Received: 29 March 2018 | Revised: 24 June 2018 | Accepted: 2 July 2018

DOI: 10.1002/gepi.22151

REVIEW ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

The evidential statistical paradigm in genetics

Lisa J. Strug

Program in Genetics and Genome Biology, The Hospital for Sick Children, The Centre for Applied Genomics, The Hospital for Sick Children, Division of Biostatistics and Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Correspondence
Lisa J. Strug, PhD, The Centre for Applied Genomics and the Program in Genetics and Genome Biology, The Hospital for Sick Children Research Institute, 686 Bay Street, Room 12-9705, Toronto, ON, Canada M5G 0A4.
Email: lisa.strug@utoronto.ca

Funding information
Cystic Fibrosis Canada, Grant/Award Number: 2626; Canadian Institutes of Health Research, Grant/Award Number: MOP-258916; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: 2015-03742; US CF Foundation, Grant/Award Number: STRUG17PO

Abstract

Concerns over reproducibility in research has reinvigorated the discourse on *P*-values as measures of statistical evidence. In a position statement by the American Statistical Association board of directors, they warn of *P*-value misuse and refer to the availability of alternatives. Despite the common practice of comparing *P*-values across different hypothesis tests in genetics, it is well-appreciated that *P*-values must be interpreted alongside the sample size and experimental design used for their computation. Here, we discuss the evidential statistical paradigm (EP), an alternative to Bayesian and Frequentist paradigms, that has been implemented in human genetics studies. Using applications in Cystic Fibrosis genetic association analyses, and describing recent theoretical developments, we review how to measure statistical evidence using the EP in the presence of covariates, model misspecification, and for composite hypotheses. Novel graphical displays are presented, and software for their computation is highlighted. The implications of multiple hypothesis testing for the EP are delineated in the analyses, demonstrating a view more consistent with scientific reasoning; the EP provides a theoretical justification for replication that is a requirement in genetic association studies. As genetic studies grow in size and complexity, a fresh look at measures of statistical evidence that are sensible amid the analysis of big data are required.

KEYWORDS

foundations of statistics, inference, likelihood paradigm, multiple hypothesis testing, statistical evidence

1 | INTRODUCTION

On February 2016, the American Statistical Association (ASA) Board of Directors published a position statement on "Statistical Significance and *P*-values" (Wasserstein & Lazar, 2016). They determined that silence on the misunderstanding and misuse of *P*-values for statistical inference was no longer an option. The board saw these misuses contributing to the "reproducibility crisis," a topic that was occupying pages of some of the most highly cited scientific journals (e.g., Nuzzo, 2014). The statement was intended to be accessible to nonstatisticians, but did not include any new information beyond the original arguments made by Fisher (1926) and Neyman and Pearson (1933) as they advocated for their significance and hypothesis testing paradigms, respectively. Statistical inference practiced today using hypothesis testing and confidence interval estimation, referred to as Frequentist statistics, is a patchwork of Fisherian *P*-value calculations and Neyman–Pearson hypothesis testing error concepts without any foundational justification, and was opposed to by both individual schools of thought (Fisher, 1956; Neyman & Pearson, 1933; Goodman, 2016). This approach has been made popular through textbooks intended to guide statistical practice and through application in research articles, and ultimately gave rise to "bright line" present day practice where $P < 0.05$ is equated with importance (Nuzzo, 2014).

The ASA statement, which has been summarized in the popular literature (http://fivethirtyeight.com/features/statisticians-found-one-thing-they-can-agree-on-its-time-to-stop-misusing-p-values/), emphasized that the *P*-value does not "tell you the size of an effect, the strength of the evidence, or the importance of a result." The statement was accompanied by more than 20 additional commentaries, many of which highlight the misuse rather than the index itself as the problem, where a focus on achieving some bright line value such as $P < 0.05$ compromises the integrity of scientific findings. Details on alternative approaches to the *P*-value were limited, although the statement noted that some statisticians augment or replace *P*-values with "alternative measures of evidence, such as likelihood ratios or Bayes Factors" which (among other alternatives) may more directly address the goal of measuring the strength of statistical evidence.

Until recently, the limitations of *P*-values have largely been overlooked by the scientific research community despite theoretical advances in the foundations of statistics beyond Fisher and Neyman and Pearson (e.g., Savage, 1972; Royall, 1997; Evans, 2015; Vieland, 2017). But *P*-value practices are now increasingly being scrutinized due to scientific reproducibility concerns. Goodman (2016) points to genomics as providing an example of a discipline that has deviated from convention, adopting a more stringent threshold for statistical significance in genome-wide association studies ($P < 5 \times 10^{-8}$; Dudbridge & Gusnanto, 2008). Indeed, this deviation from convention was even prominent in the genome-wide linkage literature, where in 1995, Lander and Kruglyak (1995) warn that "*Adopting too lax a standard guarantees a burgeoning literature of false positive linkage claims...Scientific disciplines erode their credibility when a substantial proportion of claims cannot be replicated...*" The size of genomic data and the nature of genome-wide studies has necessitated this alternative threshold. However, the alternative thresholds remain predicated on the *P*-value as the measure of evidence (Royall, 1997) and is based on conventional multiple hypothesis testing arguments for the purpose of maintaining a family-wise error rate < 0.05 (Dudbridge & Gusnanto, 2008). The threshold has been implemented broadly in the field because the number of statistical tests carried out in a genome-wide association study (GWAS) are in plain sight. This is in contrast to other fields where the number of analyses implemented are not always obvious, and restricted reporting results in a lack of reproducibility.

In current statistical practice, our field has two seemingly independent requirements: Adjusting for multiple hypothesis testing and independent replication (Vieland, 2001; Chanock et al., 2007). The replication requirement is consistent with Fisher's original recommendation that "significance" represents observations worthy of follow-up (Fisher, 1926). But Fisher also saw statistical analysis with "significance testing" as a fluid exercise with each particular observation contextualized against current evidence and ideas, strongly opposing the use of *P*-values for "automatic inference" (Fisher, 1956; Goodman, 2016). The ASA statement echoes Fisher's concerns: "By itself, a *P*-value does not provide a good measure of evidence for a model or hypothesis."

P-values are not comparable under different experimental conditions (Royall, 1997), and this is relevant in genomics as well. *P*-values need to be interpreted alongside the sample size used to calculate them, yet, for example, this has been largely dismissed in gene-based testing or meta-analysis of variants across different genotyping platforms. Similarly, *P*-values for SNPs with different minor allele frequency (MAF) are compared at face value, although the SNPs do not provide the same amount of information a priori. Unrelated to sample size, information on direction of effect inform replication studies to compute one-sided tests of significance, but not all investigators have the same information highlighting the *P*-value dependence on factors extraneous to the observed data itself (Li et al., 2015). Frequentist statistics conflates *P*-value practices with Type I error and decision theoretic concepts (Section 1.4, Box 1) and limits our ability to explore the data (Section 3.4). As genomic data analyses grow in size and complexity—genome-wide interaction studies, integration of data across several experiments, large population-based cohorts—alternative metrics that provide a more objective first-line quantification of statistical evidence that includes only what the data themselves supply (Blume, 2002), may be more suitable. The limitations of *P*-value procedures have been reviewed elsewhere (e.g., Royall, 1997). The focus here will be to review an alternative method for evidence measurement in genetic studies directly from likelihood ratios (LRs; Royall, 1997), coined the evidential paradigm (EP) by Vieland and Hodge (1998).

The outline of this review is as follows: Section 1.1 will discuss the interpretation of LRs for simple hypotheses and Section 1.2 will demonstrate how LRs are used to measure evidence in the EP. Section 1.3 provides the statistical properties that justify direct evidence measurement from LRs. Section 2 contrasts evidence measurement between the Frequentist and EP paradigms. Lastly, Section 3 applies the EP to genetic association analysis in cystic fibrosis (CF).

BOX 1 The EP *decouples* evidence measurement (LR) from error probabilities, whereas the Frequentist paradigm couples the concepts.

	Evidential paradigm	Frequentist paradigm
Evidence for two simple hypothesized values of θ	Observed LR, $\frac{L_{\text{obs}}(\theta_1)}{L_{\text{obs}}(\theta_0)}$	P-value $= P_0\left\{\frac{L(\theta_1)}{L(\theta_0)} \geq \frac{L_{\text{obs}}(\theta_1)}{L_{\text{obs}}(\theta_0)}\right\}$
Error favoring θ_1 when θ_0 is true	$M_0(n, k) = P_0\left\{\frac{L(\theta_1)}{L(\theta_0)} \geq k\right\} \leq \frac{1}{k}\ \forall\ n$	$\alpha = P_0\left\{\frac{L(\theta_1)}{L(\theta_0)} \geq c\right\}$
Strong evidence	$\frac{L_{\text{obs}}(\theta_1)}{L_{\text{obs}}(\theta_0)} \geq k$	$p \leq \alpha$ $\Leftrightarrow$ $\frac{L_{\text{obs}}(\theta_1)}{L_{\text{obs}}(\theta_0)} \geq c_{\alpha,n}$
Intervals[a] (e.g., mean of a normal distribution)	$1/k$ Likelihood Interval $\bar{x} \pm \sqrt{2 \log k}\,\sigma/\sqrt{n}$ $\frac{1}{8}$LI $\Leftrightarrow$ 95.9% CI	$(1 - \alpha)100\%$ Exact Confidence Interval $\bar{x} \pm Z_{\alpha/2}\sigma/\sqrt{n}$ 95% CI $\Leftrightarrow \frac{1}{6.67}$LI
Other errors to minimize for study planning	$M_1(n, k) = P_1\left\{\frac{L(\theta_1)}{L(\theta_0)} \leq \frac{1}{k}\right\}$ $W_1(n, k) = P_1\left\{\frac{1}{k} < \frac{L(\theta_1)}{L(\theta_0)} < k\right\}$ $W_0(n, k) = P_0\left\{\frac{1}{k} < \frac{L(\theta_1)}{L(\theta_0)} < k\right\}$	Type II error $\beta = P_1\left\{\frac{L(\theta_1)}{L(\theta_0)} \leq c_{\alpha,n}\right\}$[b]
Relationships	Strong evidence, k, and error $M_0(n, k)$ are decoupled	Strong evidence, $p \leq \alpha$ and error α are coupled

[a] There is a relationship, beyond the normal distribution, between exact confidence intervals and likelihood intervals but confidence intervals are also coupled with the Type I error probability.

[b] The Type II error, β, is analogous to $M_1(n, k) + W_1(n, k)$ but for fixed α, n. Power $(1 - \beta)$ is defined for a fixed α, therefore, although similar in spirit, is always greater than the probability of strong evidence at conventional Type I error levels and therefore the two represent different quantities. There is no concept of controlling weak evidence in the Frequentist paradigm.

1.1 | The role of likelihood ratios in statistics

For a comprehensive review of the EP, see Blume (2002). Briefly, given a set of observations and a probability model for the data, the likelihood function provides a mathematical representation of the evidence within the data. (For now we assume that the likelihood corresponds to a model that contains the true generating distribution of the data, although this assumption will be relaxed in Section 3.) The EP is the statistical paradigm that measures the strength of evidence directly from the ratio of likelihoods at different hypothesized values for a parameter of interest, therefore incorporating *only* data and a probability model into the strength of evidence assessment. A natural question arises: Why would one want to restrict evidence measurement to interpretation through data and a probability model?

To begin to answer that question, consider the simple example introduced in Royall (1997) of the diagnostic test, where one observes a positive test result, $x = 1$, for a given disease. Suppose there are two models representing the probability of disease, $\{P(X|A), P(X|B)\}$, as in Table 1, where A represents the presence of disease, B the absence and $P(B) = 1 - P(A)$.

Observation of a positive test result is empirical evidence of which model this realization is coming from. Royall suggests after observing a positive test result a physician may conclude: (a) The patient probably does not have the disease; (b) the subject should be treated for the disease; or (c) this positive test result is evidence that the patient has the disease. All three of these questions are in the realm of statistics, but are they all correct conclusions given the positive test result and the model? Conclusion (a) may be correct depending on the prior probability of having the disease, $P(A)$, which can be determined using Bayes' Theorem, where $P(A|$

TABLE 1 Properties of a diagnostic test for disease

	$x = 1$	$x = 0$
$P(X\|A)$	0.95	0.05
$P(X\|B)$	0.02	0.98

Under hypothesis A the disease is present, under hypothesis B the disease is absent, and the observations $x = 1$ or 0 represent a positive or negative test result, respectively.

$X = 1) = 0.95P(A)/[0.95P(A) + 0.02(1 - P(A))]$. For small $P(A)$, the patient probably does not have the disease, for large $P(A)$ the conclusion would be wrong. Whether the patient should be treated for the disease (Conclusion [b]) would also depend on $P(A)$ as well as other factors such as costs (loss functions). As for Conclusion (c), regardless of $P(A)$ or losses, interpreting a positive test result as evidence that the disease is not present would be wrong because it violates statistical reasoning, namely the *Law of Likelihood*.

The Law of Likelihood (Hacking, 1965) states:

If the hypothesis A implies that a random variable X takes the value x is $P(x|A)$, while hypothesis B implies that the probability is $P(x|B)$, then the observation $X = x$ is evidence supporting A over B if and only if $P(x|A) > P(x|B)$ and the likelihood ratio (LR), $\frac{P(x \mid A)}{P(x \mid B)}$, measures the strength of that evidence.

Since $P(x = 1|A) > P(x = 1|B)$ and the LR is $0.95/0.02 = 47.5$, this indicates that a positive test result provides support for the presence of disease that is 47-fold greater than no disease. In the Bayesian framework, this likelihood ratio is the factor by which the prior odds are updated to produce the posterior odds. Given the prior probabilities $P(A)$ and $P(B)$, the data produces posterior probabilities $P(A|X)$ and $P(B|X)$. Kass and Raftery (1995) note that since the posterior is a transformation of the prior through the data (for any prior specified), the function by which the prior is updated is a representation of the evidence from the data. Converting to the odds scale demonstrates that for simple versus simple hypotheses

$$\frac{P(A|X)}{P(B|X)} = \frac{P(x|A)}{P(x|B)} \times \frac{P(A)}{P(B)} \tag{1}$$

and therefore the LR is the factor by which the prior odds are updated to produce the posterior odds, and has been referred to as "the weight of the evidence" (Good, 1985).

This simple example does not capture the complexity involved in most real data analysis problems where there are multiple parameters in the model. To weigh the evidence for different hypothesized values of the parameter of interest in the presence of additional unknown parameters, these *nuisance* parameters must be accounted for. If they are integrated out with specification of prior distributions, then the ratio is the Bayes' factor. To use the Bayes' factor therefore, prior probability distributions must be chosen and then one must determine how sensitive the Bayes' factor is for measuring evidence strength in the data to that choice, limiting its desirability for some (Kass & Raftery, 1995). For an illustration of justifying prior probability choice in genetics, see the supplementary information from (Burton et al., 2007). Alternatively, nuisance parameters can be eliminated through maximization procedures (Pawitan, 2001; Royall, 2000). Nonetheless, the relationship in Equation (1) highlights that the LR has meaning unto itself and is the quantity that represents *what the data say*, alone, in the absence of prior probabilities or loss functions. The methodology that enables direct inference from the LR is the EP, and several recent advances in methodology development have made the EP paradigm accessible and applicable in genomics.

1.2 | Measuring evidence for simple versus simple hypotheses

Formally, we will use the following notation x as a realization of a random variable X with a probability distribution $f(.;\theta)$. (Assume $\theta \in \Theta$ is a scalar although all results are generalizable to $\boldsymbol{\theta}$ being a fixed-dimensional multiparameter vector.) For two simple hypothesized values about the unknown parameter θ, H_1: $\theta = \theta_1$ and H_0: $\theta = \theta_0$, and for $L(\theta) \propto f(x;\theta)$, the Law of Likelihood (Hacking, 1965) specifies that the $\text{LR} = L(\theta_1)/L(\theta_0)$ measures the strength of evidence in favor of H_1: $\theta = \theta_1$ relative to H_0: $\theta = \theta_0$. Evidence can be generated in favor of either hypothesis rather than having inference centered on disproving H_0: $\theta = \theta_0$. If $L(\theta_1)/L(\theta_0) \geq k$, one has strong evidence favoring H_1 over H_0, if $L(\theta_1)/L(\theta_0) \leq 1/k$ one has strong evidence favouring H_0 over $\text{H}_{1,}$ and for $1/k < L(\theta_1)/L(\theta_0) < k$ one has weak evidence where the data does not produce sufficiently strong evidence in favor of either hypothesis.

Plotting the likelihood function provides a graphical display to examine all possible pairwise comparisons for θ, demonstrating which hypotheses are best supported by the data. Consider an example in CF, an autosomal recessive disease caused by mutations in the *cystic fibrosis transmembrane conductance regulator* (*CFTR*). Suppose a pharmaceutical company is interested in the proportion, θ, of CF individuals within a given country who carry the most common *CFTR* genotype, p.Phe508del homozygosity. The proportion is thought to be 50%, but if it is greater than 50% then the pharmaceutical company sees benefit in developing a mutation-targeted therapy. Figure 1 displays the likelihood function for the proportion of p.Phe508del/p.Phe508del genotype carriers in a given CF population, with $n = 1{,}000$ CF individuals sampled and $x = 510$ observed p.Phe508del homozygotes. The likelihood is standardized such that the likelihood at the maximum likelihood estimate (MLE), $\hat{\theta}$, is 1.0. The relative evidence for the proportion of p.Phe508del

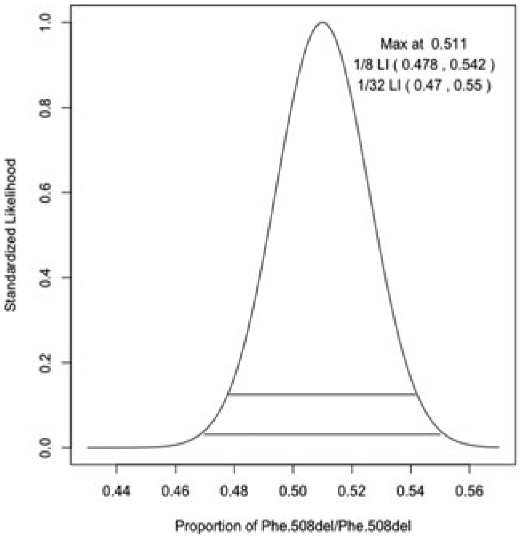

FIGURE 1 Standardized likelihood function for the proportion of p.Phe508del homozygotes, $n = 1{,}000$, $x = 510$ p.Phe508del/p.Phe508del observed. This standardized likelihood function provides a graphical representation of all possible likelihood ratios, with the 1/8 and 1/32 likelihood intervals (LI) providing the values for θ that are consistent with the data at the $k = 8$ and 32 evidence level. The value of 0.51 is more than 8-fold greater supported over values less than 0.48, but the data provides only weak evidence supporting 0.51 over 0.50

homozygotes, θ, is represented by the ratio of the likelihoods at any two points on the curve and the 1/8 and 1/32 likelihood intervals (LI) represent the values of the parameter θ that are consistent with the data at that k-evidence level, where the $1/k$ LI is defined as (Pawitan, 2001) {all θ where $L(\hat{\theta})/L(\theta) \leq k$} (Box 1).

There are several discussions on benchmarks for choosing k (Royall, 1997; Edwards, 1984) as there are for Bayes' factors (Kass & Raftery, 1995), and the choice of k can be experiment-specific. For genome-wide linkage studies, lod scores ($\log_{10}$ LR) of 3 have been justified as representing strong evidence, corresponding to a $k = 1{,}000$ (Chotai, 1984; Morton, 1955; Strug & Hodge, 2006a). Genome-wide lod score thresholds of 3 originated from arguments that assumed the LR was not maximized but rather calculated from two simple hypotheses (a recombination fraction of 0.5 vs. an alternative predetermined value near 0), with justification based on a combination of Wald's sequential testing theory (Wald, 1945) and the intent to maintain a high posterior probability of linkage when linkage was declared (Chotai, 1984; Morton, 1955). Regardless of the choice of k, a LR $= k$ represents the same evidence strength from experiment to experiment for several reasons (Royall, 1997), among which is that it is the exact factor by which the prior probability ratio is changed (Equation (1)).

Returning to our example, the data provides only weak evidence supporting the value of 0.51 over the value of 0.50, with the $L(0.51)/L(0.50) = 1.22$. If one observes weak evidence from the data, that is an undesirable result, as it is inconclusive about which hypothesis is better supported. One would want a study to be designed such that the probability of observing weak evidence is small (Strug, Rohde, & Corey, 2007). Another undesirable result is to observe strong evidence in favor of the wrong hypothesis, that is, observing misleading evidence; e.g. the $LR = L(0.51)/L(0.50) \geq k$ but the true $\theta = 0.50$. Since, on measuring evidence strength with the LR one is unaware of whether the result is misleading, it is imperative to ensure that this occurs with low probability.

The probabilities of misleading ($M_i(n, k)$), weak ($W_i(n, k)$) and strong evidence ($S_i(n, k)$) sum to 1 under the assumed hypothesis ($i = \mathrm{H}_0$ or H_1; e.g., Equation (2) under H_0), with the n and k arguments reflecting the probabilities' dependence on the sample size and evidence level for some $k > 1$ (Strug et al., 2007),

$$1 = P_0\left(\frac{L(\theta_1)}{L(\theta_0)} \geq k\right) + P_0\left(\frac{1}{k} < \frac{L(\theta_1)}{L(\theta_0)} < k\right) + P_0\left(\frac{L(\theta_1)}{L(\theta_0)} \leq \frac{1}{k}\right) = M_0(n, k) + W_0(n, k) + S_0(n, k). \tag{2}$$

$M_0(n, k)$ is analogous to the Frequentist Type I error, with one fundamental difference: The probability of misleading evidence is not fixed by design at values such as 0.05 or 5×10^{-8}. (Box 1 and Section 1.4 demonstrate why fixing the Type I error is not desirable.) Allowing $M_i(n, k)$ to vary is justified by two properties of likelihood functions that ensure the probability of misleading evidence always remains small and that both misleading and weak evidence can be further controlled by the sample size.

1.3 | Fundamental properties of likelihood functions

Likelihood functions have two fundamental performance properties that assure reliable inference from the LR (Royall, 2000; Royall & Tsou, 2003). These will be reviewed in this section and will be referred to frequently throughout the article. For simplicity again consider θ as a scalar parameter, e.g., log(Odds Ratio), but all results in this section generalize to fixed-dimensional multiparameter models, say $f(\cdot; \theta, \gamma)$ for $\gamma, \theta \in \mathbb{R}$. Assume θ_0 is the true value for θ. Then for any false value $\theta \neq \theta_0$, the first

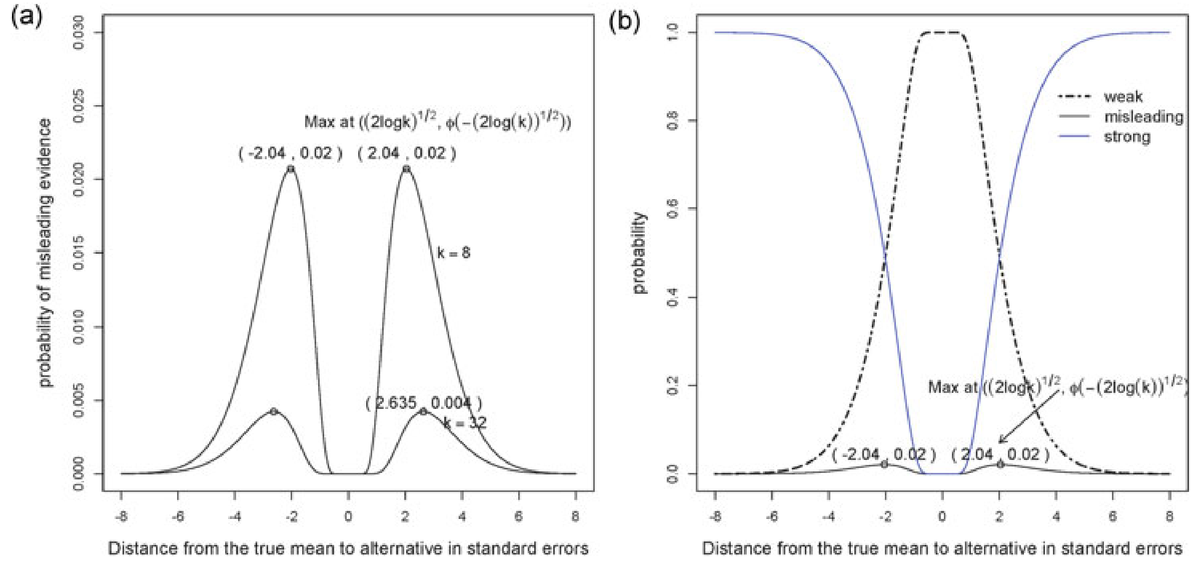

FIGURE 2 The bump function and relationships between EP error probabilities for a normal mean. Probabilities displayed as a function of the distance between the two hypothesized values for the mean in standard error units. Maximum probability of misleading evidence is 0.02 at $\Delta = 2.04$ for $k = 8$ and 0.004 at $\Delta = 2.63$ for $k = 32$, for any n. (a) Probability of misleading evidence, $k = 8$ and $k = 32$ displayed. (b) Probabilities of weak, strong and misleading evidence; $k = 8$. Weak evidence probabilities are close to 1 for small parameter differences; strong evidence probabilities increase as differences increase. EP: evidential paradigm

important property of a likelihood function is that the evidence will eventually support θ_0 over any other θ by an arbitrarily large factor:

$$\text{Property 1:} \quad P_0\left(\frac{L(\theta_0)}{L(\theta)} \rightarrow \infty \text{ as } n \rightarrow \infty\right) = 1. \quad (3)$$

Property 1 implies that the probability of eventually obtaining strong evidence in favor of the true value when interpreting evidence from the LR is assured. Moreover, this *consistency* property makes way for sample size estimation procedures since increasing sample size would therefore drive the probabilities of weak and misleading evidence to 0 (Equation (2); Strug et al., 2007).

This property implies that sample size choice can be used to ensure the LR provides small probabilities of misleading (and weak) evidence. However, bounds on misleading evidence probabilities that are independent of sample size also exist. First, for any genuine likelihood function or in other special cases (e.g., partial likelihood; Eddings, 2003) the probability of observing misleading evidence for any sample size and any specified hypothesis is bounded by $1/k$ (Royall, 1997, 2000); this feature is referred to as the *Universal Bound*.

The universal bound on the probability of misleading evidence, $M_i(n, k) \leq \frac{1}{k}$, $i = 0, 1$ is a crude bound, and in many settings one can do much better. When $f(x; \theta)$ is normally distributed, or in large samples under quite general regularity conditions (Knight, 2000), $M_i(n, k)$ can be described by a bump function (Figure 2), where for H_0 true

$$\text{Property 2:} \quad P_0\left(\frac{L(\theta_1)}{L(\theta_0)} \geq k\right) \rightarrow \Phi\left(\frac{-c}{2} - \frac{\log k}{c}\right), \quad (4)$$

where $k > 1$, Φ is the standard normal cumulative distribution function and c is proportional to the distance between hypothesized values of the mean θ_1 and θ_0, with $c = \Delta(\sqrt{n}/\sigma)$ (Royall, 2000), and the maximum of the bump function is $\Phi(-\sqrt{2 \log k})$. When the distance Δ is measured in units of the standard error, the probability of misleading evidence is independent of sample size at a fixed c.

The bump function describes the probability of misleading evidence for any sample size, n, under the normal model, or for general one-parameter models and higher dimensional models when the sample size is large (Royall, 2000). The bump function indicates that the probability of observing misleading evidence tends to 0 when the distance between the two hypothesized values increases, and the maximum of the bump function is $\Phi(-\sqrt{2 \log k})$ at $\Delta = \sqrt{2 \log k}$ (Figure 2a). The probability of observing misleading evidence is also near 0 when the distance between the two hypothesized values is very small. For these small Δ, the data is insufficient to produce strong evidence in favor of either hypothesis and corresponds to an effect size difference that will produce weak evidence with high probability (Figure 2b). This second property (Equation (4)) of likelihood functions—that the probability of misleading evidence is described by the bump function, $\Phi((-c/2) - (\log k/c))$ and is bounded,—along with the first consistency property (Equation (3)), ensure that with high probability we will get evidence in favor of the true value and that the probability of strong evidence in favor of a false value is low and controllable with the sample size. This is without fixing the error probability at some "bright line" value such as 0.05 or 5×10^{-8}.

2 | CONTRASTING THE PARADIGMS

The probability of misleading evidence under H_0, $M_0(n, k)$, is similar to the concept of a Type I error but differs in a fundamental way. Box 1 defines these error concepts and highlights that the fundamental difference between the paradigms is that in the frequentist framework the concept of evidence strength is *coupled* with the value at which the Type I error rate is fixed (Blume, 2002; Strug & Hodge, 2006a); this coupling results in many of the limitations of current statistical practice (see example below). In contrast, in the EP evidence strength k and error probabilities are decoupled.

The implication of the coupling is best illustrated by a simple example. Consider again the example of the pharmaceutical company interested in the proportion, θ, of individuals with CF who carry the most common causal *CFTR* genotype, p.Phe508del/p.Phe508del. Let the null hypothesis be H_0: $\theta_0 = 0.5$ and consider two values for an alternative hypothesized value, H_1: $\theta_1 = 0.51$ and 0.501. The likelihood for θ is $L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ and the LR can thus be expressed as $\frac{L(\theta_1;x)}{L(\theta_0;x)} = (\frac{\theta_1}{\theta_0})^x(\frac{1-\theta_1}{1-\theta_0})^{n-x}$ where n is the number of individuals sampled from the CF population, and x are the number of individuals observed to be p.Phe508del/p.Phe508del. According to the EP, one concludes strong evidence favouring θ_1 over θ_0 when $\frac{L(\theta_1;x)}{L(\theta_0;x)} \geq k$ for any n; k is defined by the investigator (Edwards, 1984; Royall, 1997). With some algebra, this implies that one observes strong evidence favoring θ_1 whenever $x \geq \frac{\log k - n\log[2(1-\theta_1)]}{\log(\frac{\theta_1}{1-\theta_1})}$. Table 2 provides the values of n, x, and x/n where one can conclude the data favors either $\theta_1 = 0.51$ or 0.501 with evidence strength of $k = 8$. For $\theta_1 = 0.51$, when there is $n = 100$ individuals sampled with CF one would require $x = 102$ individuals homozygous for p.Phe508del before the data could provide strong evidence favoring $\theta_1 = 0.51$ over 0.50. That is, a sample size of 100 is insufficient to generate strong evidence (evidence of strength $k = 8$-fold) that can differentiate $\theta = 0.51$ from 0.50. As the sample size increases to $n = 10{,}000$ individuals, an observed proportion just over 0.51 can provide strong evidence favouring 0.51 over 0.5. As the difference between the two hypothesized values for the parameter of interest diminishes, it is more difficult to discriminate between the values for θ and from Table 2 one can see that even with 1,000 individuals, one cannot produce evidence of strength 8 that will support $\theta = 0.501$ over 0.50. Again, as n increases there is "more power" to discriminate between the hypothesized values. As k increases, one would require larger sample sizes to produce strong evidence for the same hypothesized parameter values, although the effect on the required proportion x/n as k increases is minimal (not shown).

Alternatively, consider the Frequentist paradigm as practiced today. As in Box 1, the Type I error is set to α and defined as

$$P_0\left(\frac{L(\theta_1;x)}{L(\theta_0;x)} \geq C_{n,\alpha}\right) = \alpha \qquad (5)$$

With some algebra, Equation (5) is

$$P_0\left(\frac{\frac{x}{n} - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \geq \frac{\frac{\log c_{n,\alpha} - n\log[2(1-\theta_1)]}{\log(\theta_1/1-\theta_1)} - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}\right) = \alpha$$

where $\frac{\frac{x}{n} - \theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \approx N(0, 1)$ using the normal approximation to the binomial and therefore

TABLE 2 EP analysis demonstrating values of sample size (n), number of observed p.Phe508del homozygotes (x) and proportion (x/n) for which strong evidence at $k = 8$ can be observed for $\theta_1 = 0.51$ versus $\theta_0 = 0.50$ (left) and $\theta_1 = 0.501$ versus $\theta_0 = 0.50$ (right)

	$\theta_1 = 0.51$		$\theta_1 = 0.501$	
n	***x***	***x/n***	***x***	***x/n***
100	102 (>n)[a]	NA	570 (>n)[a]	NA
1,000	557	0.557	1,020 (>n)[a]	NA
10,000	5,102	0.5102	5,525	0.5525
100,000	55,552	0.5055	50,570	0.5057
1,000,000	505,052	0.5050	501,020	0.5010

Note. NA indicates there is no value of x for the given n that could produce evidence of strength 8.

EP: evidential paradigm.

[a] The cells with (>n) indicate that the required x to demonstrate strong evidence is greater than n, which is not possible. As sample size increases, the required observed x/n needed to demonstrate strong evidence becomes less extreme. As θ_1 gets closer to θ_0, one needs larger n to produce evidence of k-fold.

$$\frac{\frac{\log c_{n,\alpha} - n\log[2(1-\theta_1)]}{\log(\theta_1/(1-\theta_1))} - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} = Q(1-\alpha)$$

with $Q(1-\alpha)$ representing the $100(1-\alpha)$th percentile of the standard normal distribution. From here, one can solve for $C_{n,\alpha}$ to determine the value at which $\theta_0 = 0.50$ is rejected in favor of $\theta_1 = 0.51$ or 0.501; this is in contrast to Table 2 where the required strength of the evidence, k, is fixed ($k = 8$ in Table 2). From Equation (5), to ensure the Type I error is α, $\theta = 0.50$ is rejected when $\frac{L(\theta_1;x)}{L(\theta_0;x)} \geq C_{n,\alpha}$ where $C_{n,\alpha} = \exp\{\log\frac{\theta_1}{1-\theta_1}(\frac{n}{2} + Q(1-\alpha)\frac{\sqrt{n}}{2}) + n\log 2(1-\theta_1)\}$. Table 3 demonstrates how the value at which $\theta_0 = 0.50$ is rejected in favor of θ_1, that is, $C_{n,\alpha}$, changes with the sample size and changes as the alternative hypothesis, θ_1, changes. Where for $n = 100$, values of θ_1 closer to $\theta_0 = 0.50$ require *smaller* "evidence strength" than for θ_1 values farther from $\theta_0 = 0.50$. For a given alternative θ_1, the value of $\frac{L(\theta_1;x)}{L(\theta_0;x)} = C_{n,\alpha}$ at which θ_0 is rejected increases then decreases as a function of sample size. For example, until $n = 10{,}000$ the required evidence favoring $\theta_1 = 0.51$ over $\theta_0 = 0.50$ is increasing and is greater than 1; however, by $n = 100{,}000$ the required evidence is decreasing and $\theta_0 = 0.50$ will be rejected even when the data overwhelmingly favors $\theta_0 = 0.50$ over $\theta_1 = 0.51$ ($\frac{L(\theta_1;x)}{L(\theta_0;x)} = 2.58 \times 10^{-73}$). This is a consequence of forcing the Type 1 error rate to be α and declaring strong evidence when $p \leq \alpha$, and the example highlights the counterintuitive impact of large sample sizes on inference in this paradigm.

3 | GENETIC ANALYSIS WITH THE EP

To make the EP of practical significance in genetics and beyond, several methodological advances have had to occur, a few of which will be discussed here and illustrated with examples from a genetic association study in CF. Section 3.1 addresses the fact that rarely does one have a simple model for the data that consists of a single parameter. A solution is needed to measure evidence about a parameter in the presence of nuisance parameters. Section 3.2 discusses how most models are approximations and do not represent the true generating function of the data. The theory up until this point has assumed that the chosen family of models contains the true distribution, which is an assumption that needs to be relaxed. Section 3.3 reviews recent approaches to measuring evidence when the hypotheses are not simple versus simple comparisons. Lastly, Section 3.4 acknowledges the large scale multiple hypothesis testing that is characteristic of genomic studies and provides a salient solution from the EP perspective.

TABLE 3 The required evidence strength $C_{\alpha,n}$ to reject the null hypothesis $\theta_0 = 0.50$ in favor $\theta_1 = 0.51$ (left) or $\theta_1 = 0.501$ (right)

	The $\frac{L(\theta_1;x)}{L(\theta_0;x)}$ required to favor θ_1 over $\theta_0 = 0.50$; i.e., $C_{n,\alpha}$	
n	**$\theta_1 = 0.51$**	**$\theta_1 = 0.501$**
100	1.36	1.033
1,000	2.31	1.10
10,000	3.63	1.36
100,000	6.77×10^{-5}	2.31
1,000,000	2.58×10^{-73}	3.63
10,000,000	$<10^{-100}$	2.68×10^{-73}

Required evidence changes as n or θ_1 changes, and as n increases one rejects $\theta_0 = 0.50$ when the data overwhelmingly favors $\theta_0 = 0.50$.

Individuals with CF who have the same *CFTR* mutations have variable disease severity, and other genes referred to as modifier genes contribute to this interindividual variation (Cutting, 2015). One modifier gene locus, encompassing *SLC26A9,* was identified through genome-wide association studies to contribute to CF intestinal and pancreatic disease (Miller et al., 2015; Sun et al., 2012; Blackman et al., 2013). The functional role of SLC26A9 is not completely understood, although several studies suggest, like CFTR, SLC26A9 is an anion channel that may enhance the functional expression of CFTR through physical interaction (Loriol et al., 2008; Ohana, Yang, Shcheynikov, & Muallem, 2009). In individuals with severe *CFTR* genotypes from the International CF Gene Modifier consortium ($n = 6{,}770$ including 901 sibling pairs; Corvol et al., 2015; Sun et al., 2012), Figure 3 demonstrates the association evidence using the Frequentist paradigm at the *SLC26A9* locus with intestinal obstruction at birth (meconium ileus), a CF complication present in ~16% of CF newborns (Dupuis et al., 2016). Here, generalized estimating equations with a logit link and an exchangeable covariance structure was implemented to account for sibling relationships in the data, and *P*-values are plotted from a Wald χ^2 test for β_1 from this model. In Sections 3.1–3.4, the EP is applied to the same data at this locus.

3.1 | Measuring evidence in the presence of nuisance parameters

The universal bound and the two performance properties of likelihood functions (LR consistency [Equation 3] and the bump function [Equation (4)]) define the necessary characteristics of a measure of evidence to ensure reliable evidence interpretation. With some exceptions (e.g., partial likelihood [Eddings, 2003]) the universal bound applies to

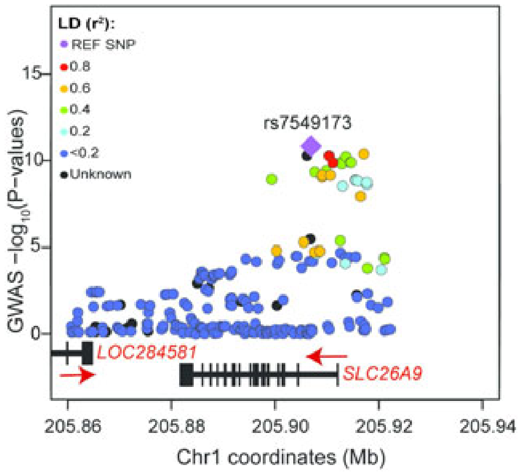

FIGURE 3 *P*-value-based analysis of the *SLC26A9* chromosome 1 CF modifier locus with meconium ileus using the LocusZoom software (http://locuszoom.sph.umich.edu/genform.php?type=yourdata). Analysis of CF participants from the International CF Gene Modifier consortium including siblings ($n = 6{,}770$). Analysis adjusted for consortium site and genotyping platform. *P*-values are from the Wald χ^2 test using generalized estimating equations with a logit link and an exchangeable covariance structure. CF: cystic fibrosis

genuine likelihoods; that is, likelihood functions based on the density function of a random variable. Consider the multiparameter model $L_n(\theta, \gamma)$, where for example the interest parameter θ is the log of the odds ratio (OR) and γ represents other covariates that are nuisance parameters. In the presence of nuisance parameters, genuine likelihoods would include marginal or conditional likelihood functions. Some likelihoods can be reparameterized such that they factor, where the interest and nuisance parameters are orthogonal (Anscombe, 1964), $L_n(\theta, \gamma) \propto L_n(\theta)L_n(\gamma)$. In this case, only the factor of the likelihood function containing the interest parameter is needed to construct the LR, which will have its probability of misleading evidence bounded by the universal bound. If such a factorization through reparameterization is achievable, a profile likelihood will provide that solution. Conditional, marginal, and these orthogonalizable likelihoods are not always available, but one can always calculate a profile likelihood. The profile likelihood maximizes the likelihood function with respect to the nuisance parameter for each fixed value of the parameter of interest (Kalbfleisch & Sprott, 1970), where $L_{pn}(\theta) = \max_\gamma L_n(\theta, \gamma) = L_n(\theta, \hat{\gamma}(\theta))$. In general, profile likelihoods are not genuine likelihood functions so the universal bound would not apply. However, in large samples, profile likelihoods have the two important performance properties of likelihood functions for evidential interpretation, where the probability of misleading evidence for likelihood ratios constructed from profile likelihoods can also be described by the bump function and are therefore bounded by $\Phi(-\sqrt{2\log k})$ (Equation (4)), and can be further sharpened by the sample size (Equation 3; Royall, 2000). Thus, profile likelihoods provide a general solution to how one should represent the strength of evidence about a parameter of interest in the presence of nuisance parameters, and they have been used in applications of the EP in genetics and genomics (Baskurt & Strug, 2018; Li et al., 2015; Strug et al., 2010; Zhong & Strug, 2018).

An R package to calculate profile likelihoods for many commonly used statistical models such as linear models, generalized linear models, proportional odds models, and mixed models, along with functions to produce their corresponding standardized likelihood plots (as in Figure 1) with likelihood intervals is available on CRAN (https://CRAN.Rproject.org/package=ProfileLikelihood). Plotting and analysis functions for linear and logistic regression tailored to genetic association studies are provided in the R package *EVIAN* (EVIdential ANalysis, https://CRAN.R-project.org/package=evian). These packages are used for EP analysis in Sections 3.1–3.4.

3.1.1 | Association of meconium ileus in CF at the *SLC26A9* locus: Profile likelihood ratios

There are 5,869 unrelated individuals with CF and 901 of their CF-affected siblings collected by the International CF Gene Modifier Consortium consisting of cohorts from France, Canada, Johns Hopkins University and the University of North Carolina/Case Western Reserve (Corvol et al., 2015). The cohort is of European origin and adjustment for population structure with principal components did not impact the conclusions so models unadjusted for population structure are presented here. Meconium ileus is a binary outcome, where $Y_i = 1$ if subject i was born with meconium ileus, and 0 otherwise. Assume a logistic regression model as in Strug et al. (2010), where $\log \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 G_{i1} + \gamma_2 Z_{i2} + \gamma_3 Z_{i3}$, $P_i = E(Y_i) = P(Y_i = 1)$, G_{i1}= 0, 1, or 2 alternative alleles at a SNP of interest for individual i; Z_{i2} is a categorical variable that indicates on which of the four genotyping platforms the patient's DNA was genotyped; and Z_{i3} is a categorical variable that represents to which of the four international collaborator cohorts the patient was recruited. Here, β_0, γ_2, and γ_3 are the nuisance parameters, and β_1 (the log(OR)) is the parameter of interest. Using the *EVIAN* R package, Figure 4 provides the likelihood intervals (LIs) across the chromosome *1 SLC26A9* region, where each LI is a summary of the likelihood function for a given SNP with the MLE for the OR noted in black on each LI. ORs that are less than 1 are inverted due to the asymmetry. The 1/100 LI is colored in green, the 1/1,000

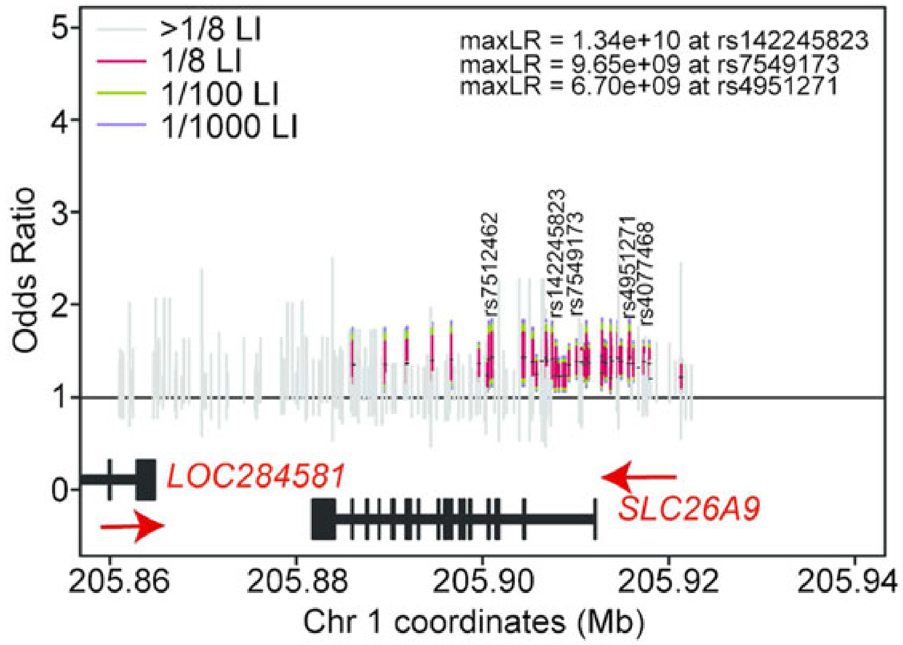

FIGURE 4 EP analysis of the *SLC26A9* chromosome 1 locus with mecomium ileus in CF. Analysis of CF participants of the International CF Gene Modifier consortium. All analyses adjusted for consortium site and genotyping platform. Analysis includes $n = 5{,}869$ unrelated individuals using a logistic regression likelihood. LIs for SNPs with max LRs> 1,000 noted in color. $1/k$ LIs, $k = 8$, 100, 1,000 displayed in red, green, and blue, respectively. MLE denoted in black on each LI. OR= 1 horizontal line noted as black solid line. Max LR and SNP name for the three SNPs with largest value noted on the figure, along with rs7512462 and rs4077488 which were identified in previous CF studies. CF: cystic fibrosis; MLE: maximum likelihood estimate

LI is colored in blue, and only SNPs for which the maximum profile likelihood ratio $\max \mathrm{LR}_{\mathrm{pn}} = \frac{L_{Pn}(\hat{\beta}_1)}{L_{Pn}(\beta = 0)} \geq 1,000$ are colored in the figure (using the benchmark from genome-wide linkage studies, $k = 1{,}000$), with all other SNPs displaying a gray 1/1,000 LI.

The LIs additionally convey an approximation to the shape of the likelihood function. Variants rs142245823, rs7549173, rs4951271, and rs4077468 have similar association evidence, with narrow LIs that are far removed from the OR = 1 line, demonstrating some of the strongest association evidence in the region (Table 4a). These variants, and others with similar strength of evidence, are located in a cluster just 5′ or in intron 1 of *SLC26A9* in a region where few variants favor OR = 1

TABLE 4 Summary statistics for the simple versus simple EP analyses for the unrelated (a) and related (b) CF samples

	MAF	max LR	MLE	1/8 LI	1/100 LI	1/1,000 LI	Robust factor *a*/*b*
a. Unrelated $n = 5{,}869$							
rs7512462	0.4089	317,462,247	1.3699	1.2345,1.5163	1.1763, 1.5953	1.1353, 1.6530	NA
rs142245823	0.4650	13,423,000,000	1.3980	1.2662,1.5474	1.2066, 1.6239	1.1645, 1.6827	NA
rs7549173	0.3951	9,652,137,218	1.3982	1.2632,1.5438	1.2037, 1.6201	1.1646, 1.6744	NA
rs4951271	0.4194	6,695,658,478	1.3980	1.2630,1.5514	1.2035, 1.6322	1.1615, 1.6870	NA
rs4077468	0.4129	4,679,117,865	1.3945	1.2598,1.5474	1.2005, 1.6280	1.1586, 1.6870	NA
b. Related $n = 6{,}770$							
rs7512462	0.4091	16,656,196	1.3221	1.1974,1.4597	1.1410, 1.5318	1.1040, 1.5872	0.975
rs142245823	0.4652	502,318,282	1.3526	1.2282,1.4896	1.1704, 1.5632	1.1324, 1.6157	0.966
rs7549173	0.3937	376,895,001	1.3494	1.2253,1.4861	1.1706, 1.5556	1.1326, 1.6078	0.970
rs4951271	0.4193	278,958,808	1.3424	1.2220,1.4783	1.1645, 1.5514	1.1266, 1.6034	1.004
rs4077468	0.4129	237,653,274	1.3458	1.2220,1.4821	1.1645, 1.5553	1.1266, 1.6075	0.995

Note. SNPs with the largest max LRs displayed, and statistics for variants rs4077468 and rs7512462 that have displayed previous association evidence with CF. Robust adjustment factor applied to the analysis of the related sample.

MAF: minor allele frequency; MLE: maximum likelihood estimate.

as a plausible value demarcating the associated region. Rs7512462 in intron 5 and rs4077468 have been demonstrated in previous *P*-value based studies as associated with CF pancreatic disease (Blackman et al., 2013; Miller et al., 2015) and therapeutic response (Strug et al., 2016).

3.2 | "All models are wrong, but some are useful." (Box, 1976)

In reality, most probability models are approximations and the data is, for example, not iid, or the wrong probability distribution has been assumed. Royall and Tsou (2003) show how (profile) likelihood functions can be made robust such that the probability of misleading evidence for the working model is asymptotically equivalent to the true model, achieving the two performance properties of true likelihood functions (Equations 3 and (4)) when the model applied to the data is different from the model that generated the data. The example in this section is the analysis of the full CF cohort, which includes siblings, introducing a departure from the independence assumption. In other settings more extreme model departures have been investigated such as the use of Poisson regression likelihood functions for the analysis of differential gene expression when the data is generated from a negative binomial distribution (Zhong & Strug, 2018).

There is a substantial body of literature on model misspecification, which is an issue for any statistical paradigm. In general, an investigator chooses a working model which enables inference about a parameter of interest, often using maximum likelihood estimation. MLEs are usually consistent when the model is correctly specified. In a working model (say *f*) for a set of observations where the observations come from another model (say *g*), the MLE obtained from *f* may not be a consistent estimate of the true parameter (θ) of interest, and the first property (Equation 3) required for EP interpretation may not hold. The MLE will be a consistent estimate of another parameter which makes *f* as close as possible to *g* (say θ*; Kent, 1982, Viraswami & Reid, 1998; White, 1982).[1] In the usual frequentist hypothesis-testing framework it is common practice to assume θ* is equal to θ and implement a robust hypothesis test by using, for example, a robust variance estimator for the MLE.

Royall and Tsou (2003) suggests it is not sufficient to assume θ* is equal to θ, rather that this assumption should be checked. Checking this assumption is equivalent to determining if the interpretation of the parameter in the working model would be the same under the true model (e.g., is one really making inference about an expected value; Blume, Su, Olveda, & Mcgarvey, 2007; Viraswami & Reid, 1998), and is analogous to ensuring property 1 (Equation 3). This condition has been confirmed for several working models and interest parameters. Royall and Tsou (2003) check this condition analytically for many commonly used one-parameter probability models, and Blume et al. (2007) showed that it holds for all generalized linear models in a regression setting when making inference about the mean parameter, as long as the mean structure is correctly specified. In this journal issue, Baskurt and Strug (2018) use simulation to ensure this condition is met for more complicated choices of *f* such as for using composite likelihood functions for genetic association with pedigrees of varying size and complexity. Therefore, several commonly used working models are available in the EP toolbox. If it is found that θ* is not equal to θ, it is recommended that one changes the working model (Freedman, 2006).[2] Alternatively, Equations 3 and 4 hold for (profile) empirical likelihood functions (Owen, 1988), suggesting that empirical likelihoods are always available for carrying out EP analysis and may even be more efficient than robust adjusted parametric likelihood ratios (Zhang, 2009).

The second property of likelihood functions (Equation (4)), that the probability of misleading evidence is described by the bump function and bounded by $\Phi(-\sqrt{2\log k})$ (Figure 2; Royall & Tsou, 2003), must also hold when the working model is not the true model. However, the probability of misleading evidence for misspecified models are not, in general, described by the bump function. Royall and Tsou (2003) derived a robust adjustment factor a/b, which is used to exponentiate the likelihood ratio: $(\mathrm{LR})^{a/b}$. This ratio a/b, is the ratio of the expected second derivative of the likelihood function to the expected square of the score function in a one parameter model, or equivalently the ratio of the asymptotic variance of the MLE to the asymptotic sandwich variance (Godambe, 1960). For a generalized linear model, this can be approximated by the ratio of the model-based variance estimate over the sandwich variance estimate of the regression parameter of interest calculated using standard statistical packages (Blume et al. 2007). As the working model gets closer to the true distribution, this ratio gets closer to 1, and if the model is correct, these two quantities will be equal (Bartlett's second identity; Ferguson, 1996) and the ratio will be exactly 1, in theory.

[1] This is in the sense of the Kullback–Leibler divergence. The asymptotic theory of this MLE from *f* was well studied in the literature. It has an asymptotic distribution that is Normal with mean θ* and the sandwich estimator as its variance (instead of the inverse Fisher information). The Wald, score, and the likelihood ratio tests were derived under the misspecified model *f* and robust versions of these tests were suggested since the distribution of these tests would not be the usual χ^2 distribution under the null hypothesis due to the model misspecification.

[2] Freedman (2006) also recognized the importance of this property, pointing out that even if the sandwich estimators provide asymptotically correct variances for MLEs under model misspecification, they do not correct the bias (θ*–θ).

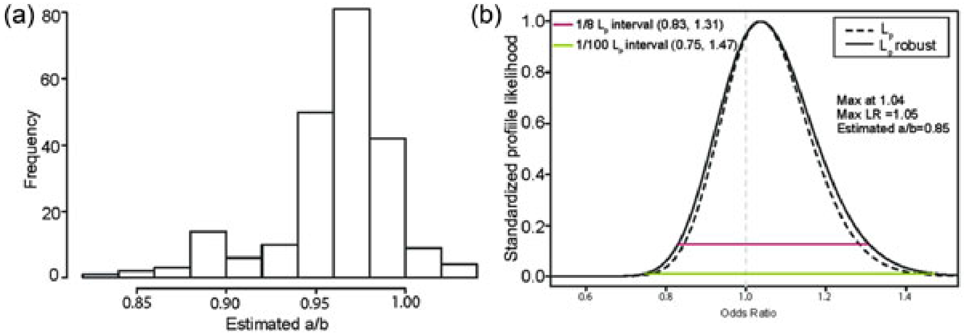

FIGURE 5 Estimated robust adjustment factor and its impact on association evidence. (a) The distribution of the estimated robust adjustment factor across the 222 variants of the *SLC26A9* locus; (b) the impact of the robust adjusted profile likelihood function on the association evidence for rs61814952. The robust adjustment factor $\frac{\hat{a}}{\hat{b}} = 0.85$ increases the tails of the likelihood function (L_p) (robust adjusted [solid line] versus unadjusted [dotted line]), widening the LIs. The likelihood function indicates that OR values around 1 are consistent with the data. OR: odds ratio

With this adjustment on the likelihood ratio, the probability of misleading evidence will behave as if one were using the true model. This robust adjustment factor corrects the tails of the likelihood function impacting the likelihood intervals, while the center of the likelihood (MLE) remains the same, highlighting why the first property is important (Figure 5). The robust adjustment factor has been implemented in the EVIAN R package for the linear and logistic regression models.

In this journal issue, Baskurt and Strug (2018) demonstrate how to compute robust adjusted composite likelihood functions for EP inference of genetic association in pedigrees, and this inference is compared to analyses using generalized estimating equations (Diggle, Liang, & Zeger, 1994). Zhong and Strug (2018) implement robust adjusted Poisson regression-based likelihood functions for differential gene expression analysis, where the robust adjustment factor protects from departures of the Poisson mean-variance assumption (i.e., over and under-dispersion). Returning to the example of meconium ileus in CF, to assess the impact of model misspecification consider applying the same logistic regression model from Section 3.1.1 that assumes independence to a data set that adds 901 siblings to the unrelated CF sample analyzed in Figure 4 (Section 3.2.1).

3.2.1 | Robust adjusted association analysis of related individuals with CF at the SLC26A9 locus

For each SNP at the chromosome 1 *SLC26A9* locus, the logistic regression model from Section 3.1.1 was fit; however, since siblings were included in the sample a robust adjustment to the LR was implemented in EVIAN to adjust for model misspecification (departure from assumed independence); this adjustment results in changes to the shape of the likelihood function and therefore all corresponding LRs and LIs. Figure 5a provides a histogram of the estimated adjustment factors, $\frac{\hat{a}}{\hat{b}}$, across the 222 SNPs analyzed in the region. The distribution for this data set shows the robust factor close to 1 for the majority of variants, indicating that the addition of the 901 siblings does not lead to a large departure from the model assumptions and consequently minimal adjustment to the likelihood. Choosing the SNP in the region with the largest model departure (rs61814952, $\frac{\hat{a}}{\hat{b}} = 0.85$; Figure 5b) demonstrates that inference without the adjustment factor would slightly overestimate the strength of the evidence, and that the adjustment factor increases the tails of the likelihood function which decreases the LRs and increases the width of the intervals. With respect to interpreting the strength of association evidence for this SNP rs61814952, an OR value of 1 is consistent with the data even at $k = 8$, with OR values around 1.04 providing similar support to values of 1; there is a lack of strong association evidence. In some circumstances, $\frac{\hat{a}}{\hat{b}}$ is greater than 1 which leads to a more concentrated likelihood function (Table 4b).

3.2.2 | Updating evidence in the EP

The robust adjusted joint analysis of the unrelated sample with their siblings (Table 4b) also highlights how to update evidence in the EP with additional data. Since evidence strength and error probabilities are decoupled (Box 1), updating the evidence with additional data simply involves multiplying the likelihood function by the likelihood of the new data sample. For a discussion on how this impacts error probabilities from the perspective of multiple hypothesis testing, see Section 3.4. Comparing Table 4a to 4b shows how the addition of the 901 patients impacts the evidence, where the MLEs are attenuated (as would be expected under winner's curse (Sun et al., 2011) and the LIs for the variants showing some of the strongest evidence strength, such as rs142245823, are narrower.

3.3 | EP analysis for composite hypotheses

Up until this point the focus has been on measuring evidence for simple versus simple hypotheses through the LR, as dictated by the law of likelihood. The likelihood function provides a graphical representation of all possible simple versus simple LRs, with the likelihood intervals providing a range of values for the parameter that are consistent with the data (Pawitan, 2001). The EP has been used in several genetic applications. Strug and Hodge (2006a, 2006b) cast linkage analysis in an EP framework, comparing the simple versus simple hypotheses of no linkage with a recombination fraction of 0.50 (H_0: $\theta = 0.50$) to linkage where H_1: $\theta = 0$, highlighting the implicit choice for two simple hypotheses in genetic linkage analysis (Morton, 1955). Gene expression provides another example in which there is a natural choice for a simple versus simple comparisons, given the convention of 2-fold differences considered as important (Bickel, 2012). The law of likelihood provides a guide for statistical practice rooted in the comparison of simple versus simple hypotheses, and Royall (2000) and Royall and Tsou (2003) has shown how this leads to reliable inference due to powerful properties for the probabilities of misleading, weak and strong evidence (Equations 3 and 4).

In current frequentist practice simple versus simple hypothesis comparisons are necessary in power calculations and sample size estimation procedures. However, for statistical testing one generally ignores the simple alternative hypothesis specified for power analysis, and compares a simple null hypothesis, H_0: $\theta = \theta_0$ to a composite alternative such as H_1: $\theta > \theta_0$ or H_1: $\theta \neq \theta_0$. Whether this is an informative and necessary comparison beyond all the simple versus simple comparisons that the likelihood provides, or whether this has just been popularized by practice and convention is debated (Blume, 2002, 2013; Royall, 1997).

Independently, Bickel (2012), and Zhang and Zhang (2013) tackled the problem of how evidence measurement could be accomplished for composite hypotheses in the EP, generalizing the Law of Likelihood (GLL), and concluding that it was only achievable when *both* hypotheses represent an interval or set of parameter values. Zhang and Zhang (2013) used the relevance of composite hypotheses in clinical trials as motivation for the GLL. The GLL states that the strength of evidence for one composite hypothesis over another should be measured by the ratio of the two likelihood functions each maximized over the set of parameter values defined by the two composite hypotheses, referred to as the generalized likelihood ratio (GLR). When one simple hypothesis is pitted against the entire sample space, e.g. $\frac{L(\hat{\theta})}{L(\theta_0)}$ (such as the case for classical frequentist likelihood ratio testing), then the probability of misleading evidence is not characterized by the bump function, and does not converge to 0 with increasing n, but rather equals the fixed value of twice the maximum of the bump function

$$\lim_{n\to\infty} P_0\left(\frac{L(\hat{\theta})}{L(\theta_0)} \geq k\right) = 2\Phi(-\sqrt{2\ln k})$$

for all $k > 1$ (Bickel, 2012; Blume & Choi, 2017). In contrast, Bickel (2012) showed that the GLR has the first important property of likelihood functions for the EP—that is, assuming some general regularity conditions, in large samples the GLR will eventually support the correct hypothesis and the probability of misleading evidence approaches 0. Li (2016) investigated the second important property of likelihood functions for the GLR motivated by genetic association studies with a null hypothesis region centered around the parameter value of no association, and the alternative hypothesis region representing the complement of the null hypothesis; that is H_0: $\Theta_0 = [\theta_0 - \varepsilon, \theta_0 + \varepsilon]$ and H_1: $\Theta_1 = \Theta_0^c = (-\infty, \theta_0 - \varepsilon) \cup (\theta_0 + \varepsilon, \infty)$, where ε is a small and positive constant $0 < \varepsilon < |\theta_0 - \theta_1|$. Deriving the probability of misleading evidence for the GLR under the null and alternative hypotheses, $M_0^c(n, k)$ and

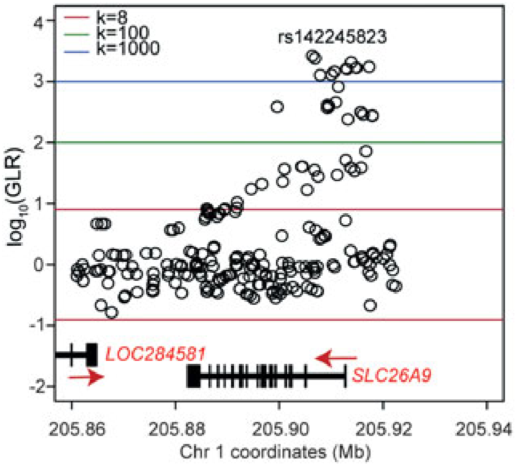

FIGURE 6 EP analysis with composite hypotheses in the unrelated CF participants of the International CF Gene Modifier consortium as in Figure 4 at the *SLC26A9* chromosome 1 locus with mecomium ileus. Analysis uses the generalized likelihood ratio (GLR) with null and alternative hypotheses defined as H_0: OR ∈ [0.87, 1.15] and H_1: OR ∈ H_0^c; $\log_{10}(\text{GLR}) < 0$ represents evidence favoring the null hypothesis. All analyses adjusted for consortium site and genotyping platform. CF: cystic fibrosis; OR: odds ratio

$M_1^c(n, k)$, for these nonoverlapping composite hypotheses that span the full parameter space, Li (2016) shows that $M_0^c(n, k)$ is bounded by $2\Phi(-\sqrt{2\log k})$ (twice the bound under the conventional simple hypothesis EP framework), while $M_1^c(n, k) < \Phi(-\sqrt{2\log k})$ (the same bound as in the simple hypothesis case).

3.3.1 | Genetic association of composite hypotheses for meconium ileus at the SLC26A9 locus

In Figure 6, the GLR is applied to the genetic analysis of meconium ileus with the 222 SNPs annotated to the *SLC26A9* region of chromosome 1 using the unrelated sample of 5,869 individuals with CF as in Figure 4 for the simple versus simple hypothesis analysis; H_0: OR $\in$ [0.87, 1.15] and H_1: OR $\in H_0^c$. Each dot on the plot corresponds to the GLR for a given SNP, with horizontal lines indicating evidence strength of $k = 8$, 100 and 1,000. GLRs < 1 provide evidence favoring the null region, while GLRs > 1 provide evidence favoring the alternative set. Comparing Figure 4 to Figure 6 shows association evidence in the same region of chromosome 1, although the strength of the evidence for a given SNP is smaller for the GLR than for the simple versus simple hypothesis comparisons. Figures 3 and 6 provide association evidence in the same region; however, analysis with the Frequentist paradigm cannot take advantage of the intuitive EP properties of updating evidence, error probability bounds and consequences for multiple hypothesis testing (Section 3.4). Moreover, unlike the *P*-value, the GLR can generate evidence that favors the null region, providing a way to demarcate the region of association; in Figure 6, by 205.88 Mb 3′ of *SLC26A9,* the majority of variants are providing evidence favouring the null region over the alternative (i.e., GLR < 1). The arbitrary nature of choosing the null region would be seen as a limitation by some, although these choices are required when alternative hypothesized values are specified for the parameter of interest in power calculations. For further discussion, see Li (2016).

3.4 | Study planning and multiple hypothesis testing

3.4.1 | Sample size and error probability estimation

Sufficient sample size should be planned under the EP to generate strong evidence, with correspondingly low probabilities of weak and misleading evidence (Figure 2). Estimating sample size by controlling Type I and II error probabilities results in insufficient estimates for EP inference (Strug et al., 2007). Power, although similar in spirit, is always greater than the probability of strong evidence at conventional Type I error levels and therefore the two represent different quantities (Box 1). If the goal of a study is to generate evidence about a parameter of interest—rather than simply choosing to reject (or not) the null hypothesis—then the sample size estimation procedure needs to reflect that goal and larger estimates are required. Formulae for EP sample size estimation, or error probabilities for a given sample size, are available for several data types (Strug et al., 2007). For the simple versus simple hypothesis case, these formulae are available for normal one and two sample comparison of means, linear combinations of regression coefficients, effects in repeated measures designs, one sample proportions, logistic regression coefficients, and for rate and survival data. Sample size estimation for composite hypotheses using the GLR are available in Li (2016), with methods for simple versus simply hypotheses of time to event end-points in Blume and Choi (2017).

3.4.2 | The impact of multiple hypothesis testing in the EP

A fundamental difference between the EP and frequentist paradigm is the decoupling of the evidence measure (the LR) from the error probabilities (Box 1). When one fixes the Type I error rate, one fixes the *P*-value (critical value) at which "importance" is concluded. The Type I error is adjusted by the effective number of hypothesis tests in, say a whole genome scan, to protect from Type I error inflation. This adjustment, in turn, adjusts the *P*-value (incorrectly interpreted as evidence strength) needed for significance (e.g., $P < 5 \times 10^{-8}$; Dudbridge & Gusnanto, 2008). By this reasoning, a *P*-value of 0.05 represents different evidence strength depending on whether one SNP is analyzed for association or 222 SNPs, while an LR of k has the same interpretation across different experimental designs (Equation (1)). The decoupling of evidence strength and error probabilities in the EP (Box 1) achieves an independence between study planning and evidence measurement and, since error probabilities are not fixed by design, allows one to ask what the impact of multiple testing is on the probability of obtaining misleading evidence. See Strug and Hodge (2006b) for a comprehensive discussion of multiple hypothesis testing for the EP in linkage analysis, which is generalizable to genetic association studies (Strug et al., 2010).

In essence, one could have multiple tests of a single hypothesis or single tests of multiple hypotheses. When comparing the evidence strength for rs142245823 in Table 4 calculated pre- and post- the addition of the 901 siblings, this was a multiple test of a single hypothesis. The theory of sequential testing (Wald, 1945) is concerned with how to adequately "spend" Type I error as one updates their inference over accumulating data; in practice, this is challenging to implement and is largely

TABLE 5 The probabilities of misleading, weak and strong evidence for alternative ORs (OR_1) of 1.10, 1.20, and 1.30 compared to OR = 1 as a function of k; $n = 5{,}869$

	$M_0(n, k)$			$W_0(n, k)$			$S_0(n, k)$		
OR_1	1.10	1.20	1.30	1.10	1.20	1.30	1.10	1.20	1.30
$k = 8$	0.0204	0.0085	0.0013	0.5384	0.0988	0.0122	0.4411	0.8927	0.9865
$k = 32$	0.0028	0.0028	0.0005	0.8075	0.1924	0.0253	0.1897	0.8048	0.9742
$k = 100$	0.0004	0.0010	0.0002	0.9302	0.2919	0.0417	0.0695	0.7070	0.9580
$k = 1{,}000$	0.0000	0.0001	0.0000	0.9965	0.5357	0.0989	0.0035	0.4642	0.9010

OR: odds ratio.

absent from the applied genetics literature. Single tests of multiple hypotheses occur when multiple SNPs are evaluated for association across, say a region (e.g., Figures 3 and 4) or a genome. In contrast to the probability of misleading evidence, both types of multiple hypothesis testing "spend" fixed Type I error.

For multiple tests of a single hypothesis if one does not fix the Type I error, the probability that an investigator will eventually observe strong misleading evidence if that investigator continues to collect data until the evidence supports a favorite incorrect hypothesis, is bounded by $1/k$ (Robbins, 1970)

$$P_0\left(\prod_{i}^{n}\frac{f_1(x_{i;}\theta_1)}{f_0(x_{i;}\theta_0} \geq k \ \text{ for some } \ n = 1, 2, \ldots\right) \leq \frac{1}{k}. \quad (6)$$

This bound holds whether one looks at the data a few or many times, or as in Wald's sequential sampling framework one looks at the data until a strong result is observed in either direction; Equation (6) is a scientific safeguard. In reality the actual probability of misleading evidence in a given situation is much smaller than $1/k$ (Strug & Hodge, 2006b). This result supports the analysis comparison of rs142245823 in Table 4a,b. Moreover, in combination with property one of likelihood functions (Equation 3), Equation (6) supports the collection and addition of more data to refine initial signals.

For genome-wide or regional association studies (single tests of multiple hypotheses), one is generally concerned with the family-wise error rate (FWER), the probability of observing at least one misleading result (or rejecting one truly null hypothesis) among all the variants assessed for association when none are true. Define the FWER across N SNPs in terms of the probability of misleading evidence,

$$\mathrm{FWER} = M_0(n, N, k) = P_0[(LR_1 \geq k) \cup (LR_2 \geq k) \cup \ldots \cup (LR_N \geq k)]$$

$\leq \sum_{j=1}^{N} M_0^{(j)}(n, k), j = 1, \ldots, N$ variants (Strug & Hodge, 2006b) with $M_0^{(j)}(n, k)$ being $M_0(n, k)$ for the jth variant. Given $M_0(n, k)$ is a planning probability, one in general would take $M_0^{(j)}(n, k)$ equal for all N and therefore

$$\mathrm{FWER} \leq NM_0(n, k) \quad (7)$$

which is a conservative upper bound and assumes independence across all SNPs. (For SNPs in linkage disequilibrium the bound is smaller: $\mathrm{FWER} \leq N_{\mathrm{eff}} M_0(n, k)$, where N_{eff} is the effective number of independent tests.) From Equation (7), the bound on the FWER is larger for more variants, but is a function of $M_0(n, k)$ which is not fixed and can be made as small as necessary by increasing the sample size (Equation 3).[3] In some situations, $M_0(n, k)$ is naturally very small such that multiplication by N results in a bound on the FWER that remains sufficiently small.

Returning to the CF example of Section 3.1, from the bump function in Equation (4) the maximum probability of misleading evidence for one of the $N = 222$ SNPs is $\Phi(-\sqrt{2\log(1000)}) = 0.0001$ for $k = 1{,}000$, and therefore the FWER for the region is bounded by $N \times \Phi(-\sqrt{2\log(1000)}) = 222 \times 0.0001 = 0.022$ when the SNPs are assumed independent. (Note here that the genome-wide linkage threshold of $k = 1{,}000$ was used although in reality k can be any experimenter-defined value, greater than 1.)

In reality, Property 1 (Equation 3) implies that the probability of misleading evidence is much smaller for this large $n = 5{,}869$. Using the bump function formula for profile likelihoods with an estimate of the variance from a previous study of rs7512462 with meconium ileus (Sun et al., 2012), the probability of misleading evidence comparing ORs of 1.10, 1.20, and 1.30 to 1.0 is provided in Table 5 as a function of k. By $k = 1000$, $M_0(n = 5{,}869, k = 1{,}000)$ is essentially 0 for a single SNP association analysis for any of the three alternative OR values considered (OR_1). Table 5 also provides the corresponding probabilities of weak and strong evidence under the null hypothesis, $W_0(n, k)$ and $S_0(n, k)$, respectively. As k gets larger, the probability of weak evidence also gets larger and the probability of strong evidence decreases, and

[3] $M_0(n, k)$ is also a function of k, but, in general, increasing k can be counterproductive since it only slightly reduces $M_0(n, k)$ while increasing the probability of weak evidence substantially (Table 5).

they do so similarly assuming the null or alternative hypothesis is true (not shown). Where, for $k = 1{,}000$ and OR_1 of 1.10 or 1.20, $W_0(n, k)$ and $S_0(n, k)$ are not at acceptable levels, demonstrating the trade-off between choosing a large initial k value.

For studies where the FWER is not sufficiently small, one can decrease the bound on the FWER by increasing the sample size (Equation (7)). Whether a larger sample size is chosen to reduce the bound on the FWER before conducting the study, or whether the additional data is added in a follow-up analysis as in a joint (two-stage) design, e.g., Table 4a,b (Skol, Scott, Abecasis, & Boehnke, 2006), results in the same LR (evidence strength), while two-stage joint designs result in equal or smaller probabilities of misleading and weak evidence (Strug & Hodge, 2006b; Strug et al., 2010). Although adjustment for multiple hypothesis testing and replication studies are independent concepts in standard Frequentist statistical practice and both are required in the genetics field for seemingly different reasons, in the EP increasing the sample size is the multiple hypothesis testing adjustment.

4 | DISCUSSION

The American Statistician's special issue call, "Statistical Inference in the 21st Century: A World Beyond $P < 0.05$" highlights the field's focus on change. Recent theoretical developments for the EP that were outlined here alongside their application to CF, provide an alternative to *P*-value procedures that are available for implementation in genetics. In summary, for EP analysis any reliable evidence function must demonstrate the two properties of likelihood functions: That the probability the evidence function will eventually favor the "true" parameter value over a false value is 1; and that the probability of misleading evidence is described by the bump function and/or is bounded. First, profile likelihood functions, although they are pseudo-likelihood functions, were shown to have the two properties of genuine likelihood functions. Consequently, profile likelihoods provide a general solution to measuring evidence in the presence of nuisance parameters. Second, likelihood functions can be made robust, such that when the working model is misspecified, the two properties of likelihood functions required for reliable evidence measurement are recapitulated. Third, despite significant debate concerning the relevance, recent approaches to measuring evidence for composite hypotheses with corresponding bounds on misleading evidence were presented. Lastly, the impact of multiple hypothesis testing on the EP was reviewed, showing that additional data provide an adjustment for multiple hypothesis testing. Arguably, the implications of multiple hypothesis testing in the EP are more consistent with scientific reasoning and the availability of big data.

EP methodology has found several applications in biomedical research, and the methodology tailored to these problems could readily be repurposed for genetic studies. These include applications in bioequivalence trials for the purpose of approving generic drugs (Du & Choi, 2015); noninferiority analyses in clinical trials (Wang & Blume, 2011); a general framework for clinical trials measuring evidence for composite hypotheses (Zhang & Zhang, 2013); and EP approaches for survival analysis using the theory of partial likelihood (Cox, 1975). In the latter case, Eddings (2003) demonstrated that LRs constructed from partial likelihoods satisfy properties 1 and 2 (Equations 3 and 4) for reliable EP inference and satisfy the universal bound.

Several areas of EP analysis in genetics require further attention. Use of the EP in applications where evidence must be compared across units of differing sample sizes (such as gene-based testing or tests based on sequence read counts) may have some of its most important contributions, and this requires further investigation. Given the growth of data in the field, estimating and controlling other error probabilities besides the FWER need to be delineated. Likewise, methods when the number of (nuisance) parameters (p) are large, and especially for $p > n$, require attention. Methodology development for rare variant analysis, as well as other set-based genomic analyses such as testing collapsed over functional annotations is required. For rare variant association analysis using classical testing procedures, collapsing methods across multiple variants or Fisher's exact test for single variants is generally used. Burden testing (e.g., Li & Leal, 2008) would be straightforward to incorporate into a given likelihood function through a covariate and carry out an EP analysis, although little work has been done in this area. Li et al. (2015) derive a conditional likelihood based on a logistic regression model for a 2×2 table of a single rare genetic variant. The conditioning achieves a likelihood function that is free of the intercept nuisance parameter, and this approach has been generalized to include additional covariates. The conditional likelihood has the same formulation as Fisher's noncentral hypergeometric distribution (Li et al., 2015).

The size and complexity of genetic data is growing, demanding a fresh look at the methods we use to measure statistical evidence. The next step, beyond data visualization and summary statistics, should be the measurement of what the data say. The EP fulfills this objective and may be more suitable for the automatic exploratory analyses that are necessitated by big data. Although *P*-value procedures and EP-based analyses agree qualitatively for some of the examples considered here, these similarities end as one starts to compare evidence strength across scenarios of different sample sizes, and as one starts to grapple with the multiple hypothesis testing implications of standard statistical practice. Intuition and concerns about reproducibility has led the genetics field to require independent replication beyond the multiple hypothesis test adjustments implicit in conventional practice. The EP provides a theoretical justification for this intuition while ensuring that evidence strength remains constant irrespective of different experimental designs; arguably, a fundamental requirement of a reliable measure of statistical evidence.

ACKNOWLEDGMENTS

This study is based on a course taught in the Biostatistics Division of the Dalla Lana School of Public Health at the University of Toronto; a special thank you to all the students who contributed to the material and discussions over the years, and to Zeynep Baskurt, Weili Li, and Bowei Xiao for assisting with the applications to CF. Thank you to Lei Sun and Susan Hodge for a critical read of the article. Thank you to the CF patients and their families who participate in the International CF Gene Modifier studies, to the US and Canadian CF Foundations for the genotyping and clinical data of the International CF Gene Modifier Consortium, and to members of the consortium Johanna Rommens, Garry Cutting, Michael Knowles, Mitchell Drumm and Harriet Corvol.

CONFLICTS OF INTEREST

The author declares that there are no conflicts of interest.

ORCID

Lisa J. Strug http://orcid.org/0000-0003-0503-9740

REFERENCES

Anscombe, F. J. (1964). Normal likelihood functions. *Annals of the Institute of Statistical Mathematics*, *26*, 1–19.

Baskurt, Z., & Strug, L. J. (2018). The analysis of pedigree data for genetic association studies: Direct inference using the composite likelihood ratio. *Genetic Epidemiology*. (in press).

Bickel, D. R. (2012). The strength of statistical evidence for composite hypotheses: Inference to the best explanation. *Statistica Sinica*, *22*, 1147–1198.

Blackman, S. M., Commander, C. W., Watson, C., Arcara, K. M., Strug, L. J., Stonebraker, J. R., … Cutting, G. R. (2013). Genetic modifiers of cystic fibrosis-related diabetes. *Diabetes*, *62*, 3627–3635.

Blume, J. D. (2002). Tutorial in biostatistcs: Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, *21*, 2563–2599.

Blume, J. D. (2013). Likelihood and composite hypotheses [Comment on "A Likelihood Paradigm for Clinical Trials"]. *Journal of Statistical Theory and Practice*, *7*, 183–186.

Blume, J. D., & Choi, L. (2017). Likelihood based study designs for time-to-even endpoints. *arXiv*, *1711*. 01527v1.

Blume, J. D., Su, L., Olveda, R. M., & Mcgarvey, S. T. (2007). Statistical evidence for GLM regression parameters: A robust likelihood approach. *Statistics in Medicine*, *26*, 2929–2936.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*, 791–799.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., … Worthington, J. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*, 661–678.

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., … Collins, F. S. (2007). Replicating genotype-phenotype associations. *Nature*, *447*, 655–660.

Chotai, J. (1984). On the lod score method in linkage analysis. *Annals of Human Genetics*, *48*, 359–378.

Corvol, H., Blackman, S. M., Boëlle, P. Y., Gallins, P. J., Pace, R. G., Stonebraker, J. R., … Knowles, M. R. (2015). Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nature Communications*, *6*, 8382.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *63*, 269–276.

Cutting, G. R. (2015). Cystic fibrosis genetics: From molecular understanding to clinical application. *Nature Reviews Genetics*, *16*, 45–56.

Diggle, P., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. New York: Clarendon Press

Du, L., & Choi, L. (2015). Likelihood approach to evaluating bioequivalence of highly variable drugs. *Pharmaceutical Statistics*, *14*, 82–94.

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*, 227–234.

Dupuis, A., Keenan, K., Ooi, C. Y., Dorfman, R., Sontag, M. K., Naehrlich, L., … Gonska, T. (2016). Prevalence of meconium ileus marks the severity of mutations of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. *Genetics in Medicine*, *18*, 333–340.

Eddings, W. (2003). *An evidential alternative to the log-rank test* (PhD thesis). Johns Hopkins University.

Edwards, A. W. F. (1984). *Likelihood*. Cambridge University Press

Evans, M. (2015). *Measuring statistical evidence using relative belief*. Chapman and Hall.

Ferguson, T. S. (1996). *A course in large sample theory*. London: Chapman & Hall

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, *33*, 503–513.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Freedman, D. A. (2006). On the so-called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, *60*, 229–302.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, *31*, 1208–1211.

Good, I. J. (1985). *"Weight of evidence: A brief survey" in bayesian statistics*. New York: Elsevier.

Goodman, S. N. (2016). STATISTICS. Aligning statistical and scientific reasoning. *Science*, *352*, 1180–1181.

Hacking, I. (1965). *Logic of statistical inference*. New York: Cambridge University Press.

Kalbfleisch, J. D., & Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *32*, 175–208.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, *69*, 19–27.

Knight, K. (2000). *Mathematical statistics*. Chapman and Hall.

Lander, E., & Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, *11*, 241–247.

Li, B., & Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics*, *83*, 311–321.

LI, W. (2016). Pure likelihood-based methods for genetic association studies (PhD thesis). University of Toronto.

Li, W., Dobbins, S., Tomlinson, I., Houlston, R., Pal, D. K., & Strug, L. J. (2015). Prioritizing rare variants with conditional likelihood ratios. *Human Heredity*, *79*, 5–13.

Loriol, C., Dulong, S., Avella, M., Gabillat, N., Boulukos, K., Borgese, F., & Ehrenfeld, J. (2008). Characterization of SLC26A9, facilitation of Cl(-) transport by bicarbonate. *Cellular Physiology and Biochemistry*, *22*, 15–30.

Miller, M. R., Soave, D., Li, W., Gong, J., Pace, R. G., Boëlle, P. Y., … Strug, L. J. (2015). Variants in solute carrier SLC26A9 modify prenatal exocrine pancreatic damage in cystic fibrosis. *Journal of Pediatrics*, *166*, 1152–1157 e6.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics*, *7*, 277–318.

Neyman, J., & Pearson, E. S. (1933). On the problems of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, *231A*, 289–338.

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*, 150–152.

Ohana, E., Yang, D., Shcheynikov, N., & Muallem, S. (2009). Diverse transport modes by the solute carrier 26 family of anion transporters. *Journal of Physiology*, *587*, 2179–2185.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*, 237–249.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford: Clarendon Press.

Robbins, H. (1970). Statistical methods related to the law of the interated logarithm. *Annals of Mathematical Statistics*, *41*, 1397–1409.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.

Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, *95*, 760–768.

Royall, R., & Tsou, T. S. (2003). Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *65*, 391–404.

Savage, L. J. (1972). *The foundations of statistics*. New York: Dover Publications, Inc.

Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics*, *38*, 209–213.

Strug, L. J., Gonska, T., He, G., Keenan, K., Ip, W., Boëlle, P. Y., … Rommens, J. M. (2016). Cystic fibrosis gene modifier SLC26A9 modulates airway response to CFTR-directed therapeutics. *Human Molecular Genetics*, *25*, 4590–4600.

Strug, L. J., & Hodge, S. E. (2006a). An alternative foundation for the planning and evaluation of linkage analysis. I. Decoupling “error probabilities” from “measures of evidence”. *Human Heredity*, *61*, 166–188.

Strug, L. J., & Hodge, S. E. (2006b). An alternative foundation for the planning and evaluation of linkage analysis. II. Implications for multiple test adjustments. *Human Heredity*, *61*, 200–209.

Strug, L. J., Hodge, S. E., Chiang, T., Pal, D. K., Corey, P. N., & Rohde, C. (2010). A pure likelihood approach to the analysis of genetic association data: An alternative to Bayesian and frequentist analysis. *European Journal of Human Genetics*, *18*, 933–941.

Strug, L. J., Rohde, C. A., & Corey, P. N. (2007). An introduction to evidential sample size calculations. *American Statistician*, *61*, 207–212.

Sun, L., Dimitromanolakis, A., Faye, L. L., Paterson, A. D., Waggott, D., & Bull, S. B. (2011). BR-squared: A practical solution to the winner’s curse in genome-wide scans. *Human Genetics*, *129*, 545–552.

Sun, L., Rommens, J. M., Corvol, H., Li, W., Li, X., Chiang, T. A., … Strug, L. J. (2012). Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nature Genetics*, *44*, 562–569.

Vieland, V. J. (2001). The replication requirement. *Nature Genetics*, *29*, 244–245.

Vieland, V. J. (2017). Measurement of statistical evidence: Picking up where hacking and others left off. *Philosophy of Science*, *84*, 853–865.

Vieland, V. J., & Hodge, S. E. (1998). Book review of statistical evidence: A likelihood paradigm. *The American Journal of Human Genetics*, *63*, 283–289.

Viraswami, K., & Reid, N. (1998). A note on the likelihood ratio statistic under model misspecification. *The Canadian Journal of Statistics*, *26*, 161–181.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*, 117–186.

Wang, S. J., & Blume, J. D. (2011). An evidential approach to noninferiority clinical trials. *Pharmaceutical statistics*, *10*, 440–447.

Wasserstein, R. L., & Lazar, N. A. (2016). ASA statement on statistical significance and P-values. *American Statistician*, *70*, 131–133.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1.

Zhang, Z. (2009). Interpreting statistical evidence with empirical likelihood functions. *Biometrical Journal*, *51*, 710–720.

Zhang, Z., & Zhang, B. (2013). A likelihood paradigm for clinical trials. *Journal of Statistical Theory and Practice*, *7*, 157–177.

Zhong, L. and Strug, L. J. (2018). *A novel framework for differential gene expression analysis using robust profile likelihood ratios*. Joint Statistical Meetings, Vancouver, Canada.

How to cite this article: Strug LJ. The evidential statistical paradigm in genetics. *Genet. Epidemiol.* 2018;42:590–607. https://doi.org/10.1002/gepi.22151