Chapter 20
Statistical Considerations in the Analysis of Rare Variants

Debashis Ghosh and Santhosh Girirajan

Abstract Recently, whole-genome and whole-exome sequencing has begun to demonstrate success in the identification of disease-causing genes. Many of these genes exhibit abnormal genetic behavior and low prevalence in the population; these molecules are commonly referred to as rare variants. In this chapter, we provide an overview of rare variants and their scientific relevance in medicine and public health. We then provide a review of existing methods for association, primarily focusing on the sequence kernel association test (SKAT) and related methods. These procedures are related to kernel machines, which we will also describe. Finally, we discuss the implications of rare variants in terms of multiple testing.

20.1 Introduction

Large-scale genomics has been at the forefront of science and medicine over the last decade. The advent of high-throughput technologies including single nucleotide polymorphism (SNPs) microarrays, array comparative genomic hybridization and genome sequencing have enabled rapid discovery of genetic variants varying in size and frequency [18]. Copy number variants are deletions and duplications in the genome that constitute the most genetic variation, in total base pairs, between individuals [35]. Classically, disease-association studies involved evaluation of either variants of high frequency in the population, also termed common variants, or variants of low frequency or rare variants. In this chapter, we will consider analysis of rare variants with specific focus on copy number variants. One of the key statistical challenges in the analysis of rare variants is that they have small population prevalences. If we view the rare variants as predictors that we

D. Ghosh (✉) • S. Girirajan
Penn State University, University Park, State College, PA 16801, USA
e-mail: ghoshd@psu.edu; sxg47@psu.edu

S. Datta and D. Nettleton (eds.), *Statistical Analysis of Next Generation Sequencing Data*, Frontiers in Probability and the Statistical Sciences, DOI 10.1007/978-3-319-07212-8_20,

wish to associate with a phenotype, then they in fact contain very little statistical information. To illustrate the idea, suppose we wish to regress the phenotype on a rare variant that we treat as binary, where zero indicates absence and one indicates presence. We assume that the regression model is linear. Then it can be shown analytically that the information about the regression coefficient in such a setup is maximized when half of the subjects have the rare variant and half do not. However, by definition, for rare variants, a majority of subjects will not have the rare variant, the implication being that we are in an inherently low-power situation. Thus, it is necessary to begin to think about pooling information in various ways; this will be one of the themes explicated on in the chapter.

The structure of this chapter is as follows. In Sect. 20.2, we provide some biological background to rare variants. Section 20.3 reviews association methods for the analysis of rare variants and in particular focuses on the sequence kernel association test (SKAT) [56] and its extensions. The SKAT methodology is based on the kernel machine framework originally proposed by Liu et al. [33, 34], so we also expand on this. Finally, we discuss the multiple comparisons problem and how its consideration needs to be modified for the rare variant problem in Sect. 20.4. This chapter concludes with some discussion in Sect. 20.5.

20.2 Biological Background

Association of disease genes to phenotypic traits or overt disease has been carried out with the discovery, characterization or genotyping of variants. Genetic studies have relied upon identifying causative genes by finding genetic variants, common ($>1\,\%$ or $>5\,\%$ in the population) or rare, and whether they are enriched in cases compared to controls. Common variants are contributed by alleles that originated during the development of humans and are therefore shared between different human populations [39]. These variants constitute most of the human genetic variation, in frequency, and are also represented as SNPs that tag specific haplotypes mapped by the HapMap project [10, 11]. While technologies and genetic methods have concentrated on implicating common or rare variants of extreme size for disease etiology, identification and characterization of variants of intermediate size and frequency remains a challenge [50].

The basis for rare variants can be best understood in a historical context. In the field of human genetics of complex traits, the dominant school of thought in the early 2000s was based on the so-called common disease-common variant (CDCV) hypothesis [45, 46]. This framework postulated that for many diseases, multiple SNPs would be needed to explain a large percentage of variation in the phenotype. Identification of SNPs in linkage disequilibrium or functional variants in the neighborhood of causative genes is the basis of genome-wide association studies (GWAS) [22]. This thought very much influenced the design of GWAS and the technology used to measure DNA variation. The dominant platform for measuring SNPs was the microarray platform, which was being used simultaneously

for measuring transcript mRNA expression. The major company that developed the SNP microarray platform was Affymetrix (http://www.affymetrix.com), and the DNA variations selected to be on the chip primarily represented variants that satisfied the CDCV hypothesis, i.e., all of the variants had to be sufficiently present in the population. In particular, what tended to be excluded from the SNP microarrays were DNA variants where the less prevalent form had a population prevalence (termed minor allele frequency) that was less than 5 %.

Currently, there have been over 2000 GWAS studies that have been conducted in humans (genome.gov, 2013) with a major finding that DNA variations in the form of SNPs can only explain a limited amount of variation for several human disease associated phenotypes [37, 38]. GWAS has been only successful in studies on type-2 diabetes, age-related macular degeneration, coronary artery disease, and Crohn disease as well as for obesity and height. These studies were not successful for a majority of common complex diseases including neurodevelopmental disorders such as autism, schizophrenia, and epilepsy. This has led to consideration of reasons for the missing heritability [38]. The difficulty of achieving statistical power to identify multiple loci of small effect sizes is considered as a major factor. Other factors, not considered in traditional GWAS studies [5], such as gene-gene interactions and gene-environment interactions, have also been proposed [57].

Rare variants, on the other hand, tend to have much bigger effects than the DNA variants identified from first-generation GWAS. This alternative model is termed common disease rare variant (CDRV). From an evolutionary point of view, these variants are under strong selection and their frequency in the population is maintained by de novo mutations. In fact, new germline mutations arise constantly, based on the underlying sequencing architecture or age of the parents, at a rate of about 61 base pairs for single nucleotide mutations [4] and 16–50 kbp per diploid genome [24]. Genetic associations for copy number variants have met with higher success for variants of low frequency. These variants were classically associated with clinically recognizable syndromes such as 7q11.2 deletions in individuals with features of Williams syndrome, 22q11.2 deletions in individuals with features of DiGeorge syndrome, and 17p11.2 deletions in Smith-Magenis syndrome [19].

Developments in microarray technology and rapid incorporation of high-throughput genotyping in diagnostic laboratories have resulted in the identification of about two dozen CNVs that are strongly enriched in affected cases with neurodevelopmental impairments compared to controls. However, extensive phenotypic heterogeneity even in individuals carrying the same CNV has complicated further analysis. For example, the 16p11.2 deletion was originally associated with autism, but was later identified to be enriched in individuals with intellectual disability, epilepsy, schizophrenia, and obesity [40, 47, 55]. Comparison of CNV load (measured as the proportion of population carrying a deletion or duplication of a particular size) across cohorts of affected population suggests that the CNV load correlates with the severity of neurodevelopmental disorders [17]. Similarly, phenotypic variability and severity associated with a specific disease-associated CNV can be also explained by rare variants in the genetic background [19, 21]. These variants modulate the ultimate phenotypic expression either by additive

or synergistic effect, in genetic terms, in a digenic or oligogenic manner [53] Genome sequencing has made tremendous strides in finding the missing heritability. Sequencing of the protein coding sequences in the genome for neurodevelopmental disorders has identified several rare, de novo variants that cluster in pathways related to nervous system development, maturation, and maintenance [16,41,42,48]. These studies have, however, revealed a complex genetic basis for common diseases; for example, recent estimates suggest that a minimum of 1000 genes as causal for autism. These disorders can be explained by an infinitesimal model consistent with the role of multiple rare variants in complex disease [14]. According to this model the genetic etiology can be explained by a hybrid of the two models. The challenge therefore lies in understanding how these variants work together in causing the disease rather than if they are rare or common [14].

The implications of rare variants for medicine and public health are potentially quite paradigm shifting. Both disciplines have placed a tremendous emphasis on evidence that has been gathered from consideration of population-based analyses of biomedical data. However, rare variants are predictor variables that by definition are quite individual-specific. Simply based on their prevalence, standard population-based analyses will have low power to detect them. The rare variant paradigm also is quite in tune with the notion of personalized medicine, where treatments and/or interventions would be tailored to the particular variant present in the individual. Very broadly speaking, this is consistent with the patient-centered/patient-oriented paradigm in medicine that has been developing over the last few years. Figure 20.1 describes the genetic spectrum for disease that analysts must contend with.

20.3 Kernel Machine Methodology

20.3.1 Setup and Review of Methods

We now describe tests of associations between rare variants and a phenotype. To make the ideas concrete in this chapter, we will suppose that we have available $(Y_i, \mathbf{G}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$ for n subjects, which is a random sample from $(Y, \mathbf{G}, \mathbf{Z})$. Here, Y denotes the phenotype, $\mathbf{G}$ is a $p-$dimensional vector for the genotypes for the p variants within a region, and $\mathbf{Z}$ is a $q-$dimensional vector of confounding variables to adjust for. Here and in the sequel, we will assume that each component of $\mathbf{G}$ will count the number of minor alleles. We can postulate a class of regression models for Y given $\mathbf{G}$ and $\mathbf{Z}$; a standard one would be to postulate a generalized linear model:

$$h(E[Y_i|\mathbf{G}_i, \mathbf{Z}_i]) = \alpha_0 + \alpha^T \mathbf{G}_i + \beta^T \mathbf{Z}_i, \tag{20.1}$$

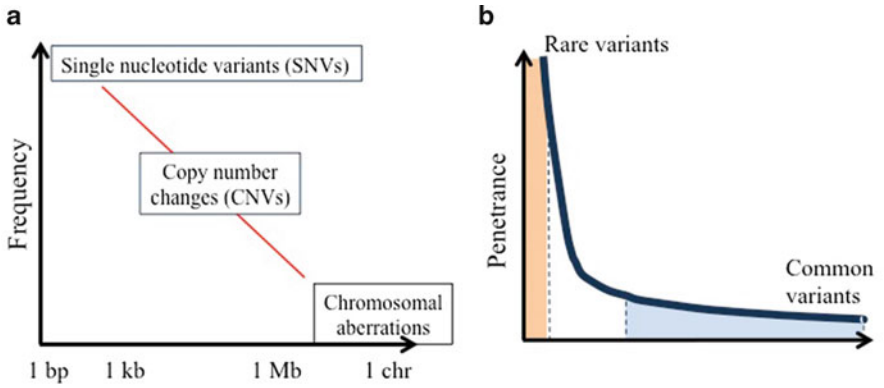

Fig. 20.1 (**a**) Size of variant. Genetic variants are ordered by size on the horizontal axis versus frequency on the vertical axis. Note that single nucleotide variants or more specifically single nucleotide polymorphisms (used for GWAS studies) are more frequent than copy number variants (i.e., deletions and duplications) in the human genome. The large chromosomal aberrations such as trisomies and monosomies are rarer and are the cause for severe developmental disabilities. (**b**) Frequency of Variants. Variants can be classified by the frequency (on the horizontal axis) and its effect, i.e. penetrance (proportion of individuals carrying a variant also manifesting a phenotype) on the vertical axis. Note that rare variants (typically $<0.1\,\%$ to $<5\,\%$) are highly penetrant, associated with severe developmental disorders, while common variants have modest effect. Variants of intermediate frequency are currently missed by most studies. Current studies also suggest that multiple rare alleles interacting in common or related pathways are responsible for several human disorders

where $(\alpha_0, \alpha, \beta)$ are the regression coefficients to be estimated, and h is a link function. Note that the current model (20.1) can allow for both continuous and binary phenotypes.

While model (20.1) is quite standard in the statistical literature, new issues arise when attempting to apply it to rare variant data. First, due to the sparsity of $\mathbf{G}$ the components of α will not be estimated very well. Due to this as well as for computationally feasibility, there has been a reliance on the use of score-based tests, which will be less sensitive to this type of sparsity relative to a Wald test, for example. A second problem is one of power. Models such as (20.1) that treat the genetic effects as fixed effects will have lower power due to the number of degrees of freedom for jointly testing $\alpha = \mathbf{0}$. To circumvent this issue, two classes of approaches have been developed. The first includes methods that can broadly interpreted as collapsing methods [31, 32, 36, 44]. These tests effectively reduce $\mathbf{G}$ into a scalar quantity $\mathbf{G}^*$ and to fit model (20.1), where $\alpha^T \mathbf{G}_i$ is replaced by $\gamma \mathbf{G}_i^*$. The reduction to a one-dimensional quantity leads to a reduction in the number of parameters and a potential gain in power.

Collapsing approaches will work in situations in which the components of $\mathbf{G}$ have effects on Y that are in the same direction. However, it might be the case that this assumption is not true. The SKAT methodology of Wu et al. [56] then becomes quite useful in this regard. In particular, it generalizes (20.1):

$$h(E[Y_i|\mathbf{G}_i, \mathbf{Z}_i]) = \alpha_0 + f(\mathbf{G}_i) + \beta^T \mathbf{Z}_i, \tag{20.2}$$

where now f is a flexible non-linear function of the rare variants. This is a special case of the kernel machine framework originally proposed by Liu et al. [33, 34]. We will describe the technical details of the approach in the next section. We point out here that the rare variant effects are allowed to be much more flexible than in (20.1). Further, the test of the genetic effect in (20.2) is identical to testing for a random effect being zero for a certain linear mixed effects model. This amounts to an effective shrinking of the degrees of freedom and allows for pooling of information across the rare variants. The score test amounts to a quadratic form that takes deviations of the individual rare variant effects and squares them.

20.3.2 Kernel Machines: Technical Details

In this section, we review the technical details behind the SKAT model in the case of h in (20.2) being the identity link. This material is intended for mathematically minded readers and can be skipped upon initial reading of this chapter. Recall the model from the previous section with $\alpha_0 = 0$:

$$Y_i = \beta^T \mathbf{Z}_i + f(\mathbf{G}_i) + e_i, \tag{20.3}$$

where β is a $q \times 1$ vector of regression coefficients, $f(\mathbf{G}_i)$ is an unknown centered smooth function, and the errors e_i are assumed to be independent and follow $N(0, \sigma^2)$. Here, we are centering the response so that there is no intercept term as in (20.2). Note that when $f(\cdot) = 0$, (20.3) reduces to the standard linear regression model.

20.3.2.1 Function Space of $f(\mathbf{G})$: Specification

We assume the nonparametric function $f(\mathbf{G})$ lies in a function space $\mathscr{F}$ spanned by a set of basis functions $\{\phi_1(\mathbf{G}), \cdots, \phi_j(\mathbf{G}), \cdots, \phi_J(\mathbf{G})\}_{j=1}^J$ such that any function in the space $\mathscr{F}$ can be written as $f(\mathbf{G}) = \sum_{j=1}^J \omega_j \phi_j(\mathbf{G})$ for some constants $\{\omega_j\}_{j=1}^J$. Note that the set of basis functions can be finite ($J < \infty$) or infinite ($J = \infty$). In the machine learning literature, such basis functions are called *features*.

Specification of a function space using basis functions or features might be complicated since explicit expressions of features are required and the number of features might be high or even infinite. An alternative way to conveniently specify a function space is to use a kernel function $K(\mathbf{G}, \mathbf{G}')$ instead of the basis functions. Specifically, a kernel function $K(\mathbf{G}, \mathbf{G}')$ is a bounded, symmetric, positive function satisfying

$$\int K(\mathbf{G}, \mathbf{G}') f(\mathbf{G}) f(\mathbf{G}') d\mathbf{G} d\mathbf{G}' \geq 0, \tag{20.4}$$

for any arbitrary square integrable function $f(\mathbf{G})$ and all $\mathbf{G}, \mathbf{G}' \in R^p$. The kernel function can be viewed as a measure of similarity between two values of the covariate vector $\mathbf{G}$ and $\mathbf{G}'$. Following from the Mercer Theorem (e.g., see p. 33 of [6]), any kernel function satisfying some regularity conditions implicitly specifies an unique function space spanned by a particular set of basis functions (features), and vice versa. Before formally defining such a function space, we give a few examples.

1. *The dth degree Polynomial Kernel*: $K(\mathbf{G}, \mathbf{G}') = [\mathbf{G} \cdot \mathbf{G}' + 1]^d$, where $\mathbf{G} \cdot \mathbf{G}' = \sum_{k=1}^{p} g_k g'_k$ denotes the dot product. Recall that g represents components of the vector $\mathbf{G}$ in (20.3). This dth degree polynomial kernel generates the function space $\mathscr{F}$ spanned by all possible dth order monomials of the components of $\mathbf{G}$. For example, if $d = 1$, the first polynomial kernel generates the linear function space with basis functions $\{z_1, \cdots, z_p\}$. If $d = 2$, the second polynomial kernel corresponds the quadratic function space with basis functions $\{z_k, z_k z'_k\}$ $(k, k' = 1, \cdots, p)$, i.e., the main effects, all two-way interactions and quadratic main effects. Note that the function space determined by the dth degree polynomial kernel is of finite dimension.
2. *The Gaussian Kernel*: $K(\mathbf{G}, \mathbf{G}') = \exp\{-||\mathbf{G} - \mathbf{G}'||^2/\rho\}$, where $||\mathbf{G} - \mathbf{G}'|| = \sum_{k=1}^{p} (g_k - g'_k)^2$. The Gaussian kernel generates the function space spanned by radial basis functions, whose nice properties can be found in Bühmann [3]. The function space determined by the Gaussian kernel is of infinite dimension.
3. *The identity by state kernel*: Kwee et al. [26] propose the use of the concept of identity by state to define a new kernel. The kernel is given by

$$K(\mathbf{G}, \mathbf{G}') = \frac{\sum_{s=1}^{p} IBS(\mathbf{G}_s, \mathbf{G}'_s)}{2p},$$

   where the IBS function denotes the number of alleles shared identically by state at position s.

The above examples suggest that the choice of a kernel function determines which function space one would like to use to approximate $f(\mathbf{G})$. The dimension of the function space defined by a kernel function $K(\cdot, \cdot)$ is determined by the dimension of the eigenfunctions of $K(\cdot, \cdot)$. The use of a kernel to specify a function space avoids specifications of complicated basis functions (features) and inner products. One will see in the next section that it has significant computational advantages in high dimensional problems. It should be noted that the term "kernel" here has a rather different meaning from that used in the kernel smoothing literature. A commonly used function space defined by a kernel is a Reproducing Kernel Hilbert Space (RKHS), which we label as $\mathscr{F}_K$. Technical details on RKHS can be found in Wahba [54] or Chapter 3 of Cristianini and Shawe-Taylor [6].

20.3.2.2 Primal and Dual Representations of $f(\mathbf{G})$

Any function $f(\mathbf{G})$ in the function space $\mathscr{F}_K$ defined by a kernel $K(\cdot,\cdot)$ can have a primal representation directly using the basis functions (features) of $\mathscr{F}_K$, and it can equivalently have a dual representation using the kernel function $K(\mathbf{G},\mathbf{G}')$ directly. Specifically, for an arbitrary function $h(\mathbf{G}) \in \mathscr{F}_K$, its primal representation takes the form

$$f(\mathbf{G}) = \sum_{j=1}^{J} \omega_j \phi_j(\mathbf{G}) = \phi(\mathbf{G})^T \omega, \tag{20.5}$$

where $\phi(\cdot) = \{\phi_1(\cdot), \cdots, \phi_J(\cdot)\}^T$ is a $J \times 1$ vector of the standardized orthogonal basis functions (features), i.e., standardized Mercer features of the function space $\mathscr{F}_K$, and $\omega \equiv (\omega_1, \cdots, \omega_J)'$ is a vector of some constants. The square norm of $f(\cdot)$ can be written as

$$\|f\|^2_{\mathscr{F}_K} = \sum_{j=1}^{J} \omega_j^2 = \omega^T \omega. \tag{20.6}$$

Alternatively, the same $f(\mathbf{G})$ can be equivalently written in a dual representation using the kernel function $K(\cdot,\cdot)$ directly as

$$f(\mathbf{G}) = \sum_{l=1}^{L} \alpha_l K(\mathbf{G}_l^*, \mathbf{G}), \tag{20.7}$$

for some integer L, some constants $\alpha_1, \ldots, \alpha_L$ and some $\{\mathbf{G}_1^*, \cdots, \mathbf{G}_L^*\} \in R^p$. For justifications of these results and more details about the RKHS, see Cristianini and Shawe-Taylor (2000[6], Chapter 3).

Estimation of β and $f(\cdot)$ proceeds by maximizing the scaled penalized likelihood function

$$-\frac{1}{2}\sum_{i=1}^{n}\{Y_i - \beta^T \mathbf{Z}_i - f(\mathbf{G}_i)\}^2 - \frac{1}{2}\lambda \|f\|^2_{\mathscr{F}_K}, \tag{20.8}$$

where λ is a tuning parameter and controls the tradeoff between goodness of fit and complexity of the model. When $\lambda = 0$, the model interpolates the data, whereas when $\lambda = \infty$, the model reduces to a simple linear model.

While the function (20.8) is hard to optimize directly, we introduce the Lagrangian multiplier (also called the dual parameter) γ to obtain

$$\mathscr{L}(\omega, \beta, e, \gamma) = -\frac{1}{2}\sum_{i=1}^{n} e_i^2 - \frac{1}{2}\lambda \omega^T \omega + \sum_{i=1}^{n} \gamma_i \{\beta^T \mathbf{Z}_i + \phi(\mathbf{G}_i)^T \omega + e_i - Y_i\}. \tag{20.9}$$

The dual problem is formulated by constructing an objective function by removing the high-dimensional primal coefficient vector ω and the constraint parameters e from $\mathscr{L}(\omega,\beta,e,\gamma)$ and writing $\mathscr{L}(\omega,\beta,e,\gamma)$ as a function of β and the dual parameter vector γ only. We will see that the resulting estimators $\hat{\beta}$ and $\hat{\gamma}$ can be expressed as a function of some kernel function $K(\cdot,\cdot)$. One can then conveniently obtain the maximizer of the original primal problem $\hat{\omega}$ and then $\hat{f}(\mathbf{G})$ at any arbitrary $\mathbf{G}$ as a function of the kernel function $K(\cdot,\cdot)$.

Specifically, the dual problem to minimizing (20.8) is

$$\min_{\beta,\gamma} \mathscr{Q}(\beta,\gamma) \tag{20.10}$$

where $\mathscr{Q}(\beta,\gamma) = \sup_{\omega,e} \mathscr{L}(\omega,\beta,e,\gamma)$. Note that (20.10) is an unconstrained optimization problem, and the number of unknown parameters depends only on β and the dual parameters γ, whose dimension is equal to the sample size n, often much smaller than J, the dimension of the primal vector ω. Therefore the dual formulation (20.10) effectively transforms the often infinite-dimensional optimization problem (20.8) into a finite-dimensional problem.

To obtain $\mathscr{Q}(\beta,\gamma)$, one differentiates $\mathscr{L}(\omega,\beta,e,\gamma)$ with respect to e and ω and sets the derivatives to zero. We have

$$\begin{aligned} \hat{e} &= \gamma \\ \hat{\omega} &= \lambda^{-1} \sum_{i=1}^{n} \gamma_i \phi(\mathbf{G}_i). \end{aligned} \tag{20.11}$$

Substituting $\hat{\omega}$ and $\hat{e}$ into $\mathscr{L}(\cdot)$, some calculations give

$$\mathscr{Q}(\beta,\gamma) = (Y - \beta^T \mathbf{Z})^T \gamma - \frac{1}{2}\gamma^T \left(I + \lambda^{-1} K\right) \gamma \tag{20.12}$$

where $Y = (Y_1,\cdots,Y_n)^T$ and $\mathbf{Z} = (\mathbf{Z}_1,\cdots,\mathbf{Z}_n)^T$, K is an $n \times n$ matrix whose (i,i')th element is $K(\mathbf{G}_i,\mathbf{G}_{i'})$, the kernel function evaluated at the pair of the design points $(\mathbf{G}_i,\mathbf{G}_{i'})$. Note that the kernel matrix K measures the similarity among the covariate values $(\mathbf{G}_1,\cdots,\mathbf{G}_n)$. One can see that even when p (the dimension of $\mathbf{G}$) or J (the dimension of the feature space) is high, the dimension of K is not affected by p and J and remains the same as the sample size n.

Differentiating $\mathscr{Q}(\beta,\gamma)$ with respect to γ and β, some calculations give

$$\hat{\beta} = \left\{\mathbf{Z}^T (I + \lambda^{-1} K)^{-1} \mathbf{Z}\right\}^{-1} \mathbf{Z}^T (I + \lambda^{-1} K)^{-1} Y \tag{20.13}$$

$$\hat{\gamma} = (I + \lambda^{-1} K)^{-1} (Y - \hat{\beta}^T \mathbf{Z}). \tag{20.14}$$

Plugging (20.14) into (20.11), we have

$$\hat{\omega} = \lambda^{-1}\{\phi(\mathbf{G}_1), \cdots, \phi(\mathbf{G}_n)\}\hat{\gamma} = \lambda^{-1}\{\phi(\mathbf{G}_1), \cdots, \phi(\mathbf{G}_n)\}(I + \lambda^{-1}K)^{-1}(Y - \hat{\beta}^T\mathbf{Z}).$$

It follows that the nonparametric function $f(\cdot)$ evaluated at the design points $(\mathbf{G}_1, \cdots, \mathbf{G}_n)^T$ is estimated as

$$\hat{f} = \lambda^{-1}K\hat{\gamma} = \lambda^{-1}K(I + \lambda^{-1}K)^{-1}(Y - \hat{\beta}^T\mathbf{Z}). \tag{20.15}$$

The estimator of the nonparametric function $f(\cdot)$ at an arbitrary $\mathbf{G}$ is

$$\hat{f}(\mathbf{G}) = \phi(\mathbf{G})^T\hat{\omega} \tag{20.16}$$

$$= \lambda^{-1}\{K(\mathbf{G}, \mathbf{G}_1), \cdots, K(\mathbf{G}, \mathbf{G}_n)\}(I + \lambda^{-1}K)^{-1}(Y - \hat{\beta}^T\mathbf{Z}). \tag{20.17}$$

Note that the estimators $\hat{\beta}$ and $\hat{f}(\cdot)$ in (20.13) and (20.15) are the maximizer of the original primal problem. Examination of equations (20.13) and (20.17) suggests that the estimators $\hat{\beta}$ and $\hat{f}(\cdot)$ are both conveniently evaluated using the kernel function $K(\cdot, \cdot)$ and do not require specifying the high (maybe infinite) dimensional basis functions (features) $\{\phi(\mathbf{G})\}$. This means one simply summarizes the similarity of high-dimensional covariates $(\mathbf{G}_1, \cdots, \mathbf{G}_n)$ using a kernel matrix K, then calculates $\hat{\beta}$ and $\hat{f}(\cdot)$ by inventing an $n \times n$ matrix involving the kernel matrix K, which is of the dimension of sample size and is often small in high dimensional problems, e.g., microarray problems. Using (20.14), one can easily see that $\hat{f}(\mathbf{G})$ can be rewritten as

$$\hat{f}(\mathbf{G}) = \sum_{i=1}^{n} \lambda^{-1}\hat{\gamma}_i K(\mathbf{G}, \mathbf{G}_i).$$

A comparison of this equation with equation (20.7) suggests that $\hat{f}(\mathbf{G})$ takes exactly a dual representation with $L = n$, $(\mathbf{G}_1^*, \cdots, \mathbf{G}_n^*) = (\mathbf{G}_1, \cdots, \mathbf{G}_n)$ and $\alpha = \lambda^{-1}\hat{\gamma}$. Hence the estimated Lagrangian multiplier $\hat{\gamma}$ serves as the coefficients in the dual representation of $\hat{f}(\mathbf{G})$, apart from a scale factor.

In Liu et al. [33], it is shown that the estimates of f and β can be derived as estimates from a random effects model of the following form:

$$Y = \beta^T\mathbf{Z} + f + e, \tag{20.18}$$

where β is a $q \times 1$ vector of regression coefficients, f is an $n \times 1$ vector random effects following $f \sim N\{\mathbf{0}, \tau K(\rho)\}$, ρ is a scale parameter, and $e \sim N(\mathbf{0}, R = \sigma^2 I)$. Because of this equivalence, all regression parameters in the model can be estimated by maximum likelihood, while the variance component parameters can be estimated by restricted maximum likelihood. If we assume $f(\mathbf{G}) \in \mathscr{F}_K$, one can easily see from the linear mixed model representation (20.18) of the least squares kernel

machine that $H_0 : f(\mathbf{G}) = 0$ is equivalent to testing the variance component $\tau = 0$. The null hypothesis $H_0 : \tau = 0$ places τ on the boundary of the parameter space. Liu et al. [33] developed a score test for testing H_0.

20.3.3 SKAT Extensions

Since the seminal work of Wu et al. [56] on this topic, there have been several notable extensions of the SKAT methodology. One extension was by Lee et al. [28, 29], which made the observation that the collapsed approaches and SKAT could be combined into a unified framework based on a prior distribution for the linkage disequilibrium between rare variants within a genomic region of interest. An application of the SKAT statistics to meta-analysis has been developed by Lee et al. [27]. Finally, we note that Ionita-Laza et al. [23] have extended the SKAT approach to simultaneously incorporate common and rare variants.

20.3.4 SKAT Example

We now describe the application of the SKAT methodology to data from Girirajan et al. [20], in which the role of structural variants in autism was explored. The data come from the Simons Simplex Complex Foundation. For the purposes of this chapter, we will assume that the rows of the data matrix below represent statistically independent observations. A sample of the data is given below:

```
     chrom     start       end   size pheno
1     chr1   6191784   6494317 302533       0
2     chr1 108655067 108718023  62956       0
3     chr1 143636400 143700636  64236       0
4     chr1 143636400 143701095  64695       0
5     chr1 143639096 143701095  61999       0
```

In this file, start and end denote the beginning and end of the structural variant, and size denotes the length of the variant and is the difference between start and end. Finally, pheno is a coding of the phenotype as zero for control and one for case (i.e., autism). Our analyses using SKAT will use start, end and pheno.

We consider data from chromosome 1, which has been considered to be a hotspot for structural variations in autism. We have measurements from 99 cases and 76 controls. We note that the hotspots have variable length, which is why the size column shows variation. In order to implement the SKAT method, we need to convert each row of the dataset into a vector of zeroes and ones. The zero represents absence of a structural variant while one indicates its presence. We partitioned chromosome 1 into 2000 nonoverlapping windows of equal size and determined for each row of the dataset how many windows the alteration overlapped with. This is

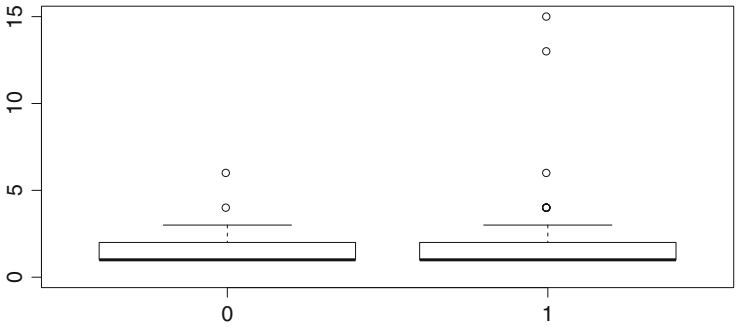

Fig. 20.2 Boxplot of distribution of copy number burden for chromosome 1 in controls (*left* boxplot) versus autism cases (*right* boxplot). The data represent the total number of structural variants from 50 windows that had at least one variant across the 175 samples

done by comparing both the start and end to the window in question. Note that this will give us a 175 by 2000 matrix with zeroes and ones. However, of the 2,000 columns, only 50 have at least one nonzero entry. This means that for each subject, we have a 50-dimensional vector of counts. It is not easy to perform descriptive statistics on this type of data. As in [20], we can define a concept of copy number burden, which means to simply add the up the counts over the 50 dimensions for each subject. A plot of the distribution of copy number burden between cases and controls is given in Fig. 20.2.

Based on Fig. 20.2, we find almost no difference between the copy number burden distribution of cases and controls, aside from two high outliers among the autism cases. However, the SKAT methodology may be able to identify differences between the controls and cases when examining the 50-dimensional count vectors that cannot be seen in the copy number burden data. To illustrate our method, we simulated a covariate $\mathbf{Z} = Z$ from a standard normal distribution and used the following R code to run SKAT.

```
# y.b = pheno variable from the dataset; Z: simulated
# normal(0,1) covariate;
# G: structural variant data, here a 50-dimensional
# vector of counts
# kernel specifies the kernel matrix needed to run
# SKAT; options include linear, IBS,  quadratic
# and 2 way interaction; the first three have the
# option of being weighted by the
# inverse of the variance of the estimated
# proportion of the rare variant, as described
# in Madsen and Browning (2009)
#
# Here, we use the weighted linear kernel.
#
```

```
obj = SKAT_Null_Model(y.b~Z,out_type="D",
          kernel="linear.weighted")
skat1 = SKAT(G,obj)
```

Further details about the code can be found in the SKAT manual. We note that the default procedure of Wu et al. [56] is recommended for a sample size greater than 2000. Given that our example has a sample size of 175, SKAT performs an adjustment in terms of using higher-order approximations in order to estimate the null distribution of the test statistic. Using this adjustment, the p-value from SKAT is 5.27×10^{-5}. Thus there is strong evidence of structural variants in chromosome one being associated with autism.

20.4 Multiple Testing

Next, we discuss the impact of multiple comparisons on the analysis of rare variant data from sequencing studies. While genomics has experienced an explosion in the literature on multiple testing, there are two unique issues in the sequencing context. First, because these variations are rare by definition, the number of single variant hypothesis tests that need to be performed are actually quite small relative to numbers of tests in other problems (e.g., number of tests in common-variant GWAS studies). What is more challenging, however, is the fact that there is an inherent discreteness in the data structure. For a given rare variant, we can represent the data as in Table 20.1, where the cell entries represent the number of samples in each of the groups. We wish to test for independence of the rows and columns, and many methods exist for testing the null hypothesis of no association between presence of rare variant and group label. If the expected cell count is greater than five in all the cells, then one can safely use chi-squared statistics. However, when the cell counts are small, we then use Fisher's exact test, where the p-value is computed using a hypergeometric distribution.

While there has been a lot of work on extensions and generalizations of the FDR estimation methodology, most of the literature in this area has used the fact that under the null distribution, the p-values are uniformly distributed on (0,1) or more generally, that the test statistics have a continuous distribution. This will not apply in the case of rare variant data with respect to the presence/absence calls. The literature on multiple testing with discrete p-values is much more limited. An initial procedure was proposed by Tarone [51] which involves only considering hypotheses

Table 20.1 Rare variant presence/absence analysis

	Rare variant present	Rare variant absent	Total
$Y = 0$	a	b	a+b
$Y = 1$	c	d	c+d
	a+c	b+d	a+b+c+d

where a sufficiently small rejection probability is possible and to then perform a Bonferroni test on those selected hypotheses. This procedure has been modified to the false discovery setting in Gilbert [15], where the Bonferroni adjustment was replaced by the Benjamini-Hochberg [2] procedure. Theoretical aspects of the B-H procedure with discrete test statistics have been addressed by Ferreira [9]. An FDR-based estimation procedure in the spirit of the q-value methodology of Storey [49] was developed in Pounds and Cheng [43]. In Kulinskaya and Lewin [25], the B-H procedure was applied to so-called fuzzy p-values, whose behavior under the null hypothesis is identical to that of a Uniform(0,1) random variable so that the usual methods apply. Applications of discrete multiple testing ideas to a cancer genomics problem can be found in Ghosh [12, 13]. Some recent work of Bancroft et al. [1] uses a novel sequential permutation p-value approach to estimate FDR that would be applicable in this setting as well.

Finally, an open problem in this area is the incorporation of dependence into multiple testing procedure. While there has been a lot of recent work in the area on multiple comparisons with dependent data [8, 30], almost all of this work again assumes that the p-values are derived from continuous distribution, which is not the case here. However, the argument that rare variants operate with a network structure is less plausible than for phenomena such as gene expression, so a case could be made that dependence is not as big of an issue as in other genomic settings. Again, this topic is definitely worthy of future exploration.

20.5 Discussion

This chapter has attempted to discuss issues in the analysis of rare variant data for a statistical audience. One of the major messages from this chapter is that the phenomenon being described is one with a low probability of occurring, but given its occurrence, it can have a large effect.

One of the major challenges in this area will be development of methods that will have high power of detecting these events. A major statistical lesson that has been used here is that the score method of testing has definite merits. While classical statistical theory teaches us that the behavior of the likelihood ratio test, Wald test and score test will be identical as the sample size tends to infinity, it is also the case that we are definitely in a small-sample scenario where asymptotic theory will not hold. The score statistic provides many advantages, one of the major ones being that of avoiding having to estimate rare variant effects.

An area not discussed in this chapter is meta-analysis. This has become the *de rigueur* method for identifying candidate genes from genomewide studies. We point the reader to the recent review by Evangelou and Ioannidis [7] and note the SKAT approach to this problem that was described in Lee et al. [27].

While this area is relatively new, we should also be wise to lessons that have been learnt in many other settings. For example, it is well-known that selected variables

or SNPs suffer from the so-called 'winner's curse' so that estimated effects will be biased. This will also be the case for the rare variants and is inherent to the statistical task at hand.

Finally, we believe that a tactic that will be useful in the future is what we term 'pooling information.' One of the major reasons that SKAT methods have had such a major impact in this area is that the equivalence with variance components models and the introduction of random effects models leads to the ability to pool information across estimated parameters. Statistically, this can be conceptualized using shrinkage theory, Empirical Bayes and more generally, Bayesian methods. Given the increasing availability of genomewide information from different data sources, pooling information using 'vertical integration' techniques [52] will be needed to identify and to elucidate the functionality of rare variants in the foreseeable future.

Acknowledgements The research of the authors is supported by NIH R01 CA 129102 and NSF ABI-1262538.

References

[1] Bancroft, T., Du, C., Nettleton, D.: Estimation of false discovery rate using sequential permutation p-values. Biometrics **69**, 1–7 (2013)
[2] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. Ser. B **57**, 289–300 (1995)
[3] Bühmann, M.D.: Radial basis functions: theory and implementation. Cambridge University Press, Cambridge (2003)
[4] Campbell, C.D., Eichler, E.E.: Properties and rates of germline mutations in humans. Trends Genet. **29**, 575–584 (2013)
[5] Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., Park, J. H.: Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. **45**, 400–405 (2013)
[6] Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (2000)
[7] Evangelou, E., Ioannidis, J. P.: Meta-analysis methods for genome-wide association studies and beyond. Nat. Rev. Genet. **14**, 379–389 (2013)
[8] Fan, J., Han, X., Gu, W.: Estimating false discovery proportion under arbitrary covariance dependence. J. Am. Stat. Assoc. **107**, 1019–1035 (2012)
[9] Ferreira, J. A.: The Benjamini-Hochberg method in the case of discrete test statistics. Int. J. Biostat. **3** (1), Article 11 (2007)
[10] Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., Pasternak, S., Wheeler, D.A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S.B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R.C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M.M., Tsui, S.K., Xue, H., Wong, J.T., Galver, L.M., Fan, J.B., Gunderson, K., Murray, S.S., Oliphant, A.R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.F., Phillips, M.S., Roumy, S., Sallée, C., Verner, A., Hudson, T.J., Kwok, P.Y., Cai, D., Koboldt, D.C., Miller,

R.D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.C., Mak, W., Song, Y.Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C.P., Delgado, M., Dermitzakis, E.T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B.E., Whittaker, P., Bentley, D.R., Daly, M. J., de Bakker, P.I., Barrett, J., Chretien, Y.R., Maller, J., McCarroll, S., Patterson, N., Peér, I., Price, A., Purcell, S., Richter, D.J., Sabeti, P., Saxena, R., Schaffner, S.F., Sham, P.C., Varilly, P., Altshuler, D., Stein, L.D., Krishnan, L., Smith, A.V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D.J., Kashuk, C.S., Lin, S., Abecasis, G.R., Guan, W., Li, Y., Munro, H.M., Qin, Z. S., Thomas, D.J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D.M., Morris, A.P., Weir, B.S., Tsunoda, T., Mullikin, J.C., Sherry, S.T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D.R., Suda, E., Rotimi, C.N., Adebamowo, C.A., Ajayi, I., Aniagwu, T., Marshall, P.A., Nkwodimmah, C., Royal, C.D., Leppert, M.F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I.F., Knoppers, B.M., Foster, M.W., Clayton, E.W., Watkin, J., Gibbs, R.A., Belmont, J.W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G.M, Wheeler, D.A., Yakub, I., Gabriel, S.B., Onofrio, R.C, Richter, D.J., Ziaugra, L., Birren, B.W., Daly, M.J., Altshuler, D., Wilson, R.K., Fulton, L.L., Rogers, J., Burton, J., Carter, N.P., Clee, C.M., Griffiths, M., Jones, M.C., McLay, K., Plumb, R. W., Ross, M.T., Sims, S.K., Willey, D.L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J.C., L'Archeveque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A.L., Brooks, L.D., McEwen, J.E., Guyer, M.S., Wang, V.O., Peterson, J.L., Shi, M., Spiegel, J., Sung, L.M., Zacharia, L.F., Collins, F. S., Kennedy, K., Jamieson, R., Stewart, J.: A second generation human haplotype map of over 3.1 million SNPs. Nature **449**, 851–861 (2007)

[11] Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., Altshuler, D.: The structure of haplotype blocks in the human genome. Science **296**, 2225–2229 (2002)

[12] Ghosh, D.: Discrete nonparametric algorithms for outlier detection with genomic data. J. Biopharm. Stat. **20**, 193–208 (2010)

[13] Ghosh, D: Genomic outlier detection in high-throughput data analysis. Methods Mol Biol **972**, 141–53 (2013)

[14] Gibson, G.: Rare and common variants: twenty arguments. Nat. Rev. Genet. **13**, 135–145 (2012)

[15] Gilbert, P.B.: A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. Appl. Stat. **54**, 143–158 (2005)

[16] Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., Spiegelman, D., Henrion, E., Diallo, O., Thibodeau, P., Bachand, I., Bao, J. Y., Tong, A. H., Lin, C.H., Millet, B., Jaafari, N., Joober, R., Dion, P.A., Lok, S., Krebs, M.O., Rouleau, G.A.: Increased exonic de novo mutation rate in individuals with schizophrenia. Nat. Genet. **43**, 860–863 (2011)

[17] Girirajan, S., Brkanac, Z., Coe, B.P., Baker, C., Vives, L., Vu, T.H., Shafer, N., Bernier, R., Ferrero, G.B., Silengo, M., Warren, S.T., Moreno, C.S., Fichera, M., Romano, C., Raskind, W.H., Eichler, E.E.: Relative burden of large CNVs on a range of neurodevelopmental phenotypes. PLoS Genet **7**, e1002334 (2011)

[18] Girirajan, S., Campbell, C.D., Eichler, E.E.: Human copy number variation and complex genetic disease. Annu. Rev. Genet. **45**, 203–226 (2011)

[19] Girirajan, S., Eichler, E.E.: Phenotypic variability and genetic susceptibility to genomic disorders. Hum. Mol. Genet. **19**, R176–187 (2010)

[20] Girirajan, S., Johnson, R.L., Tassone, F., Balciuniene, J., Katiyar, N., Fox, K., Baker, C., Srikanth, A., Yeoh, K.H., Khoo, S.J., Nauth, T.B., Hansen, R., Ritchie, M., Hertz-Picciotto, I., Eichler, E.E., Pessah, I.N., Selleck, S.B.: Global increases in both common and rare copy number load associated with autism. Hum. Mol. Genet. **22**, 2870–80 (2013)

[21] Girirajan, S., Rosenfeld, J.A., Coe, B.P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R. A., McConnell, J.S., Angle, B., Meschino, W.S., Nezarati, M.M., Asamoah, A., Jackson, K.E., Gowans, G.C., Martin, J.A., Carmany, E.P., Stockton, D.W., Schnur, R.E., Penney, L.S., Martin, D.M., Raskin, S., Leppig, K., Thiese, H., Smith, R., Aberg, E., Niyazov, D.M., Escobar, L.F., El-Khechen, D., Johnson, K.D., Lebel, R.R., Siefkas, K., Ball, S., Shur, N., McGuire, M., Brasington, C.K., Spence, J.E., Martin, L.S., Clericuzio, C., Ballif, B. C., Shaffer, L.G., Eichler, E.E.: Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N. Engl. J. Med. **367**, 1321–31 (2012)
[22] Hirschhorn, J.N. , Daly, M.J.: Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. **6**, 95–108 (2005)
[23] Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., Lin, X.: Sequence kernel asociation tests for the combined effect of rare and common variants. Am. J. Hum. Genet. **92**, 841–853 (2013)
[24] Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J., Eichler, E.E.: De novo rates and selection of large copy number variation. Genome Res **20**, 1469–148 (2010)
[25] Kulinskaya, E., Lewin, A: On fuzzy familywise error rate and false discovery rate procedures for discrete distributions. Biometrika **96**, 201–211 (2009)
[26] Kwee, L.C., Liu, D., Lin, X., Ghosh, D., Epstein, M. P.: A powerful and flexible multilocus association test for quantitative traits. Am. J. Hum. Genet. **82**, 386–397 (2008)
[27] Lee, S., Teslovich, T.M., Boehnke, M., Lin, X.: General framework for meta-analysis of rare variants in sequencing association studies. Am. J. Hum. Genet. **93**, 42–53 (2013)
[28] Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Lin, X.: Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. **91**, 224–237 (2012)
[29] Lee, S., Wu, M.C., Lin, X.: Optimal tests for rare variant effects in sequencing association studies. Biostatistics **13**, 762–775 (2012)
[30] Leek, J.T., Storey, J.D.: A general framework for multiple testing dependence. Proc. Natl. Acad. Sci. **105**, 18718–18723 (2008)
[31] Li, B., Leal, S.M.: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. **83**, 311–321 (2008)
[32] Lin, D.Y., Tang, Z.Z.: A general framework for detecting disease associations with rare variants in sequencing studies. Am. J. Hum. Genet. **89**, 354–367 (2011)
[33] Liu, D., Lin, X., Ghosh, D.: Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machines and linear mixed models. Biometrics **63**, 1079–1088 (2007)
[34] Liu, D., Ghosh, D., Lin, X.: Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinform. **9**, 292 (2008)
[35] Lupski, J.R.: Genomic rearrangements and sporadic disease. Nat. Genet. **39**, S43–S47 (2007)
[36] Madsen, B.E., Browning, S.R.: A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. **5**, e1000384 (2009)
[37] Maher, B.: The case of the missing heritability. Nature **456**, 18–21 (2008)
[38] Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., Visscher, P.M.: Finding the missing heritability of complex diseases. Nature **461**, 747–753 (2009)
[39] McClellan, J., King, M.C.: Genetic heterogeneity in human disease. Cell **141**, 210–217 (2010)
[40] Mefford, H.C., Muhle, H., Ostertag, P., von Spiczak, S., Buysse, K., Baker, C., Franke, A., Malafosse, A., Genton, P., Thomas, P., Gurnett, C.A., Schreiber, S., Bassuk, A.G., Guipponi, M., Stephani, U., Helbig, I. and Eichler, E.E.: Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. PLoS Genet. **6**, e1000962 (2010)
[41] Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., Lin, C.F., Stevens, C., Wang, L. S., Makarov, V., Polak, P., Yoon, S., Maguire, J., Crawford, E.L., Campbell, N.G., Geller, E.T., Valladares, O., Schafer, C., Liu, H., Zhao, T., Cai, G., Lihm, J., Dannenfelser, R.,

Jabado, O., Peralta, Z., Nagaswamy, U., Muzny, D., Reid, J.G., Newsham, I., Wu, Y., Lewis, L., Han, Y., Voight, B.F., Lim, E., Rossin, E., Kirby, A., Flannick, J., Fromer, M., Shakir, K., Fennell, T., Garimella, K., Banks, E., Poplin, R., Gabriel, S., DePristo, M., Wimbish, J.R., Boone, B.E., Levy, S.E., Betancur, C., Sunyaev, S., Boerwinkle, E., Buxbaum, J.D., Cook, E.H. Jr, Devlin, B., Gibbs, R.A., Roeder, K., Schellenberg, G.D., Sutcliffe, J.S., Daly, M.J.: Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature **485**, 242–245 (2012)

[42] O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., Turner, E.H., Stanaway, I.B., Vernot, B., Malig, M., Baker, C., Reilly, B., Akey, J.M., Borenstein, E., Rieder, M.J., Nickerson, D.A., Bernier, R., Shendure, J., Eichler, E.E.: Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature **485**, 246–250 (2012)

[43] Pounds, S., Cheng, C.: Robust estimation of the false discovery rate. Bioinformatics **22**, 1979–1987 (2006)

[44] Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S. M., Staples, J., Wei, L.J., Sunyaev, S.R.: Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. **86**, 832–838 (2010)

[45] Pritchard, J.K., Cox, N.J.: The allelic architecture of human disease genes: common diseaseÐcommon variantÉ or not? Hum. Mol. Genet. **11**, 2417–2423 (2002)

[46] Reich, D.E., Lander, E.S.: On the allelic spectrum of human disease. Trends Genet. **17**, 502–510 (2001)

[47] Rosenfeld, J.A., Coppinger, J., Bejjani, B.A., Girirajan, S., Eichler, E.E., Shaffer, L.G., Ballif., B.C.: Speech delays and behavioral problems are the predominant features in individuals with developmental delays and 16p11.2 microdeletions and microduplications. J. Neurodevelop. Disord. **2**, 26–38 (2010)

[48] Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A. J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., Walker, M.F., Ober, G.T., Teran, N. A., Song, Y., El-Fishawy, P., Murtha, R.C., Choi, M., Overton, J.D., Bjornson, R.D., Carriero, N.J., Meyer, K. A., Bilguvar, K., Mane, S.M., Sestan, N., Lifton, R.P., Günel, M., Roeder, K., Geschwind, D.H., Devlin, B., State, M.W.: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature **485**, 237–241 (2013)

[49] Storey, J.D.: A direct approach to false discovery rates. J. Roy. Stat. Soc. Ser. B **64**, 479–498 (2002)

[50] Sullivan, P.F., Daly, M.J., O'Donovan, M.: Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat. Rev. Genet. **13**, 537–551 (2012)

[51] Tarone, R.E.: A modified Bonferroni method for discrete data. Biometrics **46**, 515–522 (1990)

[52] Tseng, G.C., Ghosh, D., Feingold, E.: Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. **40**, 3785–99 (2012)

[53] Veltman, J.A., Brunner, H.G.: Understanding variable expressivity in microdeletion syndromes. Nat. Genet. **42**, 192–193 (2010)

[54] Wahba, G.: Spline Models for Observational Data. SIAM, Philadelphia (1990)

[55] Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T., Platt, O.S., Ruderfer, D.M., Walsh, C.A., Altshuler, D., Chakravarti, A., Tanzi, R.E., Stefansson, K., Santangelo, S.L., Gusella, J. F., Sklar, P., Wu, B.L., Daly, M. J.: Association between microdeletion and microduplication at 16p11.2 and autism. N. Engl. J. Med. **358**, 667–675 (2008)

[56] Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X.: Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. **89**, 82–93 (2011)

[57] Zuk, O., Hechter, E., Sunyaev, S.R., Lander, E.S.: The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA **109**, 1193–1198 (2012)