of great interest medically. Finally, their observations of fine-scale variation in recombination rate add of a level of complexity and distinctiveness to a typical lecture on recombination. Although there are strong selection pressures on recombination as an essential process for chromosome segregation during meiosis, many researchers see variation at different scales. Variation in recombination rate, its causes, and evolutionary consequences make a good addition to any genetics and/or evolution course.

Questions/Activities/Discussion Topics

1. Imagine that another laboratory did the same cross and came up with this combination of progeny: 360 + *sd*; 370 *g* +; 4590 *g sd*; 4680 + +. What is the recombination fraction in this example?
2. Why couldn't the authors use more visible marker mutations to pinpoint the crossovers within the 2 million base region of interest?
3. Why didn't the authors sequence all 92,105 testcross progeny to study recombination rate in this region? What would they have gained by doing all of them rather than the 6716 recombinants?
4. Given that the authors studied 6716 recombinant offspring using markers in 450 windows (spanned by 451 markers), *on average*, how many recombination events were observed per window?
5. The authors estimated the recombination rate for each window with an average sample size of *X* (*X* representing the answer to the problem above). There is necessarily some imprecision in these estimates, given the finite sample size. By what fraction would a recombination estimate for a given window change with the addition or subtraction of two recombinant offspring to that window? Why is this important to note?
6. If the authors had sequence data from the whole genome (*i.e.*, if they had not done an enrichment) from the same set of testcross progeny, would they have been able to examine recombination rates outside the 2.1-Mb region between these markers? Why would these other estimates have been less precise than the ones they obtained within the focal 2.1-Mb region?
7. What would have been different if the second-generation cross in Figure 2 used wild-type males rather than *g sd* males?
8. What was the advantage of using the combined "pools" of DNA and estimating SNP frequencies rather than individually genotyping each fly? Are there disadvantages? Why do you think the authors used 100 flies per pool (rather than 10 or 1000)?
9. A "sequence motif" is a particular nucleotide sequence found in many places in the genome, such as ATGGAAA. Singh *et al.* (2013) looked for whether particular sequence motifs were more common in 10,000-bp windows having very high rates of crossing over relative to 10,000-bp windows having very low rates of crossing over. What did they find? What would a difference in the abundance of a motif like ATGGAAA between such high- and low-recombination windows suggest? Why is this of interest both to researchers studying recombination in *Drosophila* and researchers studying recombination in other organisms?
10. Singh *et al.* (2013) report that many previous studies observed a relationship between the average fraction of bases in a window differing between two individuals of the same species (nucleotide polymorphism) and recombination rate. Basically, researchers find that regions of the genome that experience a lot of crossing over have, for example, 4% of nucleotide sites differing in sequence between individuals, whereas regions of the genome that experience almost no recombination are typically identical in sequence between individuals. What does this observation suggest? Did Singh *et al.* (2013) find the same result? Why or why not?

Literature Cited

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober *et al.*, 2010 PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836–840.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Morgan, T. H., 1910 Sex limited inheritance in Drosophila. Science 32: 120–122.

Morgan, T. H., 1911 The application of the conception of pure lines to sex-limited inheritance and to sexual dimorphism. Am. Nat. 45: 65–78.

Singh, N. D., E. A. Stone, C. F. Aquadro, and A. G. Clark, 2013 Fine-scale heterogeneity in crossover rate in the garnet-scalloped region of the Drosophila melanogaster X chromosome. Genetics 194: 375–387.

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

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, 2001 The sequence of the human genome. Science 291: 1304–1351.

Communicating editor: Elizabeth A. De Stasio