Bias–Variance Tradeoffs in Recombination Rate Estimation

Eric A. Stone[1] and Nadia D. Singh[1]
Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina 27612

ABSTRACT In 2013, we and coauthors published a paper characterizing rates of recombination within the 2.1-megabase *garnet-scalloped* (*g-sd*) region of the *Drosophila melanogaster X* chromosome. To extract the signal of recombination in our high-throughput sequence data, we adopted a nonparametric smoothing procedure, reducing variance at the cost of biasing individual recombination rates. In doing so, we sacrificed accuracy to gain precision—precision that allowed us to detect recombination rate heterogeneity. Negotiating the bias–variance tradeoff enabled us to resolve significant variation in the frequency of crossing over across the *garnet-scalloped* region.

IN 2013, we published a paper characterizing rates of recombination within the 2.1 megabase *garnet-scalloped* (*g-sd*) region of the *Drosophila melanogaster X* chromosome. We identified male progeny inheriting crossovers within the region and pooled them into groups for DNA sequencing to recover allelic proportions. These proportions were used to estimate rates of recombination under the logic that the allele frequency of a SNP should equal the proportion of males that recombined upstream of that SNP.

Gilliland (2015) criticized our approach to estimating rates of recombination. In brief, our approach was to select a subset of high-quality SNPs for recording empirical allele frequencies from the sequence data and then to infer the proportion of recombinants upstream of each SNP as the median across a symmetric window of flanking empirical frequencies. Any two SNPs define a genomic interval, and the frequency of recombination within that interval can be computed as the difference in allele frequencies between them. This procedure leads to biased estimates because the selected SNPs are not uniformly spaced. The bias is a by-product of our strategy for variance reduction—a strategy that proved successful in resolving heterogeneity in recombination rates.

Why adopt a biased estimation strategy? It is notoriously challenging to estimate allele frequencies from high-throughput sequencing read counts. This is obviously true at low coverage due to binomial sampling variation, but it is also true at higher coverage due to additional sources of variation. In our study, this challenge was apparent in plots of sample allele frequency against genomic position: while the experimental design forces the relationship between allele frequency and positon to be monotonic, in sample data, this was visible only at coarse scales. Our approach was to smooth the scatterplot—a canonical application of the bias–variance tradeoff. In doing so, we aimed to distinguish coarse patterns in the data at the expense of resolving fine details.

Figure 1 in Gilliland (2015) emphasizes the bias concomitant with our tradeoff. The smoothed allele frequency estimates at successive SNPs are medians of windows that almost completely overlap; where those windows differ, and how the medians change, is nearly independent of interval defined by the focal SNPs. As a consequence, recombination rate estimates (*y*-axis) between successive SNPs scale inversely with the physical distance between them. However, these "estimates" are not worth considering; irrespective of the analytical approach taken for this study, one should not expect to infer recombination rates between successive SNPs. Importantly, the trend accentuated by Gilliland (2015) is absent at more reasonable inter-SNP distances. Figure 1, which recapitulates figure 1 in Gilliland (2015) but includes data for all SNP pairs, shows that dependence on physical distance attenuates at 1 kb, and by 10 kb has all but vanished. This is not to say that our reported rates of recombination are free of bias (see below),

doi: 10.1534/genetics.115.185561
[1]Corresponding authors: Department of Biological Sciences, North Carolina State University, 112 Derieux Place, 3510 Thomas Hall, Raleigh, NC 27695; E-mail: eastone2@ncsu.edu; ndsingh@ncsu.edu

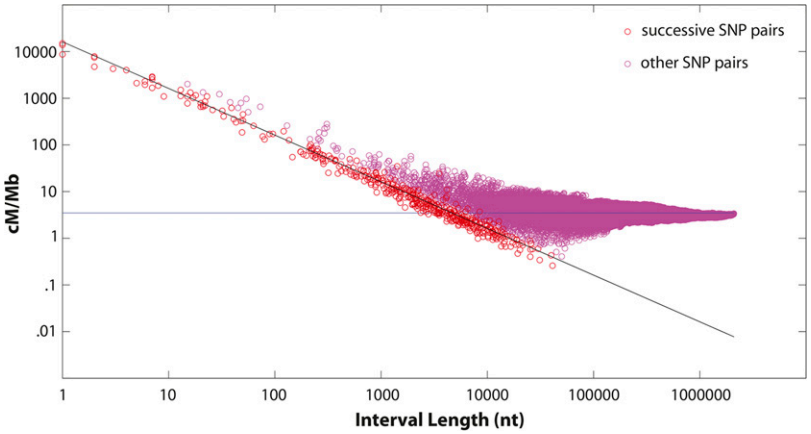

Figure 1 Plot of estimated recombination rate between SNP pairs (cM/Mb) as a function of distance between SNPs in nucleotides. Following Gilliland (2015), red circles (n = 451) denote the values derived from successive SNPs, as compared to a linear trend (*i.e.*, rate scales with distance) shown in black and a constant trend (*i.e.*, rate independent of distance) shown in blue. Included here as purple circles are the values derived from all other SNP pairs (n = 101,024).

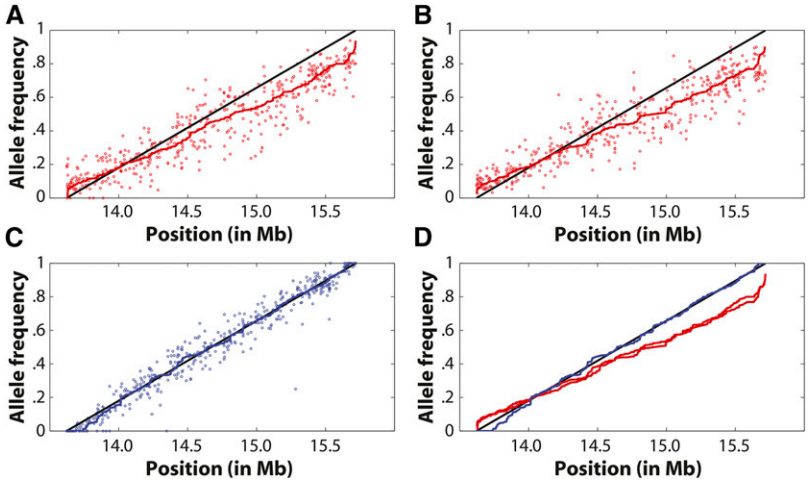

Figure 2 (a) Plot of allele frequency against genomic position in one pool of recombinants. Red circles denote empirical allele frequencies computed directly from read counts. The red line indicates the result of median smoothing. The black line is included for reference and shows a linear increase in allele frequency with genomic position. (b) Plot of allele frequency against genomic position in a second, independent pool of recombinants. (c) Plot of allele frequency against genomic position in a simulated dataset, assuming a constant rate of recombination across the *garnet-scalloped* interval. Blue circles denote empirical allele frequencies computed directly from simulated read counts. The blue line indicates the result of median smoothing. The black line is as in previous panels, but here represents the theoretical allele frequencies as well. (d) Overlay of median smoothing results from panels a–c.

but emphasis on the red circles in Figure 1 engenders a narrative that is misleadingly provocative.

The upside of bias is reproducibility; what is lost in accuracy may be compensated by gains in precision. This is apparent in Figure 2, in which our approach has been applied to both real and simulated data. Each of panels a and b considers an independent pool of recombinants from Singh *et al.* (2013). Empirical allele frequencies, calculated as the fraction of supporting reads, are shown as red circles; both recombinant pools show the same noisy, not-quite-linear relationship between allele frequency and genomic position. In each case, the trend has been captured by our median smoothing approach as indicated by the red line. Panel c, by contrast, considers the null case in which read data are simulated from a simulated pool of recombinants, assuming the rate of recombination to be constant. As highlighted in panel d, it is apparent that results from a and b are more similar to each other than either is to that of c. The nonuniform distribution of SNPs, when coupled with our median smoothing, inflates the degree of similarity observed in panels a–c. That notwithstanding, the pattern in the real data is unequivocal, and the nonuniformity of recombination rates is apparent.

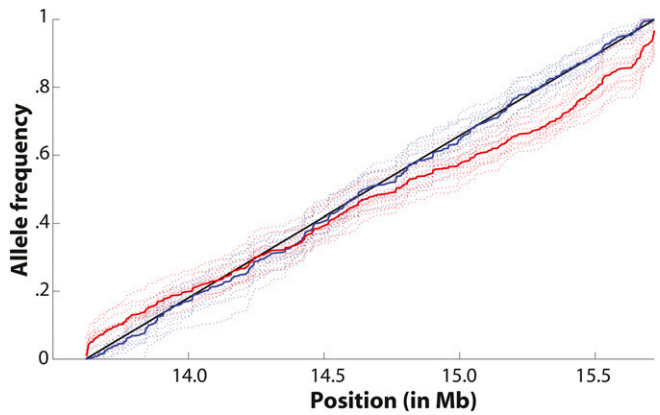

Figure 3 Median smoothed results for 13 recombinant pools (red dashed lines = "+" direction CDFs) in comparison to analogous results for 13 similar pools simulated under a uniform rate of recombination (blue dashed lines). Solid lines denote the aggregated results obtained by averaging across pools.

Our median smoothing approach uses the data to estimate an empirical cumulative distribution function (CDF) (Figure 2). That is to say, the smoothed value at a genomic position estimates the probability that the position lies downstream of a recombination event between *garnet*

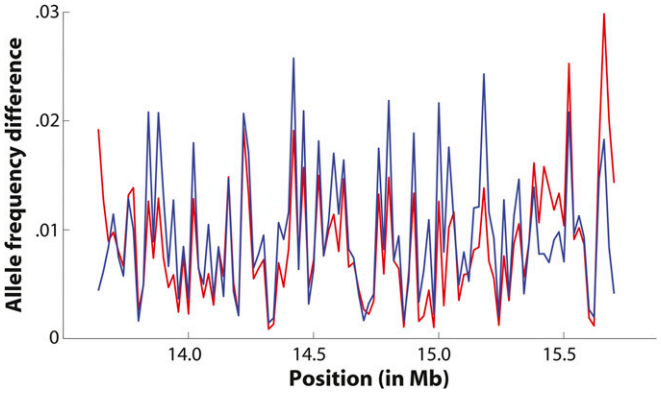

Figure 4 Empirical probability density functions (PDFs) calculated in 20-kb windows from the aggregated results in Figure 3. The difference in allele frequency across the window is plotted against the genomic position of the window center. Despite the sizable and significant difference in empirical CDFs from Figure 3, their corresponding empirical PDFs appear strongly correlated. This granularity magnifies the bias inherent to our median smoothing approach.

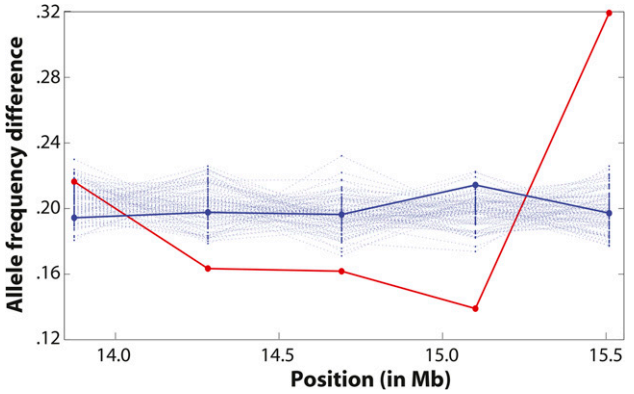

Figure 5 Substantial rate heterogeneity at modest granularity. The *garnet-scalloped* region is partitioned into five windows of equal size. As in Figure 4, the difference in allele frequency across each window is plotted against the genomic position of the window center. Results for the observed data are shown in solid red; results for the simulated data from the previous two figures in shown in solid blue. The dashed blue lines are the results of 100 additional simulations from the null case.

and *scalloped*. A uniform rate of recombination should therefore generate a linear CDF, but this is not what we observe. Rather, the empirical CDFs constructed from simulated data (*e.g.*, the blue line in Figure 2d) too often deviate from the uniform expectation (*i.e.*, the black line in Figure 2d). Nevertheless, this deviation is minimal compared to the deviation observed in real data (*e.g.*, the red lines in Figure 2d). Kolomogorov–Smirnov statistics are not needed to quantify what is so qualitatively obvious: the landscape of recombination events in our data is decidedly nonuniform (Figure 3).

The bias–variance tradeoff, of course, depends on the window size within which crossover density is reported. Finer partitioning of the *g-sd* interval results in greater apparent heterogeneity, but this is, in part, due to a second bias–variance tradeoff: though the variance in recombination rate surely increases at finer scales, so too does the contribution of bias (Figure 4). With coarse partitioning of the *g-sd* interval, the contribution of bias is minimal (Figure 5). At more intermediate, arguably more reasonable scales, the contribution of bias is also intermediate but we nevertheless find that recombination rate heterogeneity persists.

We appreciate Gilliland's attention to our paper and the opportunity to elaborate on our results and rationale. We agree that it is important to be cognizant of limits to resolution, and readers should be aware that the contribution of bias increases with increased granularity. However, as we have explained, his equivalence between bias and artifact is false. Sometimes bias is the key to pushing the limit of resolution, and such was the case in our study. Indeed, it was our use of a biased estimator that empowered us to resolve significant variation in the frequency of crossing over across the *garnet-scalloped* region.

Literature Cited

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.

Gilliland, W. D., 2015 A comment on fine-scale heterogeneity in crossover rate in the garnet-scalloped region of the *Drosophila melanogaster* X chromosome. Genetics 201: 1275–1277.

Communicating editor: M. Johnston