THE FAULTY FALSE DISCOVERY RATE

Hoa Nguyen's Thesis Proposal

Abstract

The false discovery procedure introduced by Benjamini and Hochberg in 1995 has become a mainstream method for large scale simultaneous inference in a variety of bioinformatics problems. The procedure controls the false discovery rate (FDR) at a specified level α assuming that the distribution function F_0 of null p-values P_i is $U(0,1)$. In a recent paper, Efron (2004) brought to attention that, often, the empirical null p-values do not conform to the theoretical $U(0,1)$ and the biased distribution of nulls can affect the FDR. Indeed, linear regression settings aimed for genome-wide association study provide good examples of a biased F_0. Under these scenarios, the number of covariates p is much greater than the sample size n, which eliminates the option of fitting the full regression model. Nevertheless, a resolution of fitting an abundant number of partial models permits an empirical estimation of the distribution of null p-values.

In addressing the bias in F_0, it is more convenient to study the bias in the distribution function G_0 of z-values: $Z_i = \Phi^{-1}(P_i)$. Estimating the deviation of Z_i from the $N(0,1)$ is tantamount to estimating the departure of P_i from the $U(0,1)$. Efron (2004) proposed a location-scale correction to the empirical distribution G_0. In this proposal, we show that the bias in G_0 can not be represented by a location-scale alteration alone. We propose a skewness adaptation to G_0. We show that variants of a skewed G_0 can lead to better control of FDR compared with the default $N(0,1)$. To illustrate the procedure, we examine data which are generated using a stochastic process that creates polymorphisms on chromosomal regions. The data can be analyzed using regression models.

INTRODUCTION

Multiple hypothesis testing is a classical problem which has received renewed interests in the recent statistical literature. These interests are due largely to the advancement of scientific technologies in various areas of bioinformatics. For example, in gene expression studies, a typical microarray experiment requires testing the expression levels of thousands of genes simultaneously (Lander, 1999; Brown and Botstein, 1999; Dudoit et al., 2003). At a finer genomic scale, epidemiologists test hundred thousands of single nucleotide polymorphisms (SNP's) or blocks of SNP's for disease associated loci (Altshuler et al., 2001; Daly et al. 2001; Patil et al., 2001; Gabriel et al. 2002; Botstein and Risch, 2003).

From a scientific perspective, problems such as analyzing microarray and SNP data for disease association studies entail identification of a small percentage of interesting cases for further investigation. As such, while the primary statistical task is minimizing the false positive rate (controlling type I error), due to a large number of hypotheses tested, minimizing the expected ratio of false positives to the total number of rejections (controlling the false discovery rate FDR) is of greatest interest (Storey 2003).

The false discovery procedure, introduced by Benjamini and Hochberg in 1995, is a distribution free, finite sample method for choosing a p-value rejection threshold to control FDR. Instead of adhering faithfully to family-wise error rate control (Simes 1986, Hommel 1988, Hochberg 1988 and Rom 1990), FDR controls the proportion of false positives among rejected hypotheses. Apart from the scientific relevance of the procedure, FDR was proven by Benjamini and Hochberg (1995) to have greater power than the traditional Bonferoni method. These appealing characteristics of FDR have drawn momentous attention in the research community in recent years, engendering a rich FDR literature. We mention the following key extensions of the original framework.

Benjamini and Yekutieli (2001) relaxed the assumption of independent test statistics and extended FDR to a class which possessed positive regression dependence. Efron et al. (2001) considered a Bayesian model approach to obtain multiple testing procedures that control FDR. Storey (2002) reversed the testing process: first, fix the rejection region, then estimate the corresponding error rate. Storey's proposal gave rise an FDR testing procedure of increased power and accuracy. Genovese and Wasserman (2002, 2003) developed a stochastic framework and large sample theory for FDR, enabling a deeper understanding of the original procedure and how it compares to the traditional Bonferoni method.

The above mentioned articles, however, are content with the assumption that the p values of null hypotheses are uniformly distributed. In a recent paper, Efron (2004) brought to attention that, often, the distribution of the empirical nulls do not conform to the theoretical U(0,1). Efron's results concur with our investigation of the distribution of null p-values from simulated data sets.

Our simulation study is based on a regression framework: p values are obtained from testing whether particular coefficients in a linear model are significant. In two recent papers, Bunea et al. (2003) and Devlin et al. (2003), the connection between regression models and FDR is established. Specifically, using FDR to obtain the set of significant covariates leads to a consistent estimator of this set. Such a finding strengthens the established applicability of FDR, motivating our regression framework simulation study.

While the bias of the distribution of null p-values is apparent in our examples, it is not entirely clear how to correct for such a bias. One attempt is to assume a certain parametric model for the null p-values, estimate its distribution (by either a parametric or non-parametric method) and adjust the calculation of FDR accordingly.

Let the collection of null hypotheses be H_i, $i = 1, \ldots, N$ and the corresponding p-values P_i, $i = 1, \ldots, N$. Following Efron, we prefer to work with a transformed version of the P_i, namely Z_i:

$$Z_i \;=\; \Phi^{-1}(P_i),\;\; i = 1, \ldots, N,$$

Φ is the cumulative distribution of the standard normal. Understanding the deviation of Z_i from the $N(0,1)$ is tantamount to comprehending the departure of P_i from the $U(0,1)$.

From simulations, the biased distribution of the nulls results in empirical FDR as high as seventy percent. Efron (2004) proposed a location and scale correction for the distribution of Z_i while keeping the symmetric assumption. We found that Efron's correction gives good reduction of the empirical FDR.

We take a step further, allowing the presence of skewness in the distribution of the nulls and analyzing that skewness from both a parametric and a non-parametric point of view. We give a step-by-step partial bias correction for the null p-values. For the parametric approach, we use the skew normal density introduced by Azzalini in 1985. We find that a skew normal fit to the data gives considerable improvement of the empirical FDR over Efron's location and scale correction. For the non-parametric approach, we attempt to estimate the skewness locally by using the third derivative of the log-density at the mode. We then correct for the skewness by a proper transformation, and calculate the empirical FDR accordingly. Thus far, the non-parametric correction of the skewness shows little improvement compared to no correction, and deserves further research.

In this paper, we first discuss in greater detail the simulation method for the linear regression model and the underlying scientific motivations. We then propose specific steps to obtain the skewness parameter for the distribution of null z-values, and report preliminary results on the improvement of FDR control. Finally, we discuss limitations of the current methodology and propose specific future research steps.

SCIENTIFIC MOTIVATION AND SYNTHETIC DATA

It has been estimated that any two copies of the human genome differ from one another by approximately 0.1% of the total number of base pairs (Gibbs et al. 2003). These differences occur mostly at sites where a single historical mutational event took place. For instance, some chromosomes in the population may have a G (G ”allele”) at a specific site while others have an A (A ”allele”); these alleles are termed single nucleotide polymorphisms (SNP’s). There are approximately three million SNP’s on the human genome.

Since the set of SNP’s captured 90% of the genetic variation in the population, an international SNP mapping project (HapMap) has been launched (Gibbs et al. 2003). At present, there are limited SNP data available on the public domain, and most of these data sets do not register enough chromosomes for large scale association studies. Alternatingly, Hudson (2002)’s MS program can be used to simulate a set of SNP’s in a genomic region of a particular length. The program generates independent samples of SNP sets using the standard coalescent approach described in Kingman (1982), Hudson (1990) or Norborg (2001).

Recent studies (e.g. Reich et al., 2001; Gibbs et al., 2003; Botstein and Risch, 2003) have speculated that common diseases such as cancer and diabetes are caused by multiple genetic and environmental factors. As such, the search for genetic variants affecting liability to complex diseases demands substantial knowledge of both marginal and combined effect of risk factors. Our goal to use regression framework to analyze SNP data stems from an aspiration to ultimately understand better the combined genetic effects on phenotypic traits.

We use a modified version of Hudson (2002)’s program (Wall and Pritchard, 2003) to generate SNP’s on genomic regions where recombination ”cold spots” and ”hot spots” are present.

For each synthetic genomic zone, we create two cold spots of length 20,000 base pairs, separated by a hot spot of length 10,000 base pairs. The mutation rate is $\theta = 4N_e\mu = 5.6 \times 10^{-4} \times$ {#of base pairs in the region}; μ is the mutation rate per base pair, per generation; N_e=10,000 is the effective population size.

Each simulation registers one thousand chromosomes. Each chromosome has a set of initial SNP's and we retain only those SNP's which have minor allele frequency ≥ 0.10 for further analyses. From simulation, we see that pairwise correlation between adjacent SNP's can be high (Figure 1). Thus, using the complete set of SNP's for a linear model can lead to substantial redundancy and co-linearity problem. We use a method described in Rinaldo et al. (2004), namely H-clust, to choose tagging SNP's (tagSNP). A good set of tagSNP's will capture essential information about the genomic region under investigation (Zhang et al., 2002; Ackerman et al., 2003; Ke and Cardon, 2003; Sebastiani et al., 2003). The H-clust method uses the correlation matrix among all SNP's as the dissimilarity matrix for the hierarchical clustering method to identify tagSNP's.

REGRESSION AND FDR FRAMEWORKS

Regression Settings Each tagSNP serves as a covariate in the linear model, labeled from X_1 to X_m. Since a SNP is a bi-allelic marker, X_j only takes on two values; we assign 1 to the major allele and 0 to the other. Each response variable $Y_i, i = 1, \ldots, n$, is generated by a linear combination of the X_j's, altered by some stochastic fluctuations ϵ_i; n is the number of chromosomes.

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} + \epsilon_i.$$

The vector of β's is chosen such that the proportion of non-zero β reflects the proportion of significant loci. Let the number of non-zero β_j's be N; let $a = \frac{N}{m}$.

Assume that m is large and we are unable to fit the full model (with main effects and interaction terms). We then begin with fitting a marginal model for each X_j:

$$\widehat{Y}_{ij} = \hat{\beta}_{j0} + \hat{\beta}_{j1} X_{ij}.$$

Since Y is simulated with N main effects, least square estimators of β_j's for the marginal models are biased, leading to biased test statistics and p-values. This simple framework reflects actual linear regression scenarios when lurking variables severely bias regression results.

FDR Setting We obtain p-values P_i's from testing $\beta_{1j} = 0$. Let H_i's be the collection of tests corresponding to p-values P_i's, $i = 1, \ldots, m$. Let I_0 be the set of indices of null hypotheses, and I_1 be that of alternative hypotheses.

The FDR is defined to be the expected value of the false discovery proportion FDP, where FDP is the number of false rejections over the total number of rejections (Benjamini and Hochberg 1995). Congruent with the regression settings, a is the fraction of false nulls, $H_i \sim Bernoulli(a)$. Furthermore, let $P_i|H_i = 0 \sim F_0$, $P_i|H_i = 1 \sim F_1$. The sequential p-values rejection procedure (Benjamini and Hochberg 1995) to control FDR at level α includes the following steps:

(i) Order the p-values, $P_{(1)} \leq P_{(1)} \leq \ldots \leq P_{(m)}$

(ii) Choose $k = \max\{i : P_{(i)} \leq \frac{i}{m}\alpha\}$

(iii) Reject all $H_{(i)}, i = 1, \ldots, k$.

Write the marginal distribution of the p-values as:

$$\begin{aligned} P(P_i \leq t) &= P(P_i \leq t \;\&\; H_i = 1) + P(P_i \leq t \;\&\; H_i = 0) \\ &= P(P_i \leq t|H_i = 1)P(H_i = 1) + P(P_i \leq t|H_i = 0)P(H_i = 0) \\ &= aF_1(t) + (1 - a)F_0(t) = F(t). \end{aligned}$$

We can obtain the FDR from the distributions of P_i, for $i \in I_0$ and $i \in I_1$,

$$\begin{aligned} FDP(t) &= \frac{\sum_i (1 - H_i)I(P_i < t)}{\sum_i I(P_i < t)} \\ FDR(t) = E(FDP(t)) &\approx \frac{E(1/m \sum_i (1 - H_i)I(P_i < t))}{E(1/m \sum_i I(P_i < t))} \\ &\approx \frac{(1 - a)F_0(t)}{F(t)} \equiv R(t). \end{aligned}$$

The sequential p-values procedure is equivalent to choosing a threshold t^* such that $R(t^*) = \alpha$, which implies:

$$F(t^*) = \frac{(1 - a)F_0(t^*)}{\alpha}$$

(Genovese and Wasserman, 2003). Working with the normal scale, $Z_i = \Phi^{-1}(P_i)$, the original p-value settings translate to finding a threshold z^* such that

$$G(z^*) = \frac{(1 - a)G_0(z^*)}{\alpha},$$

where G is the marginal distribution function of Z_i.

Based on the analogy above, calculation of the empirical FDR amounts to specification of the marginal distributions G_0 and G for Z_i, $i \in I_0$ & $i \in I_1$, respectively. We use the empirical $\hat{G}$ for G.

SPECIFYING A SKEWED G_0

From simulations, we observe that the distribution F_0 of null p-values is biased away from the $U(0,1)$; consequently, the distribution G_0 of null z-values departs from the $N(0,1)$. Figure 2 shows the distribution of the biased nulls for an exemplary synthetic data set.

Efron (2004) suggests a location-scale correction for the distribution of the null z-values while keeping the symmetric assumption prior to the calculation of the FDR. However, as seen in Figure 2, the realized bias cannot be characterized by a location and scale alteration alone. The apparent skewness in the distribution of null z-values can play a significant role in misrepresenting the FDR (see Figure 3). We propose a skewness adaptation to the distribution of the null z-values prior to FDR calculation. Estimation of the skewness parameter can be carried out via either parametric or non-parametric approaches.

Local estimation of skewness, a non-parametric approach By assuming a normal fit to the distribution, Efron (2004) estimated the location and scale parameters using a non-parametric procedure. Natural estimators of the parameters are:

$$\mu = arg\,max\{\hat{f}(z)\} \text{ and } \sigma = \left[-\frac{d^2}{dz^2} log \hat{f}(z)\right]_{\mu}^{-1/2},$$

where $\hat{f}$ is an estimator of f. Efron's calculations amount to assuming that local behavior of the second derivative of the log-density at the mode reflects the true spread of the distribution. Following Efron, we propose using the third derivative of the log-density at the mode to estimate the amount of skewness in the distribution. For the estimation of the derivatives, we propose using the kernel method.

The kernel method has been studied extensively; for a review, see Scott (1992). The kernel density estimator of f at a point x is defined as:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

where $K_h(u) = K(u/h)/h$. Based on this definition, we can estimate the ν^{th} derivative of f in a similar manner:

$$\hat{f}^{(\nu)} = \frac{1}{nh^{\nu}} \sum_{i=1}^{n} K_h^{(\nu)}(x - X_i).$$

Loader (1999) established the connection of the kernel estimator with the class of local polynomial density estimation. Writing the log-likelihood function as:

$$L(f) = \sum_{i=1}^{n} log(f(X_i)) - n(\int_{\mathcal{X}} f(u)du - 1),$$

the localized version of the log-likelihood in a neighborhood of x is:

$$L_x(f) = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) log(f(X_i)) - n\int_{\mathcal{X}} K\left(\frac{u - x}{h}\right) f(u)du,$$

where K is a symmetric weight function. Assume that in a neighborhood of x the log likelihood can be approximated by a polynomial of degree p:

$$logf(u) = a_0 + a_1(u - x) + \frac{a_2}{2!}(u - x)^2 + \ldots + \frac{a_p}{p!}(u - x)^p.$$

Denote this polynomial by $P_x(a, u)$ the local likelihood becomes:

$$L_x(a_0, a_1, \ldots, a_p) = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) P_x(a, X_i) - n\int_{\mathcal{X}} K\left(\frac{u - x}{h}\right) exp(P_x(a, u))du.$$

Let $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_p)$ be the MLE's of the local likelihood, $\hat{a}_i$ is then the local estimate of the i^{th} derivative of the log of the density f.

When the local polynomial is a constant (p=0), the local likelihood density estimate coincides with the kernel density estimate:

$$\hat{f}(x) = exp(\hat{a}_0) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}).$$

The density estimator, as a consequence, is part of the family of local polynomial density estimators. Thus far, we have only worked with the kernel estimator.

We propose using the third derivative at the mode as a measure of local skewness. Let W_i's be the centered Z_i's. The local skew parameter can guide us to choose a proper power transformation of the $V_i = h(W_i)$ such that the distribution of V_i is approximately symmetric. Once the skewness is removed, i.e. V_i's are specified, we can carry out a scale correction for the distribution of V_i prior to calculating the empirical FDR.

Global estimation of skewness, a parametric approach Azzalini (1985) introduced a class of skew-normal distributions which allows the presence of skewness in the normal distribution. In this report, we will mention the univariate skew normal, the multivariate version of the skew normal can be found in Azzalini and Capitanio (1999).

A random variable Z is said to have a skew-normal distribution with parameter λ if its density is:

$$\varphi(z;\lambda) \;=\; 2\phi(z)\Phi(\lambda z)$$

where ϕ and Φ are the standard normal density and distribution functions, λ is the skew parameter. In practice, we often work with the family of distributions generated by a linear transformation:

$$Y = \xi + \omega Z.$$

Z can be viewed as a standard skew normal with mean $E(Z) = \sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}$ and variance $Var(Z) = 1 - E^2(Z)$. Y can be viewed as a skew normal with location, scale parameters (ξ, ω^2). The density of Y is then:

$$2\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left(\lambda\frac{y-\xi}{\omega}\right).$$

Azzalini (1985), Azzalini and Capitanio (1999) discussed issues relating to estimation of the skew normal parameters. In particular, the Fisher information matrix becomes singular near $\lambda = 0$. This problem can be remedied by a re-parametrization with (μ, σ, λ) (Azzalini 1985, Azzalini and Capitanio 1999, Chiogna 1997):

$$Y = \mu + \sigma\frac{Z - \mu_z}{\sigma_z}.$$

For the estimation of the MLE's, a gradient-based method can be employed. Azzalini and Capitanio (1999) discussed an EM algorithm which entailed the introduction of a fictitious unobserved variable. The algorithm offers reliable estimates, especially when the initial values for the parameters are chosen by the method of moments.

In estimating the parameters of the skew-normal to obtain the marginal distribution $\hat{G}_0$ of Z_i, we face another challenge due to the fact that we do not have the set of $Z_i, i \in I_0$. If all $Z_i, i = 1, \ldots, m$ are used, we end up having large bias in the tail of the distribution, which leads to an overly conservative empirical FDR. This problem is not severe when the local estimator of the skewness at the mode is used since local estimators give high weights to observations in a neighborhood of the mode. To remedy the problem of unknown I_0 for the global estimate of the skewness, we use a pilot estimate of the mode and the spread to select a pilot set of $Z_i, i \in I_0$. Let g_0 be the density of $Z_i, i \in I_0$, here are the specific steps:

(i) Get pilot estimate of $\hat{g}_0(z_i)$ using the kernel density

(ii) Obtain the mode $\hat{\mu}_0 = arg\, max(\hat{g}_0(z))$

(iii) Get estimate of the second derivative of $\frac{d^2}{dz^2} log(g)$ using the second derivative c_2 of a smoothing spline of $(z, log(\hat{g}_0(z_i))$

(iv) Obtain the spread $\hat{\sigma}_0 = \sqrt{(-1/c_2)}$

(v) Use $Z_i \in (\hat{\mu}_0 - 3\hat{\sigma}_0, \hat{\mu}_0 - 3\hat{\sigma}_0)$ as pilot set $Z_i, i \in I_0$ for the estimation of the skew normal parameter for G_0.

FDR CONTROL WITH SKEWNESS CORRECTION

In the simulation study, we set α to 0.05. Due to various sources of bias in the choice of the distribution of null Z_i's, the empirical FDR is not controlled at the desired level α. In some cases, the empirical FDR can be extreme (see Figure 4, panel (a)). However, with skewness adjustment in the choice of the distribution of null Z_i's, we can reduce the bias prior to carrying out FDR procedure. Given the marginal distribution $\hat{G}_0$ of $Z_i, i \in I_0$, FDR is calculated as follows:

(i) Calculate $R(z) = \frac{\widehat{G}_0(z)}{\widehat{G}(z)}$

(ii) Choose $z^* = max\ z$ such that $R(z^*) \leq \alpha$

(iii) Reject all H_i for which $Z_i \leq z^*$

(iv) For rejection threshold z^*, calculate the actual FDR using knowledge of I_0 and I_1.

Figure 4 shows FDR reduction with skewness correction in the distribution of null z-values using the global skewness estimator approach. FDR can be as high as 70% with an average of 58% if the chosen G_0 is N(0,1). Using Efron's location-scale correction, FDR drops to an average of 35%. The skew-normal specification of G_0 reduces FDR to an average of 16%. The skewness correction using a local estimator such as the kernel shows little improvement over no correction (figure not shown), and awaits further research.

DISCUSSION AND FUTURE WORK

The false discovery procedure has become a mainstream method for multiple testing in a variety of bioinformatics problems. Efron (2004) brought to attention that FDR calculation can be misled by a choice of the distribution of null z-values. Particularly, when that distribution is not N(0,1), it is hard to quantify the rejection region. We have addressed the same problem here.

While the bias in the distribution of null z-values G_0 is apparent, it is not entirely clear how to correct for such a bias. Standard practice has taken the N(0,1) as a default for G_0. Efron (2004) proposed keeping the symmetric assumption of G_0 while correcting for the location and scale parameters. In this paper, through simulations in genetics and linear model settings, we show that deviations of the distribution of null z-values from the N(0,1) can not be quantified by a location-scale shift alone. Rather, the discernable skewness can misdirect calculation of the FDR.

We proposed using a skew G_0 for the FDR procedure. The skewness parameter can be estimated using a global or local approach. So far, we have achieved better control of FDR using the global skewness correction: representing G_0 by a skew-normal distribution. The local approach to estimation of the skewness parameter remains a challenge. We deem it be critical to explore further the local polynomial approach to improve estimation of the skewness so as to obtain better FDR control. In the immediate future, we would like to examine further the following directions:

- Even though we have achieved better FDR control with the global fit of a skew-normal to the data, FDR is still higher than the preset $\alpha = .05$ level. We would like to attain α level control of FDR by (1) examining whether the features of the tail in the distribution of null z-values plays a significant role in determining FDR and (2) studying how the level of separation between the distribution of null z-values and that of alternative z-values affects FDR calculation.

- The skew normal approach to obtain a global skewness parameter can potentially further improve FDR control. In this paper, we use a pilot index set I_0 of null z-values to obtain the skew parameter. The pilot I_0 is based on local estimation of the mode and the spread of the distribution. These estimators are biased in their own way (see technical appendix which gives details on the bias of these estimators). We would like to get a better pilot estimate of the set I_0 by studying the choice of δ for the window $(\hat{\mu}_0 - \delta\hat{\sigma}_0, \hat{\mu}_0 - \delta\hat{\sigma}_0)$.

- Alternatively, instead of choosing the pilot set I_0 using a window around the mode, we can use the one-sided interval $(z_m, +\infty)$. The choice of the threshold z_m reflects a bias-variance tradeoff in estimating the skew normal parameters.

- Assuming that the true distribution of the null z-values is N(0,1), we would like to study power loss resulting from applying the skew-normal procedure.

- The FDR is known to be conservative (Genovese and Wasserman, 2002). Recent work by Storey (2003) and Benjamini et al. (2003) revealed that estimating the proportion m_0 of null hypotheses and incorporating such information into the FDR procedure could improve power. We would like to examine if such suggestion can balance the power loss due to applying the skew-normal procedure when the actual distribution of null z-values is N(0,1).

- This research is motivated by the genetics problem of analyzing SNP data to identify SNP's associated with complex diseases. Thus far, we have analyzed the marginal model for each locus. Ultimately, we would like to study models in which interactions among SNP's are present.

- Population substructure can complicate disease association studies. In that respect, we would like to explore admixture mapping in our simulation study and how it affects the ability to detect liability genetic variants.

TECHNICAL APPENDIX

The technical appendix is available on the Department of Statistics Private area.

References

Ackerman, H., Usen, S. Mott, R., Richardson, A., Sisay-Joof, F., Katundu, P., Taylor, T, Ward, R., Molyneux, M., Pinder, M., Kwiatkowski, D. (2003). Haplotypic analysis of the TNF locus by association efficiency and entropy. *Genome Biology*, **2003**, 4(4)-R24.

Altshuler, D.A., V.J. Pollara, C. Cowles, W.J. Van Etten, J. Baldwin, et al. (2001). A SNP map of the human genome generated by reduced representation shortgun sequencing. Nature **407**: 513-516.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171-178.

Azzalini, A. and Capitanio A. (1999). Statistical applications of the multivariate skew-normal distribution. *J. Roy. Statist. Soc., B*, **61**, 579-602.

Benjamini Y. and Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289-300.

Benjamini Y. and Yekutieli D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165-1188

Benjamini, Y., Krieger, M. and Yekutieli, D. (2003). Adaptive linear step-up procedures that control the false discovery rate.

Botstein, D., and N. Risch (2003). Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nature Genetics Supplement **33**: 228-237.

Brown, P.O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics Supplement.* **21**, 33-37.

Bunea F., Nu X. and Wegkamp M. (2003). The consistency of FDR estimator. Technical Report, Dept. of Statistics, Florida State University.

Chiogna, M. (1997). Notes on estimation problems with scalar skew-normal distributions. Technical report 1997.15, Department of Statistical Sciences, University of Padua.

Daly, M.J., J.D. Rioux, S.F. Schaffner, T.J. Hudson, and E.S. Lander (2001). High resolution haplotype structure in the human genome. Nature Genetics **29**: 229-232.

Devlin B., Roeder K. and Wasserman L. (2003). Analysis of multilocus models of association. *Genetics Epidemiology* **25**, 36-47.

Dudoit S., Popper S., Boldrick J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18** 71-103.

Efron, B., Tibshirani R., Storey J. and Tusher V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, **96**, 1151-1160

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Associations* **99**, 96-104

Gabriel, S.B., S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, et al. (2002). The structure of haplotype blocks in the human genome. Science **296**: 2225-2229.

Genovese C. and Wasserman L. (2002). Operating characteristics and extensions of the FDR procedure. *J. Roy. Statist., Soc. B* **64**, 499-518.

Genovese C. and Wasserman L. (2003). A stochastic approach to false discovery rates. Technical Report 762, Dept. of Statistics, Carnegie Mellon University.

Gibbs, R., Belmont, J., Hardenbol, P., Willis, T., Yu, F., Yang, H., Ch'ang, L., et al. (2003). The international HapMap project. *Nature* **426**, 789-796.

Henze, N. (1986). A probabilistic representation of the skew-normal distribution. *Scand. J. Statist.* **13**, 271-275.

Hommel, G. (1988). A stagewise rejection multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383-386

Hocberg, Y. (1988). A sharper bonferoni procedure for multiple tests of significance. *Biometrika* **75**, 800-803.

Hudson, R. (1990). Gene genealogies and the coalescent process. In Futuyma, D. and Antonvics, J., editors, *Oxford Surveys in Evolutionary Biology, Vol. 7*, 1-43. Oxford University Press: Oxford, UK.

Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-338.

Ke, X., Cardon, L. (2003). Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19**, 287-288,

Kingman, J. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19A**, 27-43.

Lander, E.S. (1999). Array of hope. *Nature Genetics Supplement.* **21**, 3-4.

Loader C. (1999). Local Regression and Likelihood. Springer-Verlag: New York.

Miller, R. (1981). Simutaneous statistical inference. Springer-Verlag: New York.

Nordborg, M. (2001). Coalescent theory. In Balding, D., Bishop, K. and Cannings, C., editors, *Handbook of Statistical Genetics*, 179-212. John Wiley and Sons: Chichester, UK.

Reich, D., Cargill, M., Bolk, S., Ireland, J., Sabeti, P., Richter, D., Lavery, T., Kouyoumjian, R., Farhadian, S., Ward, R., Lander, E. (2001). Linkage disequilibrium in the human genome. *Nature* **411**, 199-204.

Rinaldo, A., Bacanu, S., Devlin, B., Sonpar, V., Wasserman, L., and Roeder, K. (2004). Characterization of multilocus linkage disequilibrium.

Rom, D. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77**, 663-665.

Sebastini, P., Lazarus, R., Weiss, S., Kunkel, L., Kohane, I., Ramoni, M. (2003). Minimal haplotype tagging. *Proceedings of the National Academy of Sciences* **100**, 9900-9905.

Scott, D. (1992). Multivariate density estimation: theory, practice and visualization. Wiley-Interscience: New York.

Simes, J. (1986). An improved Bonferoni procedure for multiple tests of significance. *Biometrika* **73**, 751-754.

Storey J. and Tibshirani R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences* **100** 9440-9445.

Storey J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64** 479-498.

Wall, J., and Prichard, J. (2003). Assessing the performance of the haplotype block model of linkage disequilibrium. *The American Journal of Human Genetics* **73**, 502-516.

Westfall P. and Young S. (1993). Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley-Interscience: New York.

Zhang, K., Deng, M., Chen, T., Waterman, M., Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences*, **99(11)**, 7335-7339.

Zhang, K., Peter, C., Nordborg, M., and Sun, F. (2002). Haplotype block structure and its application to association studies: power and study designs. *American Journal of Human Genetics* **71**, 1386-1394.

FIGURE CAPTION

Figure 1: Image plot of the correlation matrix of 79 ,SNP's in a genomic region. Lighter colors signify high correlation.

Figure 2: Histograms of null p-values and the corresponding z-values for an exemplary data set.

Figure 3: Histogram of null z-values overlaid by fitted density functions. Blue curve is the $N(0,1)$; brown curve is $N(-0.3, 1.36)$ (Efron's location-scale correction); red curve is the $SN(0.97, 2.18, -2.53)$. Parameters for Efron's location-scale and skew-normal corrections are estimated from the data.

Figure 4: Empirical FDR resulting from (a) No correction (b) Efron's location and scale correction (c) Skew normal correction. The plots are generated from 300 simulations.

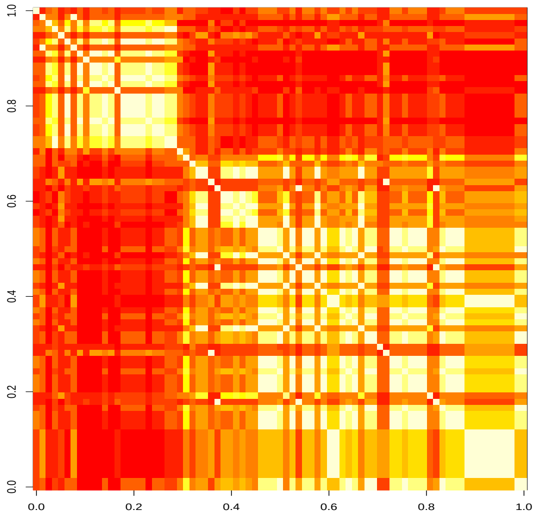

Figure 1:

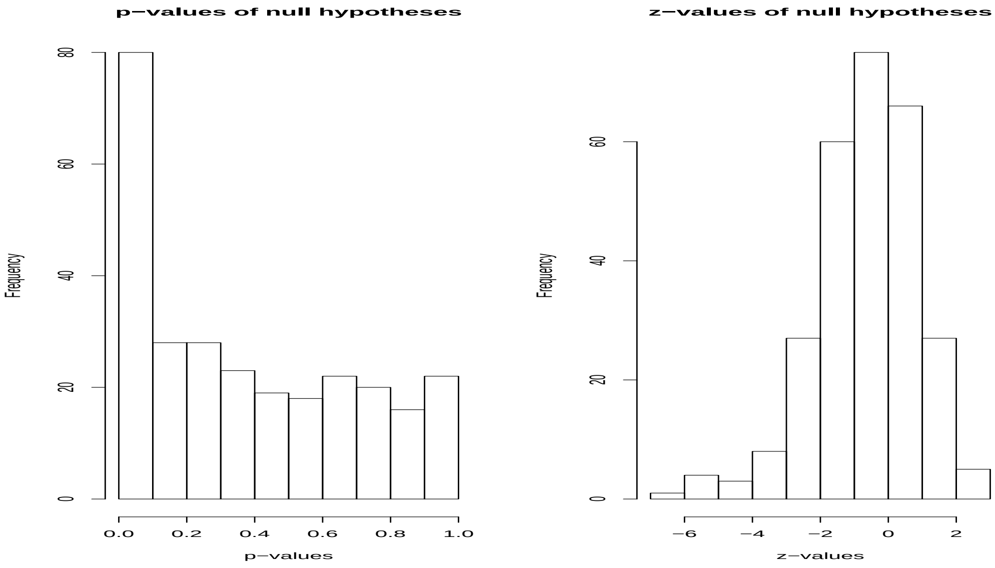

Figure 2:

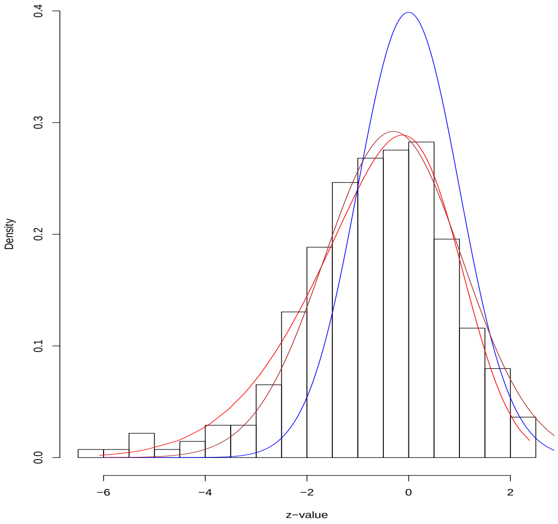

Figure 3:

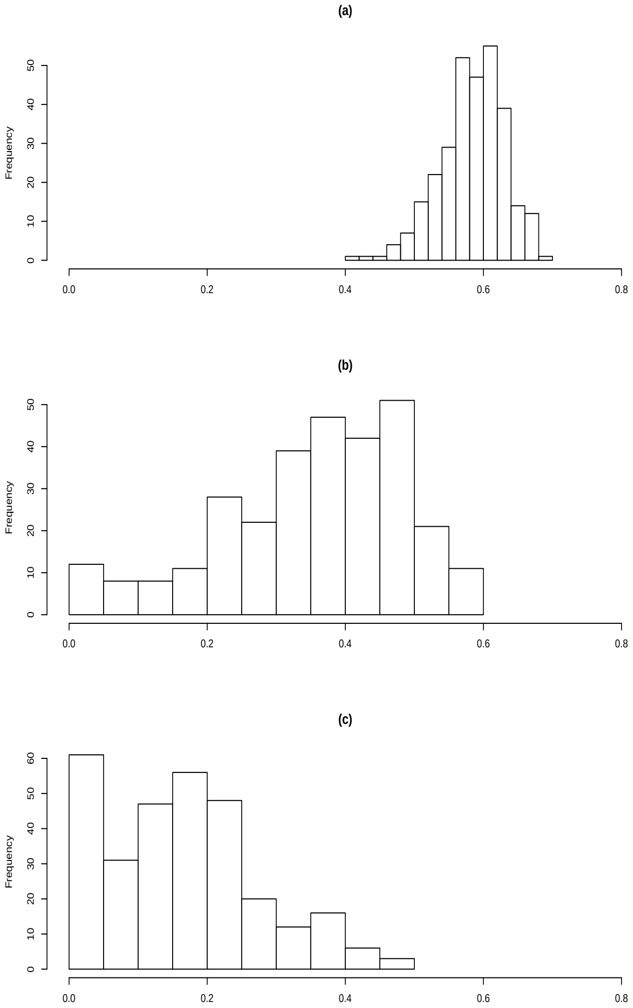
(a)
Frequency
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
(b)
Frequency
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
(c)
Frequency
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8

Figure 4: