frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 31 May 2013
doi: 10.3389/fgene.2012.00202

Integrative genomics: quantifying significance of phenotype-genotype relationships from multiple sources of high-throughput data

Eric R. Gamazon[1], R. Stephanie Huang[1], M. Eileen Dolan[1], Nancy J. Cox[1,2] and Hae Kyung Im[3]*

[1] Department of Medicine, University of Chicago, Chicago, IL, USA
[2] Department of Human Genetics, University of Chicago, Chicago, IL, USA
[3] Department of Health Studies, University of Chicago, Chicago, IL, USA

Edited by:
Barbara E. Stranger, Brigham and Women's Hospital, USA

Reviewed by:
Ali Torkamani, University of California at San Diego, USA
Marylyn D. Ritchie, The Pennsylvania State University, USA
Qingying Meng, University of California Los Angeles, USA

***Correspondence:**
Hae Kyung Im, Department of Health Studies, University of Chicago, 5841 S Maryland Avenue MC 2007, Chicago, IL 60637, USA.
e-mail: haky@uchicago.edu

Given recent advances in the generation of high-throughput data such as whole-genome genetic variation and transcriptome expression, it is critical to come up with novel methods to integrate these heterogeneous datasets and to assess the significance of identified phenotype-genotype relationships. Recent studies show that genome-wide association findings are likely to fall in loci with gene regulatory effects such as expression quantitative trait loci (eQTLs), demonstrating the utility of such integrative approaches. When genotype and gene expression data are available on the same individuals, we and others developed methods wherein top phenotype-associated genetic variants are prioritized if they are associated, as eQTLs, with gene expression traits that are themselves associated with the phenotype. Yet there has been no method to determine an overall *p*-value for the findings that arise specifically from the integrative nature of the approach. We propose a computationally feasible permutation method that accounts for the assimilative nature of the method and the correlation structure among gene expression traits and among genotypes. We apply the method to data from a study of cellular sensitivity to etoposide, one of the most widely used chemotherapeutic drugs. To our knowledge, this study is the first statistically sound quantification of the overall significance of the genotype-phenotype relationships resulting from applying an integrative approach. This method can be easily extended to cases in which gene expression data are replaced by other molecular phenotypes of interest, e.g., microRNA or proteomic data. This study has important implications for studies seeking to expand on genetic association studies by the use of omics data. Finally, we provide an R code to compute the empirical false discovery rate when *p*-values for the observed and simulated phenotypes are available.

Keywords: eQTLs, FDR, gene expression, genomics, GWAS, integrative genomics, permutation, phenotype

INTRODUCTION

The availability of genome-wide datasets is facilitating unprecedented insights into various aspects of cellular processes. Technological advances (Metzker, 2010) in high-throughput methods are contributing to new approaches in genomics, transcriptomics (Wang et al., 2009), proteomics (Farnham, 2009), and epigenomics (Laird, 2010; Zhou et al., 2011), allowing in-depth interrogation of diverse biological processes. A primary challenge from the tremendously heterogeneous and increasingly massive datasets is data integration – a challenge that is inevitably bound to intensify with the deluge of these high-throughput datasets. Nevertheless, among the many exciting promises, integrative approaches are likely to yield a comprehensive map of genome function (Degner et al., 2012) as well as a high-resolution view into the complex logic of biological systems (Hawkins et al., 2010).

Indeed, while genome-wide association studies (GWAS) have identified thousands of common genetic variants associated with diseases and other complex human traits (Hindorff et al., 2009), functional understanding of many of the variants remains elusive. Integrating other omics datasets into genome-wide analyses offers the potential to provide systematic insight into the mechanisms underlying the observed genotype-phenotype relationships. One common approach to the integration of functional data into GWAS is the use of expression quantitative trait loci (eQTL; Stranger et al., 2007a; Duan et al., 2008; Schadt et al., 2008) information to expand on the nature of the genetic component to complex phenotypes (Gamazon et al., 2010a; Nicolae et al., 2010). Such an integrative approach is clearly extensible to the use of protein (Garge et al., 2010) or microRNA quantitative trait loci (Gamazon et al., 2012), indeed other functionally relevant features of the genome, to improve identification of functional variants.

Our group (Huang et al., 2007a; Welsh et al., 2009; Nicolae et al., 2010) and others (Cheung et al., 2003; Correa and Cheung, 2004; Stranger et al., 2007b; Nica et al., 2010) have used the HapMap lymphoblastoid cell lines (LCLs) as a model for human genotype-phenotype relationships. The cell lines have been the subject of several whole-genome gene expression profiling studies (Montgomery et al., 2010; Pickrell et al., 2010; Stranger et al., 2012) to

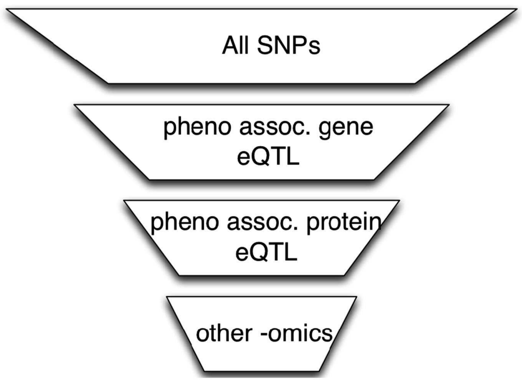

FIGURE 1 | The use of omics data to perform SNP filtering. A series of functional filters is applied to the original SNP set to perform data dimensional reduction. The reduced multiple testing burden may improve power to detect genotype associations with phenotype.

identify functional loci (e.g., eQTLs) with potentially important links to SNP associations emerging from genome-wide studies. Furthermore, the cell lines have been utilized to identify the molecular consequences associated with various exposures (Dermitzakis, 2012), such as drugs (Huang et al., 2007b), small molecules, or pathogens (Ko et al., 2009). For example, a three-way "triangle" model, correlating genotype, gene expression, and phenotype data, has been devised to identify genetic variants that contribute to chemotherapeutic-induced cytotoxicity through their effects on gene expression (Huang et al., 2007b). Nevertheless, quantifying the significance of a finding from such an integrative approach remains to be fully addressed.

MATERIALS AND METHODS

FUNCTIONAL INTEGRATION

A simple approach to integrate high-throughput functional datasets (e.g., from studies of the transcriptome, proteome, or microRNAome) with genome-wide genotype data obtained from microarray- or sequencing-based studies is to select SNPs that meet certain functional criteria as illustrated in the example in **Figure 1**. In the first step of this example, SNPs are filtered by requiring that they be associated with genes whose expression levels are associated with the phenotype (Zhong et al., 2010). In the next step, we further reduce the number of SNPs by requiring that they be associated with protein levels that are themselves associated with the phenotype. This process can continue using other omics datasets.

To simplify the description, we focus on the case in which only the gene (mRNA) expression data are integrated, which is depicted with the diagram in **Figure 2**. This triangle approach and variations thereof were proposed by Huang et al. (2007b) and others (Zhong et al., 2010) and applied to an array of cellular phenotypes. The first step of this method aims to identify a set of gene expression traits associated with the given phenotype at an arbitrarily set p-value threshold, $p < p_{\text{gene-phenotype}}$. It is important to emphasize that this threshold, as in the subsequent thresholds to be defined below, is generally set arbitrarily. In practice, these thresholds are used to prioritize genes or SNPs for downstream analyses. Indeed, one aim of our study is to quantify the significance of an association from a triangle method regardless of the choice of thresholds used during the integrative process. The second step of the method is to identify SNPs that are associated with the selected gene expression traits again at an arbitrarily set threshold, $p < p_{\text{SNP-gene}}$. At a stringent threshold, this step maps the gene expression traits to genomic loci; this step thus identifies the eQTLs for the corresponding genes. Finally, in the last step of the triangle, the resulting SNPs are interrogated for association with the phenotype. Our primary aim is to describe a method to quantify the significance of the SNPs resulting from this multi-step "triangle" approach.

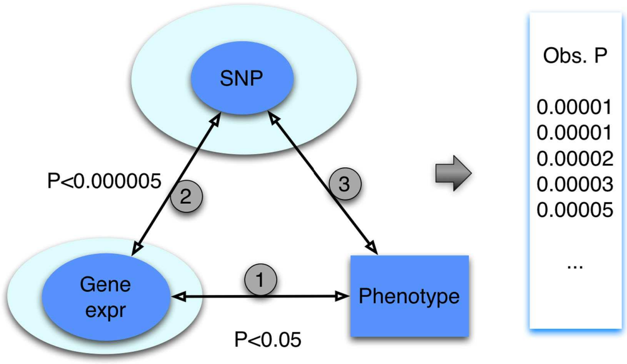

FIGURE 2 | The "triangle" method for integration of genotype, gene expression, and phenotype data. Through a series of steps, heterogeneous datasets, involving SNPs, gene expression and trait are integrated. At each step, a p-value threshold is applied. In general, the p-value threshold used is arbitrary; in practice, the choice allows for prioritization of genes or SNPs. The result of the triangle method is a set of SNP association p-values (represented by the "obs p" in the figure).

NAÏVE FDR OF SELECTED SNPs

Since the triangle method is a multi-step approach that derives a final SNP set from a series of (potentially) increasingly stringent thresholds, it is reasonable to expect that such an approach should yield a final set with substantially reduced false discovery rates (FDRs) for association with the phenotype. A simple approach to assess the significance of the findings for this subset of SNPs would be to compute the FDR for them (Storey and Tibshirani, 2003). We illustrate the problem of this approach in **Figure 3** in which we show the QQ plot of the associations after applying the triangle method to a simulated phenotype, which has no association with genotype. In this particular example, the first threshold $p_{\text{gene-phenotype}}$ was set at 0.05 while $p_{\text{SNP-gene}}$ was set at 5×10^{-6}. Circles above the red line represent SNPs with FDR < 0.05. (Strictly speaking, circles with p-values less than the one with the largest p-value that goes above the red line has FDR < 0.05.) As the figure indicates, the triangle method may yield several spurious associations, if we rely on a "naïve" FDR approach. This example shows the need to develop a more sophisticated approach to estimate the significance of results in this integrative context.

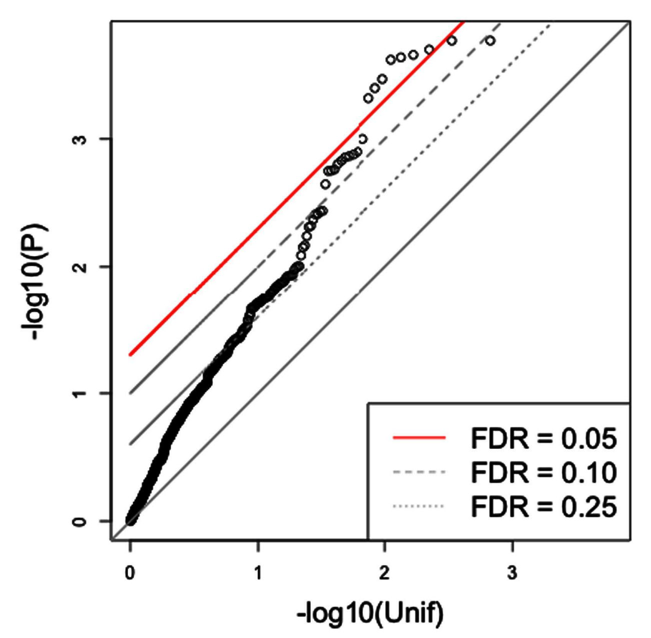

FIGURE 3 | Traditional FDR applied to the triangle method for simulated data. The triangle method may yield numerous highly significant SNP associations, based on the traditional FDR approach, from a simulated phenotype. In this case, several significant SNPs are obtained from simulated data although none should be observed.

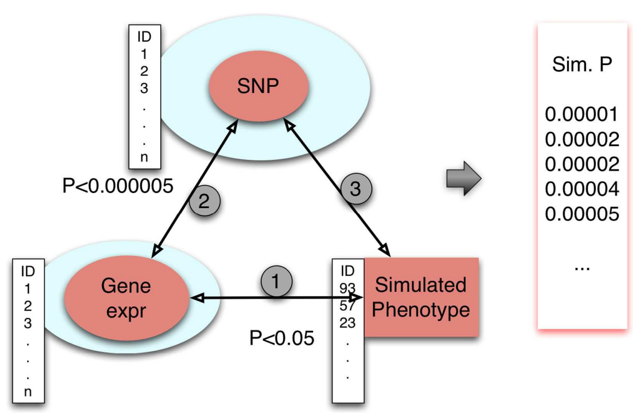

FIGURE 4 | Simulating the null distribution. Simulated phenotypes were obtained by permuting phenotype values. Associations between gene expression traits and phenotype were done conditional on preserved correlation structure between the gene expression traits. The eQTLs for selected genes were retrieved from our eQTL database SCAN. Finally, the eQTLs were tested for association with phenotype. For each of $n = 1000$ replicates, a set of SNP association p-values is generated.

SIMULATING THE NULL DISTRIBUTION

We describe here our approach to generating an empirical null distribution of p-values (**Figure 4**). First, let $Y_1, Y_2, Y_3, \ldots, Y_n$ be simulated phenotypes obtained from permuting the phenotype data. (Typically, $n = 1000$.) In case covariates are used, they should be relabeled in sync with the phenotype. For each simulated phenotype, we apply the same triangle method. For each Y_i, we derive the set of gene expression traits g_{ij} that meet the threshold, p-value $< p_{\text{gene-phenotype}}$, where the associations between the phenotype Y_i and gene expression traits are calculated while preserving the correlation structure of all gene expression phenotypes. For each g_{ij}, we retrieve the set of eQTLs, S_{ijk}, associated with the gene at the pre-defined threshold, p-value $< p_{\text{SNP-gene}}$. The subset of these eQTL SNPs that satisfy p-value $< p_{\text{SNP-phenotype}}$ provides a set of p-values $\{P_{ijk'}\}$, for each simulated phenotype Y_i. Note that each such set $\{P_{ijk'}\}$ of p-values may differ in count between simulated phenotypes. Note that i indexes simulations, j indexes genes, and k indexes eQTLs.

We utilize these sets of p-values derived from simulated phenotypes to estimate the null distribution of p-values. Having shown the limitation of the use of the traditional FDR for the integrative triangle method, we derive a simple formula to estimate the FDR using this empirical null distribution.

EMPIRICAL FDR

We closely follow Storey's approach (Storey and Tibshirani, 2003) to estimate the FDR. The difference in our approach is that we do not assume that the null distribution of p-values is uniform. Instead, we use the empirical distribution generated by simulating the phenotype and performing the integrative analysis. We define the significance level t and reject the null hypothesis of no association for all p-values smaller than t. We use the actual values in the observed vector of p-values as cutoff. Thus, for each p-value, t, in the observed vector of p-values, we compute the FDR of the strategy of rejecting all p-values less than or equal to t. Let the number of falsely significant SNPs be denoted as $F(t) = \#\{\text{null } p_i \le t, i = 1, \ldots, m\}$ and the number of significant SNPs be denoted as $S(t) = \#\{p_i \le t, i = 1, \ldots, m\}$ with m the total number of SNPs after applying the integrative approach. We estimate the FDR as follows:

$$\begin{aligned} FDR(t) &= E\left[\frac{F(t)}{S(t)}\right] \\ &\approx \frac{E[F(t)]}{E[S(t)]} \qquad (1) \\ &= \frac{mP(p \le t \text{ and null})}{mP(p \le t)} \\ &= \frac{P(p \le t \text{ and null})}{P(p \le t)} \qquad (2) \end{aligned}$$

where $E[.]$ is the expectation operator. The approximate equality in Eq. 1 is proven by Storey (2003).

The denominator is estimated using the observed number of significant SNPs $p \le t$,

$$\#\{p_{\text{obs},i} \le t,\ i = 1,\ \ldots, m\}/m$$

The numerator can be factored as $P(p \le t \text{ and null}) = P(p \le t \mid \text{null}) \cdot P(\text{null})$. The first factor $P(p \le t \mid \text{null})$ is estimated using the empirical distribution: $\#\{p_{\text{sim},i} \le t, i = 1, \ldots, M_0\}/M_0$ where the p_{sim}'s are the p-values generated with the simulated phenotypes and M_0 is the sum (across all simulations, $M_0 = \Sigma m_{o,s}$, where $m_{o,s}$

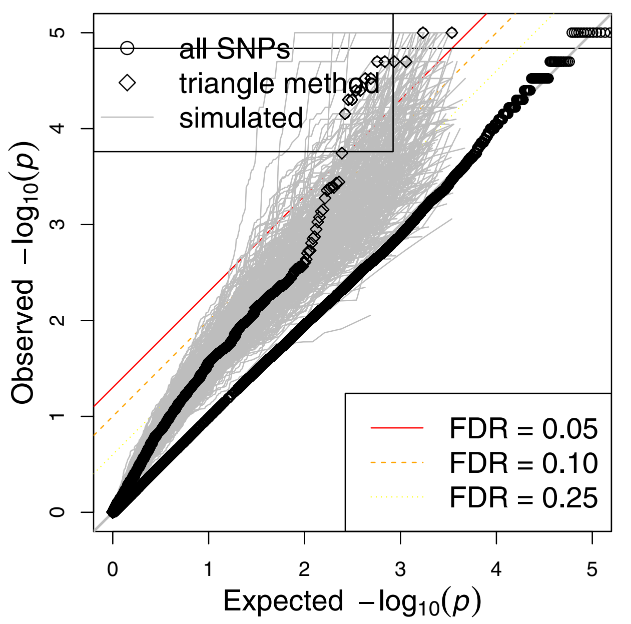

FIGURE 5 | Distribution of *p*-values for all SNPs and triangle method-derived SNPs from actual phenotype as well as simulated phenotypes. QQ plots are shown for all SNPs as well as the SNPs that come of the triangle method. Gray dots represent the QQ plots for the triangle method-derived SNPs from 1000 simulated phenotypes. Note that the triangle method may yield spurious associations if we rely on the traditional FDR.

corresponds to the number of eQTLs selected after applying the triangle method to the simulated phenotype Y_s) of the total number of SNPs selected using the simulated phenotypes. Note that for uniformly distributed p-values, we would have $P(p \leq t \mid \text{null}) = t$. We know, however, that when the set of SNPs are derived from the integrative approach, the null p-values may not be distributed uniformly, as illustrated in **Figure 5**. The second factor $P(\text{null})$ is the proportion of SNPs that are unrelated to the phenotype and may be estimated as the ratio

$$\hat{\pi}_0 = \frac{\#\left\{p_{\text{obs},i} > t,\ i = 1,\ \ldots, m\right\}/m}{\#\left\{p_{\text{sim},i} > t,\ \ i = 1,\ \ldots, M_0\right\}/M_0}$$

or may be set to 1 to yield a more conservative estimate of FDR.

In summary, we estimate the FDR based on the empirical distribution as follows:

$$\widehat{\text{eFDR}}(t) = \frac{\#\left\{p_{\text{sim}} \leq t\right\}/M_0}{\#\left\{p_{\text{obs}} \leq t\right\}/m} \cdot \frac{\#\left\{p_{\text{obs}} > \lambda\right\}/m}{\#\left\{p_{\text{sim}} > \lambda\right\}/M_0} \qquad (3)$$

$$= \frac{\#\left\{p_{\text{sim}} \leq t\right\}}{\#\left\{p_{\text{obs}} \leq t\right\}} \cdot \frac{\#\left\{p_{\text{obs}} > \lambda\right\}}{\#\left\{p_{\text{sim}} > \lambda\right\}} \qquad (4)$$

with $\lambda = 0.5$. We can also use the more conservative estimate

$$\widehat{\text{eFDR}}(t) \leq \frac{\#\left\{p_{\text{sim}} \leq t\right\}/M_0}{\#\left\{p_{\text{obs}} \leq t\right\}/m}$$

In order to ensure increasing FDR for increasing p-values, we define q-values as

$$\hat{q}(t) = \min_{p \geq t} \widehat{\text{eFDR}}(p) \qquad (5)$$

ETOPOSIDE PHARMACOGENOMICS

A triangle method, similar to the one described here, had been originally applied to cellular sensitivity data for etoposide (Huang et al., 2007b), one of the most widely used anti-cancer agents. Using our empirical FDR approach, we re-analyzed the same phenotype data from the original experiments, which had sought to quantify the cytotoxic effect of the drug on the cell lines using a colorimetric-based assay, as previously described (Huang et al., 2007b). We conducted our study on the 90 HapMap cell lines of European descent (CEU). The quantitative trait used here was IC_{50}, defined as the concentration required for 50% cell growth inhibition.

RESULTS

R FUNCTION FOR CALCULATING EMPIRICAL FDR

We provide an R function for estimating the empirical FDR that can be used once the observed and the simulated p-values are generated (http://www.scandb.org/newinterface/empiricalFDR.R). The way these p-values are generated will depend on the specific integration method used, the eQTL mapping database, and the number of components in the "genetic machinery."

Computation time

For step 1 (see **Figure 4**) we need to compute about 10,000 (the number of transcripts) linear regressions. This can be achieved in a few seconds using R and the fast linear regression computation in R such as implemented by us and made available in http://www.scandb.org/newinterface/fastlm.R. For step 2, we only need to query the eQTLs for the new set of genes from step 1, which takes a fraction of a second. For step 3: after applying steps 1 and 2 only a few SNPs are left (typically around 1000 or less). This can also be done in a fraction of a second. Adding up all three steps, the method with 1000 permutations would take a couple of hours of computing time on a typical desktop available in 2012.

TRADITIONAL GWAS AND SNP SELECTION VIA eQTLs

The GWAS of etoposide IC_{50} did not yield any significant signals, as perhaps expected from the small sample size. **Figure 5** shows a QQ plot of the distribution of p-values (as circles). However, we found a highly significant enrichment for gene regulatory signals among the etoposide-associated variants relative to frequency-matched SNPs (Gamazon et al., 2010a). This finding raises the possibility of the use of eQTL annotation to increase the power to detect true associations. We therefore proceeded to incorporate eQTL functional annotation through the integrative triangle method.

GENETIC VARIATION ASSOCIATED WITH ETOPOSIDE CYTOTOXICITY IDENTIFIED THROUGH THE TRIANGLE METHOD

Expression levels had been generated by our group on these cell lines for more than 10,000 genes, allowing us to perform associations between etoposide IC_{50} and gene expression traits; those genes meeting $p < 0.05$ (see Table S1 in Supplementary Material) were carried forward in the triangular analysis. We then utilized SCAN (Gamazon et al., 2010b), a public repository for the results of our eQTL studies on the HapMap cell lines, to annotate the selected genes showing association with etoposide IC_{50} with expression-associated SNPs ($p < 10^{-4}$; see Table S2 in Supplementary Material). Finally, the selected eQTLs were tested for association with etoposide IC_{50}. **Figure 5** shows the QQ plot of association *p*-values for all SNPs, the QQ plot for the final SNP set derived from the triangle method, and the QQ plot for the triangle method-prioritized SNPs from each of 1000 simulated phenotypes. The figure illustrates that certain eQTL SNPs from this triangle method-derived SNP set attained a (traditional) FDR < 0.05, but also that the triangle method may yield spurious associations using the traditional FDR.

EMPIRICAL FDR IDENTIFIES SIGNIFICANT ASSOCIATIONS WITH CELLULAR SENSITIVITY TO ETOPOSIDE

We applied our proposed empirical FDR method to the observed set of *p*-values from the triangle analysis-derived set of SNPs. To generate an empirical null distribution of *p*-values, we conducted simulations (see Materials and Methods). Table S3 in Supplementary Material lists the most significant etoposide-associated SNPs based on our empirical FDR method. Note the comparison between traditional FDR and eFDR for the most highly ranked SNPs prioritized by the triangle method (based on unadjusted *p*-value), showing that traditional FDR inflates the significance of selected SNPs.

DISCUSSION

Integrative approaches to diverse genomics datasets promise to resolve some important biological problems and, perhaps as importantly, generate novel hypotheses. Here we developed a *computationally feasible* permutation method to *quantify the significance* of findings arising from an integrative approach. The triangle method, a highly plausible approach to SNP prioritization and an example of how diverse high-throughput datasets may be integrated, requires an assessment of the resulting findings. This integrative method incorporates genotypic and expression data to identify trait-correlated genes that are under the regulation of eQTLs, yielding a set of candidate SNPs potentially important for the genetic etiology of the trait. Our proposed empirical FDR approach not only takes into account the integrative nature of the triangle method, but the approach also accounts for the correlation structure among gene expression traits and among genotypes. Our empirical FDR approach aims to provide a sound quantification of the significance of the prioritized SNPs from the integrative method.

It should be noted that our approach separates the phenotype from what we are calling the "genetic machinery" (e.g., genotype, gene expression, protein expression, methylation). Only the phenotype is permuted and the relationships within the genetic machinery are preserved. Consequently, we avoid having to perform multiple eQTL mappings (the most computationally costly permutation) because *p*-values in each arm are used for prioritization and not for determining the significance of the associations. Importantly, our approach differs from other approaches wherein the permutation is conducted on each arm of the triangle. In the latter approach, the threshold for significance can be arbitrary or unnecessarily conservative. A well-chosen set of thresholds will determine the performance of the integrative approach. In our method, we provide a measure of significance that is well-calibrated regardless of the set of thresholds used. Furthermore, in contrast to approaches that apply a threshold (e.g., Bonferroni) at each step of the integrative process, our method provides an overall measure of significance for the results of the integrative analysis.

Our quantification approach can easily accommodate hub eQTL analysis (SNPs associated with multiple genes, also referred to as master regulators). In the filtering procedure we require that the SNPs be eQTLs for a number of phenotype-associated gene expression traits. As long as the permutation steps follow the same filtering algorithm as the one used for the observed data, our method will yield the right FDR. Likewise, our method can be applied to both quantitative and binary outcomes.

In this study, we also explored the limitations of the traditional FDR when applied to an integrative approach such as the triangle method. In particular, we found that traditional FDR may yield spurious associations from simulated phenotypes. Furthermore, while the use of eQTL information may improve power to detect true associations, traditional FDR may still inflate the significance of the selected SNPs.

We applied our empirical FDR approach to a study of cellular sensitivity to etoposide. Etoposide is a topoisomerase II inhibitor (Sinha et al., 1988) widely used against lung cancer, non-Hodgkin's lymphoma, myelogenous leukemia, and Kaposi's sarcoma. As in the case of other chemotherapeutic agents, the drug is associated with serious toxicities, including bone marrow suppression, diarrhea, and fatigue as well as treatment-induced acute myeloid leukemia (Mistry et al., 2005). Thus, the identification of predictors of response or potentially debilitating toxicities associated with etoposide, including genetic variations, is key to the implementation of an effective treatment regimen and, longer-term, to the realization of an individualized approach to therapy. Based on cell lines derived from large pedigrees, it has been reported that a significant genetic component contributes to cellular sensitivity to etoposide (Peters et al., 2011).

Here, using our empirical FDR method, we identified 12 SNPs showing significant association (eFDR < 0.15) with cellular sensitivity to etoposide through their effect on gene expression. The 12 SNPs represent four independent genomic loci (on chromosome 8q12, 2p24, 10q23, and 16q24), of which the 10q23 SNPs are located in the glutamate receptor ionotropic delta-1 subunit (*GRID1*) gene. The expression target genes of rs9808546 (on chromosome 2) show a highly significant enrichment for *acetylation* [$n = 27$, Benjamini–Hochberg (BH) FDR = 0.0018] and *phosphoprotein* ($n = 51$, BH FDR = 0.0027; Huang da et al., 2009), consistent with studies that have shown that histone deacetylase inhibitors sensitize cells to the cytotoxic effects (particularly) of topoisomerase II agents such as etoposide (Kurz et al., 2001; Marchion et al., 2004; Hajji et al., 2008, 2010). Importantly, having provided a sound quantification of the significance of the genotype-phenotype associations, the gene expression targets of the identified eQTLs provide a set of candidate genes for functional validation and a plausible mechanism for how the genetic variation may mediate their phenotypic effect.

The R code we provide can be used to compute the empirical FDR for any case in which empirical null *p*-values are available regardless of the method used to generate them. Thus, it should prove useful for other integrative approaches.

In case there are confounding factors that yield more gene expression traits associated with the phenotype, our method yields a conservative estimate of FDR. The effect of the confounders is to increase the number of noisy genes in the first step and consequently to generate more null eQTLs than there should be in the final set. This fact decreases the overall significance of real associations and our method still provides an unbiased estimate of the significance.

In summary, we have developed a computationally feasible approach to assess the significance of genotype-phenotype associations prioritized by an integrative genomic method. As omics datasets become routinely integrated to address important biological problems, the issue our study sought to address becomes increasingly more relevant.

ACKNOWLEDGMENTS

This work was funded in part by PAAR (Pharmacogenomics of Anti-cancer Agents Research; U01 GM61393), by the University of Chicago Cancer Center Support grant P30 CA014599-36 from the National Cancer Institute, and by the Genotype-Tissue Expression project (GTeX; R01 MH090937). R. Stephanie Huang received support from NIH/NIGMS grant K08GM089941, University of Chicago Cancer Center Support Grant (#P30 CA14599), and Breast Cancer SPORE Career Development Award.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Statistical_Genetics_and_Methodology/10.3389/fgene.2012.00202/abstract

Table S1 | Top gene expression-trait correlations.

Table S2 | Top SNPs from etoposide GWAS.

Table S3 | Top associations from the eFDR approach.

REFERENCES

Cheung, V. G., Conlin, L. K., Weber, T. M., Arcaro, M., Jen, K. Y., Morley, M., et al. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33, 422–425.

Correa, C. R., and Cheung, V. G. (2004). Genetic variation in radiation-induced expression phenotypes. *Am. J. Hum. Genet.* 75, 885–890.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394.

Dermitzakis, E. T. (2012). Cellular genomics for complex traits. *Nat. Rev. Genet.* 13, 215–220.

Duan, S., Huang, R. S., Zhang, W., Bleibel, W. K., Roe, C. A., and Clark, T. A. (2008). Genetic architecture of transcript-level variation in humans. *Am. J. Hum. Genet.* 82, 1101–1113.

Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* 10, 605–616.

Gamazon, E. R., Huang, R. S., Cox, N. J., and Dolan, M. E. (2010a). Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9287–9292.

Gamazon, E. R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E. O., Nicolae, D. L., et al. (2010b). SCAN: SNP and copy number annotation. *Bioinformatics* 26, 259–262.

Gamazon, E. R., Ziliak, D., Im, H. K., LaCroix, B., Park, D. S., Cox, N. J., and Huang, R. S. (2012). Genetic architecture of microRNA expression: implications for the transcriptome and complex traits. *Am. J. Hum. Genet.* 90, 1046–1063.

Garge, N., Pan, H., Rowland, M. D., Cargile, B. J., Zhang, X., Cooley, P. C., et al. (2010). Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. *Mol. Cell Proteomics* 9, 1383–1399.

Hajji, N., Wallenborg, K., Vlachos, P., Füllgrabe, J., Hermanson, O., and Joseph, B. (2010). Opposing effects of hMOF and SIRT1 on H4K16 acetylation and the sensitivity to the topoisomerase II inhibitor etoposide. *Oncogene* 29, 2192–2204.

Hajji, N., Wallenborg, K., Vlachos, P., Nyman, U., Hermanson, O., and Joseph, B. (2008). Combinatorial action of the HDAC inhibitor trichostatin A and etoposide induces caspase-mediated AIF-dependent apoptotic cell death in non-small cell lung carcinoma cells. *Oncogene* 27, 3134–3144.

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* 11, 476–486.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic, and functional implications of genome-wide association loci for human diseases, and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367.

Huang, R. S., Duan, S., Shukla, S. J., Kistner, E. O., Clark, T. A., Chen, T. X., et al. (2007a). Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am. J. Hum. Genet.* 81, 427–437.

Huang, R. S., Duan, S., Bleibel, W. K., Kistner, E. O., Zhang, W., Clark, T. A., et al. (2007b). A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc. Natl. Acad. Sci. U.S.A.* 104, 9758–9763.

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57.

Ko, D. C., Shukla, K. P., Fong, C., Wasnick, M., Brittnacher, M. J., Wurfel, M. M., et al. (2009). A genome-wide in vitro bacterial-infection screen reveals human variation in the host response associated with inflammatory disease. *Am. J. Hum. Genet.* 85, 214–227.

Kurz, E. U., Wilson, S. E., Leader, K. B., Sampey, B. P., Allan, W. P., Yalowich, J. C., et al. (2001). The histone deacetylase inhibitor sodium butyrate induces DNA topoisomerase II alpha expression and confers hypersensitivity to etoposide in human leukemic cell lines. *Mol. Cancer Ther.* 1, 121–131.

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* 11, 191–203.

Marchion, D. C., Bicaku, E., Daud, A. I., Richon, V., Sullivan, D. M., and Munster, P. N. (2004). Sequence-specific potentiation of topoisomerase II inhibitors by the histone deacetylase inhibitor suberoylanilide hydroxamic acid. *J. Cell. Biochem.* 92, 223–237.

Metzker, M. L. (2010). Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11, 31–46.

Mistry, A. R., Felix, C. A., Whitmarsh, R. J., Mason, A., Reiter, A., Cassinat, B., et al. (2005). DNA topoisomerase II in therapy-related acute promyelocytic leukemia. *N. Engl. J. Med.* 352, 1529–1538.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., et al. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 464, 773–777.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., et al. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6:e1000895. doi:10.1371/journal.pgen.1000895

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6:e1000888. doi:10.1371/journal.pgen.1000888

Peters, E. J., Motsinger-Reif, A., Havener, T. M., Everitt, L., Hardison, N. E., Watson, V. G., et al. (2011). Pharmacogenomic characterization of US FDA-approved cytotoxic drugs. *Pharmacogenomics* 12, 1407–1415.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464, 768–772.

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., et al. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6:e107. doi:10.1371/journal.pbio.0060107

Sinha, B. K., Haim, N., Dusre, L., Kerrigan, D., and Pommier, Y. (1988). DNA strand breaks produced by etoposide (VP-16,213) in sensitive and resistant human breast tumor cells: implications for the mechanism of action. *Cancer Res.* 48, 5096–5100.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* 31, 2013–2035.

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., et al. (2007a). Relative impact of nucleotide, and copy number variation on gene expression phenotypes. *Science* 315, 848–853.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., et al. (2007b). Population genomics of human gene expression. *Nat. Genet.* 39, 1217–1224.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., et al. (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* 8:e1002639. doi:10.1371/journal.pgen.1002639

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63.

Welsh, M., Mangravite, L., Medina, M. W., Tantisira, K., Zhang, W., Huang, R. S., et al. (2009). Pharmacogenomic discovery using cell-based models. *Pharmacol. Rev.* 61, 413–429.

Zhong, H., Beaulaurier, J., Lum, P. Y., Molony, C., Yang, X., Macneil, D. J., et al. (2010). Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.* 6:e1000932. doi:10.1371/journal.pgen.1000932

Zhou, V. W., Goren, A., and Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* 12, 7–18.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 May 2012; accepted: 20 September 2012; published online: 31 May 2013.

*Citation: Gamazon ER, Huang RS, Dolan ME, Cox NJ and Im HK (2013) Integrative genomics: quantifying significance of phenotype-genotype relationships from multiple sources of high-throughput data. Front. Genet. **3**:202. doi:10.3389/fgene.2012.00202*

This article was submitted to Frontiers in Statistical Genetics and Methodology, a specialty of Frontiers in Genetics.