Heritability jointly Explained by Host Genotype and Microbiome:Will Improve Traits Prediction?

Denis Awany [1], and Emile R. Chimusa [1,⋆]

[1]Division of Human Genetics, Department of Pathology, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa.

⋆ Correspondence to Emile R. Chimusa: emile.chimusa@uct.ac.za

Abstract

As we observe the 70^{th} anniversary of the publication by Robertson that formalized the notion of 'heritability', geneticists remain puzzled by the problem of missing/hidden heritability, where heritability estimates from genome-wide association studies (GWAS) fall short of that from twin-based studies. Many possible explanations have been offered for this discrepancy, including existence of genetic variants poorly captured by existing arrays, dominance, epistasis, and unaccounted-for environmental factors; albeit these remain controversial. We believe a substantial part of this problem could be solved or better understood by incorporating the host's microbiota information in the GWAS model for heritability estimation; ultimately also increasing human traits prediction for clinical utility. This is because, despite empirical observations such as (i) the intimate role of the microbiome in many complex human phenotypes, (ii) the overlap between genetic variants associated with both microbiome attributes and complex diseases, and (iii) the existence of heritable bacterial taxa, current GWAS models for heritability estimate do not take into account the contributory role of the microbiome. Furthermore, heritability estimate from twin-based studies does not discern microbiome component of the observed total phenotypic variance. Here, we summarize the concept of heritability in GWAS and microbiome-wide association studies (MWAS), focusing on its estimation, from a statistical genetics perspective. We then discuss a possible method to incorporate the microbiome in the estimation of heritability in host GWAS.

Keywords: Heritability, Missing Heritability, Host genetics, Microbiome, Genome-wide association study, Microbiome-wide association study.

1 Introduction

Over a century ago, Weinberg [1], cognizant of the fact that phenotypic variation results from a combination of genetic and environmental factors, suggested methods of delineating genetic from environmental components of total phenotypic variability. This and subsequent works on statistical separation of the environmental and genetic variation in general populations, culminated to the specification of the fraction of the total phenotypic variance due to the genetic variance - a measure eventually termed 'degree of heritability' or simply 'heritability' in the genetic community [2]. A distinction is, however, necessary between *total* (or *broad sense*) and *additive* (or *narrow sense*) heritability. The former measures the full contribution of genes, which includes additive, dominance and epistasis components, while the latter captures only the additive contribution of genes to phenotypic variance.

It is now known that many common human diseases and traits are complex, resulting from the joint effect of host genetic and environmental factors. Indeed, genome-wide association studies (GWAS), which assays hundreds to millions of genetic markers - commonly single nucleotide polymorphisms (SNPs) - in thousands of individuals, have uncovered hundreds of genetic variants associated with many common polygenic inherited diseases and traits; revealing scores of previously unknown key biological pathways, and providing valuable insights into the complexities of their genetic architecture [3–5]. Despite this, however, GWAS has been puzzled by the apparent rather low proportion of the estimated heritability, which is far less than that obtained from familial studies - the difference being referred to as *missing/hidden* heritability. A classic, often cited, example is the human height where whereas the estimated heritability is 80%, the (narrow sense) heritability estimate with tens of thousands of people is only about 5% [6]. Many possible, and debated, explanations have been offered for this discrepancy, including sub-optimal sample size, poor detection of variants by genotyping arrays, dominance, epistasis, and shared environment [see references [4, 6] for excellent reviews]. While many investigators have argued that a considerable part of the missing/hidden heritability may be attributed to non-additive effects such as dominance and epistasis, several recent empirical studies have found no strong effects from them [7–9]. A similar observation has been made for epigenetic effects. While epigenetic variation, including methylation that has been suggested as another possible source of missing/hidden heritability, a recent study of body mass index found genetic predictors and methylation to be non-overlapping, suggesting the latter represented the environmental effects on this phenotype; for human height, methylation profiles did not explain any variation [9, 10]. Although

the volume and scope of these studies are certainly not optimal, and hence the conclusions may not be entirely generalizable, they do corroborate the significant contribution of non-genetic factors to inflating heritability estimates from GWAS.

On the other hand, in parallel to GWAS, Microbiome-wide Association Studies (MWAS), have been successful in identifying bacterial taxa that are associated with a variety of conditions, such as obesity, major depression, colorectal cancer, and inflammatory bowel disease [11]. Interesting, however, recent twin-based studies have reported the heritability of the human microbiome. For example, in the largest twin cohort to date, Goodrich et. al. (2014) [12], using the gut microbiome samples, found a number of bacterial families to be heritable, with *Christensenellaceae* and having the highest heritability ($h^2 = 0.39$). These interesting findings have raised enthusiasm, in as much as questions, among researchers on the implication of the microbiome on human health and the degree to which the human genotype versus the microbiome and the environment determines phenotypic variability.

Although GWAS and MWAS have been viewed as parallel fields, it has become increasingly apparent that time is ripe to shift away from the unidirectional host-centric and microbiome-centric interpretation to a more comprehensive view in which both host genetic and microbiome are considered as integral unit in analysis of phenotypic variability. The main reason for this is that if everything external to the human host is defined as the "environment", then the environment in this case is another living organism. From ecological viewpoint, fluxes between biotic and abiotic components in an environment relies almost entirely on the abilities of the biotic components to extract and use the abiotic [13]. This is, however, not the case for the host's microbiota, where this exchange is highly regulated by the host, for example through immune system [13, 14] and metabolic pathways [15, 16]. This associative trajectory, involving the host and microbes together with their collective genomes, greatly influences host biochemistry [16]; the ultimate result of which is the modulation of the host's phenotypic expression. Thus, whether or not the host's genome and microbial components both explain the same phenotypes, it is clear, from a statistical genetics perspective, that inclusion of both would improve statistical power to detect truly associated causative variants.

Apart from enabling the detection of causal variants, this comprehensive view, as also pointed out by other authors [13, 17], has the potential to narrow or provide insights into the missing/hidden heritability gap in GWAS for two reasons. First, while, by definition, heritability measures phenotypic variance attributable to genetic variance, GWAS only take into account the genetic variance in human cells and does not consider all the contributory role of the microbiome on the phenotype [13, 17]. Second, as a benchmark, the heritability estimates from familial studies, in which identity is inferred by kinship, are inflated because the observed phenotype is the resultant effect of the host's genotype and microbiota (and of course, in addition to other external factors) [17]. Therefore, incorporating the host's microbiota information with the genotypes will likely improve the estimates of contributors to

heritability, and eventually facilitate the determination of either additional genetic variants to explain further proportions of heritability or the proportion of genetic variance that is already explainable by the already known variants. This add will additional anable the improvement of polygenic risk score for potential clinical utility.

In this review, we discuss the prospects for unifying the estimate of heritability expalned from GWAS and MWAS in uncovering the host's genetic basis of human phenotypes. We focus on heritability estimation, from a statistical genetics perspective, and summarize the methodological approach to estimate (narrow sense) heritability. Finally, we suggest a method to incorporate the microbiome in the classical estimation of the heritability in human genome-wide association studies and conclude with a discussion of research areas where further work on both heritabilty and human traits prediction are needed.

2 Heritability in GWAS

Thousands of reports of GWAS mostly of European-ancestry encompassing larger samples, with some studies reaching up to million subjects have enabled the development of various heritability models to predict the genetic liability of human traits. Therefore, the clinical utility of the heritability has largely been explored in populations of European-ancestry and enabled applications in polygenic risk score and both genetics testing and counselling.

2.1 Broad-sense and narrow-sense heritability

Heritability is a measure of the relative contribution of genetics to a phenotypic expression. Its estimation centers around the measure of variability, which makes sense only if the phenotype is quantitative. For categorical phenotypes, therefore, one typically postulates it to be resulting from some underlying quantitative (continuous) variable, often called *liability*, which has a *threshold* that defines the intervals corresponding to the different states of the categorical variable [18]. The basic idea of heritability estimation is simple, at least in theory: partition the variance of a phenotype into components attributable to the different factors that are known to affect the phenotype, and determine the ratio of the genetic variance component (assuming genetics modulates the phenotype) to the phenotypic variance.

Suppose a quantitative trait is modulated by its overall genotype G and the exposure environment E, where G can be partitioned into additive (σ^2_g), dominance (σ^2_D), and interaction (σ^2_I) components.

That is,

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 = \sigma_g^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2\,,$$

where σ_I^2 may be refer to additive-by-additive, dominance-by-dominance, additive-by-dominance as well as many higher-order interaction terms [19]. Broad sense heritability H^2 is the ratio of the total genetic variance to the phenotypic variance: $H^2 = \frac{\sigma_G^2}{\sigma_P^2}$, and expresses the degree to which genotype determine phenotype of individuals. Narrow sense heritability h^2, on the other hand, is the ratio of the total genetic variance to the phenotypic variance: $h^2 = \frac{\sigma_g^2}{\sigma_P^2}$, and expresses the extent to which individual's phenotypes are determined by genes transmitted from parents. It therefore determines the degree of phenotypic resemblance between relatives, the observable genetic properties of a population and of the response of a population to external forces such as selection [20]. Broad sense heritability is of more theoretical interest than practical importance as it neither provide an understanding of the genetic properties of a population nor reveal the cause of phenotypic resemblance between relatives. Henceforth, as with all GWAS studies, we refer to narrow sense heritability (or simply heritability) in all subsequent discussions, unless stated otherwise.

2.2 Estimation of heritability in GWAS

We are typically interested in the *explained* heritability - defined as the ratio of heritability explained by a set of variants known definitely to be associated with the trait to the heritability explained by all genetic variants [those known (discovered) plus those not yet discovered] that are associated with the trait: $h^2_{\text{explained}} = h^2_{\text{known}}/h^2_{\text{all}}$; missing heritability being defined as $h^2_{\text{missing}} = 1 - h^2_{\text{explained}}$ [4].

From a statistical genetics perspective, the additive variance σ_g^2 at a single locus is the genetic variance explained by the regression of the expected value of the phenotypic mean in each genotypic class on the genotype [7, 21], or put differently, heritability is the coefficient of the regression obtained from the regression of additive genetic effect on the phenotype. Therefore, as we detail below, h^2_{known} can be readily estimated from observed genotype-phenotype data using regression in a 'bottom-up' approach. The estimation of h^2_{all} is, however, not straight forward because we do not know the complete repertoire of genetic variants associated with the trait; all we can do therefore is to infer it from phenotypic correlations obtained from population data in a 'top-down' approach.

Given a GWAS, let y_i, $i \in \{1, \cdots, n\}$ be the quantitative phenotypes measured on n individuals; $g_i = \{g_{i1}, \cdots, g_{im}\}$ the genotype of the i^{th} individual for the m typed SNPs, with minor allele frequencies p_j, $j \in \{1, \cdots, m\}$.

Employing the additive model, we have

$$y_i = \sum_{j=1}^{m} \beta_j z_{ij} + e_i \,,$$

where $z_{ij} = (g_{ij} - 2p_j)/\sqrt{2p_j(1-p_j)}$ is the normalized genotype and y is normalized phenotype, having mean 0 and variance 1.

If S is the subset of statistically-associated (assumed here to be causal) variants obtained from the GWAS, then, as the phenotype is normalized (i.e. $\sigma_P^2 = 1$), the additive variance is $\sigma_g^2 = \sum_{j \in S} \beta_j^2$ and the heritability is $h^2_{\text{known}} = h^2_{\text{GWAS}} = \sigma_g^2/\sigma_P^2 = \sum_{j \in S} \beta_j^2$, the sum of squared effect sizes for the normalized genotypes over the causal variants [4].

It is important to note, however, that the full set of causal variants are unknown; that is, the causal variants identified by the GWAS here is only a subset of causal variants. Consequently, h^2_{GWAS} represents only the lower bound of the true heritability h^2_{all}. The difference between h^2_{GWAS} and h^2_{all} is termed the *missing* heritability of the phenotype. This difference can be attributed to several factors. First, the non-additive genetic variance such as epistasis that is not included in estimation of heritability; the presence of such non-additive variations have been shown to inflate heritability estimates [4, 22]. Second, the exclusion of causal variants due to, say stringent GWAS significance threshold or low effect sizes, for example, can lead to underestimation of the heritability. Likewise, false positive results would inflate observed estimates.

In practice, having estimated h^2_{GWAS}, it is often of interest to know the proportion of the explained heritability. In other words, we would like to answer the following question: *what proportion of the heritability do all SNPs that contribute to the trait explain?* This question requires us to estimate h^2_{all}.

The methodological estimation of h^2_{all} is a 'top-down' approach that hinges on recognizing the equivalence between then classical definition of heritability [$h^2 = \sigma_g^2/\sigma_P^2$] and the intuitive interpretation of the proportion of phenotypic variance explained by all causal variants [$\sigma_g^2/(\sigma_P^2 = \sigma_g^2 + \sigma_\epsilon^2)$].

If

$$\boldsymbol{y} = w + g + \boldsymbol{e} \,, \tag{1}$$

where w denote the fixed effects (including candidate SNP and optional covariates), g denote genetic effects assumed to subsume any genetic effects on the trait other than at the candidate SNP, $\boldsymbol{\epsilon} = \text{N}(0, I\sigma_\epsilon^2)$; I being the identity matrix, then by treating $g = X\boldsymbol{\beta}$ as a random effect with $g \sim \text{N}(0, \sigma_g^2 A)$,

the variance-covariance matrix of $\boldsymbol{y}$ can be expressed as

$$\text{Cov}(\boldsymbol{y}) = \frac{XX'}{m}\sigma_g^2 + I\sigma_\epsilon^2 = A\sigma_g^2 + I\sigma_\epsilon^2\,, \tag{2}$$

where A is the kinship (genetic relationship) matrix between pairs of individuals, defined over all causal loci [23], m is the number of causal variants, and σ_g^2 is the variance explained by the SNPs. Since A is defined over *all causal loci*, σ_g^2 denotes the *variance explained by all causal SNPs*.

The parameters to be estimated are σ_g^2 and σ_ϵ^2, which can be done by using (2) in (1) and obtaining parameters optimization using the restricted maximum likelihood (REML) method. With these obtained, $\sigma_P^2 = \sigma_g^2 + \sigma_\epsilon^2$ can subsequently be derived. In theory, this appears a trivial task. In practice, however, this is not the case because the estimation requires that we first obtain the matrix A, defined at the causal SNPs, but we do not know the causal SNPs. The traditional, and still used, approach involves using genetically related individuals from known pedigrees (family/twins) to estimate a kinship coefficient Φ, where, for example, Φ_{ij} is taken as 0.25 for siblings and 0.5 for twins; A is then taken to be equal to 2Φ [17, 24]. Clearly, this assumption does not necessarily hold since phenotypic resemblance may be influenced by other heritable factors, other than genotype; for example epigenetic modifications and the host's microbiome. Indeed, the true covariance has been observed to vary around this assumed value [25]. Consequently, this can lead to inflation of the corresponding estimated heritability.

With the unveiling of genotype data, prodded by the advent of next generation sequencing, methods have been devised to estimate A from genotype data of unrelated individuals. This is done [26, 27] by postulating that the ungenotyped causal SNPs are tagged by the genotyped ones, and therefore although the set of causal variants is unknown, one can use all the SNPs genotyped in GWAS to estimate A, and use it as proxy for A_{causal}. The key point to note here is that all genome-wide SNPs are used, not only the genome-wide significant SNPs. Indeed, this methodology proceeds without conducting any test of association between individual SNPs and the phenotype. The genetic relationship between individuals i and j is estimated using standardized genotype by

$$A_{ij} = \frac{1}{m}\sum_{k=1}^{m}\frac{(g_{ik} - 2p_k)(g_{jk} - 2p_k)}{2p_k(1-p_k)}\,,$$

where p_k is the allele frequency of the k^{th} SNP. With A, defined at all causal SNPs, now obtained, (1) can be solved to estimate σ_g^2 and σ_ϵ^2, from where the variance heritability explained by all causal SNPs (h^2_{all}) can be determined from $h^2_{\text{all}} = \sigma_g^2/(\sigma_g^2 + \sigma_\epsilon^2)$.

In the estimation of both h^2_{all} and h^2_{GWAS} above, it is assumed that all causal SNPs have been genotyped or at least, are tagged by the genotyped SNPs. Accordingly, if the SNP array used does not fully cover

the set of common genetic variants in the GWA study population, the resulting heritability estimates are likely to be smaller than the actual value, however large the sample size maybe.

Finally, it must be noted because the set of causal variants are unknown and we have to rely on SNPs being tagged by Linkage Disequilibrium (LD), and yet LD strength declines with increasing difference in minor allele frequency (MAF) between SNPs [28, 29], some causal variants in the low frequency spectrum may not be tagged; and, as a result will not play part in heritability estimates. Moreover, MAF of a disease allele can be population-specific [30, 31]. Because of this, the heritability estimate can be population specific. Indeed, it has recently been shown, for admixed population, that the narrow-sense heritability vary according to the local ancestry of the study population [32].

We point out that the general Linear Mixed effect Model (LMM) in Eq.(1) remains the fundamental method for heritability estimation; albeit, several different variants of it have been proposed in a bid to improve performance or allow estimation in different contexts; for example, to address environmental variations across samples [33], to refine the model for categorical traits [34], or to perform estimation in context-specific scenarios (such as genotype-environment, and genotype-sex contexts) [35]. Besides the REML-based methods implemented in LMMs, regression of phenotype correlations on genotype correlations (LDSC) and regression of phenotype correlations on genotype correlations (PCGC) are the other broad categories of statistical frameworks for heritability estimation (see **Box 2**). While each general method has its inherent strength and limitation, it is important to highlight the limitations to avoid potential pitfalls when applying a particular method. Currently implemented REML-based methods underestimates heritability in case-control studies [36–38], possibly due to due to case-control ascertainment biases [38]. LDSC methods can produce biased estimates in the presence of binary covariates with strong effects [36]. Meanwhile, the PCGC methods, although effective for case-control studies, can suffer from loss when there is ascertainment bias or when the genetic correlation is not constant (that is, inhomogeneous) across the allelic frequency spectrum at the risk loci. These being said, all methods suffer power loss when applied to cohorts from ancestrally divergent populations [36].

2.3 *cis* and *trans* heritability of transcriptional regulation

While the pursuit for genetic variants underlying complex human diseases continues, results from the decade-long GWAS showed that over 90% of disease-associated variants lie in non-protein coding regions of the genome, for example in promoter regions, enhancers, and structural elements [39–43]. Since these non-coding DNA elements does bind proteins and RNA molecules which cooperate to regulate the function and expression of protein-coding genes [44], a prevailing hypothesis that human genetic variants impact traits via regulation of gene expression levels [40, 45–48]. This has motivated

expression quantitative trait loci (eQTL) studies using genome-wide gene expression and genotype data, to explore the genetic basis of variability in gene expression; serving to potentially illuminate the bridge between statistical association and biological mechanism of a genetic variant on a phenotype.

To this end, quantifying the heritability of gene expression is central to understanding its genetic basis and, ultimately, its contribution to host phenotypic diversity. Gene expression is known to be controlled by both *cis* eQTLs (defined as eQTLs located close, say within 250 kb - 1 Mb, to the gene it regulates) and *trans* eQTLs (defined as eQTLs located far, outside the cut-off distance, from the gene it regulates) [49], and therefore, when studying the heritability of gene expression, it is of biological interest [50] to express heritability in terms of *cis* heritability, h^2_{cis}, (heritability due to genetic component close to the regulated gene) and *trans* heritability, h^2_{trans}, (heritability due to genetic component far from the regulated gene); so that the total heritability of a gene expression, h^2_{expr}, is given by $h^2_{\text{expr}} = h^2_{\text{cis}} + h^2_{\text{trans}}$. That said, *cis*-variation is often considered the primary driver of phenotypic variation; albeit, it is also more difficult to detect *trans*-acting eQTLs due to limitations in statistical power as their effect sizes are small [51, 52].

If we take the 250 kb 'cis window', for example, then the h^2_{cis} (narrow-sense) would be formally defined as the proportion of 'gene expression phenotype' explained by the additive effect of SNPs in a 250-kb window of the gene, whereas h^2_{trans} (narrow-sense) would be the proportion of 'gene expression phenotype' explained by the additive effects of SNPs outside a 250-kb window of the gene [53]. Methodologically, similar to the heritability estimation for complex traits, the heritability of gene expression can be estimated by two general approaches. The first involves using genetically-related individuals in the classical twin-based study design [54], where identity-by-descent (IBD) sharing across the genome is assumed to be 0.5 and 1 for dizygotic and monozygotic twins, respectively. The second alternative uses unrelated individuals where SNPs (measured plus tagged) within the *cis* window are used to define genetic-relatedness among individuals, and the estimation proceeds using the linear random effects model [27]; similar to classic heritability estimation of human complex traits. As the number of SNPs considered within *cis* window is small, h^2_{trans} can be estimated with high precision [49, 55].

A number of recent studies have estimated heritability of gene expression across different human tissues. In a total of 856 female twins recruited from the TwinsUK resource, Grundberg and others [51] estimated h^2_{cis} of gene expression for adipose, lymphoblastoid cell lines (LCLs) and skin tissues; obtaining, respectively, 26%, 21% and 16%. Importantly, having also estimated h^2_{trans}, they reported that h^2_{cis} constituted between 30-36% of the total heritability, but up to 40% of h^2_{cis} is missed when only common SNPs (MAF $> 5\%$) are used in *cis* eQTL mapping. The important implication of this finding for host GWAS are that low frequency and rare variants may account for a substantial proportion of the unexplained *cis* heritability (for transcriptional regulation) and h^2_{GWAS} (for complex human

traits), and that the action of host genetic polymorphisms on human diseases may be mediated by gene regulation, as the estimated h^2_{expr} is enriched in genes previously identified via GWAS in a broad range of diseases. In another recent study to characterize the genetic basis of human gene expression [53], narrow-sense *cis* heritability of *LCL* gene expression was estimated to be approximately 8.2%. The authors found that singletons accounted for the vast majority (25% compared with al other MAF bins) of this heritability, and over 90% of this was due to alleles of ultra-low frequencies (MAF$<$ 0.01%). Taken together, these findings suggest much of the missing (or unexplained) heritability of complex traits may be due to variants in the low-frequency spectrum, and transcriptional regulation represent at least one intermediary bridge between host genotype and phenotype.

3 Heritability in MWAS

Although the role of microbes in health and physiology has been known for over a century, only recently have the roles of these microbes together with their collective genome - the microbiome - in the pathogenesis of many common human diseases and traits become apparent, through microbiome-wide association studies (MWAS). MWAS in which the compositional and functional diversity of the microbiome is assessed at various taxonomic ranks (e.g species or genus level) in tens or hundreds individuals, represent a powerful new tool for investigating the microbiome basis of complex traits and diseases. To date, these studies have identified several microbiome-disease/trait associations [56]. The success of GWAS provided an optimistic outlook for MWAS and the observation of host genotype-microbiome interaction led to works on the heritability of the human microbiome.

Recent studies have shown that the microbiota is vertically transmitted from mother to offspring; albeit, the role, importance, and transmission mode of prenatal microbial colonization are still unclear [57, 57–59]. However, extensive colonization begins postpartum [57, 60]. Vertical transmission via breast milk, and horizontal transmission through factors such as mode of delivery (vaginal or caesarean section), feeding method (formula or breastfeeding), and social interactions are among the crucial factors in the development of the infant microbiome [57, 60, 61]. The transmission of the microbiota across humans is corroborated by the congruence of phylogenetic tree of intestinal bacterial microbiota and humans [62, 63]. Since microbial information can be transferred to offsprings and microbes have co-evolved with their human host for millions of years, it is reasonable to expect the former to hold information on latter's phenotypic plasticity [62].

To estimate heritability of the microbiome, one can apply the standard Additive Genetics, Common Environment, Unique Environment (ACE) model, treating the abundance of each human-associated microbe as a quantitative trait. Heritability in then is estimated by determining variation in micro-

bial taxon abundances (as measured by within-community alpha diversity measures such as observed species and Shannon diversity or by between-communities beta diversity measures such as UniFrac and Bray-Curtis metrics) that is attributable to human genetics. To date, twin studies invoking Falconer's formula

$$h^2 = 2(r_{mz} - r_{dz}),$$

where r_m and r_d are the correlation between pairs of mono-zygotic and di-zygotic twins respectively, has been the basis of heritability calculation [12, 64]. These studies have provided clues into the nature and extent of host-microbiome association: bacterial taxa observed to be consistently heritable include *Christensenellaceae*, *Actinobacteria*, *Firmicutes*, and *Tenericutes*, while *Bacteroidetes* phylum were generally not heritable [12, 64]. It is important to note, however, that the volume of such research is still small and further data will lend insight into this link.

4 Incorporating the microbiome in heritability estimation

It is now known that the majority of the common complex phenotypes are the result of the contributory role of host genetics, the microbiome, and other environmental factors. How these components do combine to determine phenotypic expression is certainly unknown. The simplest model is to assume either the contribution of the environment and the genetic variants that act additively or the environment and the additive effect of the microbiome (see *Box 1*). However, multiple lines of evidence suggest that host genetics and the microbiome do not act independently to shape observed phenotype [56, 65]. Indeed, several recent studies have reported weak effect of host genetics on the microbiome, both host genetics and microbiome have been independently implicated in the etiology of the same diseases/traits. Therefore, as also noted by [17], it would be useful to integrate host genetics and the microbiome in the same analytical model. The classical definition of heritability is limited. Although host genetics do contribute to phenotypic expression, the definition is based on the premise that host genetics and the environment as an integral component do have a contributory role on the phenotype. In light of host-microbiome symbiosis, it is pertinent that the host genetics and microbiome be viewed as a single unit representing 'host community genotype'. Indeed, the shift towards the view of organisms as an ecosystems has been advocated [17] (see Figure 1).

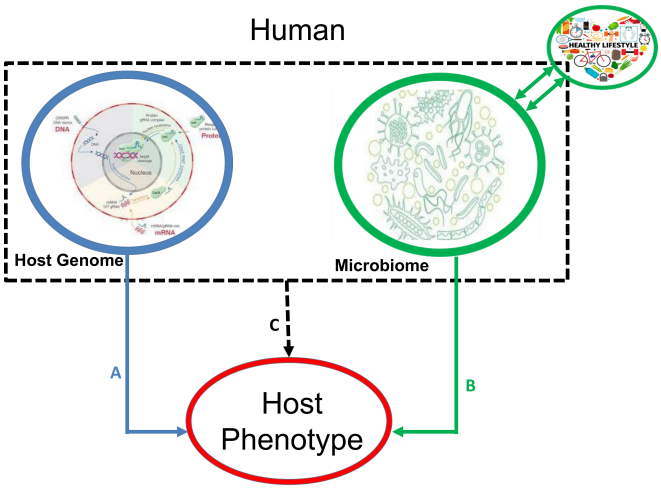

Figure 1: Conceptualization of host genetics and microbiome in heritability estimation for host phenotype. A: genetic heritability from host GWAS; B: microbiome heritability from MWAS; and C: heritability jointly explained by host genetics and microbiome.

In this community view, the microbiome can be integrated at various levels; for example, species, transcripts, metabolites, proteins, genes or their functional diversity. More generally, assume the phenotype (y) is contributory role of host genetics (g), the microbiome (b), and the environment (e). The model becomes

$$y = g + b + e \ , \tag{3}$$

where $g = Z\beta$, $b = W\alpha$, $g_i \sim N(0, \sigma_g^2)$, and $b_i \sim N(0, \sigma_b^2)$. Whilst it would be possible to model interactions effects, we concentrate on the main effects part of the model, partly for simplicity of exposition as it is very difficult to identify interactions terms with a reasonable accuracy in such high dimensional setting [66–68] and partly because the first order of Taylor expansion of the model function can accurately approximate interaction effects, which are essentially encoded in lower order terms [68]. The phenotypic variance-covariance matrix is expressed as

$$\text{Cov}(y) = \frac{ZZ'\sigma_g^2}{m} + \frac{WW'\sigma_b^2}{l} + I\sigma_e^2 = A\sigma_g^2 + B\sigma_b^2 + I\sigma_e^2 \, ,$$

where A is a host genetic relationship matrix defined on the causal variants, B is the the microbial taxa similarity matrix defined over the associated taxa, σ_g^2 is the total total additive genetic effect and σ_b^2 is the total total additive microbial effect, m and l is the total number of causal host genetic

variants and microbial taxa, respectively.

Adapting the classical GWAS methodology, A can be estimated using the all SNPs genotyped in the study, as described above. The microbiome similarity matrix can, however, be defined in two ways. First, phylogenetic distance measures, which accounts for the phylogenetic relationship among microbial taxa, could be used to define sample similarity matrix B. This approach, that has a solid foundation in the field of microbial ecology, would allow B to incorporate the degree of divergence between sequences, thereby estimating similarity among individuals based on phylogenetic relatedness of microbial communities in their bodies. This idea is further supported by the observation that host humans have co-evolved with their microbiome [13, 63, 69] and microbiome-related phenotypes can be transmitted between phylogenetically close humans [13]. Second, for each microbial specie, the abundance data can be discretized into 'categorie' such as 0,1,2 corresponding to low, medium, high abundance respectively, based on some biologically-plausible scale. The frequency of each category can be calculated based on population values. This is, however, not straightforward as it involve knowledge of community-composition of each microbial specie. In the absence of this information, the complete set of individuals in the study samples may be used as proxy for the population. Once this information has been obtained, the sample similarity between the individuals can be obtained as follows:

Consider model (3) and let R be an incidence matrix that maps different categories of microbial taxa to each subject. In the above case, the elements of B are 0, 1, or 2. Following the usual definition of Euclidean distance similarity, if we $K = RR'$ then the diagonals of K give the subject's relationship to itself while the off-diagonal elements gives the number of elements shared by the subjects. We are interested in investigating over-representation of microbial taxa. Accordingly, as in GWAS, it is possible to define a 'reference category' and determine the corresponding frequency. For our example, if we suppose category "2" is the reference category, then define F to be a matrix containing the frequencies of each category. The j^{th} column of F is $1f_j$, where f_j is the expected value of frequency of the reference category in the j^{th} taxon. With this, the matrix $(R - F)$ becomes the mean-centered form of R, which can essentially be interpreted as setting the mean value of taxa effects to zero. The microbial relationship matrix B is then calculated using

$$B = \frac{(R - F)(R - F)'}{\sum_{j=1}^{m} \mathrm{Var}(f_j)}.$$

The normalization by $\sum_{j=1}^{m} \mathrm{Var}(f_j)$ scales B in a way similar to the usual kinship matrix in GWAS.

Finally, the phenotypic variance explained by additive variation at all common SNPs, which we denote

by h^2_{geno}, is calculated from

$$h^2_{\text{geno}} = \sigma^2_g/(\sigma^2_g + \sigma^2_b + \sigma^2_e)\,.$$

h^2_{geno} is the narrow-sense host's genetic heritability of the trait, after accounting for the additive effects of host genetics and the microbiome. We call this 'geno' heritability.

Given that the microbiome co-evolved with its human host for millions of years, an emerging view is that of the 'holobiont' in which the human host and its microbiome is regarded as a single entity [13]. In light of this, one can define 'genobiome' heritability as the proportion of phenotypic variance due to both host genetics and microbiome variances. That is,

$$h^2_{\text{genobiome}} = \frac{\sigma^2_g + \sigma^2_b}{\sigma^2_g + \sigma^2_b + \sigma^2_e}\,. \tag{4}$$

The genobiome-heritability, so defined, is the narrow-sense host's genetic and microbiome heritability of the trait, and represents the heritability of the trait jointly explained by the host's genetics and microbiome.

5 Concluding Remarks

Predicting the heritability of human traits is one of the critical goals in biomedical research and precision medicine. Today, thousands of reports of GWAS, encompassing larger samples mostly generated European ancestry. These efforts enabled the development of various models of heritability to predict the genetic liability of human traits. However, current heritability models developed using large-scale European-ancestry genomic data still misestimate the predictive power of heritability of most human traits and, they additionally suffer power loss when applied to cohorts from ancestrally divergent populations [36]. The clinical utility of predictive heritability of traits is still in its infancy stage and its application in genetics testing and counselling in real-world clinical populations is limited. Due to the differences in disease/traits prevalence, linkage disequilibrium (LD), genetics ancestry, environmental factors, microbiome profiles, causal or marginal effect sizes and, epistatic or gene-environment interactions between populations, heritability of trait derived from GWAS of European-ancestry samples can potentially misestimate the predictive risk power when applied to non-European populations [10]. In addition, most of non-European populations such as Africans exhibit significantly higher risk allele frequencies, of which ancestral risk alleles is higher than derived risk alleles commonly observed to populations of European-ancestry [5, 9], therefore new heritability approaches that leverage population-specific characteristic including epigenetics, genetics ancestry, host-genetics interaction

with microbiomes are needed to improve the predictive power of the heritability of human traits.

The age of the microbiome is upon us, and the invaluable potential of the microbiome for host GWAS cannot not be overstated. Classical GWAS is based on the premise that the environment and disease are homogeneous among the study subjects. Insights gained from genetic and microbial epidemiological studies make it clear that this assumption does not generally hold, and can consequently reduce the power to detect truly associated causative variants. To this end, it is crucial that GWAS leverages the deterministic and stochastic factors that have known contributory role to phenotypic variability. In particular, given the association of host genetics and microbiome with the same phenotypes, the overlap of host genetic variants associated with the same traits, and the fact that the microbiome, unlike other abiotic environmental factors, is heritable and its variability has a genetic basis [70], it is pertinent that the microbiome be viewed as an integral part of the host rather than an external environmental factor. In this framework, the association mapping is performed on the host community, comprised of host genotype and its microbial community.

Moving forward, considering the additive effects of the microbiome in heritability calculation will be worthwhile as we seek to explain the 'dark matter' of missing/hidden heritability. Narrowing the missing/hidden heritability gap is of more than just an academic interest: knowing the heritability of a phenotype provides geneticists with the upper limit of the degree with which a phenotype can be predicted by identified variants. This is inevitable if we are to illuminate the dark path from genome-wide significant association to biological and medical application.

Beyond the missing heritability esoteric, the delineation of heritability in association mapping will be key in bridging the gap between statistical association and clinical translation in two broad ways. First, by quantifying the variance attributable to host genetics and microbiome, it will expand our understanding of complex disease architecture, which, ultimately, would guide design of experiments to fully dissect genetic and/or microbiome basis of disease aetiology. Second, is disease population risk stratification. With knowledge of the upper limit of risk stratification, disease risk models can be used to predict population-level risk of disease. The immediate benefit of this would be improved diagnosis, risk stratification, and disease management.

Key Points

1. Heritability estimates from human genome-wide association study (GWAS) and microbiome-wide association study (MWAS) provide, respectively, the extent of host genetics and microbiome contributions to host phenotype.
2. The involvement of the microbiome on host phenotypes makes it apparent that the microbiome be integrated with host genotype in host trait association mapping.
3. A substantial portion of the unexplained (aka missing) heritability in GWAS could be accounted for if the microbiome variation is taken into consideration.
4. In light of the holobiont theory, the narrow-sense heritability jointly explained by host genetics and microbiome can be determined.
5. The clinical utility of heritability estimates, which include traits prediction, disease risk stratification and characterization of disease architecture, necessitates its pursuit.

Acknowledgments

The author thank Delesa Damena and Imane Allali for helpful comments in an early draft of the paper. The authors thank CHPC (https://www.chpc.ac.za/) for providing computing facility.

Funding

The authors are supported in part by DAAD, the German Academic Exchange Programme, under funding reference # 91653117, and the National Institutes of Health Common Fund under grant number U41HG006941 and National Research Foundation of South Africa for funding (NRF) [grant # RA171111285157/119056]. The content of this article does not reflect the official opinion of the funders. Responsibility for the information and views expressed in the article lies entirely with the authors.

Box 1: Statistical Models for Heritability of Quantitative Human Traits

Quantitative genetics theory is traditionally developed for quantitative traits. Nevertheless, the theory can still be applied to a categorical trait by assuming it to be governed by some underlying quantitative latent variable, often called a liability, whose thresholds delimit the categories [18]. Let Y denote the random variable for the quantitative trait, and suppose G, M and E are the random variables for genotype, microbiome and environment respectively.

Model 1: Host's Phenotype is Influenced by the Host's Genotype and Environment.
If we consider Model 1 and assume additive genetic effects and ignore dominance and epistasis effects, then the overall phenotypic variance can be decomposed as

$$Var(Y) = Var(G) + Var(E) + 2Cov(G, E)\,, \tag{5}$$

where $Var(\cdot)$ is the variance of $(\cdot)$, and $Cov(\cdot, \cdot)$ is the covariance of $(\cdot, \cdot)$.

In practice in heritability estimation, it is implicitly assumed that G and E are independent, that is, $Cov(\cdot, \cdot) = 0$. The narrow-sense heritability of the host's phenotype, h^2_g, is then defined by

$$h^2_g = \frac{Var(G)}{Var(G) + Var(E)}\,. \tag{6}$$

h^2_g is the proportion of phenotypic variance attributable to genetic variance in the host.

Model 2: Host's Phenotype is Influenced by the Host's Microbiome and Environment.
For this model, if we similarly assume additive effects of bacterial taxa on phenotype, and ignore between-taxa, and taxa-environment interactions, then, as above, the narrow-sense heritability of the host's phenotype, h^2_m, is then defined by

$$h^2_m = \frac{Var(M)}{Var(M) + Var(E)}\,. \tag{7}$$

Analogous to h^2_g, h^2_m is the proportion of phenotypic variance attributable to variability in the host's microbiome.

Model 3: Host's Phenotype is Influenced by the Host's Genotype, Microbiome and Environment.
A perhaps more realistic model would be as in Model 3. In this case, assuming additive effects of SNPs and microbiome, we have

$$Var(Y) = Var(G) + Var(M) + Var(E)\,, \tag{8}$$

where we have imposed the assumption that $Cov(G, E) = Cov(M, E) = Cov(G, M) = 0$.
There are two possible heritability measures of interest that can be defined from this model, each with a different interpretation. First, is the 'geno' heritability which is the proportion of phenotypic variance explained by host genetic variance. That is,

$$h^2_{\text{geno}} = \frac{Var(G)}{Var(G) + Var(M) + Var(E)}\,. \tag{9}$$

In other words, geno-heritability is the narrow-sense host's genetic heritability of the trait. As with the classical interpretation of heritability, h^2_{geno} is taken to be obtained after accounting for the factors known to modulate the phenotype; the factors, in this case, being host's genotype, microbiome and environment.

Second, in light of the holobiont theory of humans host and its microbiome [13], one can define 'genobiome' heritability as the proportion of phenotypic variance due to both host genetics and microbiome variances. That is,

$$h^2_{\text{genobiome}} = \frac{Var(G) + Var(M)}{Var(G) + Var(M) + Var(E)}\,. \tag{10}$$

Box 1 (continued): Statistical Models for Heritability of Quantitative Human Traits

Methodological Implementation.
Estimation of heritability can be performed by fitting a linear mixed model (LMM) or Haseman-Elston (H-E) regression. The LMM has been the standard tool for heritability estimation for various host quantitative traits. It has also recently been applied to heritability estimation for transcriptional expression traits, including gene expression, methylation level, and other molecular traits [71, 72]. In the standard LMM implementation, the phenotype of each individual is modeled as the sum of two sets of random effects; one based on the covariate(s) of interest (e.g individual's genotype, or microbiome) and one based on environmental factors (see Eq.(1) in Main text for more detail). The parameters of the model are typically then fitted by maximizing the restricted maximum likelihood (REML) of the data [27], from where the desired heritability can be calculated from the estimated variance parameters.

Alternatively, especially when the sample size is small, as often seen with transcriptional expression traits, the H-E regression [36, 73] may be opted for, given its robustness for small sample sizes [53]. The idea here is to regress the phenotypic covariance on the genotypic covariance so that the resulting effect sizes, which will actually be the variance components, can be used to obtain the heritability of the phenotype under consideration. Again, in the practical implementation of H-E regression with transcriptional data, the SNPs are usually partitioned into K disjoint subsets based on minor allele frequency (MAF). The overall heritability of the trait is then the sum of heritability due to SNPs from each partition; this estimation, after partitioning SNPs, has been shown to correct for over- or under-estimation of heritability [53, 74].

Typically, in the transcriptional expression trait, the phenotypic covariance (denote it by, say, Y) is taken to be the upper triangle of the outer product of quantile-normalized $\log_2$(FPKM) [FPKM; Fragments Per Kilobase of transcript per Million mapped reads], and genotypic covariance (denote it by, say, X) defined as the upper triangle of a genomic-relationship matrix generated from all SNPs in the partition. For the k^{th} partition, $X_k = \frac{G_k G_k'}{M_k}$, where G_k and M_k are, respectively, the standardized genotype and number of SNPs in the k^{th} partition. The mapping is then carried out with the usual linear regression; viz

$$Y \sim X_1 + X_2 + \cdots + X_K$$

In this regression, the effect size for the k^{th} SNP partition represents the genetic variance of that partition; that is, $\beta_k = \sigma_k^2$. Thus, the total genetic variance due to all SNPs is $\sigma_g^2 = \sum_{k=1}^{K} \sigma_k^2$. As the phenotype is normalized to unit variance, the (narrow-sense) heritability of the transcriptional expression trait, h^2, is then equal to σ_g^2.

Box 2: Common Tools to Estimate Heritabilty

Software Tool	**Statistical Model**	**Trait**	**Note**	**Link**
GCTA	Linear Mixed Model (LMM)	Quantitative	Most routinely used tool for complex traits. Has been applied to a diverse array of traits, including human complex traits, transcriptional expression traits, and microbiome data. Generally effective for large sample size.	[27]
GxEMM	LMM	Quantitative and binary	Specifically designed for estimation in context-specific scenarios, including gene-environment interaction. Can accommodate modest sample sizes.	[35]
OpenMx	Structural Equation Modeling (SEM)	Quantitative and binary	Specific to classic twin-based study. Current impplementation more appropriate for quantitative traits only.	[75]
GEAR	Phenotype correlations on genotype correlations regression	Quantitative and binary	The method originally proposed for linkage studies, achieves good performance with small samples.	[76]
MetaSex	LMM	Quantitative	Specifically designed to account for potential sex difference in genetic architectures.	[77]
StructLMM	Structured LMM	Quantitative	Purposely designed to model gene-environment interaction.	[78]
LDSC	LD score regression	Quantitative and binary	Estimating and partitions SNP heritability by functional annotations. Uses summary statistics as input data	[79]
SumHer	LMM	Quantitative and binary	Estimating and partitions SNP heritability by functional annotations. Uses summary statistics as input data. Key difference from LDSC is that it allows the user to specify the heritability model.	[80]
PCGC	Phenotype correlations on genotype correlations regression	Quantitative and binary	Performs between than GEAR in case-control studies.	[36]
LAMatrix	LMM	Quantitative	Estimates heritability by local ancestry, global ancestry, and degree of population differentiation at causal regulatory variants for gene-expression traits in admixed populations.	[81]

References

[1] W Weinberg. Über vererbungsgesetze beim menschen ii. *Z. indukt. Abstamm. Vererbungsl*, 2:276–330, 1909.

[2] Jay L Lush. Family merit and individual merit as bases for selection. part i. *The American Naturalist*, 81(799):241–261, 1947.

[3] Joel N Hirschhorn et al. Genomewide association studies–illuminating biologic pathways. *New England Journal of Medicine*, 360(17):1699, 2009.

[4] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.

[5] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[6] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747, 2009.

[7] Zhihong Zhu, Andrew Bakshi, Anna AE Vinkhuyzen, Gibran Hemani, Sang Hong Lee, Ilja M Nolte, Jana V van Vliet-Ostaptchouk, Harold Snieder, Tonu Esko, Lili Milani, et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. *The American Journal of Human Genetics*, 96(3):377–385, 2015.

[8] Noah Zaitlen, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L Price. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS genetics*, 9(5):e1003520, 2013.

[9] Ilja M Nolte, Peter J van der Most, Behrooz Z Alizadeh, Paul IW de Bakker, H Marike Boezen, Marcel Bruinenberg, Lude Franke, Pim van der Harst, Gerjan Navis, Dirkje S Postma, et al. Missing heritability: is the gap closing? an analysis of 32 complex traits in the lifelines cohort study. *European Journal of Human Genetics*, 25(7):877, 2017.

[10] Sonia Shah, Marc J Bonder, Riccardo E Marioni, Zhihong Zhu, Allan F McRae, Alexandra Zhernakova, Sarah E Harris, Dave Liewald, Anjali K Henders, Michael M Mendelson, et al. Improving phenotypic prediction by combining genetic and epigenetic associations. *The American Journal of Human Genetics*, 97(1):75–85, 2015.

[11] Jack A Gilbert, Robert A Quinn, Justine Debelius, Zhenjiang Z Xu, James Morton, Neha Garg, Janet K Jansson, Pieter C Dorrestein, and Rob Knight. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 535(7610):94, 2016.

[12] Julia K Goodrich, Jillian L Waters, Angela C Poole, Jessica L Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, William Van Treuren, Rob Knight, Jordana T Bell, et al. Human genetics shape the gut microbiome. *Cell*, 159(4):789–799, 2014.

[13] Santiago Sandoval-Motta, Maximino Aldana, and Alejandro Frank. Evolving ecosystems: Inheritance and selection in the light of the microbiome. *Archives of medical research*, 48(8):780–789, 2017.

[14] Jean-Christophe Simon, Julian R Marchesi, Christophe Mougel, and Marc-André Selosse. Host-microbiota interactions: from holobiont theory to analysis. *Microbiome*, 7(1):5, 2019.

[15] Lora V Hooper, Dan R Littman, and Andrew J Macpherson. Interactions between the microbiota and the immune system. *Science*, 336(6086):1268–1273, 2012.

[16] Jeremy K Nicholson, Elaine Holmes, James Kinross, Remy Burcelin, Glenn Gibson, Wei Jia, and Sven Pettersson. Host-gut microbiota metabolic interactions. *Science*, 336(6086):1262–1267, 2012.

[17] Santiago Sandoval-Motta, Maximino Aldana, Esperanza Martínez-Romero, and Alejandro Frank. The human microbiome and the missing heritability problem. *Frontiers in genetics*, 8:80, 2017.

[18] Douglas S Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of human genetics*, 29(1):51–76, 1965.

[19] Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998.

[20] Douglas Scott Falconer et al. Introduction to quantitative genetics. *Introduction to quantitative genetics.*, 1960.

[21] Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[22] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446, 2010.

[23] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[24] Kenneth Lange. *Mathematical and statistical methods for genetic analysis*. Springer Science & Business Media, 2003.

[25] Peter M Visscher, Sarah E Medland, Manuel AR Ferreira, Katherine I Morley, Gu Zhu, Belinda K Cornes, Grant W Montgomery, and Nicholas G Martin. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, 2(3):e41, 2006.

[26] Gustavo De Los Campos, Daniel Gianola, and David B Allison. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, 11(12):880, 2010.

[27] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[28] Naomi R Wray. Allele frequencies and the r 2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Research and Human Genetics*, 8(2):87–94, 2005.

[29] Jian Yang, Jian Zeng, Michael E Goddard, Naomi R Wray, and Peter M Visscher. Concepts, estimation and interpretation of snp-based heritability. *Nature genetics*, 49(9):1304, 2017.

[30] John P Klein and Melvin L Moeschberger. Statistics for biology and health. *Stat. Biol. Health, New York*, 27238, 1997.

[31] Ethan Linck and CJ Battey. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, 19(3):639–647, 2019.

[32] Noah Zaitlen, Bogdan Pasaniuc, Sriram Sankararaman, Gaurav Bhatia, Jianqi Zhang, Alexander Gusev, Taylor Young, Arti Tandon, Samuela Pollack, Bjarni J Vilhjálmsson, et al. Leveraging population admixture to characterize the heritability of complex traits. *Nature genetics*, 46(12):1356, 2014.

[33] David Heckerman, Deepti Gurdasani, Carl Kadie, Cristina Pomilla, Tommy Carstensen, Hilary Martin, Kenneth Ekoru, Rebecca N Nsubuga, Gerald Ssenyomo, Anatoli Kamali, et al. Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences*, 113(27):7377–7382, 2016.

[34] Shiquan Sun, Jiaqiang Zhu, Sahar Mozaffari, Carole Ober, Mengjie Chen, and Xiang Zhou. Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. *Bioinformatics*, 35(3):487–496, 2019.

[35] Andy Dahl, Khiem Nguyen, Na Cai, Michael J Gandal, Jonathan Flint, and Noah Zaitlen. A robust method uncovers significant context-specific heritability in diverse complex traits. *The American Journal of Human Genetics*, 106(1):71–91, 2020.

[36] David Golan, Eric S Lander, and Saharon Rosset. Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

[37] Omer Weissbrod, Jonathan Flint, and Saharon Rosset. Estimating snp-based heritability and genetic correlation in case-control studies directly and with summary statistics. *The American Journal of Human Genetics*, 103(1):89–99, 2018.

[38] Tristan J Hayeck, Noah A Zaitlen, Po-Ru Loh, Bjarni Vilhjalmsson, Samuela Pollack, Alexander Gusev, Jian Yang, Guo-Bo Chen, Michael E Goddard, Peter M Visscher, et al. Mixed model with correction for case-control ascertainment increases association power. *The American Journal of Human Genetics*, 96(5):720–730, 2015.

[39] Hector Giral, Ulf Landmesser, and Adelheid Kratzer. Into the wild: Gwas exploration of non-coding rnas. *Frontiers in cardiovascular medicine*, 5:181, 2018.

[40] Stacey L Edwards, Jonathan Beesley, Juliet D French, and Alison M Dunning. Beyond gwass: illuminating the dark road from association to function. *The American Journal of Human Genetics*, 93(5):779–797, 2013.

[41] Aashiq H Mirza, Simranjeet Kaur, Caroline A Brorsson, and Flemming Pociot. Effects of gwas-associated genetic variants on lncrnas within ibd and t1d candidate loci. *PLoS One*, 9(8), 2014.

[42] Barbara Hrdlickova, Rodrigo Coutinho de Almeida, Zuzanna Borek, and Sebo Withoff. Genetic variation in the non-coding genome: Involvement of micro-rnas and long non-coding rnas in disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1910–1922, 2014.

[43] Brian S Gloss and Marcel E Dinger. Realizing the significance of noncoding functionality in clinical genomics. *Experimental & molecular medicine*, 50(8):1–8, 2018.

[44] Gautam Mehta, Rajiv Jalan, and Rajeshwar P Mookerjee. Cracking the encode: from transcription to therapeutics. *Hepatology (Baltimore, Md.)*, 57(6):2532–2535, 2013.

[45] Douglas W Yao, Luke J O'connor, Alkes L Price, and Alexander Gusev. Quantifying genetic effects on disease mediated by assayed gene expression levels. *BioRxiv*, page 730549, 2019.

[46] Claartje Aleid Meddens, Amy Catharina Johanna Van Der List, Edward Eelco Salomon Nieuwenhuis, and Michal Mokry. Non-coding dna in ibd: from sequence variation in dna regulatory elements to novel therapeutic potential. *Gut*, 68(5):928–941, 2019.

[47] Michael Bulger and Mark Groudine. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144(3):327–339, 2011.

[48] Alex Wells, David Heckerman, Ali Torkamani, Li Yin, Jonathan Sebat, Bing Ren, Amalio Telenti, and Julia di Iulio. Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nature communications*, 10(1):1–9, 2019.

[49] Klaasjan G Ouwens, Rick Jansen, Michel G Nivard, Jenny van Dongen, Maia J Frieser, Jouke-Jan Hottenga, Wibowo Arindrarto, Annique Claringbould, Maarten van Iterson, Hailiang Mei, et al. A characterization of cis-and trans-heritability of rna-seq-based gene expression. *European Journal of Human Genetics*, 28(2):253–263, 2020.

[50] Yoav Gilad, Scott A Rifkin, and Jonathan K Pritchard. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in genetics*, 24(8):408–415, 2008.

[51] Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, et al. Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 44(10):1084–1089, 2012.

[52] Hung-ying Lin, Qiang Liu, Xiao Li, Jinliang Yang, Sanzhen Liu, Yinlian Huang, Michael J Scanlon, Dan Nettleton, and Patrick S Schnable. Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by erd-gwas. *Genome biology*, 18(1):192, 2017.

[53] Ryan D Hernandez, Lawrence H Uricchio, Kevin Hartman, Chun Ye, Andrew Dahl, and Noah Zaitlen. Ultra-rare variants drive substantial cis-heritability of human gene expression. *bioRxiv*, page 219238, 2019.

[54] Dorret Boomsma, Andreas Busjahn, and Leena Peltonen. Classical twin studies and beyond. *Nature reviews genetics*, 3(11):872–882, 2002.

[55] Peter M Visscher, Gibran Hemani, Anna AE Vinkhuyzen, Guo-Bo Chen, Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Jian Yang. Statistical power to detect genetic (co) variance of complex traits using snp data in unrelated samples. *PLoS genetics*, 10(4), 2014.

[56] Denis Awany, Imane Allali, Shareefa Dalvie, Sian Hemmings, Kilaza S Mwaikono, Nicholas E Thomford, Andres Gomez, Nicola Mulder, and Emile R Chimusa. Host and microbiome genome-wide association studies: current state and challenges. *Frontiers in genetics*, 9, 2018.

[57] Pamela Ferretti, Edoardo Pasolli, Adrian Tett, Francesco Asnicar, Valentina Gorfer, Sabina Fedi, Federica Armanini, Duy Tin Truong, Serena Manara, Moreno Zolfo, et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell host & microbe*, 24(1):133–145, 2018.

[58] Ryan W Walker, Jose C Clemente, Inga Peter, and Ruth JF Loos. The prenatal gut microbiome: are we colonized with bacteria in utero? *Pediatric obesity*, 12:3–17, 2017.

[59] Maria Elisa Perez-Muñoz, Marie-Claire Arrieta, Amanda E Ramer-Tait, and Jens Walter. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome*, 5(1):48, 2017.

[60] Shaopu Wang, C Anthony Ryan, Patrick Boyaval, Eugene M Dempsey, R Paul Ross, and Catherine Stanton. Maternal vertical transmission affecting early-life microbiota development. *Trends in microbiology*, 28(1):28–45, 2020.

[61] Linda Wampach, Anna Heintz-Buschart, Joëlle V Fritz, Javier Ramiro-Garcia, Janine Habier, Malte Herold, Shaman Narayanasamy, Anne Kaysen, Angela H Hogan, Lutz Bindl, et al. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nature communications*, 9(1):1–14, 2018.

[62] Maria Gloria Dominguez-Bello, Filipa Godoy-Vitorino, Rob Knight, and Martin J Blaser. Role of the microbiome in human development. *Gut*, 68(6):1108–1114, 2019.

[63] Howard Ochman, Michael Worobey, Chih-Horng Kuo, Jean-Bosco N Ndjango, Martine Peeters, Beatrice H Hahn, and Philip Hugenholtz. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS biology*, 8(11):e1000546, 2010.

[64] Julia K Goodrich, Emily R Davenport, Michelle Beaumont, Matthew A Jackson, Rob Knight, Carole Ober, Tim D Spector, Jordana T Bell, Andrew G Clark, and Ruth E Ley. Genetic determinants of the gut microbiome in uk twins. *Cell host & microbe*, 19(5):731–743, 2016.

[65] Siegfried Ussar, Shiho Fujisaka, and C Ronald Kahn. Interactions between host genetics and gut microbiome in diabetes and metabolic syndrome. *Molecular metabolism*, 5(9):795–803, 2016.

[66] Asko Mäki-Tanila and William G Hill. Influence of gene interaction on complex trait variation with multilocus models. *Genetics*, 198(1):355–367, 2014.

[67] Wen Huang and Trudy FC Mackay. The genetic architecture of quantitative traits cannot be inferred from variance component analysis. *PLoS genetics*, 12(11):e1006421, 2016.

[68] Daphna Rothschild, Omer Weissbrod, Elad Barkan, Alexander Kurilshikov, Tal Korem, David Zeevi, Paul I Costea, Anastasia Godneva, Iris N Kalka, Noam Bar, et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature*, 555(7695):210, 2018.

[69] Ruth E Ley, Micah Hamady, Catherine Lozupone, Peter J Turnbaugh, Rob Roy Ramey, J Stephen Bircher, Michael L Schlegel, Tammy A Tucker, Mark D Schrenzel, Rob Knight, et al. Evolution of mammals and their gut microbes. *Science*, 320(5883):1647–1651, 2008.

[70] Lucas P Henry, Marjolein Bruijning, Simon KG Forsberg, and Julien F Ayroles. Can the microbiome influence host evolutionary trajectories? *bioRxiv*, page 700237, 2019.

[71] Christine S Cheng, Rachel E Gate, Aviva P Aiden, Atsede Siba, Marcin Tabaka, Dmytro Lituiev, Ido Machol, Meena Subramaniam, Muhammad Shamim, Kendrick L Hougen, et al. Genetic determinants of co-accessible chromatin regions in t cell activation across humans. *bioRxiv*, page 090241, 2017.

[72] Allan F McRae, Joseph E Powell, Anjali K Henders, Lisa Bowdler, Gibran Hemani, Sonia Shah, Jodie N Painter, Nicholas G Martin, Peter M Visscher, and Grant W Montgomery. Contribution of genetic variation to transgenerational inheritance of dna methylation. *Genome biology*, 15(5):R73, 2014.

[73] JK Haseman and RC Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior genetics*, 2(1):3–19, 1972.

[74] Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *The American Journal of Human Genetics*, 91(6):1011–1021, 2012.

[75] Michael C Neale, Michael D Hunter, Joshua N Pritikin, Mahsa Zahery, Timothy R Brick, Robert M Kirkpatrick, Ryne Estabrook, Timothy C Bates, Hermine H Maes, and Steven M Boker. Openmx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, 81(2):535–549, 2016.

[76] Guo-Bo Chen. Estimating heritability of complex traits from genome-wide association studies using ibs-based haseman–elston regression. *Frontiers in genetics*, 5:107, 2014.

[77] Eun Yong Kang, Cue Hyunkyu Lee, Nicholas A Furlotte, Jong Wha J Joo, Emrah Kostem, Noah Zaitlen, Eleazar Eskin, and Buhm Han. An association mapping framework to account for potential sex difference in genetic architectures. *Genetics*, 209(3):685–698, 2018.

[78] Rachel Moore, Francesco Paolo Casale, Marc Jan Bonder, Danilo Horta, Lude Franke, Inês Barroso, and Oliver Stegle. A linear mixed-model approach to study multivariate gene–environment interactions. *Nature genetics*, 51(1):180–186, 2019.

[79] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228, 2015.

[80] Doug Speed and David J Balding. Sumher better estimates the snp heritability of complex traits from summary statistics. *Nature genetics*, 51(2):277–284, 2019.

[81] Yizhen Zhong, Minoli A Perera, and Eric R Gamazon. On using local ancestry to characterize the genetic architecture of human traits: Genetic regulation of gene expression in multiethnic or admixed populations. *The American Journal of Human Genetics*, 104(6):1097–1115, 2019.