Primer in Genetics and Genomics, Article 5—Further Defining the Concepts of Genotype and Phenotype and Exploring Genotype–Phenotype Associations

Fay Wright, PhD, RN1, and Kristen Fessele, PhD, RN2,3

Abstract
As nurses begin to incorporate genetic and genomic sciences into clinical practice, education, and research, it is essential that they have a working knowledge of the terms foundational to the science. The first article in this primer series provided brief definitions of the basic terms (e.g., genetics and genomics) and introduced the concept of phenotype during the discussion of Mendelian inheritance. These terms, however, are inconsistently used in publications and conversations, and the linkage between genotype and phenotype requires clarification. The goal of this fifth article in the series is to elucidate these terms, provide an overview of the research methods used to determine genotype–phenotype associations, and discuss their significance to nursing through examples from the current nursing literature.

Keywords
genotype, genomics, phenotype, symptom science, nursing research

As nurses begin to incorporate genetic and genomic sciences into clinical practice, education, and research, it is essential that they have a working knowledge of the terms foundational to the science. The first article in this primer series provided brief definitions of the basic terms (i.e., genetics and genomics) and introduced the concept of phenotype during the discussion of Mendelian inheritance. These terms, however, are inconsistently used in publications and conversations, and the linkage between genotype and phenotype requires clarification. The goal of this fifth article in the series is to further elucidate these terms, discuss the significance of genotype–phenotype associations, describe the process of determining these associations, and provide examples of studies that have explored potential genotype–phenotype associations to inform nursing practice, education, and research. (For a more detailed description of nucleotides, see Dorman, Schmella, & Wesmiller, 2016.) We have come to understand that a person’s DNA comprises a fixed, measurable sequence of molecules that can be observed at multiple levels, ranging from a single base up through genes and to the unit encompassing the entire available complement of information, that is, the genome. Genes are generally defined as segmented sequences of nucleotides along the DNA strand that contain the code that, when transcribed and translated, results in the production of functional proteins (Stotz, Griffiths, & Knight, 2004).

The term genome refers to the total DNA sequence that an organism possesses; in humans, this includes approximately 3 billion nucleotides. Only about 1% of this complex structure corresponds to protein-coding exons within genes, while an additional 0.5% comprises accompanying regulatory, but

Defining Gene, Genome, Genotype, Phenotype, and Related Terms
The terms gene, genome, and genotype describe closely related concepts that are primarily focused on describing the sequence of DNA within an organism’s cells. Our understanding of these terms has expanded exponentially since the successful completion of the Human Genome Project in 2003. Nucleotides, also referred to as bases, make up the structure and sequence of DNA. Each nucleotide is composed of a nitrogenous base, a five-carbon sugar molecule, and a phosphate. The sequence of the bases (e.g., adenine, cytosine, guanine, thymine) differentiates strands of DNA. (For a more detailed description of nucleotides, see Dorman, Schmella, & Wesmiller, 2016.) We have come to understand that a person’s DNA comprises a fixed, measurable sequence of molecules that can be observed at multiple levels, ranging from a single base up through genes and to the unit encompassing the entire available complement of information, that is, the genome. Genes are generally defined as segmented sequences of nucleotides along the DNA strand that contain the code that, when transcribed and translated, results in the production of functional proteins (Stotz, Griffiths, & Knight, 2004).

The term genome refers to the total DNA sequence that an organism possesses; in humans, this includes approximately 3 billion nucleotides. Only about 1% of this complex structure corresponds to protein-coding exons within genes, while an additional 0.5% comprises accompanying regulatory, but

1 Yale School of Nursing, Orange CT, USA
2 Scientific Project Leader, Flatiron Health, New York, NY, USA
3 Post-doctoral Research Fellow, University of Utah College of Nursing, Salt Lake City, UT, USA

Corresponding Author:
Fay Wright, PhD, RN, Postdoctoral Research Fellow, Yale School of Nursing, 400 West Campus Drive, Orange CT, 06516-0972, USA.
Email: fay.wright@yale.edu

Series Information:
The “Primer in Genetics and Genomics” series is a collaboration between Biological Research for Nursing and the International Society of Nurses in Genetics (ISONG). Sheila A. Alexander, PhD, RN, FCCM, serves as guest editor for the series.
noncoding, sequence necessary for gene expression (Marian, 2012). Because nurses and other researchers and clinicians are most frequently interested in the phenomena associated with gene expression, such as the risk for development of disease or drug metabolism, genetic testing is generally targeted toward this small percentage of the overall genome, assessing the inherited sequences at specific locations within protein-coding segments of genes.

There is tremendous genetic diversity among humans, though our genomes are 99.9% similar. In a recent review of the National Center for Biotechnology Information (NCBI) resource Single-Nucleotide Polymorphism Database (dbSNP, build 149, November 9, 2016), Marian (2012) identified over 34 million variants, also referred to as polymorphisms, among the genomes of humans included in the database. The term genotype is defined most simply as “the version of a DNA sequence that a person has” (National Institutes of Health. National Human Genome Research Institute, n.d.). Wilhelm Johansen coined the term genotype in 1909, prior to the discovery of DNA, in response to his observations of physical traits passed down during generations of plant breeding and building on Mendel’s original work (Johansen, 1909). In current usage, genotype typically describes the sequence of genetic material present in a particular location of interest within the overall DNA sequence (e.g., “What is that person’s genotype at DNA location XYZ?”).

Genotype determines an individual’s phenotype. A phenotype is an individual’s observable traits, such as height, eye color, and blood type, which are the outward expression of the genotype (National Institutes of Health [NIH], n.d.). Historically, the signs and symptoms collected on exam by clinicians as patients presented for care formed the basis for what could be considered classical phenotyping—using direct, predominantly visual observations about variations in an individual’s anatomy, physiology, and behavior to place them in a possible diagnostic group and guide further confirmatory workup.

When discussing genotype–phenotype associations, it is important to understand a number of other terms related to genotypes. Locus refers to the expected location of a gene on either of the two chromosome copies (i.e., whether maternally or paternally contributed), whereas allele refers to the variant of the gene sequence present at either chromosome locus (Figure 1). Chromosomes are structurally identified with “long” arms (i.e., q arms) and “short” arms (i.e., p arms). The locus includes the location of the gene on either arm of the chromosome. For example, the PAH gene, responsible for the production of the enzyme phenylalanine hydroxylase, is located on the long arm (i.e., q arm) of chromosome 12, specifically at locus 12q23.2. A mutation within this gene can reduce the function of the enzyme, resulting in the metabolic disease phenylketonuria (PKU). However, as PKU is an autosomal recessive disorder, clinical manifestation of the disease requires that both PAH gene alleles residing at the 12q23.2 locus, that is, the alleles on both the maternally and paternally contributed copies, must be mutated (National Library of Medicine, 2016).

**Figure 1.** Locus, alleles, and genotype example. In this hypothetical chromosome pair, the colored bands indicate the specific locus. The letters labeling these bands (A and T) indicate different nucleotides present in the alleles at that locus, illustrating an “AT” genotype. A = adenine; T = thymine.

**Evolution of the Phenotype Concept**

Though the definition of genomics has evolved tremendously over the past 100 years, the classical “appearance-type” definition of phenotype that Johansen developed in 1909 remains essentially relevant. In the biological and medical literature, authors often apply the term phenotype as a simple diagnostic label (e.g., patients with elevated blood glucose or hypertension). More recently, however, Robinson (2012) defined phenotype as “the collection of observable traits of an organism, comprising its morphology, its physiology at the level of the cell, the organ and the body, and its behavior, comprising even characteristics such as the gene expression profiles in response to environmental cues” (p. 777). As understanding of the complexity of diseases grows, including the underlying biologic and environmental factors that contribute to the risk of developing and severity of these conditions, it is apparent that more robust phenotypic descriptions, such as those that Robinson described, would facilitate comparison across studies.

The phrase “deep phenotyping,” which increasingly appears in the literature, involves “complete descriptions of the physical state of individuals (and, by aggregation, groups of individuals) using the concurrent collection of clinical signs, symptoms and multiple ‘omic’ methods to detect underlying biologic activity particular to the individual at a specific point in time, such as genotyping, mRNA transcription, protein production and cellular metabolism” (Tracy, 2008, p. 151). This more complex, highly granular level of description fits with the goals of the precision medicine movement, where careful classification of patients into subpopulations with a common underlying biological basis for their disease and/or response to treatments facilitates the selection of the most effective therapies (Robinson, 2012). Evaluation and replication of these granular descriptions across populations to confirm the most precise and refined phenotypes associated with risk of developing disease or negative outcomes associated with proposed
therapies may enable better proactive risk stratification and shared decision-making.

**How Does a Genotype Affect a Phenotype?**

As technology advances and DNA sequencing methods become more accessible and affordable, the addition of diagnostic and pretreatment genotype testing to other traditional laboratory methods to enable formation of a more robust phenotypic description is becoming more commonplace. We now know that variation in inherited genes is a major driver of differences in observed phenotypes. For example, as we described above, a mutation in the *PAH* gene changes the genotype and results in the PKU phenotype.

Significant differences in health may result from genotypic variations as minor as a single substitution of a nucleotide base in the DNA sequence called a single-nucleotide polymorphism (SNP). For example, hemoglobin A (HgbA) represents the normal form of the protein produced by the hemoglobin-β (*HBB*) gene. When a person inherits a variation in *HBB* such that the nucleotide A is replaced by T at Position 6 (an SNP), it changes the amino acid produced during RNA translation from glutamic acid to valine, resulting in the mutated hemoglobin S (HgbS) form associated with sickle cell disease (Saraf et al., 2013). Depending on the combination of alleles inherited from each parent, the possible genotypes a person might possess include HgbA/A, HgbA/S, or HgbS/S; the disease phenotypes associated with these genotypes would likely show a person unaffected by the disease, a carrier of sickle cell trait, and a patient with sickle cell disease, respectively (Yawn et al., 2014).

**Utilization of Genotype–Phenotype Associations in Clinical Practice and Research**

**Inherited Risk of Disease**

Whether involving an SNP or a mutation involving a larger sequence of DNA, variations in genotype may be associated with inherited risk of disease. (NIH, 2013) Readers should consult Aiello and Chiatti’s (2017) discussion in the fourth paper in this primer series for an excellent overview of the topic of inherited risk. Early work to define gene–disease associations focused on “simple” Mendelian models in which a single gene mutation resulted in high penetrance of the disease phenotype in a family, such as with cystic fibrosis or Huntington’s disease. An important next step in defining these associations, already underway, involves deeper insight into more complex, polygenic diseases, especially those with a heavy environmental exposure component (Botstein & Risch, 2003). Diabetes provides an interesting exemplar that potentially bridges both models. Though research has shown both type 1 and type 2 diabetes to be associated with multiple genetic variants, recent work has uncovered a subgroup of patients (i.e., those with monogenic diabetes of the young and neonatal diabetes mellitus, comprising approximately 2–5% of cases worldwide) that have a monogenic form of the disease where mutation of genes affects nonautoimmune function of the pancreatic β cells (National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, 2014; Thomas & Philipson, 2015).

Since the 1950s, an increasing number of genetic tests have become available to predict the risk of developing heritable diseases or to confirm a diagnosis (NIH, 2013). One example, which serves to illustrate the use of a genetic test to predict risk based on a genotype–phenotype connection, involves the *BRCA1* and *BRCA2* genes. The wild-type forms of these two genes serve a protective function, coding for a tumor suppressor. Mutations in one or both of these genes are common in certain populations. For example, researchers first identified the genotypes *BRCA1* 187delAG (where an adenine and guanine nucleotide are deleted from the DNA sequence), *BRCA1* 5385insC (where a cytosine nucleotide is inserted into the sequence), and *BRCA2* 6147delT (where a thymine nucleotide is deleted) in Ashkenazi Jews (Meric-Bernstam et al., 2013). Phenotypically, individuals with these genotypes, especially those involving *BRCA1* mutations, are at an increased risk of being diagnosed with breast cancer at a younger age than those with a wild-type genotype, and their tumors are more likely to be estrogen, progesterone, and HER2 negative, which confers challenges in treatment selection, as hormonal blockade and biological therapy with trastuzumab are not indicated in these “triple negative” cases (Tung et al., 2014).

**Symptom Science**

Beyond disease susceptibility, genotype identification can also help to elucidate symptom phenotypes. Nurse scientists are working to describe genotype–phenotype associations that may help to predict which patients may experience more severe symptoms along a disease trajectory or after certain therapies. For example, a growing body of evidence links variations in cytokine genes, including *IL-1, IL-10*, and others, with a higher incidence and severity of fatigue in patients with breast cancer. Researchers recently found that in patients followed for 6 months postsurgery for breast cancer, those with the *IL-1B* SNP rs16944 (the number following “rs” is the reference sequence number identifying the specific SNP) with two A alleles (genotype AA) had a 2.98-fold higher risk of belonging to the “higher fatigue” class in their analysis. The phenotypic characteristics of this class included younger age, lower Karnofsky Performance Status score, higher self-administered Comorbidity Questionnaire score, higher fatigue score at study enrollment, and higher number of lymph nodes removed (Kober, Smoot, et al., 2016).

**Pharmacogenetics**

An increasingly widespread application of genotype–phenotype associations in clinical practice relates to pharmacogenetics. Pharmacogenetics combines knowledge of drug pharmacology and genomics to tailor dosing and monitoring to a patient’s specific genetic makeup (NIH, 2017). Research
has identified specific genotypes associated with phenotypes of differential drug metabolism. Pretreatment testing for genetic variations that might impact drug metabolism is becoming more common, and efforts are underway by organizations such as the Clinical Pharmacogenetics Implementation Consortium to provide guidance to clinicians to support evidence-based implementation of pharmacogenetically indicated dose adjustments (Relling & Klein, 2011).

For example, codeine, a prodrug of morphine, is metabolized by the hepatic cytochrome P450 isoenzyme 2D6 (CYP2D6), which is involved in the metabolism of about one fourth of all medications. The rate at which this metabolism occurs can vary significantly based on an individual’s genotype for the CYP2D6 gene. The wild type, or most typically occurring genotype, is the CYP2D6*1/*1 form, indicating that the individual possesses one copy of the gene per chromosome, with the *1 allele found in each of the maternal and paternally contributed chromosomes at the expected locus. Along with this form, the *2, *33, and *35 alleles are associated with generally normal rates of metabolic activity when a person’s genome contains only one CYP2D6 gene copy per chromosome (Dean, 2012). As with many other genes associated with metabolism, however, changes in the inherited genotype or in the number of CYP gene copies per chromosome within a person’s genome can alter function. Extra copies (also known as a copy number variation) of even the wild-type *1 or *2 alleles in the CYP2D6 gene per chromosome can result in higher than normal production of the isoenzyme, resulting in what is termed ultrarapid metabolism (Gaedigk et al., 2008). In the case of codeine, patients who are ultrarapid metabolizers are at high risk of inadvertent opiate overdose, as the codeine prodrug is converted more quickly into morphine, resulting in a rapid rise in serum drug levels and potential deleterious effects such as depressed respiration, somnolence, and confusion (Roxane Laboratories, 2014). In contrast, patients with the *4, *5, or *6 alleles may have low- to non-functioning CYP2D6 genes, resulting in little to no conversion of codeine to morphine. These patients, called poor metabolizers, may experience suboptimal pain relief despite what appears to be adequate drug administration (Dean, 2016).

It is important for nurses to be able to understand the basic implications of pharmacogenomic results. For example, patients who receive a genotype result for CYP2D6 indicating that they are poor metabolizers for codeine are also at risk of negative drug outcomes with the tricyclic and selective serotonin reuptake inhibitor (SSRI) antidepressant drug classes, among others that are also metabolized via this hepatic isoenzyme (Whirl-Carrillo et al., 2012).

**Determination of Genotype–Phenotype Associations**

An up-to-date understanding of the terms *genotype* and *phenotype* is necessary to make sense of the translational literature in which researchers are seeking associations between the two related to a particular disease, symptom, or treatment response. While current technology is making the determination of a genotype a more straightforward process of measurement and reporting, the process of reporting accurate and robust phenotypic descriptions in clinical and research settings must continue to evolve to facilitate focused study of patient cohorts most likely to have true underlying biologic similarity. Investigators may use many different paths to identify genotype–phenotype associations. Figure 2 illustrates one path for cases in which there are no known associations.

**Figure 2.** Example path for the discovery of genotype–phenotype associations when there are no currently known genotypic associations for a particular phenotype.
Table 1. Useful Databases for the Exploration of Relationships Among Genetic Variations and Specific Phenotypes.

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
<th>Use in Research</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALlele FREquency Database (ALFRED)</td>
<td>Database of allele frequency data in human populations</td>
<td>Determine allele frequencies of STRPs, VNTRs, SNPs, INDELs, and RFLPs</td>
<td><a href="http://alfred.med.yale.edu">http://alfred.med.yale.edu</a></td>
</tr>
<tr>
<td>dbSNP</td>
<td>Database of single nucleotide polymorphisms within and across different species</td>
<td>Determine if the short genetic variations SNPs, small-scale INDELs, and STRPs have documented associations</td>
<td><a href="https://www.ncbi.nlm.nih.gov/snp">https://www.ncbi.nlm.nih.gov/snp</a></td>
</tr>
<tr>
<td>Human Phenotype Ontology (HPO)</td>
<td>Database that provides a standardized vocabulary of phenotypic abnormalities encountered in human disease</td>
<td>Search for an identified phenotype’s association with diseases and genotypes</td>
<td><a href="http://human-phenotype-ontology.github.io/">http://human-phenotype-ontology.github.io/</a></td>
</tr>
<tr>
<td>GeneCards</td>
<td>Cataloged information about all annotated and predicted human genes</td>
<td>Explore a specific gene’s function, expression, proteomics, clinical, and functional information</td>
<td><a href="http://www.genecards.org">http://www.genecards.org</a></td>
</tr>
<tr>
<td>Online Mendelian Inheritance in Man (OMIM)</td>
<td>Compendium of human genes and genetic phenotypes with a focus on the relationship between phenotype and genotype</td>
<td>Search for known associations between phenotypes and genotypes</td>
<td><a href="http://omim.org">http://omim.org</a></td>
</tr>
<tr>
<td>PhenX tool kit</td>
<td>Catalog of standard measures of phenotypes and environmental exposures</td>
<td>Identify validated measures relevant to phenotype of interest and related environmental exposures</td>
<td><a href="https://www.phenxtoolkit.org">https://www.phenxtoolkit.org</a></td>
</tr>
</tbody>
</table>

Note. INDELs = insertion/deletion polymorphisms; RFLPs = restriction fragment length polymorphisms; SNPs = single-nucleotide polymorphisms; STRPs = short tandem repeat polymorphisms; VNTRs = variable number tandem repeat polymorphisms; ALFRED = ALlele FREquency Database; dbSNP = Single-Nucleotide Polymorphism Database; HPO = Human Phenotype Ontology; OMIM = Online Mendelian Inheritance in Man.

Genome Epidemiology literature review of studies examining the association between HLA-DQ polymorphisms and type 1 diabetes found that multiple reports included type 2 diabetics who used insulin (Dorman & Bunker, 2000). The authors concluded that, to clearly identify the genetic underpinnings of type 1 diabetes, the phenotypic inclusion criteria must center not only on insulin-dependence criteria but more critically on the specific diabetes diagnosis.

The first step on the path to discovering a genotype–phenotype association, then, is to define the phenotype. Phenotype definition can comprise demographic characteristics (e.g., age, sex, and education), clinical assessments (e.g., blood pressure measurements, height, and weight), clinical characteristics (e.g., disease diagnosis, laboratory values, and diagnostic test results), symptom characteristics (e.g., self-reported fatigue, pain, or depression levels), a specific response to treatment (e.g., chemotherapy-induced symptoms and analgesia responsiveness), or a combination of these characteristics. Researchers may define the phenotype based on their research question and what is currently known about the condition they are studying. Once there is a well-defined phenotype of interest, it is possible to access multiple databases to determine if prior research has already produced evidence of a genotypic association (see Table 1 for descriptions of these databases). Some of the databases have overlapping information because they collate genomic information from common sources such as the NIH gene database or NCBI. For example, the Online Mendelian Inheritance in Man (OMIM) database allows a user to type in a phenotype in basic English terms (e.g., sickle cell) and returns a list of results (in the case of sickle cell, 16,598 entries) from which to choose the specific phenotype of interest (e.g., sickle cell anemia). Once the user chooses the specific phenotype, the OMIM database returns a description and clinical synopsis of the phenotype and any known genotypic relationships. The OMIM database is thus a great place to start to determine whether there are any known associations between the phenotype of interest and specific genotypes.

**Linkage Analysis**

Historically, when there was no known link between a phenotype of interest and specific genotypes, investigators turned to linkage analysis to discover possible associations (Ott, Wang, & Leal, 2015). Linkage analysis was primarily used to identify specific genes associated with disease phenotypes within a family lineage (e.g., Huntington disease). A family pedigree would be developed to identify affected and nonaffected family members. Investigators assumed that family members of an affected patient would inherit the same polymorphism and the disease gene together. Linkage analysis involved literally
hunting for a gene whose locus was close to that of the polymorphism associated with the phenotype on the same chromosome and then calculating a logarithm of the odds (LOD) score that represented the likelihood of a linkage between the polymorphism and the disease gene (Dueker & Pericak-Vance, 2014). An LOD score of ≥3 is traditionally an adequate indication of linkage. Linkage analysis studies were limited because the sample comprised only families with members affected with the specific condition. The linkage analysis, therefore, was only as strong as the phenotype definition for both the affected and unaffected family members.

With the advent of genome-wide scanning (GWS) technology, linkage analysis can now identify rare variants that have large effects on the heritability (i.e., the chances of inheriting an increased risk of a specific condition) of specific disease phenotypes (Ott et al., 2015). For example, while coronary artery disease (CAD) is associated with more than 50 common genetic variants, the heritability of CAD is <20% collectively for these variants (Girelli, Martinelli, Peyvandi, & Olivieri, 2009). In one study, a GWS linkage analysis of families with early-onset CAD identified two CAD loci with LOD scores ≥5.4 on chromosome 3, 3p25.1, and 3q29, which is highly significant evidence of linkage and can be used to guide further identification of specific genotypes associated with the early-onset CAD phenotype (Gao et al., 2014).

**Genome-Wide Association Studies**

The approach most widely used currently to find associations between specific common genetic variations with a specific condition/phenotype when there are no known genotype–phenotype associations is the genome-wide association study (GWAS). In a GWAS study, investigators define a phenotype and then scan the genomes of many different people, with and without the phenotype, to look for genetic polymorphisms that could be used to predict the presence of the phenotype. By comparing the genome of people with and without the phenotype, researchers can identify differences between the two groups. Once genotypic differences (i.e., polymorphisms) are identified, they can be used to understand how genes contribute to the disease and to develop better prevention and treatment strategies.

To conduct a GWAS, investigators use microarray technology, which can analyze hundreds of thousands of SNPs to identify differences between the groups. Microarray technology involves placing thousands of reference gene sequences in known locations on a glass slide called a gene chip. A sample that contains DNA from the study participants is placed in contact with the gene chip. Complementary base pairing between the participant’s sample and the reference gene sequences on the chip produces light that is measured to determine if there are gene sequences that match. The data are visualized on a plot (commonly called a Manhattan plot) according to the participant sample DNA and the reference sequence associations (i.e., matches) that meet a predetermined level of significance (see Figure 3).

The goal of a GWAS is to identify functional variations (i.e., SNPs) associated with the phenotype of interest to provide a valid path for future research based on the biological plausibility that a specific gene or multiple genes are associated with the phenotype (Ngeow & Eng, 2015). The GWAS identifies the loci of candidate genes, and future studies can confirm or refute these associations. For example, to explore genetic associations for the type 1 diabetes (T1D) phenotype, researchers performed a GWAS and combined these results in a meta-analysis involving two prior studies (Barrett et al., 2009). The complete

![Figure 3. Genome-wide association study (GWAS) results plot. This sample plot identifies the results of a GWAS analysis. Data resulting from microarray analyses are visualized on a plot with genomic position on the horizontal scale and the level of association between the reference genes and the participants’ samples on the vertical scale. The dotted horizontal reference line indicates the threshold level of statistical significance of the association. In this example, five loci met the significance threshold. Image was obtained from the Creative Commons and was downloaded from Wikimedia Commons on January 4, 2017 (https://commons.wikimedia.org/wiki/File: Manhattan_Plot.png). Original source for image: Ikram et al. (2010).](https://commons.wikimedia.org/wiki/File: Manhattan_Plot.png)
sample for the meta-analysis included the genomes of 7,514 patients with T1D and 9,045 people unaffected by T1D. T1D was phenotyped through standard clinical diagnosis. The GWAS identified 18 distinct loci of candidate genes associated with T1D. Other researchers then targeted the loci identified in this GWAS and others to determine the effect size of the genotype–phenotype analysis, demonstrating how a GWAS identifies initial genotypes for further study (Ge et al., 2016).

Candidate–Gene Association Studies

Once a genotype–phenotype association has been identified, researchers may use candidate–gene association studies to further explicate the association. For example, in one candidate–gene association study, investigators identified the associations between four genotypes determined by SNPs (i.e., rs8179526, rs10177833, rs6731545, and rs6726450) on the 4, sodium bicarbonate cotransporter member 5 gene (SLC4A5), and hypertension in African American women (Taylor, Maddox, & Wu, 2009). They chose the SLC4A5 gene a candidate for further analysis because prior GWS and linkage analysis studies had identified an association between the gene and increased systolic blood pressure (SBP). Phenotyping comprised measures of subjects’ physical activity, sodium levels, body mass index, and blood pressure. Investigators completed genotyping for the presence of the four candidate SNPs via the polymerase chain reaction (PCR) technique using the TaqMan Allelic Discrimination assay. (Note that PCR is a laboratory technique that makes multiple exact copies of DNA to use as a template for analysis. TaqMan is a method used to identify the presence of SNPs in a genetic sample.) Findings indicated that one SNP (rs10177833) was significantly related to SBP. Specifically, when cytosine (C) was replaced by adenine (A) on one genotype (C/A) or two alleles (AA), the participants had higher SBP (p = .030 and p = .046, respectively) compared to women with a wild-type (C/C) genotype. This study thus added to the body of evidence pointing to an association between SLC4A5 and a high-SBP phenotype.

In another candidate–gene study, researchers investigated SNPs on genes previously identified as having an association with inflammation (i.e., cytokines [Interleukin 1a (IL1a), IL4, IL6, IL8, IL10, and IL13], tumor necrosis factor-α, and VEGF-C and VEGF-D) as genotypes potentially associated with lymphedema. Investigators determined the lymphedema phenotype with a perometer measurement that identifies an increase of ≥5% in limb volume from the presurgery baseline assessment of the arm. While this phenotype was not associated with any of the tested genotypes, an SNP in IL4 was significantly associated with a revised phenotype further defined by the addition of impaired limb mobility and fluid accumulation (Fu et al., 2016).

Gene Expression Analysis

Another step in genotype–phenotype association analysis is to examine how the genotype effects gene expression. In brief, gene expression analysis evaluates the transcription of the gene to produce protein products. Researchers can use an extension of PCR methodology called reverse transcription PCR to determine how gene expression changes over time or under different conditions through the creation of complementary DNA transcripts from RNA (VanGuilnder, Vrana, & Freeman, 2008). Researchers use expression analysis to test the hypothesis that variation in the amount of the gene’s product plays a role in the phenotype of interest. As discussed in the “Symptom Science” section above, Kober, Smoot, and colleagues (2016) identified genotypes associated with a specific phenotype of fatigue in women after breast cancer surgery. These genotypes were associated with inflammatory processes (i.e., IL1B rs16944). However, researchers cannot determine the mechanism by which a genotype affects a phenotype, in this case inflammatory responses and thus the fatigue phenotype, using candidate–gene association studies. Rather, to determine the effect of the SNP on gene function, investigators use expression analysis. Thus, for example, investigators phenotyped women with breast cancer undergoing chemotherapy based on fatigue scores, physical function, and demographic and clinical characteristics to identify high- and low-fatigue groups (Kober, Dunn, et al., 2016). Using RNA extracted from participants’ whole-blood samples, the researchers conducted an analysis of gene expression, identifying differential expression by fatigue group in genes involved in inflammation as well as those involved in neurotransmitter regulation and energy metabolism. These expression results expand earlier findings regarding genotype–phenotype association and add to the store of knowledge about the multifactorial nature of fatigue.

Epigenetics

Epigenetic changes can regulate gene expression and are involved in the development of different phenotypes (e.g., T1D, CAD; Dorman et al., 2016). Epigenetics–gene interaction studies potentially elucidate variability in genotype–phenotype associations and provide insight into the underlying mechanisms of the phenotype. In the candidate–gene study of hypertension in African American women discussed above, the nurse researchers also examined the effects of interactions between SLC4A5 and sodium status and physical activity on blood pressure (BP) (Taylor et al., 2009). One of the results of this study’s gene interaction analysis serves as an illustration of how epigenetic-and-gene interaction studies elucidate genotype–phenotype associations. Researchers found that the SNP they identified in the candidate–gene association study as being significantly related to systolic BP (SBP) (i.e., SLC4A5 rs1017783) also interacted with physical activity to differentially affect SBP. Specifically, they noted that participants who carried the C/A and A/A genotypes of rs10177831 and reported engaging in <30 min of physical activity daily had increased SBP relative to those reporting <30 min of physical activity who carried the C/C genotype (Taylor et al., 2009). Diastolic BP (DBP) was also associated with interactions between SLC4A5 rs1017783 and physical activity, with participants with the A/A genotypes of rs10177831 who reported <30 min...
Density of the genotyping Low-density arrays may not offer adequate structure of LD Might not adequately capture information

Study design platform Prospective studies are free of potential confounding differences between cases and controls

Population admixture Increase the risk of spurious results

Phenotype Phenotypic admixture and phenocopy conditions dilute the power to detect an effect size

Study design platform Prospective studies are free of potential confounding differences between cases and controls

Population admixture Increase the risk of spurious results

Phenotype Phenotypic admixture and phenocopy conditions dilute the power to detect an effect size

Study design platform Prospective studies are free of potential confounding differences between cases and controls

Structure of LD Might not adequately capture information content of the true causative alleles

Density of the genotyping Low-density arrays may not offer adequate cover for the common haplotypes

Table 2. Study Design Issues to Be Considered in Studies of Complex Genetic Traits.

<table>
<thead>
<tr>
<th>Determinant</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>Direct correlation between the sample size and the power to detect the causative alleles (within a limit)</td>
</tr>
<tr>
<td>Effect sizes of the</td>
<td>Inverse correlation between effect size and power to detect an association</td>
</tr>
<tr>
<td>causative alleles</td>
<td></td>
</tr>
<tr>
<td>Minor allele frequency</td>
<td>GWAS by design detect only common alleles (MAF &gt; .05)</td>
</tr>
<tr>
<td>Proximity of the</td>
<td>More powerful for detecting an effect on proximal than distal phenotypes</td>
</tr>
<tr>
<td>phenotype to the genotype</td>
<td></td>
</tr>
<tr>
<td>Population characteristics</td>
<td>Presence of other competing factors dilute the power to detect an effect</td>
</tr>
<tr>
<td>Phenotype</td>
<td>Increase the risk of spurious results</td>
</tr>
<tr>
<td>Study design platform</td>
<td>Prospective studies are free of potential confounding differences between cases and controls</td>
</tr>
<tr>
<td>Structure of LD</td>
<td>Might not adequately capture information content of the true causative alleles</td>
</tr>
<tr>
<td>Density of the genotyping</td>
<td>Low-density arrays may not offer adequate cover for the common haplotypes</td>
</tr>
</tbody>
</table>

Note. GWAS = genome-wide association studies; LD = linkage disequilibrium; MAF = minor allele frequency. Reprinted from Marian (2012), with permission from Elsevier.

of physical activity having increased DBP relative to those with the C/C genotype. These results could inform nurses’ specific advice to patients regarding the use of exercise as a method to regulate BP, providing insight into the development of precision interventions based on genotype–phenotype associations.

Challenges and Promise of Research Exploring Genotype–Phenotype Associations

Underpinning the National Institute of Nursing Research Innovative Questions initiative is the discovery of genotype–phenotype associations to inform the development of precision interventions. Well-designed studies that identify and quantify genotype–phenotype associations through data mining and new investigations will provide opportunities for nurses to develop and test precision interventions to treat specific conditions or symptoms.

Genotype–phenotype associations, however, are complex. Many phenotypes have no known genotype associations, and for the known associations, research has yet to fully elucidate the effect of the genotype on phenotype variability. A genotype–phenotype association is just that a connection between a specific genotype and an individual’s phenotype. There may not be a definitive causality to the association and, in many cases, there may be additional factors, such as environmental exposures, that add variability. Marian (2012) identified key factors that determine the “robustness” of genetic association studies (see Table 2). While some of the factors (e.g., sample size and population characteristics) are challenges common to all forms of research design, other factors (e.g., proximity of the phenotype to the genotype) are specific to genomic research.

A critical element necessary for the success of precision medicine is the reduction in the segregation of clinical and genomic data on an individual patient level and, importantly, across large populations. Numerous groups are working toward secure, comprehensive linkage of sequenced DNA data from biorepositories with the clinically enriched information from electronic health record (EHR) systems to allow large-scale data mining for the discovery and elucidation of genotype–phenotype associations (Boland, Hirpacsak, Shen, Chung, & Weng, 2013; Ritchie et al., 2010). There are numerous methodological and infrastructure-related challenges to this approach, such as interoperability between different health systems’ databases (i.e., ability to exchange data), database security and the protection of patient confidentiality, and the high prevalence of unstandardized, unstructured clinical data present in an EHR (e.g., scanned PDF documents, narrative clinical notes rather than structured data collected through data entry into preselected drop-down fields). However, the promise of developing true learning health systems is the ability to reveal associations between clinical phenotype and possibly actionable genotype information at the population level for researcher and clinician use for developing tests and/or early interventions and personalizing treatment for individual patients (Grossman, Powers, & McGinnis, 2011).

Conclusion

As technology evolves, investigators will develop new methods of genomic analysis. This primer is not intended to be a conclusive overview of basic methods to examine genotype–phenotype associations but rather is meant to provide a common understanding of the terms genotype and phenotype and a working knowledge of the process used to determine these associations. Armed with this information, nurses can more fully incorporate knowledge of genotype–phenotype associations into clinical practice, education, and research.

Authors’ Contribution

F. Wright and K. Fessele contributed to conception, design, and interpretation; drafted and critically revised the manuscript; gave final approval; and agreed to be held accountable for all aspects of work, ensuring integrity and accuracy.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Fay Wright is funded by the National Institute of Nursing Research's postdoctoral training program, T32NR008346, at Yale University School of Nursing. Kristen Fessele received funding from the National Institute of Nursing Research's postdoctoral training program, T32NR013456, at the University of Utah College of Nursing.

References
Ikram, M. K., Xueling, S., Jensen, R. A., Cotch, M. F., Hewitt, A. W., Ikram, M. A., ... Wong, T. Y. (2010). Four novel loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo. PLoS Genetics, 6 (10), e1001184. doi:10.1371/journal.pgen.1001184.g001
Molecular and General Genetics MGG, 2, 136–137.


