INTRODUCTION
The aim of this chapter is to review and discuss the designs and methods of behavior genetics and genetic epidemiology. We define these fields broadly as the study of the role of genes and environment in phenotypic individual differences, where the term phenotype is employed for any measured (observed) characteristic or trait, i.e., ranging from biomarkers in blood, MRI or EEG data, and blood pressure, to personality traits, psychiatric disorders, and lifestyle. While behavior genetics often is associated with psychology, we define it to encompass any human phenotype which is measurable and can be characterized by appreciable individual differences. We limit this chapter to human behavior genetics, but acknowledge the common foundation of animal and human genetics in the biometrical model (Fisher, 1918; Falconer & Mackay, 1996; Lynch & Walsh, 1998; Plomin, DeFries, Knopik, & Neiderhiser, 2013) and quantitative genetic theory. Individual differences, as far as they are inherited, often depend on genes with on average small effects and are usually influenced by many genetic variants.

The methods of quantitative genetics are characterized by studies of large populations and families/twins and by the fact that the nature of quantitative differences requires the measurement (and not just the classification) of individuals. Quantitative genetic theory is concerned with the genetic properties of populations (population genetics) and the inheritance of measurements (biometrical genetics) and is now merging with molecular genetics with the realistic aim of localization and characterization of genes causing quantitative variation. We focus on human phenotypes that are complex, i.e., subject to the influence of many genes and environmental exposures.

Before 1990 behavior genetics was dominated by family, adoption, and twin studies, in which the contributions of unmeasured (or “latent”) genetic and environmental factors to phenotypic variation were inferred from the phenotypic resemblance among family members (Martin & Eaves, 1977; Posthuma et al., 2003). This work established beyond reasonable doubt the importance of genetic and environmental factors in a variety of human phenotypes (McGue & Bouchard Jr., 1998; Bouchard & McGue, 2003; Polderman et al., 2015) and set the stage for addressing follow-up issues, including questions about comorbidity and multivariate traits, the analyses of longitudinal data and time-series, the interaction and covariance of genes and environment, and the identification of genetic variants at the DNA sequence level. Increasingly, questions about the regulation of gene expression, through, for example, epigenetic processes, are part of the field of behavior genetics (Van Dongen, Slagboom, Draisma, Martin, & Boomsma, 2012).

From the late 1990s onwards, the scope of behavior genetics was expanded by the advent of high throughput genotyping technologies, which enabled researchers to measure DNA variants at a genome-wide level. Initially, highly polymorphic genetic variants comprising many alleles were used in linkage analyses of complex phenotypes such as neuroticism (Wray et al., 2008) or EEG power (Anokhin et al., 1992) to identify chromosomal regions harboring a causal genetic variant. From the early 2000s, advances in genotyping technology enabled researchers to measure single nucleotide polymorphisms (SNPs) in the hundreds of thousands and conduct genome-wide association studies (GWAS), which led to numerous discoveries (Visscher, Brown, McCarthy, & Yang, 2012). In linkage studies, a within-family design is employed and biologically related family members are required. In association studies, the measured genetic variants are tested directly by regression of the phenotype on the variant in a sample of individuals, who can be related or unrelated. Association studies largely superseded linkage studies, because they are generally more powerful and do not require related individuals. In addition, association studies may identify causal variants, or the regions harboring a causal genetic variant which are appreciably smaller than those identified in linkage analysis.

The availability of vast amounts of measured genetic variants, and their use in GWAS, has also given rise to techniques to estimate the contribution of the entire set of measured SNPs to phenotypic variation, without identifying the association of the phenotype with any
individual SNP (Yang, Lee, Goddard, & Visscher, 2011; Visscher, Yang, & Goddard, 2010; So, Li, & Sham, 2011; Lubke et al., 2012). A second technique to test predictions from subsets of SNPs involves aggregating a subset of SNPs into a weighted polygenic score (Purcell et al., 2009; Dudbridge, 2013; Wray et al., 2014), where the weighting of SNP variants is based on the outcome of large-scale GWA studies. For example, polygenic scores based on the GWAS for educational attainment in adults predicted ADHD in children (de Zeeuw et al., 2014).

This chapter follows the historical outline. We first discuss twin and family studies to estimate heritability and then genetic linkage and genetic association analysis to localize and identify DNA variants and next introduce analysis of polygenic scores, and the estimation of SNP-based heritability. Throughout we consider the different methods in terms of a basic regression model, allowing the reader to gain insight into the communalities between the methods.

FAMILY AND TWIN DESIGNS FOR HERITABILITY ESTIMATION OF COMPLEX HUMAN TRAITS

Family and twin designs provide the means to infer, under explicit assumptions, the contributions of genetic and environmental influences to phenotypic variation. The statistical model employed in family and twin studies involves the regression of a measured phenotype (y) on unobserved or latent genetic (G) and latent total environmental (T) variables:

\[ y_{ij} = b_0 + g \cdot G_{ij} + t \cdot T_{ij} \] (15.1),

where i denotes family or twin pair, j denotes family member, and b0 is an intercept. Equation 15.1 represents the linear regression; given a binary or discrete phenotype generalized linear regression is used (e.g., probit regression). It is assumed that interaction among genes and environment is absent; we return to this assumption below. The regression coefficient g quantifies the contribution of G to the phenotype y and the regression coefficient t quantifies the contribution of all non-genetic factors. G includes all effects of a possibly large, but unknown, number of genes. Given that genetic variants occupy specific chromosomal locations, the term “locus” is sometimes used in reference to a gene. To contribute to variance in y, genes must be polymorphic. Functional genetic loci can be monomorphic and since they do not vary between individuals, their contribution to individual differences in the phenotype is mediated by factors that regulate their expression.

As genes come in pairs, except for the X-chromosome in men, a di-allelic genetic variant with alleles A and B gives rise to three genotypes: BB, AB, AA (assuming that BA and AB cannot be distinguished), which can be coded 0, 1, 2 (reflecting the presence of the number of A alleles). The effect of alleles may be additive; this is the case if the presence of one “A” allele increases the phenotype value by a given value a and genotypes coded 0, 1, and 2 correspond to effects 0, 1a, and 2a, respectively. Genetic non-additivity implies a deviation from linearity due to intra-locus allelic interaction (e.g., 0, 1, 2 correspond to 0, 0, 2a). Such intra-locus non-additivity is referred to as genetic dominance. Interaction between alleles at different loci is referred to as epistasis. To accommodate additive and non-additive genetic effect, the G in the regression model is replaced by A and D, representing additive and dominance genetic effects. We discard epistatic effects as they are hard to distinguish statistically from dominance effects (Falconer & Mackay, 1996).

All effects on the phenotype that are not attributable to genetic variation are referred to as environmental effects. Environmental effects are broad in nature and include prenatal exposures, and experiences and exposures during childhood and throughout life. Relevant environmental variables are often unidentified and unmeasured, and their effects are inferred from twin and family correlations. Often two classes of environmental influences are distinguished: common environmental factors (C) that are shared among twin pairs or family members, and unique environmental factors (E) that are not shared among family members. When studies focus on older participants, whose phenotypes are assessed simultaneously with the phenotypes of cohabitants (spouses or other adults with whom they share a household), shared environment is sometimes referred to as “household effects.” Both terms emphasize that resemblance among relatives, whether they are biological relatives or not, can arise from sharing an environment, in addition to sharing of genes. Given the additive (A) and non-additive (D) genetic effects, and common (C) and unshared (E) environmental effects we arrive at the following regression model:

\[ y_{ij} = b_0 + a \cdot A_{ij} + d \cdot D_{ij} + c \cdot C_{ij} + e \cdot E_{ij} \] (15.2),

where a, d, c, and e are regression parameters. Given that the predictors are latent, we have to impose some scale on them, as is standard in latent variable modeling (Bollen, 1989). We assume that the predictors are standardized (have unit variance and zero mean). By definition A and D are uncorrelated (Falconer & Mackay, 1996), and C and E are uncorrelated. Assuming that genetic and environmental factors also are uncorrelated we arrive at the following decomposition of phenotypic variance:

\[ \sigma_y^2 = a^2 \sigma_A^2 + d^2 \sigma_D^2 + c^2 \sigma_C^2 + e^2 \sigma_E^2 - a^2 + d^2 + c^2 + e^2 \] (15.3),

as the latent factors have unit variance. The relative influence of genetic factors on phenotypic variation, called the “heritability,” is defined as the percentage of total phenotypic variance that can be attributed to genetic effects. “Broad-sense” heritability includes both additive and non-additive genetic variance (h2b = [a2 + d2] / s2y), “narrow-sense” heritability is limited to additive genetic variance (h2n = [a2] / s2y). Note that large heritability estimates imply
that genetic differences contribute substantially to phenotypic variance, but provide no information concerning the number or location of the relevant genes. Measurement error often cannot be distinguished from unshared environmental effect, and is usually included in \(e^2\). The variance of environmental effects may also include effects that are not strictly environmental, such as the effects of private genetic mutations, and so-called “developmental noise” (Molenar, Boomsma, & Dolan, 1993; Freund et al., 2013).

Family (including adoptive families) and twin designs are required to estimate the variance components in equation 15.3. In such designs, the correlation among family members is modeled as a function of their known genetic and environmental relations (Eaves, 1987; Boomsma, Busjahn, & Peltonen, 2002; Martin, Boomsma, & Machin, 1997). As phenotypic resemblance is summarized in correlation or covariance matrices, the statistical analysis of such family data is essentially covariance structure or structural equation modeling, in which the predictors A, D, C, and E are treated as latent variables. As a large number of genetic studies in psychophysiology employed the classical twin design, we base our explanation of these models on this design.

Let us reconsider equation 15.2 for monozygotic (MZ) and dizygotic (DZ) twin pairs (subscript \(j\) is now \(j = 1\) or \(j = 2\) and \(y\) is the deviation from the population mean).

\[
y_{i1} = a * A_{i1} + d * D_{i1} + c * C_{i1} + e * E_{i1}
\]

\[
y_{i2} = a * A_{i2} + d * D_{i2} + c * C_{i2} + e * E_{i2}
\]

The covariance between twin 1 (\(y_{i1}\)) and twin 2 (\(y_{i2}\)), \(\sigma_{y_{i1}y_{i2}}\), equals \(\sigma_{y_{i1}y_{i2}} = a^2 r(A_1, A_2) + d^2 r(D_1, D_2) + c^2 r(C_1, C_2)\), and \(r()\) denotes correlation. The twin design is based on the fact that we know the values of these correlations, given that we know the zygosity of twin pairs. If shared environmental influences for MZ twins are the same as for DZ twins, \(r(C_1C_2) = 1\). However, monozygotic (MZ) twins share 100 percent of their genes (barring de novo mutations (Ehli et al., 2012)), implying that \(r(A_1, A_2) = r(D_1, D_2) = 1\). If the genetic correlation among their parents is zero (random mating), dizygotic (DZ) twins share on average 50 percent of their segregating genes (Visscher et al., 2006; Mather & Jinks, 1977) and so \(r(A_1, A_2) = 0.5\). The dominance correlation is 0.25 (Mather & Jinks, 1977).

Thus, we arrive at the following expected covariance matrices \(\Sigma\) in MZ and DZ twins:

\[
\Sigma_{MZ} = \begin{bmatrix}
a^2 + d^2 + c^2 + e^2 \\
ad^2 + d^2 + c^2 \\
a^2 + d^2 + c^2 + e^2 \\
\end{bmatrix}
\]

(15.5a)

\[
\Sigma_{DZ} = \begin{bmatrix}
a^2 + d^2 + c^2 + e^2 \\
\frac{1}{2} a^2 + \frac{1}{4} d^2 + c^2 \\
a^2 + d^2 + c^2 + e^2 \\
\end{bmatrix}
\]

(15.5b)

Usually, we assume no differences between first- and second-born twins or between zygosities in means or variances, but note that these assumptions are easy to test. Also note that two observed covariances and one total phenotypic variance cannot identify the four parameters a, d, c, and e. Depending on the pattern of MZ and DZ correlations (Keller & Coventry, 2005) either an ACE or an ADE model is considered. When D contributes significantly to a phenotype, we expect the correlation in MZ pairs to be larger than twice the correlation in DZ pairs. If C contributes significantly to a phenotype, we expect the correlation in MZ pairs to be less than twice as large as DZ correlations. As a rule of thumb, if \(2 r_{DZ} > r_{MZ}\), this is indicative of an ACE model, and a quick estimate of the narrow-sense heritability is obtained as \(2 r_{MZ} – r_{DZ}\). For example, if \(r_{MZ} = 0.6\) and \(r_{DZ} = 0.4,\) heritability is \(2(0.6-0.4)=0.4\), the proportion of variance explained by C is \(2 r_{DZ} r_{MZ} (2^*0.4 – 0.6 = 0.2)\) and by E \(1 – r_{MZ} (1–0.6 = 0.4)\). If \(2 r_{DZ} < r_{MZ}\), this is indicative of the effects of non-additive genetic effects (dominance and/or epistasis).

The fact that we are limited to ACE or ADE can be overcome by adding data from additional family members, such as half-siblings growing up in the same household, or parents of twins. Recent work indicates that dominance variation at common SNPs tends to explain only a small fraction of phenotypic variation for human complex traits (Zhu et al., 2015).

Genetic covariance structure modeling (GCSM) is often used to fit genetic models to data from MZ and DZ twin pairs, to obtain estimates of the parameters a, c (or d), and e. Often this is done in programs for structural equation modeling with maximum likelihood (ML) estimation such as LISREL (Boomsma & Molenar, 1986; Fulker, Baker, & Bock, 1983), Mplus (Prescott, 2004), Mx (Maes et al., 2009), and OpenMx (Boker et al., 2011). Both Mx and OpenMx were written specifically to facilitate twin and family modeling. ML estimation has the advantages of providing goodness of fit indices to evaluate overall model fit (Bollen, 1989), standard error of parameter estimates, and nested model comparison using the likelihood ratio test. Covariance structure models, including twin and family models, can be represented graphically in a path diagram. An example of the path model that corresponds to the covariance model in equation 15.5a and 15.5b is shown in Figure 15.1. Path diagrams provide insightful and intuitive representations, which are mathematically complete.

So far, we have considered the linear regression model, in which the phenotype is continuously and (conditional on fixed covariates such as sex and age) approximately normally distributed. Non-normality can be handled by transforming the data to approximate normality, or by adopting a robust estimator. Phenotypes of interest such as the prevalence and/or absence of disease may be measured on a discrete scale. Fitting models to discrete data can be done by applying a suitable least squares estimator (Browne, 1973; Bollen, 1989) or by full information maximum likelihood estimation, which is the main method in OpenMx. Regardless of the estimator, the approach is closely related to probit regression analysis, where the dependent phenotype is binary. Alternatively, models for discrete phenotypes often assume an underlying, latent,
normally distributed variable, called liability (Plate 19). The frequencies (or prevalences) of the discrete values (say 0 and 1 for unaffected and affected) determine the position of a threshold on the liability scale and define affected (when an individual crosses the threshold) and unaffected status. This approach can be generalized to discrete phenotypes comprising more than two values. In this “liability-threshold model,” the liability is the variable that is subject to genetic covariance structure modeling using twin data (Falconer & Mackay, 1996).

**Multivariate and Longitudinal Genetic Analyses**

Above we considered a single phenotype, measured in twins. As twin pair (rather than individual twin) is the sampling unit, the data are by definition bivariate, as is the covariance structure model. Data arising from psychophysiological studies often are multivariate and genetic covariance structure modeling is easily generalized to multivariate data, where $y$ in equation 15.2 becomes a vector of $p$ observations and $A$, $D$, $C$, and $E$ also are $p$-dimensional vectors.

Figure 15.2 shows a path diagram that can be used to decompose the covariance between two traits into components due to correlated genetic and correlated environmental effects, i.e., here the multivariate twin model is used to determine the contributions of genetic and environmental effects to the phenotypic variances and covariances. Multivariate data arise naturally in longitudinal studies, where the same phenotype(s) is (are) measured repeatedly. The longitudinal genetic model can reveal contributions of genetic and environmental effects to the stability over time or age. The genetic and environmental covariance structures can be modeled using well-established models for repeated measures, such as the autoregressive model (Boomsma & Molenaar, 1987; Hewitt, Eaves, Neale, & Meyer, 1988; Kan et al., 2013), which emphasizes genetic and environmental contributions to stability and change, and growth curve models, which emphasize the contributions to growth (McArdle, 2006).

The classical twin model is based on various explicit assumptions. These include random mating, absence of interaction and correlation among the latent variables, and the assumption of equal environmental (including prenatal intra-uterine environment) effects in MZ and DZ twins. A lot of work has been devoted to extending the classical twin model to arrive at designs in which these assumptions can be tested. For instance, assortative mating (a correlation among phenotypes of spouses) can result in spurious estimates of common environmental variance in the classical twin design. By extending the twin design to include parents, assortative mating can be included in the model, and its effects accounted for (Cloninger, Rice, & Reich, 1979; Keller et al., 2009). Below we discuss gene–environment interaction and correlation in more detail.
Gene–Environment (G×E) Interaction: Moderation of Genetic and Environmental Effects

G×E interaction is conceptualized in terms of genetic control of sensitivity to the environment, i.e., the effect of the environment depends on the genetic makeup of the individual, or alternatively as the impact of a particular genotype depending on the environment in which it is expressed (Kendler & Eaves, 1986). We can cast this in terms of moderation of environmental or genetic effects by any exposure or moderator. As such, we can consider G×E interaction in the twin model, where the moderation of environmental effects by genetic effects is detectable as heteroskedasticity where the variance of E varies as a function of the level of G. This is complicated by the fact that G is a latent variable, but various approaches have been developed that can detect such heteroskedasticity (Jinks & Fulker, 1970; Eaves, 1987; van der Sluis, Dolan, Neale, Boomsma, & Posthuma, 2006; Molenaar et al., 2013; Molenaar & Dolan, 2014). Alternatively we can consider interaction in terms of the effects of a measured moderator on genetic and environmental effects (Purcell, 2002; Medland, Neale, Eaves, & Neale, 2009). In this case the moderator can be any variable, and may itself be subject to genetic and environmental effects. A relatively simple case is a moderator that has the same value in both twins, such as age, religious upbringing (Boomsma, De Geus, Van Baal, & Koopmans, 1999), or socioeconomic status. In this case, one can estimate the effects of genotype and environment on the phenotype of interest conditional on the moderator. In the absence of any moderation, the genetic and environmental effects, as expressed in terms of variance components (a², c² or d², and e²), are the same regardless of the value of the moderator, for example heritability does not differ as a function of age or socioeconomic status. Note that the absence of moderation does not imply the absence of a main effect. For instance, if the genetic and environmental effects on height are the same in 12- and 18-year-olds, this means that the size of the variance components are identical, but it does not rule out a main effect of age on height. Clearly, on average, 12-year-olds and 18-year-olds are likely to differ in height.

Sex is another example of a moderator. Introducing a subscript for sex (f for female; m for male), in the univariate model we test a²f = a²m, etc. However, the presence of DZ opposite-sex twins (DZOS) offers the unique possibility of further investigating the nature of sex differences in a² (a²f ≠ a²m) or c² (c²f ≠ c²m). Given a² ≠ a²m there are two possibilities which can be distinguished by testing the DZOS genetic correlation. Either the same genes are active in males and females, but the effect of the genes is moderated by sex (a quantitative sex difference), or different genes are active in males and females (a qualitative sex difference). The DZOS additive genetic correlation is expected to be 0.5 in the former case, but less than 0.5 in the latter (Vink et al., 2012). The presence of DZOS twins, in combination with male and female MZ and DZ twins, allows one to test this. The qualitative sex differences model can also be applied in the context of an environmental hypothesis: instead of fixing the correlation between C factors at 1 in DZOS twins, it can be estimated as a free parameter. If it is judged to be significantly lower than 1.0, this indicates that the influence of the shared environment differs in the two sexes. However, note that, as there is only one group of DZOS twins (there are no MZ twins of opposite sex) this analysis is limited to either the genetic or the common environment correlation, as they cannot be estimated simultaneously.

Gene–Environment (GE) Correlation

Genetic and environment variables may be correlated (Scarr & McCartney, 1983; Kendler & Eaves, 1986). Various plausible processes are expected to give rise to GE correlation. For instance, the contributions of a parent to the home environment, as experienced by the offspring, may depend on the parent’s genotype. This process can be investigated in models including data from parents and twins. Similarly, siblings (including twins) who grow up together may contribute (negative or positively) to each other’s environment (Eaves, 1976). If the phenotypes involved in such contribution are subject to genetic effects (e.g., rowdiness, aggression), this will give rise to GE correlation. This process can be studied by including the regression of the twins phenotypes on each other in a cross-sectional (Eaves, 1976) or longitudinal twin study (Carey, 1986; Dolan, de Kort, Van Beijsterveldt, Bartels, & Boomsma, 2014).

Purcell (2002) developed a general model to investigate G×E interaction in the presence of gene–environment correlation, i.e., where moderation is by a continuously distributed moderator (G×E), which is itself possibly subject to genetic and environmental effects. For instance, general parental support (encouragement, taking an active interest, helping with homework) may moderate genetic and environmental influences on offspring intelligence. But parental support, while contributing to the environment of the child, is itself likely to be subject to genetic influences. Gene–environment correlation could arise if the amount of support was a function of parental intelligence.

Causal Modeling in Genetic Data

In the co-twin control design (Cederlof, Friberg, & Lundman, 1977; Kendler et al., 1993), MZ and DZ twins who are discordant for an environmental exposure or a disease are studied, sometimes along with unrelated individuals. In addition to disease status a risk factor is measured (say smoking), which is related to the condition. Assuming genetic and environmental influences are uncorrelated, and assuming a direct causal effect of the risk factor on the condition, the strength of the association
will be the same in MZ and DZ pairs (e.g., the affected twin smokes more often), and in pairs of unrelated individuals (affected individuals smoke more often). If the association between a risk factor and the condition is due to pleiotropic genetic effects (the same genes affect both phenotypes), the strength of the association will be greatest in the unrelated individuals, but smaller in the DZ, and zero in the MZ pairs (as the MZ twins within a pair are completely matched for genetic influences). Groen-Blokhuys and colleagues (Groen-Blokhuys, Middeldorp, van Beijsterveldt, & Boomsma, 2011) investigated the association between low birth weight and attention problems. In MZ pairs, DZ pairs, and unrelated pairs of children, the child with the lowest birth weight scored higher on attention problems at age 3, 7, 10, and 12 years. Thus, the association is causal: that is, a lower birth weight directly causes increased attention problems. Other extensions of causal modeling using twin and family data are discussed by De Moor and colleagues (De Moor, Boomsma, Stubble, Willemsen, & De Geus, 2008) who concluded that the well-established association of lack of exercise and depression is not a causal one.

**LINKAGE ANALYSIS**

The advent of high throughput genotyping technologies enabled researchers to measure DNA variants in unprecedented volumes. These data can be exploited in the hunt for the causal genetic variants contributing to the variance of complex phenotypes. Initially, the focus was on highly polymorphic genetic markers such as microsatellites for which many alleles are found in the population. Such markers were exploited in genetic linkage analysis to locate chromosomal regions associated with a phenotype. Such regions were typically large and included many genes. Linkage analysis relies on the fact that genes in close proximity are transmitted together (linked). Such linkage is disrupted by recombination. That is, when gametes (sperm and egg cells) are produced during meiosis, the paired homologous chromosomes separate so that each gamete contains only one of the pair of alleles at a locus. During the first division of meiosis, sections near the ends of chromosomes commonly exchange parts of their chromatids with the other chromosome of their homologous pair. The probability of the linkage between two loci being disrupted, i.e., recombination, depends on the distance between the loci.

If we observe that family members who share a particular stretch of a genome, defined on the bases of sharing polymorphic markers, also share a phenotype, this indicates that this stretch contains one or multiple loci influencing the phenotype. Linkage analysis can be done in multigeneration pedigrees (parametric linkage analysis) or in sibling (or DZ) twin pairs. Here we consider this last type of non-parametric linkage analysis. In non-parametric linkage, variation in the proportion of alleles at a marker locus that family members share identically by descent (IBD, i.e., from the same ancestor) may identify the contribution of the marker locus to the phenotypic differences, if the marker locus is close to the causal locus (sometimes called a quantitative trait locus or QTL). We introduce this type of analyses within the context of genetic structural equation modeling (GCSM). The model employed in linkage can be written in terms of regression of the phenotype on a QTL (q), the latent genetic variance (G), and the total environmental variance (T).

\[ y_{ij} = b_0 + q \cdot Q_{ij} + g \cdot G_{ij} + t \cdot T_{ij} \]  
(15.6)

As applied to full siblings, the model is often limited to the QTL, an additive genetic variable (A), and unshared environmental effects (E), i.e., \( y_{ij} = b_0 + q \cdot Q_{ij} + a \cdot A_{ij} + e \cdot E_{ij} \). This regression model implies the following decomposition of phenotypic variance:

\[ \sigma_y^2 = q^2 \sigma_Q^2 + a^2 \sigma_A^2 + e^2 \sigma_E^2 = q^2 + a^2 + e^2 \]  
(15.7)

as scaling of the latent variables (Q, A, and E) implies that their variances equal one. We can derive standardized variance components in the model, i.e., the total heritability equals \((q^2 + a^2) / \sigma_y^2\), and the variance explained by the QTL is \(q^2 / \sigma_y^2\).

The covariance between siblings is \( \sigma_{yi} = q^2 \cdot r(Q_1, Q_2) + a^2 \cdot r(A_1, A_2) \), where \( r(A_1, A_2) = 0.5 \). The correlation between QTL factors of DZ twins or siblings \( r(Q_1, Q_2) \), which is often denoted pi-hat (\( \hat{\pi} \)), is obtained from measured genotypic markers. IBD status for the marker data determines this correlation. IBD status at a given locus equals 0, 1, or 2 in siblings, depending on the exact configuration of parental alleles that the siblings have inherited and we can specify the covariance between siblings as:

\[ \sigma_{yi} = q^2 \cdot \pi_i + a^2 + 0.5 \]  
(15.8)

To illustrate this in the simplest case, suppose that the parental genotypes at a locus are A1A2 and A3A4, in mother and father, respectively. If the siblings have identical genotypes (e.g., both A1A3), they share two alleles IBD (namely A1 and A3). If the siblings have no alleles in common (e.g., A1A3 and A2A4), they share zero alleles IBD. Finally if they share one and the same allele from a given parent (e.g., A1A3 and A1A4), they share one allele IBD (namely A1). Often IBD status cannot be established with certainty (when parental genotypes are A1A1 and A1A2, then siblings with genotypes A1A2 and A1A2 may be IBD2 or IBD1). It is, however, always possible to assign IBD probabilities (Haseman & Elston, 1972; Kruglyak & Lander, 1995).

Note that pi-hat is indicative of genetic resemblance in the region of the marker. Suppose that the marker happens to be the QTL (i.e., causal variant). In that case, the QTL contributes to the phenotypic resemblance of the siblings as a direct function of pi-hat. For example, pi-hat equals 1 means that the siblings are genetically identical at
the QTL. If pi-hat is zero, the siblings are essentially genetically unrelated at the QTL, and so the QTL contributes nothing to their phenotypic resemblance. Of course, the further away the marker is from the QTL, the more the IBD relationship will be diluted by recombination, i.e., the less indicative the pi-hat value is of genetic resemblance at the QTL. This specification of the within sib pair covariance allows us to construct two models for the observed covariance for sibships and DZ twin pairs, namely $r(\text{sib}) = \frac{\pi^2}{2} + 0.5 \sigma^2$ versus $r(\text{sib}) = 0.5 \sigma^2$ (i.e., no effect of the QTL). The models can be fitted using maximum likelihood and compared based on a likelihood-ratio test which is distributed as $\chi^2$. In a linkage analysis, results often are summarized in the form of an LOD score. There is a straightforward correspondence between $\chi^2$ and LOD scores: $LOD = \frac{\chi^2}{2\ln 10}$ (Sham, Zhao, & Curtis, 1997).

Evidence for linkage is present when the maximum LOD score exceeds a predefined threshold, which is based on the size of the genome and the number of measured markers. A commonly used threshold is an LOD score of 3, corresponding to the data being 1,000 times more likely given the model including the QTL than the null model. While the usage of linkage has largely been abandoned in favor of association analysis, recent papers based on linkage techniques have been used to determine the heritability of complex phenotypes without relying on common assumptions associated with the twin model (Visscher et al., 2006). Linkage can further be combined with association to fine map a specific region of the genome for application (see Van Dongen et al., 2014).

ASSOCIATION ANALYSIS

In twin and linkage studies, the significance of genetic factors is tested on the covariance structure among family members. Association studies, in contrast, focus on the direct effect of genetic variants on the trait itself which are investigated in related or unrelated individuals. If the phenotype is continuous and conditionally (on the predictor) normally distributed, association can be cast in terms of a simple linear regression analysis:

$$y_j = b_0 + b_x \cdot \text{SNP}_j + e_j$$  \hspace{1cm} (15.9),

where $y_j$ is the phenotype, SNP$_j$ is measured (usually coded 0, 1, 2 with the code indicating the number of reference alleles present in individual $j$), $e_j$ is the residual, and $j$ denotes the individual ($j = 1, \ldots, N$). Note that in this model the residual includes all effects not attributable to the SNP. These include environmental effects, genetic effects (attributable to all other genetic variants), and measurement error. Association can be tested statistically by testing the null hypothesis $b_x = 0$ vs. $b_x \neq 0$. If the phenotype is binary (i.e., disease status in a case-control design), generalized linear regression is used, but the test is the same. Association analysis is statistically more powerful than (sib pair) linkage analysis, because the test $b_x = 0$ concerns first-order statistics (conditional mean or prevalence), whereas the linkage test concerns second-order statistics (covariances).

In association studies, the observed association between a trait and a genetic variant may be due to population stratification. If a population comprises several subpopulations that differ in allele frequencies and phenotypic values, then an association may express mainly these differences, not any true relationship between a given genetic variant and the phenotype. Hamer and Sirota (2000) illustrated this by a (fictional) study in which genetic variants were related to the use of chopsticks in Asian and Caucasian students, who differ in the frequency of chopstick use, and in allele frequencies at various loci. The association between any of these loci and chopstick use is spurious, as it is driven by subpopulation differences, not by the effect of genes. If one corrects for this type of population stratification, the association disappears. For example, one might consider testing the association separately in the Asian and Caucasian students. Alternatively, to correct for stratification one may adopt a within-family association design. These designs are not subject to the effect of stratification, because they base the test of association on the association as observed within families, where the phenotypic differences are not subject to stratification, and the family members are matched with respect to many variables.

Association analysis was initially targeted at biologically plausible candidate loci, i.e., candidate gene association studies. These studies are hypothesis driven: i.e., they focus on a gene (possibly encompassing many genetic variants), which is judged to be of special interest on prior grounds. Several strategies can be used in selecting candidate genes. Genes that are part of physiological systems known to influence the trait may be suitable candidates. Genes (or chromosomal regions) that are known to influence the trait in animal models can inform the selection of candidate genes or regions in humans. Or, genes can be selected in regions of the genome, which were identified as being of interest in linkage analysis.

To a large extent, candidate gene association studies have now been replaced by genome-wide association studies (GWAS). GWAS are possible because advances in genotyping technology allow the assessment of hundreds of thousands of variants, which can then (with knowledge from sequenced reference genomes) be imputed up to millions of variants. GWAS are exploratory in that they test the association between a phenotype and many hundreds of thousands of genetic variants, usually single nucleotide polymorphisms (SNPs). SNPs are di-allelic, and so give rise to three genotypes, which are chosen to cover the whole genome. In this exploratory approach, a significant association may concern a causal genetic variant (with a biological interpretation), but is more likely to be a genetic variant that is in linkage disequilibrium with this causal variant. Linkage disequilibrium (LD) refers to the fact that with genetic variants that are
located so closely on a given chromosome, recombination between them is a rare event. As a consequence, the loci co-segregate in the population over many generations.

Both GWAS and candidate locus association studies require a correction of the alpha (significance threshold) for multiple testing. In a GWAS, comprising a large number of tests (>500K to >7000K), the corrected alpha level is usually set at $5 \times 10^{-8}$, which corresponds to a testing burden of one million independent tests (Hoggart, Clark, De Iorio, Whittaker, & Balding, 2008; Pe’er, Yelensky, Iorio, Whittaker, & Balding, 2008). GWAS and candidate gene studies have to control for population stratification. The presence of many measured genetic variants in a GWAS allows one to control for population stratification by means of a principal component (PC) analysis of the genetic data (Price et al., 2006).

Note that in equation 15.9 we included no family index, i.e., we assumed that the sample consists of unrelated individuals. The standard test of the SNP effect (i.e., the test of the null hypothesis $b_k = 0$) in (logistic) linear regression assumes that, conditional on the predictors, the residuals ($e_j$) are independent. However, the sample may include individuals clustered in families or there may be relatedness in the sample that the researcher is unaware of (cryptic relatedness). Recent methodological advances in association studies have seen the development of integrated approaches based on the linear mixed model to testing association given population stratification, the presence of known closely related individuals, and cryptic relatedness (Lippert et al., 2011; Yang, Zaitlen, Goddard, Visscher, & Price, 2014). Alternatively, one can forgo the modeling of the background covariance, treat $e_j$ as independent or as conforming to a simple covariance structure, and correct the standard error of the estimate of $b_k$ for this misspecification by means of a sandwich correction (Minicâ, Dolan, Kampert, Boomsma, & Vink, 2014).

**Meta- and Mega-Analysis of Genome-Wide Association Studies**

Given the potential confounders discussed above, and the large number of tests, GWAS require replication in independent samples. The results of multiple studies can be combined in a meta-analysis to arrive at a single test of association based on all available results. However, as different studies often use different genotype arrays, and the arrays measure different sets of SNPs, it is possible that only a small number of SNPs are genotyped in all replication samples. However, SNPs in close proximity on the genome are generally in strong LD (i.e., strongly correlated). This information can be leveraged to impute SNPs to arrive at a set of SNPs, common to all studies. In specific reference samples, all SNPs are measured using whole genome sequencing (International HapMap Consortium, 2005; 1000 Genomes Project Consortium, 2012). Given that all SNPs in the reference population are characterized, and the LD between these SNPs is known, one can impute the SNPs in the set that are not directly measured on a given genotyping platform (Marchini, Howie, Myers, McVean, & Donnelly, 2007; De Bakker et al., 2008; Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012). After imputing all SNPs in the reference set for all samples, the association test is performed for all SNPs that are imputed with acceptable accuracy. This yields effect size estimates for a homogeneous set of SNPs across all cohorts. These results can then subsequently be meta-analyzed.

Software is available to perform such genome-wide meta-analysis (for example: METAL (Willer, Li, & Abecasis, 2010)). The need to harmonize phenotypes and the need to impute SNPs in individual cohorts to a common reference set before performing primary analysis requires close cooperation of many labs and groups in large consortia, involving hundreds of collaborators. Some consortia go beyond meta-analysis and combine the raw genotype data to perform mega-analysis (Sullivan, 2010). This allows across cohort quality control before imputation and allows centralized analysis of the complete dataset. However, not all cohorts are free to store genetic data offline. The use of consortium driven meta- and mega-analysis has enabled identification of loci for a wide range of phenotypes (Manolio et al., 2007).

**Estimation of Heritability and Genetic (Co)Variance Based on Measured Genotypes**

Genome-wide SNPs arrays are designed to economically capture a substantial portion of genetic variation in the human genome. Yang and colleagues (2011) developed a method to estimate the variance in a phenotype explained by all measured SNPs, when genotype and phenotype data are available in (large) numbers of unrelated, or very distantly related subjects. So and colleagues (2011) developed methods that can be applied to the results of GWAS and meta-analysis studies. Both methods estimate the (narrow-sense) heritability based on information from measured genome-wide typed SNPs. The two approaches differ substantially, with the first approach resembling the variance decomposition methods as used in twin studies, and the second one based on density estimation (DE) methods. The Yang et al. approach (often referred to as the GCTA method after the software package) requires raw genotype data and uses these to obtain a measure of genetic similarity between all possible pairs of (unrelated) individuals in the study. In a second step this genetic relatedness matrix (GRM) is used to predict the phenotype similarity between individuals (just as the different similarity of MZ and DZ twin pairs predicts their different phenotype resemblances). The DE method can be applied after the genome-wide association study has been done. Here, the distribution of z-statistics of the association measure between SNPs and the phenotype in a GWAS is compared to the theoretical null distribution of z-statistics representing no effects. Explained variance will differ

Downloaded from https://www.cambridge.org/core. Vrije Universiteit Bibliotheek, on 05 Mar 2018 at 10:24:36, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781107415782.015
from zero if more z-statistics from the GWAS have larger values than expected under the null. Heritability based on twin data compared to those based on GRM and DE methods for major depression, smoking, and continuous measures of fasting glucose and height found that a substantial proportion of the twin-based heritability estimates is recovered by the GRM and DE methods (Lubke et al., 2012).

Various extensions of the GRM-based model have been developed, and some are included in GCTA. One useful extension is to conduct the analysis for the SNPs on the individual chromosomes or to partition the SNP variance over different functional categories of SNPs, e.g., SNPs in the portion of genome that is expressed (i.e., exonic SNPs), SNPs that regulate genes (e.g., DNaseI hyperactivity sites, gene promoter region, untranslated regions), SNPs that alter gene function (i.e., coding variants), and SNPs in intergenic regions (Gusev et al., 2014). A second extension implemented in GCTA generalizes the GRM-based model to the bivariate phenotypic case. In this model, the genetic covariance between traits attributable to the SNPs is estimated (Lee, Yang, Goddard, Visscher, & Wray, 2012). While such genetic covariance terms can be estimated readily in multivariate twin and family studies, the present estimate has the virtue of pertaining to the measured SNPs. The bivariate (co)variance decomposition is possible if each phenotype is measured in a distinct sample, but the precision of the covariance estimate improves if the phenotypes are measured in the same sample. Application of this method to data collected by the psychiatric genetics consortium demonstrated an SNP-based genetic covariance between schizophrenia, bipolar disorder, and depression.

As the primary goal of GRM-based analyses is to obtain an estimate of the variance explained by measured SNPs, closely related individuals are generally excluded. A recently proposed extension allows for the estimation of the variance attributable to SNPs in the presence of related individuals (Zaitlen et al., 2013). Specifically it allows for the estimation of the variance attributable to SNPs and the total variance attributable to genetic influences. To fit this model we require individuals who are closely and distantly related. Finally, the GCTA software suite also allows for the assessment of genetic-environment interaction, given a binary environmental variable, coded 0/1 (Yang et al., 2011).

A recent development is the possibility to accurately estimate the variance explained by all measured SNPs based on GWAS summary statistics (Bulik-Sullivan et al., 2015b). The LD score regression method assumes that complex traits are highly polygenic, i.e., thousands of SNPs have a causal effect on a complex phenotype. A variant in linkage disequilibrium with a causal genetic variant will also be associated with the trait. Any variant in stronger than average linkage disequilibrium with its neighbors will also have a higher than average chance to tag one of the many causal variants in the genome. Population stratification and other forms of confounding would increase the effects of all SNPs, not specifically SNPs in strong LD with their neighbors. Under this model one can derive the variance explained by all measured SNPs by regressing the chi-squared statistic of the SNPs, as obtained in a GWAS, on the linkage disequilibrium associated with each SNP.

\[ X^2_j = b_0 + b_1 * L_j + e_j \]  \hspace{1cm} (15.10),

where \( L_j \) is the sum of squared correlations between variant \( j \) and all other genetic variants: \( \sum_k r_{jk}^2 \). \( X^2 \) is the chi-squared statistic, i.e., the squared standardized regression coefficient, \( b_0 \) is the regression intercept and \( b_1 \) the regression slope. The slope in this regression reflects the average heritability explained per SNP and can be rewritten as:

\[ b_1 = N * \sigma^2_{u}/M \]  \hspace{1cm} (15.11),

where \( N \) is the sample size and \( M \) is the number of SNPs. As confounding factors do not correlate with \( L_j \), the intercept \( b_0 \) can be rewritten as:

\[ b_0 = N * a \]  \hspace{1cm} (15.12),

where the intercept captures bias and the effects of population stratification. A significant positive departure from 1 for the intercept indicates confounding or population stratification. To estimate the genetic correlation between two traits using LD score regression, the GWAS effect sizes for both traits are converted to Z scores and the product of Z scores per SNP is regressed on the LD score (Bulik-Sullivan et al., 2015a). In the model used to estimate the genetic correlation, the intercept accounts for possible confounding due to sample overlap between the two meta-analyses. LD score regression can and has been used to partition the total heritability in a trait over multiple functional categories of SNPs (Finucane et al., 2015). Functional categories can, for example, be SNPs with known effects on the protein structure, SNPs in specific regulatory parts of the genome, or SNPs in genes that are expressed in the brain. These analyses can provide insight into which regions in the genome, or which functional unit of the genome contain SNPs enriched for effects on a specific phenotype.

To assess the overall genetic overlap between the two phenotypes, SECA (SNP effect concordance analysis) combines the summary statistics from two GWAS analyses to determine the concordance in effect size and concordance in effect direction between the summary statistic for both traits (Nyholt, 2014). Using an exact binomial test, SECA tests whether SNPs for which the test statistic is associated with a p-value below a certain threshold \( t_1 \) for trait 1 have a p-value below a second threshold \( t_2 \) for trait 2 significantly more often than expected by chance. SECA performs this binomial test for 12 values of \( t_1 \) and \( t_2 \) in each set, resulting in 144 separate binomial tests. SECA further uses a Fisher’s exact test to calculate whether there is an excess of SNPs with concordant
directions of effect across the two primary GWAS analysis. This concordance is again calculated on sets of SNPs restricted to p-values below $t_1$ for trait 1 and $t_2$ for trait 2. This analysis results in 144 Fisher’s exact tests of effect concordance. Additionally, SECA can produce conditional QQ plots where the p-values for trait 1 are plotted conditionally on the p-value for trait 2. These techniques were first described by Andreassen and colleagues (Andreassen et al., 2013) who showed that a conditional analysis of the bipolar disorder and schizophrenia GWAS performed in the psychiatric genetic consortium massively increased discovery of associated genetic variances. These analyses go beyond finding evidence for genetic correlations between traits, but harness genetic correlation to improve the chance to detect variants associated with either of two correlated traits.

**POLYGENIC RISK SCORES**

GCTA can be applied to demonstrate that measured SNPs explain an appreciable portion of variance in a trait and bivariate GCTA can be applied to determine whether a shared set of SNPs is associated with two distinct traits. GCTA can be applied to find evidence of genetic effects (heritability) even if no individual SNP is significant, but GCTA requires access to measured genotypes and both phenotypes. An alternative method to demonstrate the presence of signal in genetic markers, or pleiotropy between two traits, is by means of polygenic (risk) scores. This involves the selection of a subset of SNPs which satisfy a given alpha level (not necessarily the genome-wide alpha of $5 \times 10^{-8}$), and the calculation of the weighted linear combination of the SNPs in the set, where the weights are set to equal the regression coefficients associated with the individual SNPs (i.e., the parameter $b_s$). This linear function of the SNPs is called the polygenic (risk) score or PGS. The regression coefficients used are derived from a genome-wide meta-analysis of a phenotype of interest (i.e., the discovery sample). The polygenic scores are then calculated for individuals that are not included in the discovery sample (i.e., the target sample). To determine the presence of signal in the genetic markers, the phenotype, measured in the target sample, is regressed on the polygenic risk score derived from the discovery sample. To determine genetic overlap between two traits, one can use the weights from a meta-analysis of, e.g., schizophrenia GWAS and use the polygenic scores to predict, e.g., bipolar disorder (Purcell et al., 2009). The regression of the phenotype on the risk score is expected to be significant if the set of SNPs is associated with the phenotype of interest. The significance level for inclusion of SNPs in the polygenic score may be set at varying values to assess their effect on the explained variance. The discovery and target sample need to be independent, as dependency may result in overestimation and false positives. While polygenic risk scores are often found to be predictive, the predictive power, even if based on the best available discovery samples, has been found to be too low to be clinically relevant (Dudbridge, 2013; Wray et al., 2013). Software packages to handle data management and computation of polygenic risk scores include Plink (Purcell et al., 2007) or PRSice (Euesden, Lewis, & O’Reilly, 2014).

**DISCUSSION**

There have been several reviews that focused on the applications of techniques that we outlined in this chapter specifically to phenotypes in the realm of psychophysiology (e.g., Boomsma & Gabrielli, 1985; Anokhin, 2014). A special issue of the *Journal of Psychophysiology* (December 2014) is completely devoted to a series of 17 comprehensive studies (Iacono, 2014) of electrodermal activity, startle eye blink, antisaccade error, and electroencephalographic spectral characteristics and ERP carried out in the Minnesota Twin Family study.

Traits studied in psychophysiology tend to show moderate (e.g., blood pressure and heart rate) to high heritability (e.g., EEG power). Wang et al. (2015) reviewed the genetic contribution to blood pressure and heart rate, and Wu and colleagues (Wu, Snieder, & De Geus, 2010) discuss how genetic influences on cardiovascular parameters tend to be more strongly expressed under conditions of stress reactivity. Linkage and association studies for blood pressure (e.g., Hottenga et al., 2007; Wain et al., 2011) and heart rate (den Hoed et al., 2013) have reported the first loci for these phenotypes and explored their effects on, for example, cardiac conduction and rhythm disorders. Reviews and meta-analyses of the heritability of total brain volume and subcortical brain volumes in adults and children (Blokland, de Zubicaray, McMahon, & Wright, 2012; den Braber et al., 2013; Swagerman, Brouwer, de Geus, Hulshoff Pol, & Boomsma, 2014) conclude that these structural MRI measures have (very) high heritability, also in childhood, and linkage (Kochunov et al., 2009) and GWAS (Hibar et al., 2015) have identified the first common genetic variants influencing human brain structures. Heritability studies of the human electroencephalogram (EEG) conclude that EEG power is among the mostheritable human phenotypes (van Beijsterveldt & Boomsma, 1994) whereas the analysis of event-related brain potentials (ERPs) indicates that these measures tend to be less heritable. Still, linkage studies (Begleiter et al., 1998) of P3 voltage and visual ERPs (Towne, Almasy, Siervogel, & Blangero, 1999) reported some suggestive findings. The first GWAS (Hodgkinson et al., 2010) of alpha, beta, and theta EEG power in only 322 participants identified three genes (SGIP1, ST6GALNAC3, and UGDH) with nominal association to variability of theta or alpha power.

The reviews and meta-analyses and this current chapter cover the methods and techniques from behavior genetics and genetic epidemiology, to linkage and association...
studies of candidate genes and genome-wide variants. They focus on association of genetic variants which have a reasonable frequency in the population, i.e., the alleles at these loci are not very rare. To detect and analyze rare genetic variants, DNA sequence data rather than variant data measured on genotyping arrays are measured and their association with complex traits requires suitable statistical techniques (an overview of these is given in Franic et al., 2015). Other genetic variants than di-allelic variants are common in the human genome, for example copy-number variants (CNV) form a class of structural variation which have been created (see, e.g., Pennisi, 2015). These layers of complexity may even differ between identical twins (Ehli et al., 2012).

The genome is more than a set of codes for proteins. It also contains sequences that control gene activity. The epigenome and the transcriptome describe how the DNA is regulated and expressed and can be studied through assessment of, for example, methylation probes and RNA expression data. These layers of complexity complicate searches for the genetic basis of complex phenotypes, but in the past few years catalogues and several “user manuals” for the genome, mapping the locations of control switches, the genes they control, and in which tissue of the body they are turned on or off have been created (see, e.g., Pennisi, 2015). In Table 15.1 a short list of websites is provided to guiding the reader to some of the catalogues and websites for genetic analyses.

Table 15.1 Useful online resources for genetic analysis

<table>
<thead>
<tr>
<th>Resource</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behavioral genetics interactive educational modules for analysis of twin data, genetic linkage studies, and other topics.</td>
<td><a href="http://pngu.mgh.harvard.edu/~purcell/bgim/">http://pngu.mgh.harvard.edu/~purcell/bgim/</a></td>
</tr>
<tr>
<td>dbSNP, a database containing all known single nucleotide polymorphisms in the human genome.</td>
<td><a href="http://www.ncbi.nlm.nih.gov/SNP/">www.ncbi.nlm.nih.gov/SNP/</a></td>
</tr>
<tr>
<td>Ensembl: contains a genome database for human, vertebrates and eukaryotic species.</td>
<td><a href="http://www.ensembl.org/index.html">www.ensembl.org/index.html</a></td>
</tr>
<tr>
<td>GCTA, tools to carry out genetic association studies, including designs that include relatives, to estimate the proportion of phenotypic variance explained by all genotyped SNPs and to estimate power to detect SNP heritability.</td>
<td><a href="http://cnsgenomics.com/software/gcta/">http://cnsgenomics.com/software/gcta/</a></td>
</tr>
<tr>
<td>Genetics of Personality Consortium (GPC) report of the meta-analysis for neuroticism:</td>
<td><a href="http://www.tweelingenregister.org/GPC">www.tweelingenregister.org/GPC</a></td>
</tr>
<tr>
<td>LD score regression, software to compute genetic correlations based on summary statistics derived from GWA meta-analysis:</td>
<td><a href="https://github.com/bulik/ladder">https://github.com/bulik/ladder</a></td>
</tr>
<tr>
<td>METAL, software to perform genome-wide association studies:</td>
<td><a href="http://csg.sph.umich.edu/abecasis/metal/">http://csg.sph.umich.edu/abecasis/metal/</a></td>
</tr>
<tr>
<td>Michigan Imputation server, a service to impute genotypes against contemporary reference panels:</td>
<td><a href="https://imputationserver.sph.umich.edu/index.html">https://imputationserver.sph.umich.edu/index.html</a></td>
</tr>
<tr>
<td>OpenMX, R package that enables the estimation of parameters in twin models:</td>
<td><a href="http://openmx.psyc.virginia.edu/">http://openmx.psyc.virginia.edu/</a></td>
</tr>
<tr>
<td>Plink software to perform genome-wide association studies and compute polygenic scores:</td>
<td>Ver 1.07: <a href="http://pngu.mgh.harvard.edu/~purcell/plink/anal.shtml">http://pngu.mgh.harvard.edu/~purcell/plink/anal.shtml</a></td>
</tr>
<tr>
<td>Plink2, R package for association studies:</td>
<td>Ver 1.9: <a href="http://www.cog-genomics.org/plink2">www.cog-genomics.org/plink2</a></td>
</tr>
<tr>
<td>Psychiatric Genomics Consortium (PGC) results download page. Here you find the summary statistics associated with GWAS meta-analysis results for a wide range of psychiatric disorders and other traits:</td>
<td><a href="http://www.med.unc.edu/pgc/results-and-downloads">www.med.unc.edu/pgc/results-and-downloads</a></td>
</tr>
<tr>
<td>SECA, SNP effects concordance analysis to detect pleiotropy between traits based on summary statistics from GWA meta-analysis:</td>
<td><a href="http://neurogenetics.qimrberghofer.edu.au/SECA/">http://neurogenetics.qimrberghofer.edu.au/SECA/</a> USCS genome browser, extensive annotation and visualization of human and animal genomes:</td>
</tr>
</tbody>
</table>

REFERENCES


