BOOK REVIEW

Handbook of statistical genetics, 2nd edition

Edited by D. J. Balding, M. Bishop and C. Cannings

Wiley; 2003; 1308 pages in two volumes; £230–£290; ISBN 0-470-84829-4 (hardcover)

Reviewed by J. H. EDWARDS*

The problems created by the massive data available since the decipherment of the first human genome, followed by *in situ* quantitation of RNA in the thousand or so cell varieties, in health and disease, are far from resolved. Unlike astronomy, with its equivalent problems of evolution and data-drowning, biology and medicine involve fewer units but a far greater variety: only sequence data can claim exemption from the fuzzy nouns on which most other descriptions depend. To make matters worse few observers of living organisms—well, sick or dead—can approach the average mathematical ability of astronomers. Unlike astronomy, biology is more than an observational study: we depend on plants and animals to concentrate and package our foods and medical services to ease the problems of birth, illness and death.

Big problems need big books. This one, or rather pair—the first edition has twinned—certainly qualifies, weighing in at almost three kilograms and costing £230–£290 pounds (the amount seems to vary for new books according to supplier identified on the Internet, and some suppliers quote lower, changing prices for used copies). Blackwells decline to display such expensive books, possibly explaining their solitary order when last consulted. While the reputation of its contributors will ensure good sales to major libraries, its cost will deprive them of the readership they deserve. It will rarely enter the shelves of small departmental libraries, where they exist, or are allowed to exist. Advances in genetics, as in its subject matter, are largely dependent on small departments: larger departments are essential for utilizing these advances.

Additional chapters in the second edition are on 'Evolutionary quantitative genetics', 'Bayesian methods in genomics' and 'Analysis of microarray gene expression data': only the last is justified by technical advances since the first edition. There is more extensive indexing in the second edition. The reference index is deficient in giving the page of the reference rather than the page needing the reference. The glossary has many omissions and some unacceptable errors. There seems no justification for the publishers producing a second edition when the changes in the 35 chapters in the first edition are minimal. What was needed was a companion volume that could include brief supplements, including errata, to the first and second editions, and complementary chapters on the application of computer programs to the analysis of real data, excluding sequence, farmyard and phylogenetic data adequately covered in the first edition. I will use the term book to cover the two-volume second edition.

Is it worth it? The answer is yes for all institutions that have not purchased the first edition and are spending more than a thousand times the cost of the second on studies relating to the mammalian genome, especially the human genome, and inborn factors influencing, but not determining, disease. It costs less than typing a small part of the genome of just one individual, and many thousands of times less than the cost of a misdirected study or an inappropriately specialized or staffed institute. The presentation is good, but includes only four well-chosen colour photographs, all for the new microarray article, and most impressive: barely enough given the price. It lacks a few obvious needs, such as the karyotypes of man, mouse, rat, fly and worm, to use the genetic vernacular. As all editors are from the UK, most authors from Europe, India or Australia, and the publisher from the UK, there is no excuse for not writing in the fairly standard non-American English of Europe and countries, including Canada, that were once British colonies. Some authors appear to have had the richness and subtlety of their usual style dumbed—a possible casualty of Microsoft's aids to simple prose. If true, a serious loss: this is not a book for children.

It is a book with most chapters by, and for, highly numerate statisticians: it documents and extends the theoretical foundations, most with little direct application to real problems. But it is in the application to real problems, especially those involving common disorders, that new developments are urgently needed, and must be based on the sound foundations established by the formal statistical analysis of imaginary populations as well as the established mechanisms relating phenotype and genotype. The wealth of raw data from HapMap and other major studies—past, present and future—needs better procedures. The data must be translated into

*Emeritus Professor of Genetics, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK.
Email: jheox@ntlworld.com.

words and diagrams, with supporting estimates of relevant measures and their precision, rather than yes and no solutions based on tests of significance.

It is, as the editors point out, not a book to be 'read'. It is a collection of chapters with a commendable degree of overlap, with the inevitable casualties of differing symbols for the same words. The final dates of many chapters testify to the remarkable enforcement of editorial discipline—perhaps explaining the absence of some expected authors. The production is good, if faulted by transatlantic spelling. The classification of groups of chapters differs from that in the first edition, but the attempt to impose order creates its own chaos. It would have been better to retain the order of the first book, with added chapters being placed at the end. Or to use the alphabet, as in *Biostatistical genetics and genetic epidemiology* by Elston, Polson and Palmer (2002) with its numerous short articles—an equally valuable complementary volume. They should be shelved together and, if either can be afforded, preferably purchased together.

Few chapters have more than one or two authors, and most maintain the difficult but essential aim of what the introduction calls 'serving as both a tutorial and a reference book'. The reviewer, a hunter-gatherer, is inadequately tutored to follow many 'tutorials' but can at least claim to have read all the words and some of the formulae, perhaps the only person to have done so.

The foundation of population genetics necessarily lies in the analysis of imaginary populations with assumptions on imaginary organisms simple enough to make mathematical analysis possible. The benefits of applying numbers to the 'tangled web' of biological reality, with its fluffy nouns for the individually unique units observed in field, farm and clinic, have had distinguished critics. Mayr questioned the existence of any concepts requiring anything that could not be handled by simple arithmetic, a challenge Haldane responded to with his 'defence of beanbag genetics'. Those whose power of words may have benefited by freedom from the distractions and temptations of numeracy include Erasmus and Charles Darwin, Wallace and Bateson. Wallace, who used his experience as a surveyor to good effect, limited his key inferences to words. There is room for both imaginary and real populations—the problem is that they only enjoy mutual benefit if rigorously distinguished.

The book is largely restricted to the analysis of imaginary populations that are large, mate at random, at least for the first time, and have genomes whose disruptions from recombination and mutation occur independently between, or within, any units of inheritance considered. Further assumptions include uniform mutation and evenly spaced recombinants unrelated to mutation or its obverse, conversion, which is usually associated with recombination, and a restriction to point mutations—no deficiencies or highly mutable segments with repeats.

It is useful to distinguish imaginary Populations, which I will distinguish by a capital P, and real populations. The ideal of Platonic Populations and the world of real populations are essential partners and their distinction was usually implicit in the writings of Fisher, Wright, Haldane, Bernstein, Penrose, Hogben, Morgan, Sturtevant, Waddington, most of whom used mathematics, often including advanced mathematics, when, and only when, words were inadequate, in distinction to Pearson and his biometric school who expected nature to follow art.

There were exceptions, but relatively few, as in Haldane's preoccupation with infinite series, and possibly Fisher's use of Bessel functions in his book on inbreeding, but they were distractions rather than obstructions to their words. Others of that vintage developed the basis of most practical applications, building on the plan first explicitly advanced by Erasmus Darwin over two centuries ago, applying the hand of chance as defined by Mendel and maintaining the exacting vocabulary initiated by Mendel, Galton, Bateson, Morgan and others, especially Wright, Fisher and Muller, who added new words only when necessary.

Erasmus Darwin's *The Temple of Nature* (1803) starts:

> *By firm immutable immortal laws*
> *Impress'd on Nature by the GREAT FIRST CAUSE,*
> *Say, MUSE! how rose from elemental strife*
> *Organic forms, and kindled into life;*
> ....

This book provides an anthology of answers, but their relevance to unravelling the 'tangled bank' described by his grandson still awaits adequate development in integrating numbers with words and Populations with populations.

Some major subjects escape mention, including the EM algorithm, and its limitations; altruism, discussed by Fisher and Haldane, and advanced, with personal observations and deep mathematical treatment by Hamilton; and Maynard Smith's application of 'games theory'. However, these major contributions to concept formation were largely based on Platonic Populations, as opposed to real populations, and cannot easily be applied to any large mammal, especially ourselves with our rapid evolution, with successful societies however small depending on maintaining variety in the powers to observe, act, hunt, gather, fight, infer and communicate, unlike the 'standard phenotype' evolving to conserve similarity in most vertebrates. From the great apes to urbanization we survived with sparse localized populations within which mate selection could rarely extend beyond a dozen or so possible candidates, few beyond third cousins. Human populations were genetically 'enriched' by conquest with merging, the capture of women and children, and later slavery, but this can hardly be simulated. The tree model of Populations is far from the disorderly lattice of populations.

The generations since our common ancestry with the chimpanzee can probably be estimated to within a factor of three, the effective population size to within a factor of 10, and the mutation rates at different loci, and recombination between equally spaced loci, to within a factor of 100. Duplications, deficiencies and conversions, the latter usually as-

sociated with recombination, offer even greater divergence from a model Population. This is not a problem provided the distinction between Platonic Populations and the real world of populations, so dependent on the insights and vocabulary generated and honed on both, is appreciated.

The book provides major contributions in the form of reviews to the problems of Populations, but has little on how to apply the results to populations excepting phylogeny and agricultural and forensic genetics and a major contribution on human evolution. The glossary, a new and important addition, has many omissions—including conversion, chromatid, deficiency, duplication, insertion, inversion, MLOD, resistance and susceptibility, polymorphism, and translocation, including reciprocal, balanced and unbalanced forms. And there are unfortunate errors. In spite of authorial variety and expertise there are some gaps even where historical introductions have been included. Bernstein, who introduced what is now termed LD analysis in 1924 to define the first human linkage, and later demonstrated the feasibility of using two generations, documenting the first human non-linkage, is not mentioned. Nor is Hogben's development of the latter procedure in 1934—indeed Hogben escapes notice altogether.

The glossary should have provided a much-needed and authoritative verbal foundation. It includes some very clear and short definitions, such as 'homology', but has many omissions, including those noted above, and some errors. They include:

centiMorgan: It is not 'because of variability in recombination rates, genetic distance differs from physical distance'. It is the length of a segment with a 1% chance of recombination.

Clones occur naturally as well as artificially.

Epistasis usually implies affecting a single gene.

Exons 'do (not *may*) code for a specific part' of a protein.

Fixation involves alleles not loci.

Genotype: A genotype of two or more loci is 'the product of' and not 'equivalent to' two haplotypes.

Haemoglobin and *haploid* entries are unduly brief in view of their importance.

Haplotype: This is correctly defined as 'the alleles at different loci on a chromosome' but the gratuitous addition 'in the presence of strong linkage disequilibrium haplotypes may be inferred from genotypes with few errors' should be omitted unless supported by information on the length of inferred haplotype and 'few': but this would need a chapter—indeed this urgently needed chapter is a sad omission.

Hardy–Weinberg's law, now often termed an equilibrium, is defined via a cumbrous double negative.

Inbreeding: Does not necessarily result from endogamy. It can be avoided by a preference against close relatives, and is in most societies. Once stabilized it will not lead to an 'increase in the prevalence of recessive traits' although they will be concentrated within the more inbred families.

Linkage: 'two genes are said to be linked if they are close together on the same chromosome'—a difficult problem of definition: if loci A,B and B,C can be linked can A,C be unlinked? Perhaps 'on the same chromosome', often restricted to within range of detection by family studies. 'Close' usually implies less than five centiMorgans. Up to 30 cM is usually within range.

LOD score: Barnard introduced the word 'lod' to statistics, Morton to genetics, and Ott used it in his first book. There is no reason to change this established lower-case usage, which has the advantage of being less confused with MLOD, which lacks the key feature of additivity.

Mutation can change more than an allele—deficiencies, duplications, insertions and translocations qualify—words missing from the glossary.

Polygenic is more than 'more than one gene'. In normal genetic usage, as in Greek, ancient and modern, it implies 'many many, usually a dozen or so'. The Latin 'multifactorial' has the advantage of being a more appropriate adjective, covering smaller numbers, and being well established in this context. It is not in the glossary.

Polymorphic and *monomorphic* are rare terms, compared to *polymorphism* which is not included.

The first chapter, on chromosome maps, appropriately if fortuitously placed, reviews the history, the technique, its consequences, its analysis and its results in the form of chromosomes annotated by linkage: it covers a wide range of organisms, including fungi. Attention is drawn to the neglect of recent work on combinatorial optimization in attempting to order loci from family data. The second chapter, on the problems of sequence comparisons, succeeds in the even more difficult problem of explaining the mathematical basis of the major advances in fast and sound interpretations of different sequences of DNA and amino acids, enlivened by the elegant and simple statement that the expectation of a run of heads of length k in n tosses is $n/2^k$, a solution attributed to Erdös.

After this lucid start, that few geneticists could read without benefit, a third chapter on Bayesian methods applied to sequence analysis involves advanced mathematics but provides no empirical evidence of the benefits of relative speed or security of inference. Bayesian is not restricted to its usual sense of expectations based on evidence prior to analysis. These three chapters are typical of the book, and should be—a mixture of topics—, some necessarily beyond being tutorials for the untutored, most assuming a very numerate readership. The next chapter, on within-gene detection, exon prediction and allied problems, returns to tutorial status with only necessary mathematics, mainly simple, complemented by powerful diagrams. The fifth explores the formal and, in practice, largely neglected, problem relating different species by tree structures. But no examples. Visual displays relating pairs of species, in which the first author made major advances using and extending ACeDB, are not discussed or exhibited. Not a chapter for the untutored. After this section of five chapters with common themes—three for tutored and untutored alike, and two for the well-tutored—, the rest of

the book follows a similar pattern of related batches of chapters appropriately mixed for deep mathematicians and others. Most chapters provide extensive references: one on 'Phylogenetics: parsimony and distance methods' is somewhat bereft of the past, with only eight references, four including an author, the latest from 1989. The general pattern continuing, the book closes on a valuable paper on human evolution that summarizes the position with minimal, but necessary, mathematics.

What of the future? Massive books with impressive titles and authorships relevant to the genome are purchased by librarians for drug companies, universities and large research institutes. But cost limits purchase by individuals or small departments, and even the largest booksellers hesitate to purchase for display, only ordering on demand. Blackwells, a major bookseller in Oxford, sold one copy after displaying none when last asked. While such massive books may be economically viable, hopefully with appropriate royalties, the authors are not rewarded with the readership they deserve. After seventy years, if not a hundred, opinions differ, they will be available on the Web, a goldmine for historians of this difficult field in which, at the time of publication, the production of data had outrun both adequate means for their analysis and a realization of this obstacle to sound investment of research funds.

In the evolution of genetics, as of its subject matter, conceptual advances usually take place in small departments, or departments with other primary interests. There is a need for such a book, and its hoped-for successor on application, to be on the Web with electronic cross-indexing and appropriate arrangements for payment, profit and royalties for viewing, with higher costs for printing and legal restraints on further copies. There is a market for encyclopaedias and large books, but while there are economic benefits to publishers and, one hopes, substantial royalties to editors and authors, they fail in the main purpose of a book: to be read. Unless the major publishers in this expanding field appreciate the risks of overproduction of hard copy, and the competition from the increasing influence and standards of such organizations as www.wikipedia.org and the long-term plans of Google, their days will be numbered. If either the foundations of words and algorithms, or the superstructure of their application, at present a major limitation on both strategy and analysis, are to advance, radical changes in ease of access are necessary.

Otherwise our successors may well be surprised at so wide a potential readership, and so expert an authorship, being as ineffective as the many copies of Mendel's work on inheritance in the edible pea.