Routledge Handbook of Sport and Exercise Systems Genetics

J. Timothy Lightfoot, Monica J. Hubal, Stephen M. Roth

Human Systems Genetic Modeling Used in Exercise

Publication details
Jaakko Kaprio
Published online on: 03 Apr 2019

How to cite :- Jaakko Kaprio. 03 Apr 2019, Human Systems Genetic Modeling Used in Exercise from: Routledge Handbook of Sport and Exercise Systems Genetics Routledge
Accessed on: 26 Sep 2019

Please scroll down for document

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Introduction

Nearly all human characteristics and behaviors run in families, that is to say, relatives and people living together tend to resemble each other more than two individuals chosen at random from the same population. Such characteristics can be structural (height, bone structure, obesity) or physiological (muscle strength), behavioral (physical activity), or life-course related (such as number of children or lifespan). Correspondingly, abnormal conditions and disease often also run in families. Familial aggregation is thus a well-established observation, but the nature and causes of it depends on the trait in question. Familial aggregation is also well known for physical activity and exercise characteristics (2, 5, 21; see also Chapter 6). Given that there are multiple modes of physical activity and exercise, this chapter uses physical activity traits as the primary phenotypes in the examples described below.

Genetic modeling is a method to test hypotheses about the causes of familial aggregation and to seek understanding of its mechanisms. Until recently, the role of genetic factors was inferred on the basis of indirect evidence from family relationships – greater resemblance between genetically close persons was taken as evidence for genetic evidence but was not proof of it. Indeed, among humans, almost all the literature on the role of genetic factors in physical activity comes from modeling of family data, including special family designs such as twins. Modern molecular genetic methods are only now, in the past few years, providing robust, direct evidence for genetic factors to account for interindividual differences in human exercise and physical activity in the population at large.

This chapter will not go deeply into the theoretical framework or detailed modeling of quantitative genetics or molecular genetics. For this the reader is referred to textbooks and review articles that cover these aspects more thoroughly and some suggested further reading is provided at the end of the chapter. Further, Table 3.1 provides commonly accepted definitions of the most frequently used terms in this area. The chapter will cover family and twin designs, as they have been integral in providing evidence for genetic factors, while the chapter will conclude by summarizing molecular genetic designs, particularly the genome-wide association study (GWAS) and its recent developments. A couple of very new studies are reviewed to show how GWAS is being applied to exercise and physical activity.
Study of families

What are families?

The core biological and social unit for studies of families consists of two parents, who have one or more offspring. They are thus biologically related due to transmission of shared chromosomes and genes from parents to their children. Parents and their children are defined as first-degree relatives, and through other biological relatives, more extensive biological relationships can be identified. These include grandparents, aunts and uncles, and cousins as second-degree relatives. By sharing genes in common, family members can be assumed to resemble each other due to the nature of the time they spend together. To further characterize the nuclear family as consisting of parents and children who share the same environment for varying amounts of time. In an idealized situation, both parents raise their children together, creating a family environment with common material conditions (such as a shared home and finances), time spent together, and common rearing and family values. This family environment then acts throughout the development of the child into adolescence and at least until the child moves out of the parental home. One can then assume that the children would adopt behaviors, such as sports, that their parents engage in. The children would also adopt common parental values (e.g., the value of exercise). Thus, the parents and children resemble each other due to the nature of the time they spend together. Together with the effect of shared genes, shared social influences within families can result in familial aggregation. Overall, nongenetic environmental influences are more important during the time family members share a common household, i.e., during childhood and adolescence, but these influences can be maintained in later life either socially or even through epigenetic mechanisms. If we study nuclear families (parents and children), we can observe familial aggregation, but cannot be confident in ascribing where it arises from – is it common genes, shared exposures and experiences, or both?

In human society, there is a lot of variation in the structure and function of family units. This variation has also provided opportunities for research designs that help to distinguish the effects of genes from the rearing environment. The most used of these has been the twin study and its

Table 3.1 Definitions of selected terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive genetic variance</td>
<td>The component of variance of a phenotype that is attributable to the additive effect of both alleles at all relevant loci</td>
</tr>
<tr>
<td>Allele</td>
<td>One of two or more states in which either copy of a gene can exist. This can be a single base-pair difference or a more extensive change in the genomic sequence</td>
</tr>
<tr>
<td>Assortative mating</td>
<td>A tendency for individuals with similar genotypes to mate</td>
</tr>
<tr>
<td>Genetic variance due to dominance</td>
<td>The component of variance of a phenotype that is attributable to the interactions between alternative alleles at a locus over all relevant loci</td>
</tr>
<tr>
<td>Epistatic genetic variance</td>
<td>The component of variance of a phenotype that is attributable to the interactive effect of two or more genes</td>
</tr>
<tr>
<td>Shared environmental variance</td>
<td>The component of variance of a phenotype that is attributable to the experiences and exposures shared by family members making them more similar over the expected similarity based on their degree of gene relatedness</td>
</tr>
<tr>
<td>Unshared environmental exposure</td>
<td>The component of variance of a phenotype that is attributable to all environmental factors specific to the individual considered, and includes also measurement error</td>
</tr>
</tbody>
</table>

extensions. However, there are increasing numbers of diverse family relationships, such as single-parent families in which the other parent has been present for only a short period of the child’s life, if at all. On the other hand, blended families consist of a couple, the children they have had together, and their children from previous relationships. Such families offer opportunities but also considerable challenges for study of the behavioral determinants of physical activity and engagement in sports.

Family studies focused on first-degree relatives assess only overall “familiality,” i.e., the proportion of variance attributable to both genetic and nongenetic familial influences as well as their interactions (8). Family studies can include nuclear families (parents and their offspring) or more-extended pedigrees (grandparents, parents, offspring, aunts, uncles, cousins, etc.). Exceptionally large multigenerational pedigrees can be studied with very distantly related persons descendant from early ancestors even some centuries back (6) – such as studies conducted among the Amish, Mormons, Icelanders, and other geographically defined isolates with very little immigration until recent times. In larger pedigrees, there is more scope for distinguishing genetic and nongenetic factors, but most have focused on diseases and other conditions and there appears to be a dearth of studies of exercise characteristics in these large pedigrees.

**Quantifying family data**

For quantitative, continuous traits, familial correlations can be estimated for pairs of relatives. The intraclass correlation coefficient estimates the degree of resemblance between two family members of the same generation (such siblings or cousins), while interclass correlation estimates the degree of resemblance between two family members from different generations (such as parents and children, and grandparents and children). Higher values of these correlations are evidence of more familial resemblance; first-degree relatives are expected to show larger correlations than second-degree relatives.

Categorical traits such as the presence or absence of participation in a selected sport, or ordered categories such as the level of activity (e.g., asked as inactive, moderately active, or very active) can be analyzed using a threshold model of liability (1). In the model, the liability to the behavior is assumed to be normally distributed, and there are certain latent cut points that distinguish one category from the next. The latent liability is assumed to arise from multiple causative factors, each with a small effect, either genetic or environmental, which then give rise to the variation in liability in the population. The model precludes major genes with large effects, but these have not been found for physical activity or other exercise characteristics. The assumption of bivariate normality (for a relative pair) can be tested when there are three or more categories of the study variable, and the familial resemblance is then computed using a polychoric correlation. For a binary trait, the assumption cannot be tested.

In order to understand the roots of family aggregation, it is necessary to distinguish between effects arising from the rearing environment, and those arising from shared genes. By studying nuclear families alone, this is very difficult. For example, a recent major meta-analysis (26) of 112 studies of parent–child physical activity derived estimates ranging from 0.19 for mother–son pairs to 0.29 for father–son pairs, with parent–daughter correlations taking intermediate values. As parents and children share 50% of their segregating genes, the parent–offspring correlation can be doubled to yield an upper-bound estimate of the proportion of variance ascribed to genetic factors (i.e., 38–58% of the population variance ascribed to genetic factors) under the assumption that nongenetic factors play no role. However, the actual role of genetic factors may be much smaller, and social modeling, as proposed by the authors, might be more meaningful.
Without further information from other relative types, we cannot estimate the true proportion of variance accounted for by genetic effects in nuclear families. One option is to extend family studies to second-degree relatives. However, biologically less-related family members (such as aunts and nephews or cousins) are also less likely to have spent time in a common rearing environment, not permitting incisive study of relative contribution of genes and rearing environment. Until the advent of modern molecular genetics and the ability to genotype millions of genetic variants in large samples, adoption, twin, and twin-family designs were almost the only way to provide more insight about the relative contribution of genetic and nongenetic contributions to familiality. Therefore, in order to disentangle genes and experience, we have studied special family groups: twins who share experiences but differ in shared genes, or adoptees and their biological and foster parents who differ in their shared experience.

**Twin studies**

There is a massive twin literature on human behaviors and traits of all kinds as summarized in a recent comprehensive review article (17) indicating that nearly all studied traits have some degree of genetic influence, but environmental influences are also ubiquitous. Thus, twin studies have been the workhorse of behavioral genetic studies of exercise and physical activity.

**Biology of twins and twinning**

The biology of twinning and twins is complex. Genetically, two types of twins exist in humans. First, monozygotic (MZ) twins, as the result of an early division of the zygote into two individuals share the same genomic sequence and hence are genetically identical, thus often called identical twins. Whole-genome sequencing indicates that at most only a handful of base-pair differences are seen between twins in young MZ pairs, but with age and in certain conditions they do accumulate somatic mutations and other changes that may account for discordance between twins (12, 27). Other external influences can act on them in utero to create phenotypic differences. A notable potential source is the variation in pregnancy experiences. MZ twins, depending on the timing of the division of the zygote, may develop with their own chorion and placenta, or share the same placenta and chorion either with or without a common amnion. There is little available evidence that these placentation differences affect the results of twin studies of exercise behaviors, but few studies have rigorously assessed this due to a lack of reliable placentation information. After birth, multiple factors can generate both similarities and differences between the MZ twins.

Dizygotic (DZ) twinning is more heritable than MZ twinning and arises from the simultaneous release and fertilization of two eggs. Thus, DZ twins are genetically full siblings and each twin has their own placenta during pregnancy. Twin pregnancies result in the birth of the two individuals, generally with lower birth weight and shorter gestation than singletons. Despite the lower birth weight, twins catch up with singletons in development quickly, and as adults are very similar to singletons in mortality and virtually all behavior measures that have been examined. This supports the generalizability of twin study results to the general population.

**The classical twin model**

The realization of the existence of two types of twins more than a century ago has led to comparison of the similarity of MZ versus DZ twins for estimates on the role and relative contribution of genetic factors to interindividual differences in behavior. If both types of pairs are overall
more similar than two individuals chosen at random, this is evidence for familial aggregation. The absence of differences in mean similarity of MZ versus DZ co-twins suggests the absence of genetic influences. As MZ twins share the same genomic sequence and DZ twins share only 50% of their segregating genes on average, then in the presence of genetic factors, MZ twins would be expected to resemble each other for the trait in question more than DZ twins.

This inference about the role of genetic factors from the twin model holds under certain basic assumptions. The first is that the environmental variances in MZ and DZ pairs are equal, i.e., that MZ and DZ twins are equally correlated in the exposure to environmental experiences and factors that are relevant for the behavior being studied. This assumption has been found to hold generally (9). Differences could arise from placentation and in utero effects, or from differential parental treatment of MZ versus DZ twins. Assessment of relevant exposures and their similarity in MZ and DZ pairs is needed to test the assumption, and violation of the assumption may give rise to spuriously high estimates of genetic effects. The analysis is particularly challenging as it has been shown that measures of what are often considered environmental exposures may also have a genetic component to them. For example, smoking is considered to be an environmental exposure, and actual cigarette smoke is a true environmental toxicant exposure. However, there are interindividual differences in the amount smoked by smokers that relate to genetic differences (3).

The other key assumption of the twin model is the random mating of the parents with respect to the behavior being studied. It is assumed that the DZ pairs, like full siblings, share 50% of their segregating genes. However, if the parents resemble each other more than expected in a particular characteristic and the characteristic has a genetic basis, then this would make the siblings more similar than expected under random mating. For example, if both parents are elite athletes, the biological characteristics underlying that elite ability may be expected to be enriched in their children. This assortative mating then biases the family and twin models.

**Basic concepts of the twin model**

The twin model assumes that we are dealing with behaviors in which the genetic component is due to multiple genes, each of quite small effect. For example, new studies with molecular genetic approaches indicate that this is indeed the case for physical activity behaviors and major genes with large effects are rare or nonexistent (25). Twin study designs, basic analyses, and modeling approaches are reviewed elsewhere (4, 18, 23).

As indicated above, the genetic similarity of MZ twins is expected to be twice that of DZ twins, and so the expected genetic correlation for additive effects of MZ pairs is unity (1.0) and that of DZ twins is 0.5; for a derivation of the expectation and the underlying genetics, please see, for example, Neale and Cardon (1992) or Thomas (2003) (see “Further reading”). This source of variation in the phenotype is known as additive genetic variance (A) and it is that part of genetic effects that is transmitted from parents to offspring. Genetic effects may also be due to dominance (D), i.e., the sum of all nonlinear effects of alleles at a locus. The expected genetic correlation reflecting dominance effects is unity in MZ but only 0.25 in DZ and sibling-pairs, and this is known as genetic variance due to dominance; dominance effects do not contribute to parent–offspring similarity. Finally, there may be gene–gene interactions (i.e., epistasis) affecting phenotypes, but the classical twin model assumes that these are not present.

Nongenetic variance in a trait is divided into that shared by the twin siblings, i.e., those experiences and exposures that make them similar – termed environmental effects in common (C), and those that are not shared, i.e., unique to each twin (E). These are distinguished by whether the effects of the experiences and exposures are shared and have equal effects on both
twins, not by the actual environmental factor. Also, measurement error and random effects are part of E (i.e., unique environment exposure), so E is included in all models.

### The univariate twin model

Figure 3.1 provides a path diagram of a twin model for estimation of A, C, and E effects, commonly known as the ACE model. Based on the expectations listed above, it is possible to model data from MZ and DZ twin pairs to derive estimates of the relative contribution of genetics. This model also yields an estimate of the heritability of a trait, defined as the proportion of total variance for a trait accounted for by genetic factors. This heritability estimate may be $A/A+C+E$ but sometimes also $A+D/A+D+E$. Heritability is a relative measure, so changes in the environmental variance can change it even when no genetic changes occur. Also, heritability is not a fixed characteristic of the behavior, but rather a population-level estimate taken at a given time in a given population. For an extensive discussion of the concept of heritability, there are two excellent reviews (14, 24).

Modeling permits evaluation of which models best account for the observed variance in a trait, providing the best statistical fit. Thus, we can evaluate which of several models fits the data when a single behavior is looked at. The simplest model is unique environment (E), thus rejecting all evidence for familial effects; this model rarely happens. An AE model would specify that the pattern of twin similarity in MZ and DZ models fits a purely polygenic additive model with environmental effects unique to each twin. It would have no shared environmental effects (C) and no genetic effects due to dominance (D). The alternative models (CE, ACE, and ADE) can also be specified and tested. By comparing the fit of two models, such as ACE and AE, one

![Diagram of the univariate twin model](image-url)
can decide whether shared environment (C) effects are statistically needed to account for the data. Very often the pattern of MZ and DZ correlations is used as a starting point to guide the modeling – in a univariate case the choice is often straightforward, but not for multivariate models. Customized software for conducting twin modeling is available, but it is recommended to seek the advice and guidance of an experienced statistical geneticist knowledgeable about the model, particularly if one aims to conduct multivariate models. A number of postgraduate courses in statistical genetics, behavioral genetics, or genetic epidemiology are regularly offered that include twin modeling.

**Multivariate twin models**

At present, many types of multivariate and longitudinal quantitative genetic models are possible. The simplest form is to run the univariate model in strata of the data, say, separately among young and old participants or among women and men, to see if the variance components A, C, and E differ between strata. Twin models can address sources of sex differences in more detail if data on opposite-sex twins have been collected. The so-called sex-limitation models ask whether different sets of genes influence the behavior in men and women. If the observed correlation for the behavior is much lower in opposite-sex (male–female) DZ pairs than in same-sex DZ pairs of either sex, this result can indicate that there are sex-specific effects. For example, such effects were observed in the very large twin analysis of physical activity in the GenomEUtwin study by Stubbe et al. (21).

Multivariate models permit answering questions about the degree of shared genetic or environmental effects across related traits or over time. For example, do the same genetic and environmental effects affect physical activity in adolescence and adulthood, providing more detail on the contributors to stability and change of the behavior than can be obtained from standard longitudinal surveys and follow-up studies of unrelated individuals? While genes do not change in structure over time, their expression and activity do. These alterations in expression and activity permit novel genetic effects to arise as people age and develop. Another multivariate question may address whether two or more facets of exercise behavior are correlated due to shared genes or due to shared nongenetic influences. Such genetic correlations have been estimated until recently solely by family and twin data (15), but now molecular genetic data can provide more information on the genetic structure of related phenotypes through GWAS studies (see later section). This information can then guide phenotype development, in terms of combining or otherwise constructing variables with the most genetic information.

A basic starting point for multivariate modeling is the Cholesky decomposition or lower-triangle model. Figure 3.2 provides a schematic of the model, and the legend to the figure provides more details. An advantage of the model is that it is solvable and provides a saturated model against which models with fewer parameters and/or paths can be tested. Also, it is easy to apply to data sets if variables or measurement points of the same variable are relatively few. The model produces genetic and environmental correlation matrices of the study variables, which can then be compared to the genetic correlations derived from GWAS analyses. Because the model decomposes variances in each component from left to right into more restricted factors, the ordering of variables is important to consider.

This schematic illustrates a full model and can be used to derive matrices of the genetic and environmental correlations between variables. These may represent different variables, such as levels of physical activity during work, leisure, and sports (9) or the same variable at different points in time; for an illustrative full model of four measures of physical activity from adolescence to young adulthood see Figure 1 in Aaltonen et al. (1)
In addition, the full model can be tested against models that specify selected values for the paths. For example, setting path $a_{12}$ to zero would test whether there is a significant genetic correlation between VAR1 and VAR2.

For longitudinal models using the Cholesky approach, time/age is the natural ordering given that all factors impact only on current and later time points. Given its nature, the Cholesky decomposition can accommodate different patterns of change but is not directly falsifiable. In longitudinal contexts, the Cholesky model provides information about the relative contribution of genetic and environmental factors to the tracking of trait. Thus, it is a useful starting point for analyses when little is known about the longitudinal process and contributing variables. From this exploratory tool one can expand to other longitudinal models such as the simplex model or the longitudinal growth model (7). The longitudinal growth model allows us to ask whether genetic factors contribute to variation in both initial level as well as rate of change of the study variable and are those correlated. In other words, do the same genes influence both initial level (say, initial fitness) and rate of change (improvement in fitness over a training period)? In contrast to the Cholesky decomposition, the longitudinal growth model enables predictions on future time points.

Multivariate models may also be much more specific in their construct. Among the best known are the common pathways model and the independent pathways model. The common pathways model specifies that there is one common latent trait through which the genetic and environment variances act. For example, there may be four or more correlated physical activity measures. In the common pathway model, the latent variable has regression paths to all measured variables, and this common latent variable is then decomposed into genetic and environmental factors. Thus, unlike the Cholesky, the genetic and environmental components do not act directly on the variables but run through the common latent variable. In addition, each variable has residual genetic and environmental components to them. The common pathway model is a useful tool for measurement construction and theory development (for more details, see Neale and Cardon (1992) in “Further reading”). The independent pathway model can be

---

**Figure 3.2** Schematic diagram of the essential feature of the Cholesky decomposition multivariate model. The figure shows an AE model applied to three variables (VAR1, VAR2, VAR3) for which twin data are available. The univariate twin model in Figure 3.1 is extended so that paths from the latent genetic ($A$) and environmental ($E$) variation also cover relationships between variables. Paths $a_{11}$, $a_{22}$, and $a_{33}$ represent variable-specific effect of genes on the trait (variable), and correspondingly $a_{12}$ denotes the degree to which genetic effects affecting VAR1 also affect VAR2, and so on. The corresponding relationships hold for the environmental effects.
considered a special case of the common pathway model. It posits three latent variables for A, C, and E effects, such that they all correlate with the measured variables. Other models that have been developed for use with twin data include the direction of causation model, a sibling interaction model, and models to assess rater bias (for more details, see Neale and Cardon (1992) in “Further reading”). The latter can be used to assess the effect of having two more raters provide information on several family members, typically parents reporting on both twins.

Finally, these models permit assessment of gene–environment interactions, asking whether a known exposure modifies the impact of genetic variation. An example is the well-replicated finding that physical activity buffers the impact of genes on obesity; among sedentary persons, genes account for a much larger fraction of variance in body mass index (BMI) than among physically active persons. This observation from twin studies (16) has now been extensively corroborated using measured genotypes associated with BMI (11, 20).

MZ discordant pairs to study causal associations

When studying the association of a putative risk factor or exposure on an outcome, the association may be causal, i.e., implying that reducing the exposure would lead to a reduction in the outcome. Alternatively, it can be due to confounding. Measurement of known confounders and adjustment for them in statistical models has been the standard approach in observational epidemiology, but they cohort or case–control studies. However, not all confounders are known or can be measured. Genetic factors underlying athletic ability may be shared with genetic effects on obesity. Thus, the association of exercise with weight gain may be causal or it may be accounted for by known and unknown confounders. As exercise itself is, in part, heritable and now that genes for various aspects of exercise behavior are beginning to be identified, there is potentially confounding due to shared genes, as illustrated by the very latest molecular genetic studies.

A study design to examine nongenetic (possibly causal) associations that control for genetic variation uses exposure-discordant twin pairs. As MZ twins share the same genomic sequence, all difference between the twins arise from nongenetic causes, taken very broadly to include all small random events. If we can identify twin pairs in which one exercises and the other does not, then a test of the causal hypothesis of the association between exercise and weight gain would be to study weight development in a large number of such pairs discordant for exercise. If the MZ co-twins who are physically inactive have significantly more weight gain than their co-twins who are active, strong evidence would be provided to support a causal hypothesis. The design controls for genetic background, but also for sex and age effects as well as the exposures that both twins have shared, such as numerous childhood and adolescent exposures from their common childhood home. The challenge in such studies is often finding sufficient numbers of pairs who are truly discordant. Exposing one twin to an experimental intervention while keeping the other twin as a control is a strong design that combines twins with an intervention. Another design is to examine the children of discordant pairs, to see if the discordance is transmitted in equal measures to the children of DZ pairs (biologically cousins) and of MZ pairs (biologically half-siblings but socially cousins).

Adoption studies

Theoretically, adoption studies are a very powerful design for disentangling genetic and nongenetic influences. A biological parent–adoptivee correlation, when the adopted child has been raised since birth by unrelated adoptive parents, is strong evidence for genetic effects.
However, in practice the adoption may not have occurred immediately after birth, and exposures during pregnancy may be transmitted epigenetically later into life. Adoptive parent-adoptiveee correlations are indicators of the effects of the rearing environment. Adoption studies of physical activity and related traits are rare.

**Twin family designs**

A central limitation of the twin design is that it does not directly tell about the transmission of effects from one generation to the next. Therefore, there are a number of designs combining family and twins, where the twins can be either in the offspring generation or as parents. Including two generations permits one to not only model the transmission of the study trait from parents to offspring, but also to include the relationship of the parents. As discussed above, the twin model assumes that there is random mating with respect to the trait being studied. When this is not the case, the twin model results may be biased. Parents may resemble each other more than expected if they come from a microenvironment that is more homogeneous than in the general population. This could, for example, be a religious affiliation, social strata (of deprivation, for example), or geographical locality. This results in what is known as social homogamy, whereby the parents are more alike for social reasons rather than genetic. If spouses have chosen their mates on the basis of their actual personal characteristics, this phenotypic assortment means that there is also some degree of genetic relatedness. For example, spouses have more similar heights than expected, i.e., height shows phenotypic assortment and this would increase the expected genetic relatedness with respect to height in their offspring. One concern in two-generational studies results from the fact that the parents and offspring are often studied at different ages, and the actual behavior may have different genetic and environmental determinants at different ages; simple adjustment for age may not suffice. Ideally, the study behavior should be assessed at the same age in that case, but that may mean waiting for several decades in a longitudinal study of such families. Three-generation studies are even rarer (2).

**Genome-wide association studies**

In contrast to family studies, molecular genetic studies permit the study of the role of genetic factors in unrelated persons. Until the sequencing of the human genome, the available tools were limited and findings regarding the genetic basis of physical activity sparse and inconsistent. Some candidate genes have been put forward and studied, but on their basis, it has not been possible to assess how much human genetic variation actually contributes to interindividual differences in exercise ability, physical activity, and related phenotypes (19). In the past decade, the experimental design of GWAS has led to increasing findings on the role of specific genes and genetic variation overall in many diseases as well as normal physiological and behavioral traits. Peter Visscher and his colleagues reviewed recently the first 10 years of GWAS discoveries (25).

The design of GWAS is fairly straightforward as current genotyping technologies permit the efficient and cheap genotyping of hundreds of thousands of genetic variants, mostly SNPs, across the human genome. These serve as markers for blocks of DNA that are mostly inherited together and therefore there is no need to fully sequence all the genomes under study. GWAS is informative of a large fraction of human genetic variation and permits identification of small regions of association. Advances in technology, statistical genetics, and genomic biology have permitted researchers to make strong advances in the use of GWAS for understanding population and complex-trait genetics and the biology of the conditions. Two central advances have been the availability of ever larger sample sizes on the one hand, and better characterized
phenotypes on the other. Unlike studies done on obesity, diabetes, height, and education which have yielded hundreds of genetic loci linked to these traits, the study of physical activity has been challenging, and until recently very little progress in GWAS of physical activity was achieved (19).

Using data on nearly half a million persons from the UK Biobank (22), Tikkanen et al. identified 64 loci associated with handgrip strength that replicated using 223,315 individuals for discovery and 116,610 for replication. The strongest association was with FTO, a locus first associated with obesity. The handgrip loci findings extended those reported earlier in the CHARGE consortium (13) and will help to understand the reasons for variation in handgrip strength in the population. We can expect that as other objectively measured indices of exercise characteristics such as time spent on various physical activities become available for genomic studies, more loci will be found due to more accuracy in the studied phenotype. The genome-wide significant loci for handgrip strength accounted for only 1.7% of variance in measured physical activity, but overall common variants accounted for some 13% of variance. This is substantial but considerably less than implied by twin and family studies.

The UK Biobank data was also at the core of another large GWAS study (10) that extended the analyses with data from the ARIC study, with robust findings for eight loci for moderate to vigorous physical activity. The strongest associations were seen with APOE, a lipid and dementia-associated gene, and with CADM2, previously associated with risk-taking behaviors. The authors also found that the physical activity measures show genetic correlations with educational attainment and various obesity-related traits. In summary, large-scale GWAS of self-reported and objectively measured physical activity traits are now producing the first consistent and replicated associations with a handful of genes. They also show that the common genetic variants overall account for as little as 10% of the variance in these traits, a share that may rise somewhat as more variants are genotyped and imputed more accurately and phenotyping improves. Finally, the examined physical activity GWAS analyses reveal expected genetic correlations with other traits such as education, obesity, and risk-taking behaviors.

The future in the era of molecular genetics

When considering overall genetic influences, it should be kept in mind that heritability of a trait in a population is not necessarily static. For example, if important, influential environmental and behavioral exposures change over time and explain more or less of the population variance in a trait. As such, with all else being equal, there will be inverse variations in the heritability of the trait. In addition, heritability of a trait may vary by sex and age and other covariates. While heritability is an important concept related to the overall magnitude of genetic versus environmental influences on a trait at the population level, it must be recognized that there is a highly complex interplay between genetic and environmental factors, as is seen in epigenetics and gene expression, for example.

Beyond heritability estimates, classic twin studies with multivariate analyses have allowed the examination of shared genetic influences between phenotypes to test hypotheses regarding shared etiologies or possible pathways of genetic influences on clinical phenotypes. Finally, molecular genetics are now bringing a host of new approaches and tools to better understand the genetic architecture of exercise characteristics and gain insights to relevant biological processes.
Human systems genetic modeling

Further reading


References


