The Summer Institute in Statistical Genetics

Bruce S. Weir[1]
Department of Biostatistics, University of Washington, Seattle, Washington 98195
ORCID ID: 0000-0002-4883-1247 (B.S.W.)

The Elizabeth W. Jones Award for Excellence in Education recognizes an individual or group that has had significant, sustained impact on genetics education at any level, from K-12 through graduate school and beyond. Bruce Weir (University of Washington) is the 2019 recipient in recognition of his work training thousands of researchers in the rigorous use of statistical analysis methods for genetic and genomic data. His contributions fall into three categories: the acclaimed Summer Institute in Statistical Genetics, which has been held continuously for 23 years and has trained > 10,000 researchers worldwide; the popular graduate-level textbook Genetic Data Analysis; and the training of a growing number of forensic geneticists during the rise of DNA evidence in courts around the world.

The 2019 Elizabeth W. Jones Award for Excellence in Education allows me this opportunity to reflect on the symbiotic relationship between statistics and genetics, as well as to recall my pleasant interactions with Beth Jones on the National Institutes of Health Genetics Study Section and on the *GENETICS* Editorial Board. The award recognizes the success of the Summer Institute in Statistical Genetics (SISG), now in its 24th year, and here I will focus on a couple of themes running through SISG planning.

The SISG provides short courses in statistical genetics, primarily for graduate students in the biological sciences. It presents modern statistical methods for current genetic data, with an emphasis on adding to understanding of the biological processes leading to those data. The impetus for SISG came from the explosive growth we have seen in genetic data. As a newly minted Ph.D., I joined the laboratory of Robert W. Allard who was collecting genetic marker data at a new scale for plant populations. My year in Davis led to a pair of papers in this journal (Allard *et al.* 1972; Weir *et al.* 1972), describing our findings for esterase allozyme data in an experimental population of barley, where we proudly showed our analyses for 30,000 plants scored at four loci. Fifty years after my postdoc I am associated with the project of the National Heart, Lung, and Blood Institute (www.nhlbiwgs.org), where there are now 140,000 whole-genome sequences for individuals in > 80 populations and over 1 billion variants have been revealed.

Fortunately, the growth in data has been accompanied by growth in computing power and in statistical software. We did not need to provide the punch-card facilities I used as a postdoc for SISG participants when we started in 1996, but we did need to set up computing laboratories. Now participants bring their own laptops with preloaded software, and we provide some of them with cloud access. As recently as 2015, the SISG offered three courses in R, at different levels of complexity, and these were fully subscribed, but for 2019 we had trouble filling a single course. Students across the biological sciences now come to the SISG with R skills, and all our courses use R scripts and packages as a routine matter. The drop in numbers for R courses has been offset by a rise in numbers for statistical mixed model instruction and steady numbers in the Markov chain Monte Carlo course.

The SISG focuses on statistical genetics, and progress in this field may be viewed through the lens of testing for Hardy–Weinberg equilibrium (HWE). This foundational result featured in the first issue of *GENETICS* (Jennings 1916) as the basis for predicting genotype proportions in populations under various mating schemes, and by 1948 testing for consistency with HWE was being used to infer the action of selection in natural populations (Dobzhansky and Levene 1948). Statistical methods for detecting selection, including HWE testing, became of great interest during the debate over the neutral theory of evolution, and HWE testing features in the continuing "validation" of forensic frequency databases

doi: https://doi.org/10.1534/genetics.119.302506
[1] Corresponding author: Department of Biostatistics, University of Washington, Campus Box 359461, Seattle, WA 98195-9461. E-mail: bsweir@uw.edu

(*e.g.*, Ghiani *et al.* 2019) following some analyses surrounding the first US court case where DNA evidence was successfully challenged (Lander 1989). This forensic activity could be regarded as one instance of the use of HWE testing for data cleaning, most often seen in genome-wide association studies (GWAS) (Gogarten *et al.* 2012).

Dobzhansky and Levene (1948) gave the standard goodness-of-fit test statistic that has a chi square distribution when HWE holds, and it previewed the use of the binomial distribution by Levene (1949). Levene attached normal distribution approximations to the exact *P*-values allowed by the binomial, and it was not until later that computational resources allowed complete enumeration of all possible sets of genotypes for a given set of allele counts (Vithayasai 1973) or Monte Carlo methods for loci with multiple alleles (Guo and Thompson 1992). It is now common to conduct HWE testing with publicly available software, such as HardyWeinberg by Graffelman (2015). The ease of applying these tests to data with millions of genetic variants has meant that the limitations of frequentist hypothesis tests are exposed, not least among these limitations being the perceived need to adopt the *ad hoc* "genome-wide significance level" of 5×10^{-8} as the *P*-value below which a SNP is declared not to be in HWE and may thus not be used in further analyses. It is unlikely that a fixed threshold is appropriate for all sample sizes, numbers of tests, and variant frequencies (Fadista *et al.* 2016), although similar comments hold for the traditional 5% threshold for single tests. My forensic science colleague Ian Evett (personal communication) asks why population geneticists acknowledge that HWE does not hold in natural populations, proceed to test for HWE, and then fail to reject HWE when $P > 0.05$. I make the standard response about the convenience of replacing genotype proportions by products of allele proportions when these two quantities are similar, but I see the attraction of Bayesian alternatives to HWE testing (Wakefield 2010) and there is a Bayesian option in HardyWeinberg. Frequentist and Bayesian methods are both covered in the SISG.

A variant that occurs only once or twice in a sample, as happens for over one-half of the variants revealed by whole-genome sequencing, offers very little chance of detecting significant population-level departures from HWE. The same problems of low power arise in GWAS, where the remedy has been to summarize rare variant information within a gene or a region by a single genetic score, or genetic burden, before performing an association test (*e.g.*, Chen and Wang 2019). Another method for aggregating single-variant test statistics is to employ variance component methods (*e.g.*, Dutta *et al.* 2019). Two SISG modules address these approaches.

HWE testing is applied to single loci, but many biological questions benefit from analyses of genomic regions. In our current work, following on from Weir and Goudet (2017), we have been struck by how well inbreeding levels for individuals are estimated from quite naive implementation of methods using runs of homozygosity when compared to a range of single-SNP-based likelihood and moment estimators. Single-SNP-based methods for estimating distant kinship for pairs of individuals cannot compete with methods based on identity-by-descent (IBD) regions as reviewed by Browning and Browning (2012). We have all been impressed by the recent emergence of forensic methods that identify distant relatives of the source of an evidential profile by the degree of inferred IBD sharing with profiles in a genealogy database, following procedures such as those described by Henn *et al.* (2012) and Erlich *et al.* (2018). The SISG covers kinship estimation, the use of kinship in GWAS, and forensic applications of kinship.

SISG courses present statistical methods that recognize current genetic understanding. Statistical models are described with parameters that have genetic interpretations, such as allelic contributions to a quantitative trait, the probability that an individual with a particular genotype contributes to the next generation, or the probability that an individual carries IBD alleles at a locus. Now that we are well into the "small-large *p*" era, where the number (p) of parameters is far greater than the number (n) of observations, there is movement to the use of machine-learning approaches. Kinship estimation, for example, could be addressed by regressing kinship values on SNP genotypes for many pairs of individuals in a training set with known relationship and then evaluating the predictor on an independent set of pairs of individuals. There would be no appeal to an IBD model to identify a function of SNP genotypes that allowed kinship prediction. Two recent reviews of progress in this direction have been given by Li *et al.* (2018) for economic traits in cattle and by Ho *et al.* (2019) for precision medicine. The SISG spun off a sister institute, the Summer Institute in Statistics for Big Data, in 2015, where machine-learning topics are covered.

The SISG has a large alumni group and now plans to extend its service to that group, and to others with an interest in the field, by hosting a Virtual Institute in Statistical Genetics. This will preserve online versions of course notes and it will provide a forum for exchanging new methods, posting queries and their solutions, and providing a general service to an important community. Details will appear on the SISG website (www.biostat.suminst/sisg).

Acknowledgments

This work was supported in part by award GM-075091 from the National Institutes of Health (NIH). The Summer Institute in Statistical Genetics has been supported by NIH award GM-059733, and by National Science Foundation awards DEB-9900388, DEB-0352235, and DEB-1049804.

Literature Cited

Allard, R. W., A. L. Kahler, and B. S. Weir, 1972 The effect of selection on esterase allozymes in a barley population. Genetics 72: 489–503.

Browning, S. R., and B. L. Browning, 2012 Identity by descent between distant relatives: detection and applications. Annu. Rev. Genet. 46: 617–633. https://doi.org/10.1146/annurev-genet-110711-155534

Chen, Z., and K. Wang, 2019 Gene-based sequential burden association test. Stat. Med. 38: 2353–2363. https://doi.org/10.1002/sim.8111

Dobzhansky, T., and H. Levene, 1948 Genetics of natural populations; proof of operation of natural selection in wild populations of Drosophila pseudoobscura. Genetics 33: 537–547.

Dutta, D., L. Scott, M. Boehnke, and S. Lee, 2019 Multi-SKAT: general framework to test for rare-variant association with multiple phenotypes. Genet. Epidemiol. 43: 4–23. https://doi.org/10.1002/gepi.22156

Erlich, Y., T. Shor, I. Pe'er, and S. Carmi, 2018 Identity inference of genomic data using long-range familial searches. Science 362: 690–694. https://doi.org/10.1126/science.aau4832

Fadista, J., A. K. Manning, J. C. Florez, and L. Groop, 2016 The (in)famous *P*-value threshold revisited and updated for low-frequency variants. Eur. J. Hum. Genet. 24: 1202–1205. https://doi.org/10.1038/ejhg.2015.269

Ghiani, M. E., A. Mameli, R. Robledo, and C. M. Calò, 2019 Allele frequency and forensic efficiency of 15 autosomal STR loci in the Sardinian population (Italy). Forensic Sci. Int. Genet. 41: e26–e29. https://doi.org/10.1016/j.fsigen.2019.04.002

Gogarten, S. M., T. Bhangale, M. P. Conomos, C. A. Laurie, C. P. McHugh *et al.*, 2012 GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics 28: 3329–3331. https://doi.org/10.1093/bioinformatics/bts610

Graffelman, J., 2015 Exploring diallelic genetic markers: the HardyWeinberg package. J. Stat. Softw. 64: 1–23. https://doi.org/10.18637/jss.v064.i03

Guo, S. W., and E. A. Thompson, 1992 Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48: 361–372. https://doi.org/10.2307/2532296

Henn, B. M., L. Hon, J. M. Macpherson, N. Eriksson, S. Saxonov *et al.*, 2012 Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One 7: e34267. https://doi.org/10.1371/journal.pone.0034267

Ho, D. S. W., W. Schierding, M. Wake, R. Saffery, and J. O'Sullivan, 2019 Machine learning SNP based prediction for precision medicine. Front. Genet. 10: article e267. https://doi.org/10.3389/fgene.2019.00267

Jennings, H. S., 1916 The numerical results of diverse systems of breeding. Genetics 1: 53–89.

Lander, E. S., 1989 DNA fingerprinting on trial. Nature 339: 501–505. https://doi.org/10.1038/339501a0

Levene, H., 1949 On a matching problem arising in genetics. Ann. Math. Stat. 20: 91–94. https://doi.org/10.1214/aoms/1177730093

Li, B., N. X. Zhang, Y. G. Wang, A. W. George, and A. Reverter, 2018 Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front. Genet. 9: Article e237. https://doi.org/10.3389/fgene.2018.00237

Vithayasai, C., 1973 Exact critical values of the hardy-weinberg test statistic for two alleles. Commun. Stat. 1: 229–242. https://doi.org/10.1080/03610927308827020

Wakefield, J., 2010 Bayesian methods for examining Hardy-Weinberg equilibrium. Biometrics 66: 257–265. https://doi.org/10.1111/j.1541-0420.2009.01267.x

Weir, B. S., and J. Goudet, 2017 A unified characterization of population structure and relatedness. Genetics 206: 2085–2103. https://doi.org/10.1534/genetics.116.198424

Weir, B. S., R. W. Allard, and A. L. Kahler, 1972 Analysis of complex allozyme polymorphisms in a barley population. Genetics 72: 505–523.