Biostatistics (2016), **0**, 0, *pp.* 1–17
doi:10.1093/biostatistics/article˙Epistasis˙suppl

A powerful and efficient two-stage method for detecting gene-to-gene interactions in GWAS (Supplementary Materials)

Jakub Pecanka[*,1,2], Marianne A. Jonker[3,4], International Parkinson's Disease Genomics Consortium (IPDGC)[5], Zoltan Bochdanovits[6], Aad W. van der Vaart[7]

[1] *Leiden University Medical Center, Department of Medical Statistics and Bioinformatics, Leiden, The Netherlands*

[2] *VU University, Department of Mathematics, Amsterdam, the Netherlands*

[3] *VU University Medical Center, Department of Epidemiology and Biostatistics, Amsterdam, The Netherlands*

[4] *Radboud university medical center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands*

[5] *A full list of authors and affiliations is provided in Appendix B of the Supplementary Materials*

[6] *VU University Medical Center, Department of Clinical Genetics, Amsterdam, The Netherlands*

[7] *Leiden University, Mathematical Institute, Leiden, The Netherlands*

j.pecanka@lumc.nl

[*]To whom correspondence should be addressed.

APPENDIX A

A.1 Error control in two-stage testing with independent stages

Denote by $\mathcal{K}_0 \subset \mathcal{K}$ the set of true null hypotheses H_{0k}^{epi}. Given levels $\alpha_1, \alpha_2 \in (0,1)$ the FWER of the two-stage procedure described in Section 3.3 is $\lambda_n = \mathsf{P}(\exists k \in \mathcal{K}_0 \text{ s.t. } T_{nk} \geqslant \tau_{\alpha_1}, |R_{nk}| \geqslant \xi_{\alpha_2})$. The procedure controls FWER by $\alpha \in (0,1)$ if $\lambda_n \leqslant \alpha$. In agreement with our two-stage testing scheme, we use level $\alpha_2 = \alpha/K_1$, where K_1 is the number of rejections in S1. Then, assuming the independence of $(R_{nk}, k \in \mathcal{K})$ and $(T_{nk}, k \in \mathcal{K})$, and denoting $A_k = \{T_{nk} \geqslant \tau_{\alpha_1}\}$, we get

$$
\begin{aligned}
\lambda_n &= \mathsf{P}\big(\exists k \in \mathcal{K}_0 \text{ s.t. } |R_{nk}| \geqslant \xi_{\alpha/K_1}, A_k\big) \\
&= \textstyle\sum_{C \subset \mathcal{K}} \mathsf{P}(\mathcal{K}_1 = C)\, \mathsf{P}\big(\exists k \in \mathcal{K}_0 \text{ s.t. } |R_{nk}| \geqslant \xi_{\alpha/|C|}, A_k \,\big|\, \mathcal{K}_1 = C\big) \\
&= \textstyle\sum_{C \subset \mathcal{K}} \mathsf{P}(\mathcal{K}_1 = C)\, \mathsf{P}\big(\exists k \in \mathcal{K}_0 \cap \mathcal{K}_1 \text{ s.t. } |R_{nk}| \geqslant \xi_{\alpha/|C|}, A_k \,\big|\, \mathcal{K}_1 = C\big) \\
&= \textstyle\sum_{C \subset \mathcal{K}} \mathsf{P}(\mathcal{K}_1 = C)\, \mathsf{P}\big(\exists k \in \mathcal{K}_0 \cap \mathcal{K}_1 \text{ s.t. } |R_{nk}| \geqslant \xi_{\alpha/|C|} \,\big|\, \mathcal{K}_1 = C\big) \\
&\leqslant \textstyle\sum_{C \subset \mathcal{K}} \mathsf{P}(\mathcal{K}_1 = C) \sum_{k \in \mathcal{K}_0 \cap C} \mathsf{P}\big(|R_{nk}| \geqslant \xi_{\alpha/|C|} \,\big|\, \mathcal{K}_1 = C\big) \\
&= \textstyle\sum_{C \subset \mathcal{K}} \mathsf{P}(\mathcal{K}_1 = C) \sum_{k \in \mathcal{K}_0 \cap C} \alpha/|C| \;\leqslant\; \alpha \sum_{C \subset \mathcal{K}} \mathsf{P}(\mathcal{K}_1 = C) \;=\; \alpha.
\end{aligned}
$$

Thus, under independence between the two stages the procedure with $\alpha_2 = \alpha/K_1$ controls FWER in the strong sense for any α_1. Regarding the assumption of independence between the two stages, we note that the independence is *exact* when D_n is used in S2 in conjunction with either T_n^A or T_n^B, while it is *asymptotic* when A_n combined with either T_n^A or T_n^B or when C_n is combined with either T_n^C or T_n^D.

A.2 Asymptotic theoretical results

Unlike the classical random sample design (i.e. independent and identically distributed (iid) individuals), the case-control design consists of two random samples from each subpopulation with a fixed sample size ratio between them. However, from the perspective of inference about

the parameters of the logistic regression model this difference between the two designs is of little importance. In fact, it is well known that the coefficients of all non-trivial regressors in the logistic regression model are identifiable under both designs and (profile) likelihood functions for the two designs are proportional for these coefficients. It is only the penetrance parameter β_0 that is not estimable under the case-control design, which, however, is irrelevant for our purposes. Therefore, we can use this "near-equivalence" of the two designs to conveniently prove the theoretical results below using the iid design.

DEFINITION A.1 (Asymptotic independence) Fix $m \in \mathbb{N}$ and let $X_n^1, \ldots, X_n^m$, $n = 1, 2, \ldots$, be random vectors of dimensions $k_1, \ldots, k_m \in \mathbb{N}$. The random sequences $(X_n^1)_{n=1}^\infty, \ldots, (X_n^m)_{n=1}^\infty$ are said to be *(weakly) asymptotically independent* as $n \to \infty$ if there exist sequences $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ both of dimension $k = \sum_{i=1}^m k_i$, where the coordinates of a_n are all positive, such that the sequence $(a_n \otimes (X_n^{1\prime}, \ldots, X_n^{m\prime})' - b_n)_{n=1}^\infty$, where $\otimes$ denotes coordinate-wise product, converges in distribution to a random vector with non-degenerate distribution function F such that $F(x_1, \ldots, x_m) = \prod_{i=1}^m F_i(x_i)$ for all $x_i \in \mathbb{R}^{k_i}$, $i = 1, \ldots, m$, and for some distribution functions $F_1, \ldots, F_m$.

Theorem A.2 below gives an asymptotic linear relationship which links the score statistic C_n with the (vector) score statistic evaluated at the true values of the LRM nuisance parameters β_0, β_1, β_2. Even though the theorem is formulated for the score statistic in the LRM, in the proof we take a more general approach using the asymptotic machinery found in Van der Vaart (1998). Consequently, the theorem could easily be reformulated for many other sufficiently regular statistical models. For details see Van der Vaart (1998), Sections 7.4 and 16.3. In the proof we parameterize the model by $\theta \in \mathbb{R}^k$, which in the LRM corresponds to $\beta \in \mathbb{R}^4$. We also use the concept of convergence of sets defined in Section 7.4 of Van der Vaart (1998), which postulates that a sequence of sets H_n is said to converge to H if H is the set of all limits $\lim_{n\to\infty} h_n$ of

converging sequences h_n with $h_n \in H_n$ for every n and, more over, the limit $h = \lim_{i\to\infty} h_{n_i}$ of every converging sequence h_{n_i} with $h_{n_i} \in H_{n_i}$ for every i is contained in H.

THEOREM A.2 (Asymptotic representation of the score statistic) Assume the LRM of (3.4) holds, let C_n be defined by (3.5), and let S_n be the (vector) score statistic of the LRM at the true values of $\beta_0, \beta_1, \beta_2$ defined as $S_n = n^{-1/2}\sum_{i=1}^n (\Delta_i - \Psi_i^0)\,\mathsf{z}(x_i, y_i)$, where $\Psi_i^0 = (1+\exp(-\beta_0 - \beta_1 x_i - \beta_2 y_i))^{-1}$ and $\mathsf{z} = (1, x_i, y_i, z(x_i, y_i))'$. If $\beta_3 = 0$, then $C_n = \mathsf{e}_4'\mathsf{M}S_n + o_P(1)$, where $\mathsf{M} = \mathsf{I}_\beta^{1/2}(\mathsf{1} - \mathsf{\Pi})\,\mathsf{I}_\beta^{-1/2}$, and I_β is the Fisher information matrix (FIM), and $\mathsf{e}_4 = (0,0,0,1)'$, and $\mathsf{1}$ is the unit matrix in $\mathbb{R}^4$, and $\mathsf{\Pi}$ is a projection onto the space $\mathsf{I}_\beta^{1/2}\mathsf{H}_0$, and $\mathsf{H}_0 = \mathbb{R}^3 \times \{0\}$ is the null hypothesis parameter space. Consequently, the statistic C_n is asymptotically zero-mean normal with asymptotic variance matrix $\mathsf{e}_4'\mathsf{M}\mathsf{I}_\beta\mathsf{M}'\mathsf{e}_4$.

Proof. Given a statistical model $\{P_\theta \text{ s.t. } \theta \in \Theta\}$, where $\Theta \in \mathbb{R}^k$ is an open subset of $\mathbb{R}^k$ and P_θ is a probability distribution for every $\theta \in \Theta$, the k-dimensional vector score statistic of a sample of iid random variables $X_1, \ldots, X_n$ is defined as $S_{n,\theta} = n^{-1/2}\,\dot\ell_n(X;\theta)$, where $X = (X_1, \ldots, X_n)'$ and $\ell_n(X;\theta) = \sum_{i=1}^n \log p_\theta(X_i)$ and $\dot\ell_n = (\partial/\partial\theta)\,\ell_n$ is the score function. Then, the probability model for X seen as a single variable is $\mathcal{M}_\theta^n = \{P_\theta^n \text{ s.t. } \theta \in \Theta\}$. Denote by ϑ the (unknown) true value of the parameter θ. The goal is to test the null hypothesis $H_0{:}\vartheta \in \Theta_0$ for some $\Theta_0 \subset \Theta$ within such model. We estimate the unknown value of θ by $\widehat\theta_n^0 = \arg\max_{\theta\in\Theta_0} \ell_n(X;\theta)$ and use it to calculate the null hypothesis (vector) score statistic $C_n^0 = n^{-1/2}\,\dot\ell_n(X;\widehat\theta_n^0)$.

Instead of working with $\mathcal{M}_\theta^n$ it is convenient to work with the equivalent experiment $\mathcal{M}_{\vartheta,h}^n$ parameterized by $h = \sqrt{n}(\theta - \vartheta)$. In $\mathcal{M}_{\vartheta,h}^n$ the null hypothesis local parameter space is $\mathcal{H}_{n,0} = \sqrt{n}(\Theta_0 - \vartheta)$ and the null hypothesis ML estimator of h is $\widehat h_n^0 = \arg\max_{h\in\mathcal{H}_{n,0}} \Lambda_n^\vartheta(h)$, where $\Lambda_n^\vartheta(h)$ is the log-likelihood within $\mathcal{M}_{\vartheta,h}^n$. Denote the FIM of $\mathcal{M}_{\vartheta,h}^n$ by $\mathcal{I}_\vartheta$. Under sufficiently regular model with sufficiently smooth $\Lambda_n^\vartheta(h)$ (such as the LRM), the score statistic can be Taylor expanded as

$$C_n^0 \;=\; S_{n,\vartheta} - \mathcal{I}_\vartheta\,\widehat h_n^0 + o_{P_\vartheta}(1). \tag{A.1}$$

Moreover, using arguments analogous to those employed in the proofs of Theorems 7.12 and 16.7 in Van der Vaart (1998) together with Lemma 7.13 therein, it can be shown that

$$\widehat{h}_n^0 \;=\; \arg\min_{h\in\mathcal{H}_0} \| \mathcal{I}_\vartheta^{1/2} h - \mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta}\|^2 + o_P(1), \tag{A.2}$$

where $\mathcal{H}_0$ is the limit set of $\mathcal{H}_{n,0}$. Combining (A.1) and (A.2) then yields

$$C_n^0 \;=\; S_{n,\vartheta} - \mathcal{I}_\vartheta \arg\min_{h\in\mathcal{H}_0} \| \mathcal{I}_\vartheta^{1/2} h - \mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta}\|^2 + o_{P_\theta}(1).$$

Simple arithmetics applied to the right-hand side of the equality above yields

$$\begin{aligned} C_n^0 \;&=\; \mathcal{I}_\vartheta^{1/2}\big(\mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta} - \mathcal{I}_\vartheta^{1/2} \arg\min_{h\in\mathcal{H}_0} \| \mathcal{I}_\vartheta^{1/2} h - \mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta}\|^2\big) + o_{P_\vartheta}(1) \\ &=\; \mathcal{I}_\vartheta^{1/2}\big(\mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta} - \arg\min_{g\in\mathcal{I}_\vartheta^{1/2}\mathcal{H}_0} \| g - \mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta}\|^2\big) + o_{P_\vartheta}(1). \end{aligned} \tag{A.3}$$

The term $g^n_{\min} = \arg\min_{g\in\mathcal{I}_\vartheta^{1/2}\mathcal{H}_0} \| g - \mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta}\|^2$ above is a solution to the minimization problem of finding an element in the space $\mathcal{I}_\vartheta^{1/2}\mathcal{H}_0$ that is closest to $\mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta}$, or in other words, finding an orthogonal projection of $\mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta}$ onto $\mathcal{I}_\vartheta^{1/2}\mathcal{H}_0$. Denoting as $\Pi_{\mathcal{I}\mathcal{H}}$ the projection operator onto $\mathcal{I}_\vartheta^{1/2}\mathcal{H}_0$, we get $g^n_{\min} = \Pi_{\mathcal{I}\mathcal{H}}(\mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta})$. Finally, plugging this into (A.3) yields

$$C_n^0 \;=\; \mathcal{I}_\vartheta^{1/2}(1 - \Pi_{\mathcal{I}\mathcal{H}})\,\mathcal{I}_\vartheta^{-1/2} S_{n,\vartheta} + o_{P_\vartheta}(1), \tag{A.4}$$

where 1 is the identity matrix (of the same dimension as ϑ). Given (A.4), the definition of M in the theorem is obvious, while e_4 is added because C_n is the last (fourth) coordinate of the vector score statistic. Finally, the asymptotic normality of C_n^0 and the variance follow by Slutsky lemma from (A.4) and the asymptotic normality of $S_{n,\vartheta}$ given by the central limit theorem (CLT). □

The following theorem yields joint asymptotic distribution of the adjusted score statistic A_n and the case-based partial sample Pearson chisquare statistic T_n^A or the trend test statistic T_n^B.

THEOREM A.3 (Normality and independence) Denote $\delta = m_1/m$ and let C_n be defined by (3.5), and let A_n, U_n, $\widetilde{U}_n$ be defined by (3.6). If H_0^{epi} holds, then the random vector $E_n = (C_n', (U_n -$

$\mathsf{E}U_n)', (\widetilde{U}_n - \mathsf{E}\widetilde{U}_n)')'$, as $n \to \infty$, converges in distribution to the zero-mean normal distribution with variance

$$\mathsf{W} = \begin{pmatrix} \mathsf{M}\mathsf{I}_\beta\mathsf{M}' & \mathsf{M}\mathsf{C}_{SU} & \sqrt{(1-\delta)/\delta}\,\mathsf{M}\mathsf{C}_{SU} \\ \mathsf{C}_{SU}'\mathsf{M}' & \mathsf{V}_U & 0 \\ \sqrt{(1-\delta)/\delta}\,\mathsf{C}_{SU}'\mathsf{M}' & 0 & \mathsf{V}_U \end{pmatrix},$$

where V_U is the asymptotic variance matrix of U_n and C_{SU} is the asymptotic covariance matrix of S_n and U_n. Moreover, for any $\delta \in (0,1)$ the random vectors A_n and U_n are asymptotically independent, thus making A_n asymptotically independent with both T_n^A and T_n^B. Finally, A_n is asymptotically normal with zero expectation and variance matrix $\mathsf{V}_A = \mathsf{M}(\mathsf{I}_\beta + \frac{\delta}{1-\delta}\,\mathsf{C}_{SU}\mathsf{V}_U^{-}\mathsf{C}_{SU}')\mathsf{M}'$.

Proof. We first derive the asymptotic distribution of $F_n = (S_n', (U_n - \mathsf{E}U_n)', (\widetilde{U}_n - \mathsf{E}\widetilde{U}_n)')'$. Define the random variables $B_i \in \{0,1\}$, $i = 1, \ldots, n$ as indicators of whether the i-th individual belongs to the subsample of m_1 cases from which U_n is calculated. Consequently, $B_i = 0$ for all controls ($\Delta_i = 0$) and $\sum_{i=1}^n B_i = m_1$. Using these indicators, we write $m_{kl} = \sum_{i=1}^n B_i\Delta_i I_{\{x_i=k,\, y_i=l\}}$, $m_{k\cdot} = \sum_{i=1}^n B_i\Delta_i I_{\{x_i=k\}}$ and $m_{\cdot l} = \sum_{i=1}^n B_i\Delta_i I_{\{y_i=l\}}$, while analogous equalities hold for the counts based on the sample of the remaining m_2 controls used to calculate $\widetilde{U}_n$. Using CLT we get the asymptotic normality of $n^{-1/2}\sum_{i=1}^n Z_i$, where

$$Z_i = \begin{pmatrix} (\Delta_i - \Psi_i^0)\,\mathsf{z}(X_i, Y_i) \\ B_i\Delta_i\,(\,I_{\{X_i=k,\,Y_i=l\}} - p_{kl},\, I_{\{X_i=k\}} - p_k,\, I_{\{Y_i=l\}} - q_l)_{k,l}' \\ (1-B_i)\Delta_i\,(\,I_{\{X_i=k,\,Y_i=l\}} - p_{kl},\, I_{\{X_i=k\}} - p_k,\, I_{\{Y_i=l\}} - q_l)_{k,l}' \end{pmatrix},$$

where p_{kl}, p_k, q_l, were defined in Section 3.1. Then, using the delta method, we get the asymptotic normality of $(S_n', Z_{n1}', Z_{n2}')'$, where

$$\begin{aligned} Z_{n1} &= (\delta\gamma n)^{1/2}\big((\widehat{p}_{kl} - p_{kl} - p_k(\widehat{q}_l - q_l) - q_l(\widehat{p}_k - p_k))\big)_{k,l}, \\ Z_{n2} &= ((1-\delta)\gamma n)^{1/2}\big((\widetilde{p}_{kl} - p_{kl} - p_k(\widetilde{q}_l - q_l) - q_l(\widetilde{p}_k - p_k))\big)_{k,l}, \end{aligned} \tag{A.5}$$

with $\widehat{p}_{kl}$, $\widehat{p}_k$, $\widehat{q}_l$ and $\widetilde{p}_{kl}$, $\widetilde{p}_k$, $\widetilde{q}_l$ denoting the ML estimators of p_{kl}, p_k and q_l based on the m_1 and m_2 cases, respectively. Since $F_n = (S_n', Z_{n1}', Z_{n2}')' + O_P(n^{-1/2})$, Slutsky lemma yields the

asymptotic normality of F_n with zero mean under H_0^{epi} and variance

$$\begin{pmatrix} \mathsf{I}_\beta & \mathsf{C}_{SU} & \sqrt{(1-\delta)/\delta}\,\mathsf{C}_{SU} \\ \mathsf{C}'_{SU} & \mathsf{V}_U & 0 \\ \sqrt{(1-\delta)/\delta}\,\mathsf{C}'_{SU} & 0 & \mathsf{V}_U \end{pmatrix},$$

where the diagonal block elements of the matrix above come from the asymptotic variance of S_n, which is the Fisher information matrix I_β, and the fact that the asymptotic variance matrices of U_n, $\widetilde{U}_n$, Z_{n1} and Z_{n2} are all equal. The zero covariance blocks of U_n and $\widetilde{U}_n$ (or Z_{n1} and Z_{n2}) come from their independence, while the covariance blocks of S_n with U_n and $\widetilde{U}_n$ (or with Z_{n1} and Z_{n2}) again follow from the symmetry of U_n and $\widetilde{U}_n$ and the fact that going from $\mathsf{cov}(C_n, U_n)$ to $\mathsf{cov}(C_n, \widetilde{U}_n)$ requires only scaling by $\sqrt{(1-\delta)/\delta}$ to account for the difference in sample sizes (see Lemma A.5 below). The asymptotic normality of E_n now follows from the asymptotic representation $C_n = \mathsf{M}S_n + o_P(1)$ given by Theorem A.2 by Slutsky lemma. Moreover, since the vector $(A'_n, (U_n - \mathsf{E}U_n)')'$ is asymptotically equal to a linear transformation of E_n, the joint asymptotic normality of A_n and U_n follows by the delta method and Slutsky lemma. Furthermore, the asymptotic variance matrix of $(A'_n, (U_n - \mathsf{E}U_n)')'$ is block diagonal, which yields the asymptotic independence of A_n and U_n. Since both T_n^A and T_n^B are smooth functions of only U_n, the asymptotic independence of A_n with T_n^A and T_n^B follows. □

A.3 Evaluating the adjusted score statistic A_n

Theorem A.3 gives the asymptotic distribution of the adjusted score statistic A_n defined in (3.6). In order to use A_n in practice we need to be able to evaluate the matrix M defined in Theorem A.2 as well as the asymptotic variance of U_n and the asymptotic covariance of S_n and U_n.

By definition $1 - \Pi_{\mathcal{IH}}$ of Theorem A.2 is a projection onto the orthogonal complement space of $\mathsf{I}_\beta^{1/2}\mathcal{H}_0$. For a space generated by a matrix $\mathsf{X} \in \mathbb{R}^{m\times p}$, which we denote by $\mathcal{S}(\mathsf{X})$, the projection matrix H of a vector $y \in \mathbb{R}^m$ onto $\mathcal{S}(\mathsf{X})$ can be written using the matrix X as $\mathsf{H} = \mathsf{X}(\mathsf{X}'\mathsf{X})^-\mathsf{X}'$. To find the projection matrix onto the space $\mathsf{I}_\beta^{1/2}\mathcal{H}_0$ we have to find a base of the limiting local

parameter space $\mathcal{H}_0 = \{(\beta_0, \beta_1, \beta_2, \beta_3)' \in \mathbb{R}^4 : \beta_3 = 0\}$, which is the column space of the matrix $\mathsf{J}_4 = \mathrm{diag}((1,1,1,0))$. Thus, the space $\mathsf{I}_\beta^{1/2}\mathcal{H}_0$ is generated by the matrix $\mathsf{X} = \mathsf{I}_\beta^{1/2}\mathsf{J}_4$. Since both I_β and J_4 are symmetric, plugging X into the hat matrix $\mathsf{H} = \mathsf{X}(\mathsf{X}'\mathsf{X})^-\mathsf{X}'$ yields

$$\mathsf{\Pi}_{\mathcal{I}\mathcal{H}} \;=\; \mathsf{I}_\beta^{1/2}\mathsf{J}_4(\mathsf{J}_4\mathsf{I}_\beta\,\mathsf{J}_4)^-\mathsf{J}_4\,\mathsf{I}_\beta^{1/2} \;=\; \mathsf{I}_\beta^{1/2}(\mathsf{J}_4\mathsf{I}_\beta\,\mathsf{J}_4)^-\mathsf{I}_\beta^{1/2}. \tag{A.6}$$

The second equality above follows from the fact that a (pseudo)inverse of a block-diagonal matrix is obtained by (pseudo)inverting the individual blocks. Consequently, the multiplication of $(\mathsf{J}_4\mathsf{I}_\beta\,\mathsf{J}_4)^-$ by J_4 from left and right does not modify $(\mathsf{J}_4\mathsf{I}_\beta\,\mathsf{J}_4)^-$. Defining an estimator $\hat{\mathsf{I}}_\beta$ for I_β is completely straight forward, and so is calculating its square-root and a square-root of its (pseudo)inverse. Plugging them into (A.6) yields an estimator $\widehat{\mathsf{\Pi}}_{\mathcal{I}\mathcal{H}} = \hat{\mathsf{I}}_\beta^{1/2}(\mathsf{J}_4\hat{\mathsf{I}}_\beta\mathsf{J}_4)^-\hat{\mathsf{I}}_\beta^{1/2}$, which in turn yields $\widehat{\mathsf{M}} = \hat{\mathsf{I}}_\beta^{1/2}(1 - \widehat{\mathsf{\Pi}}_{\mathcal{I}\mathcal{H}})\,\hat{\mathsf{I}}_\beta^{-1/2}$.

Lemma A.4 gives an expression for the asymptotic variance of U_n, while Lemma A.5 gives the asymptotic covariance of S_n and U_n. Combining Lemma A.5 and Theorem A.2 yields the asymptotic covariance of C_n and U_n. In the definition of U_n in (3.2) we indexed the elements by two indices $k, l = 0, 1, 2$. As explained in Section , throughout this paper we use the same row-wise ordering of the elements of U_n given by $i = 3k+l$. The same ordering is used in Lemmas A.4 and A.5.

LEMMA A.4 Let U_n be defined by (3.2) and let its elements be ordered according to $3k+l$, where $k, l = 0, 1, 2$. Then, the elements of the asymptotic variance matrix $\mathsf{V}_U = (v_{ij})_{i,j=0,\ldots,8}$ of U_n are

$$v_{ii} \;=\; p_{kl}(1 + p_k q_l - p_k - q_l) + (p_k + q_l + p_{kl})(p_k q_l - p_{kl}) + 4p_k q_l(p_{kl} - p_k q_l), \qquad \text{for } \; k = r,\; l = s, \tag{A.7}$$

$$v_{ij} \;=\; p_{kl}(2p_r q_s - p_{rs}) + 2p_k q_l(p_{rs} - p_r q_s) + p_r q_l(p_{ks} - p_k q_s) + p_k q_s(p_{rl} - p_r q_l), \qquad \text{for } \; k \neq r,\; l \neq s, \tag{A.8}$$

$$\;=\; p_{rl}(2p_k q_l - p_{kl}) + p_r q_l(3p_{kl} - p_k q_l) + p_k q_l(p_{rl} - 3p_r q_l) + p_k(p_r q_l - p_{rl}) - p_r p_{kl}, \qquad \text{for } \; k \neq r,\; l = s, \tag{A.9}$$

$$= \; p_{ks}(2p_kq_l - p_{kl}) + p_kq_s(3p_{kl} - p_kq_l) + p_kq_l(p_{ks} - 3p_kq_s) + q_s(p_kq_l - p_{kl}) - q_lp_{ks}, \quad \text{for } \; k = r, \; l \neq s, \tag{A.10}$$

where $i = 3k + l$ and $j = 3r + s$ for $k, l, r, s = 0, 1, 2$.

Proof. For the terms d_{kl} inside U_n defined in (3.2) it holds

$$d_{kl} \; = \; \sqrt{m_1}\big[(\widehat{p}_{kl} - p_{kl}) - p_k(\widehat{q}_l - q_l) - q_l(\widehat{p}_k - p_k) + (p_{kl} - p_kq_l)\big] + O_P(n^{-1/2}).$$

Denote elements of the (finite sample) variance matrix of U_n by v^n_{ij}, where $i = 3k + l$ and $j = 3r + s$, with $k, l, r, s = 0, 1, 2$. Just like any other variance terms, the elements v^n_{ij} have the form $v^n_{ij} = \mathsf{E}(d_{kl} - \mathsf{E}\, d_{kl})(d_{rs} - \mathsf{E}\, d_{rs})$, where the terms $(d_{kl} - \mathsf{E}\, d_{kl})(d_{rs} - \mathsf{E}\, d_{rs})$ equal

$$\begin{aligned} & m_1\big[(\widehat{p}_{kl} - p_{kl})(\widehat{p}_{rs} - p_{rs}) - p_k(\widehat{q}_l - q_l)(\widehat{p}_{rs} - p_{rs}) - p_r(\widehat{p}_{kl} - p_{kl})(\widehat{q}_s - q_s) \\ & - q_l(\widehat{p}_k - p_k)(\widehat{p}_{rs} - p_{rs}) - q_s(\widehat{p}_{kl} - p_{kl})(\widehat{p}_r - p_r) + p_kp_r(\widehat{q}_l - q_l)(\widehat{q}_s - q_s) \\ & + q_lq_s(\widehat{p}_k - p_k)(\widehat{p}_r - p_r) + p_rq_l(\widehat{p}_k - p_k)(\widehat{q}_s - q_s) + p_kq_s(\widehat{q}_l - q_l)(\widehat{p}_r - p_r)\big] + O_P(n^{-1/2}). \end{aligned} \tag{A.11}$$

Working out the expectations of the individual terms in (A.11) by considering the different combinations of k, l, r, s and combining the results yields (A.7) – (A.10). □

LEMMA A.5 Let S_n and U_n be defined by (3.5) and (3.2), respectively. Denote by C_{SU} the asymptotic covariance matrix under H_0^{epi} of S_n and U_n and let $m/n \to \gamma \in (0, 1)$ and $m_1/m \to \delta \in (0, 1)$ as $n \to \infty$. The columns c_{3k+l}, $k, l = 0, 1, 2$, of C_{SU} are

$$\mathsf{c}_{3k+l} \; = \; (\delta/\gamma)^{1/2}\big(\mathsf{d}_{kl} - p_kq_l^{1/2}\textstyle\sum_j \mathsf{d}_{kj}q_j^{-1/2} - q_lp_k^{1/2}\sum_i \mathsf{d}_{il}p_i^{-1/2}\big), \tag{A.12}$$

where $\mathsf{d}_{kl} \; = \; \Psi_1'(\beta_0 + \beta_1k + \beta_2l)\,\mathsf{z}(k, l)\,\pi_{kl}$, $\mathsf{z}(a, b) = (1, a, b, z(a, b))'$, $\Psi_1'(x) = \mathrm{e}^{-x}/(1 + \mathrm{e}^{-x})^2$, $p_k = \mathsf{P}(X = k \,|\, \Delta = 1)$, $q_l = \mathsf{P}(Y = l \,|\, \Delta = 1)$ and $\pi_{kl} = \gamma\mathsf{P}(X = k, Y = l \,|\, \Delta = 1) + (1-\gamma)\mathsf{P}(X = k, Y = l \,|\, \Delta = 0)$. In other words, π_{kl} are the genotype frequencies in a case-control population with the prevalence of the case phenotype ($\Delta = 1$) equal to γ.

Proof. Since $(S_n', (U_n - \mathsf{E}U_n)')'$ and $(S_n', Z_{n1}')'$ have the same asymptotic distribution, we can determine C_{SU} by calculating $\mathsf{cov}(S_n, Z_{n1})$, which in turn can be determined by calculating $\mathsf{cov}(S_n, \widehat{p}_{kl})$, $\mathsf{cov}(S_n, \widehat{p}_k)$ and $\mathsf{cov}(S_n, \widehat{q}_l)$. Similarly to the proof of Theorem A.3, write $S_n = n^{-1/2}\sum_{i=1}^n (\Delta_i - \Psi_i^0)\,\mathsf{z}(X_i, Y_i)$ and $\widehat{p}_{kl} = m_1^{-1}\sum_{j=1}^n B_i\Delta_j\, I_{\{X_j=k,\, Y_j=l\}}$. By taking the expectation when calculating $\mathsf{cov}(S_n, \widehat{p}_{kl})$ the cross-product terms with $i \neq j$ drop due to independence of individuals. Since also $\mathsf{E}S_n = 0$ under H_0^{epi}, the calculation of the covariance boils down to

$$
\begin{aligned}
\mathsf{c}_{kl} &= \mathsf{E}\big((\Delta_1 - \Psi_1^0)\,\mathsf{z}(X_1, Y_1)\,\Delta_1\, I_{\{X_1=k,\, Y_1=l\}}\big) \\
&= \mathsf{E}\big(\mathsf{E}\big((\Delta_1 - \Psi_1^0)\,\mathsf{z}(X_1, Y_1)\,\Delta_1\, I_{\{X_1=k,\, Y_1=l\}} \mid X_1, Y_1\big)\big) \\
&= \mathsf{E}\big(\mathsf{E}\big((\Delta_1 - \Psi_1^0)\Delta_1 \mid X_1, Y_1\big)\,\mathsf{z}(X_1, Y_1)\, I_{\{X_1=k,\, Y_1=l\}}\big) \\
&= \mathsf{E}\big(\,\Psi_1^0(1 - \Psi_1^0)\,\mathsf{z}(X_1, Y_1)\, I_{\{X_1=k,\, Y_1=l\}}\big) \\
&= \mathsf{E}\big(\,\Psi_1'(\beta_0 + \beta_1 X_1 + \beta_2 Y_1)\,\mathsf{z}(X_1, Y_1)\, I_{\{X_1=k,\, Y_1=l\}}\big) \\
&= \Psi_1'(\beta_0 + \beta_1 k + \beta_2 l)\,\mathsf{z}(k, l)\,\mathsf{E}\, I_{\{X_1=k,\, Y_1=l\}} \\
&= \Psi_1'(\beta_0 + \beta_1 k + \beta_2 l)\,\mathsf{z}(k, l)\,\pi_{kl},
\end{aligned}
$$

where we used $\Psi_1'(x) = \Psi_1^0(x)(1 - \Psi_1^0(x))$. Finally, since $\widehat{p}_k = \sum_l \widehat{p}_{kl}$ and $\widehat{q}_l = \sum_k \widehat{p}_{kl}$, the columns of the covariances $\mathsf{cov}(S_n, \widehat{p}_k)$ and $\mathsf{cov}(S_n, \widehat{q}_l)$ are $\mathsf{c}_{k\cdot} = \sum_j \mathsf{c}_{kj}$ and $\mathsf{c}_{\cdot l} = \sum_i \mathsf{c}_{il}$, respectively. Combining these results with (A.5) finally yields (A.12). □

A.4 Choice of S1 level and sample split ratio

In this section we present a theory based method for choosing the input parameters α_1 and δ in our two-stage methods by maximizing the overall power. For the sake of simplicity, we focus on testing $\beta_3 = 0$ against the one-sided alternative $\beta_3 > 0$ using a two-stage procedure based on T_n^A and D_n, which we refer to as *Pca*$_4$*–DS*. We consider a multiple testing scenario where a vast majority of the interaction hypotheses $H_{01}^{\text{epi}}, \ldots, H_{0K}^{\text{epi}}$ are true. Consequently, K_1 is approximately equal to $\alpha_1 K$. For the theory below we also assume no background dependence, which means

that all of the dependence between the genotypes is attributable to the presence of interaction. Such assumption is realistic for loci far enough from each other on the genome (such as located on different chromosomes).

Focusing on the power function of the Pearson chisquare statistic T_n^A, according to Cohen (1988) for large sample size the distribution of T_n^A is well-approximated by the non-central chisquare distribution with non-centrality parameter $\eta_\delta = \|\mu\|^2$, where $\mu = (\mu_{kl})_{k,l}$ and $\mu_{kl} = \sqrt{\delta m}\,(p_{kl} - p_k q_l)/\sqrt{p_k q_l}$. Its limiting power function is $\Pi^{(1)}(\alpha_1, \delta) = 1 - G(G^{-1}(1-\alpha_1, 4, 0), 4, \eta_\delta)$, where $G(\cdot, 4, \eta)$ is the non-central chisquare-four distribution function with non-centrality parameter η. Given that η_δ is a simple function of p_{kl}, under the LRM it is a matter of straightforward arithmetics to express it as a function of β_3. Doing so yields

$$\eta_\delta \;=\; \delta m \sum\nolimits_{k,l} \frac{\left[\pi_{kl}\Psi_{kl} - \left(\sum_{i,j}\pi_{ij}\,\Psi_{ij}\right)^{-1}\sum_{i,j}\pi_{kj}\pi_{il}\Psi_{kj}\Psi_{il}\right]^2}{\sum_{i,j}\pi_{kj}\pi_{il}\Psi_{kj}\Psi_{il}},$$

where $\Psi_{kl} = \Psi(\beta_0 + \beta_1 k + \beta_2 l + \beta_3 kl)$ and $\pi_{kl} = \mathsf{P}(X = k,\, Y = l)$, $\pi_{k\cdot} = \mathsf{P}(X = k)$, $\pi_{\cdot l} = \mathsf{P}(Y = l)$, which are the genotype frequencies in a hypothetical case-control population with the same prevalence of $\Delta = 1$ as in the case-control sample at hand.

Next we determine the power of the score test based on D_n. We focus on the local alternative of type $\beta_3 = h/\sqrt{n_2}$ for some $h > 0$, where $n_2 = n - m_1$ is the size of the sample size underlying D_n. The limiting power function of D_n is $\Pi^{(2)}(\alpha_1, \delta) = 1 - \Phi(\Phi^{-1}(1 - \alpha/(\alpha_1 K)) - Fh)$, where Φ denotes the standard normal distribution function and F is the slope of D_n. The slope of the score statistic D_n can be determined using Theorem 15.4 and Addendum 15.5 in Van der Vaart (1998) as $F = \left(\mathsf{e}_4' \mathsf{I}_{\beta_0}^{-1} \mathsf{e}_4\right)^{-1/2}$, where I_{β_0} is the null hypothesis FIM and $\mathsf{e}_4 = (0, 0, 0, 1)'$. Due to the independence of two tests the limiting power function of the combined test is $\Pi(\alpha_1, \delta) = \Pi^{(1)}(\alpha_1, \delta)\;\Pi^{(2)}(\alpha_1, \delta)$. By maximizing this power function with respect to α_1 and δ we can obtain suitable values for the two input parameters. The optimization requires as input the distribution of genotypes in the general population (via π_{kl}), the values of β_0, β_1, β_2 and finally the value of β_3 towards which the method is tuned, which can all be reasonably specified by the user.

A.5 Fisher-type p-value combination method

In the real data analysis of Section 5 instead of merging the four cohorts directly we instead combined the p-values using the weighted Fisher combination method (Fisher (1932), Box (1954)). For a general set of k p-values $p_1, \ldots, p_k$ the combination method utilizes the statistic $F = -2\sum_{i=1}^{k} w_i \log p_i$, where the weights w_i are based directly on the sample sizes underlying each p-value. The null hypothesis distribution of F is given by Theorem 2.4 of Box (1954).

APPENDIX B: LIST OF IPDGC MEMBERS AND AFFILIATIONS

Mike A Nalls (Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA), Vincent Plagnol (UCL Genetics Institute, London, UK), Dena G Hernandez (Laboratory of Neurogenetics, National Institute on Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK), Manu Sharma (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Una-Marie Sheerin (Department of Molecular Neuroscience, UCL Institute of Neurology), Mohamad Saad (INSERM U563, CPTP, Toulouse, France; and Paul Sabatier University, Toulouse, France), Javier Simón-Sánchez (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, and DZNE, German Center for Neurodegenerative Diseases, Tübingen, Germany), Claudia Schulte (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research), Suzanne Lesage (INSERM, UMR_S975 [formerly UMR_S679], Paris, France; Université Pierre et Marie Curie-Paris, Centre de Recherche de l'Institut du Cerveau et de la Moelle épinière, Paris, France; and CNRS, Paris, France), Sigurlaug Sveinbjörnsdóttir (Department of Neurology, Landspítali University Hospital, Reykjavík, Iceland; Department of Neurology, MEHT Broomfield Hospital, Chelmsford, Essex, UK; and Queen Mary College, University of London, London, UK), Sampath Arepalli (Laboratory of Neurogenetics, National Institute on Aging), Roger Barker (Department of Neurology, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK), Yoav Ben-Shlomo (School of Social and Community Medicine, University of Bristol), Henk W Berendse (Department of Neurology and Alzheimer

Center, VU University Medical Center), Daniela Berg (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research and DZNE, German Center for Neurodegenerative diseases), Kailash Bhatia (Department of Motor Neuroscience, UCL Institute of Neurology), Rob M A de Bie (Department of Neurology, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands), Alessandro Biffi (Center for Human Genetic Research and Department of Neurology, Massachusetts General Hospital, Boston, MA, USA; and Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA), Bas Bloem (Department of Neurology, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands), Zoltan Bochdanovits (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Michael Bonin (Department of Medical Genetics, Institute of Human Genetics, University of Tübingen, Tübingen, Germany), Jose M Bras (Department of Molecular Neuroscience, UCL Institute of Neurology), Kathrin Brockmann (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research and DZNE, German Center for Neurodegenerative diseases), Janet Brooks (Laboratory of Neurogenetics, National Institute on Aging), David J Burn (Newcastle University Clinical Ageing Research Unit, Campus for Ageing and Vitality, Newcastle upon Tyne, UK), Elisa Majounie (Laboratory of Neurogenetics, National Institute on Aging), Gavin Charlesworth (Department of Molecular Neuroscience, UCL Institute of Neurology), Codrin Lungu (National Institutes of Health Parkinson Clinic, NINDS, National Institutes of Health), Honglei Chen (Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, NC, USA), Patrick F Chinnery (Neurology M4104, The Medical School, Framlington Place, Newcastle upon Tyne, UK), Sean Chong (Laboratory of Neurogenetics, National Institute on Aging), Carl E Clarke (School of Clinical and Experimental Medicine, University of Birmingham, Birmingham, UK; and Department of Neurology, City Hospital, Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK), Mark R Cookson (Laboratory of Neurogenetics, National Institute on Aging), J Mark Cooper (Department of Clinical Neurosciences, UCL Institute of Neurology), Jean Christophe Corvol (INSERM, UMR_S975; Université Pierre et Marie Curie-Paris; CNRS; and INSERM CIC-9503, Hôpital Pitié-Salpêtrière, Paris, France), Carl Counsell (University of Aberdeen, Division of Applied Health Sciences, Population Health Section, Aberdeen, UK), Philippe Damier (CHU Nantes, CIC0004, Service de Neurologie, Nantes,

France), Jean-François Dartigues (INSERM U897, Université Victor Segalen, Bordeaux, France), Panos Deloukas (Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK), Günther Deuschl (Klinik für Neurologie, Universitätsklinikum Schleswig-Holstein, Campus Kiel, Christian-AlbrechtsUniversität Kiel, Kiel, Germany), David T Dexter (Parkinson's Disease Research Group, Faculty of Medicine, Imperial College London, London, UK), Karin D van Dijk (Department of Neurology and Alzheimer Center, VU University Medical Center), Allissa Dillman (Laboratory of Neurogenetics, National Institute on Aging), Frank Durif (Service de Neurologie, Hôpital Gabriel Montpied, Clermont-Ferrand, France), Alexandra Dürr (INSERM, UMR_S975; Université Pierre et Marie Curie-Paris; CNRS; and AP-HP, Pitié-Salpêtrière Hospital), Sarah Edkins (Wellcome Trust Sanger Institute), Jonathan R Evans (Cambridge Centre for Brain Repair, Cambridge, UK), Thomas Foltynie (UCL Institute of Neurology), Jing Dong (Epidemiology Branch, National Institute of Environmental Health Sciences), Michelle Gardner (Department of Molecular Neuroscience, UCL Institute of Neurology), J Raphael Gibbs (Laboratory of Neurogenetics, National Institute on Aging; and Department of Molecular Neuroscience, UCL Institute of Neurology), Alison Goate (Department of Psychiatry, Department of Neurology, Washington University School of Medicine, MI, USA), Emma Gray (Wellcome Trust Sanger Institute), Rita Guerreiro (Department of Molecular Neuroscience, UCL Institute of Neurology), Clare Harris (University of Aberdeen), Jacobus J van Hilten (Department of Neurology, Leiden University Medical Center, Leiden, Netherlands), Albert Hofman (Department of Epidemiology, Erasmus University Medical Center, Rotterdam, Netherlands), Albert Hollenbeck (AARP, Washington DC, USA), Janice Holton (Queen Square Brain Bank for Neurological Disorders, UCL Institute of Neurology), Michele Hu (Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Xuemei Huang (Departments of Neurology, Radiology, Neurosurgery, Pharmacology, Kinesiology, and Bioengineering, Pennsylvania State University Milton S Hershey Medical Center, Hershey, PA, USA), Isabel Wurster (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research and German Center for Neurodegenerative diseases), Walter Mätzler (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research and German Center for Neurodegenerative diseases), Gavin Hudson (Neurology M4104, The Medical School, Newcastle upon Tyne, UK), Sarah E Hunt (Wellcome Trust Sanger Institute), Johanna Huttenlocher (deCODE genetics),

Thomas Illig (Institute of Epidemiology, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany), Pálmi V Jónsson (Department of Geriatrics, Landspítali University Hospital, Reykjavík, Iceland), Jean-Charles Lambert (INSERM U744, Lille, France; and Institut Pasteur de Lille, Université de Lille Nord, Lille, France), Cordelia Langford (Cambridge Centre for Brain Repair), Andrew Lees (Queen Square Brain Bank for Neurological Disorders), Peter Lichtner (Institute of Human Genetics, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany), Patricia Limousin (Institute of Neurology, Sobell Department, Unit of Functional Neurosurgery, London, UK), Grisel Lopez (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI, National Institutes of Health), Delia Lorenz (Klinik für Neurologie, Universitätsklinikum Schleswig-Holstein), Codrin Lungu (National Institutes of Health Parkinson Clinic, NINDS, National Institutes of Health), Alisdair McNeill (Department of Clinical Neurosciences, UCL Institute of Neurology), Catriona Moorby (School of Clinical and Experimental Medicine, University of Birmingham), Matthew Moore (Laboratory of Neurogenetics, National Institute on Aging), Huw R Morris (National Hospital for Neurology and Neurosurgery, University College London, London, UK), Karen E Morrison (School of Clinical and Experimental Medicine, University of Birmingham; and Neurosciences Department, Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK), Valentina EscottPrice (MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff, UK), Ese Mudanohwo (Neurogenetics Unit, UCL Institute of Neurology and National Hospital for Neurology and Neurosurgery), Sean S O'Sullivan (Queen Square Brain Bank for Neurological Disorders), Justin Pearson (MRC Centre for Neuropsychiatric Genetics and Genomics), Joel S Perlmutter (Department of Neurology, Radiology, and Neurobiology at Washington University, St Louis), Hjörvar Pétursson (deCODE genetics; and Department of Medical Genetics, Institute of Human Genetics, University of Tübingen), Pierre Pollak (Service de Neurologie, CHU de Grenoble, Grenoble, France), Bart Post (Department of Neurology, Radboud University Nijmegen Medical Centre), Simon Potter (Wellcome Trust Sanger Institute), Bernard Ravina (Translational Neurology, Biogen Idec, MA, USA), Tamas Revesz (Queen Square Brain Bank for Neurological Disorders), Olaf Riess (Department of Medical Genetics, Institute of Human Genetics, University of Tübingen), Fernando Rivadeneira (Departments

of Epidemiology and Internal Medicine, Erasmus University Medical Center), Patrizia Rizzu (Department of Clinical Genetics, Section of Medical Genomics, VU University Medical Centre), Mina Ryten (Department of Molecular Neuroscience, UCL Institute of Neurology), Stephen Sawcer (University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's hospital, Cambridge, UK), Anthony Schapira (Department of Clinical Neurosciences, UCL Institute of Neurology), Hans Scheffer (Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands), Karen Shaw (Queen Square Brain Bank for Neurological Disorders), Ira Shoulson (Department of Neurology, University of Rochester, Rochester, NY, USA), Joshua Shulman (Baylor College of Medicine, Houston, Texas), Ellen Sidransky (Section on Molecular Neurogenetics, Medical Genetics Branch, NHGRI), Colin Smith (Department of Pathology, University of Edinburgh, Edinburgh, UK), Chris C A Spencer (Wellcome Trust Centre for Human Genetics, Oxford, UK), Hreinn Stefánsson (deCODE genetics), Francesco Bettella (deCODE genetics), Joanna D Stockton (School of Clinical and Experimental Medicine), Amy Strange (Wellcome Trust Centre for Human Genetics), Kevin Talbot (University of Oxford, Department of Clinical Neurology, John Radcliffe Hospital, Oxford, UK), Carlie M Tanner (Clinical Research Department, The Parkinson's Institute and Clinical Center, Sunnyvale, CA, USA), Avazeh Tashakkori-Ghanbaria (Wellcome Trust Sanger Institute), François Tison (Service de Neurologie, Hôpital HautLévêque, Pessac, France), Daniah Trabzuni (Department of Molecular Neuroscience, UCL Institute of Neurology), Bryan J Traynor (Laboratory of Neurogenetics, National Institute on Aging), André G Uitterlinden (Departments of Epidemiology and Internal Medicine, Erasmus University Medical Center), Daan Velseboer (Department of Neurology, Academic Medical Center), Marie Vidailhet (INSERM, UMR_S975, Université Pierre et Marie Curie-Paris, CNRS, UMR 7225), Robert Walker (Department of Pathology, University of Edinburgh), Bart van de Warrenburg (Department of Neurology, Radboud University Nijmegen Medical Centre), Mirdhu Wickremaratchi (Department of Neurology, Cardiff University, Cardiff, UK), Nigel Williams (MRC Centre for Neuropsychiatric Genetics and Genomics), Caroline H Williams-Gray (Department of Neurology, Addenbrooke's Hospital), Sophie Winder-Rhodes (Department of Psychiatry and Medical Research Council and Wellcome Trust Behavioural and Clinical Neurosciences Institute, University of Cambridge), Kári Stefánsson (deCODE genetics), Maria Martinez (INSERM UMR 1043; and Paul

Sabatier University), Nicholas W Wood (UCL Genetics Institute; and Department of Molecular Neuroscience, UCL Institute of Neurology), John Hardy (Department of Molecular Neuroscience, UCL Institute of Neurology), Peter Heutink (DZNE, German Center for Neurodegenerative Diseases and Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany), Alexis Brice (INSERM, UMR_S975, Université Pierre et Marie Curie-Paris, CNRS, UMR 7225, AP-HP, Pitié- Salpêtrière Hospital), Thomas Gasser (Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research, and DZNE, German Center for Neurodegenerative Diseases), Andrew B Singleton.

References

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics* **25**(2), 290–302.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, second edition. Hillsdale, New Jersey, USA: Lawrence Erlbaum.

Fisher, R.A. (1932). *Statistical Methods for Research Workers*, fourth edition. London, UK: Oliver and Boyd.

Van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.