AN EMPIRICAL BAYES TESTING PROCEDURE FOR DETECTING VARIANTS IN ANALYSIS OF NEXT GENERATION SEQUENCING DATA[1]

BY ZHIGEN ZHAO[*,2], WEI WANG[†] AND ZHI WEI[†]

Temple University* and New Jersey Institute of Technology[†]

Because of the decreasing cost and high digital resolution, next-generation sequencing (NGS) is expected to replace the traditional hybridization-based microarray technology. For genetics study, the first-step analysis of NGS data is often to identify genomic variants among sequenced samples. Several statistical models and tests have been developed for variant calling in NGS study. The existing approaches, however, are based on either conventional Bayesian or frequentist methods, which are unable to address the multiplicity and testing efficiency issues simultaneously. In this paper, we derive an optimal empirical Bayes testing procedure to detect variants for NGS study. We utilize the empirical Bayes technique to exploit the across-site information among many testing sites in NGS data. We prove that our testing procedure is valid and *optimal* in the sense of rejecting the maximum number of nonnulls while the Bayesian false discovery rate is controlled at a given nominal level. We show by both simulation studies and real data analysis that our testing efficiency can be greatly enhanced over the existing frequentist approaches that fail to pool and utilize information across the multiple testing sites.

1. Introduction. The per-base cost of DNA sequencing has plummeted by 100,000-fold over the past decade because of the dramatic development in sequencing technology in the past few years [Lander (2011)]. As a result, this new or "next generation" sequencing (NGS) technology becomes much more affordable today. With high digital resolution, NGS is expected to replace the traditional hybridization-based microarray technology [Mardis (2011)]. For genetics studies, NGS holds the promise to revolutionize genome-wide association studies (GWAS). In the microarray era, GWAS mainly addresses common Single Nucleotide Polymorphisms (SNPs) with minor allele frequency $>5\%$, based upon the common disease/common variant (CD/CV) hypothesis [Manolio et al. (2009)]. However, the identified common variants explain only a small proportion of heritability [Hindorff et al. (2009)]. Rare variants therefore have been hypothesized

to account for the missing heritability [Bodmer and Bonilla (2008), Frazer et al. (2009)]. To identify rare variants, a direct and more powerful approach is to sequence a large number of individuals [Li and Leal (2009)]. This line of thought also implicitly motivates the recent 1000 Genomes Project, which will sequence the genomes of 1200 individuals of various ethnicities by NGS [Hayden (2008)]. It is expected to extend the catalogue of known human variants down to a frequency near 1%. Besides human genetics, NGS is also revolutionizing genetics in other species. For example, NGS has been used for genotyping in maize, barley [Elshire et al. (2011)] and rice [Huang et al. (2009)], accessing allele frequencies genome-wide in Drosophila [Turner et al. (2011), Zhu et al. (2012)], and quantifying strain abundance in yeast [Smith et al. (2010)]. Because of the small sizes of their genomes, whole-genome sequencing data for tens or hundreds of samples can be feasibly generated by one single sequencing run [Smith et al. (2010), Zhu et al. (2012)]. Finally, in cancer genomics, it is interesting to study the subclonal architecture of tumors. Within a single tumor, that is, just one individual, there often exists subclones of various sizes that have distinct somatic mutations. In the case of smaller subclones, their distinct variants can be present at low frequency when one sequences the tumor as a whole. To resolve these subclones, one must be able to accurately identify such low frequency variants and use them to make inferences about cellular frequency and, thus, subclonal composition. For such applications, even if one tumor (one sample) is sequenced as a whole, it actually consists of a pool of heterogeneous cells from which rare variants are sought.

Thousands of samples need to be sequenced for securing the chance of finding most rare variants with a frequency <1% [Li and Leal (2009)]. A cost-effective strategy is needed in order to afford very large sample sizes for finding rare variants. Similar issues of cost and labor were confronted in the early expensive stage of GWAS and were circumvented by focusing on small candidate regions and the use of genomic DNA pooling [Norton et al. (2004), Sham et al. (2002)]. Borrowing the same idea, many targeted resequencing applications utilizing pooling have been seen in the past few years [Calvo et al. (2010), Momozawa et al. (2011), Nejentsev et al. (2009), Out et al. (2009)].

Current NGS can generate up to several hundred million reads per run, which may lead to oversampling with little gain in data quality when analyzing one sample with a small genome or small targeted genomic regions. To fully exploit the high-throughput of NGS, nucleotide-based barcodes have been used to multiplex individual samples [Craig et al. (2008)]. Different from the aforementioned pooling strategy, this methodology allows to sequence multiple samples in a single flow cell while keeping sample identities. However, it should be noted that, despite the more efficient use of sequencing throughput, multiplexing techniques still require a large number of individual DNA extractions, manipulations of reagents, barcoding oligos, PCR reactions and sequencing library constructions [Zhu et al. (2012)]. For example, in one of our ongoing projects targeted resequencing 6 Mb genomic regions of 960 human samples, the cost for the library preparation kit

(TruSeq Library Prep + NimbleGen Custom EZ Seq Cap Panel) is \$405 per sample (labor cost not included). We might multiplex 96 samples on one Illumina HiSeq 2000 lane and get enough sequencing depth per sample ($>$40X/sample). Although the cost for the sequencing step is then restrained to \$2200 (one lane), the library preparation would cost dominantly as much as $96 * 405 = \$38{,}880$, which is not reduced by multiplexing/barcoding. The library preparation step is cheaper for whole genome sequencing, as there is no need for capturing targeted regions. However, the total library preparation cost for multiplexing tens or more of samples on one lane is still much higher than that for the sequencing step. In contrast, pooling individuals prior to DNA extraction and sequencing the pooled DNA without barcodes are very cost-effective by reducing library preparation cost. As a result, for population studies where identifying variants and frequencies is the primary interest rather than knowing which sample the variant came from, nonindexed multi-sample pools are being widely used to discover rare variants and/or assess allele frequencies at population level in Drosophila [Kolaczkowski et al. (2011), Turner et al. (2011), Zhu et al. (2012)], Anopheles gambiae [Cheng et al. (2012)], Arabidopsis [Turner et al. (2010)], pig [Amaral et al. (2011)] and human [Margraf et al. (2011)], among others.

A schematic example of pooled NGS data is illustrated in Figure 1 assuming there are M pools with N samples in each pool. For most species, the genetic material DNA is identical at most bases in a population apart from variations at a small proportion of loci. Single Nucleotide Variants (SNVs) are the most common DNA sequence variations occurring when a single nucleotide (A, T, C or G) in the genome differs between members of a biological species or paired chromosomes in an individual. SNVs generally exhibit two alleles in a population. In this particular example, the two alleles, reference (major) allele and alternative (minor) allele, are A and G, respectively. Each nucleotide site in each individual chromosome is sequenced a random number of times. When pooling $N > 1$ individuals, the information of which individual chromosome is represented in a particular read is lost. In addition, sequencing errors may flip the original allele into different ones that are observed. It is noted that when there is only $N = 1$ individual in a "pool", it represents so-called (individually sequenced) multiple-sample variant call. Finally, in the aforementioned cancer genomics studies, because of the heterogeneity of tumor cell population, the effective N for one individual tumor sample is believed to be larger than 1.

Identification of genomic variants has become routine after NGS DNA data are generated. Quite a few tools have been implemented to identify SNVs. Formally, for a genomic locus, if its minor allele frequency (MAF) in a population is larger than 0, then we call it a SNV. SNV detection is a relatively straightforward problem in analysis of individual data, because the frequency of a candidate allele can be only 0 (nonvariant), 0.5 (heterozygous) or 1 (alternate homozygous) for a diploid genome. Several similar conventional Bayesian models have been used in existing popular tools [Li, Ruan and Durbin (2008), Li et al.
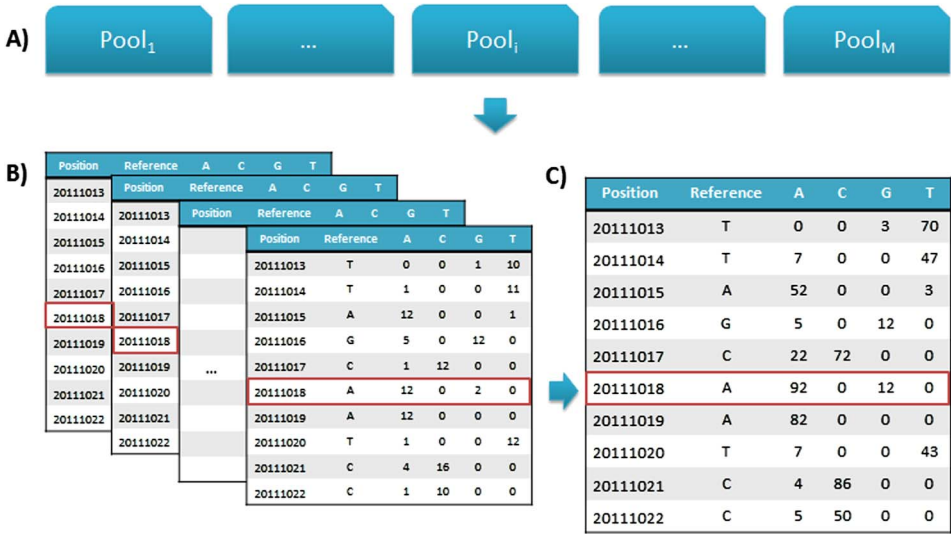
A)

| Pool$_1$ | ... | Pool$_i$ | ... | Pool$_M$ |

B)

Position	Reference	A	C	G	T
20111013	T	0	0	1	10
20111014	T	1	0	0	11
20111015	A	12	0	0	1
20111016	G	5	0	12	0
20111017	C	1	12	0	0
20111018	A	12	0	2	0
20111019	A	12	0	0	0
20111020	T	1	0	0	12
20111021	C	4	16	0	0
20111022	C	1	10	0	0

C)

Position	Reference	A	C	G	T
20111013	T	0	0	3	70
20111014	T	7	0	0	47
20111015	A	52	0	0	3
20111016	G	5	0	12	0
20111017	C	22	72	0	0
20111018	A	92	0	12	0
20111019	A	82	0	0	0
20111020	T	7	0	0	43
20111021	C	4	86	0	0
20111022	C	5	50	0	0

FIG. 1. *Schematic illustration of pooled next generation sequencing data.* (A) *Suppose M pools are designed for sequencing and each pool contains N samples. There are two scenarios for pooled data. When $M > 1$, multiple pool sequencing data are generated. It is possible that $M > 1$ and $N = 1$, representing so-called (individually sequenced) multiple-sample variant call. If $M = 1$, it becomes single pool data. It noted that for "one" heterogenous cancer sample, the effective N is larger than 1.* (B) *For the ith pool of N samples, each nucleotide site is sequenced a random number of times, which yields different counts of four nucleic acid bases (A, C, G, T) that make up DNA.* (C) *An example of pooling N samples at a particular site. There are two types of alleles: the reference (major) allele A and the alternative (minor) allele G. The information is combined from the entire pool of n individuals.*

(2009a, 2009b), McKenna et al. (2010)]. The multiplicity issue has been largely ignored in these conventional Bayesian approaches. Identifying variants from pooled NGS data is more challenging in that pooled DNA are sampled from a number of individuals, which consequently will give rise to variant allele frequencies other than simply 0, 0.5 or 1. Driven by the need for analysis of increasing amount of pooled NGS data, quite a few statistical models for the detection of variants from pooled sequencing data have been developed [Altmann et al. (2011), Bansal (2010), Druley et al. (2009), Vallania et al. (2010), Wei et al. (2011)]. Most existing methods, however, are based on statistical tests from a frequentist point of view. For example, Wei and colleagues propose a binomial–binomial model for testing the existence of variants from a single-pool data [Wei et al. (2011)]. Their binomial–binomial model provides a unified likelihood function for both pooled and individual data and has addressed the multiplicity issue. When there is more than one pool, they employ the partial conjunction test [Benjamini and Heller (2008)] that at least $u = 1$ out of the M hypotheses is false for testing whether a locus is a variant site. Alternatively, one can also combine individual pool p-values by conducting meta analysis. These frequentist approaches, despite

making few assumptions, fail to pool and utilize information across the multiple sites that are being tested. Although these approaches are valid in terms of controlling the FDR at the nominal level, they are not optimal and powerful in detecting variants of interest. We call an FDR procedure *valid*, if it controls the FDR at the nominal level and *optimal*, if it has the smallest false negative rate [FNR, Genovese and Wasserman (2002)] among all valid FDR procedures [Wei et al. (2009)]. The optimality issue in multiple testing has received more and more attention in the past few years [Sun and Cai (2007, 2009), He, Sarkar and Zhao (2012), Sun and Wei (2011), Wang, Wei and Sun (2010), Wei et al. (2009), Xie et al. (2011)].

Hundreds of thousands or more sites are tested in typical NGS data. Such high dimensionality imposes great challenges, but can also be a blessing for inference if handled properly. Empirical Bayesian approaches, a hybrid of frequentist and Bayesian methods, become increasingly popular in modern high-dimensional data inference [Efron (2005)]. It enables the frequentists to achieve the Bayesian efficiency in solving high-dimensional problems [Efron (2010)]. Assume that the high-dimensional parameters follow some distribution governed by, for instance, a few hyperparameters. These hyperparameters can be estimated reliably via a classical frequentist way. In addition, empirical Bayesian approaches eliminate the subjective selection of priors and are generally more robust.

In this article we propose a parametric empirical Bayes testing procedure for detecting variants in the analysis of high-dimensional NGS data. When deriving our empirical Bayes procedure, we start from assuming the hyperparameters are known. Given the known hyperparameters, we derive a Bayesian decision rule which is optimal in the sense of detecting the maximum number of variants while the Bayesian false discovery rate [Sarkar, Zhou and Ghosh (2008)] is controlled at a given nominal level. To avoid a subjective choice of the hyperparameters, we estimate the hyperparameters consistently by using the method of moments, followed by an empirical Bayes procedure. Asymptotically, it is guaranteed that the empirical Bayes procedure mimics the oracle procedure uniformly for all the hyperparameters.

In this article we introduce our empirical Bayes testing procedure in Section 2. We present results from simulation studies in Section 3 to demonstrate the superiority of the proposed procedures in comparison with existing methods. In Section 4, for a case study, we apply the data-driven procedure to analyze a recent real NGS data set. We present a brief discussion in Section 5. The proof of the theorems are provided in the supplemental article [Zhao, Wang and Wei (2013)].

The methods developed in this paper have been implemented using Java in a computationally efficient and user-friendly software package, EBVariant, as well as an R package, available from http://ebvariant.sourceforge.net/.

2. Statistical models and methods. To discover (rare) variants in a cost-effective way, we consider a sequencing procedure by pooling a normalized amount of DNA from multiple samples. Because of a capacity issue, samples may be distributed and sequenced independently in more than one pools. Without loss of generality, we assume that there are M pools, each pool with N individuals (haploids). It is noted that the following proposed model assumes a general framework and does not require $N > 1$. As a result, when $N = 1$ ($N = 2$ for a diploid genome), implying each pool has only one sample, the proposed model is still applicable and will make an individually sequenced multiple-sample variant call. Suppose that sequencing covers p sites that are to be tested for variant candidates. We expect p to be tens or hundreds of thousands for targeted resequencing, millions for whole-exome sequencing, and billions for whole-genome sequencing (human). We assume that K_{ij} short reads cover locus i in pool j, out of which we observe X_{ij} reads carry alternative alleles. If there were no sequencing and mapping errors, we might easily identify variant loci as those with $X_{ij} > 0$. We assume a general sequencing/mapping error ε, under which the alternative allele will be flipped to one of the other three alternate alleles, and vice versa. Our goal is to identify single nucleotide variants (SNVs) that have nonzero minor allele frequencies in the population.

2.1. *Oracle testing procedure for multiple pools.* We assume that θ_{ij} is the minor (alternative) allele frequency (MAF) at the ith site in the jth pool. Let $\mu_i \in \{0, 1\}$ be the hidden state of whether the ith locus is a SNV. Given $\mu_i = 0$, then $\theta_{ij} = 0, \forall j = 1, 2, \ldots, M$. If $\mu_i = 1$, then θ_{ij}'s are nonzero but may vary across different pools. Following a binomial–binomial model proposed by Wei et al. (2011), we assume that the unknown MAF θ_{ij} governs n_{ij}, the number of haploids in a pool carrying the alternative alleles, by a binomial model; and that the unobserved n_{ij} governs its proxy X_{ij} by another binomial model. Unlike the frequentist approach in Wei et al. (2011), we put a prior for θ_{ij} as $\psi(\theta_{ij})$ when it is nonzero. We therefore have a hierarchical model as follows:

$$
(2.1) \quad
\begin{cases}
X_{ij}|n_{ij} \sim b\left(K_{ij}, \dfrac{n_{ij}}{N}(1 - \varepsilon) + \left(\dfrac{N - n_{ij}}{N}\dfrac{\varepsilon}{3}\right)\right), \\
n_{ij}|\theta_{ij} \sim b(N, \theta_{ij}), \\
\theta_{ij}|\mu_i \sim (1 - \mu_i)\delta_0 + \mu_i\psi(\theta_{ij}), \\
\mu_i \sim \text{Bernoulli}(\pi_0).
\end{cases}
$$

When there are millions of parameters to be inferred, a common strategy is to assume that these parameters are drawn from a certain distribution. We take the parametric approach and assume that θ_{ij} follows a uniform distribution $U(0, a)$ with $0 < a < 1$ when $\mu_i = 1$. The corresponding likelihood function of X_{ij} ($i =$

$1, 2, \ldots, p, j = 1, 2, \ldots, M)$ is

$$f(X_{ij}|\theta_{ij}, \mu_i = 1) = \sum_{n_{ij}=0}^{N} \binom{K_{ij}}{X_{ij}} \left(\frac{n_{ij}}{N}(1 - \varepsilon) + \frac{N - n_{ij}}{N}\frac{\varepsilon}{3}\right)^{X_{ij}}$$

$$(2.2) \qquad\qquad \times \left(1 - \left(\frac{n_{ij}}{N}(1 - \varepsilon) + \frac{N - n_{ij}}{N}\frac{\varepsilon}{3}\right)\right)^{K_{ij} - X_{ij}}$$

$$\times \binom{N}{n_{ij}} \theta_{ij}^{n_{ij}} (1 - \theta_{ij})^{N - n_{ij}}.$$

When $\mu_i = 0$, the likelihood function becomes

$$(2.3) \qquad f(X_{ij}|\mu_i = 0) = \binom{K_{ij}}{X_{ij}} \left(\frac{\varepsilon}{3}\right)^{X_{ij}} \left(1 - \frac{\varepsilon}{3}\right)^{K_{ij} - X_{ij}}.$$

To identify the variants, we test the hypothesis $H_i : \mu_i = 0, i = 1, 2, \ldots, p$. In this multiple-pool scenario, a question remains on how to combine the data from multiple pools together. Wei and colleagues test each single pool separately and combine the single-pool p-values using the Simes' method for testing a partial conjunction hypothesis [Wei et al. (2011)]. Alternatively, one can conduct the meta-analysis using, for instance, Fisher's combined probability test [Fisher (1925)]. However, none of these methods is optimal. We will show in Section 3 that these two approaches are conservative in detecting the variants. The goal of this paper is to construct an optimal multiple testing procedure by using the Bayesian decision theory [He, Sarkar and Zhao (2012), Sun and Cai (2007)].

Let δ_i be the 0–1 decision rule corresponding to the ith hypotheses, that is, we reject the hypothesis H_i if $\delta_i = 1$. We consider the loss function

$$(2.4) \qquad L(\boldsymbol{\delta}, \boldsymbol{\mu}) = \sum_i \lambda(1 - \mu_i)\delta_i + \mu_i(1 - \delta_i),$$

where the tuning parameter λ controls the trade-off between the Type I error and the Type II error. Then to minimize the Bayes risk $EL(\boldsymbol{\delta}, \boldsymbol{\mu})$, we have the Bayesian decision rule $\boldsymbol{\delta}^B = (\delta_1^B, \ldots, \delta_p^B)$ with

$$(2.5) \qquad \delta_i^B = I\left(P(\mu_i = 0|\mathbf{X}) < \frac{1}{\lambda + 1}\right).$$

Let $fdr_i(\mathbf{X}) = P(\mu_i = 0|\mathbf{X})$ be the posterior probability of μ_i being zero, which is the local fdr score as given in Efron et al. (2001), Efron (2008, 2010). It can be written as

$$(2.6) \quad fdr_i(\mathbf{X}) = \frac{\pi_0 \prod_{j=1}^{M} f(X_{ij}|\mu_i = 0)}{\pi_0 \prod_{j=1}^{M} f(X_{ij}|\mu_i = 0) + \pi_1 \prod_{j=1}^{M} \int f(X_{ij}|\theta_{ij})\psi(\theta_{ij})\,d\theta_{ij}}.$$

Unlike the two aforementioned approaches, the local fdr score combining the information across multiple pools proves optimal in the decision theoretical framework.

The Bayesian decision rule (2.5) depends on the tuning parameter λ which, however, is not trivial to set. In many real applications, of interest is to control certain type I error rates. False discovery rate (FDR) [Benjamini and Hochberg (1995)] is one of the most popular ones for high-dimensional data. Its recent extensions include mFDR, which equals $FDR + O(1/p)$ under weak conditions [Genovese and Wasserman (2002)], and positive FDR [Storey (2003)]. Following Sarkar, Zhou and Ghosh (2008), we consider the Bayes version of FDR and FNR (false nondiscovery rate) in the Bayesian framework as follows.

Let $R = \sum_{i=1}^{p} \delta_i$ and $A = \sum_{i=1}^{p}(1 - \delta_i)$ be the total number of rejections and acceptances, respectively. Let $V = \sum_{i=1}^{p} \delta_i(1 - \mu_i)$ and $U = \sum_{i=1}^{p} \mu_i(1 - \delta_i)$ be the number of false rejections and false acceptances, respectively. Define BFDR and BFNR as

$$BFDR = E_{\mathbf{X},\boldsymbol{\mu}} \frac{V}{R \vee 1}, \qquad BFNR = E_{\mathbf{X},\boldsymbol{\mu}} \frac{U}{A \vee 1}.$$

Let $t = \frac{1}{\lambda+1}$ and we rewrite the decision Bayes rule as $\boldsymbol{\delta}^B(t) = (\delta_1^B(t), \ldots, \delta_p^B(t))$ with

$$(2.7) \qquad \delta_i^B(t) = I\big(P(\mu_i = 0|\mathbf{X}) < t\big).$$

Then

$$BFDR\big(\boldsymbol{\delta}^B(t)\big) = E \frac{\sum_i I(fdr_i(\mathbf{X}) < t) fdr_i(\mathbf{X})}{\sum_i I(fdr_i(\mathbf{X}) < t) \vee 1},$$

which is increasing with respect to t. As $t \to 0$, it converges to 0. When $t \to +\infty$, then

$$\lim_{t \to +\infty} BFDR\big(\boldsymbol{\delta}^B(t)\big) = \frac{1}{p} E_{m(\mathbf{X})} \sum_i fdr_i(\mathbf{X}) = \pi_0.$$

Consequently, when $\pi_0 > \alpha$, there exists a value $t(\alpha)$ such that the decision Bayes rule controls the BFDR at α and the BFDR is greater than α for any $t > t(\alpha)$. Sun and Cai (2007) and He, Sarkar and Zhao (2012) have shown that this procedure is optimal in the sense that it yields the minimal BFNR among all procedures that can control the BFDR at level α. This optimal rule relies on the cut-off $t(\alpha)$, which depends on α implicitly. After deriving the empirical Bayes version of the local fdr scores in Section 2.2, we introduce a data driven procedure to choose this cutoff in Section 2.3.

2.2. *Empirical Bayes estimators.* The oracle testing procedure defined in Section 2.1 assumes that the hyperparameters π_0, π_1 and a are known. To avoid a subjective choice of these hyperparameters, we estimate them using an empirical Bayes approach. To simplify our discussion, we first explain the estimators for the

hyperparameters for single-pool data. Taking out the pool index j, the hierarchical model for single-pool data becomes

$$
(2.8) \quad
\begin{cases}
X_i | n_i \sim b\left(K_i, \dfrac{n_i}{N}(1 - \varepsilon) + \left(\dfrac{N - n_i}{N} \dfrac{\varepsilon}{3} \right) \right), \\
n_i | \theta_i \sim b(N, \theta_i), \\
\theta_i | \mu_i \sim (1 - \mu_i)\delta_0 + \mu_i U(0, a), \\
\mu_i \sim \text{Bernoulli}(\pi_0).
\end{cases}
$$

Define two statistics

$$
(2.9) \quad m_1 = \frac{\sum_i (X_i / K_i - \varepsilon/3)}{p}
$$

and

$$
m_2 = \frac{1}{p} \sum_i (X_i^2 - K_i^2(\varepsilon^2/9) - K_i(\varepsilon/3)(1 - (\varepsilon/3)) - K_i(1 - (2\varepsilon/3))m_1
$$

$$
(2.10) \hspace{6cm} - K_i^2(2\varepsilon/3)m_1)
$$

$$
/((K_i^2 - K_i)(1 - 4\varepsilon/3)^2).
$$

THEOREM 2.1. *Assume the model* (2.8) *and the definitions of* m_1 *and* m_2 *in* (2.9) *and* (2.10), *then*

$$
Em_1 = \left(1 - \frac{4\varepsilon}{3} \right) \pi_1 \frac{a}{2}
$$

and

$$
Em_2 = \frac{N-1}{N} \pi_1 \frac{a^2}{3} + \frac{1}{N} \pi_1 \frac{a}{2}.
$$

By using the method of moments, we can estimate a, π_0 and π_1 as

$$
(2.11) \quad
\begin{cases}
\hat{a} = \dfrac{3(N(1 - 4\varepsilon/3)m_2 - m_1)}{2m_1(N - 1)}, \\
\hat{\pi}_1 = \dfrac{2m_1}{(1 - 4\varepsilon/3)\hat{a}}, \hspace{2cm} \hat{\pi}_0 = 1 - \hat{\pi}_1.
\end{cases}
$$

THEOREM 2.2. *Assume that the empirical Bayes estimators of* a, π_0 *and* π_1 *are given by* (2.11), *then* $\hat{a} \overset{P.}{\to} a$, $\hat{\pi}_0 \overset{P.}{\to} \pi_0$ *and* $\hat{\pi}_1 \overset{P.}{\to} \pi_1$, *for all* $0 < a < 1, 0 < \pi_1 < 1$.

The estimation of these hyperparameters borrows information across all loci and is thus consistent when the number of loci goes to infinity. This can be viewed as the blessing of the high dimensionality. It is noted that the estimation may result in

negative estimates of a and π_1 when p is finite. For NGS data analysis, people may have certain knowledge about these unknown parameters. For example, genome-wide π_1 is believed to be greater than 0.1%. We then can set $\hat{\pi}_1$ as 0.1% if it is less than 0. Similarly, we may estimate a as 0.01 if $\hat{a} < 0$. Therefore, we have the truncated estimators for the hyperparameters as

$$(2.12) \quad \begin{cases} \hat{a}^T = \hat{a} I(\hat{a} > 0) + 0.01 I(\hat{a} < 0), \\ \hat{\pi}_1^T = \hat{\pi}_1 I(\hat{\pi}_1 > 0) + 0.001 I(\hat{\pi}_1 < 0), \qquad \hat{\pi}_0^T = 1 - \hat{\pi}_1^T. \end{cases}$$

These truncated estimators are still consistent for $\pi_0 \in (0, 1)$ and $a \in (0, 1)$.

For the multiple-pool scenario as described in model (2.1), we assume the observations X_{ij}, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, M$, share the same marginal distribution. Treating $\{X_{ij}\}$ and $\{K_{ij}\}$ as $p \times M$-dimensional vectors, we can estimate π_1 and a by (2.12) similarly. Such estimators converge even faster because of the larger sample size.

2.3. *An empirical Bayes testing procedure.*   Section 2.1 has developed an optimal oracle testing procedure. Section 2.2 has provided the empirical Bayes estimators for the parameters π_0 and a in the testing procedure when they are unknown. In this section we propose an empirical Bayes testing procedure as follows.

DEFINITION 2.1 [An Empirical Bayes Testing Procedure (emBayes)].

1. Estimate π_0 and a according to (2.12).
2. For the ith locus, calculate the local fdr $\widehat{fdr_i(X)}$ by plugging the $\hat{\pi}_1$ and $\hat{a}$ into (2.6).
3. Order $\widehat{fdr_i(X)}$ as $\widehat{fdr_{(1)}(X)} \leq \widehat{fdr_{(2)}(X)} \leq \cdots \leq \widehat{fdr_{(p)}(X)}$.
4. Find the maximum J such that $\frac{1}{J} \sum_{i=1}^{J} \widehat{fdr_{(i)}(X)} \leq \alpha$.
5. Reject hypothesis $H_{(1)}, H_{(2)}, \ldots, H_{(J)}$ and accept the rest.

THEOREM 2.3.   *Assume the model* (2.1) *and the hyperparameters are estimated as described in Section* 2.2. *Let* $\widetilde{BFDR}$ *and* $\widetilde{BFNR}$ *be the Bayes FDR and FNR of the empirical Bayes procedure. Then*

$$\widetilde{BFDR} = BFDR_{OR} + o(1), \qquad \widetilde{BFNR} = BFNR_{OR} + o(1)$$

for any $\pi_1 \in (0, 1)$ *and* $a \in (0, 1)$, *where* $BFDR_{OR}$ *and* $BFNR_{OR}$ *are the Bayes FDR and FNR of the oracle optimal multiple testing procedure.*

The empirical Bayes procedure was first introduced by Robbins (1951, 1956), and is also known as a nonparametric empirical Bayes procedure because the prior is completely unspecified. Recently, Sun and Cai (2007) and He, Sarkar and Zhao (2012) constructed optimal nonparametric empirical Bayes multiple testing proce-

dures in the normal mean setting. In our study, the observation follows a binomial–binomial model. We put a family of priors with a few hyperparameters for governing the high-dimensional parameters. The resultant approach is a parametric empirical Bayes procedure, first proposed by Efron and Morris (1971, 1973, 1975). Asymptotically, the procedure controls the Bayes FDR uniformly for all hyperparameter settings. This control is less stringent than that in the frequentist procedure which requires that the Bayes FDR be controlled for the class of all point priors on θ [Morris (1983)]. Our empirical Bayes procedure is more robust than the conventional Bayesian approach which takes a subjective choice of the hyperparameters. For instance, when setting π_1 as 0.4%, the conventional Bayesian procedure may not control the BFDR if the true π_1 is less than 0.4%, and it may lack power if the true π_1 is greater than 0.4%.

3. Simulation. We first investigate the numerical performance of the proposed empirical Bayes procedure (emBayes) using simulated data. Simulation design follows Wei et al. (2011), with the settings: $M = 5$ pools, $N = 20$ subjects in each pool, the proportion of alternatives π_1 varying among 1%, 0.7%, 0.3% and 0.1%, the MAF $\psi(\theta_{ij}) \sim U(0, a)$ with a being 0.01, 0.02, 0.03 or 0.05, the number of loci $p = 1$ million (1M) or 2 millions (2M), the sequencing error $\varepsilon = 0.01$, and the sequencing coverage K_{ij} following a gamma distribution with mean 30 [Prabhu and Pe'er (2009)].

We compare emBayes with its oracle version, where we use the true values of a and π_0, and two frequentist approaches, SNVer and META. Both SNVer and META test each single pool separately using the binomial–binomial model. SNVer [Wei et al. (2011)] combines the single-pool p-values using the Simes' method for testing a partial conjunction hypothesis in order to get multiple-pool p-values. META conducts meta-analysis and obtains multiple-pool p-values as

$$p^{\text{Pool}} = P\left(\chi^2_{2M} > -2\sum_{j=1}^{M} \ln p_j\right),$$

where χ^2_{2M} is the chi-squared random variable with $2M$ degrees of freedom. Both approaches then employ the BH procedure [Benjamini and Hochberg (1995)] to control FDR.

We evaluate these methods by the number of total rejections (ER), the number of false rejections (EV) and the FDR, averaged over 100 replications, at the nominal FDR level 0.05. The results are summarized in Table 1. Compared with SNVer, META is more conservative and dominated, as indicated by its smaller FDR, fewer total rejections and fewer true rejections. The results for META are thus not included in the table.

From Table 1 we can see that the FDR levels of all three procedures are controlled at 0.05 asymptotically under all settings while SNVer is conservative. The power of emBayes is greatly improved over SNVer. For instance, when $p = 1M$, $\pi_1 = 0.4\%$ and $a = 0.02$, the numbers of correctly rejected hypotheses for these

TABLE 1
The power and FDR comparison of emBayes, SNVer and the oracle procedure at the nominal FDR level 5%. ER: the number of total rejections; EV: the number of false rejections; FDR: false discovery rate

π_1	a	p	emBayes		Oracle		SNVer	
			ER/EV	FDR	ER/EV	FDR	ER/EV	FDR
$\pi_1 = 1\%$	0.01	$1M$	467/20	0.039	541/27	0.05	277/3.3	0.012
		$2M$	1058/52	0.049	1088/55	0.05	563/6.7	0.012
	0.02	$1M$	1464/73	0.05	1467/74	0.05	850/11	0.013
		$2M$	2931/144	0.049	2943/147	0.05	1702/22	0.013
$\pi_1 = 0.7\%$	0.01	$1M$	295/12	0.038	341/17	0.049	178/2.2	0.012
		$2M$	632/28	0.042	682/33	0.049	351/4.2	0.012
	0.02	$1M$	959/48	0.05	962/48	0.05	533/6.6	0.012
		$2M$	1917/94	0.049	1931/97	0.05	1063/13	0.012
$\pi_1 = 0.4\%$	0.01	$1M$	132/4.3	0.029	160/7.4	0.046	83/0.9	0.010
		$2M$	292/12	0.04	325/16	0.049	170/2.1	0.012
	0.02	$1M$	470/22	0.047	487/24	0.049	257/3.1	0.012
		$2M$	971/48	0.049	985/49	0.05	520/6.2	0.012
$\pi_1 = 0.1\%$	0.01	$1M$	22/1.1	0.041	26/1.4	0.051	13/0.14	0.01
		$2M$	44/1.8	0.032	55/2.4	0.044	26/0.18	0.0068
	0.02	$1M$	73/3	0.036	88/4.4	0.05	45/0.6	0.013
		$2M$	153/6	0.035	177/8.3	0.047	90/0.91	0.0099

two approaches are 470 and 257, respectively. The number of true rejections is almost doubled. The emBayes has very comparable, if not the same, performance, compared with the oracle procedure. The discrepancy is more noticeable when π_1 and a are smaller. The reason is that the empirical Bayes estimators of the hyperparameters converge slowly near the boundary of the parameter space.

In all these simulations, SNVer proves conservative as indicated by extremely low FDR. It is tempting to conjecture that the higher power of emBayes is gained at the price of a higher FDR level. In other words, these two methods might actually yield similar rankings of the candidate loci and would demonstrate comparable power at the same empirical FDR level. To clarify the superiority in terms of prioritizing candidate loci, we employed ROC curves to illustrate ranking efficiency. Specifically, we calculated sensitivity as the average proportions of the total number of true rejections to the total number of nonnulls over the 100 replications. We varied the significance thresholds for identifying up to 10,000 variants and calculated corresponding FDRs and sensitivities. The resultant ROC curves of sensitivity versus FDR for emBayes, SNVer, META and the oracle procedure under the setting of $p = 1M$ and $a = 0.02$ are shown in Figure 2. It is clearly seen that emBayes dominates SNVer and META. Our proposed empirical Bayes ap-
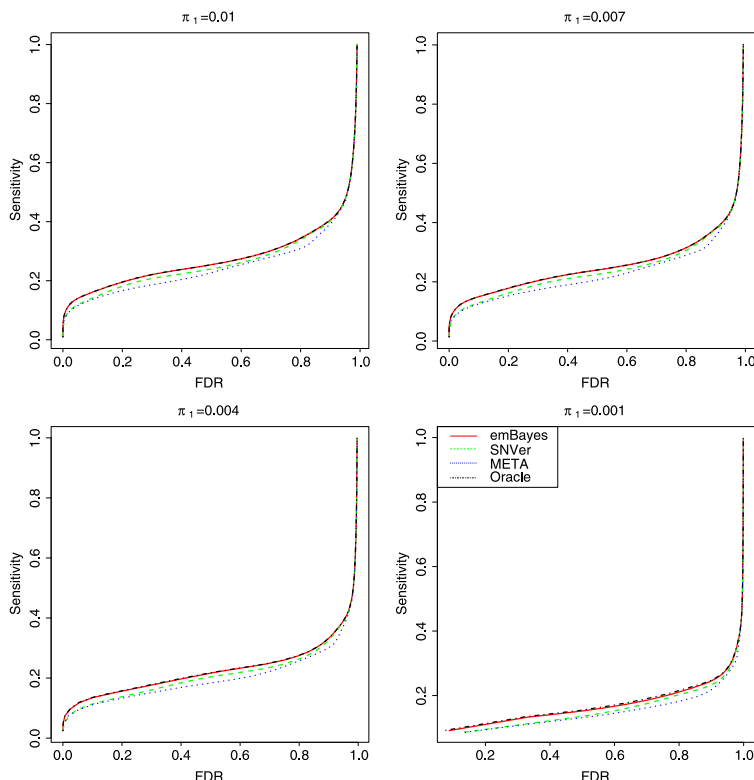
FIG. 2. *ROC curves to compare ranking efficiency of emBayes* (*red solid*), *SNVer* (*green dashed*), *META* (*blue dotted*) *and oracle procedure* (*black-dot dashed*) *under the setting of* $p = 1M$ *and* $a = 0.02$ *with different proportions of nonnulls.*

proach can identify more true variants than the frequentist competitors at the same FDR levels. For example, when $a = 0.02$, $\pi_1 = 0.1\%$ and the FDR level of 0.1, the numbers of true rejections for emBayes, SNVer, META and the oracle procedure are 98, 80, 81 and 98, respectively. The improvement of emBayes over SNVer is as large as $(98 - 80)/80 = 22.5\%$.

In summary, our simulation studies show that not only can emBayes control FDR at nominal level, but, more importantly, it also proves optimal in terms of power and can detect more variants than its frequentist alternatives.

4. Real data analysis. We also assess the performance of our proposed approach by analyzing a real NGS data set. In a recent pooled sequencing study, Zhu and colleagues conducted whole-genome resequencing pools of nonbarcoded Drosophila melanogaster strains [Zhu et al. (2012)]. The library A (SRR353364.1) in their study was constructed from a pool of 220 flies (10 females per strain) and sequenced on a single lane of Illumina GAIIx platform with 100 bp paired-end

reads, leading to an averaged sequencing depth of 10X. This library was also independently sequenced by the Drosophila Population Genomics Project (DPGP) (http://www.dpgp.org/). Following the authors, we utilized this library to evaluate variant call performance. Specifically, we extracted the genotypes of those 22 strains in the Library A from the Drosophila Genetic Reference Panel (DGRP) (http://dgrp.gnets.ncsu.edu) and used them as gold standard for estimating False Discovery Rate (FDR).

We downloaded the given bam file, based on which we then called variants using emBayes and SNVer at the nominal FDR level 0.05. Because of the large size of Drosophila genome, we analyzed the data separately for each chromosome. The variant call results are displayed in Figure 3. The emBayes called significantly more varaints than SNVer across all five chromosomes, with an average of 97,000
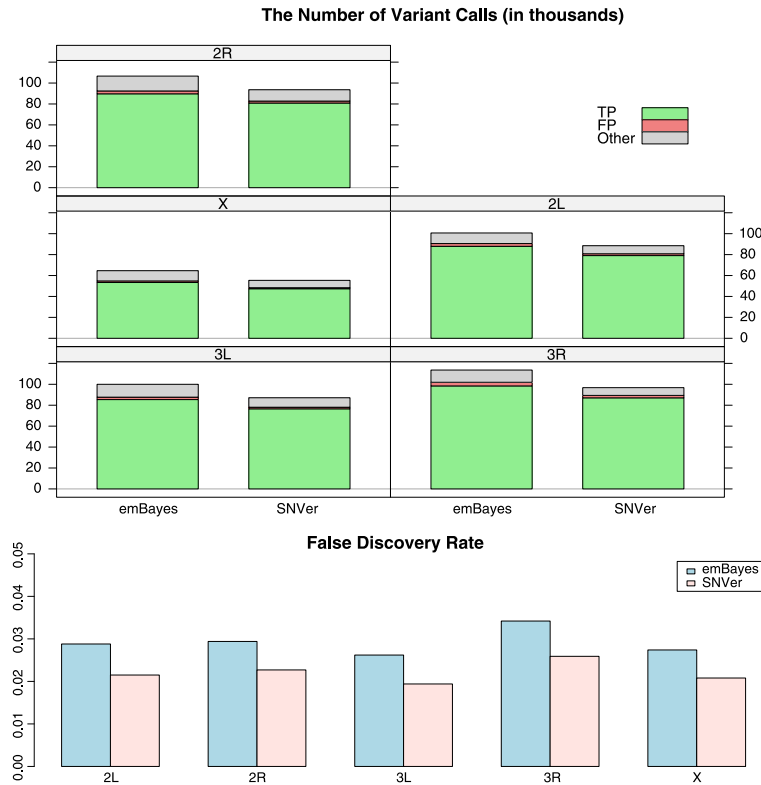

FIG. 3.　*Variant call performance. For both methods emBayes and SNVer, we call variants at the nominal level $\alpha = 0.05$. TP: True Positive; FP: False Positive. TP and FP are the variants that are called by the method and also have genotype information available from another data source (DGRP). Other: the variants called by the method but without genotype information available from DGRP. FDR: estimated false discovery rate equal to FP/(TP + FP). emBayes calls more variants (>10%) than SNVer across all five chromosomes. Both methods can control FDR at the nominal level, while SNVer is more conservative than emBayes.*

variants per chromosome and the improvement ranging from 13.78% (Chromosome 2L) to 17.4% (Chromosome 3R). Although, as expected, emBayes identified more variants than SNVer, it is also important to check if these two methods can control FDR at the prespecified nominal level. The majority of the called variants (89%) were found to have their genotype information available from DGRP, which were then used for estimating FDR. As we can see from Figure 3, both of the two methods controlled FDR at the nominal level, while SNVer revealed a little more conservative than emBayes. Consistent to the simulation studies, the larger numbers of variants called by emBayes therefore support its improved power over SNVer.

In summary, the real data analysis confirms that the proposed empirical Bayesian method, while addressing the multiplicity issue by controlling FDR, is a more powerful approach by utilizing the global information than the frequentist approach in detecting variants in NGS study.

5. Conclusion and discussion. This paper has derived an optimal empirical Bayes testing procedure for detecting variants in analysis of the increasingly popular NGS data. We utilize the empirical Bayes technique to exploit the across-site information among the vast amount of testing sites in the NGS data. We prove that our testing procedure is valid and *optimal* in the sense of rejecting the maximum number of nonnulls while the marginal FDR is controlled at a given nominal level. We show by both simulation studies and real data analysis that our testing efficiency can be greatly enhanced over the existing frequentist approaches that fail to pool and utilize information across the multiple testing sites.

The existing approaches for variant call in NGS study are either conventional Bayesian models or frequentist tests. Our empirical Bayes approach can be viewed as a hybrid of the frequentist and Bayesian methods. It thus enjoys the pros of both and overcomes the cons of each. Compared to the frequentist approaches, it enjoys the Bayesian advantage of its capability of pooling information across testing sites, and therefore is more powerful. In addition, its output local fdr scores can be used as variant call quality that may be useful in downstream association analysis [Daye, Li and Wei (2012)]. Compared to the conventional Bayesian approaches, it avoids any subjective choice of prior parameters and estimates them reliably via a classical frequentist way; it gains multiplicity control by controlling the Bayes FDR at any designated level uniformly for all the hyperparameters. This is particularly desirable because tens of thousands or millions of loci are simultaneously examined in typical NGS experiments. Each user can choose the false-positive error rate threshold he or she considers appropriate, instead of just the dichotomous decisions of whether to "accept or reject the candidates" provided by most existing methods.

Our current empirical Bayes testing procedure can be extended and improved in several ways. First, sequencing/mapping error in NGS data is much more complicated. Due to the heterogeneity of DNA, such as repeats, duplication and GC

content, there could be distinct error profiles for different genomic regions even if they are sequenced under the same experimental condition. Instead of assuming a global and general error rate, we may take and estimate specific and local error rates empirically from the data for further improving variant call efficiency. Second, strand bias is an issue observed in many sequencing platforms but not yet considered in our testing model. We may count and model ACGT for the forward strand and reverse strand separately, so as to detect the strand bias and/or allele imbalance issues introduced by inaccurate mapping or sequencing error. Third, besides single nucleotide variants (SNVs), there exist small insertions and deletions (indels). The prevalence and distribution of these indels are quite different from SNVs. A similar empirical Bayes model but with different priors may be developed. How to combine them for an overall multiplicity control while maintaining optimality is not clear. The recent pooled analysis idea for multiple-testing in GWAS [Wei et al. (2009)] may be borrowed and worthy of further research. We are currently working on these extensions.

SUPPLEMENTARY MATERIAL

Supplement to "An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data" (DOI: 10.1214/13-AOAS660SUPP; .pdf). This file contains the technical proof of the theorems.

REFERENCES

ALTMANN, A., WEBER, P., QUAST, C., REX-HAFFNER, M., BINDER, E. B. and MÜLLER-MYHSOK, B. (2011). vipR: Variant identification in pooled DNA using R. *Bioinformatics* **27** i77–i84.

AMARAL, A. J., FERRETTI, L., MEGENS, H.-J., CROOIJMANS, R. P. M. A., NIE, H., RAMOS-ONSINS, S. E., PEREZ-ENCISO, M., SCHOOK, L. B. and GROENEN, M. A. M. (2011). Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS ONE* **6** e14782.

BANSAL, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* **26** i318–i324.

BENJAMINI, Y. and HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215–1222. MR2522270

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 289–300. MR1325392

BODMER, W. and BONILLA, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40** 695–701.

CALVO, S. E., TUCKER, E. J., COMPTON, A. G., KIRBY, D. M., CRAWFORD, G., BURTT, N. P., RIVAS, M., GUIDUCCI, C., BRUNO, D. L., GOLDBERGER, O. A., REDMAN, M. C., WILTSHIRE, E., WILSON, C. J., ALTSHULER, D., GABRIEL, S. B., DALY, M. J., THORBURN, D. R. and MOOTHA, V. K. (2010). High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nat. Genet.* **42** 851–858.

CHENG, C., WHITE, B. J., KAMDEM, C., MOCKAITIS, K., COSTANTINI, C., HAHN, M. W. and BESANSKY, N. J. (2012). Ecological genomics of Anopheles gambiae along a latitudinal cline: A population-resequencing approach. *Genetics* **190** 1417–1432.

CRAIG, D. W., PEARSON, J. V., SZELINGER, S., SEKAR, A., REDMAN, M., CORNEVEAUX, J. J., PAWLOWSKI, T. L., LAUB, T., NUNN, G., STEPHAN, D. A., HOMER, N. and HUENTELMAN, M. J. (2008). Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5** 887–893.

DAYE, Z. J., LI, H. and WEI, Z. (2012). A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Res.* **40** e60.

DRULEY, T. E., VALLANIA, F. L. M., WEGNER, D. J., VARLEY, K. E., KNOWLES, O. L., BONDS, J. A., ROBISON, S. W., DONIGER, S. W., HAMVAS, A., COLE, F. S., FAY, J. C. and MITRA, R. D. (2009). Quantification of rare allelic variants from pooled genomic DNA. *Nat. Methods* **6** 263–265.

EFRON, B. (2005). Bayesians, frequentists, and scientists. *J. Amer. Statist. Assoc.* **100** 1–5. MR2166064

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B. and MORRIS, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators. I. The Bayes case. *J. Amer. Statist. Assoc.* **66** 807–815. MR0323014

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

EFRON, B. and MORRIS, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* 311–319.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

ELSHIRE, R. J., GLAUBITZ, J. C., SUN, Q., POLAND, J. A., KAWAMOTO, K., BUCKLER, E. S. and MITCHELL, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6** e19379.

FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

FRAZER, K. A., MURRAY, S. S., SCHORK, N. J. and TOPOL, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10** 241–251.

GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 499–517. MR1924303

HAYDEN, E. C. (2008). International genome project launched. *Nature* **451** 378–379.

HE, L., SARKAR, S. K. and ZHAO, Z. (2012). Capturing the severity of type II errors in high-dimensional multiple testing. Technical report.

HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., RAMOS, E. M., MEHTA, J. P., COLLINS, F. S. and MANOLIO, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106** 9362–9367.

HUANG, X., FENG, Q., QIAN, Q., ZHAO, Q., WANG, L., WANG, A., GUAN, J., FAN, D., WENG, Q., HUANG, T., DONG, G., SANG, T. and HAN, B. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Res.* **19** 1068–1076.

KOLACZKOWSKI, B., KERN, A. D., HOLLOWAY, A. K. and BEGUN, D. J. (2011). Genomic differentiation between temperate and tropical Australian populations of Drosophila melanogaster. *Genetics* **187** 245–260.

LANDER, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature* **470** 187–197.

LI, B. and LEAL, S. M. (2009). Discovery of rare variants via sequencing: Implications for the design of complex trait association studies. *PLoS Genet.* **5** e1000481.

LI, H., RUAN, J. and DURBIN, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18** 1851–1858.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. and 1000 GENOME PROJECT DATA PROCESSING SUBGROUP (2009a). The sequence alignment/map format and SAMtools. *Bioinformatics* **25** 2078–2079.

LI, R., LI, Y., FANG, X., YANG, H., WANG, J., KRISTIANSEN, K. and WANG, J. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19** 1124–1132.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F. C., MCCARROLL, S. A. and VISSCHER, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

MARDIS, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature* **470** 198–203.

MARGRAF, R. L., DURTSCHI, J. D., DAMES, S., PATTISON, D. C., STEPHENS, J. E. and VOELKERDING, K. V. (2011). Variant identification in multi-sample pools by illumina genome analyzer sequencing. *J. Biomol. Tech.* **22** 74–84.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. and DEPRISTO, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20** 1297–1303.

MOMOZAWA, Y., MNI, M., NAKAMURA, K., COPPIETERS, W., ALMER, S., AMININEJAD, L., CLEYNEN, I., COLOMBEL, J.-F., DE RIJK, P., DEWIT, O., FINKEL, Y., GASSULL, M. A., GOOSSENS, D., LAUKENS, D., LÉMANN, M., LIBIOULLE, C., O'MORAIN, C., REENAERS, C., RUTGEERTS, P., TYSK, C., ZELENIKA, D., LATHROP, M., DEL-FAVERO, J., HUGOT, J.-P., DE VOS, M., FRANCHIMONT, D., VERMEIRE, S., LOUIS, E. and GEORGES, M. (2011). Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. *Nat. Genet.* **43** 43–47.

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849

NEJENTSEV, S., WALKER, N., RICHES, D., EGHOLM, M. and TODD, J. A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324** 387–389.

NORTON, N., WILLIAMS, N. M., O'DONOVAN, M. C. and OWEN, M. J. (2004). DNA pooling as a tool for large-scale association studies in complex traits. *Ann. Med.* **36** 146–152.

OUT, A. A., VAN MINDERHOUT, I. J. H. M., GOEMAN, J. J., ARIYUREK, Y., OSSOWSKI, S., SCHNEEBERGER, K., WEIGEL, D., VAN GALEN, M., TASCHNER, P. E. M., TOPS, C. M. J., BREUNING, M. H., VAN OMMEN, G.-J. B., DEN DUNNEN, J. T., DEVILEE, P. and HES, F. J. (2009). Deep sequencing to reveal new variants in pooled DNA samples. *Hum. Mutat.* **30** 1703–1712.

PRABHU, S. and PE'ER, I. (2009). Overlapping pools for high-throughput targeted resequencing. *Genome Res.* **19** 1254–1261.

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950* 131–148. Univ. California Press, Berkeley and Los Angeles. MR0044803

ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I* 157–163. Univ. California Press, Berkeley and Los Angeles. MR0084919

SARKAR, S. K., ZHOU, T. and GHOSH, D. (2008). A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. *Statist. Sinica* **18** 925–945. MR2440399

SHAM, P., BADER, J. S., CRAIG, I., O'DONOVAN, M. and OWEN, M. (2002). DNA pooling: A tool for large-scale association studies. *Nat. Rev. Genet.* **3** 862–871.

SMITH, A. M., HEISLER, L. E., ONGE, R. P. S., FARIAS-HESSON, E., WALLACE, I. M., BODEAU, J., HARRIS, A. N., PERRY, K. M., GIAEVER, G., POURMAND, N. and NISLOW, C. (2010). Highly-multiplexed barcode sequencing: An efficient method for parallel analysis of pooled samples. *Nucleic Acids Res.* **38** e142.

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657

SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 393–424. MR2649603

SUN, W. and WEI, Z. (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *J. Amer. Statist. Assoc.* **106** 73–88. MR2816703

TURNER, T. L., BOURNE, E. C., WETTBERG, E. J. V., HU, T. T. and NUZHDIN, S. V. (2010). Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nat. Genet.* **42** 260–263.

TURNER, T. L., STEWART, A. D., FIELDS, A. T., RICE, W. R. and TARONE, A. M. (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS Genet.* **7** e1001336.

VALLANIA, F. L. M., DRULEY, T. E., RAMOS, E., WANG, J., BORECKI, I., PROVINCE, M. and MITRA, R. D. (2010). High-throughput discovery of rare insertions and deletions in large cohorts. *Genome Res.* **20** 1711–1718.

WANG, W., WEI, Z. and SUN, W. (2010). Simultaneous set-wise testing under dependence, with applications to genome-wide association studies. *Stat. Interface* **3** 501–511. MR2754747

WEI, Z., SUN, W., WANG, K. and HAKONARSON, H. (2009). Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics* **25** 2802–2808.

WEI, Z., WANG, W., HU, P., LYON, G. J. and HAKONARSON, H. (2011). SNVer: A statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* **39** e132.

XIE, J., CAI, T. T., MARIS, J. and LI, H. (2011). Optimal false discovery rate control for dependent data. *Stat. Interface* **4** 417–430.

ZHAO, Z., WANG, W. and WEI, Z. (2013). Supplement to "An empirical Bayes testing procedure for detecting variants in analysis of next generation sequencing data." DOI:110.1214/13-AOAS660SUPP.

ZHU, Y., BERGLAND, A. O., GONZÁLEZ, J. and PETROV, D. A. (2012). Empirical validation
    of pooled whole genome population re-sequencing in Drosophila melanogaster. *PLoS ONE* **7**
    e41901.

Z. ZHAO   W. WANG
DEPARTMENT OF STATISTICS   Z. WEI
TEMPLE UNIVERSITY   DEPARTMENT OF COMPUTER SCIENCE
346 SPEAKMAN HALL   NEW JERSEY INSTITUTE OF TECHNOLOGY
1810 N. 13TH STREET   GITC 4400, UNIVERSITY HEIGHTS
PHILADELPHIA, PENNSYLVANIA 19122   NEWARK, NEW JERSEY 07102
USA   USA
E-MAIL: zhaozhg@temple.edu   E-MAIL: ww42@njit.edu
                                                         zhiwei@njit.edu
                                                 URL: http://ebvariant.sourceforge.net