ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

Genetic Influences on Disease Subtypes

Andy Dahl[1,2,3] and Noah Zaitlen[2,3]

[1]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA; email: andyd@uchicago.edu

[2]Department of Neurology, University of California, Los Angeles, California 90024, USA; email: nzaitlen@ucla.edu

[3]Department of Computational Medicine, University of California, Los Angeles, California 90095, USA

Annu. Rev. Genom. Hum. Genet. 2020. 21:413–35

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-120319-095026

Keywords

subtypes, genetic heterogeneity, precision medicine, clustering, genetic architecture

Abstract

Disease classification, or nosology, was historically driven by careful examination of clinical features of patients. As technologies to measure and understand human phenotypes advanced, so too did classifications of disease, and the advent of genetic data has led to a surge in genetic subtyping in the past decades. Although the fundamental process of refining disease definitions and subtypes is shared across diverse fields, each field is driven by its own goals and technological expertise, leading to inconsistent and conflicting definitions of disease subtypes. Here, we review several classical and recent subtypes and subtyping approaches and provide concrete definitions to delineate subtypes. In particular, we focus on subtypes with distinct causal disease biology, which are of primary interest to scientists, and subtypes with pragmatic medical benefits, which are of primary interest to physicians. We propose genetic heterogeneity as a gold standard for establishing biologically distinct subtypes of complex polygenic disease. We focus especially on methods to find and validate genetic subtypes, emphasizing common pitfalls and how to avoid them.

1. INTRODUCTION

Identifying and refining classifications of human disease are central goals in medicine and science. Precisely characterizing the nature of each person's illness improves our ability to predict outcomes, optimize medical interventions, and discover and correct for the causal disease mechanisms. To achieve these objectives, disease subtypes are continually being proposed, tested, and adopted into medical and scientific practice. Recent examples of proposed subtypes span diverse disease domains, including asthma (10, 141), type 2 diabetes (2, 69, 101, 133), breast cancer (13, 42, 57, 84, 88, 96, 112, 118), chronic kidney disease (74), autoimmune diseases (21), and neuropsychiatric diseases, including autism (9, 16, 59, 92, 99, 100, 136, 140), bipolar disorder and schizophrenia (4, 5, 56, 110), and major depression (38, 62, 102). In each case, disease subtypes have been used to improve prognosis, treatment recommendations, or knowledge of disease etiology. In this review, we focus primarily on the use of genetics to inform disease subtyping and how this can improve our ability to model and more deeply understand the architecture of complex disease.

Because subtyping is such a basic and useful concept, a diverse range of researchers spanning several disciplines and disease domains are currently proposing new subtypes. A major downside of this diversity, in our view, is that no unifying definitions, methods, or statistical criteria are being developed and broadly adopted—e.g., subtypes can be based on physician experience, computational clustering of genomic or clinical features, or clustering of genetic variants themselves. Moreover, the purpose of subtyping, along with the implicit definition of a subtype, varies markedly among communities. Therefore, before discussing the relationship between genetic variation and disease subtypes, we propose three definitions of a subtype that, we feel, capture existing usage:

- Descriptive subtypes: All subtyping efforts begin with a descriptive classification of individuals into groups. One could, for example, group breast cancer patients into those with and without estrogen receptor (ER) proteins on the surfaces of their tumor cells (ER+/−) or divide diabetics by the color of their hair (76). This initial step of defining subtypes is crucial and can be driven by prior hypotheses or putative differences in relevant features of the disease. However, just because a division of samples is descriptively interesting or plausible, there is no guarantee that it is clinically or scientifically meaningful. Moreover, there is no guarantee that important subtypes will have obvious post hoc descriptions, especially when initially discovered, and hence beginning the search for subtypes with obvious patient descriptions will not always succeed.
- Pragmatic subtypes: To go beyond merely descriptive subtypes, we define pragmatic subtypes as those that are medically relevant to disease. For example, ER status defines pragmatic subtypes of breast cancer because it has a substantial impact on prognosis and treatment recommendation (57, 96). Hair color, by contrast, does not currently define a pragmatic diabetes subtype; although hair color may eventually prove pragmatically important, we do not yet suspect that it is relevant to diabetes. Despite their utility, pragmatic subtypes could be a consequence, rather than a cause, of the disease process, and therefore may not be relevant to mechanisms causing disease. For example, some factors influence treatment response or disease progression independently of disease risk (37, 77, 87), which is of enormous pragmatic importance for patient care but not useful for studying causal disease mechanisms.
- Biological subtypes: This review focuses primarily on biological subtypes, defined as those that have distinct causal disease etiologies. In particular, we focus on subtyping complex polygenic diseases, for which genome-wide association studies (GWASs) have found hundreds of genetic risk factors with small effect sizes (80). Genetic effect size heterogeneity across subtypes is uniquely well suited to prove differential etiology, as genetic effects must

be causal (modulo uncorrected confounding structure). For example, ER status defines a biological, as well as pragmatic, subtype of breast cancer because some genetic risk factors have different effect sizes for ER+ and ER− breast cancer (13, 42, 84, 88, 112). However, hair color does not define a biological subtype of diabetes; although there certainly are genetic variants that affect hair color, these are not expected to be relevant to diabetes risk, and we are not aware of any diabetes risk variants with effects that are modified by hair color. In other words, there are genuinely distinct underlying disease components for ER+ and ER− breast cancer, despite their many common clinical presentations, while current evidence suggests the same basic disease process exists for diabetes regardless of hair color. In general, not all biological subtypes have known pragmatic significance. For example, mutations in several different genes are known to cause pragmatically similar ataxia phenotypes (58). Nonetheless, we generally expect that biological subtypes will have pragmatic importance, at least eventually; for example, drugs may be developed that affect only a subset of ataxia-causing genes.

In this review, we first present a mathematical model for polygenic subtypes of complex traits and enumerate biomedical phenomena that generate such subtypes. We then discuss approaches to define novel subtypes, focusing on the strengths and limitations of methods based on computational clustering. Next, we discuss how to statistically validate proposed subtypes and how to use validated subtypes to better understand genetic architecture, as well as to better recommend treatments and predict progression. Finally, we examine approaches to defining Mendelian disease subtypes, emphasizing the similarities and differences with the polygenic setting. We discuss breast cancer subtypes in detail in the sidebar titled Breast Cancer Has Several Descriptive, Pragmatic, and Biological Subtypes, which describes extensively studied subtypes with different levels of significance.

2. A MODEL FOR POLYGENIC SUBTYPES OF COMPLEX TRAITS

We make the broad problem of polygenic subtype inference concrete with a linear model that parsimoniously encapsulates genetic subtype heterogeneity. Specifically, we assume there are K categorical, mutually exclusive subtypes, where $z_i = k$ indicates that sample i has subtype k. Then the linear model for heterogeneous effects across subtypes is

$$y_i | z_i = k \overset{\text{ind}}{\sim} \sum_{s=1}^{S} G_{is}\gamma_{sk} + \epsilon_i. \qquad 1.$$

This model assumes that each sample i has a quantitative phenotype y_i. G represents the putatively heterogeneous covariates, which may be genetic and/or nongenetic. γ_{sk} defines the subtype-specific effect of covariate s in G for samples in subtype k. For simplicity, we have omitted homogeneous effects from Equation 1. Finally, ϵ_i is the pure noise term, which can differ in distribution across subtypes but is often assumed to be independent and identically distributed and Gaussian.

For simplicity, we have written Equation 1 by assuming that y is quantitative, to emphasize the connection between pragmatic/biological subtypes and nonzero effect heterogeneity for nongenetic/genetic columns of G. Nonetheless, Equation 1 can be naturally extended to a generalized linear model (GLM) to accommodate binary disease phenotypes. Such models can be related to Equation 1 by treating the quantitative trait as a latent liability, where samples receive the disease if the liability exceeds some threshold (29, 36, 86). The extension to case-only subtypes requires a more complex GLM, where subtypes are the outcome in a multinomial logistic regression, as discussed in Section 3.5.

BREAST CANCER HAS SEVERAL DESCRIPTIVE, PRAGMATIC, AND BIOLOGICAL SUBTYPES

Breast cancer has many partially overlapping proposed and known subtypes with varying levels of established descriptive, pragmatic, and biological significance. As discussed in Section 1, the most studied subtypes are defined by ER status. ER+/− subtypes have obvious descriptive significance and have been observed for more than 50 years (39). As the description of ER+/− subtypes proliferated, studies increasingly demonstrated their unequivocal pragmatic significance in prognosis and response rates for drugs that interact with ER, such as tamoxifen (79). Decades later, large, well-powered genetic association studies established that ER+/− subtypes have distinct causal biology (13, 42, 84, 88, 112).

Furthermore, breast cancer has several other important subtypes that variably overlap with ER status. For example, trastuzumab specifically improves prognosis when tumors express HER2 (104). Therefore, HER2 status defines a pragmatic subtype, and future work may demonstrate a biological basis to this subtype, as with ER+/−. A related pragmatic subtype with particularly poor prognosis is defined by triple-negative tumors, which do not express ER, HER2, or progesterone receptors.

Complementing these subtypes defined by specific genes, a broader gene expression signature has been developed to bifurcate patients into pragmatic subtypes with different prognoses, and these subtypes are often used to inform treatment choices (135). In another example of a subtype with a complicated description, people with high polygenic risk constitute a significant fraction of breast cancer cases and may be useful for prioritizing early screening in the population (63, 83, 103). Finally, highly penetrant genetic variants in *BRCA1* and *BRCA2* define well-known genetic subtypes of breast cancer cases that are comparatively easy to describe and already have established pragmatic significance. For example, highly penetrant *BRCA* mutations are so prevalent in some populations as to arguably warrant screening in healthy women (41, 48).

We say that covariate s is heterogeneous when $\gamma_{sk} \neq \gamma_{sk'}$ for any two subtypes k and k'. Assuming a linear model for phenotype, such an interaction between some covariate and the proposed subtypes is necessary and sufficient to produce real subtypes; otherwise, Equation 1 reduces to a standard linear model, $y_i \sim \sum_{s=1}^{S} G_{is}\beta_s + \epsilon_i$, where $\beta_s = \gamma_{sk}$ for all subtypes k and covariates s. In other words, we say the subtypes are real unless all genetic and environmental risk factors affect all individuals in exactly the same way. Note that a study finding a covariate that has a significant association in one subtype but not another is insufficient to prove that the effects differ between subtypes. Although this phrasing is statistically formal and clear, it is not immediately connected to biologically plausible mechanisms for disease subtypes. Therefore, we will enumerate several such mechanisms, which are independently interesting and demonstrate the breadth of expressivity in the model represented by Equation 1.

The first setting that leads to disease subtypes is when multiple distinct biological processes lead to a common disease diagnosis. We can stylize this process as a limiting pathway model (148), where samples carry liability along several pathways and obtain the disease if any liability exceeds its threshold (**Figure 1*a***):

$$L_{ik} = \sum_{s=1}^{S} G_{is}\beta_s^{(k)} + \epsilon_{ik} \qquad 2.$$

$$z_i = \begin{cases} k & \text{if } L_{ik} > \tau_k, 0 \quad \text{otherwise.} \end{cases} \qquad 3.$$

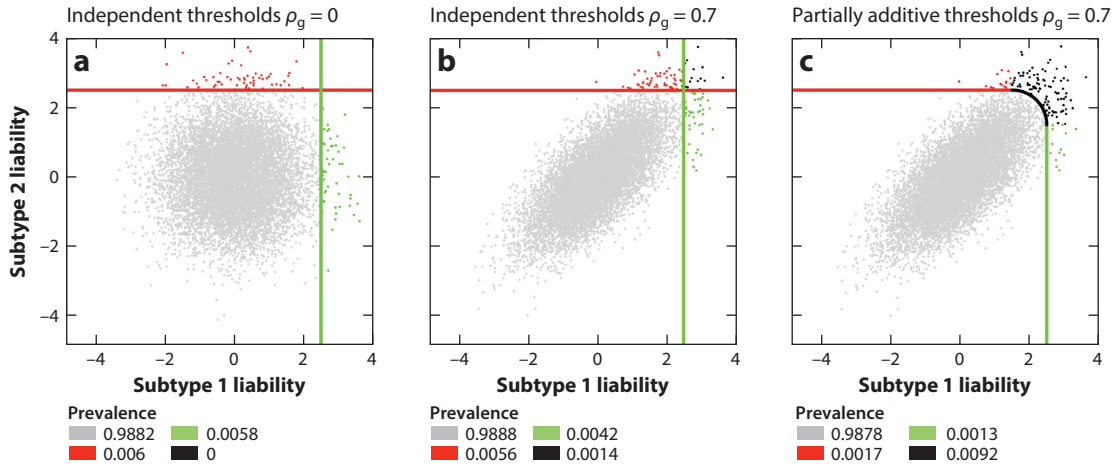

Figure 1

The limiting pathway model, which generates disease subtypes (*red* and *green*), controls (*gray*), and cases with both subtypes (*black*). Simulations draw bivariate Gaussian liabilities independently for 10,000 individuals. (*a*) The disease pathways are genetically uncorrelated, and the disease is obtained if one liability exceeds its threshold. Subtypes are naturally defined by which pathway exceeds its threshold. Because the disease is rare (∼1%), the chance of having both subtypes is negligible—and it never happens in this simulation. (*b*) The disease subtypes now have substantial genetic correlation, representing a partially overlapping subtype etiology, and ∼10% of cases have both subtypes. (*c*) The limiting pathway model is relaxed to accommodate partially additive pathway effects on disease, where only ∼20% of cases can be ascribed to a single pathway.

Here, L_{ik} is the liability carried by person i along pathway k, τ_k controls the prevalence of each disease subtype, and ϵ_{ik} is the noise in person i's pathway k–specific liability.

Equation 3 assumes the disease is rare enough that no samples exceed multiple thresholds. However, if the genetic effects in β or the noise in ϵ is correlated across subtypes, many people will have multiple disease subtypes (**Figure 1*b***). Furthermore, real diseases may often result when samples have multiple near-threshold pathways, which also increases the prevalence of multi-subtype disease (**Figure 1*c***). Nonetheless, a key feature of the limiting pathway model is that subtypes remain meaningful—so long as the pathways are not perfectly correlated, in which case the disease model has effectively one pathway—for two reasons. First, there will be patients with only one subtype, implying that pathway-specific treatments will be completely ineffective for some people [i.e., $\nabla_{L_k} P(\text{disease}) = 0$ for some k along nonnegligible portions of the disease boundary]. Second, it remains important to understand each pathway to disease, both for scientific purposes, such as genetic mapping, and for medical purposes, such as developing new treatments.

Biologically, distinct pathways could represent different disease-relevant tissues, where the limiting pathway model assumes that disease results whenever any tissue is impaired. For example, endocrine-relevant diseases often broadly result from miscoordination between tissues, suggesting that defects in any tissue will cause systemic problems. Second, pathways may represent distinct genomic pathways to a shared phenotype—e.g., mutations in diverse genes converge on low-density lipoprotein (LDL) cholesterol level (54, 65). Third, pathways may represent endogenous and exogenous disease components—e.g., alcoholic fatty liver disease is a biologically important liver disease subtype, but the direct causal mechanism is environmental (albeit heritable). Fourth, one pathway may contain Mendelian variants and another may contain phenocopying polygenic variants—e.g., hypercholesterolemia may result from highly penetrant genetic variants in the LDL receptor (45) or many variants of small effect that act through body mass index (BMI) and/or other components of cholesterol metabolism (44). Fifth, some diseases may be secondary

to several distinct primary disorders—e.g., chronic kidney diseases may result from hypertension, obesity, and/or type 2 diabetes (74). Broadly, pathway-driven subtypes can differ in treatment response, progression, genetic architecture, tissue and genomic feature enrichment, and comorbidities whenever the underlying pathways differ in these properties.

A second important source of subtypes is gene–gene interaction (G×G). One particularly prominently studied form of G×G is interaction between autosomes and sex chromosomes (G×Sex). For example, sex plays a significant role in modifying the functional genomic basis of gene expression, which may propagate in important ways to complex traits (46, 64, 128). Furthermore, G×Sex is clearly established for many diseases and disease-related phenotypes, including asthma (94), autism (89), immune response (66), and metabolic traits, particularly adipose tissue distribution (117, 120). Although differences in disease prevalence or heritability between sexes may be important or descriptively interesting, they can arise under purely additive models and are not sufficient to establish sexes as biological disease subtypes.

Third, gene–environment interaction (G×E) is a classically studied phenomenon that can induce subtypes. G×E is well known for many organisms and phenotypes and has recently attracted attention across a wide range of environments (*E*) for genomic and complex traits in humans (6, 27, 35, 72, 90, 109, 145, 146). The fact that G×E can induce subtypes is easily seen when *E* is a categorical environment—if we set $z = E$, Equation 1 immediately becomes a G×E model. For example, G×E has been established for smoking status as *E* for lung function (51) and lipid profiles (7). When *E* is continuous, it instead induces a continuous subtype gradient, which can still be approximated by the discrete subtype model, albeit with a loss of information and parsimony. This link to G×E is important because tests and intuition for G×E are relatively mature and largely carry over to testing subtype heterogeneity. We will return to this point when we discuss tests for subtype validation (Section 3.3), many of which are applications or modifications of G×E tests. We note that the line between G×G and G×E is conceptually useful but usually blurry, as the environments used as *E* are often heritable (32).

A fourth source of subtypes is disease misdiagnosis, which is particularly important for rare or difficult-to-diagnose diseases. For example, many psychiatric diseases are difficult to distinguish clinically, especially in early stages of the disease or diagnostic process. While precise misdiagnosis rates are difficult to estimate and vary across time and countries and as a function of patient demographic features, misclassification rates on the order of 50% are not uncommon (1, 23, 93). High rates hold even for severe psychiatric diseases with seemingly obvious delineations, such as bipolar disorder and schizophrenia, where misdiagnosis rates are on the order of 10% (14, 71, 142). Therefore, we can imagine that labeled schizophrenia cases are a mixture of at least two subtypes: bona fide schizophrenia cases and misdiagnosed bipolar disorder cases. Clearly, covariates that differentially affect true schizophrenia and bipolar disorder will have heterogeneous effects across these subtypes. These subtypes can be considered merely statistical, as they disappear under perfect diagnoses; on the other hand, such subtypes are likely to persist in practice and have significant implications for clinical treatment and scientific estimates, such as genetic correlation (Section 3.5). Misdiagnosis can also be considered a special case of the limiting pathway model, with one pathway being the ordinary pathway to the undiagnosed disease.

3. APPROACHES TO IDENTIFY AND VALIDATE POLYGENIC SUBTYPES

3.1. Methods for Proposing Subtypes

In this section, we examine methods for proposing new subtypes of complex polygenic diseases. By construction, subtyping is a clustering problem, where like patients are grouped by taking

relevant features as input and producing putative subtypes as output. Commonly used features for subtyping include standard clinical traits (e.g., age, BMI, and LDL cholesterol levels), gene expression [e.g., for breast cancer prognosis (135)], relevant molecular biomarkers (e.g., white blood cell count), or genetic and genomic properties [e.g., number of de novo copy number variants (CNVs) or genetic load in conserved regions]. We focus on computational clustering approaches that exploit large and/or high-dimensional modern data sets but emphasize that the basic concept of clustering similar patients has always guided nosology.

3.1.1. Clustering clinical phenotypes. One reasonable hypothesis is that distinct disease subtypes manifest with distinct clinical features, comorbidities, or progression. This is the classical approach to define diseases and their subtypes, historically achieved through cumulative physician experience (116). This history has motivated several recent studies to propose subtypes by computational clustering on clinical, transcriptomic, metabolomic, and other high-dimensional phenotypes. These approaches aim to go beyond traditional nosology by using more numerous, precisely measured, and/or disease-relevant features; sample sizes much larger than any physician sees in a lifetime; and sophisticated decompositions that can uncover patterns that are invisible to human eyes and intuition. Most studies apply off-the-shelf clustering tools, including both relatively simple approaches, such as *k*-means (2, 138) or hierarchical clustering (131), and more complex approaches, such as topological data analysis (53, 75, 95). These tools are valuable for their simplicity and reliably find the dominant structure in a data set (**Figure 2*a***). Therefore, these tools succeed when disease subtypes are the primary source of heterogeneity. Conversely, these tools cannot discern between meaningful and meaningless structures in the data, and in particular will reliably uncover only confounders when they are strong (**Figure 2*b***). For example, the five type 2 diabetes subtypes hypothesized by Ahlqvist et al. (2) were driven partially by age differences, which is as expected because many of the features used differ by age (e.g., BMI and age itself); however, these age-correlated disease subtypes are necessarily transient and are not exclusively capturing intrinsic organismal biology (119).

Dahl et al. (24) recently developed a method for subtyping clinical traits, RGWAS (reverse GWAS), primarily to address this issue of confounding. By carefully incorporating covariates to remove heterogeneity induced from confounders such as age and population structure, RGWAS increases the odds that the discovered subtypes are biologically meaningful (**Figure 2*c***). This is essential for complex traits, as it is well known that causal polygenic effects are often swamped by confounding population structure (106, 134). Note that standard approaches to confounder

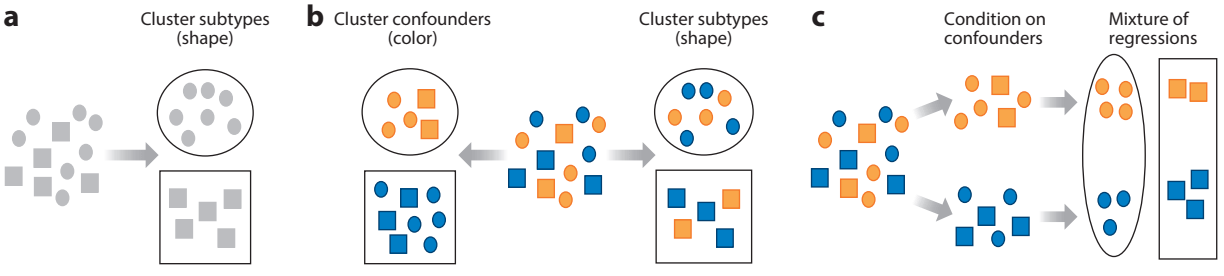

Figure 2

Clustering clinical traits to define subtypes. (*a*) When the only structure in the data consists of real subtypes (*circle* and *square shapes*), any reasonable approach will recover those subtypes. (*b*) When additional confounders are present (*blue* and *orange colors*), standard clustering will choose whichever feature is strongest. (*c*) By conditioning on confounders as covariates, methods such as RGWAS (reverse GWAS) prioritize the desired subtype structure over spurious signals.

correction, such as clustering on the residuals from a homogeneous regression on the confounders, will produce biased clusters (24, 27). Going forward, tools that jointly infer a mixture of partially overlapping clusters in clinical traits would be useful. For example, proposed breast cancer subtypes are often profiled across both ER+ and ER− subtypes to test for subtype–subtype interactions (see the sidebar titled Breast Cancer Has Several Descriptive, Pragmatic, and Biological Subtypes). Generally, clinical traits are likely affected simultaneously by demographic, genetic, and environmental subtypes, and fully understanding each component requires also learning the others; in **Figure 2**, this amounts to jointly learning both the shape- and color-level structures.

Finally, we note that confounder correction is not necessary when meaningful subtypes are much stronger than confounders [e.g., for gene expression signatures in breast cancer tumors (135)] or when likely confounders are controlled by design [e.g., by excluding patients on anti-inflammatory drugs when seeking an inflammation signature in asthma (141)]. However, such successes rely heavily on strong subtypes or prior expertise and are likely to dwindle over time: Unknown subtypes will generally be more subtle than known subtypes and/or more obscured by complex, uncontrollable confounders.

3.1.2. Clustering genetic risk. Another plausible hypothesis is that subtypes are driven by dysregulation in different tissues, cell types, or genomic pathways, consistent with the limiting pathway model. Under this scenario, it may be impossible to subtype based on clinical features if the limiting pathways converge on an identical organismal phenotype. However, approaches based on pathway-specific risk factors can distinguish subtypes, provided that we can approximate these underlying pathways. This may be possible with pathway-specific measurements, e.g., using expiratory measurements as a proxy for lung function (51) or metabolites such as albumin to profile kidney function (74). One genetic approach to approximate these risks uses functional genomic annotations. This method has been proposed for neurodevelopmental disorders as an extension of genotype-based approaches for Mendelian disease (9, 33, 52, 61, 123) to a molecular subtyping approach that aggregates genetic effects at the level of genes, pathways, or other functional categories (43, 68, 97, 124, 125). These ideas motivate directly clustering single-nucleotide polymorphism (SNP) effect sizes as a function of genomic annotations or pleiotropy to secondary traits. For example, a variant of nonnegative matrix factorization (NMF) has been applied to the matrix of z-scores for dozens of type 2 diabetes risk SNPs and dozens of type 2 diabetes–relevant traits to infer continuous subtypes of SNP effects (133). The SNPs in each factor can be tested post hoc for enrichment in genomic features. It is also possible to calculate each person's genetic risk per pathway post hoc, though these scores will not generally inherit a subtype structure even when SNP effects have subtypes (133).

We caution that differential genomic enrichments or comorbidities across putative subtypes can be observed even under a purely additive disease model because patterns of pleiotropy vary along the genome and can be annotation dependent (115). Hence, under complete homogeneity, inferred SNP or sample subtypes will have differential genomic enrichments and comorbidities by construction. For example, cases harboring genetic disease risk in genes specifically expressed in the liver will likely have different secondary liver-related phenotypes than samples harboring primarily lung-based risk even absent any real subtypes, because liver and lung have different patterns of pleiotropy with secondary traits. Nonetheless, these examinations can be useful for proposing and characterizing subtypes.

3.1.3. Clustering genotypes. A final set of approaches are based on the hypothesis that minor allele frequencies will differ among subtypes at causal variants. Under this assumption, clustering genotypes of cases could produce meaningful subtypes. For example, Arnedo et al. (4) performed

NMF on schizophrenia-relevant SNPs and interpreted the resulting factors as distinct schizophrenia subtypes. Unfortunately, this approach is unlikely to be useful for highly polygenic diseases and subtypes, as differences in minor allele frequency between true subtypes will be small and dispersed over many hard-to-detect variants. While this may suggest applying NMF to a large number of risk SNPs, such approaches are liable to detect confounding population structure rather than disease subtypes; in fact, when all SNPs in the genome are included, NMF is an established way to estimate latent population structure (34).

3.2. Pitfall 1: Population Structure Produces Replicating, Seemingly Heritable Subtypes

In an ordinary GWAS, uncorrected population structure causes spurious genetic associations when population-scale axes of genetic variation are confounded by nongenetic effects (107). Population structure in genetic data may derive, for example, from distinct continental populations, isolation by distance within a continent, or fine-grained structure within geographic regions (82). Therefore, a key step for calibrated statistical tests in genetic association studies is correcting for this structure via proxies for ancestry, such as genetic principal components.

In an exactly analogous way, subtype-specific population structure will confound downstream genetic heterogeneity tests (see also Section 3.3). This means that both main and subtype-specific effects of population structure must be adjusted (24), as has been proposed for G×E tests (129). For example, it may be the case that population A has much higher risk than population B for disease subtype 1, but they have essentially the same risk for subtype 2. In this case, heterogeneity tests that adjust only for the population main effects will have false positives: Any disease-irrelevant variant that happens to have a higher minor allele frequency in population A will spuriously predict subtype 1 over subtype 2. Worse yet, this spurious heterogeneity will replicate in independent data sets with similar forms of population structure.

However, population structure presents an even more severe problem for inferring subtypes. As discussed above, almost all clustering approaches simply aim to find the largest structures in the data, regardless of whether they are confounders or biomedically useful (**Figure 2**). In the case of polygenic subtyping, confounding by population structure is likely to be much stronger than causal genetic subtype structure. As a result, clustering approaches that ignore population structure are liable to learn populations themselves as so-called subtypes (2, 4, 55, 67, 75). Indeed, there is evidence that published inferred subtypes of clinical or imaging traits in multiethnic populations may be confounded by ancestry, which, furthermore, led to spurious genetic associations; for example, the top subtype-specific SNP associations in papers by Krishnan et al. (67) and Li et al. (75) have dramatically different allele frequencies in the Genome Aggregation Database (gnomAD) across Africans, East Asians, and/or Europeans.

Of course, there are known cases of important subtypes—and diseases—that differ in prevalence among populations. For example, some patients have a severe, life-threatening adverse response to the anticonvulsant drug carbamazepine, a condition known as Stevens–Johnson syndrome and toxic epidermal necrolysis (SJS/TEN). This pragmatic subtype is most common in Asian populations, motivating recommendations for genetic screening in these populations (37). In particular, screening for the main risk allele in Taiwan reduced the incidence of carbamazepine-induced SJS/TEN to zero in a study of approximately 5,000 people (20). This example also illustrates the importance of studying subtypes across diverse populations [carbamazepine-induced SJS/TEN is less common in Europeans (85)].

Despite the importance of such examples, we emphasize that, until proven otherwise, population-differential subtypes should be assumed to be confounded rather than causal,

especially given the history and horrific consequences of falsely assuming population specificity for certain diseases. For example, cystic fibrosis was long assumed to be specific to Europeans by many scientists and doctors, leading to historical and current biases in research and diagnosis (126, 127), which in turn led to dramatic differences in detection and treatment among different ethnicities and countries, including threefold-higher mortality rates in Hispanic compared with non-Hispanic patients in California after adjustment for "socioeconomic status and clinical risk factors" (17, p. 380). Moreover, as was the case for tests of polygenic selection (8, 121, 134), we anticipate that polygenic tests for subtype heterogeneity will be especially sensitive to uncorrected population structure.

Finally, we note that regressing out genetic principal components a priori from the clustered features is not generally sufficient to yield calibrated downstream heterogeneity tests. Although this method will adjust for homogeneous confounding and likely improve the utility of the inferred subtypes, it is insufficient to control false positives (24). Intuitively, projecting out the additive effect of principal components from the phenotypes in the model represented by Equation 1 may remove the confounding effect on the across-subtype average effect (a weighted sum over the γ_k), but it will not appropriately remove confounding from each subtype-specific effect (γ_k).

3.3. Tests to Validate Proposed Subtypes

In this section, we discuss approaches to demonstrate that a putative clustering of samples into subtypes is meaningful. In other words, these approaches take descriptive subtypes as input and test them for pragmatic or biological significance. We focus on genetic tests for biological significance, though we also mention genomic-based tests and pragmatic significance tests.

Conceptually, the goal is to evaluate whether subtypes have distinct genetic architectures, which demonstrates causally distinct subtype biology. This can be done by testing genetic variants for heterogeneous effects, be they SNPs, de novo mutations, CNVs, polygenic scores, large-effect rare variants, or fully Mendelian variants. However, because putative subtypes have subtle and polygenic architectures, single-SNP tests are often underpowered. Furthermore, even successful GWASs for G×E usually discover only a handful of variants, which is only modest evidence that subtypes warrant further investigation.

Nonetheless, complex traits can harbor so many small SNP effects that, cumulatively, they explain a substantial portion of trait variation (143). In the ordinary, additive case, two established approaches to aggregate this signal across all SNPs in the genome are polygenic risk scores (PRSs) and genome-based restricted maximum likelihood (GREML). Both approaches sacrifice SNP-level resolution in exchange for added power to detect and quantify the overall genetic effect. Therefore, extensions of PRSs and GREML to subtype-specific SNP effects are invaluable for validating genetic subtypes: It is necessary only to robustly demonstrate that the subtypes differ genetically, not to finely partition this difference. We discuss methods that are suited to this latter task in Section 3.6, which is useful for refining our understanding of biologically validated subtypes.

Broadly, PRSs are constructed by adding the estimated effects of many relevant SNPs. Regardless of how exactly the PRS is built, it can be used as a univariate summary of genome-wide effects and hence can immediately be used in many preexisting heterogeneity tests. For example, PRSs can be used to test for genetic interaction with covariates such as age (83) or stress (102) or to evaluate comorbidity from pleiotropy and/or misclassification (110). PRSs can also be readily applied in newly developed SNP-level tests for added power, as in the TreeWAS test for heterogeneity across International Classification of Diseases (ICD) codes (22).

GREML also aims to aggregate SNP effects across the genome. Unlike PRSs, which explicitly estimate and combine individual SNP effects, GREML directly learns their aggregate size with a

random effect model that integrates out individual SNP effects. Consequently, it has more power to detect the existence of polygenic effects and can estimate the total contribution in an unbiased way (60, 122). Like PRSs, GREML can also be extended to accommodate genetic interactions (27, 109, 144), which can be used to model subtype-specific heritability (24) and more general polygenic interactions with environmental or other genetic covariates. Some (but not all) such tests for interaction are generally calibrated even when the disease and subtype status itself (or environment) share some genetic basis (i.e., gene–environment correlation) (3, 19, 27). Similarly, some (but not all) tests for interaction are dependent on the chosen scale for quantitative phenotypes and chosen link function for categorical phenotypes (130, 146).

We also note that many nongenetic covariates have large effects and therefore can be invaluable for powerfully validating and characterizing subtypes, even though these tests do not establish biological subtypes. For example, pharmacogenomic variants can define bona fide subtypes of disease that do not respond to standard treatments and even have genetic bases (37, 77, 87)—but these variants are not generally risk factors for disease itself. In another example, the elderly have dramatically different prognoses and treatments for common infections like pneumonia, even though the causal disease process is often thought to be largely homogeneous. Of course, pragmatic subtypes are incredibly important in their own right, but this is not our focus.

3.4. Pitfall 2: Clustering Methods Make Clusters

Subtypes produced by clustering are constructed to have distinctive features. This holds both for physician-intuited subtypes and for computational algorithms. For example, imagine we care about a disease that has nothing to do with height, but we nonetheless cluster people into shorter and taller groups—perhaps by using *k*-means on many biometric measurements. These height-based subtypes will significantly differ across all height-correlated traits, have significant genetic differences because height is heritable, and pass ordinary reliability metrics, such as clustering strength and external replication. Although all three properties seem to indicate that height is a subtype, they are in fact spurious metrics.

This issue is essentially just overfitting: The clustering process creates clusters, and downstream tests that are ignorant of this fact will falsely validate the clusters. Indeed, this is not a failure of the downstream test. However, it absolutely is a failure of the analyst, because they are testing the wrong hypothesis. The relevant question is whether the putative subtypes are genetically heterogeneous conditional on the fact that they were constructed from potentially confounded and overfit knowledge. This strong null hypothesis is not evaluated by ordinary downstream tests. Similar care is needed for other two-step tests in genetics, as when imputing genotypes (81) or phenotypes (26) or when correcting for estimated confounders in gene expression (25, 73, 137).

We note that many published studies have used computational clustering to infer subtypes and then, downstream, used *t*-tests to argue that the clusters differ on the clustered variables. Such examples are widespread in the electronic health record literature and are increasingly common in other fields. This approach is statistically invalid. Nonetheless, the miscalibration is negligible when clustering uncertainty is negligible. For example, this approach is common in single-cell sequencing studies, where scientists wish to evaluate which genes distinguish inferred cell types. This clustering may be nearly perfect for broad cell types, yielding nearly calibrated differential expression tests. However, as focus shifts to rarer, poorly understood cell types, the overfitting issue will inevitably become important. Nonetheless, even when the tests are miscalibrated, they can be used to prioritize likely relevant features.

One approach to avoid this overfitting is to specifically treat the putatively heterogeneous variable as strictly additive during subtyping (24). This approach is inspired by score tests, which test

an alternative hypothesis while evaluating only the null hypothesis. Conceptually, this restriction means the clustering algorithm cannot overfit the putatively heterogeneous variable. This restriction is practically important, as it is very easy to define simulations where subtypes are nonexistent, tests based on *k*-means and Gaussian mixture models have an ∼100% false-positive rate, and the RGWAS tests described are calibrated (24).

3.5. Testing Case-Only Subtypes

So far, the models considered mostly assume subtypes that span both cases and controls. For example, the subtypes may be smoking status and the phenotype lung cancer, or the subtypes may be defined by the presence of a major stressful life event and the phenotype psychiatric disease. In this section, we consider disease-specific subtypes, where only cases are separated into subtypes and all controls are grouped into a single different subtype, as in the limiting pathway model (**Figure 1**). Case-only subtypes are also natural when classifying with disease-specific features, such as the gene expression of a tumor, or features that are too expensive or unethical to measure in controls, such as treatment side effects.

The interaction model in Equation 1 is not meaningful for case-only subtypes; for example, a perfect fit is obtained by predicting the disease if and only if samples are in a case-only subtype. Instead, we consider multinomial logistic regression (MLR) a sensible default model for case-only subtypes. MLR is a classical GLM (86) that has previously been proposed for testing genetic heterogeneity in disease subtypes (18, 91). However, it is not common in practice, and many studies instead use meta-analysis-inspired approaches—for example, by separately comparing each case subgroup with controls and using permutations to account for the shared controls.

MLR uses a linear model for the log odds ratio that compares disease subtype *k* with a reference group, which we take to be the controls (i.e., subtype 0):

$$\log\left(\frac{p_{ik}}{p_{i0}}\right) = \sum_{s=1}^{S} G_{is}\beta_s^{(k)}. \qquad 4.$$

Here, p_{ik} is the probability that sample i is in subtype k, which is governed by the covariates in G and the subtype k–specific effect sizes $\beta^{(k)}$. When there is only one case subtype, MLR reduces to logistic regression, and $\beta^{(1)}$ reduces to the ordinary case/control odds ratios. We note that MLR cannot accommodate noncategorical case subtypes, i.e., fractional proportions or continuous case-only features.

MLR can test for genetic heterogeneity by testing whether all SNP effect sizes are equal. The null hypothesis is that $\beta_s^{(1)} = \cdots = \beta_s^{(K)}$, which allows the SNP to affect disease but requires the effect to be equal across subtypes. However, this heterogeneity test is not sufficient to establish genetic subtypes because it does not show that the SNP is disease relevant, in the sense of having a nonzero marginal effect on disease. This is an important caveat because most putative disease subtypes will be defined based on heritable factors, which in turn induces genetic heterogeneity regardless of disease relevance. For example, dividing diabetes patients by hair color likely does not define a biologically meaningful diabetes subtype, but hair color SNPs will nonetheless have subtype-specific associations. For the same reason, case–case tests are susceptible to false positives from disease-irrelevant differences among subtypes. We therefore require that SNPs additionally have nonzero main effects on disease before determining that they define genetic subtypes. Nonetheless, this restriction reduces power to detect biological subtypes, and the optimal combination of significance thresholds for the homogeneous and heterogeneous tests remains an open issue.

One approach to solve this difficulty, Subtest, uses a mixture model to partition the genetic differences among subtypes into disease-relevant and disease-irrelevant components across the genome (76). By contrast, standard estimates of genetic correlation are susceptible to false positives from heritable disease-irrelevant subgroup differences (e.g., as with diabetes and hair color). In other words, genetic correlation below 1 is insufficient to establish disease subtypes, because this quantity does not require genetic differences among subtypes to be rooted in disease biology. Future methods in this vein would be useful, as Subtest is computationally challenging and accommodates only two case subtypes.

Another limitation of MLR is its symmetric treatment of all subtypes. This symmetry is unlikely to hold for controls and case subtypes, as case subtypes are almost always more similar to each other than they are to controls. One alternative is ordinal logistic regression, a GLM that assumes that subtypes are ordered (including controls), which is natural for diseases with an ordered progression, such as tumor grades and chronic kidney disease stages (74). Another common setting that violates the MLR symmetry assumption is where subtypes are defined by combinations of discrete features; for example, breast cancer tumors can be divided by the presence of ER, human epidermal growth factor receptor 2 (HER2), and progesterone receptors, defining $2^3 = 8$ possible case subtypes (see the sidebar titled Breast Cancer Has Several Descriptive, Pragmatic, and Biological Subtypes). Complex variants of MLR have been developed for this setting by hierarchically testing for heterogeneity within and between discrete features, with an emphasis on scaling to large numbers of subtypes, given the combinatorial subtype explosion as features are added (18). In the future, it may be helpful to pursue richer discrete choice models from the econometrics literature to accommodate these and other case subtype structures (see chapter 21 of Reference 47). It would also be helpful to develop polygenic versions of these heterogeneity tests—beyond simply using PRSs—though it will be nontrivial to ensure that the polygenic distinctions are disease relevant (76).

One particularly important nonexchangeable subtype structure is hierarchically organized subtypes. The most prominent example is the ICD classification system, which is common for electronic health records and continually (re)evaluated for descriptive, biological, and pragmatic relevance. For example, ICD-10 posits that cervical and lumbar ankylosing spondylitis—both of which are ankylosing spondylopathies—are more related to each other than to other spondylopathies. TreeWAS is built for exactly such subtype hierarchies (22). Inspired by models of mutation along a pedigree, TreeWAS partitions genetic association across a hierarchy of disease subtypes using a Markovian inheritance model and can be used for SNPs or PRSs. Interestingly, it can be reversed to evaluate the hierarchy itself, assuming that genetic variants generally cluster under a bona fide hierarchy. Nonetheless, TreeWAS is nascent, and it will be important in the future to accommodate misdiagnosis, nongenetic correlations across subtypes, and covariates—especially population structure (see Section 3.2)—as well as to provide frequentist coverage.

Finally, we note that misdiagnosis causes special concerns for case-only subtypes and, more broadly, when comparing similar diseases. Although measurement error is meaningful more generally, diseases are often diagnosed exclusively—de facto, if not explicitly—which creates complicated dependencies among disease labels even for independent diseases. These issues are particularly salient in psychiatry, where diagnostic criteria remain in flux even for long-known disorders; a prominent example is the recent replacement of categorical autism subgroups with an autism spectrum in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (70). Generally, discussing misdiagnosis is complicated by the lack of gold-standard diagnoses. Nonetheless, when external gold (or near-gold) standard diagnoses are available, they can be leveraged to inform common patterns of misdiagnosis and their impact on genetic architecture. One such approach, BUHMBOX (Breaking Up Heterogeneous Mixture Based on Cross-Locus

Correlations) (50), uses GWAS results from a carefully phenotyped cohort to detect contamination of cases from a different disease. For example, rheumatoid arthritis can be subtyped based on the presence of one or more anti-citrullinated protein antibody (ACPA) biomarkers (49), and BUHMBOX detected contamination from ACPA+ cases in an ACPA− cohort. The key feature of BUHMBOX is distinguishing pleiotropy from misclassification: The former manifests as genetic correlation between all ACPA− cases and the ACPA+ GWAS, while the latter results in a mixture, with some so-called ACPA− cases correlating essentially perfectly with ACPA+ cases and the rest having much lower correlation. Despite its novelty, BUHMBOX is limited by modest power and its reliance on a gold-standard external GWAS.

3.6. Testing SNP Effect Heterogeneity with Subtypes

In this section, we discuss how validated subtypes can be used to learn about genetic and nongenetic heterogeneity. This approach is deeply related to the goal in the previous sections: validating subtypes by demonstrating covariate heterogeneity. However, we distinguish these complementary goals primarily because different methods are relevant. Identifying and validating subtypes require genome-wide aggregate effects and large-effect covariates for power and robustness. But SNP-level tests are useful for providing more fine-grained characterization of the genetic architecture of subtypes once they have been validated and, moreover, can increase power to detect genetic associations.

When subtypes are learned jointly across all samples in a data set, Equation 1 provides natural GLM tests for SNP heterogeneity. For case-only subtypes, models like MLR are instead more natural, as given in Equation 4, though there remain several open problems in this area. Crucially, these GLMs easily allow covariates, which are needed to control for population structure and increase power. For example, in GWASs, it is well established that genetic principal components (or an analog, such as linear mixed models) must be used to adjust for confounding structure (see Section 3.2). In G×E testing, this remains necessary but is no longer sufficient: It is essential to also correct for subtype-specific principal component effects [or subtype-specific kinship matrices in a GREML setting (129)].

We also note that testing SNPs for variance effects can effectively prioritize likely-interacting SNPs, because unmodeled interactions induce heteroscedasticity (15, 98, 146). This approach has the benefit that the interacting variable(s) need not be specified. A recent approach, StructLMM (Structured Linear Mixed Model), goes further when many putatively interacting variables are available. StructLMM is particularly useful for high-dimensional covariates or subtypes, as it collapses information across all these groups to derive a powerful SNP-level test. StructLMM and GxEMM (Gene–Environment Interaction Mixed Model) (27) are complementary tests for high-dimensional G×E, with the former integrating out the environmental dimension and the latter integrating out the gene dimension.

4. MENDELIAN EFFECTS AND RARE GENETIC SUBTYPES

While not the primary focus of this work, we outline general practices for subtyping Mendelian diseases to describe their conceptual relationships to subtyping complex polygenic phenotypes. At a high level, many of the methods and goals are similar, but the relative strengths and weaknesses can be very different. Also, borrowing intuition from Mendelian subtypes can inform fruitful paths forward for polygenic subtyping, as Mendelian subtypes are generally better understood—mirroring the history of complex diseases generally being more statistically difficult to parse than Mendelian diseases.

GWAS results have established that complex diseases are highly polygenic, with at least hundreds of individual small-effect loci contributing to risk (80). Formally, each possible genetic profile represents a subtype with unique biology, but it is not scientifically or medically useful to consider these subtypes—even if a complex disease has only 30 causal variants, there will be hundreds of billions of possible genetic configurations.

Mendelian diseases, on the other hand, are amenable to genotype-based subtyping because they are caused by rare, penetrant mutations. Highly penetrant or large-effect variants are similarly useful for genotype-based subtyping. For example, some CNVs are sufficiently penetrant and prevalent to constitute important subtypes of common complex diseases (147), such as the 16p11.2 CNV that explains ~1% of autism cases (140). Because the biological disease mechanisms are clearly distinct, at some level, for these CNV-based subtypes, they are biologically significant. And, due to the high penetrance, some CNV or large-effect variant subtypes are sufficiently prevalent to be pragmatically useful for diagnostic purposes, even if they do not cause distinct phenotypes.

There has been long-standing debate about the biomedical utility of genotype-based subtypes. In psychiatry, these causation-based somatoetiological subtypes have long been preferred over the symptom-based symptomatological subtypes in principle; however, they have a worse track record for replication, which is due at least partially to incorrect approaches for proposing and validating causal mechanisms driving subtypes (116). These arguments roughly correspond to our discussion of clustering polygenic traits based on genomic risk versus clinical phenotypes (Section 3.1). Clustering genomic-annotation-specific PRSs is similar to Mendelian genotype-based subtyping because it collapses close polygenic genotypic subtypes to increase power and utility, not dissimilar to burden tests that aggregate over likely-similar rare variants in a region.

Recently, Eichler and others (9, 33, 43, 52, 61, 68, 97, 123–125) have published a series of works demonstrating the trade-offs involved in subtyping based on genotypes, genomic risk factors, or phenotypes; although focused largely on neurodevelopmental disorders, the concepts discussed in these works are fundamental and shared across many rare diseases. A similar story is emerging for schizophrenia, where rare and common risk SNPs harbor partially additive effects that can converge at the gene or pathway levels (113, 139). Such genetic architectures are hypothesized to be relevant for many other complex diseases (113) and dovetail with the omnigenic model (11).

We outline several benefits and complexities of genotype- and phenotype-based Mendelian subtyping below. Throughout, we focus on the hereditary ataxia phenotype as an example, because our current understanding of ataxia includes both genotypic and phenotypic subtypes and suggests that many biomedically significant subtypes remain unknown. We emphasize that each approach has complementary strengths, depending on the specific setting, but that genetic subtypes have a unique ability to inform causal mechanisms and guide scientific study.

4.1. Genotype-Based Subtypes

The first and primary argument for genotype-based subtypes is that they are based directly on causation and are therefore biological subtypes. Relatedly, genotype-based subtypes naturally suggest tractable, testable hypotheses about causal disease architecture. Instead of scanning for genome-wide signals in a genetically heterogeneous disease, scientists can anchor their understanding to simpler, biologically grounded, genotype-based hypotheses and then dig more efficiently and deeply into causal mechanisms, as has been done for candidate genes prioritized by GWASs (e.g., 113, 132).

A second, pragmatic argument is that different genotype-based disease subtypes often present similar symptoms and hence cannot easily be phenotypically subtyped. For example, many subtypes of ataxia are clinically indistinguishable (58). In another example, amyotrophic lateral sclerosis is diagnosed by exclusion, but early diagnosis of a known subtype, such as C9ORF72 (28, 108), can prevent a long, painful, and expensive diagnostic odyssey. Moreover, some genotypic subtypes may differ in prognosis or treatment response, and therefore genetically resolving subtypes can provide substantial and immediate value even for subtypes without obvious clinical differences.

There are also limitations to genotype-based subtypes. First, there is the practical matter that not all large-effect genes are known for all Mendelian and rare diseases. Nonetheless, this is often a reason to care more, not less, about genetic subtypes: Globally characterizing the diverse genetic variants that broadly converge on cystic fibrosis phenotypes is pragmatically (and ethically) important for early detection, which is important, in turn, for the early interventions that can dramatically improve outcomes (see Section 3.2). Second, as for biological subtypes of complex disease, some genotype-based Mendelian subtypes may not have any (known) clinical relevance. Third, an opposite situation can occur: Instead of several genotype-based subtypes having a shared clinical presentation, a single genotype-based subtype may have diverse clinical presentations. For example, cystic fibrosis is monogenic, yet genetic and nongenetic modifiers can lead to diverse presentations (31). In these cases, the genotypic subtypes may have reduced clinical utility.

4.2. Phenotype-Based Subtypes

Before the development of efficient technologies for sequencing DNA, all subtypes were identified through phenotypic clustering. Friedreich's ataxia, for example, was discovered in 1863 and was shown to be hereditary in 1876, but characterization of the causal gene lagged behind by more than a century, as this characterization required DNA sequencing (114). The main benefit to phenotypic subtyping is that, almost by definition, the subtypes will be clinically relevant, as they are defined based on physician experience and disease-relevant measurements. Another benefit is that phenotypes are often easier and cheaper to measure than genotypes. Finally, even when phenotypic categorization oversimplifies causal genetic heterogeneity, this simplification can be preferable from a clinical perspective, especially if clinical features predict drug response more accurately than genotypes do. However, developments of new treatments may benefit from considering genetic subtypes, as phenotypic subtypes are generally backward looking, in that they are defined based on established phenotypic differences rather than potential therapeutic differences. This also holds for scientific analyses, as collapsing patients with distinct causal architectures will reduce power and muddy interpretation.

Finally, we acknowledge that division of disease into Mendelian and polygenic is an oversimplification, as many (or perhaps all) diseases have components of each architecture. In fact, some patients may have a disease only when both polygenic and large rare-variant contributions are present, and mounting evidence suggests that complex diseases are driven by effects across the allele frequency spectrum that are at least partially additive (68, 113, 139) and even partially colocalize in genomic regions (40). In some domains, people harboring large-effect variants may phenotypically resemble controls due to other protective factors—often young age—but even in these cases, it can be medically useful to identify these at-risk patients for preventive treatment and monitoring (30). Historically, such diseases have been subdivided into familial and sporadic cases based on family history, because at the population level, Mendelian diseases more obviously evince heritability than polygenic diseases. However, this distinction is eroding for many traits as our understanding of complex genetic architecture advances (44, 45, 54, 65).

5. CONCLUSION

Modeling genetic subtypes of disease is biologically important for understanding basic disease architecture and medically important for optimizing treatment and prognosis. Historically, this is well known for rare, familial diseases, where identifying disorders (with or without knowledge of causal genes) has guided treatment and prognosis for decades or centuries. Similar concepts apply to polygenic subtypes, but the problem is statistically and conceptually more challenging because all people will carry a variety of risk factors even when strong subtypes are present. We have discussed several approaches to uncover these complex genetic subtypes, which extend two families of approaches from Mendelian disease: aggregating patients based on clinical phenotypic presentation and aggregating them based on causal genes. We have also discussed the downstream statistical tests that are needed to validate putative subtypes, focusing particularly on common pitfalls that result from imprecisely defining subtypes, ignoring confounding by population structure, overfitting, and the special models needed for case-only subtypes. Carefully identifying and testing subtypes offers the opportunity to make genuine, lasting discoveries that improve nosology and treatment.

The field of polygenic subtyping remains nascent, and there are several obvious next steps. Methodologically, TreeWAS, Subtest, and RGWAS are novel but young and unrefined approaches that would benefit from extensions. Furthermore, there are few tools to perform subtyping based on genomic features in complex traits, and using PRSs constructed from SNPs in specific genomic annotations seems a natural path forward. Perhaps more important, however, is identifying common diseases that likely harbor subtypes and, especially, the features that are likely to distinguish biomedically important subtypes. Psychiatry presents an obvious opportunity to define subtypes because phenotypes are difficult to measure and spectral, resulting in large genetic correlations among diseases (12), high misdiagnosis rates, and significant variation in clinical diagnostic standards both historically and recently. Furthermore, metabolic traits may have interesting subtypes given long-standing evolutionary trade-offs among metabolic programs that are optimal in different environments. Finally, drug development may benefit from evaluating polygenic subtype-specific efficacy and side effects, as patients with a common disease may respond very differently if their diseases are driven by fundamentally different genomic processes. For example, statin is an important drug for lowering LDL cholesterol levels and provides established cardiovascular benefits, but it causes myopathy and likely increases type 2 diabetes risk (105, 111). Identifying patient subtypes without these side effects would be enormously beneficial but has been historically difficult; nonetheless, recent evidence suggests that genetic subtypes may be used to suggest novel predictors and mechanisms for myopathy (78) and/or type 2 diabetes risk (24).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Adeponle AB, Thombs BD, Groleau D, Jarvis E, Kirmayer LJ. 2012. Using the cultural formulation to resolve uncertainty in diagnoses of psychosis among ethnoculturally diverse patients. *Psychiatr. Serv.* 63:147–53
2. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, et al. 2018. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* 6:361–69

3. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. 2001. Limitations of the case-only design for identifying gene-environment interactions. *Am. J. Epidemiol.* 154:687–93
4. Arnedo J, Svrakic DM, del Val C, Romero-Zaliz R, Hernández-Cuervo H, et al. 2015. Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. *Am. J. Psychiatry* 172:139–53
5. Bansal V, Mitjans M, Burik CAP, Linnér RK, Okbay A, et al. 2018. Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia. *Nat. Commun.* 9:3078
6. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. 2012. Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection. *PNAS* 109:1204–9
7. Bentley A, Sung YJ, Brown MR, Winkler TW, Kraja AT, et al. 2019. Multi-ancestry genome-wide gene–smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nat. Genet.* 51:636–48
8. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, et al. 2019. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* 8:e39725
9. Bernier R, Golzio C, Xiong B, Stessman HAF, Coe BP, et al. 2014. Disruptive *CHD8* mutations define a subtype of autism early in development. *Cell* 158:263–76
10. Bønnelykke K, Ober C. 2016. Leveraging gene-environment interactions and endotypes for asthma gene discovery. *J. Allergy Clin. Immunol.* 137:667–79
11. Boyle EA, Li YI, Pritchard JK. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169:1177–86
12. Brainstorm Consort., Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, et al. 2018. Analysis of shared heritability in common disorders of the brain. *Science* 360:eaap8757
13. Broeks A, Schmidt MK, Sherman ME, Couch FJ, Hopper JL, et al. 2011. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. *Hum. Mol. Genet.* 20:3289–303
14. Bromet EJ, Kotov R, Fochtmann LJ, Carlson GA, Tanenberg-Karant M, et al. 2011. Diagnostic shifts during the decade following first admission for psychosis. *Am. J. Psychiatry* 168:1186–94
15. Brown AA, Buil A, Vinuela A, Lappalainen T, Zheng HF, et al. 2014. Genetic interactions affecting human gene expression identified by variance association mapping. *eLife* 3:e01381
16. Bruining H, Eijkemans MJC, Kas MJH, Curran SR, Vorstman JAS, Bolton PF. 2014. Behavioral signatures related to genetic disorders in autism. *Mol. Autism* 5:11
17. Buu MC, Sanders LM, Mayo JA, Milla CE, Wise PH. 2016. Assessing differences in mortality rates and risk factors between Hispanic and non-Hispanic patients with cystic fibrosis in California. *Chest* 149:380–89
18. Chatterjee N. 2004. A two-stage regression model for epidemiological studies with multivariate disease classification data. *J. Am. Stat. Assoc.* 99:127–38
19. Chatterjee N, Carroll RC. 2005. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika* 92:399–418
20. Chen P, Lin JJ, Lu CS, Ong CT, Hsieh PF, et al. 2011. Carbamazepine-induced toxic effects and HLA-B*1502 screening in Taiwan. *N. Engl. J. Med.* 364:1126–33
21. Cho JH, Feldman M. 2015. Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. *Nat. Med.* 21:730–38
22. Cortes A, Dendrou CA, Motyer A, Jostins L, Vukcevic D, et al. 2017. Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. *Nat. Genet.* 49:1311–18
23. Coulter C, Baker KK, Margolis RL. 2019. Specialized consultation for suspected recent-onset schizophrenia: diagnostic clarity and the distorting impact of anxiety and reported auditory hallucinations. *J. Psychiatr. Pract.* 25:76–81
24. Dahl A, Cai N, Ko A, Laakso M, Pajukanta P, et al. 2019. Reverse GWAS: using genetics to identify and model phenotypic subtypes. *PLOS Genet.* 15:e1008009
25. Dahl A, Guillemot V, Mefford J, Aschard H, Zaitlen N. 2019. Adjusting for principal components of molecular phenotypes induces replicating false positives. *Genetics* 211:1179–89

26. Dahl A, Iotchkova V, Baud A, Johansson Å, Gyllensten U, et al. 2016. A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* 48:466–72
27. Dahl A, Nguyen K, Cai N, Gandal MJ, Flint J, Zaitlen N. 2020. A robust method uncovers significant context-specific heritability in diverse complex traits. *Am. J. Hum. Genet.* 106:71–91
28. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, et al. 2011. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* 72:245–56
29. Dempster ER, Lerner IM. 1950. Heritability of threshold characters. *Genetics* 35:212–36
30. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, et al. 2016. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* 354:aaf6814
31. Drumm ML, Ziady AG, Davis PB. 2012. Genetic variation and clinical heterogeneity in cystic fibrosis. *Annu. Rev. Pathol.* 7:267–82
32. Dudbridge F, Fletcher O. 2014. Gene-environment dependence creates spurious gene-environment interaction. *Am. J. Hum. Genet.* 95:301–7
33. Earl RK, Turner TN, Mefford HC, Hudac CM, Gerdts J, et al. 2017. Clinical phenotype of ASD-associated *DYRK1A* haploinsufficiency. *Mol. Autism* 8:105
34. Engelhardt BE, Stephens M. 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLOS Genet.* 6:e1001117
35. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, et al. 2014. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Nature* 343:1246949
36. Falconer DS. 1967. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Ann. Hum. Genet.* 31:1–20
37. Ferrell PB, McLeod HL. 2008. Carbamazepine, *HLA-B*1502* and risk of Stevens–Johnson syndrome and toxic epidermal necrolysis: US FDA recommendations. *Pharmacogenomics* 9:1543–46
38. Flint J, Kendler KS. 2014. The genetics of major depression. *Neuron* 81:484–503
39. Folca P, Glascock R, Irvine W. 1961. Studies with tritium-labelled hexoestrol in advanced breast cancer: comparison of tissue accumulation of hexoestrol with response to bilateral adrenalectomy and oophorectomy. *Lancet* 278:796–98
40. Freund MK, Burch KS, Shi H, Mancuso N, Kichaev G, et al. 2018. Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. *Am. J. Human Genet.* 103:535–52
41. Gabai-Kapara E, Lahad A, Kaufman B, Friedman E, Segev S, et al. 2014. Population-based screening for breast and ovarian cancer risk due to *BRCA1* and *BRCA2*. *PNAS* 111:14205–10
42. Garcia-Closas M, Couch FJ, Lindström S, Michailidou K, Schmidt MK, et al. 2013. Genome-wide association studies identify four ER negative–specific breast cancer risk loci. *Nat. Genet.* 45:392–98
43. Geisheker MR, Heymann G, Wang T, Coe BP, Turner TN, et al. 2017. Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat. Neurosci.* 20:1043–51
44. Ghaleb Y, Elbitar S, El Khoury P, Bruckert E, Carreau V, et al. 2018. Usefulness of the genetic risk score to identify phenocopies in families with familial hypercholesterolemia? *Eur. J. Hum. Genet.* 26:570–78
45. Goldstein JL, Brown MS. 2009. The LDL receptor. *Arterioscler. Thromb. Vasc. Biol.* 29:431–38
46. Grath S, Parsch J. 2016. Sex-biased gene expression. *Annu. Rev. Genet.* 50:29–44
47. Greene WH. 2003. *Econometric Analysis*. New York: Pearson. 5th ed.
48. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, et al. 1990. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 250:1684–89
49. Han B, Diogo D, Eyre S, Kallberg H, Zhernakova A, et al. 2014. Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity. *Am. J. Hum. Genet.* 94:522–32
50. Han B, Pouget JG, Slowikowski K, Stahl E, Lee CH, et al. 2016. A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. *Nat. Genet.* 48:803–10

51. Hancock DB, Artigas MS, Gharib SA, Henry A, Manichaikul A, et al. 2012. Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLOS Genet.* 8:e1003098
52. Hanson E, Bernier R, Porche K, Jackson FI, Goin-Kochel RP, et al. 2015. The cognitive and behavioral phenotype of the 16p11.2 deletion in a clinically ascertained population. *Biol. Psychiatry* 77:785–93
53. Hinks TSC, Brown T, Lau LCK, Rupani H, Barber C, et al. 2016. Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and chitinase 3–like protein 1. *J. Allergy Clin. Immunol.* 138:61–75
54. Hoffmann TJ, Theusch E, Haldar T, Ranatunga DK, Jorgenson E, et al. 2018. A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* 50:401–13
55. Howrylak JA, Moll M, Weiss ST, Raby BA, Wu W, Xing EP. 2016. Gene expression profiling of asthma phenotypes demonstrates molecular signatures of atopy and asthma control. *J. Allergy Clin. Immunol.* 137:1390–97.e6
56. Huckins LM, Dobbyn A, Ruderfer DM, Hoffman G, Wang W, et al. 2019. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nat. Genet.* 51:659–74
57. Iqbal J, Ginsburg O, Rochon PA, Sun P, Narod SA. 2015. Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. *JAMA* 313:165–73
58. Jayadev S, Bird TD. 2013. Hereditary ataxias: overview. *Genet. Med.* 15:673–83
59. Jeste SS, Geschwind DH. 2014. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nat. Rev. Neurol.* 10:74–81
60. Jiang J, Li C, Paul D, Yang C, Zhao H. 2016. On high-dimensional misspecified mixed model analysis in genome-wide association study. *Ann. Stat.* 44:2127–60
61. Katayama Y, Nishiyama M, Shoji H, Ohkawa Y, Kawamura A, et al. 2016. CHD8 haploinsufficiency results in autistic-like phenotypes in mice. *Nature* 537:675–79
62. Kendler KS, Kessler RC, Walters EE, MacLean C, Neale MC, et al. 1995. Stressful life events, genetic liability, and onset of an episode of major depression in women. *Am. J. Psychiatry* 152:833–42
63. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50:1219–24
64. Khramtsova EA, Davis LK, Stranger BE. 2019. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* 20:173–90
65. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, et al. 2018. Genetics of blood lipids among ~300,000 multi-ethnic participants of the million veteran program. *Nat. Genet.* 50:1514–23
66. Klein SL, Flanagan KL. 2016. Sex differences in immune responses. *Nat. Rev. Immunol.* 16:626–38
67. Krishnan ML, Wang Z, Aljabar P, Ball G, Mirza G, et al. 2017. Machine learning shows association between genetic variability in PPARG and cerebral connectivity in preterm infants. *PNAS* 114:13744–49
68. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, et al. 2015. Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* 47:582–88
69. Laakso M. 2019. Biomarkers for type 2 diabetes. *Mol. Metab.* 27:S139–46
70. Lai MC, Lombardo MV, Chakrabarti B, Baron-Cohen S. 2013. Subgrouping the autism "spectrum": reflections on DSM-5. *PLOS Biol.* 11:e1001544
71. Laursen TM, Agerbo E, Pedersen CB. 2009. Bipolar disorder, schizoaffective disorder, and schizophrenia overlap: a new comorbidity index. *J. Clin. Psychiatry* 70:1432–38
72. Lee MN, Ye C, Villani AC, Raj T, Li W, et al. 2014. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* 343:1246980
73. Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genet.* 3:e161
74. Levey AS, Coresh J. 2012. Chronic kidney disease. *Lancet* 379:165–80
75. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, et al. 2015. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* 7:311ra174

76. Liley J, Todd JA, Wallace C. 2016. A method for identifying genetic heterogeneity within phenotypically defined disease subgroups. *Nat. Genet.* 49:310–16
77. Lynch T, Price A. 2007. The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am. Fam. Physician* 76:391–96
78. Mangravite LM, Engelhardt BE, Medina MW, Smith JD, Brown CD, et al. 2013. A statin-dependent QTL for *GATM* expression is associated with statin-induced myopathy. *Nature* 502:377–80
79. Manna S, Holz MK. 2016. Tamoxifen action in ER-negative breast cancer. *Signal Transduct. Insights* 5:1–7
80. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–53
81. Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11:499–511
82. Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* 44:243–46
83. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, et al. 2019. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Human Genet.* 104:21–34
84. McClellan J, King MC. 2010. Genetic heterogeneity in human disease. *Cell* 141:210–17
85. McCormack M, Alfirevic A, Bourgeois S, Farrell JJ, Kasperavičiūtė D, et al. 2011. HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. *N. Engl. J. Med.* 364:1134–43
86. McCullagh P, Nelder J. 1989. *Generalized Linear Models*. Boca Raton, FL: Chapman & Hall/CRC. 2nd ed.
87. Mega JL, Simon T, Collet JP, Anderson JL, Antman EM, et al. 2010. Reduced-function *CYP2C19* genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. *JAMA* 304:1821–30
88. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, et al. 2017. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* 49:1767–78
89. Mitra I, Tsang K, Ladd-Acosta C, Croen LA, Aldinger KA, et al. 2016. Pleiotropic mechanisms indicated for sex differences in autism. *PLOS Genet.* 12:e1006425
90. Moore R, Casale FP, Bonder MJ, Horta D, BIOS Consort., et al. 2019. A linear mixed-model approach to study multivariate gene-environment interactions. *Nat. Genet.* 51:180–86
91. Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, et al. 2010. A powerful approach to sub-phenotype analysis in population-based genetic association studies. *Genet. Epidemiol.* 34:335–43
92. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, et al. 2008. Identifying autism loci and genes by tracing recent shared ancestry. *Nature* 321:218–23
93. Mukherjee S, Shukla S, Woodle J, Rosen AM, Olarte S. 1983. Misdiagnosis of schizophrenia in bipolar patients: a multiethnic comparison. *Am. J. Psychiatry* 140:1571–74
94. Myers RA, Scott NM, Gauderman WJ, Qiu W, Mathias RA, et al. 2014. Genome-wide interaction studies reveal sex-specific asthma risk alleles. *Hum. Mol. Genet.* 23:5251–59
95. Nicolau M, Levine AJ, Carlsson G. 2011. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *PNAS* 108:7265–70
96. Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. 2009. Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. *Clin. Med. Res.* 7:4–13
97. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, et al. 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485:246–50
98. Paré G, Cook NR, Ridker PM, Chasman DI. 2010. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLOS Genet.* 6:e1000981
99. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, et al. 2013. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155:1008–21
100. Parikshak NN, Swarup V, Belgard TG, Irimia M, Ramaswami G, et al. 2016. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* 540:423–27

101. Patel CJ, Chen R, Kodama K, Ioannidis JPA, Butte AJ. 2013. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. *Hum. Genet.* 132:495–508
102. Peterson RE, Cai N, Dahl AW, Bigdeli TB, Edwards AC, et al. 2018. Molecular genetic analysis subdivided by adversity exposure suggests etiologic heterogeneity in major depression. *Am. J. Psychiatry* 175:545–54
103. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. 2008. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* 358:2796–803
104. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, et al. 2005. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N. Engl. J. Med.* 353:1659–72
105. Preiss D, Seshasai SRK, Welsh P, Murphy SA, Ho JE, et al. 2011. Risk of incident diabetes with intensive-dose compared with moderate-dose statin therapy: a meta-analysis. *JAMA* 305:2556–64
106. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–9
107. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. *Am. J. Hum. Genet.* 67:170–81
108. Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, et al. 2011. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* 72:257–68
109. Robinson MR, English G, Moser G, Lloyd-Jones LR, Triplett MA, et al. 2017. Genotype-covariate interaction effects and the heritability of adult body mass index. *Nat. Genet.* 49:1174–81
110. Ruderfer DM, Ripke S, McQuillin A, Boocock J, Stahl EA, et al. 2018. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell* 173:1705–15.e16
111. Sattar N, Preiss D, Murray HM, Welsh P, Buckley BM, et al. 2010. Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials. *Lancet* 375:735–42
112. Schmidt MK, Hogervorst F, van Hien R, Cornelissen S, Broeks A, et al. 2016. Age- and tumor subtype–specific breast cancer risk estimates for *CHEK2**1100delC carriers. *J. Clin. Oncol.* 34:2750–60
113. Schrode N, Ho SM, Yamamuro K, Dobbyn A, Huckins L, et al. 2019. Synergistic effects of common schizophrenia risk variants. *Nat. Genet.* 51:1475–85
114. Schulz JB, Pandolfo M. 2013. 150 years of Friedreich ataxia: from its discovery to therapy. *J. Neurochem.* 126:1–3
115. Shi H, Mancuso N, Spendlove S, Pasaniuc B. 2017. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* 101:737–51
116. Shorter E. 2015. The history of nosology and the rise of the *Diagnostic and Statistical Manual of Mental Disorders*. *Dialogues Clin. Neurosci.* 17:59–67
117. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, et al. 2015. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* 518:187–96
118. Skol AD, Sasaki MM, Onel K. 2016. The genetics of breast cancer risk in the post-genome era: thoughts on study design to move past *BRCA* and towards clinical relevance. *Breast Cancer Res.* 18:99
119. Sladek R. 2018. The many faces of diabetes: addressing heterogeneity of a complex disease. *Lancet Diabetes Endocrinol.* 383:1084–94
120. Small KS, Todorčević M, Civelek M, Moustafa JSES, Wang X, et al. 2018. Regulatory variants at *KLF14* influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat. Genet.* 50:572–80
121. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, et al. 2019. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* 8:e39702
122. Steinsaltz D, Dahl A, Wachter KW. 2018. Statistical properties of simple random-effects models for genetic heritability. *Electron. J. Stat.* 12:321–58
123. Stessman HAF, Bernier R, Eichler EE. 2014. A genotype-first approach to defining the subtypes of a complex disease. *Cell* 156:872–77
124. Stessman HAF, Turner TN, Eichler EE. 2016. Molecular subtyping and improved treatment of neurodevelopmental disease. *Genome Med.* 8:22

125. Stessman HAF, Xiong B, Coe BP, Wang T, Hoekzema K, et al. 2017. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat. Genet.* 49:515–26
126. Stewart C, Pepper MS. 2016. Cystic fibrosis on the African continent. *Genet. Med.* 18:653–62
127. Stewart C, Pepper MS. 2017. Cystic fibrosis in the African diaspora. *Ann. Am. Thorac. Soc.* 14:1–7
128. Stone G, Choi A, Meritxell O, Gorham J, Heydarpour M, et al. 2019. Sex differences in gene expression in response to ischemia in the human left ventricular myocardium. *Hum. Mol. Genet.* 28:1682–93
129. Sul JH, Bilow M, Yang WY, Kostem E, Furlotte N, et al. 2016. Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLOS Genet.* 12:e1005849
130. Sverdlov S, Thompson E. 2017. Combinatorial methods for epistasis and dominance. *J. Comput. Biol.* 24:267–79
131. Terao C, Brynedal B, Chen Z, Jiang X, Westerlind H, et al. 2019. Distinct HLA associations with rheumatoid arthritis subsets defined by serological subphenotype. *Am. J. Hum. Genet.* 105:616–24
132. Thyme SB, Pieper LM, Li EH, Pandey S, Wang Y, et al. 2019. Phenotypic landscape of schizophrenia-associated genes defines candidates and their shared functions. *Cell* 177:478–91.e20
133. Udler MS, Kim J, von Grotthuss M, Bonàs-Guarch S, Cole JB, et al. 2018. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. *PLOS Med.* 15:e1002654
134. Uricchio LH, Kitano HC, Gusev A, Zaitlen NA. 2019. An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evol. Lett.* 3:69–79
135. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530–36
136. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, et al. 2011. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474:380–84
137. Wang J, Zhao Q, Hastie T, Owen AB. 2017. Confounder adjustment in multiple hypothesis testing. *Ann. Stat.* 45:1863–94
138. Wang L, Liang R, Zhou T, Zheng J, Liang BM, et al. 2017. Identification and validation of asthma phenotypes in Chinese population using cluster analysis. *Ann. Allergy Asthma Immunol.* 119:324–32
139. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, et al. 2017. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.* 49:978–85
140. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, et al. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* 358:667–75
141. Woodruff PG, Modrek B, Choy DF, Jia G, Abbas AR, et al. 2009. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am. J. Respir. Crit. Care Med.* 180:388–95
142. Wray NR, Lee SH, Kendler KS. 2012. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. *Eur. J. Hum. Genet.* 20:668–74
143. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42:565–69
144. Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Human Genet.* 88:76–82
145. Young AI, Wauthier FL, Donnelly P. 2016. Multiple novel gene-by-environment interactions modify the effect of *FTO* variants on body mass index. *Nat. Commun.* 7:12724
146. Young AI, Wauthier FL, Donnelly P. 2018. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nat. Genet.* 50:1608–14
147. Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genom. Hum. Genet.* 10:451–81
148. Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS* 109:1193–98

Annual Review of Genomics and Human Genetics

Volume 21, 2020

Contents