Published in final edited form as:

Per Med. 2011 January ; 8(1): 59–70. doi:10.2217/pme.10.75.

Genotype–environment interactions and their translational implications

Tesfaye M Baye^{1,†}, Tilahun Abebe², and Russell A Wilke³

¹ Cincinnati Children's Hospital Medical Center, Division of Asthma Research, Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA

² Department of Biology, University of Northern Iowa, Cedar Falls, IA, USA

³ Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

Abstract

Organisms frequently encounter different environmental conditions. The physiological and behavioral responses to these conditions depend on the genetic make up of individuals. Genotype generally remains constant from one environment to another, although occasional spontaneous mutations may occur which cause it to change. However, when the same genotype is subjected to different environments, it can produce a wide range of phenotypes. These phenotypic variations are attributable to the effect of the environment on the expression and function of genes influencing the trait. Changes in the relative performance of genotypes across different environments are referred to as genotype–environment interactions (GEI). A general argument for research on the impact of GEI in common diseases is that it provides insights into disease processes at the population, individual and molecular levels. In humans, GEI is complicated by multiple factors including phenocopies, genocopies, epigenetics and imprinting. A better understanding of GEI is essential if patients are to make informed health choices guided by their genomic information. In this article, we clarify the role of the environment on phenotype, we describe how human population structure can obscure the resolution of GEI and we discuss how emerging biobanks across the globe can be coordinated to further our understanding of genotype–phenotype associations within the context of varying environment.

Keywords

biobanks; genome; genotype–environment interaction; genotype–treatment interaction; personalized medicine; pharmacogenetics; population structure

The concept that phenotype represents the consequence of genotype–environment interactions (GEI) is universal and relates to all living organisms. Garrod was one of the first scientists to note that the effect of genes on phenotype could be modified by the environment (E) [1]. He suggested that differences in the genetic composition play a major role in variable response to drugs, and that this effect of the genotype (G) could be further

© 2011 Future Medicine Ltd

[†]Author for correspondence: Tel.: +1 513 803 2766, Fax: +1 513 636 1657, tesfaye.mersha@cchmc.org.

For reprint orders, please contact: reprints@futuremedicine.com

Financial & competing interests disclosure

This work was supported by K01HL103165 and P30HL10133 (TMB) grants, the University of Northern Iowa (TA, USA) and R01DK080007 (RAW). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

modified by diet. Similarly, Turesson demonstrated that the development of a plant is often influenced by its surroundings [2]. He articulated the existence of a close relationship between varieties of crop plants and their environment, and stressed that the presence of a particular variety in a given locality is not just a chance occurrence; rather there is a genetic component that helps the individual adapt to that area. He pointed out the role of selection in directing the 'genotypical differentiation of the population in a given locality'. Wright further elaborated the existence of a functional relationship between various biological end points and networks of genes and environmental factors in his studies of mutation, selection and breeding [3].

Additional complexity is introduced when these factors are transmitted to generations through epigenetics [4]. In response to severe environmental changes, a genome can respond by selectively regulating (increasing or decreasing) the expression of specific genes. Plants, for example, can detect and respond to specific environmental signals that affect developmental pathways and confer a wide range of adaptive capacities over time [5]. In cultivated maize, variation in genome size can reach nearly 40% [6,7]. Rayburn and Auger quantified the nuclear DNA content of 12 Southwestern USA maize populations collected at various altitudes, and they observed a significant positive correlation between genome size and altitude [8].

In humans, environmental effects are complex, and their relative importance may vary considerably according to the genetic constitution of individuals. The physical environment can include such factors as pollutants [101]. Physical, social and economic factors further conspire to drive GEI through complex environmental insults such as tobacco smoke. As a result, GEI can influence disease onset, the rate of disease progression and the clinical response to pharmacological intervention. The rapidly increasing prevalence of many common human disease traits clearly illustrates the importance of GEI on disease onset. Asthma rates in the USA have risen dramatically over the past 20 years. Since this is too short a period of time for most allele frequencies to change significantly, nongenetic factors, that is, changes in the environment, must be important. Similarly, the doubling of obesity rates seen in many countries during the past few decades did not result from changes in the underlying genetic architecture, but rather from alterations in the environment, primarily what people eat and how much they move [9]. The human genome initially evolved to handle exhausting physical activity in a lifestyle where resources were scarce and famine was common (i.e., traditional hunter-gatherer lifestyle). As we moved to a 'modern' lifestyle, our genome has not had time to adapt to sedentary living with abundant food and material goods. Thus, a large percentage of the population in the USA and other developed countries is at risk of developing complex diseases including obesity, diabetes and hypertension. Just as the genome is made up of genes controlling the phenotype, the environment consists of an amalgamation of biological and physical factors, which independently or in combination affect the genome.

Classification of G–E interactions

Genotype–environment interactions can be grouped into three broad categories (Figure 1) 'no' G–E interaction, noncrossover interaction and crossover interaction. As the number of environments and the number of genotypes increase, the number of possible G–E interactions (given by $GE!/G!E!$) increases tremendously. With only two Gs and two Es, and with only a single criterion, at least four different types of interactions are possible. Thus, with ten Gs and ten Es, 400 types of interactions are possible, which would certainly make their implications and interpretation more difficult to comprehend [10,11]. The interactions are usually higher for diseases and traits with lower heritabilities, such as asthma and other common complex diseases such as obesity. The interactions are also higher for traits such as

reproduction and feeding efficiency, whereas they are lower for traits with higher heritabilities, such as eye or skin color, sickle-cell anemia and other Mendelian diseases. Interactions may allow for specific targeting of interventions in high-risk groups. In situations where there are reduced patterns of interaction, the implications of G–E for informing targeted interventions and prevention of disease are likely to be very limited.

No G–E interaction

When there is no G–E interaction, the effects of each of the risk factors are consistent (homogeneous) across the levels of the other risk factors. A ‘no’ G–E interaction occurs when one genotype (e.g., G1) consistently performs better than the other genotype (G2) by approximately the same amount across both Es. In such a situation, SNP markers tested in one E provide universal results. When there is no noise, experimental results would be exact in identifying the best G without error, and there would be no need for replication. Within this context, one replication at one E would be sufficient to identify the best SNPs that pharmacogenomicists could rely on. Figure 1A illustrates that Gs G1 and G2 perform similarly in two Es, because their responses are parallel and stable. This type of stability, also referred to as biological stability [12], is desirable in pharmacogenomics. Figure 1B also illustrates a no G–E interaction. Genotype G1 performs better than genotype G2 in both Es. The norms of reaction (variations in trait expression across a range of Es for a given G) for the two Gs are additive. The intergenotypic variance remains unchanged in the two Es and the direction of environmental modification of Gs is the same. In Figure 1A, there is a main effect of G, and in Figure 1B there is a main effect of E.

Noncrossover (quantitative) G–E interaction

A noncrossover G–E interaction is said to occur when one G (G1) consistently outperforms another (G2) across the test E. However, unlike in Figure 1A or Figure 1B, the differential performance is not the same across the E. Figure 1C represents a noncrossover type of interaction. The Gs G1 and G2 respond differently to the two Es but their ranks remain unchanged. The response of the two Gs under different Es is not additive, the magnitude of intergenotypic variance increases, and the environmental modification of the two Gs are in the same direction.

Crossover (qualitative) G–E interaction

The differential and nonstable response of Gs to diverse Es is referred to as a crossover interaction when the ranks of G change or switch from one E to another [13]. In human disease genetic studies, the failure to replicate candidate genes or genome-wide association studies (GWAS) could be attributed to crossover interactions. Crossover interaction implies that no G is superior in multiple Es [14]. If the effect of a treatment (T) differs from trial to trial, especially when the effects are positive in some studies and negative in others, no general therapy recommendation can be made. Differences in the response of Gs to the E may necessitate the development of geographic-specific personalized medicine strategies [15].

Figure 1D represents a crossover, rank change type of interaction. The direction of environmental modification of Gs G1 and G2 is opposite: the performance of G1 increases and that of G2 decreases. The genotypic ranks change between the two Es, but the magnitude of intergenotypic variance remains unchanged. Figure 1E also represents a crossover interaction as Gs switch ranks between the two Es. It also represents a change in magnitude of intergenotypic variance changes. In E E1, the difference between Gs G1 and G2 is smaller than that in E E2, and the direction of environmental modification of the two Gs is the same. Figure 1F illustrates a crossover interaction with the environmental modification in opposite direction; performance of G1 increases but that of G2 decreases.

This situation is different from that illustrated in Figure 1D in that the magnitude of intergenotypic variance increases between Es.

Genetic structure & G–E interaction

The magnitude of a GEI is influenced by the genetic structure of the G. Gs with less heterogeneity or heterozygosity generally interact more with the E than mixtures of Gs, because of lower amounts of adaptive genes. The genetic structure of a population differs mainly in two respects: the level of heterozygosity at the population level and the amount of genetic heterogeneity within the individual [16]. In the absence of GEIs, the variance between individuals (in cases where the individuals are genetically alike) is expected to be homogeneous. In contrast with population-based studies in which the average effect of an environmental exposure is compared between groups, the identification of susceptible individuals within populations via genotyping allows a better estimation of the true magnitude of the effect of an environmental exposure on the population at risk.

Population genetic structure, a special type of confounding in allelic association studies [17], may also affect G–E interaction, possibly leading to spurious findings. Campbell and colleagues found that a SNP in the lactase (*LCT*) gene was initially strongly associated with height in a sample of European–Americans [18]. This association was later found to be due to stratification; both height and the frequency of the SNP varied widely across Europe. When subjects were rematched on the basis of the refined location of ancestry, the strength of the initial association was greatly diminished. Since population subgroups are expected to differ with respect to environmental exposures, the same type of confounding could occur in studies of G–E interaction. That is, different subgroups may have both different genetic backgrounds and different cultures or socioeconomically influenced patterns of behavior, creating a correlation between G and environmental exposure that must be controlled. Without this control, spurious associations result. Unlinked genetic markers can be used to detect such stratification and make corrections when it is present [19].

Modeling the interaction

To understand the relationship between human complex diseases and the E, we begin with the fundamental relationship of G, E, phenotype (P), and GEI model in randomized controlled trials (RCTs). If no interaction between G and E is assumed, a P can be expressed as equation $P = G + E$. However, observed phenotype is a function of G, E, and their GEI. For GEI to produce an array of phenotypes and be detected via a statistical procedure there must be at least two distinct Gs evaluated in at least two different Es. The components of GEI can be explained as follows [20]:

$$GEI = (\Delta 2 - \Delta 1) = (\Delta 4 - \Delta 3)$$

or

$$GEI = (\Delta 1 - \Delta 4) = (\Delta 2 - \Delta 3)$$

The G effect, $\Delta 3$, represents the difference between Gs in E E1, and $\Delta 4$ represents the difference between Gs in E E2. The environmental effect, $\Delta 1$, represents the change attributable to Es for G G1, and $\Delta 2$ is the change attributable to Es for G G2. Thus, the total effect is:

$$P=G+E+GE$$

or

$$GE=P - G - E$$

This model can be written from a statistical standpoint as:

$$P_{ij}=\mu+G_i+E_j+GE_{ij}+\varepsilon_{ijk}$$

Where, P_{ij} is phenotype of an individual with G_i and E_j , μ is the overall mean and ε_{ijk} is the random error for the k th patient in the group with G_i and E_j .

Expanding the model

The G–T interaction (GTI) effect can be interpreted as the response of a G to T in a given E. The GTI study helps to identify those individuals who will respond (un)favorably to the drug candidate based on their G. Finding genes that modify drug response has the potential to significantly improve drug delivery. The GTI can be partitioned using clinical trial design and analysis. Analysis methods include both parametric and nonparametric procedures – partitioning of variance, regression analysis, nonparametric methods and multivariate techniques (for details refer to [21,22]). Different models of analysis of variance (ANOVA) are used for partitioning variance. In candidate gene association, Gs are usually chosen not at random since they are deliberately selected based on biology. Similarly, E or Ts are often not randomly chosen. However, they may be considered random if there are many of them spread over a large E. If the Gs are evaluated over several collaborative Es, the effects can be random since the E is not controlled. In addition, if Gs are tested from random pool such as from a GWAS study, their effects can be assumed to be random. In the ANOVA model the resultant gene effect (X_{ijk}) is assumed to be the result of T, G and E effects and their interactions over T_k , E_j and G_i :

$$X_{ijk}=\mu+G_i+E_j+T_k+GE_{ij}+GT_{ik}+ET_{jk}+GET_{jk}+e_{ijk}$$

Where μ as the overall mean. The remaining variation is captured in the error term e_{ijk} . ANOVA model taking G, T and E as random are presented in Table 1. Variance components can be calculated from Table 1 using the following equations.

$$\begin{aligned}\sigma_g^2 &= \frac{m1-m4-m5+m7}{rte} \\ \sigma_{gt}^2 &= \frac{m4-m7}{re} \\ \sigma_{ge}^2 &= \frac{m5-m7}{rt} \\ \sigma_{gte}^2 &= \frac{m7-m8}{r} \\ \sigma^2 &= m8\end{aligned}$$

Genotype–treatment interaction has significant influence on the efficiency of drugs, largely because it confounds comparisons among Gs with the test E. It is argued that to overcome these constraints in drug evaluation, we need to develop an understanding of the differences in Gs associated with differences in drug efficacy. GTI is of interest to pharmacogenomist

for several reasons. The need to develop drugs for specific purposes is determined by an understanding of GTI. Unique drugs may be required for different races or G combinations. The need for unique drugs in different geographical areas requires an understanding of GTI. The importance of this interaction can determine if the clustering of a large geographical area into subareas is needed and justified for testing new drugs. Effective allocation of resources for testing drugs across genotypic combination is based on the relative importance of G–drug interactions. The response of Gs to variable drug doses provides an understanding of their efficacy and toxicity.

Study design

A purely DNA sequence-based approach is not sufficient to fully explain the risks of common diseases. Rather, diseases result from interactions between the individual genetic make up and environmental factors [23]. Therefore, although certain genes individually or in combination with other genes may increase the chance of developing diseases [24], the unfavorable effects due to GEIs can be partially overcome by modifying the social, behavior or environmental conditions (Figure 2). From a statistical point of view, interaction can be defined as a deviation from conditional independence [25]. This definition is entirely dependent on the measurement scale (multiplicative or additive) used. Ratio measures such as relative risks (RRs) or odds ratios assess the effects of risk factors on a multiplicative scale, because they reflect the degree to which disease risk (for RR) or odds (for odds ratio) are multiplied in individuals with the risk factor compared with those without. By contrast, risk differences assess the effects of risk factors on an additive scale, because they reflect how much disease risk is added in individuals who have the risk factor, compared with those who do not. The statistical definition of interaction differs depending on which of these measurement scales is used. For example, for factors A and B, interaction on a multiplicative scale is defined as a different RR for factor A across strata defined by factor B, while on an additive scale, interaction is defined as a different risk differences for factor A across strata defined by factor B. Interaction where G–E is greater than additive supports an approach targeting high-risk groups [25].

There are several ways in which an epidemiological study could be designed for testing models of GE and GTI. Observational and intervention are the two main types of design used to investigate interactions (for details refer to [26,27]). Most epidemiological studies are observational and include traditional case–control, cohort, practice-based cohort, adoption and twin study design [28,29], which are carried out with the implicit assumption that any variation that is not attributable to genetic factors must stem from the E [30]. The different designs offer advantages and disadvantages with respect to validity and efficiency [31]. It is difficult to objectively estimate the duration, intensity and frequency of a large variety of multidirectional environmental influences from observational epidemiological studies [32]. Even strong associations between an environmental factor and a disease do not necessarily prove that the environmental factor has caused the disease [32]. One way around this would be to carry out a RCT with replication and control subjects (for details refer to [22]). RCTs are the cornerstone of evidence-based medicine. Such trials rely on the random assignment of individuals to different Ts one of which could be a control to assess baseline patient factors that could affect outcomes. Cluster randomized trials, that is, trials which randomize intact groups of individuals (‘clusters’) instead of the individuals themselves, have become common in health and healthcare research [33].

Epidemiological and controlled experiment studies are useful to indicate the presence of GEI. However, the epidemiological method is limited by its observational studies of naturally occurring genetic polymorphisms and environmental variability and its inability to separate genetic and environmental factors. An experimental methodology is required to single out environmental/T effects independent of G through randomized clinical trials. The

use of experimental animal models has provided a great deal of information about the genetic, physiological and environmental aspects of complex disorders [25]. Whereas animal genetic studies are mainly experimental, human genetic studies are observational since the manipulation of the human genome is unethical. Moreover, human studies are expensive, and characterization of the impact of GEI on P is often quite challenging because we cannot control the subjects' E and life history. Finally, associations do not necessarily imply causation because of possible confounding due to population structure. There are two general strategies to study the role of genes in humans. The measured G approach is based on the direct measurement of genetic variation at the protein or DNA levels in an effort to assess the effect of allelic variability on phenotypic differences. The unmeasured (and perhaps unknown) G approach attempts to estimate the contribution of genetic variance (from the environmental risk factors) to differences in disease expression (phenotypic variance) and to find a quantitative evidence for the role of single genes in the development of the disease in question.

Case-control studies—Most association studies (gene or genome-wide), which do not consider familial inheritance patterns, use a case-control design based on allele or G frequency comparison of unrelated affected and unaffected individuals in the population [34]. An allele in a gene is said to be associated with a trait if it occurs at a significantly higher frequency in affected individuals compared with the control group (i.e., when the null hypothesis of equal allele frequency across groups is false). The statistical significance can be assessed by a Pearson χ^2 test of homogeneity of proportions. The strongest evidence in support of a reported association will be replication of association with the same allele, the same P and the same direction of effect in an independent population sample, with combined p-values that are low enough to survive a conservative correction for testing multiple hypotheses.

Practice-based cohorts—Observational cohort studies represent an attractive alternative. Currently, there is tremendous interest in linking DNA repositories to secure encrypted copies of comprehensive electronic medical records. One such effort has been the electronic Medical Records and Genomics (eMERGE) network, established by the NIH in 2007 [102]. By linking high-throughput genotyping technologies to electronic health care records within the context of these biobanks, investigators can study and dissect the effect of E at multiple geographic locations [35] (Figure 3). Ancestral diversity varies between biobanks; some are highly homogeneous [36], while others contain a considerable amount of admixture [37]. Furthermore, environmental heterogeneity in the various biobanks can introduce another level of analytical challenge. This problem demands increased cooperation and the standardization of data entry and phenotyping methods [38].

One approach to minimizing the impact of GEIs within and across biobank data has been to group Gs (subjects with similar ancestral background) according to their response to the E via cluster analysis [38]. The resulting data may be useful in developing predictive tools that describe expected maps of genetic variation over geographic regions such as cities, counties and states. The other more traditional approach to minimizing the impact of such interaction has been to perform stability analysis across diverse Es by analyzing and interpreting genotypic differences within the context of highly variable clinical phenotypes such as T outcome [13]. This latter approach would allow researchers to select Gs with consistent outcome measures, identify the causes of GEI and provide the opportunity to correct the problem [38]. When the cause for the unstable G is known, a variety of options present themselves: the G can either be improved by genetic means (as in the case of plant genetics), or a proper E (change in drug exposure) can be selected to optimize clinical outcome. A G that performs consistently across many Es would possibly possess broad-based, durable tolerance to environmental factors encountered during metabolism. The more providers

know about GEI, the more likely they are to efficiently implement appropriate personalized medicine strategies. Population-based and disease-oriented biobanks are essential to establish the disease relevance of human genes and provide opportunities to evaluate their interaction with lifestyle and E and for the development of personalized medicine. However, detailed disease phenotype characterization and highly specified sample collection procedures are expensive and laborious (Table 2). The recently established Public Population Project in Genomics (P³G) [103] promotes the collaboration between researchers in the field of population genomics and biobanking to ensure public access to population genomics data. It supports the construction of cross-sectional baseline questionnaires to define a core set of information that is of particular scientific relevance for a specific type of biobank. More than 25 international biobanks have contributed to the conception of Data Schema and Harmonization Platform for Epidemiological Research (DataSHaPER). P³G fosters the harmonization of nomenclature of biological, medical, demographic and social data collected from participants, mainly in the context of population-based studies to dissect GEIs.

Limitations—For studies of T outcome, confounding by indication can be a strong source of bias in observational studies. This is because medications are not prescribed at random. Presence of comorbid conditions (and greater severity of illness) impacts T choice. Newer and more costly medications are often reserved for patients with more severe illnesses. To minimize this source of bias, investigators working with observational data may need to employ propensity scores to adjust for observed patient or provider characteristics which influence medication choice or selection. Self selection can also be a strong source of bias. In a practice-based observational study, patients who have continued to take their medication over long periods of time are not representative of all patients for whom the drug has been prescribed. Adherence may reflect better self-care practices or the ability to tolerate the medication without significant adverse events. Most observational studies disproportionately capture persons successful in remaining on medications for longer periods, weighting the data to the exposure experience of these specific patients. These sources of bias must therefore be addressed, and factored into any analysis characterizing genetics, E, T and clinical outcome.

Moving forward

Genome-wide association studies have emerged as powerful approaches for identifying genetic variants influencing common diseases, and complex traits such as T outcome [39–43]. Most genetic loci discovered to date, however, only account for a small fraction of the total phenotypic variation, and most of the inherited component of risk remains unexplained [44,45]. Some of this missing inherited risk (i.e., the proportion not attributable to variants identified to date) can be attributed to GEI [46]. Nearly all GWAS conducted to date have concentrated on detecting and characterizing main effects (one SNP at a time) and have not fully explored the potential role of environmental factors in modifying genetic risk [47]. Moreover, the current attention in population-based association studies is focused almost entirely on genetic markers and etiological variants that are common (>5% frequency).

The current emphasis on common alleles is purely for practical reasons [48]. Common diseases are assumed to be influenced by many genetic and environmental factors, all with a modest effect on the trait. If the genetic influences are rare, the sample sizes required to detect the modest effects become impossibly large. Thus, it is often impractical to search for rare genetic effects using a classical allelic association design [15]. This consideration explains the current focus on gene discovery strategies aimed at common alleles and implies that real effects associated with rare alleles may go undetected [49]. Large networks of

biobanks will be particularly useful for the rapid identification of genetic markers that predict rare adverse outcomes.

Because currently, association studies are based on genotyping known SNPs in the human genome since the cost of sequencing the entire genome for large numbers of individuals is prohibitive. However, it is possible to use data from the HapMap to estimate recombination maps across the genome to accurately infer Gs for SNPs not directly assayed in the study [50,51]. Inference of Gs allows for finer mapping of regions of interest and also has utility for validation and correction of data at genotyped markers. Furthermore, imputation of Gs at markers not directly assayed also provides the possibility of combining data from multiple biobank genome-wide scans that have used different SNP sets, since all SNPs genotyped in any of the studies may be inferred in other studies. This approach has led to the identification of novel genetic factors influencing the efficacy of HMGCoA reductase inhibitors (statins), the most commonly prescribed class of drugs in the USA [43]. By increasing the number of individuals for whom G information is available, such a strategy has the potential to increase the power to detect and dissect GEI [52].

Expanding GTI

A distinction must be made between the impact of genetic variants on disease risk and the impact of genetic variants on T outcome (G–T interactions) as illustrated in (Figure 4A & 4B). The degree to which these processes overlap depends on the clinical condition under consideration. For most common diseases, GEI modifies the disease process at the population, individual and molecular level. Conversely, GTI is the primary basis for individualizing T through predictive markers. Better understanding of GTI and GEI are, thus, essential if patients are to make informed health choices guided by their E, T and genomic information.

As an example, the onset of asthma is influenced by interactions between many genes as well as the E and developmental stages. In childhood, boys are as nearly as twice as likely to develop asthma as girls but in adulthood, asthma occurs more frequently among women than men [53]. Thus one could expand and model asthma based on developmental stages adjusting for gender risk ratio or within a given gender group as:

$$\frac{\text{gene}(s) - \text{environment interaction}}{\text{development}}$$

It has also been suggested that genetics may contribute as much as 60–80% of the interindividual variability in therapeutic response to asthma T [54]. Numerous genetic studies have reported linkage or association with asthma and the asthma-associated phenotypes; atopy, elevated immunoglobulin E (IgE) levels and bronchial hyperresponsiveness. In addition, specific alleles tagging cytokine/chemokine, remodeling, or IgE regulating genes have been shown to influence risk [55]. The clinical implications of variability in these candidate asthma genes (i.e., disease genes) with respect to their impact on T outcome remain largely undetermined.

Asthma drug responses vary widely between different populations and are also highly variable among individuals within the same population. One highly informative example is the variability observed between asthma patients exposed to inhaled β_2 -agonist therapy, where up to 75% of the variability is heritable. Approximately 60% of asthmatic children who are homozygous for arginine at position 16 (Arg 16/Arg 16) of the β -adrenergic receptor (also a candidate disease gene) may respond favorably to albuterol, compared with only 13% in individuals homozygous for glycine at that position [56]. The degree to which

this response varies is different among different ethnic groups [57]. Homozygosity for arginine at position 16 predicts therapeutic response to β 2-agonists in Puerto Ricans, but not in Mexicans [58]. There is also evidence to suggest that variants in the β 2-adrenergic receptor may explain differences in airway responsiveness in smokers versus nonsmokers [59]. Approaches such as these are necessary (i.e., modeling not only age and gender, but race and E as well) if individualized healthcare is to become a reality within the context of all common complex diseases.

Expanding GEI

To date, there have only been a few replicated, biologically plausible and methodologically sound examples of the application of GEI to individualized care [60–62]. Extending the discussion of atopic clinical disorders introduced above, Finnish Karelians have a higher prevalence of allergic diseases than Russian Karelians. Yet, both populations belong to the same ethnic group. A recent study compared associations between allergic diseases and *CD14* G in Finnish and Russian Karelian women. The *CD14* -159C/T (rs2569190) risk allele for atopic phenotypes in Finnish Karelia appears to be the protective allele in Russian Karelia. The risk allele was C in Russians and T in Finns [63]. In GEI terminology, this is an example of crossover (qualitative) G–E interaction. Similarly, *CD14* tested for association with total and specific IgE has demonstrated that the rs2569190 TT G is associated with lower IgE and decreased risk of sensitization in children exposed to pets, at 4 and 8 years of age [64]. Nonexposed and age matched children showed no association with the TT G of rs2569190. These examples illustrate that the direction and magnitude of a genetic effect can vary as the E changes. In other words, genetic risk for disease is modifiable in an E-specific manner.

Conclusion

Although concerns about the role of GEIs in disease etiology have developed over the last century, prioritizing these interactions as a means to prevent complex diseases remains an emerging area of study. E-based personalized disease prevention may be considered reasonable in cases when an exposure has a negative effect in one G group and a protective effect in another. Environmental risk factors are often complex and include respiratory infections, allergens, emotions, air pollution, cigarette smoke, lifestyle, dietary and psychosocial factors. Often it is difficult to identify the relevant exposures. Therefore, it is not unreasonable to surmise that as yet undetected GEIs might contribute to the problems of disease T that still frustrates association studies. Investment in genotyping technology must therefore be matched by equally robust investment in methods necessary to accurately characterize environmental exposures.

Pharmacogenetics/genomics offers the hope of predicting an individual's response to a pharmacologic intervention. However, for most drug–gene–outcome relationships, it remains undetermined what level of evidence will be needed to translate gene-based drug dosing into routine clinical practice. Factors influencing this process include frequency of the disease (e.g., GEI), variability in drug efficacy and frequency of any corresponding adverse drug reaction [65]. For some drugs, prospective gene-based T trials will be needed before the clinical and economic impact of such an approach is fully understood. For other drugs, the benefits of gene-based dosing may only be fully understood within the context of large observational studies conducted using practice-based cohorts [35]. Drug–gene–outcome relationships strongly influenced by GEI may best be characterized through the combined analyses of genetic material and secure encrypted electronic medical records contained within the world's growing biobanks.

Acknowledgments

The authors thank A Benor for his help in drawing Figure 4.

Bibliography

Papers of special note have been highlighted as:

- of interest
- of considerable interest

1. Garrod AE. The incidence of alkatonuria: a study in chemical individuality. *Lancet*. 1902; 160:1616–1620.
2. Turesson G. The genotypical response of the plant species to the habitat. *Hereditas*. 1922; 3:211–350.
3. Wright S. The roles of nutrition, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth Annual Congress of Genetics*. 1932:356–366.
4. Gluckman PD, Hanson MA, Beedle AS. Non-genomic transgenerational inheritance of disease risk. *Bioessays*. 2007; 29:145–154. [PubMed: 17226802]
5. Scandalios JG. Response of plant antioxidant defense genes to environmental stress. *Adv Genet*. 1990; 28:1–41. [PubMed: 2239448]
6. Laurie DA, Bennett MD. Nuclear DNA content in the genera *Zea* and *Sorghum*: intergeneric, interspecific and intraspecific variation. *Heredity*. 1985; 55:307–313.
7. Rayburn AL, Price HJ, Smith JD, Gold JR. C-band heterochromatin and DNA content in *Zea mays*. *Am J Bot*. 1985; 72:1610–1617.
8. Rayburn AL, Auger JA. Nuclear DNA content variation in the ancient indigenous races of Mexican maize. *Acta Botanica Neerlandica*. 1990; 39:197–202.
9. Yang W, Kelly T, He J. Genetic epidemiology of obesity. *Epidemiol Rev*. 2007; 29:49–61. [PubMed: 17566051]
10. Allard RW, Bradshaw AD. Implications of genotype–environment interactions in applied plant breeding. *Crop Sci*. 1964; 4:503–508.
11. Allard, RW. John Wiley & Sons. *Principles Of Plant Breeding*. 2. Wiley-Blackwell; NY, USA: 1999.
12. Becker HC. Correlations among some statistical measures of phenotypic stability. *Euphytica*. 1981; 30:835–840.
13. Haldane JBS. The interaction of nature and nurture. *Ann Eugen*. 1946; 13:197–205. [PubMed: 20282564]
14. Via S. The quantitative genetics of polyphagy in an insect herbivore. II. Genetic correlations in larval performance within and across host plants. *Evolution*. 1984; 38:896–905.
15. Baye TM, Wilke RA, Olivier M. Genomic and geographic distribution of private SNPs and pathways in human populations. *Per Med*. 2009; 6:623–641. [PubMed: 20352079]
16. Schnell FW. A study of methods and categories of plant breeding. *Zeitschrift fuer Pflanzenzuechtung*. 1982; 89:1–18.
17. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst*. 2000; 92:1151–1158. [PubMed: 10904088]
18. Campbell CD, Ogburn EL, Lunetta KL, et al. Demonstrating stratification in a European American population. *Nat Genet*. 2005; 37:868–872. [PubMed: 16041375]
19. Pritchard JK, Donnelly P. Case–control studies of association in structured or admixed populations. *Theor Popul Biol*. 2001; 60:227–237. [PubMed: 11855957]
20. Simmonds NW. Genotype (G), environment (E), and GE components of crop yields. *Expl Agric*. 1981; 17:355–362.

21. Ziegler, A.; König, IR. Statistical Approach to Genetic Epidemiology: Concepts and Applications. 2. Wiley-VCH; Germany: 2010.
22. Elston, RC.; Johnson, WD. Basic Biostatistics For Geneticists And Epidemiologists: A Practical Approach. John Wiley & Sons; Hoboken, NJ, USA: 2008.
23. Vercelli D. Discovering susceptibility genes for asthma and allergy. *Nat Rev Immunol*. 2008; 8:169–182. [PubMed: 18301422]
24. Lin PI, Vance JM, Pericak-Vance MA, Martin ER. No gene is an island: the flip-flop phenomenon. *Am J Hum Genet*. 2007; 80:531–538. [PubMed: 17273975]
25. Hernandez, LM.; Blazer, DG., editors. Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. National Academies Press; Washington, DC, USA: 2006.
26. Pocock, SJ. Clinical Trials: A Practical Approach. John Wiley & Sons; Hoboken, NJ, USA: 2004.
27. Lilienfeld, DE.; Stolley, PD. Foundations of Epidemiology. 3. Oxford University Press; Oxford, UK: 1994.
28. Andrieu N, Goldstein AM. The case-combined-control design was efficient in detecting gene–environment interactions. *J Clin Epidemiol*. 2004; 57:662–671. [PubMed: 15358394]
29. Moffitt TE, Caspi A, Rutter M. Strategy for investigating interactions between measured genes and measured environments. *Arch Gen Psychiatry*. 2005; 62:473–481. [PubMed: 15867100]
30. Hemminki K, Lorenzo Bermejo J, Forsti A. The balance between heritable and environmental aetiology of human disease. *Nat Rev Genet*. 2006; 7:958–965. [PubMed: 17139327]
31. Dempfle A, Scherag A, Hein R, et al. Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. *Eur J Hum Genet*. 2008; 16:1164–1172. [PubMed: 18523454]
32. Taubes G. Epidemiology faces its limits. *Science*. 1995; 269:164–169. [PubMed: 7618077]
33. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. *Stat Med*. 2007; 26:2–19. [PubMed: 17136746]
34. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996; 273:1516–1517. [PubMed: 8801636]
35. McCarty CA, Wilke RA. Biobanks and pharmacogenomics. *Pharmacogenomics*. 2010; 11(5):637–41. [PubMed: 20415552]
36. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods, and recruitment for a large population-based biobank. *Per Med*. 2005; 2:49–79.
37. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype–phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet*. 2010; 86:560–572. Demonstrates that biobanks linked to routine clinical practice-based data can be used to characterize genetic associations previously identified in disease-based cohorts. [PubMed: 20362271]
38. Baye TM, Wilke RA. Mapping genes that predict treatment outcome in admixed populations. *Pharmacogenomics J*. 2010; 10(6):465–477. Features the utilization of ancestry information in quantifying genetic determinants and interactions of treatment in an admixed population. [PubMed: 20921971]
39. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*. 2007; 39:870–874. [PubMed: 17529973]
40. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for Type II diabetes. *Nature*. 2007; 445:881–885. [PubMed: 17293876]
41. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet*. 2007; 39:645–649. [PubMed: 17401363]
42. Daly AK, Donaldson PT, Bhatnagar P, et al. *HLA-B*5701* genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat Genet*. 2009; 41:816–819. [PubMed: 19483685]
43. Barber MJ, Mangravite LM, Hyde CL, et al. Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS ONE*. 2010; 5:e9763. [PubMed: 20339536]

44. Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am J Hum Genet.* 2009; 85:309–320. [PubMed: 19733727]
45. Willer CJ, Speliotes EK, Loos RJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet.* 2009; 41:25–34. [PubMed: 19079261]
46. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet.* 2008; 17:R156–R165. [PubMed: 18852205]
47. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am J Hum Genet.* 2010; 86:6–22. [PubMed: 20074509]
48. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol.* 2010; 8:e1000294. [PubMed: 20126254]
49. Robinson R. Common disease, multiple rare (and distant) variants. *PLoS Biol.* 2010; 8:e1000293. [PubMed: 20126253]
50. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* 2010; 11:499–511. [PubMed: 20517342]
51. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.* 2007; 39:906–913. [PubMed: 17572673]
52. McCarty CA, Wilke RA. Biobanking and pharmacogenomics. *Pharmacogenomics.* 2010; 11:637–641. Features the utilization of clinical practice-based datasets for quantifying genetic determinants of treatment outcome specifically within the context of relevant covariates. [PubMed: 20415552]
53. Schatz M, Clark S, Camargo CA Jr. Sex differences in the presentation and course of asthma hospitalizations. *Chest.* 2006; 129:50–55. [PubMed: 16424412]
54. Eder W, Ege MJ, von Mutius E. The asthma epidemic. *N Engl J Med.* 2006; 355:2226–2235. [PubMed: 17124020]
55. Weiss ST, Raby BA, Rogers A. Asthma genetics and genomics 2009. *Curr Opin Genet Dev.* 2009; 19:279–282. [PubMed: 19481925]
56. Snyder EM, Beck KC, Dietz NM, et al. Influence of β 2-adrenergic receptor genotype on airway function during exercise in healthy adults. *Chest.* 2006; 129:762–770. [PubMed: 16537879]
57. Drazen JM, Silverman EK, Lee TH. Heterogeneity of therapeutic responses in asthma. *Br Med Bull.* 2000; 56:1054–1070. [PubMed: 11359637]
58. Choudhry S, Ung N, Avila PC, et al. Pharmacogenetic differences in response to albuterol between Puerto Ricans and Mexicans with asthma. *Am J Respir Crit Care Med.* 2005; 171:563–570. [PubMed: 15557128]
59. Litonjua AA, Silverman EK, Tantisira KG, et al. β 2-adrenergic receptor polymorphisms and haplotypes are associated with airways hyperresponsiveness among nonsmoking men. *Chest.* 2004; 126:66–74. [PubMed: 15249444]
60. Ordovas JM, Mooser V. Nutrigenomics and nutrigenetics. *Curr Opin Lipidol.* 2004; 15:101–108. [PubMed: 15017352]
61. Brennan P. Gene–environment interaction and aetiology of cancer: what does it mean and how can we measure it? *Carcinogenesis.* 2002; 23:381–387. [PubMed: 11895852]
62. Gardiner SJ, Begg EJ. Pharmacogenetic testing for drug metabolizing enzymes: is it happening in practice? *Pharmacogenet Genomics.* 2005; 15:365–369. [PubMed: 15864139]
63. Zhang G, Khoo SK, Laatikainen T, et al. Opposite gene by environment interactions in Karelia for *CD14* and *CC16* single nucleotide polymorphisms and allergy. *Allergy.* 2009; 64:1333–1341. [PubMed: 19222419]
64. Bottema RW, Reijmerink NE, et al. Interleukin 13, CD14, pet and tobacco smoke influence atopy in three Dutch cohorts: the allergenic study. *Eur Respir J.* 2008; 32:593–602. [PubMed: 18417506]
65. Wilke RA, Lin DW, Roden DM, et al. Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. *Nat Rev Drug Discov.* 2007; 6:904–916. [PubMed: 17971785]

Websites

101. The United Nations Millennium Ecosystem. www.millenniumassessment.org
102. The eMERGE electronic Medical Records and Genomics. 2007. www.gwas.net
103. Public Population Project in Genomics (P³G). www.p3g.org/secretariat

Executive summary

- Gene-environmental interactions play an important role in human disease and have been relatively well studied in model organisms. Rigorous quantitative assessment of environmental influences will be necessary to elucidate gene-environment interaction in humans.
- Longitudinal data available in practice-based (e.g., longitudinal cohorts followed within chronic disease management clinics) datasets will position investigators to characterize genetic factors with small but reproducible effects on drug outcome in the context of genotype-environment interactions.
- Biobanks have become increasingly common globally, and there is an urgent need for networking and sharing samples and tools across biobanks. The electronic Medical Records and Genomics (eMERGE) network represents a novel opportunity to coordinate these investigative efforts across multiple institutions to dissect genotype-environment interactions.
- The study of interactions may allow us to identify risk factors for disease that would not be found if only main effects of exposures were examined. Studying interactions may also be essential to identify novel genetic risk variants if genotype has a strong effect when environment is also present, but has no effect if environment is absent. Such relationships are more plausible for studies of adverse treatment-genotype effects in pharmacogenetic studies.
- Interaction, especially where qualitative interactions occur, may lead to an increased understanding of biological mechanisms underlying disease etiology. Such interactions may allow for specific targeting of interventions in high-risk groups. In fact, where the relationship between two risk factors is greater than additive (i.e., multiplicative), the largest reduction in absolute risk of disease is achieved from interventions.
- Modeling genotype-environment and genotype-treatment systems could potentially help to reduce the noise in genomic research and help to quantify genotype-treatment-environment-phenotype relationships.

Figure 1. Graphical representation of the ‘no’ interaction, noncrossover interaction and crossover interaction types of genotype–environment interactions

For details of A–F, see text.

Adapted from [38].

Figure 2. Phenotype is under the control of genetic and environmental effects
 Epigenetics links environment and genotype to phenotype and disease.

Figure 3. Proposed approaches to study genotype–environment interactions’ stability using biological material linked to comprehensive electronic medical databases

GEI: Genotype–environment interactions.

Figure 4. Genotype-environment and genotype-treatment interplay: a model looking at genotype versus environment/treatment interplay to health effects.

Table 1

Analysis of variance taking genotype, treatment and environment as random into the analytical model.

Source	d.f.	m.s.	E(m.s.)
Genotype	$f1 = g-1$	m1	$\sigma^2 + r\sigma_{gte}^2 + er\sigma_{gt}^2 + rt\sigma_{ge}^2 + rte\sigma_g^2$
Treatment	$f2 = t-1$	m2	$\sigma^2 + r\sigma_{tge}^2 + er\sigma_{gt}^2 + rg\sigma_{te}^2 + reg\sigma_t^2$
Environment	$f3 = e-1$	m3	$\sigma^2 + r\sigma_{gte}^2 + rt\sigma_{ge}^2 + rg\sigma_{te}^2 + rtg\sigma_e^2$
G-T	$f4 = f1f2$	m4	$\sigma^2 + \sigma_{gte}^2 + er\sigma_{gt}^2$
G-E	$f5 = f1f3$	m5	$\sigma^2 + \sigma_{gte}^2 + rt\sigma_{ge}^2$
E-T	$f6 = f2f3$	m6	$\sigma^2 + \sigma_{gte}^2 + rg\sigma_{te}^2$
G-E-T	$f7 = f1f2f3$	m7	$\sigma^2 + \sigma_{gte}^2$
Error	$F8 = (r-1)(g-1)te$	m8	σ^2

This model estimates genetic variances including interactions and covariances.

e, g, t and r are the levels for environment, genotype, treatment and replication variables.

d.f.: Degree of freedom; E(m.s.): Expected mean square; E-T: Environment-treatment;

G-E: Genotype-environment; G-T: Genotype-treatment;

G-E-T: Genotype-environment-treatment; m.s.: Mean square.

Table 2

Analysis of genotype–environment interaction in biobanks.

Issue	Challenge
Biobank design	Heterogeneity in selection criteria – population-based, practice-based, disease-based or treatment-based
Enrollment strategy	Opt-in (informed consent) versus opt-out at point of care
Data management	Biocomputing and infrastructure
Data storage	Variability in format – structured versus unstructured
Phenotyping	Algorithms (codes, meds or labs) not yet standardized
Environment quantification	Lack of specific measurements in routine EMR and standard methods to evaluate the environmental risk factor
Multivariate phenotypes associated with complex disease	Genetic predisposition, lifestyle and diet, exposure to unmeasured environmental factors
Timing of collection for plasma and serum	Temporal link to event of interest not predetermined
Different samples such as liver, muscle, subcutaneous and visceral adipose tissue	Absence of detailed disease phenotype characterization in a highly specified sample
‘Omics data (genomics, transcriptomics, proteomics and metabolomics)	Lack of data integration approaches
Transregional and national collaborations	Heterogeneous ethical and legal framework – different structure, rules and standards
Core samples	Absence of approaches to define core sample
Statistical test	Millions of samples – false-positive or -negative results

EMR: Electronic medical records.