Received: 21 April 2021 | Revised: 21 March 2022 | Accepted: 24 March 2022

DOI: 10.1002/bimj.202100123

Biometrical Journal

RESEARCH ARTICLE

# A powerful global test for spliceQTL effects

**Renee X. de Menezes**[1,2] | **Armin Rauschenberger**[2,3] | **BIOS Consortium**[4] | **Peter A. C. 't Hoen**[5] | **Marianne A. Jonker**[6]

[1]Department of Psychosocial Research and Epidemiology, Room H.8.040, Netherlands Cancer Institute, Amsterdam, The Netherlands

[2]Department of Epidemiology and Data Science, Amsterdam UMC, location VUmc, Amsterdam, The Netherlands

[3]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg

[4]Biobank-based Integrative Omics Study Consortium, The Netherlands

[5]Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands

[6]Department for Health Evidence, section Biostatistics, Radboud University Medical Center, Nijmegen, The Netherlands

**Correspondence**
Renee X. de Menezes, Department of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Postbus 90203, 1006 BE Amsterdam, The Netherlands.
Email: r.menezes@nki.nl
Marianne A. Jonkers, Department for Health Evidence, section Biostatistics, Radboud University Medical Center, 6525 EZ Nijmegen, The Netherlands.
Email: Marianne.Jonker@radboudumc.nl

**Funding information**
Netherlands Organization for Scientific Research, Grant/Award Number: 184021007; Innovation-Oriented Research Program on Genomics, Grant/Award Numbers: IGE01014, IGE05007; Centre for Medical Systems Biology; National Institute for Healthy Ageing, Grant/Award Numbers: 05040202, 05060810

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to the data being not publicly available.

## Abstract

Statistical methods to test for effects of single nucleotide polymorphisms (SNPs) on exon inclusion exist but often rely on testing of associations between multiple exon–SNP pairs, with sometimes subsequent summarization of results at the gene level. Such approaches require heavy multiple testing corrections and detect mostly events with large effect sizes. We propose here a test to find spliceQTL (splicing quantitative trait loci) effects that takes all exons and all SNPs into account simultaneously. For any chosen gene, this score-based test looks for an association between the set of exon expressions and the set of SNPs, via a random-effects model framework. It is efficient to compute and can be used if the number of SNPs is larger than the number of samples. In addition, the test is powerful in detecting effects that are relatively small for individual exon–SNP pairs but are observed for many pairs. Furthermore, test results are more often replicated across datasets than pairwise testing results. This makes our test more robust to exon–SNP pair-specific effects, which do not extend to multiple pairs within the same gene. We conclude that the test we propose here offers more power and better replicability in the search for spliceQTL effects.

KEYWORDS
gene set testing, multinomial response, multivariate outcome, $p >> n$, score test

## 1 | INTRODUCTION

A gene is a chromosomal region defined on DNA, which may contain subunits called exons alternating with introns. After transcription of the gene into a pre-mRNA, the introns are removed and only a subset of the exons within the gene may be included in the mature mRNA. This process is referred to as mRNA splicing. Each distinct subset of exons potentially yields a different transcript of the gene, which may lead to proteins with different functions. Alternative splicing is the fact that different transcripts can be produced, as a consequence of different subsets of exons being spliced. A schematic representation of (alternative) splicing and its link to DNA can be found in Supplementary Figure 1. Alternative splicing events can be studied by considering exon-level counts from RNA-sequencing data: it can be observed if samples present different relative amounts of exons of the same gene.

Understanding what affects alternative splicing has been recognized as yet another important challenge for postgenome biology (Black, 2000; Modrek & Lee, 2002). Our knowledge of alternative splicing has grown through the use of RNA-sequencing (RNA-seq). It is now known that between 92% and 95% of multiexon genes in higher eukaryotes contain at least one alternative splicing event (Pan et al., 2008; Wang et al., 2008; Sultan et al., 2008). Single nucleotide polymorphisms (SNPs) refer to positions in the genome that have variable genotypes in the population: individuals may have different DNA nucleotides at such positions. These may be linked to a clinical or a molecular phenotype, such as gene expression or splicing. There is, thus, interest in studying how SNPs affect alternative splicing.

The wider availability and affordability of high-throughput molecular profiling technology has led to an increasing number of omics studies where each sample has multiple high-dimensional molecular profiles, including for example both SNP and mRNA expression profiles. This paves the way for studying the impact of (populational) genomic variation on alternative splicing. The challenge is to combine these two high-dimensional datasets by means of a model and develop a statistical test that is powerful enough to detect associations of interest.

Here we will focus on alternative splicing that is associated with SNPs, and we aim to find these spliceQTLs (splicing quantitative trait loci). Specifically, the problem involves finding associations between exon-level mRNA data in a given gene and SNP genotypes in a window around the gene. Methods currently available to handle this problem are based on pairwise correlation and/or testing, thus involving one pair of exon–SNP at a time. The multiple-testing correction here is severe due to the large number of exon–SNP pairs, leading to low power for finding effects. In addition, such approaches ignore effects that span many exons and/or many SNPs. A method that, per gene, takes into account data of all exons, as well as all SNPs of interest, will not only decrease the multiple-testing burden but can also benefit from considering more data (multiple exons and SNPs) at the same time. This can be particularly useful when exon–SNP associations are not very strong for any individual pair but exist at a detectable level for multiple exon–SNP pairs. Such associations cannot typically be distinguished from noise when testing pairs individually, while their occurrence in multiple exon–SNP pairs mapping to the same gene suggests that they are unlikely to be due to noise.

One way to increase the utility of the study is to decrease the multiple testing by diminishing the number of tests. Some authors have opted for first selecting features (genes or exons, depending on the method) associated with alternative splicing and subsequently use those features to examine whether they displayed associations with SNPs after applying existing statistical methods (Li et al., 2018; Zhang et al., 2020; Tian et al., 2019). For example, Li et al. (2018) first used their suggested method to estimate alternatively spliced introns, then subsequently performed a method to find eQTLs (fastqtl proposed by Ongen et al., 2016) per intron cluster, after dimension reduction of both genotype and expression data. Similarly, both Zhang et al. (2020) and Tian et al. (2019) used an existing method to first yield a set of exons associated with alternative splicing and subsequently estimated the association between each selected exon with each SNP located within a given window from it. The proposed selection procedures are based on comparisons between groups, for example, between tumor and normal samples, or between different tumor types. So, using such approaches, the search for spliceQTLs is limited to genes that display strong enough alternative splicing between two or more groups. Being critical of the choice of the set of genes and SNPs is helpful, but they should be additional to using powerful statistical methodology. In this paper, we focus on a novel powerful test that can be applied to any selection of genes and SNPs.

A method that considers all exons and many SNPs per gene needs a model to make sense of all available data at once. On the one hand, we would like to better understand the expression variation of multiple exons at the same time. On the other hand, we wish to consider how this variation is associated with genotypes of SNPs in a large window, say all SNPs located between the start and end of the gene. One possibility is to consider a multiple regression framework, where all exon expressions are considered as response variables and the SNP genotypes are explanatory variables. However, the number of SNPs can amount to a number larger than the number of samples. While penalized regression would offer a solution to fit the model (e.g., Pecanka et al., 2019), it does not yield a reliable testing tool, due to the penalty parameter estimation.

Another important issue is the SNPs' correlation structure. Genotypes for different SNPs can display strong and even full association. Such strong correlations essentially make approaches based on classic regression models unsuitable for use, as these require uncorrelated covariates and breakdown under (near) full collinearity between covariates. Thus, methods to test for spliceQTL should be able to handle not only a large number of covariates but also correlated covariates.

In the next section, we will present a statistical test that solves the aforementioned problems. It is based upon the score function, which considers the association between expression data for all of the gene's exons and all SNPs considered to be of interest, using all data referring to a given gene at once. Our test is designed to find specifically spliceQTL effects, and not eQTL effects, by means of conditioning the gene's total expression (Supplementary Figure 1). Furthermore, the empirical correlation structure between SNPs can be taken into account. In the Results section, we perform a simulation study to evaluate the test's power to find true effects, for various effect types. In the same section, the test is applied to two publicly available datasets, the GEUVADIS/1000Genomes data, as well as to BIOS data, enabling us to evaluate how reproducible the selection of genes is. Finally, we gather our findings in the Discussion.

## 2 | METHODS

### 2.1 | The model

Suppose a gene with $K$ exons is considered, with $K \geq 2$ (a gene with $K = 1$ exon only has one isoform, so it cannot display alternative splicing). As response, we take the set of $K$ exon-specific measurements, denoted as the $K \times 1$ column vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})^t$ for sample $i$, and as covariates the $M$ SNP genotypes, denoted by the $M \times 1$ column vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{iM})^t$ for sample $i$ ($i = 1, \ldots, n$). Each exon-specific measurement quantifies the amount of RNA found to map to that exon and, to reflect the fact that here we are handling RNA quantification, from now on we will refer to exon-specific measurements as "exon expression," even though strictly speaking exons are not individually expressed but rather are included in transcripts. The exact definition of the covariates depends on the assumed underlying genetic model; for an additive model, $X_{ij}$ is equal to the number of minor alleles at the corresponding SNP (so taking values in $\{0, 1, 2\}$) and for the dominant or recessive genetic model $X_{ij}$ is binary. In the simulation studies, the additive model was assumed. The relationship between exon expression levels and SNP genotypes is assumed to be

$$\mathrm{E}(\mathbf{Y}_i|\mathbf{X}_i) = h^{-1}\left(\boldsymbol{\alpha} + \mathbf{B}^t\mathbf{X}_i\right), \tag{1}$$

where $\mathbf{B}$ is an $M \times K$ matrix in which the $k$th column is the vector $\boldsymbol{\beta}_k \equiv (\beta_{1k}, \beta_{2k}, \ldots, \beta_{Mk})^t$ of regression coefficients relating exon $k$ to the $M$ SNP genotypes, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^t$ is the intercept. The function $h$ is an $\mathbb{R}^K \to \mathbb{R}^K$ link function; it links the $K$-dimensional vector of expected exon-expression and the linear combination of the coded SNP genotypes for every exon. The exact definition of this link function is determined by the distribution of $\mathbf{Y}_i$. For the derivation of the score test-statistic, it is assumed that the response vector $\mathbf{Y}_i$ follows a distribution belonging to the canonical exponential family with a canonical link function (see, for instance, McCullagh and Nelder, 1989, Table 2.1). The response, exon-expression, may either consist of raw counts or normalized values. In the first case, a multinomial distribution for $\mathbf{Y}_i$ will be assumed, and for the latter the multivariate normal distribution would be appropriate. Both distributions are members of the canonical exponential family.

Based on data of $n$ independent samples as just described, the aim is to test per gene whether there is an association between its exon-specific expression levels and the SNP genotypes. In terms of the model, this means that we want to test the null hypotheses $\mathrm{H}_0 : \beta_{m1} = \beta_{m2} = \cdots = \beta_{mK} = 0$ for all $m = 1, \ldots, M$. Note that alternative splicing may still take place if the combined null hypotheses hold, by means of different $\alpha_k$ in $\boldsymbol{\alpha}$, but this event is not associated with the SNP genotypes. However, if $\alpha_k \equiv \alpha$ for all $k$, all exons have the same expected expression level and there can be no alternative splicing under the null hypothesis. Supplementary Figure 1 illustrates how SNP genotypes and exon-level data are connected via the model.

If $M$ is much larger than $n$, this test cannot be performed with $\boldsymbol{\beta}$ nonstochastic, since in such a case the model cannot be fitted. To make testing possible, we assume each $\boldsymbol{\beta}_k$ to be a random vector with a mean zero, for $k = 1, \ldots, K$. Given the vector of covariates $\mathbf{X}_i$ for sample $i$, we define $r_{ik} := \sum_{m=1}^{M} X_{im}\beta_{mk} = \boldsymbol{\beta}_k^t\mathbf{X}_i$ as a random effect corresponding to exon $k$ of sample $i$, and $\mathbf{r}_i = (r_{i1}, \ldots, r_{iK})^t$ as the random effect column vector for sample $i$. By construction, $\mathbf{r}_i$ has mean zero and

for any pair of samples indexed by $i, j$ and pair of exons indexed by $k, l$, we have

$$\begin{aligned}\mathrm{Cov}\big(r_{ik}, r_{jl}|\mathbf{X}_i, \mathbf{X}_j\big) &= \mathrm{Cov}\big(\boldsymbol{\beta}_k^t\mathbf{X}_i, \boldsymbol{\beta}_l^t\mathbf{X}_j|\mathbf{X}_i, \mathbf{X}_j\big)\\ &= \mathbf{X}_i^t\,\mathrm{Cov}(\boldsymbol{\beta}_k, \boldsymbol{\beta}_l)\mathbf{X}_j \equiv \sigma^2\mathbf{X}_i^t\,\mathbf{W}_{kl}\mathbf{X}_j,\end{aligned} \tag{2}$$

where $\mathbf{W}_{kl}$ is an $M \times M$ symmetric matrix and $\sigma^2$ is a constant chosen so that $\mathrm{diag}(\mathbf{W}_{kk})$ has all entries equal to 1 (for all $k$); below a motivation for this assumption is given. Conditional on the random effects $\mathbf{r}_1, \ldots, \mathbf{r}_n$, the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are assumed to be independent and the link as given in (1) can be rewritten as

$$\mathrm{E}\,(\mathbf{Y}_i|\mathbf{r}_i) = h^{-1}(\boldsymbol{\alpha} + \mathbf{r}_i) \tag{3}$$

(with possibly a different intercept, notwithstanding the notation).

Testing the null hypothesis that $\boldsymbol{\beta}_k = 0$ for all $k = 1, \ldots, K$ in the original fixed effects model is now equivalent to testing $\mathrm{H}_0 : \sigma^2 = 0$ against $\mathrm{H}_a : \sigma^2 > 0$ in the random effects model. Note that when $\mathrm{H}_0$ is rejected, there is evidence that alternative splicing is observed *and* it displays association with genotypes at different SNPs, which we refer to as *spliceQTL*.

The covariance structure between random effects $r_{ik}$ and $r_{jl}$ depends on the matrix $\mathbf{W}_{kl}$. By modeling this matrix, the way the correlation between the covariates (SNPs in our case) is taken into account by the test statistic can be fine-tuned. At this point, it is assumed that $\mathbf{W}_{kk} \equiv \mathbf{W}$, for $k = 1, \ldots, K$, under $\mathrm{H}_0$. This can be interpreted as assuming that the correlation structure between any pair of effects of covariates on one response (exon $k$) does not depend on the particular exon involved. So, this assumption is equivalent to saying that the exons are exchangeable, as far as the covariances between their effects are concerned. In addition, it is assumed that $\mathbf{W}_{kl} = 0\mathbf{I}$ for all $k, l = 1, \ldots, K$ with $k \neq l$, under the null hypothesis $\mathrm{H}_0$ (for $\mathbf{I}$ the identity matrix). This essentially means that random effects relating to different exons are uncorrelated. This also seems reasonable to assume under $\mathrm{H}_0$.

In application, $\mathbf{W}$ is replaced by an estimate and either $\widehat{\mathbf{W}} = \mathbf{I}_M$ or $\widehat{\mathbf{W}} = \mathbf{X}\mathbf{X}^t$ is used, where $\mathbf{X}$ is an $M \times n$ matrix with $\mathbf{X}_i$ as its $i$th column; in particular, in the latter case a generic $[\widehat{\mathbf{W}}]_{pq}$ entry of $\widehat{\mathbf{W}}$ is given by $\sum_i X_{ip}X_{iq}$, which is related to the sample covariance between SNP genotypes $p$ and $q$ across all samples. Indeed, $\mathbf{X}\mathbf{X}^t$ represents the sample correlation matrix if the covariates in $\mathbf{X}$ are centered and scaled. In this way, the linkage disequilibrium structure of the SNPs is used to model the corresponding correlation structures between the association effects between the SNPs and a single exon.

## 2.2 | Test statistic for splice-changing events

In order to test for spliceQTL, that is, alternative splicing associated with SNP genotypes, the null hypothesis $\mathrm{H}_0 : \sigma^2 = 0$ is tested against $\mathrm{H}_a : \sigma^2 > 0$. Suppose that $\mathbf{Y}_i$ is a vector of count variables, as it would be if exon-specific expression data were generated by RNA-seq. Then $\mathbf{Y}_i|\mathbf{r}_i$ (i.e., $\mathbf{Y}_i$ given the random effect $\mathbf{r}_i$) can be assumed to follow a multinomial distribution, given the total gene expression per sample $i$, and the score-test statistic (see Section 1 of the Supporting Information for details) equals

$$S = \mathrm{trace}\big[\mathbf{Y}(\mathbf{I}_n - \mathbf{H})^t\mathbf{X}^t\widehat{\mathbf{W}}\mathbf{X}(\mathbf{I}_n - \mathbf{H})\mathbf{Y}^t\big], \tag{4}$$

where $\mathbf{Y}$ is a $K \times n$ matrix with the $i$th column given by $\mathbf{Y}_i$, so that the rows correspond to the exons and the columns to the samples. Here $\mathbf{I}_n$ represents the $n \times n$ identity matrix and $\mathbf{H} = N^{-1}\mathbf{J}(N_1, \ldots, N_n)$ with $N = \sum_{i=1}^n N_i$, $N_i = \sum_{k=1}^K Y_{ik}$ and $\mathbf{J}(N_1, \ldots, N_n)$ the $n \times n$ matrix filled with $N_j$ at row $j$. Since for each $i$, $\mathbf{Y}_i$ is assumed to follow a multinomial distribution conditional on the total gene expression $\sum_k Y_{ik}$, the test statistic does not change with changing total gene expression. This implies that the test is robust to eQTL effects, and is thus aimed at finding spliceQTL effects that cannot be explained by changes in total gene expression associated with SNP genotypes.

In case $\widehat{\mathbf{W}} = \mathbf{X}\mathbf{X}^t$, (4) becomes

$$S_W = \mathrm{trace}\big[\mathbf{Y}(\mathbf{I}_n - \mathbf{H})^t\mathbf{X}^t\mathbf{X}\mathbf{X}^t\mathbf{X}(\mathbf{I}_n - \mathbf{H})\mathbf{Y}^t\big], \tag{5}$$

while, if $\widehat{\mathbf{W}} = \mathbf{I}_M$, (4) becomes

$$S_I = \mathrm{trace}\big[\mathbf{Y}(\mathbf{I}_n - \mathbf{H})^t\mathbf{X}^t\mathbf{X}(\mathbf{I}_n - \mathbf{H})\mathbf{Y}^t\big]. \tag{6}$$

The test statistic $S_I$ in (6) involves the sample correlation matrix of the (centered and scaled) covariates in $\mathbf{X}$, while $S_W$ in (5) involves the square of the same sample correlation matrix. As such, the impact of the correlation structure on the observed value of the test statistic is smaller when $S_W$ is used, compared with when $S_I$ is used. This means that, in cases where the covariates display strong correlation, $S_I$ will tend to yield more extreme values than $S_W$.

The proposed test statistic depends on the term $(\mathbf{I}_n - \mathbf{H})\mathbf{Y}^t = \mathbf{Y}^t - \mathbf{H}\mathbf{Y}^t$. The matrix $\mathbf{H}\mathbf{Y}^t$ is an estimate of the expectation $\mathrm{E}\,\mathbf{Y}_i$ under the null hypothesis of no association (see the Supporting Information). So, the test statistic can be interpreted as the squared distance of the response and its expectation, corrected for the dependence between SNPs.

For the problem where exon-level expression data are available per gene, the multinomial distribution is a natural one to assume for the conditional distribution of the response. However, our results can also be used for other applications with a multivariate response. In fact, we derived a test statistic based upon the score function under the general assumption that the conditional distribution of $\mathbf{Y}_i$ given $\mathbf{r}_i$ belongs to a canonical exponential family (see Section 1 of the Supporting Information). We gave here the exact form of the test statistic for a multivariate response with a multinomial distribution. We also give an expression for the case of a response with a multivariate normal distribution (see Section 1 of the Supporting Information).

The distribution of (4) under $\mathrm{H}_0$ is not known, so $p$-values are computed by using permutations on the observed data. In this case, samples are permuted, and the test statistic is recomputed for each permuted data.

When the test statistic (4) is applied to many genes simultaneously, multiple testing correction needs to be applied to the computed $p$-values. Here we use the step-up false discovery rate (FDR) procedure proposed by Benjamini and Hochberg (1995).

## 2.3 | Relation with other works

The proposed test statistic (6) is related to the univariate multinomial test proposed in Goeman and Le Cessie (2006): the two test statistics can be shown to be equivalent (see Section 2.2 in the Supporting Information). However, the formulation used in Goeman and Le Cessie (2006) requires a vector with one observation per read as a response, indicating which exon that read mapped to. This is not only impractical but also requires handling very large objects during computations, making them at least inefficient. In contrast, the formulation used here is more convenient when the response variables are tabulations of the reads observed per exon, as it is the case with exon expression, requiring much smaller objects and thus allowing for considerably more efficient computation.

Note also that our multivariate multinomial formulation uses a vector of per-exon expression counts as a single, multivariate response per gene. Previously, Chaturvedi et al. (2017) proposed a test for a multivariate normal response, so the current test statistic can be seen as an extension of this earlier test statistic for the multinomial case. In addition, we here allow for explicit modeling of the covariance structure between effects via $\mathbf{W}$, which was not done by Chaturvedi et al. (2017).

## 2.4 | Software

Functions implementing the spliceQTL test proposed are available via the R package `spliceqtl`, available from https://github.com/rxmenezes/spliceQTL. Supplementary files including code and data used to generate the results in this paper can be found at https://github.com/rxmenezes/spliceQTLsuppMat.

## 2.5 | Experimental data used in examples

We applied our proposed test to study the association between per-gene variations in exon expression and the number of minor alleles on SNPs within or around the gene. For this, we used two publicly available datasets that involve unrelated samples with each RNA-seq data yielding exon-level counts. Both datasets have corresponding whole-genome SNP data.

In Example 1, we compare effects found by our proposed method with those found by a pairwise-testing approach. In the latter, each pair of exon–SNP is tested individually. Results from each analysis are corrected for multiple testing. For this, molecular profiles for samples involved in the GEUVADIS project (http://www.geuvadis.org/web/geuvadis) were used. Briefly, in the Lappalainen et al. (2013) study, RNA from HapMap samples was sequenced and analyzed, and the data are available via ArrayExpress with accession number E-GEUV-3. Exon expression data were obtained (http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/), and for these samples SNP genotype data were produced

by the 1000Genomes project, obtained from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/. Only European populations were included in the analysis, totaling 373 samples; Yoruban samples were excluded as their genetic populational structure is too different from that of the European populations. Features mapping to chromosome 1 were used for analyses.

The SNP genotype data were taken as the number of minor allele copies per SNP and per sample. We assumed an additive effect of the number of minor alleles on the exon expression. This allowed us to consider the number of minor alleles for any given SNP as numeric in the model, taking a value in $\{0, 1, 2\}$, where variables were coded so that the major allele was the one with the highest frequency in the dataset at hand. We also only kept SNPs with a minor allele larger than 0.05.

After obtaining the RNA-seq data, exons with zero variation were excluded, and genes containing only a single exon were left out, as in such cases no alternative splicing can be detected (in the former) or can take place (in the latter).

For each gene, we considered SNPs located within 2 megabases of the gene's transcription start site, in either direction, as previously done by Lappalainen et al. (2013).

In Example 2, we assess the reproducibility of the spliceQTL test by comparing results obtained using two independent datasets. The first dataset was produced in the context of the Biobank-based Integrative Omics Study (Zhernakova et al., 2017). Only the samples from the Leiden Longevity Study (LLS; Schoenmaker et al., 2006) were included in our analysis. We compare lists of genes obtained with those from the GEUVADIS data, used in Example 1. To ensure better comparability, the GEUVADIS data were aligned and preprocessed again in the same way as the LLS data. By assessing the extent to which results can be replicated across these datasets, we draw conclusions about the robustness of the spliceQTL test, relative to that of pairwise testing.

## 3 | RESULTS

### 3.1 | Simulation study

We ran a simulation study to evaluate the performance of using our multinomial-response test statistic, using either $S_W$ as in (5) or $S_I$ as in (6), and each under various scenarios. For clarity, both are spliceQTL test statistics, with $S_W$ using the square of the sample correlation matrix while $S_I$ involves the untransformed matrix. We constructed scenarios of practical interest and made various assumptions. For a full description, see Section 3 of the Supporting Information. Below we give a brief description of our method for clarity.

It is difficult to generate SNP data at random in a way that represents realistic correlation structures. To ensure that our simulation was based on a realistic situation, we used experimental SNP data produced by the 1000Genomes study to generate two sets of 100 SNP datasets each. Specifically, we used SNP genotype data included in windows around two different genes. These genes were chosen at random from those found to be significant by the spliceQTL test on chromosome 1. We then partitioned at random the 373 samples available for each dataset into five disjoint subsets of around 74 samples each and used 20 distinct partitions, yielding a total of 100 SNP genotype sets of around 74 samples each. This is done in order to generate 100 SNP datasets with a realistic correlation structure and sample size. In particular, the structure in the SNP data can accommodate the significant association with the exon-level counts. We will refer to results originating from each of those original SNP data matrices as "gene 1" and "gene 2" because they arise from two separate genes while noting that they actually refer merely to different SNP data.

Matrices of effect sizes were generated at random to represent different effect patterns, which we call "scenarios," as follows: the strong effect of just a few (2%) SNPs on the response vector (S1), weak effect of a few (2%) SNPs (S2), and weak effect of many (20%) SNPs (S3). The same matrices were applied to both genes for each given scenario. As a consequence, results for gene 1 and gene 2 only differ in that the SNPs used are different. While the effect sizes are generated at random and independently from each other, the underlying SNPs may be correlated, yielding correlated random effects. The use of independent effect sizes means that responses are simulated under the assumption that $\mathbf{W} = \mathbf{I}$. However, the underlying SNP correlation may lead to a departure from this assumption.

SNPs in the two original SNP data matrices differ noticeably how strong are their pairwise correlations. Indeed, those arising from gene 1 display stronger correlations in general than those arising from gene 2 (Supplementary Figure 2). These two genes represent somewhat extremes in terms of how correlated their SNPs are, with other sets of consecutive SNPs displaying correlation patterns in between (Supplementary Figure 3). As a consequence, results using SNPs from gene 1 are more likely to display differences between the test statistics using $\widehat{\mathbf{W}} = \mathbf{I}$ and those using $\widehat{\mathbf{W}} = \mathbf{X}\mathbf{X}^t$, than those from gene 2.

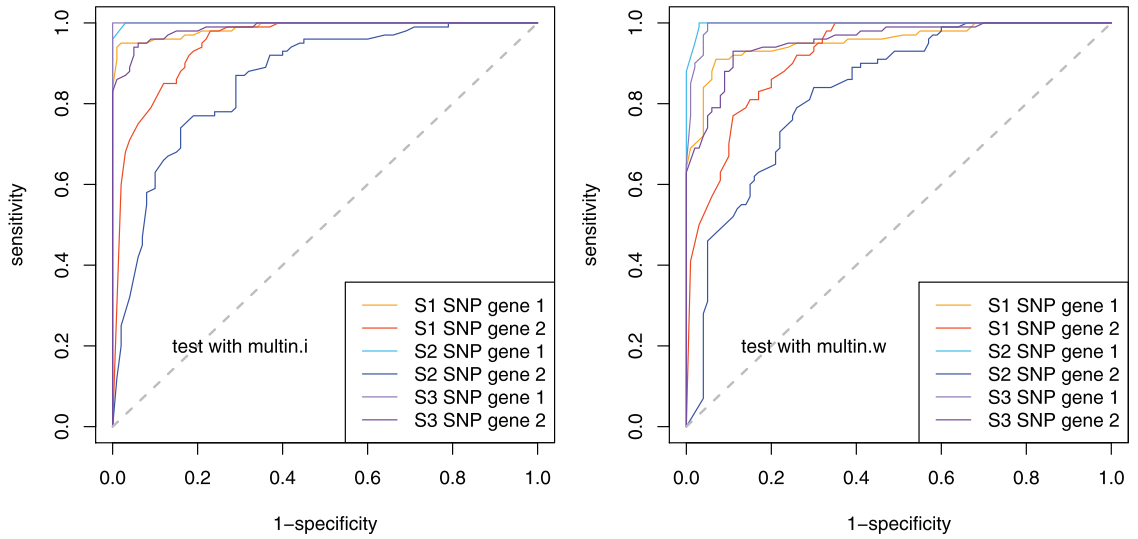


**FIGURE 1** ROC curves generated for all scenarios, using the multinomial test with either $\widehat{\mathbf{W}} = I$ (left: multin.i) or $\widehat{\mathbf{W}} = X^t X$ (right: multin.w). The different scenarios are as follows: strong effect of just a few (2%) SNPs on the response vector (S1), weak effect of a few (2%) SNPs (S2), and weak effect of many (20%) SNPs (S3). A full description of the study setup is given in Section 3 of the Supporting Information

The different scenarios are used to illustrate different patterns of association between SNPs and exons: we want to illustrate that in scenario S1 effects are not missed by our approach which takes many exons and many SNPs simultaneously into account. Scenario S3 is where our approach should perform particularly well, by taking many (correlated) variables into account. Scenario S2 is an intermediate between S1 and S3. We evaluate the performance of our spliceQTL test using a combination of genes (in fact, simply different SNP data matrices, with a somewhat different degree of correlation between SNPs involved) and scenarios, to assess the impact of these aspects on the overall power of the test. Finally, we also consider two numbers of exons in the response: 3 and 50. This helps us assess the impact of a possible dilution of effects in the presence of many (50), compared to only a few (3). As a reference, human transcripts can have up to 363 exons, with 50% having up to 9 exons and 99% of them having up to 44 exons (Piovesan et al., 2019).

After generating response vectors with either 3 or 50 exons, we used the multinomial spliceQTL test with either $S_I$ or $S_W$, on 100 datasets, per scenario, gene, and number of responses. For each dataset, half of the samples were assumed to involve an effect of SNPs on the response vector, while the remaining samples had response vectors simulated with no effect. This may represent a situation observed in practice, where a subset of samples display an association between SNPs and exon expression levels. Such inter-individual variability in the strength of the association between SNP and exon expression may arise through differences in cell-type composition and environmental signaling, which has been shown to exist for eQTLs (Zhernakova et al., 2017).

Using only the results for the response vectors with no effect, we note that the test level is on average the desired level (top-row graphs of Supplementary Figure 4 and Supplementary Table 3), with results for $S_I$ and $S_W$ approximately the same. Indeed, the variation in test level is largest between genes (larger for gene 2 than for gene 1) and scenarios (the tests tend to select more false positives in S2), than between the two test formulations. In addition, the empirical distribution of the computed $p$-values is very similar to the uniform distribution in all cases (data not shown).

Receiver-operator characteristic (ROC) curves for the different situations and genes suggest that the power to find effects varies, depending on how strongly the SNPs are correlated, as well as the scenario (Figure 1). This variability is observed both when $S_I$ was used, as well as when $S_W$ was used. A general trend is that the area under the curve (AUC) is higher when SNPs are more strongly correlated (gene 1, the left-hand side graph in Figure 1) than when they are weakly correlated (gene 2, the right-hand side graph in Figure 1). The lowest AUCs were obtained when SNPs displayed low correlation (gene 2; see Figure 1): the lowest was for a small effect, involving few SNPs (S2), followed by the one with a large effect, again involving few SNPs (S1). Note that each ROC curve combines results using 3 exons as a response vector, with those using 50 exons as a response vector.

We also observed that the AUC was in general higher when $S_I$ was used than when $S_W$ was used in all considered situations (Supplementary Figure 5). As the data were simulated using $\widehat{\mathbf{W}} = \mathbf{I}$, this was expected. This is confirmed by the proportions of response vectors selected using various thresholds (bottom-row graphs in Supplementary Figure 4).

However, the differences in AUCs are smaller between the results using $S_I$ and $S_W$ than those between genes or scenarios. This suggests that the test is robust to the misspecification of $\mathbf{W}$ in practical situations where the true matrix $\mathbf{W}$ is unknown.

## 3.2 | Example 1: Comparison with pairwise testing

We will compare results of the spliceQTL test with those obtained by testing all exon–SNP pairs individually. For this, we will use data from the GEUVADIS study, previously analyzed in Lappalainen et al. (2013). Here we use only the test statistic $S_W$ given by (5), as the true $\mathbf{W}$ is unknown but it is unlikely that the SNPs are uncorrelated. Pairwise testing involved using the linear regression to test for the association between the expression of each exon and each SNP variable, for each exon–SNP pair per gene. This yielded a matrix of $p$-values per gene.

The data were preprocessed as indicated in Section 2.5. Additionally, the exon-length effect on expression was corrected by means of a mixed-effects model (see Section 4 of the Supporting Information and Supplementary Figure 6).

Testing for spliceQTL was done per gene, as follows. We used our spliceQTL model to test for the association of exon-level counts with all SNP variables, yielding one $p$-value per gene. For comparison, we used the linear regression to test for association between each pair of exon–SNP, with the exon-level counts as response variables and the SNP (number of minor alleles) as explanatory variables. Per test type, $p$-values were corrected for multiple testing using the Bonferroni method.

As expected, the number of tests performed in the pairwise approach (53,286,610) is much larger than the number of genes on chromosome 1 (1,376), which is the same as the number of tests required by the spliceQTL approach. The obvious consequence is that the multiple testing correction in the former is much more severe than in the latter. Thus, an effect size at the exon–SNP pair level must be much larger to yield a significant test result compared to the effect size required to yield a significant test with the spliceQTL test.

In order to compare results at the exon–SNP level, we further refined the associations selected with the spliceQTL test using a sequential approach, as described in the Supporting Information (Section 5 and Supplementary Figure 7). Briefly, the sequential approach involves the following steps: (i) testing for association between all exons and all SNPs at once, per gene; (ii) testing for association between each exon and all SNPs of genes selected in step (i); and (iii) testing for association between exon–SNP pairs, for exons selected in step (ii). Multiple testing correction is done in this case by controlling the family-wise error rate considering the test tree, as suggested by Meinshausen (2008). Doing this enabled us to compare the number of selected (or not) tests for each exon–SNP pair between the two approaches (Supplementary Table 4). Specifically, out of the total 53,286,710 exon–SNP pairs, 9784 were selected by both approaches, 3259 were selected only by the spliceQTL approach, and 15,850 were selected only by the pairwise testing.

Note that using the spliceQTL test sequentially as described above requires an overall smaller number of tests, compared with a pairwise approach. This is because exon–SNP pairs within genes or exons not selected in the first two steps are not tested. As such, a less severe multiple testing correction ensues, yielding more power to find effects involving multiple exons and multiple SNPs per gene (to the extent that the gene and involved exons are selected). So, as long as the test used can be applied to multiple responses as well as more covariates than samples ($p \gg n$), the sequential approach is preferable to the pairwise approach to find such effects. The gain in power increases with the number of genes and exons not selected as significant in the first two stages. Obviously, the only situation where the two approaches yield the same number of tests is when all genes, as well as all of their exons, are selected as being associated with their corresponding SNP sets.

To illustrate this, we compare here the number of tests required by each approach. We have mentioned that the pairwise approach involves testing for all 53,286,610 exon–SNP pairs. In contrast, the sequential approach started by testing 1376 genes, yielding 34 significant tests. Testing for an association between each exon in these 34 genes and all SNPs required a further 1793 tests. By subsequently selecting only the significant exons, a further 1,617,287 exon–SNP pairs had to be tested. So, the total number of tests required (1,619,114) amounted in this case to just 3% of all 53,286,610 tests required by the pairwise approach.

While considerable overlap exists between results obtained with the two approaches, there are also findings obtained by only one of the approaches. For example, gene ENSG00000007341_12 (Figure 2) is selected by the spliceQTL test, while the linear regression does not select any of the exon–SNP pairs involved. Indeed, Spearman correlations between each exon–SNP pair are mostly between −0.2 and 0.2, so are individually too small to be selected. However, when considered as a set, they clearly display a pattern that is highly unlikely to be due to chance alone, which is why the gene is selected by

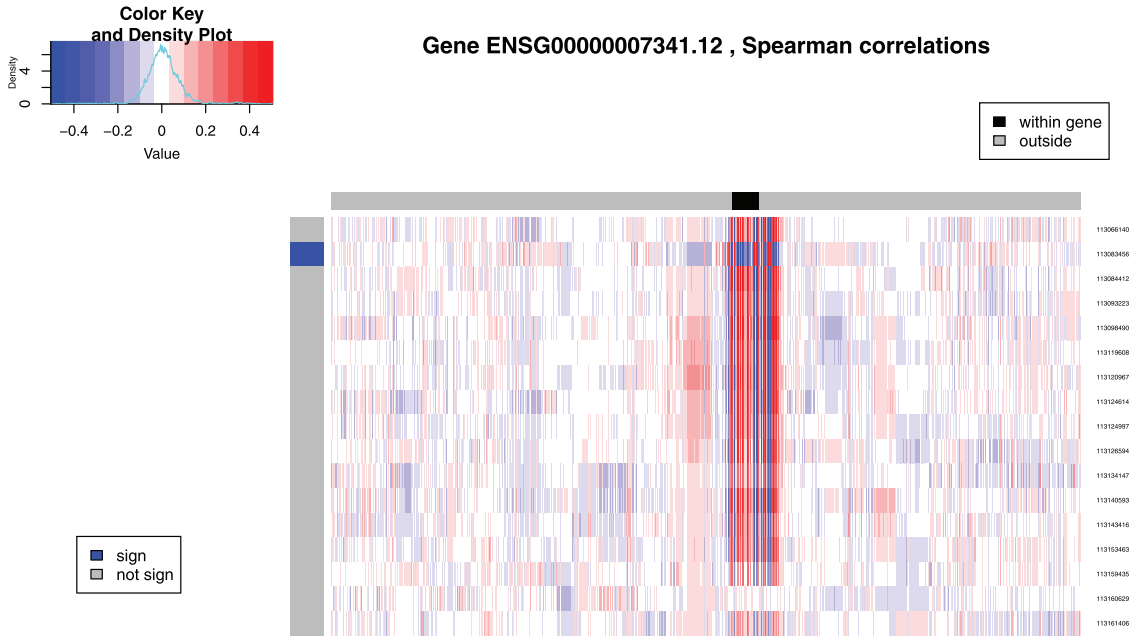


**FIGURE 2** Heatmap representing Spearman correlations between 3961 individual SNPs (columns) and 17 exons (rows) for gene ENSG00000007341_12. The top-left scale bar indicates which color corresponds to which Spearman correlation value. The left-hand side vertical bar indicates which exons are selected at the second step of sequential testing (blue) or not (gray)—see the bottom-left hand-side legend. The top-horizontal bar indicates which SNPs are located within (black) or outside (gray) the gene—see the top-right hand-side legend. This gene was selected by the spliceQTL test to have SNP profiles associated with the exon expression levels as a whole, but none of the individual exon–SNP pairs was selected as significant using the pairwise approach. SNPs in the gene 1 set to display more correlation than those in the gene 2 set

the spliceQTL test. We also notice from the heatmap in Figure 2 that most SNPs located within the gene display a stronger association with (almost all) exons, compared to SNPs located well outside the gene. In fact, only some SNPs just to the "right" of those within the gene display a comparable correlation with exon counts, as evidenced by a subset of columns displaying more intense colors for SNPs just outside the gene. Note that both positive (red tones) as well as negative (blue tones) correlations are observed, as the spliceQTL test is sensitive to their absolute size, regardless of the direction.

In contrast, gene ENSG00000131236_11 (Figure 3) was not selected by the spliceQTL test, while the linear regression selected a few of the exon–SNP pairs involved. Spearman correlations between each exon–SNP pair are again mostly between −0.2 and 0.2, but in a relatively small number of cases they are smaller than −0.4. Their number is too small relative to the total, so those correlations become diluted by the spliceQTL test. However, some of them are individually strong enough to be selected by pairwise testing.

So, we conclude that the spliceQTL test may find associations between exon-level counts and SNPs that are individually small but are observed for a "large enough" subset of exon–SNP pairs. Of course, what is large enough here depends on many aspects, including effect sizes, the correlation between SNPs, and the number of exon–SNP pairs displaying nonzero correlation, relative to the total number of pairs. In addition, the sequential approach requires a smaller number of tests, thus requiring less severe multiple testing correction. The pairwise approach may, however, potentially find exon–SNP associations that are individually large but are not observed for any other pair within the same gene.

## 3.3 | Example 2: Replication of results

In this example, we assess the reproducibility of spliceQTL test results, compared with that obtained using pairwise testing. For this, we will apply both approaches to the GEUVADIS and the LLS data, and per approach check if results are replicated. For each approach, we included in tests only SNPs located within the gene, that is, between the start and end of the gene. Data preprocessing was done as described in Section 2.5. Note that the features (exons, SNPs) measured in the two datasets did not completely overlap. Only overlapping pairs are taken into account. For the same reasons as in

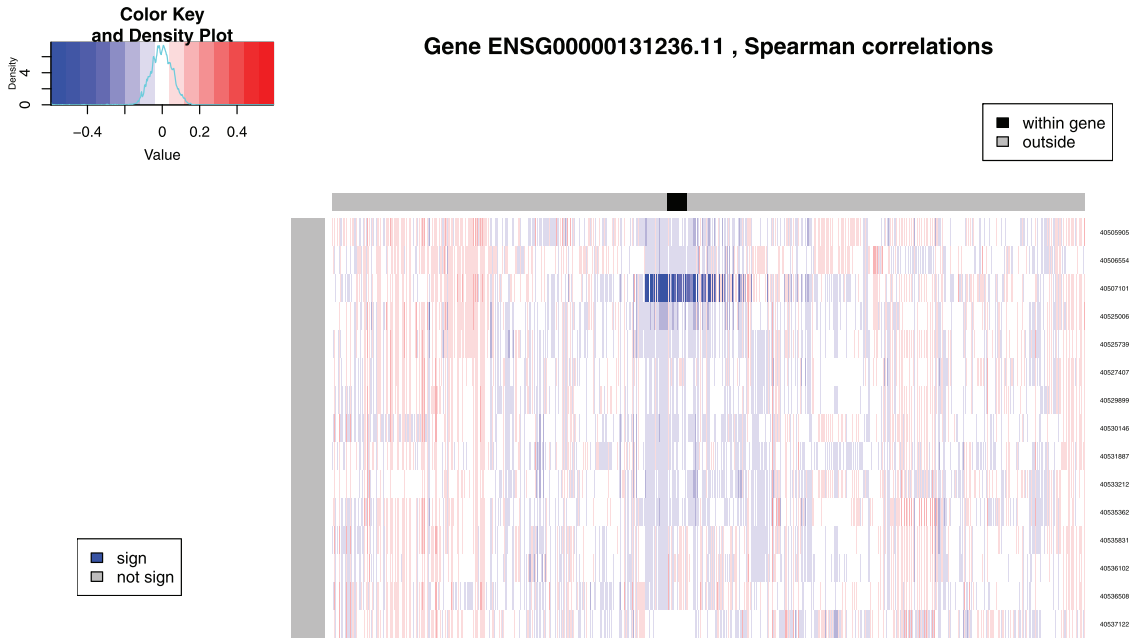


**FIGURE 3** Heatmap representing Spearman correlations between 3697 individual SNPs (columns) and 15 exons (rows) for gene ENSG00000131236_11. The top-left scale bar indicates which color corresponds to which Spearman correlation value. The left-hand side vertical bar indicates which exons are selected at the second step of sequential testing (blue) or not (gray)—see the bottom-left hand-side legend. The top-horizontal bar indicates which SNPs are located within (black) or outside (gray) the gene—see the top-right hand-side legend. This gene was not selected by the spliceQTL test to have SNP profiles associated with the exon expression levels, but some of the individual exon–SNP pairs were found to be significant using the pairwise approach

Example 1, we used here only $S_W$ defined in (5). Results of the two approaches were compared at the gene level, so no sequential testing was used.

Comparison between the spliceQTL test results used the per-gene $p$-values, based upon which we computed: (a) the Pearson correlation between $p$-values on the $-$log scale; (b) the number of genes with the $p$-value under a cutoff, for different cutoff values, without and with multiple testing correction.

Comparisons between the pairwise testing results were done using various summaries at the gene level: (i) via the Pearson correlation between the ($-$log-scale) $p$-values for all exon–SNP pairs for that gene, yielding one correlation value per gene; (ii) computing the proportion of $p$-value pairs per gene valued under a cutoff, for various cutoff values.

The density of the correlations between $p$-values from pairwise testing shows that, while many genes display a positive association between results, many displays even negative associations (Supplementary Figure 8). Indeed, 38% of the genes display relatively low correlations between $-0.2$ and 0.2 (vertical dashed-gray lines in Supplementary Figure 8), half of the genes display correlation below 0.076, 42% below 0, and 22% below $-0.2$. Such low or negative correlations may arise due to effects or noise that are data specific.

In contrast, the spliceQTL results yield a Pearson correlation of 43%, and only 24% of the per-gene correlations of the pairwise testing results had a value at least as large as that (or $\geq 0.43$; indicated by a vertical dashed-black line in Supplementary Figure 8). These arguments suggest that the association between spliceQTL test results is stronger than those between pairwise testing results.

We then examined the scatterplots of the proportions of selected exon–SNP pairs per gene, for three different cutoffs (Supplementary Figure 9). These results show that, for all cutoffs, there is an increase in proportions that are very close to zero for one dataset and much larger in the other. This pattern is behind the negative correlations obtained between gene-specific results. In contrast, the scatterplot between ($-$log) $p$-values of the spliceQTL test displayed no such pattern (Supplementary Figure 10). Note that, for both methods, using the LLS data a larger number of genes are selected: the number of genes with proportions larger than 0.5 with pairwise testing is larger, and the number of genes with small spliceQTL test $p$-values is larger.

The reason why the spliceQTL test results are more often replicated across these two datasets is that it takes multiple covariates (SNPs) and responses (exons) into account, making it more robust to fluctuations that are exon–SNP pair

specific but do not extend to multiple pairs within the same gene. While the pairwise testing approach may identify interesting local effects that the spliceQTL test has low power to find, as it is susceptible to data-specific effects its results are very often not replicated across datasets.

The strong concordance between spliceQTL test results is further illustrated by per-chromosome Manhattan plots of the results of both datasets (Supplementary Figures 11–15): for these, we computed the per-gene FDRs (computed as indicated in Section 2.2), and these yielded virtually the same pattern for the two datasets.

We conclude that the spliceQTL test results are more often replicated across datasets than pairwise testing results. This is likely due to the spliceQTL test that takes into account many variables at once, which averages out data-specific effects.

## 4 | DISCUSSION

We proposed a test based upon the score function that can at once check for the association between expression levels of all of the gene's exons as well as all SNPs considered to be of interest. It has more power to find genes where exon-level expression is associated with SNP genotypes for many pairs of exons and SNPs, compared to when testing one exon–SNP pair at a time. It can be used when the number of SNPs is larger than the number of samples, as it does not require estimating parameters for individual SNPs. It is possible to use the test as part of a three-stage procedure to look for, and better understand, spliceQTL effects, by first considering all exon–SNP pairs mapping to a gene, then looking for the exons that display significant associations with all SNPs, and finally looking for exon–SNP pairs that are significant from among the exons found to be significant.

Our test is designed to find specifically spliceQTL effects that cannot be linked to eQTL effects, and it does this by conditioning on the gene's total expression. So in our approach, we consider exon expression data arising from a multinomial distribution, with the total being equal to the total gene expression. For studying eQTL effects, for example, the global test (Goeman et al., 2004) can be used, using the total expression of each gene as a response and all SNPs of interest as a covariate set.

In our examples, we have considered SNPs with an minor allele frequency (MAF) > 0.05, as these were assumed to represent reliable signals. From a methodological point of view, this restriction is not necessary. The power to find associations between outcome and rare variants may, however, be limited as with most methods. In addition to the MAF, multiple aspects affect the power, including its effect size, correlation structure with neighboring SNPs and, therefore, also the effect size of these neighboring SNPs. Because of the latter, for rare variants the proposed statistical test is expected to be more powerful than procedures where every SNP is tested one by one.

Score-based tests can yield too much power, especially when declaring significant effects that are too small to be biologically relevant. While this is difficult to illustrate and verify due to the high-dimensional nature of the data, one possible reason for this may be that covariates are in fact correlated, while the test is developed under the assumption they are not. Correlated covariates do not invalidate the test's results (see the next paragraph), yet it is undesirable to have increased power to find effects so small as to have little biological relevance. Our method takes into account the covariates' correlation structure, which helps to avoid this issue. We illustrated this in the simulation study, where the results obtained with the extra correction for SNP data correlation yield slightly smaller AUCs for the SNP data with weaker correlation (gene 2) compared with those for the SNP data with stronger correlation (gene 1), using the same effects on the responses. By taking the SNP correlation structure into account, our test thus avoids yielding too many small effects of little biological interest.

The original formulation of a score test for association between a covariate set and a univariate response, used in Le Cessie and van Houwelingen (1995) and Goeman et al. (2004), assumes that random effects are uncorrelated under the null hypothesis. However, when this assumption does not hold the test is not invalidated, but rather it has increased power to find effects in directions where the covariates are highly correlated (Goeman et al., 2006). This differentiates this approach from classic regression models, which generate unreliable results when explanatory variables are highly correlated. In particular, classic regression models do not have unique model parameter estimators if any subset of the explanatory variables is perfectly correlated. The approach used here does not involve inverting the design matrix, and, thus, testing can be done even under collinearity, including in cases when there are more covariates than samples. Here we take this framework further, by fine-tuning the modeling of the covariates' correlation structure.

Goeman and Le Cessie (2006) proposed a test to find associations between a set of covariates (possibly with $p > n$) and a multinomial response. Their test was designed for a univariate response, which for the application at hand would be a variable representing the exon each read maps to. As such, the data would have to be formatted in very large objects, with a response as long as the number of reads per sample, for all samples. While our proposed test statistic $S_I$ can be shown to

yield the same results as the one they proposed, theirs cannot be used for studying spliceQTL due to practical constraints. Chaturvedi et al. (2017) were the first to propose a test for association between a set of covariates (with possibly $p > n$) and a multivariate response vector, for the case where the response follows a multivariate normal distribution. Here we extend this formulation to the multinomial case and allow for taking into account the correlation between covariates explicitly in the test statistic.

Through simulation studies, we showed that the spliceQTL test is particularly suitable for finding effects that are small but that are spread across multiple exon–SNP pairs. This was confirmed by our results using the GEUVADIS/1000Genomes data (Example 1). By applying the test to two independent datasets in Example 2 (GEUVADIS and LLS data), we concluded that the spliceQTL test results are more often replicated across these two datasets, compared with pairwise testing. This is likely due to the spliceQTL test taking multiple covariates (SNPs) and responses (exons) into account. This makes the test more robust to exon–SNP pair-specific fluctuations, which do not extend to multiple pairs within the same gene and therefore are also less likely to be observed for other datasets. In general, approaches that take a small subset of variables into account at a time, such as pairwise testing of each exon–SNP pair individually, are better able to pick up very strong associations observed only for a small subset of exon–SNP pairs, within a gene, as we have shown in Example 1. Such associations need to be very strong so as to be detectable after heavy multiple testing. Driven by just a single or subset of exon–SNP pairs, these are more likely to be data specific and therefore less likely to be replicated across datasets.

Our method is general for finding genes that may display differences in the distribution of reads across exons, associated with a set of covariates. Differences in the distribution of reads across exons can ensue in various situations, with alternative splicing across samples just the most obvious one. In addition, the spliceQTL test can also be applied for covariates other than SNPs. Indeed, it could equally well be applied to studying spliceMeth events, that is, changes in splicing associated with methylation changes. Furthermore, exons corresponding to multiple genes can be considered, for example, to try to find associations between differential exon usage across multiple genes and sets of covariates.

Our test statistic makes use of the assumption that random effects relating to different exons are uncorrelated. However, when cassettes of multiple consecutive exons are included or excluded as a whole, a process which is quite common in practice, random effects relating to different exons may well be correlated. The framework we developed can be extended to cover this situation.

In developing the test statistic and in examples, we have focused on the situation where the exon-level variables are counted, as often happens with RNA-seq data. However, a very similar development can be used to yield a test statistic if exon-level variables are continuous. If $\mathbf{Y}_i$ is a vector of continuous variables, then $\mathbf{Y}_i|\mathbf{r}_i$ may be assumed to follow a multivariate normal distribution. The score test statistic for testing for spliceQTL has the same form as in (4), with the only difference that the matrix $\mathbf{H}$ equals an $n \times n$ matrix filled with $1/n$ at every entry (where $n$ represents the number of samples in the data). Examples where this could be applied to include studies where exon-level expression values correspond to normalized and scaled measurements from either RNA-seq or microarrays, as well as when the response is a vector of image-derived variables, such as the ones yielded by radiomics.

Our spliceQTL test is based on the assumption that SNP-specific effects on exons can be modeled by a linear function between the number of minor alleles and the exon counts. While this is likely to be the case for many genes, in cases where it does not hold our test can still work as it uses a first-order approximation. Furthermore, the simplified illustration of spliceQTL in our Supplementary Figure 1 does not depict the realistic possibility where, when SNP1 has 1 minor allele, two transcripts are observed. Our test would in fact be able to detect such effects, as these would result in different read distributions across exons. Indeed, our test can find evidence of alternative splicing, avoiding the much more difficult problem of transcript reconstruction. It can in fact serve as a basis for transcript reconstruction, for which many methods exist (Steijger et al., 2013).

Monlong et al. (2014) have proposed a test to find associations between alternative splicing ratios and one SNP at a time but not for multiple SNPs. Our method can in fact also be applied to the sort of data they consider (a multivariate response composed of many splicing ratios, instead of exons as considered here), and as such can handle the same problems as well as those involving many SNPs, making it more general.

Some authors have tried to handle this problem by first performing dimension reduction on one or both datasets. This often involves a two-stage procedure, for example, first finding features (genes or exons, depending on the method) associated with alternative splicing and subsequently using those features to examine whether they display association with SNPs applying existing statistical methods. As such methods rely on a group comparison in the first stage, it is difficult to compare them with the spliceQTL test proposed here, which does not require a group comparison, focusing on finding significant association patterns between sets of exons and SNPs.

To conclude, we have proposed a powerful test for spliceQTL effectsthat is able to find even small effects which are spread across multiple exons and/or SNPs. Compared with methods based on pairwise tests, our test results are more likely to be replicated across datasets, as the test takes into account multiple exons and SNPs simultaneously, making it robust to exon–SNP pair-specific effects. These characteristics make it a tool to help in uncovering alternative splicing events associated with QTL effects.

## ACKNOWLEDGMENTS

We are very grateful to J. J. Goeman for discussions over the multivariate multinomial test and N. Chaturvedi for scripts to compute the G2 test statistic.

This study makes use of data generated by the Biobank-based Integrative Omics Study. A full list of the investigators is available from http://www.bbmri.nl/en-gb/activities/rainbow-projects/bios. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated July 9, 2009, and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL).

The LUMC Longevity Study was supported by a grant from the Innovation-Oriented Research Program on Genomics (SenterNovem IGE01014 and IGE05007), the Centre for Medical Systems Biology, and the National Institute for Healthy Ageing (Grant 05040202 and 05060810), all in the framework of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

For the analysis of data presented in Examples 1 and 2, a virtual machine from SURFsara (https://www.surfsara.nl) was used.

## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## AUTHOR CONTRIBUTION

RXM and MAJ proposed the model and determined the functional form of the test statistic, based upon ideas arising from discussions between RXM and PACH. RXM and AR determined the equivalence between the multinomial test statistics for univariate and multivariate responses and ran all computations. All authors read and approved the final manuscript.

## DATA AVAILABILITY STATEMENT

All scripts and some datasets are available from the public repository https://github.com/rxmenezes/spliceQTLsuppMat.git

## OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to the data being not publicly available.

## ORCID

*Renee X. de Menezes* https://orcid.org/0000-0002-1378-2721
*Marianne A. Jonker* https://orcid.org/0000-0003-0134-8482

## REFERENCES

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate—A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, *57*, 289–300.

Black, D. L. (2000). Protein diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell*, *103*(3), 367–370.

Chaturvedi, N., Menezes, R. X., & Goeman, J. J. (2017). A global × global test for testing associations between two large sets of variables. *Biometrical Journal*, *59*, 145–158.

Goeman, J. J., & Le Cessie, S. (2006). A goodness-of-fit test for multinomial logistic regression. *Biometrics*, *62*, 980–985.

Goeman, J. J., van de Geer, S. A., de Kort, F., & van Houwelingen, H. C. (2004). A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*, *20*, 93–99.

Goeman, J. J., van de Geer, S. A., & van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society Series B*, *68*, 477–493.

Lappalainen, T., Sammeth, M., Friedlander, M. R., Hoen, P., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., …, Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*, 506–511.

Le Cessie, S., & van Houwelingen, H. C. (1995). Testing the fit of a regression-model via score tests in random effects models. *Biometrics*, *51*, 600–614.

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using leafcutter. *Nature Genetics*, *50*, 151–158.

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, *95*(2), 265–278.

Modrek, B., & Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, *30*(1), 13–19.

Monlong, J., Calvo, M., Ferreira, P. G., & Guigó, R. (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature Communications*, *5*, 4698.

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., & Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, *32*, 1479–1485.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413–1415.

Pecanka, J., van de Vaart, A. W., & Jonker, M. A. (2019). Modeling association between multivariate correlated outcomes and high-dimensional sparse covariates: The adaptive SVS method. *Journal of Applied Statistics*, *46*(5), 893–913.

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, C. M., & Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, *12*, 315.

Schoenmaker, M., de Craen, A. J. M., de Meijer, P. H. E. M., Beekman, M., Blauw, G. J., Slagboom, P. E., & Westendorp, R. G. J. (2006). Evidence of genetic enrichment for exceptional survival using a family approach: The Leiden Longevity Study. *European Journal of Human Genetics*, *14*, 79–84. https://www.nature.com/articles/5201508

Steijger, T., Abrii, J. F., Engstrom, P. G., Kokocinski, F., The RGASP Consortium, Hubbard, T. J., Guig, R. ó., Harrow, J., & Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods*, *10*, 1177–1184.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., Keeffe, O. S., Haas, S., Vingron, M., Lehrach, H., & Yaspo, M-L. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, *321*(5891), 956–960.

Tian, J., Wang, Z., Mei, S., Yang, N., Yang, Y., Ke, J., Zhu, Y., Gong, Y., Zou, D., Peng, X., Wang, X., Wan, H., Zhong, R., Chang, J., Gong, J., Han, L., & Miao, X. (2019). Cancersplicingqtl: A database for genome-wide identification of splicing QTLs in human cancer. *Nucleic Acids Research*, *47*, D909–D916.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, CB. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*, 470–476.

Zhang, Y., Yang, H. T., Kadash-Edmonson, K., Pan, Y., Pan, Z., Davidson, B. L., & Xing, Y. (2020). Regional variation of splicing QTLs in human brain. *American Journal of Human Genetics*, *107*, 196–210.

Zhernakova, D. V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.-J., Bonder, M. J., van Rooij, J., Verkerk, M., Jhamai, P. M., Moed, M., Kielbasa, S. M., Bot, J., Nooren, I., …, Franke, L. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics*, *49*, 139–145.

McCullagh, P., & Nelder, J. A. (1989) *Generalized Linear Models*. Chapman & Hall, Boca Raton.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** de Menezes, R. X., Rauschenberger, A., BIOS Consortium, 't Hoen, P. A. C., & Jonker, M. A. (2022). A powerful global test for spliceQTL effects. *Biometrical Journal*, 1–14. https://doi.org/10.1002/bimj.202100123