Bayesian statistics in genetics

a guide for the uninitiated

Statistical analyses are used in many fields of genetic research. Most geneticists are taught classical statistics, which includes hypothesis testing, estimation and the construction of confidence intervals; this framework has proved more than satisfactory in many ways. What does a Bayesian framework have to offer geneticists? Its utility lies in offering a more direct approach to some questions and the incorporation of prior information. It can also provide a more straightforward interpretation of results. The utility of a Bayesian perspective, especially for complex problems, is becoming increasingly clear to the statistics community; geneticists are also finding this framework useful and are increasingly utilizing the power of this approach.

Jennifer S. Shoemaker
shoem003@
mc.duke.edu

***Ian S. Painter**
painter@talariainc.com

‡Bruce S. Weir
weir@stat.ncsu.edu

The Cancer Prevention, Detection, Control Research Program, Duke Medical Center, Box 2949, Durham, NC 27710, USA.
*Talaria, Inc., 501 Hoge Building, 705 2nd Avenue, Seattle, WA 98104, USA.
‡Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA.

Statistical analyses are pervasive in genetics. These analyses are generally conducted in a classical statistical framework, but there is a rising interest in the applications of Bayesian statistics to genetics. Bayesian methods can be especially valuable in complex problems or in situations that do not conform naturally to a classical setting; many genetics problems fall into one of these categories. In addition, Bayesian approaches can be easier to interpret and they have been employed in many genetic areas, including: the classification of genotypes and estimating relationships[1–3]; population genetics and molecular evolution[4–17]; linkage mapping (including gene ordering and human-risk analysis[18–33]); and quantitative genetics [including quantitative trait locus (QTL) mapping[34–45]]. Here, we discuss the classical and Bayesian approaches and we then illustrate the appeal of Bayesian approaches by providing examples from the literature.

A difference in the definition of probability

The views that we present here are necessarily oversimplified and are meant to capture only the essence of classical and Bayesian perspectives. A good introduction to Bayesian statistics can be found in Berry[46]; more advanced books are also available[47,48].

Classical methods, also called frequentist or standard methods, are named for their definition of probability as a long-term frequency. In other words, probability is viewed from the framework of (hypothetically) repeating an experiment many times under identical circumstances (Fig. 1). Imagine crossing two plants to determine the mode of inheritance of trait *A*. Assuming a particular mode of inheritance, the expected ratio of phenotypes can be determined *a priori*. Testing the observed ratio against a theoretical ratio leads to a *P* value, which is interpreted from the point of view of long-term frequency: if the same experiment was repeated many times, the observed result (or a more-extreme one) would be expected a proportion *P* of the time, assuming the null hypothesis of no difference is true. It is important to note that a *P* value is a long-term frequency statement and that it is specifically a statement about the data.

The Bayesian paradigm also uses probability to assess statistical confidence, but with an expanded definition of probability. The name Bayesian comes from the Reverend Bayes, who formulated Bayes' rule, which is the computational underpinning of Bayesian methods. In the Bayesian paradigm, a probability is a direct measure of uncertainty, and might or might not represent a long-term frequency. This definition of probability is closer to that in common speech: 'I think there is only a 5% chance that the earth has been visited by aliens this century'. It is hard to frame this probability in terms of a long-term frequency about aliens visiting earth this century. The statement uses probability as a direct measure of uncertainty: 'I am as sure (or as certain) that aliens have visited earth this century as I would be of obtaining a 20 the next time I roll a fair 20-sided die'.

Additionally, in a Bayesian framework, probability statements are made about the parameter. In Fig. 1, one could consider more than one mode of inheritance and calculate the probability of each mode. In a more complicated example from the literature, a Bayesian approach was used to investigate whether inheritance in the tetraploid perennial, *Astilbe biternata*, was disomic or tetrasomic[10].

Drawing conclusions based on posterior distribution

In the framework of the above definitions of probability, how are conclusions drawn? In classical statistics, conclusions can be based on a *P* value, or on a confidence interval, each of which is a long-term frequency statement. Essentially, these provide evidence against a hypothesis. Often, one assumes a null hypothesis of no difference between two quantities. One then performs the experiment

0168-9525/99/$ – see front matter © 1999 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(99)01751-5

and calculates, from the data, the sample value of the appropriate test statistic. In order to evaluate the evidence against the null hypothesis, one compares the sample value of this test statistic with the distribution of the test statistic under the null hypothesis. Extreme values of the observed result are taken as evidence against the null hypothesis. However, with a large-enough sample size, one can always reject a false null hypothesis. The framework of testing the significance of the null hypothesis confounds the amount of evidence with the degree to which the null is violated; statistical significance does not always imply biological significance. A lively discussion on the problems with *P* values in the framework of null hypothesis testing can be found in Cohen[49] and Hagen[50].

In Bayesian statistics, evidence in favor of certain parameter values, θ, is considered. Inference is based on the posterior distribution, $p(\theta|X)$ (see Box 1 and Box 2), which is the conditional distribution of the parameter, given the data, X. It is a combination of the prior information and the data. From Bayes' rule:

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)} \qquad (1)$$

The term $p(\theta)$ represents the prior distribution on the parameters. Prior information can be based on previous experiments, or on theoretical or other considerations (see below). In our example to determine the mode of inheritance of trait *A*, equal weight could be placed on each mode of inheritance. The data are phenotypic counts, so $p(X|\theta)$ (which has the same form as the likelihood) is multinomial. The denominator $p(X)$ is a normalizing factor.

In Bayesian and classical statistics we want to make inferences about a fixed, but unknown, parameter value. The difference is in how we approach this goal and in the interpretation of the results.

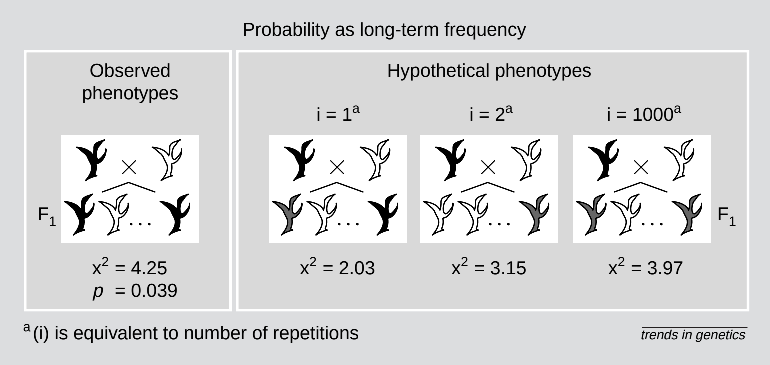
Advantages of Bayesian statistical methods

Addressing the question of interest directly

In many cases, Bayesian methods can address the question of interest more directly than a classical approach. Two examples from the literature, the first from the field of population genetics and the second from linkage analysis, illustrate this point. In a large, randomly mating population that is free of disturbing forces, allele and genotypic frequencies do not change and are related in a simple way. The population is said to conform to Hardy–Weinberg equilibrium (HWE). In a classical setting, one tests for whether the population is exactly in HWE and then looks for evidence against this null hypothesis. However, in many cases, the experimenter does not believe that the population is exactly in HWE and might fail to reject a false null hypothesis. A Bayesian approach can reflect a more relevant question, which might be 'are departures from HWE large enough to be important?' The size of departure that is important varies with the context. We addressed this question in the context of forensic science, where an important departure in human populations was suggested by the United States National Research Council (NRC)[12].

A typical objective of research in linkage analysis is to determine the extent of linkage between two loci, for example, between a trait locus and a marker locus. In a classical setting, evidence against the null hypothesis of no linkage is investigated; in a Bayesian approach, the probability of linkage is calculated, given the data from a particular experiment[20,26,30,31]. Silver and Buckler[31] contrasted the questions asked in the two frameworks concisely: 'The Bayesian approach answers the question "Given the observed results, what is the probability that two loci are separated by up to *m* centimorgans?" The more traditional analysis answers the question "If the loci were separated by *m* centimorgans, how unlikely would the observed results be?" We believe that the answer to the former question corresponds more closely to what the experimenter wants to know'.

FIGURE 1. Testing the observed against the theoretical



In an experiment to determine the mode of inheritance of a trait, a cross is performed. The ratio of the observed phenotypes in the offspring is tested against the expected (theoretical) ratio using a goodness-of-fit test. A *P* value of 0.039 means that if the experiment were repeated many times, the expected proportion of chi-square statistics as large (or larger) than the observed value, 4.25, is 0.039. The *P* value depends on a framework of hypothetical repetitions and is a long-term frequency statement about the data.

BOX 1. Glossary

Dimensionality
The number of axes; here the number of axes in a parameter space. For example, if we are interested in measuring departure from Hardy–Weinberg equilibrium in a population with two alleles at the locus of interest, there are perhaps two parameters to consider: an allele frequency and a parameter describing the departure from HWE; thus, the parameter space is two dimensional.

Markov chain Monte Carlo (MCMC)
A method for integrating by sampling from the posterior distribution; allows integration over high-dimensional spaces.

Nuisance parameter
A parameter that is needed to define the problem but is not of primary interest; in considering departures from HWE, allele frequencies are nuisance parameters.

Parameter
Unobservable quantities of interest; these can include population parameters, such as allele frequencies or location of QTL, or missing data. Here, we denote a parameter by θ.

Parameter space
The set of all possible values of the quantity of interest; the parameter space for a population allele frequency includes all values between zero and one, inclusive.

Posterior distribution (posterior)
The conditional probability distribution of the unobserved quantities of interest (parameters) given the observed data.

$p(\theta|X)$
Symbol referring to the posterior distribution; it should be read 'conditional distribution of theta given the data' or simply 'posterior distribution'.

Vague (dispersed) prior distribution
Distribution is spread out diffusely over the parameter space. An example of a vague prior could be a prior in which all possible values of the parameter have equal weight. Another way to think about a vague prior is that is has a larger variance than a prior that is not so spread out.

BOX 2. Calculating posterior probabilities in a discrete setting

As an example, consider the problem of determining which of three subpopulations (θ_1, θ_2 or θ_3) individual *I* belongs to, based on observations of genotypes at several loci and knowledge of genotype frequencies in each of the subpopulations. The context might be that a blood stain is found at a crime scene and the question is to determine which of three subpopulations, Caucasian, Maori or Western Polynesian, the contributor of the stain belongs to (assume that attention can be restricted to these three subpopulations). The New Zealand census in 1991 reported that the population in that country had the following composition: 81.9% Caucasian, 13.7% Maori, and 4.4% Western Polynesian. Probabilities of the observed genotypes (here called a DNA forensic profile) X_I can be calculated. For this example, we will suppose that the three probabilities $p(X_I/\theta_1)$, $p(X_I/\theta_2)$ and $p(X_I/\theta_3)$ have been calculated to be 3.96×10^{-9}, 1.18×10^{-8}, 1.91×10^{-7}. The prior probabilities, $p(u)$ for the three subpopulations are 0.819, 0.137, and 0.044, respectively. Using equation **(1)**, the posterior probability of belonging to each of the subpopulations is then:

$$p\left(\theta_1 \middle| X_I\right) = \frac{0.819 \times 3.96 \times 10^{-9}}{p(X_I)} = 0.25$$

$$p\left(\theta_2 \middle| X_I\right) = \frac{0.137 \times 1.18 \times 10^{-8}}{p(X_I)} = 0.12$$

$$p\left(\theta_3 \middle| X_I\right) = \frac{0.044 \times 1.91 \times 10^{-7}}{p(X_I)} = 0.63$$

where

$$\begin{aligned} p(X_I) &= 0.819 \times 3.96 \times 10^{-9} + 0.137 \times 1.18 \times 10^{-8} \\ &\quad + 0.044 \times 1.91 \times 10^{-7} \\ &= 13.26384 \times 10^{-9} \end{aligned}$$

Incorporation of prior information

Prior information is used in classical settings, for example, in planning the size of an experiment. However, in a Bayesian analysis, prior information is incorporated in a very specific way. It is combined with information from the data to generate the posterior distribution over the parameter values, according to Bayes' rule. What prior information might one use? As an example, we consider the analysis of the impact of DNA-sequencing errors on the ability to align predicted protein sequences to known protein sequences[14]. Sequencing errors, such as substitutions or frameshift errors (insertions or deletions), decrease the ability to align sequences correctly. Including prior information on these types of errors allowed the determination of how high the error rates could be, while maintaining accurate alignments. Including prior information on the location of error rates, for example, if they were not distributed uniformly over the sequence, allowed accurate alignment in the presence of even-higher error rates. In the same study, prior information on codon bias in yeast (*Saccharomyces cerevisiae*) improved detection of the correct reading frame.

It is important to distinguish between prior information about the parameters of interest, and prior information about nuisance parameters, which are parameters that are not of primary interest, but which are needed to define a problem. The Bayesian approach provides a framework for making inferences about the parameters of interest, while taking into account uncertainty in the nuisance parameters. An interesting discussion of choosing priors in human linkage analysis on the parameter of interest, the linkage parameter, and on the nuisance parameters, allele frequencies and penetrances, can be found in Thomas and Cortessis[33]. Other studies have used prior information about the population structure[1], the genome length and number of chromosomes[20,26,30,31], the distribution of allelic effects and allele frequencies[22,23], or the range of quantitative-trait values[38,45].

Even if there is only vague prior information, it is still possible to select a prior distribution in a sensible way. If different hypotheses are being considered, each can be weighted equally or weighted according to other information[10,12,13]. In any situation, a vague or dispersed prior can be used, so that any particular value does not have a high weight. Prior distributions are often chosen from classes of distributions that have a computationally convenient form, but are flexible enough to represent the desired uncertainty[35,37–40,43–45].

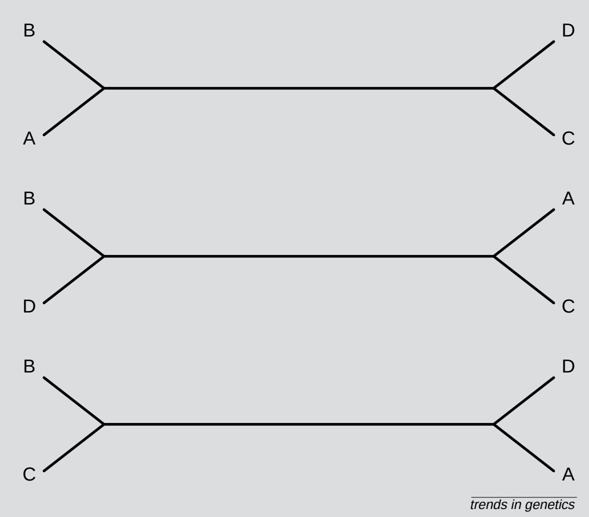
Avoiding problems with hypothesis testing

In a classical setting, assessing confidence for several hypotheses can be problematic because only two hypotheses can be compared at a time. In a Bayesian setting, the posterior probability of each hypothesis is calculated. Sinsheimer *et al.*[13] consider the case of constructing an unrooted phylogenetic topology from four taxa. There are three possible unrooted topologies to consider (Fig. 2). From a classical perspective, testing three topologies is equivalent to three pairs of hypothesis tests, while from a Bayesian perspective, it is equivalent to calculating the probability of three hypotheses. The latter results are easier to interpret. In the classical case, confidence is often assessed indirectly (by looking at evidence against a null in three pairs of hypothesis tests and trying to assess an overall level of plausibility). In the Bayesian case, confidence is assessed directly by calculating the probability that a given phylogeny is correct. Mau and Newton[9] noted that their Bayesian method for reconstructing phylogenetic topologies includes a measure of confidence; classical methods first find an optimal tree, and then repeat the analysis many times to generate confidence statements. In their example, a Bayesian method provided a level of confidence that is easy to interpret as well as a method that is more efficient.

Bayesian solutions to complex problems

In many problems in genetics, the number of parameters is large. For QTL mapping, the parameters and missing data can include: marker allele frequencies; marker map

FIGURE 2. Possible trees



The three possible unrooted trees for four taxa, represented by A, B, C and D.

positions; QTL allele frequencies; QTL map position(s); QTL effects; QTL-marker genotypes; linkage parameters; number of QTL; overall mean and other fixed effects; and polygenic and residual variances[35,38,39,43,44]. For the assessment of genetic risk, parameters include: gene order; recombination distance; mode of inheritance; pedigree configuration; penetrance; and mutation rate[25,26]. For population genetics problems, the number of parameters increases as the number of alleles increase[5,51]. For phylogenetic reconstruction, parameters can include nucleotide substitution rates, transition or transversion rates, tree topologies and speciation times[9,15].

As mentioned, some of the parameters needed to define a problem are not of primary interest and are called nuisance parameters; allele frequencies often fall into this category. A very important feature of Bayesian methods is that they provide a convenient way for accounting for uncertainty in nuisance parameters. In QTL mapping, the locations and effects of the QTL are often of primary interest, but, as in the HWE problem, allele frequencies are necessary to define the problem. Bayesian methods offer a convenient way of handling nuisance parameters. The distribution of the parameter(s) of interest, such as the degree of departure from HWE or the effect of a QTL, can be obtained by integrating over all possible values of the nuisance parameter(s). In classical statistics, the value of the nuisance parameter is often concentrated upon, instead of considering all possible values.

Finally, missing data, ubiquitous in pedigree studies for example, can be handled in a Bayesian analysis by treating them as an unknown parameter. In these cases, the Markov chain Monte Carlo (MCMC) methods described below are invaluable.

Problems with Bayesian methods

A common criticism of the Bayesian approach is that the choice of the prior distribution is too subjective. This objection is related to the fact that, in some cases, the posterior distribution is very sensitive to the choice of prior. In these cases, two researchers using the same data could reach different conclusions if they used different priors. The definition of probability as a degree of belief or uncertainty, rather than as a long-term frequency, is related to this concept. Proponents of a Bayesian approach might counter with arguments that are summarized succinctly by Gelman *et al.*[48]: 'All statistical methods that use probability are subjective in the sense of relying on mathematical idealizations of the world. Bayesian methods are sometimes said to be "subjective" because of their reliance on a prior distribution, but in most problems, scientific judgement is necessary to specify both the "likelihood" and the "prior" parts of the model. For example, linear regression models are generally at least as suspect as any prior distribution that might be assumed about the regression parameters'.

Another difficulty is that the implementation of Bayesian methods can be very complex. Prior distributions must be specified for the parameters and the posteriors integrated over the nuisance parameters. Even if convenient priors are chosen, integrating over the nuisance parameters can be complicated in practice, especially if the parameter space is complex, or the dimensionality is high, or both. However, the increase in computing power together with the use of MCMC methods have made Bayesian techniques more accessible. The use of MCMC is probably the single most important development in this field and has made Bayesian computation feasible. In this introductory review, we cannot do justice to this development, but note that it uses samples from a simulated distribution that is expected to be the posterior distribution, instead of deriving this distribution by integration. Those interested in learning more about this topic should read the excellent book by Gilks *et al.*[52] In addition, software programs are available, such as BUGS (Bayesian inference using Gibbs sampling; http://www.mrc-bsu.cam.ac.uk/bugs)[53], BAMBE (Bayesian analysis in molecular biology and evolution; http://mathcs.duq.edu/larget/bambe.html)[54], and others (http://www.wadsworth.org/resnres/bioinfo)[17].

Conclusion

Because the purpose of this review is to raise awareness of a Bayesian approach, the focus is on the advantages of this approach. Our goal is not to replace classical statistics with Bayesian methods but to emphasize areas where the latter can be particularly useful, such as when it makes sense to consider a question from the point of view of updating uncertainty about a parameter, rather than considering a question in the framework of repeated hypothetical experiments. Examples include questions about the size of departure from HWE, the probability of linkage and the probability of a given topology of a phylogenetic tree.

In general, as more information becomes available from genome and gene-expression projects, the demand for methods of analysis increases. Bayesian methods can contribute to the development of suitable methods by providing a framework in which many questions can be addressed directly, uncertainty in all parameters can be taken into account and prior information can be incorporated.

Acknowledgements

We are very grateful for a thorough reading of the manuscript by J. Thorne, J. Leips and J. Curran. In addition, we are grateful to two anonymous reviewers, whose excellent suggestions improved the manuscript. This work was supported in part by US Public Service Grants P01 GM43544 and T32 GM08443.

References

1 Alexander, H. *et al.* (1995) A Bayesian approach to the inference of diploid genotypes using haploid genotypes. *Theor. Appl. Genet.* 91, 1284–1287

2 Thompson, E.A. and Meagher, T.R. (1987) Parental and sib likelihoods in genealogy reconstruction. *Biometrics* 43, 585–600

3 Painter, I. (1997) Sibship reconstruction without parental information. *J. Agricultural, Biol. Environ. Statistics* 2, 212–229

4 Allison, L. and Wallace, C.S. (1994) The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimisation of multiple alignments. *J. Mol. Evol.* 39, 418–430

5 Ayres, K.L. and Balding, D.J. (1998) Measuring departures from Hardy–Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity* 80, 769–777

6 Gunel, E. and Weardon, S. (1995) Bayesian estimation and testing of gene frequencies. *Theor. Appl. Genet.* 91, 534–543

7 Lawrence, C.E. *et al.* (1993) Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 262, 208–214

8 Lindley, D. (1988) Statistical inference concerning Hardy–Weinberg equilibrium. *Bayesian Stat.* 143, 307–326

9 Mau, B. and Newton, M.A. (1997) Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo methods. *J. Comp. Graph. Stat.* 122–131

10 Olson, M.S. (1997) Bayesian procedures for discriminating among hypotheses with discrete distributions: Inheritance in the tetraploid *Asilbe biterna. Genetics* 147, 1933–1942

11 Pereira, C. and Rogatko, A. (1984) The Hardy–Weinberg equilibrium under a Bayesian perspective. *Rev. Brasil. Genet.* 4, 689–707

12 Shoemaker, J. *et al.* (1998) Bayesian characterization of Hardy–Weinberg disequilibrium. *Genetics* 149, 2079–2088

13 Sinsheimer, J.S. *et al.* (1996) Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics* 52, 193–210

14 States, D.J. and Botstein, D. (1991) Molecular sequence accuracy and the analysis of protein coding regions. *Proc. Natl. Acad. Sci. U. S. A.* 88, 5518–5522

15 Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.* 14, 717–724

16 Durbin, R. *et al.* (1998) *Biological Sequence Analysis,* Cambridge University Press

17 Zhu, J. *et al.* (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics* 14, 25–39

18 Blank, R.D. *et al.* (1988) A linkage map of mouse chromosome 12: Localization of Igh and effects of sex and interference on recombination. *Genetics* 120, 1073–1084

19 Churchill, G.A. *et al.* (1993) Pooled-sampling makes high resolution mapping practical with DNA markers. *Proc. Natl. Acad. Sci. U. S. A.* 90, 16–20

20 Elston, R.C. and Lange, K. (1975) The prior probability of autosomal linkage. *Ann. Hum. Genet.* 38, 341–350

21 Geburek, T. and von Wuehlisch, G. (1989) Linkage analysis of isozyme gene loci in *Picea abies* (L.) Karst. *Heredity* 62, 185–191

22 Hoeschele, I. and VanRaden, P.M. (1993) Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. *Theor. Appl. Genet.* 85, 946–952

23 Hoeschele, I. and VanRaden, P.M. (1993) Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. *Theor. Appl. Genet.* 85, 953–960

24 Mueller, B. *et al.* (1989) Problems in genetic counseling in a family with an atypical centronuclear myopathy. *Am. J. Med. Genet.* 32, 417–419

25 Neumann, P.E. (1991) Three-locus linkage analysis using recombinant inbred strains and Bayes' theorem. *Genetics* 128, 631–638

26 Ott, J. (1991) *Analysis of Human Linkage,* John Hopkins University Press

27 Renwick, J.H. (1969) Progress in mapping human autosomes. *British Med. Bull.* 25, 65–73

28 Rogatko, A. (1995) Risk prediction with linked markers: Theory. *Am. J. Med. Genet.* 59, 14–23

29 Rogatko, A. (1995) Risk prediction with linked markers: Pedigree analysis. *Am. J. Med. Genet.* 59, 24–32

30 Rogatko, A. and Zacks, S. (1993) Ordering genes: Controlling the design error probabilities. *Am. J. Hum. Genet.* 52, 947–957

31 Silver, J. and Buckler, C.E. (1986) Statistical considerations for linkage analysis using recombinant inbred strains and backcrosses. *Proc. Natl. Acad. Sci. U. S. A.* 83, 1423–1427

32 Stephens, D.A. and Smith, A.F.M. (1993) Bayesian inference in multipoint gene mapping. *Annal. Hum. Genet.* 57, 65–82

33 Thomas, D.C. and Cortessis, V. (1992) A Gibbs sampling approach to linkage analysis. *Hum. Hered.* 42, 63–76

34 Gianola, D. and Fernando, R.L. (1986) Bayesian methods in animal breeding theory. *J. Anim. Sci.* 63, 217–244

35 Hoeschele, I. *et al.* (1997) Advances in statistical methods to map quantitative traits. *Genetics* 147, 1445–1447

36 Janss, L.L.G. *et al.* (1997) Bayesian statistical analyses for presence of single genes affecting meat quality traits in a crossed pig population. *Genetics* 145, 395–408

37 Janss, L.L.G. *et al.* (1993) Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. *Theor. Appl. Genet.* 91, 1137–1147

38 Satagopan, J.M. *et al.* (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* 144, 805–816

39 Sillanpaa, M. and Arjas, E. (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred lines. *Genetics* 148, 1373–1388

40 Sorensen, D.A. *et al.* (1994) Bayesian analysis of genetic change due to selection using Gibbs sampling. *Genet. Sell Evol.* 26, 333–360

41 Tavernier, A. (1991) Genetic evaluation of horses based on ranks in competitions. *Genet. Sell Evol.* 23, 159–174

42 Thomas, D.C. *et al.* (1997) A Bayesian approach to multipoint mapping in nuclear families. *Genet. Epidemiol.* 14, 903–908

43 Uimari, P. and Hoeschele, I. (1997) Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* 146, 735–743

44 Uimari, P. *et al.* (1996) The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* 143, 1831–1842

45 Wang, C. *et al.* (1994) Response to selection for litter size in Danish landrace pigs: a Bayesian analysis. *Theor. Appl. Genet.* 88, 220–230

46 Berry, D.A. (1996) *Statistics: A Bayesian Perspective,* Wadsworth Publishing

47 Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis* (2nd edn), Springer-Verlag

48 Gelman, A. *et al.* (1995) *Bayesian Data Analysis,* Chapman & Hall

49 Cohen, J. (1994) The earth is round (P<0.5) *Am. Psychol.* 49, 997–1003

50 Hagen, R.L. (1997) In praise of the null hypothesis statistical test. *Am. Psychol.* 52, 15–24

51 Zaykin, D. *et al.* (1995) Exact tests for association between alleles at arbitrary numbers of loci. *Genetica* 96, 169–178

52 Gilks, W.R. *et al.* (1995) *Markov Chain Monte Carlo in Practice,* (Chapman & Hall)

53 Gilks, W.R. *et al.* (1994) A language and program for complex Bayesian modelling. *The Statistician* 43, 169–178

54 Larget, B. and Simon, D. (1998) Society for Molecular Biology and Evolution, abstract, annual meeting, Vancouver, B.C.

RNA-triggered gene silencing

Double-stranded RNA (dsRNA) has recently been shown to trigger sequence-specific gene silencing in a wide variety of organisms, including nematodes, plants, trypanosomes, fruit flies and planaria; meanwhile an as yet uncharacterized RNA trigger has been shown to induce DNA methylation in several different plant systems. In addition to providing a surprisingly effective set of tools to interfere selectively with gene function, these observations are spurring new inquiries to understand RNA-triggered genetic-control mechanisms and their biological roles.

Andrew Fire
fire@
mail1.ciwemb.edu

Carnegie Institution of
Washington, 115 West
University Parkway,
Baltimore, MD 21210,
USA.

As gene-transfer technologies have become commonplace, an increasing number of organisms have been shown to exhibit potent and unexpected responses to foreign nucleic acids. The ability of some transgenes to silence the expression of homologous (chromosomal) loci was first observed in plants[1] and has subsequently been seen in nematode[2], fungal[3], insect[4] and protozoan[5] systems. Homology-dependent *trans*-silencing effects (see Box 1 for glossary) have been divided into two categories based on the nature of the effect on the target. In the first category, transcription of the target locus is unaffected, whereas the half-life of target RNAs is decreased dramatically[6–9]. Such processes have been called 'PTGS' (post-transcriptional gene silencing). A second category of homology-triggered processes exert their primary effect on the chromatin template[10], and have been termed 'TGS' (transcriptional gene silencing). A striking feature of PTGS, and of a subset of TGS phenomena, has been the existence of RNA trigger molecules responsible for the long-range effect of the transgene locus on the endogenous gene. This article will attempt to describe some emerging views, first of RNA-triggered PTGS and then of RNA-triggered TGS, while highlighting the many mechanistic questions that remain to be resolved.

0168-9525/99/$ – see front matter © 1999 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(99)01818-1