Metadata of the chapter that will be visualized online

Chapter Title	Gene-Gene and Gene-Environment Interactions AU1	
Copyright Year	2018	
Copyright Holder	Springer Science+Business Media, LLC, part of Springer Nature	
Corresponding Author	Family Name	**DeWan**
	Particle	
	Given Name	**Andrew T.**
	Suffix	
	Division	Department of Chronic Disease Epidemiology
	Organization	Yale School of Public Health
	Address	New Haven, CT, USA
	Email	andrew.dewan@yale.edu
Abstract	Identifying gene–gene and gene–environment interactions may help us to better describe the genetic architecture for complex traits. While advances have been made in identifying genetic variants associated with complex traits through more dense panels of genetic variants and larger sample sizes, genome-wide interaction analyses are still limited in power to detect interactions with small effect sizes, rare frequencies, and higher order interactions. This chapter outlines methods for detecting both gene-gene and gene-environment interactions both through explicit tests for interactions (i.e., ones in which the interaction is tested directly) and non-explicit tests (i.e., ones in which an interaction is allowed for in the test, but does not test for the interaction directly) as well as approaches for increasing power by reducing the search space. Issues relating to multiple test correction, replication, and the reporting of interaction result in publications.	
Keywords (separated by '-')	Interaction - Epistasis - Environment - GWAS - Power - Replication	

Chapter 7

Gene-Gene and Gene-Environment Interactions

Andrew T. DeWan

Abstract

Identifying gene–gene and gene–environment interactions may help us to better describe the genetic architecture for complex traits. While advances have been made in identifying genetic variants associated with complex traits through more dense panels of genetic variants and larger sample sizes, genome-wide interaction analyses are still limited in power to detect interactions with small effect sizes, rare frequencies, and higher order interactions. This chapter outlines methods for detecting both gene-gene and gene-environment interactions both through explicit tests for interactions (i.e., ones in which the interaction is tested directly) and non-explicit tests (i.e., ones in which an interaction is allowed for in the test, but does not test for the interaction directly) as well as approaches for increasing power by reducing the search space. Issues relating to multiple test correction, replication, and the reporting of interaction result in publications.

Key words Interaction, Epistasis, Environment, GWAS, Power, Replication

1 Introduction

An interaction results when the effect of one factor is only evident in the presence of another. These factors could be genetic markers and/or environmental exposures. Much has been written on the topic of gene-gene (also known as epistasis) and gene-environment interactions with several comprehensive reviews of study designs and methods for analyzing both gene-gene [1–3] and gene-environment interactions [4, 5]. But why are we interested in studying interactions? Both gene-gene and gene-environment and potentially gene-gene-environment interactions allow us to better describe the underlying genetic architecture of a particular trait and as such we can begin to fill in the missing heritability [6] for a particular phenotype.

Biological interactions were originally defined as the situation when the phenotypic effect of one gene was only evident in the presence of a second gene [7]. In contrast, a statistical interaction is defined as a departure from a linear model combining two or more genetic factors (or a genetic factor and environmental factor) [8].

Evangelos Evangelou (ed.), *Genetic Epidemiology: Methods and Protocols*, Methods in Molecular Biology, vol. 1793, https://doi.org/10.1007/978-1-4939-7868-7_7, © Springer Science+Business Media, LLC, part of Springer Nature 2018

Another way to think about this is that biological interactions are observed at the individual level and statistical interactions at the population level, but this does not imply that observing evidence of one will lead to observing evidence of another [9]. The traditional method to test for statistical interactions is to use a regression-based model that includes main effect variables representing each genetic factor and an interaction term (the product of the main effect variables), and then testing for the significance of the interaction term after adjusting for the main effects. However, several other methods exist to test for statistical interactions and these will be discussed as well.

We now have the ability to conduct very dense genome-wide association studies with up to five million genetic markers genotyped at one time and millions more imputed using large reference panels from population-based sequencing projects. The combinatorial problem can be immense if you attempt to look at all possible marker combinations with little power to detect significant interactions after accounting for all interactions tested. Given the exponential improvements in computational power and the relative ease of parallel computing, the computational hurdles of examining all pairwise gene-gene interactions are not insurmountable, and exhaustive searches of higher order interactions will follow. However, the immense power constraints on currently available sample sizes, while improving with studies routinely examining 10,000 to more than 100,000 subjects, are still underpowered to detect interactions with modest effect sizes ($OR < 1.2$) and low frequency variants (minor allele frequency [MAF] < 0.01). Though marker pruning using linkage disequilibrium can reduce this problem somewhat, there are several other data reduction approaches that will be discussed based on prior biological knowledge or statistical evidence for association. These can mitigate the multiple testing burden but depend heavily on the quality of this prior information.

Similar to gene-gene interactions, gene-environment interactions arise when the effect of a genetic factor on a phenotype is dependent on the presence or absence of an environmental factor. Statistically, this can be tested in a similar fashion as for gene-gene interactions traditionally done using a regression-based model with main effects for the genetic factor and environmental factor and an interaction term and testing for the significance of the interaction term. This environmental factor may be one traditionally thought of as an environmental exposure such as smoking, indoor NO_2 levels, or sun exposure, but these could also be other potentially genetically influenced phenotypes such as obesity, blood glucose levels, or birthweight that may be influencing the effect of the genetic factor depending on the value of these secondary phenotypes.

As outlined below, direct assessment of interactions (termed as "explicit" test) is often less powerful than joint tests that include either an interaction term or allow for interaction but do not test for the significance of the interaction directly (termed as "non-explicit" test). This classification of tests for interactions will be used throughout this chapter as different approaches and tests are discussed.

2 Materials

2.1 Data

2.1.1 Genotype Data

The core of both gene-gene and gene-environment interaction analyses is obtaining high-quality genotype data. One may use data generated specifically for a particular project, but there are also many outstanding datasets available for analysis from a number of online repositories.

If samples are being genotyped for the specific project there are many options available. There are whole genome microarray panels of markers based on relatively even coverage across the genome, panels of markers that maximize the coverage for specific race/ ethnic groups, and panels that allow custom markers to be added-on to existing panels to increase coverage of specific genes of interest or previously associated markers, for example. There are also panels of markers that target specific regions of the genome such as the exonic regions, cancer-associated genes, and metabolic genes. The selection of the right microarray will depend on your specific study hypotheses, type of study, and budget. Lower throughput genotyping can be done for single variants to hundreds of markers simultaneously depending again on the hypotheses and goals of the project.

Genotyping known panels of markers is not the only choice for interaction studies. One could choose to utilize sequencing-based approaches to genotype unknown variants and/or low frequency and rare variants. Whole genome or whole exome sequencing could be utilized for hypothesis-free analyses or targeted sequencing if specific genes or pathways are hypothesized to be involved in the interaction(s).

If secondary data analysis is an option, there are a plethora of datasets with genome-wide data available as well as extensive phenotype data. Two such repositories are the database of genotype and phenotype (dbGaP) maintained by NCBI primarily of studies conducted in the United States (https://www.ncbi.nlm.nih.gov/gap) and the European Genome-phenome Archive (EGA) maintained by EBI primarily of studies conducted in Europe (https://www.ebi.ac.uk/ega/home). These databases contain hundreds of datasets accessible through an application to the respective data access committees.

2.1.2 Environmental Data

The environmental factors that could be considered in a study of gene-environment interactions are extremely broad. These include chemical (e.g., polychlorinated biphenyl (PCB)), physical (e.g., airborne particulate matter), biological (e.g., viral infections), and lifestyle (e.g., physical activity). The measurement of each of these individual environmental factors is going to vary widely depending on the environmental factor of interest. This could range from measure NO_2 levels in the air via chemiluminescence, measuring radiation exposure via a dosimeter, conducting a daily food diary to estimate saturated fat intake, or reviewing charts to collect data on BMI history. The discussion in this chapter will focus on environmental exposures at one time point, but there is some evidence that longitudinal environmental data may increase the power to detect gene-environment interactions for common diseases [10].

2.1.3 Biological/ Functional Data

For analyses that are pursuing a hypothesis-driven approach and/or filtering based on biological information, one may want to utilize prior biological or functional data. There are a variety of databases and programs that can be accessed to provide this type of information. One could simply curate information from publications in the scientific literature through systemic reviews of publications in databases such as PubMed. Data on functionality of variants can be obtained more systematically from databases such as ENCODE (Encyclopedia of DNA Elements) [11] that contains a comprehensive list of functional elements at both the RNA and protein levels and is available for viewing or downloading from the UCSC Genome Browser (www.genome.ucsc.edu/ENCODE). Direct annotation of variants could be conducted using a program such as Annovar [12] to annotate variants as to their respective genes, coding vs. noncoding, and predicted functional consequence. An alternative annotation program more directly related to filtering variants for interaction analyses is Biofilter which allows for the annotation of variants based on previous association studies and biological knowledge, filtering variants based on specific biological hypotheses, and building sets of testable variant interactions based on implication indices compiled from available data [13].

2.1.4 Previous Statistical Data

For analyses filtering variants based on prior statistical knowledge, data from one's own GWAS or single-variant association study could be used, results mined from previous publications or, alternatively, association results obtained from databases such as the GWAS Catalog (http://www.ebi.ac.uk/gwas/). While results from previous publications or the GWAS Catalog are a convenient and useful resource, they have the disadvantage of being biased toward reporting only genome-wide significant results and other nominally significant results will likely not be available and should be kept in mind when planning the analysis approach.

2.2 Software

Below are a listing of programs that can be used to conduct gene-gene and/or gene-environment interaction analyses with other programs mentioned and described throughout the chapter. This is not an exhaustive list and is only an example of programs that are commonly used to assess interactions in genetic epidemiological studies. Additional programs not directly related to the interaction analysis such as for computing eigenvalues from principal components analyses (e.g., EIGENSTRAT [14]) and imputation of variants (e.g., IMPUTE2 [15]) are not listed.

2.2.1 PLINK

This is a suite of tools designed to conduct genome association analyses, including both gene-gene and gene-environment interactions [16]. The primary interaction analyses are based on logistic and linear regression. They can accommodate both gene-gene and gene-environment interactions on the genome-wide scale or on a smaller number of variants by creating subsets of genetic variants to test again each other or the environmental factor. The program has the flexibility to conduct both explicit tests for interaction by testing for the significance of the interaction term directly in the regression model or a non-explicit joint test by testing the main and interactions effects. There is also a faster option for conducting genome-wide gene-gene interactions (*fast-epistasis*) based on the Z-score for the differences in OR for SNP-SNP combinations between cases and controls or for cases alone (case-only test).

2.2.2 CASSI

This is a software package that is specifically designed to conduct genome-wide gene-gene interaction analyses in a computationally efficient manner ([17]; https://www.staff.ncl.ac.uk/richard.howey/cassi/index.html). This package corrects a minor error in the Wu et al. statistic [18] in the calculation of the variance for estimated rather than observed haplotypes and in the *fast-epistasis* variance originally implemented in PLINK.

2.2.3 BOOST

The Boolean Operation-based Screening and Testing (BOOST) program was designed to efficiently screen and then explicitly test for genome-wide gene-gene interactions [19]. The screening phase involves a non-iterative procedure to approximate the likelihood ratio and then all variant pairs that survive this screening are subjected to a classical likelihood ratio test in the testing phase.

2.2.4 MDR

The Multifactor Dimensionality Reduction (MDR) software package [20] is designed to conduct data mining on discrete variables and can be used to detect both gene-gene and gene-environment interactions and dichotomous outcomes [21]. The traditional MDR approach is a non-explicit test for interaction as it is a non-parametric test that combines factors that may be interacting in order to best discriminate the subjects among the dichotomous outcome. An extension of the MDR has been developed that incorporates a permutation-based approach that can explicitly test for interactions [22]. A recent extension to the MDR has implemented a t-test approach that allows for quantitative outcomes [23]. The MDR method, however, is designed primarily for smaller sets of markers, but parallel computing could be utilized to conduct a genome-wide analysis.

3 Methods

3.1 Quality Control (QC)

No analysis can be successful without high-quality data. The specific steps of the genotype QA/QC will depend on the type of assay used to generate the genotypes. These range from single-variant assays based on PCR, whole genome microarray genotyping, and whole-exome and whole-genome sequencing. A brief outline of the QC steps for each is outlined below.

3.1.1 Single-Variant QC

The primary steps are to assess the overall performance of the individual genotyping assays through examination of the variant call rate (variant call rate = total number of genotype calls/total number of individuals genotyped) and Hardy-Weinberg Equilibrium (HWE). While the thresholds chosen to eliminate variants can be arbitrary, typically one would look for variant call rates >98% (which should be examined in cases and controls separately if conducting a case-control analysis to ensure no bias due to differences in call rates between cases and controls) and HWE p-values $>10^{-4}$ which if conducting a case-control study are assessed only in controls. Without the benefit of genome-wide genotype data it is impossible to assess the data for population stratification, but adjustments can be made in the analysis (if using regression-based methods) for relevant covariates that may capture potential stratification such as self-reported race/ethnicity.

3.1.2 Microarray QC

As with single-variant QC, one will examine both the individual variant call rates and HWE to ensure that each variant probe is generating high-quality genotype data with similar thresholds applied as mentioned above. However, additional steps can and should be taken into account. The individual subject call rates should be examined first to determine if there were general problems with the individual array and/or DNA. These call rate thresholds may range from 93% to 98% and are often suggested by the array manufacturer-based past performance of the array (subject call rate = total number of genotype calls for an individual subject/total number of variants attempted to be genotyped). Poorly performing subjects should be removed prior to any downstream QC steps. Population stratification should be assessed using a genome-wide procedure such as principal components analysis to determine if there are slight variations in the genotype frequencies between subpopulations within your dataset. This procedure can detect any systematic differences that may be due to differences in allele frequencies arising ancestry differences, but also due to experimental/processing differences (e.g., plate effects). If significant principal components (PCs) are detected it is suggested that these PCs be adjusted for in the analysis. If the analysis assumes unrelated subjects it is suggested that the dataset be examined for cryptic relatedness using a procedure such as estimate pairwise identity-by-decent (IBD). This pairwise measure that is often used is $\widehat{p_i}$. Again, the threshold for identifying cryptically related subjects is arbitrary one often chooses a threshold ranging from 0.125 to 0.2 and then eliminates one of the two subjects in this cryptically related pair. This can be done randomly, or one may want to eliminate the subjects based on the subject call rate (eliminating the subject with a fewer genotype calls) or if it is a case-control study and a cryptically related pair is comprised of a case subject and a control subject it may be beneficial to eliminate the control if cases are in short supply.

To assess whether or not the QC steps that have been taken are successful prior to conducting an interaction analysis, it would be beneficial to conduct a genome-wide single-variant analysis and examine QQ plots and/or estimate λ from the data after adjusting for PCs and other covariates. QQ plots can be generated using an R script such as qqman.r (https://CRAN.R-project.org/package=qqman) and λ estimated using PLINK. Deviations from the expected line on the QQ plot are not expected except at the tail (i.e., the true positives) and with deviations along much of the expected line being an indication of residual population stratification. λ estimates greater than 1.05 are routinely seen as indicators of population stratification and additional PCs should be adjusted for until the lambda value falls below this threshold. This assessment is typically done in single-variant analyses prior to any interaction analyses.

3.1.3 Sequencing QC

Specific workflows for alignment, variant calling, and variant QA/QC and filtering are described in detail elsewhere (*see* Ref. 24 for a detailed step-by-step pipeline covering the major sequencing analysis tools). Briefly, a standard analysis pipeline would start by aligning the FASTQ raw sequence reads to a reference genome using the Burrows-Wheeler Aligner (BWA, [25]).Then converted to BAM format, sorted, indexed, PCR duplicates marked and then merged into one BAM file using SAMtools [26]. Finally, the alignments in the BAM file can be locally realigned around insertion/deletions, recalibrated and variants called using HaplotypeCaller in the Genome Analysis Toolkit (GATK, [27, 28]).

Variant QC can utilize a variety of different metrics, but an example of one approach is how we conducted our QC in our whole-exome sequence analysis of a family segregating asthma [29]. Variants were flagged (and not considered further in our analysis) if they met any of the following criteria: three or more variants detected within 10 bp; four or more alignments map to different locations equally well; coverage of less than five reads; quality score <50; low quality for a particular sequence depth (variant confidence/unfiltered depth <1.5); and strand bias (Phred-scaled *p*-values using Fisher's Exact Test >200). There are many variations that can be employed in your QC pipeline, but the most important aspect is achieving the highest quality set of variants to retain in your analysis.

3.1.4 Linkage Disequilibrium Pruning

Reducing the number of variants considered in the analysis, regardless of genotyping method, can be accomplished by linkage disequilibrium (LD) pruning. LD is an indication of the correlation, or non-independence, between variants and can be measured using r^2 or D'. A typical r^2 threshold used to prune variants is 0.8, but a lower threshold can be used to eliminate more variants at the risk of excluding some that are independently informative. This procedure is particularly important when conducting a case-only gene-gene interaction analysis (discussed in Subheading 3.2.12) as the variants must not be correlated with each other.

3.2 Gene-Gene Interaction Analysis

The basic analysis is straightforward, the assessment of a deviation from an additive or multiplicative model containing two or more variants. This is traditionally assessed through regression-based modeling but many different methods are available and several different methods will be discussed below to exemplify this approach. One could divide these approaches into explicit vs. -non-explicit tests for interactions, with explicit tests determining if the null hypothesis of the sum on the additive/linear scale or the product on the multiplicative scale of the joint effects of the two variants is contributing to the outcome with the alternative hypothesis being that the joint effect of the two variants is greater or less than the expected. Non-explicit tests are able to determine if the grouping of variants is associated with the outcome, but not necessarily that the variants together deviate from an additive or multiplicative model. Different inheritance models can be imposed on the genetic variants in the interaction models (*see* **Note 1**) and the issue of outcome scale is also an important consideration when interpreting interactions (*see* **Note 2**).

Regression-based approaches are attractive not only for their ability to explicitly detect interactions in the data, but because of the ability to adjust for multiple covariates. In a genome-wide setting this is important so that significant principal components can be adjusted to remove population stratification, but other potential confounders can be adjusted as well, including age, sex, etc. depending on the outcome of interest.

3.2.1 Logistic Regression

For dichotomous outcomes, a logistic regression model is one of the most straightforward approaches for testing an interaction, either explicitly or non-explicitly. In the following logistic regression model we are modeling the probability of our dichotomous outcome (p) using an intercept (β_0), two genetic variants (SNP1 and SNP2, for example) and their respective main effects (β_1 and β_2) and their interaction effect (β_3):

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathrm{SNP1} + \beta_2 \mathrm{SNP2} + \beta_3 \mathrm{SNP1} \times \mathrm{SNP2}$$

In this model, we can test for the significance of the interaction effect of the two variants by testing the null hypothesis that $\beta_3 = 0$. This is an explicit test for interaction in that we are testing for an interaction after adjusting for the main effects (i.e., independent effects) of the two genetic variants. This model will produce an effect estimate (β_3) which can easily be converted to an OR by $\exp\beta_3$. It should be noted that the main effects should not be interpreted as the main effect of the variants (or the variant and the environmental factor) since they are adjusted for the interaction term in the model.

Alternatively, we can test for the significance of the joint effects of the two variants by using a 2 degree of freedom likelihood ratio test of the full model against a model in which there is neither an interaction term nor main effect term for one of the SNPs:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \ \beta_1 \mathrm{SNP1}$$

While this model contains an interaction term this is a non-explicit test for interaction as we are testing the significance of the main effect of one of the SNPs and interaction term combined. This may be seen as a less desirable test but this can be more powerful for detecting significant signals [30]. However, this joint test can be significant if either of the main effects is strong enough or if both main effects are strong enough without a significant interaction term (β_3). As will be discussed later, one may want to screen combinations of variants for joint effects and then follow up for explicit tests of an interaction only on the limited set of variants demonstrating significant joint effects.

When reporting the results of regression-based analyses, it is important to report not only the significance of the interaction (p-value), but also the parameter estimate of the interaction as well as those of the main effects (*see* **Note 3**).

3.2.2 Linear Regression

For continuous outcomes, linear regression can be used to test for interactions in much the same way as logistic regression for dichotomous outcomes. βs are not converted to ORs, but are rather interpreted directly with the sign of β_3, an indication of the direction of the interaction, and the p-value for the likelihood ratio test of the full model versus one without the interaction term an indication of the significance of the interaction term. A two degree of freedom joint test can also be constructed.

3.2.3 Nonparametric Analyses

The most popular nonparametric method for interaction analysis are the multidimensional reduction methods. These are based on the idea that across a contingency table of two or more genetic variant genotype combinations or genotype and discrete environmental factor combinations, each square in the contingency table can be divided into increasing risk and decreasing risk in the context of the phenotype. These collapsed contingency tables can then be tested for association with the phenotype. This is typically done by splitting the data into a training set (to build the collapsed contingency table) and a testing set to determine the classification error. The data are split multiple times and the model is assessed by the classification error and prediction error estimated from these multiple cross-validations [21]. Due to the fact that this is a nonparametric method, a p-value is obtained for the set of factors included in the model, but an effect estimate is not. The significance of the model is interpreted as the significance of the group of factors (genetic variants or genetic variant(s) and environmental factors) but cannot be interpreted as direct evidence of an interaction.

The model-based MDR (MB-MDR) methodology is an extension of the classic MDR framework that allows for the direct testing of interactions through the use of the Wald statistic on high and low-risk genotype categories [22, 31]. The significance of the interaction is then tested using permutation testing of the maximum Wald statistic and can be considered an explicit test of an interaction.

3.2.4 Multiple Test Correction

One of the biggest challenges in conducting genome-wide analysis in general, and more specifically interaction analysis, particularly gene-gene interactions, is the multiple testing problem. Given that hundreds of thousands or millions of variants are considered in most genome-wide studies, the chance of detecting a false positive is immense. This is even more apparent when multiple environmental factors are considered and most problematic when all pairwise variant interactions are considered in an exhaustive gene-gene interaction search. Traditionally, genome-wide studies have adopted a Bonferroni correction to adjust for the number of variants analyzed, but this can be overly conservative, especially for dense maps of markers in high LD. However, other procedures such as the False Discovery Rate (FDR; [32]), LD-based variant counting to account only for the number of independent tests [33] and permutation testing [34, 35] have been adopted. Permutation testing, while attractive because it accounts for the number of independent tests by maintaining the LD structure in the permuted data, is computationally intensive, even more so when all pairwise variant combinations need to be considered in thousands of permuted datasets. Therefore, the FDR or LD-based methods are more attractive alternatives.

3.2.5 Genome-Wide (Exhaustive) Approaches

A typical genome-wide association study will genotype between 100,000 and 1,000,000 singe-nucleotide polymorphisms. While exponential advances in computational power mean that the computing power to run the analysis on the pairwise tests required for a panel of 500,000 SNPs (2.5×10^{11} tests) is not insurmountable and with parallel computing relatively quick, the major hurdle is power. While large multi-center consortium studies mean that the number of subjects for many subjects has greatly increased, with studies reaching 50,000 or more subjects depending on the frequency of the disease or phenotype, these may still not be powerful enough to detect significant interactions on the genome-wide scale. In a study of genome-wide data obtained from the Resource for Genetic Epidemiology Research on Adult Health and Aging (GERA), using 45,171 subjects for ten phenotypes and conducting an exhaustive search for interactions, we failed to identify any genome-wide significant interactions, suggesting that we were underpowered to detect interactions with apparently weak effect sizes [36].

As demonstrated in Table 1, for a genome-wide gene-gene interaction study of 200,000 markers using an exhaustive approach, requires more than 30,000 cases and 30,000 controls to achieve genome-wide significance for two loci each with an MAF of 0.2 and an interaction OR of 1.2. All power calculations presented in this chapter were performed in Quanto [37]. The power calculations assumed a population risk of the disease of 0.1 and a log-additive mode of inheritance and a main effect of each variant of 1.2. For a case-control study with 5000 cases and 5000 controls, a reasonably sized case-control GWAS, and 200,000 markers genotyped, the minimal detectable OR is just over 1.6 (Table 2). For an interaction OR of 1.5, the number of markers considered would need to be reduced to around 2000 to detect a significant interaction among all pairs of interactions (4,000,000) and at an interaction OR of 1.2, only one interaction can be considered as only a nominally associated interaction is detectable (Table 3).

This lack of power can be daunting, but as detailed in the following sections there are several ways to increase power by reducing the number of tests through filtering or using biologically informed combinations of variants.

Table 1
Power of interaction analysis

	Number of cases and controls					
Interaction type	5000	10,000	20,000	30,000	40,000	50,000
G-G	0.000	0.003	0.164	0.654	0.940	0.995
G-E	0.024	0.032	0.317	0.717	0.928	0.988

Table 2
Power of case-control and case-only gene-gene interaction analyses

intOR	Case-control	Case-only
1.1	0.000	0.000
1.2	0.000	0.013
1.3	0.004	0.547
1.4	0.076	0.992
1.5	0.388	1.000
1.6	0.782	1.000
1.7	0.961	1.000
1.8	0.996	1.000
1.9	1.000	1.000
2.0	1.000	1.000

3.2.6 Data Reduction Approaches

In order to overcome the multiple testing burden of a genome-wide screen for all gene-gene interactions, it may be advantageous to focus on a subset of SNPs that may have a one or more properties that may make the interaction analysis more likely to detect a significant interaction. If we reduce our set of 200,000 SNPs down to 2000 we can now detect an interaction OR of 1.5 using a set of 5000 cases and 5000 controls (Table 3). Below I outline a series of methods to reduce the set of SNPs considered.

Filtering by Allele Frequency

The simplest method to increase power to detect interactions is to impose a filter by minor allele frequency (MAF). Depending on the sample size and subsequent power estimates, it may be advantageous to filter out all SNPs with MAFs less than the power to detect a reasonable interaction effect size (e.g., all SNPs with an MAF <0.2).

Table 3
Power of gene-gene interaction analysis by interaction OR and number of markers

# Markers	intOR = 1.2	intOR = 1.4	intOR = 1.5
200,000	0.000	0.076	0.388
100,000	0.000	0.107	0.465
20,000	0.000	0.221	0.649
10,000	0.001	0.289	0.724
2000	0.004	0.488	0.869
1000	0.008	0.584	0.914
200	0.037	0.793	0.975
100	0.066	0.864	0.988
20	0.219	0.966	0.999
10	0.337	0.986	1.000
5	0.488	0.995	1.000
2	0.713	0.999	1.000
1	0.865	1.000	1.000

Filtering by Marginal Effects

While a significant marginal and/or main effect is not required to detect a significant interaction, there are few reports of significant interactions without also having detectable main effects. Therefore, it may be advantageous to filter based on marginal effects (i.e., single-variant effect) and only include those variants that have, for example, a single-variant association *p*-value of <0.05 (nominal significance threshold). This should reduce the number of SNPs to approximately 5% of the starting number (for our 500,000 SNPs this would result in 6.25×10^{8} tests). This approach may be too stringent, as it requires both variants be nominally significant. An alternative approach would be to select a set of SNPs reaching a predefined significance threshold (e.g., $p < 10^{-4}$) and testing this set of SNPs against all other SNPs. In our example with 500,000 SNPs we would select ~50 SNPs with $p < 10^{-4}$ and test for an interaction with the remaining 499,450 for a total of 2.5×10^{-7} SNPs. This balances the requirement that there be some marginal effect of one of the SNPs with being able to detect interactions with SNPs showing no marginal effects but having a significance effect on disease only in the presence of a second SNP.

Candidate Gene Approaches

Filtering based on prior evidence of a gene's involvement in a particular disease is another approach to reducing the search space for interactions. Candidate genes could be selected by systematically reviewing the literature or for some diseases databases of candidate genes exist based on association studies, linkage analyses, and/or expression studies. For example, there are databases for preterm birth [38], preeclampsia [39], and non-syndromic hearing loss [40] for which one can obtain lists of candidate genes. By annotating variants to their respective genes using programs such as Annovar [12] only those SNPs annotated to the set of candidate genes can be selected for inclusion in the interaction analysis. As for the marginal effects, it may be advantageous to consider those SNPs within candidate genes against all other SNPs genotyped. In this way, it is possible to detect interactions with SNPs in novel genes not previously identified to be associated with the disease of interest.

Filtering by Function

Filtering by the effects of specific variants is another approach to reducing the search space. Again, variants can be annotated, but now the selection may be made based on being a coding variant or a splicing variant that is more likely to be functional. Other biological information could also be utilized such as examining interactions between SNPs in genes known to be involved in protein-protein interactions with the rationale being that SNPs in these biologically interacting genes are more likely to also show evidence of a statistical interaction. Similarly, one could examine interactions between variants within transcription factors and those within their binding sites.

3.2.7 Multistage Approaches

In order to maintain a genome-wide approach but overcome the hurdle of the immense multiple testing problem, it may be beneficial to employ a multistage approach. This is possible in situations in which a study has a large sample size, but still not adequately powered to detect significant interactions genome-wide. In a study of asthma, we first screened all pairwise interactions (9.1×10^{10}) in a small subset of the data and then carried through all interaction with a suggestive significance ($p < 10^{-5}$) to a follow-up stage of independent subjects and then attempted to replicate the top SNPs in a third set of independent subjects [41]. While this approach did not identify any genome-wide significant interactions, it did identify a candidate interaction between SNPs in two regions of the genome. The major advantage of this type of approach is that it allows for an unbiased examination of the SNPs without relying on previously reported biological and/or association data.

3.2.8 Gene-Based Interaction Tests

Another approach to reduce the multiple testing burden and combine information across multiple variants is to conduct a gene-based test of interaction. By considering each gene as a unit in the interaction rather than each individual variant, the total number of interactions considered is significantly reduced, thus increasing power. One such approach combines interaction p-values across all combinations of genetic variants in two genes into a single gene-gene interaction *p*-value that also accounts for linkage disequilibrium [42].

It should be appreciated that these gene-based tests can be applied to the setting of rare variant which are increasingly being studied. Extensions to rare variant tests to incorporate gene-gene and gene-environment interactions have been developed and include SKAT [43] which can handle gene-gene interactions and iSKAT [44] and rareGE [45] for gene-environment interactions.

3.2.9 Replication

Regardless of the analysis approach taken to identify gene-gene interactions, the gold standard is to conduct a replication analysis using an independent dataset. While fairly common for single-variant association studies, this is less routinely followed for gene-gene interaction analyses, as demonstrated in our systematic review of asthma gene-gene interactions where only 15.2% of interactions were attempted to be replicated [46]. The challenge is often identifying an appropriate replication dataset in which both variants were genotyped. Through the use of imputation the variant to be replicated could be imputed if they were not directly in the independent dataset.

A challenge for any replication of a genetic effect is the directionality of the effect. Differences due to the populations selected that can alter the minor allele frequencies and linkage disequilibrium structure can result in differences in both the magnitude and direction of effect [47]. This is amplified when looking at two or more loci as the probability of subtle differences can result in differences in the direction of effect when looking at interactions.

3.2.10 Meta-Analysis

This data analysis technique is commonly used to pool data across multiple studies and increase evidence for an association with a genetic variant. Meta-analysis is an attractive analytical technique because it can be used to increase power by substantially increasing the total sample size. While relatively straightforward for single-variant analyses using either fixed or random effects models [48] or by combining *p*-values [49], this is not always the case for gene-gene and gene-environment interactions. Differences in the analytical strategy and the way in which the results are presented may make meta-analyses more challenging. The *p*-value approach does not take into account the effect estimates, but this may be a better first approach as it can be used on a much broader set of analytical strategies such as MDR, regression-based approaches, and Random Forrest methods, for example. Meta-analysis methods that account for the effect estimate are more attractive because they can account for heterogeneity in both the effect estimates and between populations.

3.2.11 Case-Only Approach

The case-only study design to detect interactions was first described for gene-environment interactions [50, 51]. This design is premised on the idea that the variant is independent of the environmental exposure in the population. In the presence of a gene-environment interaction, there would be an association between the variant and the environmental exposure among cases only and this can be most easily tested using the regression-based approaches described above (linear or logistic depending on the environmental exposure being investigated); however, it should be noted that when using the case-only design the main effects of the variant and the environmental exposure cannot be determined as this is only possible using a case-control design.

The case-only design is more powerful than the case-control design (Table 4; 5000 cases and 5000 controls for case-control and 5000 cases for case-only, main effects ORs of 1.2 for both the environmental factor and genetic variant, a frequency of the environmental exposure of 0.25 in the population and MAF of 0.2, population risk of disease of 0.1 and 200,000 genetic markers). The major caveat is that there is an assumption that there is independence between the genetic variant and the environmental factor, i.e., that there is no association between these factors among the source population. Violation of this assumption can lead to biased estimates of the OR and there is a recommendation that a case-only gene-environmental interaction analysis only proceed for environmental factors with population-specific data [52]. This independence may be difficult to establish in population-based data, so independence could be tested for among controls (e.g., when a

Table 4
Power of case-control and case-only gene-environment interaction analyses

intOR	Case-control	Case-only
1.1	0.000	0.001
1.2	0.002	0.063
1.3	0.038	0.564
1.4	0.211	0.962
1.5	0.541	1.000
1.6	0.829	1.000
1.7	0.959	1.000
1.8	0.994	1.000
1.9	0.999	1.000
2.0	1.000	1.000

case-only analysis is being conducted among a case-control study) or replication of the finding among an independent case-control study. This is of particular importance when the environmental factor could be under genetic influence such as BMI. The substantial power gain of the case-only study makes this analysis strategy relatively attractive, but this is somewhat diminished by the strong independence assumption and the need to test this assumption in a set of population-based samples.

The case-only approach can also be utilized to detect gene-gene interactions [53]. In the presence of an interaction between two genetic variants there will be an association between these genetic variants among cases. This is most easily tested using a logistic regression model using one of the genetic markers as the dependent variable. As with case-only gene-environment interaction analyses, only the interaction can be examined and not the main effects of the genetic variants. Independence of the genetic variants in the source population is assumed, but this is much easier to achieve given the large amounts of genome-wide genotype data available for many diverse populations (e.g., HapMap [54]) or the use of LD pruning among controls, if available (*see* Subheading 3.1.4).

3.2.12 Three or More Variants

Gene-gene interactions are not limited to pairwise interactions. Higher order interactions involving three or more variants have been reported (e.g., renin-angiotensin system SNPs and hypertension [55]) and can be modeled in the regression framework. The problem is that as higher order interactions are considered the data become more and more sparse and the power to detect a significant interaction decreases. One approach to detecting higher order interactions would be to use a nonparametric, non-explicit method such as MDR to screen for potential higher order combinations of variants and then subsequently explicitly test for the interaction on a subset of the best performing combinations using a regression-based method.

3.3 Gene-Environment Interaction Analysis

At the core, the analysis of gene-environment interactions does not fundamentally differ from gene-gene interactions. The analysis can be conducted using the regression-based or nonparametric approaches described above with the same caveats for testing for explicit vs. non-explicit interactions in the dataset. Joint, or non--explicit, tests may be used to screen the data to detect potential interactions, followed by explicit tests to determine if an interaction exists between the identified variant and the environmental factor. In the logistic regression framework, one of the genetic variant variables is replaced by the environmental exposure of interest and the significance of the interaction β is tested in the model. Using the non-explicit MDR approach, the environmental variable is entered into the algorithm with the caveat that the environmental factor must be categorical in order for the reduction algorithm to work.

The power to detect gene-environment interactions suffers from the same lack of power on the genome-wide scale as gene-gene interactions. More than 30,000 cases and controls are required to detect an interaction OR of 1.2 (Table 1) under similar parameters as the gene-gene interactions analysis (main effects of OR = 1.2 for gene and environmental factors, MAF of 0.2, environmental prevalence of 0.25, population risk of disease of 0.1 and 200,000 genetic markers tested). Using 5000 cases and 5000 controls, the smallest detectable OR is just under 1.6 (Table 4). The same approaches to increase power by reducing the search space can be applied to the search for gene-environment interaction analyses, through all of the data reduction techniques described for gene-gene interactions.

It cannot be stressed enough that replication is the key to describing true positive gene-environment interactions, as was discussed for gene-gene interactions. As mentioned previously for gene-gene interaction replication, if the exact genetic variants are not directly genotyped in the independent dataset, imputation can be used to estimate the genotypes to be replicated. While this will work for the genetic variants in gene-environment interactions to be replicated, this is not the case if there is not comparable environmental data in the independent dataset. This can make the identification and selection of an appropriate dataset for replication more challenging for gene-environment interactions, but makes the replication of the findings no less important.

3.3.1 Modifiable Environmental Factors

One attractive aspect of identifying and describing gene-environment interactions for a complex disease is that this gives us the possibility of potentially modifying one contributing factor to a disease. At this point in time, inherited genetic variants are not modifiable, but if, for example, a significant gene-environment interaction were identified for obesity with a genetic variant and high saturated fat diet, individuals carrying the risk variant could be more strongly encouraged to reduce their saturated fat intake. Despite the general benefit we could all gain from reducing our saturated fat intake, this may be more effective if it were targeted to individuals based on their genetic profile and increased risk for obesity when both factors are present (beyond the additive main effects).

3.3.2 Gene-Gene-Environment Interactions

We should not think of gene-gene and gene-environment interactions as being mutually exclusive. As with higher order gene-gene interactions, interactions involving multiple genetic variants and an environmental factor are possible to model. The same issues with power due to sparsity of data and the number of unique combinations of factors apply, but given sufficient sample size these types of interaction models can be tested. It may be more important in this setting to attempt a screening step using a non-explicit approach and then apply an explicit test only to the set of interactions that surpass an initial significance threshold.

3.4 Conclusions

I have outlined several methods for conducting interaction analysis to detect both gene-gene and gene-environment interactions. However, it should be clear that there is no optimal method to detect either type of interaction. The method(s) chosen are often dictated based on the type of data you have available (e.g., case-control, case-only), the number of markers you have genotyped, and the number of subjects you have included. I strongly recommend reporting the results of explicit tests for interaction as this will greatly improve ability of other groups to attempt to replicate your results and meta-analyze where appropriate, but the use of non-explicit tests for interaction can be extremely useful to initially screen large numbers of interactions and when sample sizes are limited. We must continue to invest time and resources into identifying interactions in genome-wide data as this will help us to fill in the missing heritability gap and better understand the genetic architecture of complex traits.

4 Notes

1. Inheritance models: It should be noted that as with single-variant approaches, inheritance models can be imposed on the variants that include additive, multiplicative, dominant, recessive, and overdominant. These inheritance models can be imposed on the variants in the interaction model independently and can be considered in a combinatorial fashion for variant 1 and variant 2 (e.g., additive × additive, additive × multiplicative, additive × dominant, etc.). However, if this is done, care must be taken to account for this additional multiple testing. These models can be problematic when they are misspecified as they can reduce power which makes nonparametric approaches attractive.
2. Scale: The scale on which the outcome is measured or evaluated can influence whether or not an interaction exists between two variants or between a variant and an environmental factor. It needs to be kept in mind that there may be a monotone transformation of the outcome that could remove the interaction. For example, on the odds ratio scale, an interaction may exist between two SNPs (coded dichotomously as in a dominant model) in which the OR for having dominant alleles at both SNPs is greater than the sum of the ORs for having one dominant allele at each SNP. However, on the log(OR) scale this interaction is removed and termed a removable interaction. If there is no monotone transformation that can remove this interaction it is termed essential. The method to detect these types of interactions is described in a paper by Wu et al. [56].

3. Reporting of interaction results: As was previously outlined in our paper [46], there are several recommendations for how results of interaction analyses are reported in order to increase the interpretability and replicability of the interaction. Effect estimates should be provided so that both the strength and direction of the interaction can be assessed. If a regression-based approach is used, parameter estimates of the main effects and the interaction term should be provided. If a nonparametric approach is used, such as MDR, effect estimates are not produced, however, counts of cases and controls for the contingency table of genotype combinations should be provided. This will allow for a better assessment of the interaction and the possibility to incorporate the data into a meta-analysis.

References

1. Niel C, Sinoquet C, Dina C et al (2015) A survey about methods dedicated to epistasis detection. Front Genet 6:285
2. Ritchie MD (2015) Finding the epistasis needles in the genome-wide haystack. Methods Mol Biol 1253:19–33
3. Gusareva ES, Van Steen K (2014) Practical aspects of genome-wide association interaction analysis. Hum Genet 133(11):1343–1358
4. Tiret L (2002) Gene-environment interaction: a central concept in multifactorial diseases. Proc Nutr Soc 61(4):457–463
5. Ottman R (1990) An epidemiologic approach to gene-environment interaction. Genet Epidemiol 7(3):177–185
6. Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. Nature 461(7265):747–753
7. Bateson W (1909) Mendel's principles of heredity. Cambridge University Press, Cambridge
8. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11 (20):2463–2468
9. Moore JH (2005) A global view of epistasis. Nat Genet 37(1):13–14
10. Ma J, Thabane L, Beyene J et al (2016) Power analysis for population-based longitudinal studies investigating gene-environment interactions in chronic diseases: a simulation study. PLoS One 11(2):e0149940
11. Dunham I, Kundaje A, Aldred SF et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489 (7414):57–74
12. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164
13. Bush WS, Dudek SM, Ritchie MD (2009) Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput:368–379
14. Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909
15. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5(6): e1000529
16. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575
17. Ueki M, Cordell HJ (2012) Improved statistics for genome-wide interaction analysis. PLoS Genet 8(4):e1002625
18. Wu X, Dong H, Luo L et al (2010) A novel statistic for genome-wide interaction analysis. PLoS Genet 6(9):e1001131
19. Wan X, Yang C, Yang Q et al (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 87(3):325–340
20. Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19(3):376–382
21. Ritchie MD, Hahn LW, Roodi N et al (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69(1):138–147

22. Calle ML, Urrea V, Malats N et al (2010) mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. Bioinformatics 26 (17):2198–2199

23. Gui J, Moore JH, Williams SM et al (2013) A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. PLoS One 8(6):e66545

24. Van der Auwera GA, Carneiro MO, Hartl C et al (2013) From FASTQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10 1–11.1033

25. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26(5):589–595

26. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079

27. McKenna A, Hanna M, Banks E et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20 (9):1297–1303

28. DePristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498

29. Dewan AT, Egan KB, Hellenbrand K et al (2012) Whole-exome sequencing of a pedigree segregating asthma. BMC Med Genet 13 (1):95

30. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37(4):413–417

31. Calle ML, Urrea V, Vellalta G, Malats N, Steen KV (2008) Improving strategies for detecting genetic patterns of disease susceptibility in association studies. Stat Med 27 (30):6532–6546

32. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57 (1):289–300

33. Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74(4):765–769

34. North BV, Curtis D, Sham PC (2002) A note on the calculation of empirical P values from Monte Carlo procedures. Am J Hum Genet 71 (2):439–441

35. North BV, Curtis D, Sham PC (2003) A note on calculation of empirical P values from Monte Carlo procedure. Am J Hum Genet 72 (2):498–499

36. Murk W, DeWan AT (2016) Exhaustive genome-wide search for SNP-SNP interactions across 10 human diseases. G3 (Bethesda) 6 (7):2043–2050

37. Gauderma WJ, Morrison JM, QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies. http://hydra.usc.edu/gxe2006

38. Uzun A, Sharma S, Padbury J (2012) A bioinformatics approach to preterm birth. Am J Reprod Immunol 67(4):273–277

39. Uzun A, Triche EW, Schuster J et al (2016) dbPEC: a comprehensive literature-based database for preeclampsia related genes and phenotypes. Database (Oxford). https://doi.org/10.1093/database/baw006. pii:baw006

40. Shearer AE, Eppsteiner RW, Booth KT et al (2014) Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. Am J Hum Genet 95(4):445–453

41. Murk W, DeWan AT (2016) Genome-wide search identifies a gene-gene interaction between 20p13 and 2q14 in asthma. BMC Genet 17(1):102

42. Ma L, Clark AG, Keinan A (2013) Gene-based testing of interactions in association studies of quantitative traits. PLoS Genet 9(2):e1003321

43. Wu MC, Lee S, Cai T et al (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93

44. Lin X, Lee S, Wu MC et al (2016) Test for rare variants by environment interactions in sequencing association studies. Biometrics 72 (1):156–164

45. Chen H, Meigs JB, Dupuis J (2014) Incorporating gene-environment interaction in testing for association with rare genetic variants. Hum Hered 78(2):81–90

46. Murk W, Bracken MB, DeWan AT (2015) Confronting the missing epistasis problem: on the reproducibility of gene-gene interactions. Hum Genet 134(8):837–849

47. Greene CS, Penrod NM, Williams SM et al (2009) Failure to replicate a genetic association may provide important clues about genetic architecture. PLoS One 4(6):e5639

48. Fleiss JL (1993) The statistical basis of meta-analysis. Stat Methods Med Res 2(2):121–145

49. Fisher RA (1948) Combining independent tests of significance. Am Stat 2:30

50. Piegorsch WW, Weinberg CR, Taylor JA (1994) Non-hierarchical logistic models and

case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13(2):153–162

51. Begg CB, Zhang ZF (1994) Statistical analysis of molecular epidemiology studies employing case-series. Cancer Epidemiol Biomark Prev 3 (2):173–175
52. Hodgson ME, Olshan AF, North KE et al (2012) The case-only independence assumption: associations between genetic polymorphisms and smoking among controls in two population-based studies. Int J Mol Epidemiol Genet 3(4):333–360
53. Yang Q, Khoury MJ, Sun F et al (1999) Case-only design to measure gene-gene interaction. Epidemiology 10(2):167–170
54. The International HapMap Consortium (2003) The international HapMap project. Nature 426:789–796
55. Yang CH, Lin YD, Wu SJ et al (2015) High order gene-gene interactions in eight single nucleotide polymorphisms of renin-angiotensin system genes for hypertension association study. Biomed Res Int 2015:454091
56. Wu C, Zhang H, Liu X et al (2009) Detecting essential and removable interactions in genome-wide association studies. Stat Interface 2(2):161–170

Author Queries

Chapter No.: 7 394545_1_En

Query Refs.	Details Required	Author's response
AU1	Please check whether the affiliation and correspondence details are presented correctly.	
AU2	Please check the hierarchy of the section headings and confirm if correct.	
AU3	Please check the sentence starting: "Then converted to BAM format, sorted, indexed, PCR duplicates marked and then. . ." for completeness.	
AU4	Please provide volume number for Ref. [13].	