Received ; | Revised ; | Accepted

DOI: xxx/xxxx

Gene Set Analysis for time-to-event outcome with the Generalized Berk–Jones statistic

Laura Villain[*1,2,4] | Thomas Ferté[1,2,4,1] | Rodolphe Thiébaut[1,2,4,1] | Boris P. Hejblum[†1,2,4]

[1]Univ. Bordeaux, Inserm Bordeaux Population Health Research Center, SISTM team, UMR 1219, 146 rue Léo Saignat, Bordeaux France
[2]INRIA Bordeaux Sud Ouest, SISTM team Talence, France
[3]Vaccine Research Institute, VRI, Hôpital Henri Mondor, 51 avenue du Maréchal de Lattre de Tassigny, Créteil, France
[4]CHU Pellegrin, Groupe Hospitalier Pellegrin, Place Amélie Raba Léon, Bordeaux, France

Correspondence
[*]Laura Villain Email: laura.villain@u-bordeaux.fr
[†]Boris Hejblum Email: boris.hejblum@u-bordeaux.fr

Present Address
INSERM U1219, ISPED, Univ. Bordeaux, 146 Rue Leo Saignat, 33076 Bordeaux, France

Summary

Gene Set analysis allows to evaluate the impact of groups of genes on an outcome of interest, such as the occurrence of a disease. Through the definition of the gene sets, gene set analysis takes into account biological knowledge and makes it easier to interpret the results, while improving the statistical power compared to a gene-wise analysis. In the time-to-event context, few methods exist, but most of them do not take into account the correlation that occurs inside a gene set, which can be strong. As the Generalized Berk-Jones statistics showed great consistency and includes the correlation inside the test statistic, we adapted this method to the time-to-event context by using a Cox model. We compared our approach to other methods based on the Cox model, and showed that the Generalize Berk-Jones statistic offers great adaptability, meaning that it can be used in all kinds of data structures. We applied the different methods to two different contexts: Gliomas and Breast cancer. In terms of statistical power, we did offer similar results to the other Cox model methods, but with greater accuracy. In the breast cancer framework, we showed better statistical power than methods based on Kernel Machine score.

KEYWORDS:
RNAseq, Gene Set Analysis, Generalized Berk-Jones, Glioma, Breast cancer

1 | INTRODUCTION

The analysis of the whole transcriptome thanks to RNA-seq[1] usually relies on differential expression analysis approaches. Statistical methods such as `edgeR`[2], `DESeq2`[3], `limma-voom`[4], or `dearseq`[5] allow comparing gene expression between groups of patients or within patients over time to identify the genes involved in diseases[6] or committed by an intervention in a trial (e.g. a vaccine)[7]. Yet, gene-wise analyzes have several limitations: i) the high-dimensionality of gene expression data (tens of thousands of genes measured per sample) might lead to either no gene being detected differentially expressed after multiple testing correction, or on the contrary to a very large list of genes difficult to interpret biologically; ii) many genes interact within biological pathways, but with potentially only small individual changes, thus leading to all genes within a pathway to be non-significant after multiple-testing correction. To tackle those issues and avoid missing important biological links with the output of interest, gene Set Analysis leverages predefined sets of genes (available in databases such as MSigDB[8], KEGG[9] or Gene Ontology[10] for instance) in order to identify groups of genes differentially expressed rather than single genes. Both the diminution of the number of statistical tests to perform and the strength of a coordinated signal within a gene set increase statistical power. Numerous methods have been proposed to analyze gene set expression[11], for example GSA[12], Gene Set Enrichment Analysis

(GSEA)[13], Time-course Gene Set Analysis (TcGSA)[14], Generalized Higher Criticism (GHC)[15], dearseq[5] or the Generalized Berk-Jones (GBJ) statistic[16].

In the time-to-event context, some methods for gene expression analysis are able to tackle high dimensional data such as survival random forest[17] or penalized Cox regression[18], but only a handful methods are available to perform gene set analysis. Based on Kernel Machine, Cai *et al.*[19] proposed a score test for survival gene set analysis, that Neykov *et al.*[20] later adapted to take into account competing risk. Besides, three additional methods are building on the Cox model to perform survival gene set analysis, using different test statistics and relying on permutations to compute the associated p-values: i) the `global test`[21], ii) the `Wald test`[22], and iii) the `global boost test`[23]. In a review performed by Lee *et al.*[24], all three tests outperformed GSEA in term of statistical power. However, the `global test` rely on the strong hypothesis that all regression coefficients are sampled from the same distribution, and none of the compared methods takes into account the correlation structures between the genes in the pathway.

Gaynor *et al.*[16] compared the GBJ statistic with GSA, GSEA, and GHC to identify gene sets differentially expressed between two tumor grade, and they concluded that their proposed GBJ method was more consistent. The GBJ statistic has the advantage of not requiring any distributional assumption on the count data or the regression coefficients, while also taking into account the correlation structure between the genes within the same set. We propose to adapt the GBJ statistic in the time-to-event context using a statistic derived from the Cox model. We denote our method by `sGBJ` for "survival Generalized Berk–Jones". Section 2 describes our new method `sGBJ`, then Section 3 evaluates its performance in a simulation study, similarly to Lee *et al.*[24]. Section 4 presents two real data analyses, with brain cancer data and breast cancer data respectively. Finally Section 5 discusses our results and their limits. The sGBJ method is implemented in a R package, available on github at https://github.com/lauravillain/sGBJ

2 | METHOD

2.1 | Generalized Berk–Jones set-based testing method

The GBJ statistic can be used in a set based testing procedure to determine the effect of a set of genes on a given clinical outcome (e.g: tumor grade). Introduced by Sun and Lin[25], it was used in the context of Genome-Wide Association Studies (GWAS)[26] and its consistency was evaluated for identifying pathways whose expression is associated with either low or high grade of breast cancer[16].

Here is a 4 steps description of the GBJ approach developed by Gaynor *et al.*[16]:

1. Model and null hypothesis specification for a given gene set of d genes. Sun and Lin[25] study the association between the outcome of interest Y_i of patient i with the vector of gene expression of d genes $\mathbf{G}_i$ in a generalized linear model, with $\boldsymbol{\mu}_\mathbf{i}$ the conditional mean of Y_i and $\mathbf{X}_i$ the other covariates:

$$g(\boldsymbol{\mu}_\mathbf{i}) = \mathbf{X}_i^\intercal \boldsymbol{\alpha} + \mathbf{G}_i^\intercal \boldsymbol{\beta} \tag{1}$$

2. For each gene j of a given gene set with $j = \{1, ..., d\}$, a score value Z_j is computed (e.g. from the score test statistic as in Sun and Lin[25]) that must verify $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ under the null hypothesis.

3. The GBJ statistic[26] is computed for the whole gene set based on the Z_j values of each gene using the threshold function S, defined as $S(t) = \sum_{j=1}^{d} \mathbb{1}\left\{|Z_j| \geq t\right\}$ and computing the number of genes of the gene set that have an absolute value of their score Z_j higher than a limit t.

4. Once the GBJ statistic g is computed, its associated p-values is determined by inverting g with a root-finding algorithm to determine the boundaries points b_j, i.e. the limit value of the j^{th} greatest absolute value of $\mathbf{Z}$ (if this bound is higher than b_j, then the observed statistic should be greater than g).

2.2 | Time-to-event context

The GBJ method as defined by Sun and Lin[25] can be applied to any generalized linear model, but not for time-to-event data analysis. From the third step onward, the procedure relies on a score statistic calculated for each gene within the gene set, with a multivariate normal distribution assumption on this score centered around zero under the null hypothesis (i.e. no association

between gene expression and the dependent variable $\boldsymbol{Y}$). We propose the survival GBJ method (`sGBJ`) which uses a score adapted to time-to-event to compute the GBJ statistic.

1. Model

To deal with time-to-event outcomes, we rely on a Cox proportional hazards regression model:

$$\lambda(t, \mathbf{G}_i) = \lambda_0(t)\exp(\mathbf{G}_i^\intercal \boldsymbol{\beta}), \tag{2}$$

with $\mathbf{G}_i$ the vector of gene expressions of patient i. The null hypothesis is that there is no association between patient survival and the expression of genes within the set: $H_0 : \boldsymbol{\beta} = \mathbf{0}_{d\text{x}1}$.

2. Z score

For each gene j of the gene set, we compute the value Z_j being the square root of the Wald statistic[27]:

$$Z_j = \sqrt{W_j} = \frac{\hat{\beta}_j - \beta_0}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}. \tag{3}$$

This gives us a vector $\mathbf{Z}$ that follows $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ under the null hypothesis H_0.

$\boldsymbol{\Sigma}$ is estimated through permutations of the original survival observations: for each permutation $p \in \{1, \dots, P\}$, the Z^*_{jp} value for each gene j gives us P $\mathbf{Z}^*_p$ vectors of length d, allowing us to estimate the matrix $\hat{\boldsymbol{\Sigma}}$.

3. GBJ statistic

The GBJ statistic is computed according to the formula from Sun and Lin[25]:

$$GBJ = \max_{\{1 \geq j \geq \frac{d}{2}\}} \log \left[\frac{\Pr\left\{ \mathrm{S}(|\mathrm{Z}|_{\mathrm{d-j+1}}) = \mathrm{j} | \mathrm{E}(\mathbf{Z}) = \hat{\mu}_{\mathrm{jd}}.\mathbf{J}_{\mathrm{d}}, \mathrm{cov}(\mathbf{Z}) = \hat{\boldsymbol{\Sigma}} \right\}}{\Pr\left\{ \mathrm{S}(|\mathrm{Z}|_{\mathrm{d-j+1}}) = \mathrm{j} | \mathrm{E}(\mathbf{Z}) = 0.\mathbf{J}_{\mathrm{d}}, \mathrm{cov}(\mathbf{Z}) = \hat{\boldsymbol{\Sigma}} \right\}} \right] \mathrm{x}\mathbb{1}\left\{ 2\phi(|\mathrm{Z}|_{\mathrm{d-j+1}}) < \frac{\mathrm{j}}{\mathrm{d}} \right\}, \tag{4}$$

with $\mathbf{J}_{\mathrm{d}}^{\mathrm{T}} = (1, 1, ..., 1)_{\mathrm{dx1}}$, ϕ the survival function of a standard normal random variable, and $\hat{\mu}_{jd} > 0$ solving the following equation:

$$\frac{j}{d} = 1 - \left\{ \phi(|Z|_{d-j+1} - \hat{\mu}_{jd}) - \phi(-|Z|_{d-j+1} - \hat{\mu}_{jd}) \right\}. \tag{5}$$

The GBJ statistic is calculated only on the higher half values of $|\mathbf{Z}|$, and can, as Sun *et al.*[26] explained it, be represented as the maximum of a set of likelihood ratios on $S(t)$.

4. p-value computation

The p-value can be formulated as

$$\Pr(\mathrm{G_d} \geq \mathrm{g}) = 1 - \Pr\left\{ \forall \mathrm{j} = 1, .., \mathrm{d} : |\mathrm{Z_j}| \leq \mathrm{b_j} | \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \right\}, \tag{6}$$

and the associated rejection region is then

$$\Pr\left\{ \forall \mathrm{j} : |\mathrm{Z_j}| \leq \mathrm{b_j} | \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \right\} = \Pr\left\{ \forall \mathrm{j} : \mathrm{S(b_j)} \leq (\mathrm{d} - \mathrm{j}) | \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \right\}. \tag{7}$$

3 | SIMULATIONS AND COMPARISON

3.1 | Simulation method

`sGBJ` does not require any distributional assumption on the gene expression matrix, thus we simulated this matrix for 50 observations following a multinormal distribution $\mathcal{N}(\mathbf{0}, \mathbf{C})$, with $\mathbf{C}$ being the genes variance-covariance matrix. Following a similar procedure than what Lee *et al.*[24] proposed in their comparison of survival gene sets analysis, we simulated a gene set of 50 genes, among which a proportion p_g of genes were significantly associated with survival. p_g was either 0.05, 0.1, or 0.2 depending on the scenario. We simulated a variance of $C_j j = 0.2$ for each gene j. We generated three correlations scenarios:

- Case (I): no covariance between the genes: $C_{jk} = C_{kj} = 0$
- Case (II): covariance of $0.4 \times \rho_{jk}$ with $\rho_{jk} \sim \mathcal{N}(0.4, 0.1^2)$ between the significant genes j and k, and no correlation otherwise;

- Case (III): covariance of $0.4 \times \rho_{jk}$ with $\rho_{jk} \sim \mathcal{N}(0, 0.01^2)$ between all gene pairs j and k.

Survival times of the 50 observations were then generated using a Cox model featuring an effect β for the significant genes. We again studied three effect types:

- Type A: $\beta_j \sim \mathcal{N}(0, 0.5^2)$ for each significant genes
- Type B: for half of the significant genes $\beta_j \sim \mathcal{N}(0.1, 0.5^2)$, and for the other half $\beta_k \sim \mathcal{N}(-0.1, 0.5^2)$
- Type C: for half of the significant genes $\beta_j \sim \mathcal{N}(0.1, 0.25^2)$, and for the other half $\beta_k \sim \mathcal{N}(-0.1, 1^2)$

Finally we considered potential censoring. Censoring times were generated following an exponential distribution for 30% of the observations. For each simulation scenario, 2,000 independent Monte-Carlo repetitions were generated, allowing us to compute the statistical power for each case.

`sGBJ` was compared to three state-of-the-art methods: i) the `global test`[21], ii) the `Wald test`[22], and iii) the `global boost test`[23], that are also based on statistics derived from the Cox model. The `global test` relies on the hypothesis that all β_j are sampled from the same normal distribution centered in 0 with a common variance σ^2, so the null hypothesis can be reduced to testing $\sigma^2 = 0$. The `Wald test` uses the sum of squares of the Wald statistics from a Cox model apply individually to each gene within a gene set, possibly adjusted to other covariates. The `global boost test` combines the Cox model with a boosting algorithm to assess the additional predictive value of gene expression within a gene set. Each method relies on a different test statistic, and the corresponding p-value is evaluated by permutations for all three methods.

3.2 | Results

All methods control the type-I error adequately in all simulation scenarios (see Supplementary Figure 1). Compared to the three state-of-the-art methods, `sGBJ` exhibits the highest statistical power in almost all simulation scenarios (see Figure 1). In the rare exceptions, e.g. in type A - case (I), the difference between the best performing method and `sGBJ`'s statistical power is negligible. In contrast, the three other methods are less consistent in showing good performance (e.g. while the `global boost test` performs well in several cases, its statistical power significantly drops for the correlation structure in case (II); conversely the `Wald test` and the `global test` performs well in case (II), but show poor performance in type C - case (I)). `sGBJ` is thus a very competitive alternative to the existing state-of-the-art. In particular, in presence of effect heterogeneity within the gene set (e.g. type C scenario where both the mean and variance of the significant genes effect β_i can vary), `sGBJ` performs better than the competing methods (especially the `global test` which hypothesizes that all significant β_i are sampled from the same normal distribution with mean 0). Furthermore, `sGBJ` is almost always the highest in term of statistical power when the percentage of significant genes is low, making it the method of choice for detecting weak signals.

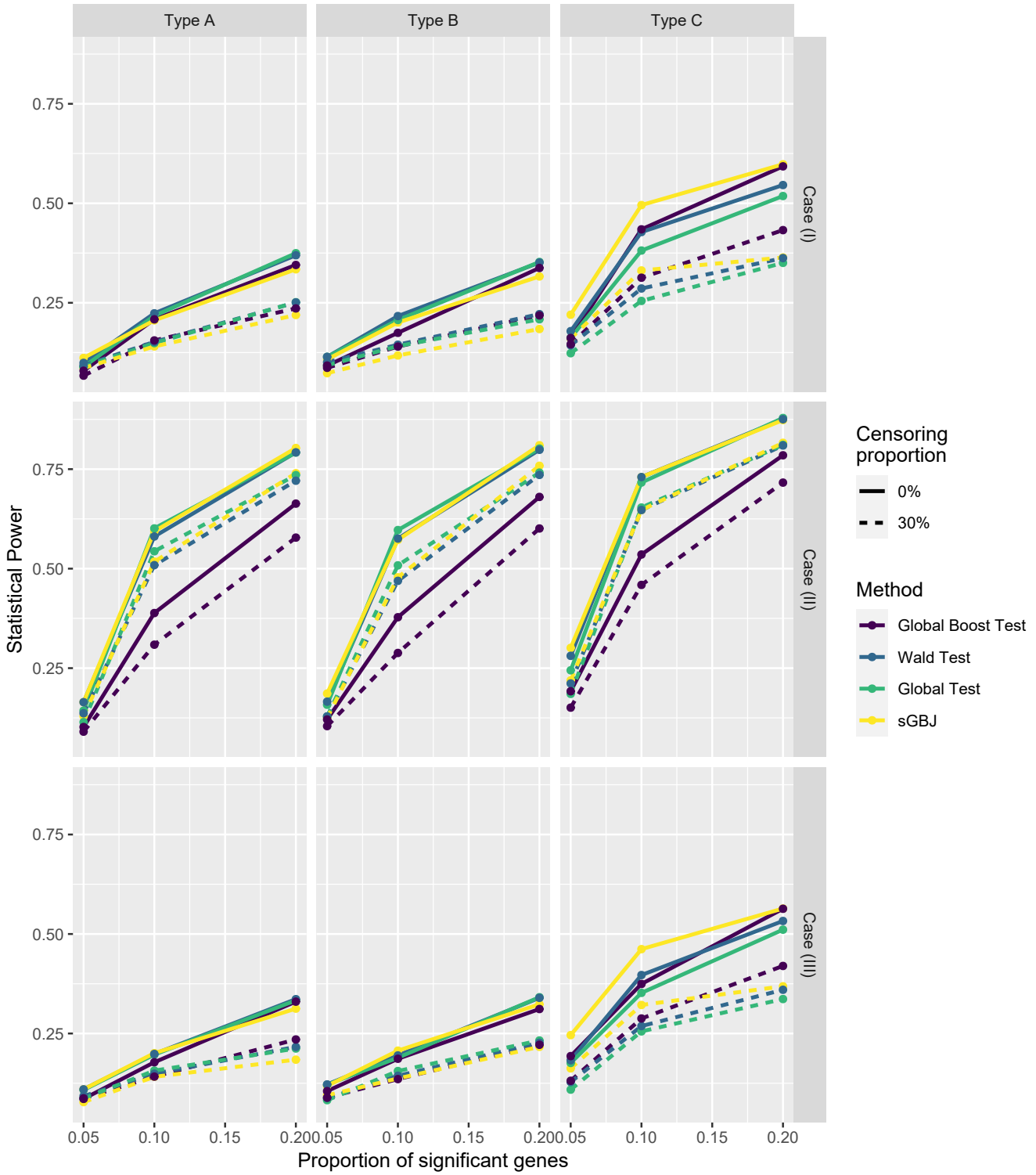

FIGURE 1 The statistical power of the four methods (`sGBJ`, `Global boost test`, `Wald test` and `Global test`) are compared over the six different combinations: three Cases of correlation (I, II and III) and three Types of effects (A, B and C, see Section 3.1 for more details). Each method is represented by a color, with the statistical power computed for different proportions of significant genes. Dotted lines represent a censoring fraction of 0.3; full lines, a censoring fraction of 0.

4 | REAL DATA APPLICATIONS

4.1 | Survival analysis in glioma subtypes using the Rembrant database

Glioma represent over 80% of malignant brain tumor, and are associated with poor survival with less than 5% of survival at 5 years for the glioblastoma (the most common type of glioma)[28]. Glioma can be classified in different types according to the World Health Organization (WHO), with mostly histological and molecular alterations criteria (a first classification in 2007 revised in 2016). The severity can increase from grade I to grade IV, being mostly the glioblastoma[29]. The high genetic heterogeneity of glioma, and the poor response to targeted treatment[30,31] with a high rate of recurrence, makes it hard to treat patients with glioma. To better understand the mechanisms underlying glioma, researchers are actively investigating the molecular and genetic processes linked with the different types of gliomas and patient survival or risk of recurrence.

Rembrandt is one of the largest public databases on Glioma, featuring clinical and genomic data on 671 patients with any types of gliomas (based on the 2007 WHO classification) collected by 14 institutions, as a bioinformatic platform for brain cancer research. They are accessible on the Georgetown University's G-DOC System[32,33], and described by Madhavan *et al.*[34]. For our application, we compared the patients with two of the main glioma types: i) astrocytoma, and ii) oligodendroglioma. We included the patients from Rembrant for whom both tumor gene expression (as measured by micro-array) and overall survival follow-up was available, with 23% of patients still alive at the end of the study. Glioblastoma patients were excluded to avoid too large genetic variation across compared glioma types. Table 1 presents the different characteristics of the total 154 patients included, and Figure 2 displays the corresponding Kaplan-Meier curves stratified on the glioma type. We observe that astrocytoma and oligodendroglioma patients present similar survival curves.

We studied the association between the patient risk of recurrence and pathway gene expression, overall and according to the two glioma types (astrocytoma and oligodendroglioma). We investigated two pathway collections from MSigDB[8]: i) the C8 collection, that includes 671 gene sets related to cellular type, and ii) the C6 collection that contains 189 gene sets related to cellular pathways often found to be dis-regulated in oncogenic studies. We used the following hazard equation to link the recurrence time with gene expression within a gene set of interest:

$$\lambda\left(t|\mathbf{G}_i, \mathbf{Z}_i\right) = \lambda_0(t)\, \exp\left(\mathbf{G}_i^\intercal \boldsymbol{\beta} + \mathbf{Z}_i^\intercal \boldsymbol{\gamma}\right), \tag{8}$$

with $\lambda_0(t)$ the baseline hazard function, G_{ij} the gene expression of the gene j for the patient i, β_j the associated effect, and Z_{il} the values of the covariates l (i.e. age and sex) with γ_l the associated effect. The null hypothesis tested by the method is that all β_j are zeros within a gene set.

While no pathway is identified as significantly associated with recurrence risk in oligodendroglioma patients, most pathways were found significant in the astrocytoma patients (98% of the oncogenic pathways and 92% of the cell type pathways), as well as in all patients without stratification (after Benjamini-Hochberg for multiple testing correction on the number of gene sets tested[35]). All methods yield similar results in the number of significant gene sets (cf Figure 3). However the accuracy of `sGBJ` is greater compared to the three competing methods which all compute p-values by permutation. We performed 1,000 permutations (the default for the `global boost test` implementation in R). This limited number of permutations (due to computation time considerations) has no impact for the oligodendroglioma patients (as p-values are well above 10^{-3}), and only little impact for the astrocytoma patients. However, this highlights the numerical precision shortcomings of the existing methods. Indeed, especially in multiple testing settings, it is critical to correctly and precisely estimate p-values[36]. As there were fewer patients in the oligodendroglioma group than the astrocytoma group (46 patients for oligodendroglioma, 108 for astrocytoma), statistical power might not have been sufficient to identify significant gene sets. Similarly to Gaynor *et al.*[16], this lead us to investigate the top pathways rather than all the significant pathways (see Table 2). For the astrocytoma and all patients, we can see that a high number of identified oncogenic pathways are well known to be linked specifically with Glioma, such as E2F1

	Astrocytoma (108)	Oligodendroglioma (46)
Male	70 (65 %)	23 (50%)
Death	89 (82 %)	37 (81 %)
Median follow-up period in years (min ; max)	2.7 (0.02 ; 17.8)	2.0 (0.08 ; 20.9)

TABLE 1 Characteristics of Astrocytoma and Oligodendroglioma patients included from the Rembrant database

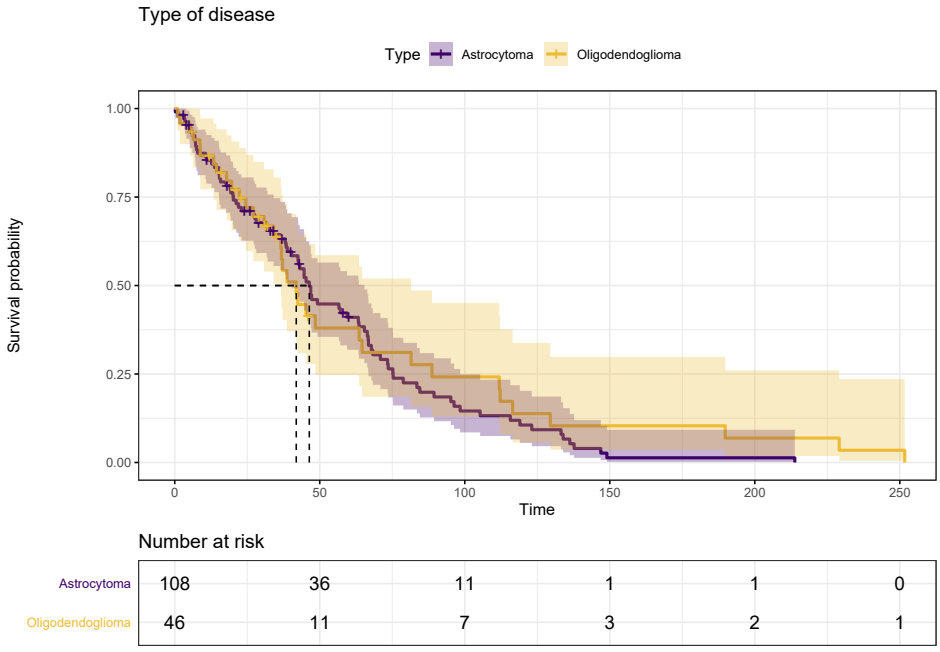

FIGURE 2 Kaplan-Meier curves of Rembrandt patients, stratified by type of glioma. Dotted lines represent the time to 50 % survival.

pathway[37], P53 pathway[38], and so are some of the identified cell type pathways[39 40]. Even in the oligodendroglioma group, some of the pathways in the top ten identified are also known to be specifically linked with glioma, such as the pathways linked with the KRAS gene[41], or pathways linked with dedifferentiation of neurons and astrocytes[42], among others[43 44]. Thus, identified pathways are consistent with other studies findings.

4.2 | Progression free survival in breast cancer

Breast cancer is the most common cancer, with a number of global death predicted to reach 11 million per year by 2030[45]. Survival is highly impacted by the presence of metastasis[46], which justifies the search for genes or pathways associated with either death or metastasis.

We analyzed the breast cancer data published by Van de Vivjier *et al.*[47], featuring both gene expression and clinical data for 260 patients having breast cancer. Gene expression was measured using microarrays and the median follow-up was 7.1 years. Two events of interest were recorded: the apparition of metastasis and the death. We studied the metastasis free survival, i.e a composite outcome considering time to either metastasis or death, and its potential association with specific gene sets or pathways. Table 3 presents the different characteristics of the patients included, and Figure 4 displays the corresponding Kaplan-Meier curves stratified on cancer severity.

We studied the association between the metastasis free survival and gene set expression adjusted on age[48] and without stratifying on severity grade using the following Cox model:

$$\lambda(t, \mathbf{G}_i) = \lambda_0(t)\exp(\mathbf{G}_i^{\intercal}\boldsymbol{\beta} + \gamma \mathrm{Age}_i), \tag{9}$$

with $\lambda_0(t)$ the baseline hazard function, G_{ij} the gene expression of the gene j for the patient i and β_j the associated effect. The null hypothesis tested by the method is that all β are nulls. We investigated 639 of pathways from the KEGG database[49], and also a manually curated subset of 70 pathways from this list selected by Cai *et al.*[19] for their known association with breast cancer patient survival, and also analyzed in Neykov *et al.*[20] to study competing risk between metastasis and death.

Among the manually curated selection of 70 pathways by Cai *et al.*[19], 52 pathways (i.e. 75%) were significantly associated with metastasis free survival according to `sGBJ`, after Benjamini-Hochberg correction (see Supplementary Figure 2). In their initial work, Cai *et al.* proposed a new method based on a Kernel Machine score relying on a flexible framework that allows non linear effects of the genes on the survival[19], but were only capable of identifying 23 out of 70 pathways among the pre-selection

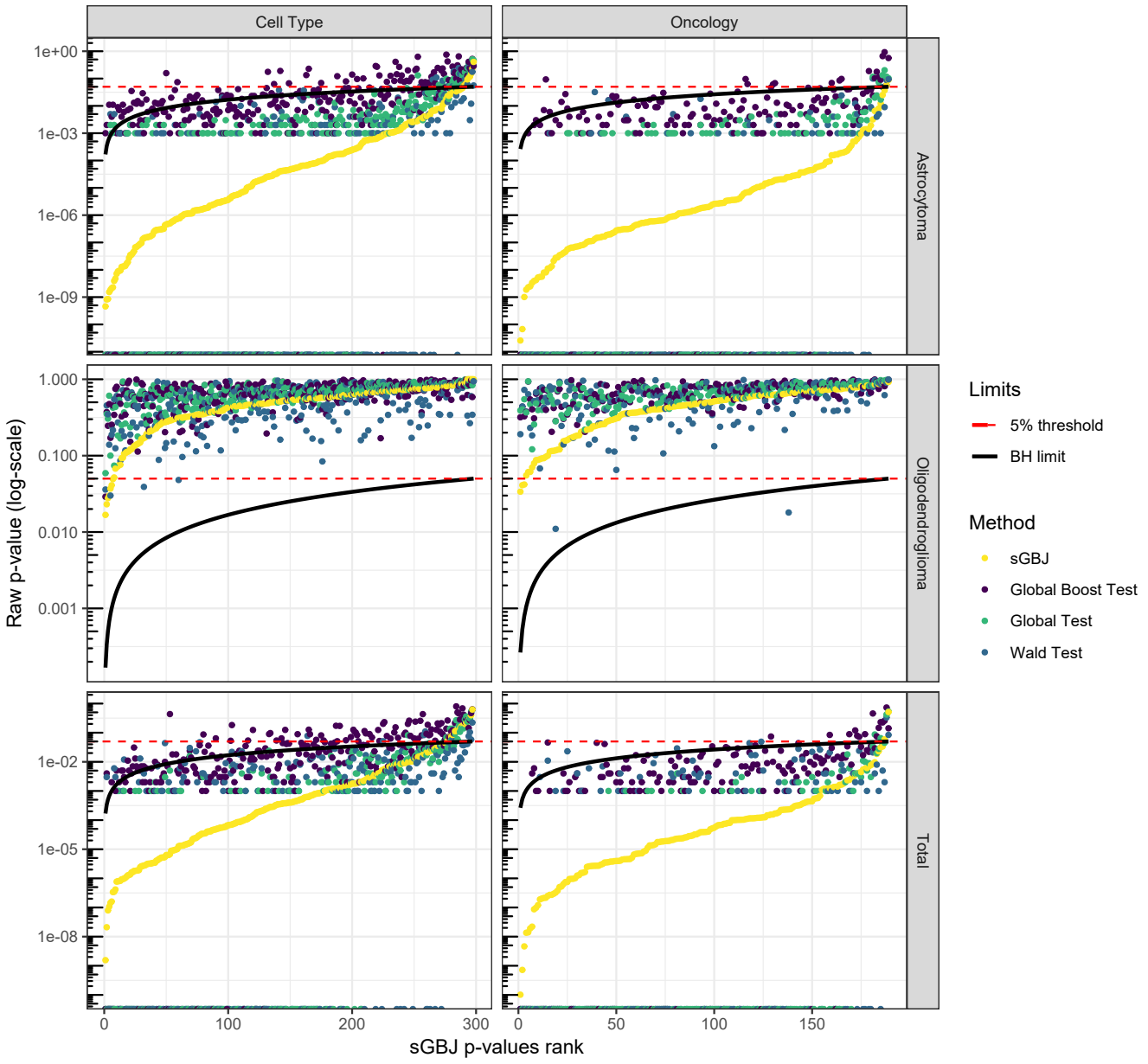

FIGURE 3 Raw p-values in function of the ordered ranks of `sGBJ` for the 4 methods (`sGBJ`, `global boost test`, `Wald test` and `global test`), with the 5% threshold and the Benjamini Hochberg limit, computed for astrocytoma, oligodendroglioma and all patients. *Nota Bene*: The Benjamini Hochberg limit only applies for the `sGBJ` method, as the ranks are computed for `sGBJ` only.

that were known to be associated with breast cancer survival. Among the top ten pathways identified from the whole 639 KEGG gene sets and reported in Table 4 , most have a link with cancer in general or directly with breast cancer: "G1 to S cell cycle reactome" and "glycosis and glucogneogenesis" have been reported to be associated with breast cancer development[50,51], while observations shown that the "pyrimidine metabolism" and the "inositol metobolism" might be involved in cancer[52,53].

	Cell type		Oncogenic	
	Pathway	p-value	Pathway	p-value
Total	hay bone marrow nk cells	1.6×10^{-9}	E2F1 UP.V1 UP	1.9×10^{-8}
	hay bone marrow naive t cell	2.1×10^{-8}	AKT UP.V1 UP	6.8×10^{-8}
	hu fetal retina rpe	7.9×10^{-8}	P53 DN.V1 DN	2.9×10^{-7}
	cui developing heart c6 epicardial cell	1.1×10^{-7}	CYCLIN D1 KE .V1 UP	5.3×10^{-7}
	cui developing heart valvar endothelial cell	1.4×10^{-7}	ESC V6.5 UP LATE.V1 DN	5.3×10^{-7}
	muraro pancreas mesenchymal stromal cell	1.6×10^{-7}	IL15 UP.V1 DN	5.9×10^{-7}
	manno midbrain neurotypes hrgl1	3.4×10^{-7}	E2F1 UP.V1 DN	5.9×10^{-7}
	manno midbrain neurotypes hperic	3.5×10^{-7}	TBK1.DF UP	2.1×10^{-6}
	la kidney c19 collecting duct intercalated cells type a medulla	4.1×10^{-7}	IL21 UP.V1 UP	2.1×10^{-6}
	hay bone marrow early erythroblast	7.7×10^{-7}	TGFB UP.V1 DN	2.2×10^{-6}
	Pathway	p-value	Pathway	p-value
Astrocytoma	cui developing heart c7 mast cell	8.4×10^{-8}	E2F1 UP.V1 UP	4.8×10^{-9}
	hay bone marrow nk cells	8.4×10^{-8}	TGFB UP.V1 DN	6.4×10^{-9}
	hu fetal retina rpe	8.4×10^{-8}	PTEN DN.V1 UP	6.4×10^{-9}
	zhong pfc major types microglia	9.7×10^{-8}	IL15 UP.V1 DN	7.8×10^{-8}
	aizarani liver c14 hepatocytes 2	9.7×10^{-8}	ESC V6.5 UP LATE.V1 DN	7.8×10^{-8}
	la kidney c17 collecting system pcs stressed dissoc subset	9.7×10^{-8}	CSR EARLY UP.V1 UP	7.8×10^{-8}
	hu fetal retina fibroblast	9.7×10^{-8}	RB DN.V1 UP	8.8×10^{-8}
	aizarani liver c11 hepatocytes 1	1.5×10^{-7}	CYCLIN D1 KE .V1 UP	8.8×10^{-8}
	muraro pancreas mesenchymal stromal cell	1.6×10^{-7}	IL21 UP.V1 UP	8.8×10^{-8}
	manno midbrain neurotypes hmgl	2.1×10^{-7}	GCNP SHH UP LATE.V1 DN	8.8×10^{-8}
	Pathway	p-value	Pathway	p-value
Oligodendroglioma	zhong pfc c1 astrocyte	0.86	CAHOY ASTROGLIAL	0.85
	zhong pfc c2 thy1 pos opc	0.86	KRAS.50 UP.V1 UP	0.85
	zheng cord blood c2 putative basophil eosinophil mast cell progenitor	0.86	CSR LATE UP.V1 DN	0.85
	durante adult olfactory neuroepithelium vascular smooth muscle cells	0.86	KRAS.DF.V1 UP	0.85
	gao stomach 24w c2 tff2pos multipotent progenitor	0.86	CSR EARLY UP.V1 DN	0.85
	cui developing heart c6 epicardial cell	0.86	RELA DN.V1 UP	0.85
	durante adult olfactory neuroepithelium respiratory ciliated cells	0.86	GLI1 UP.V1 DN	0.85
	gao large intestine 24w c7 goblet progenitor	0.86	TBK1.DN.48HRS UP	0.85
	cui developing heart c5 valvar cell	0.86	LTE2 UP.V1 DN	0.85
	manno midbrain neurotypes hnbm	0.86	MEK UP.V1 DN	0.85

TABLE 2 Top ten pathways for the Cell types (C8 collection) and Oncogenic signature (C6 collection), with the astrocytoma, oligodendroglioma and all patients

Mean age in years (sd)	44.2 (5.4)
Event (metastasis or death)	83 (32 %)
Grade	
1	70 (28 %)
2	91 (35 %)
3	99 (38 %)
Median follow-up period in years (min ; max)	7.1 (0.05 ; 18.3)

TABLE 3 Characteristics of the 260 patients included in our metastasis free survival analysis of data from Van de Vivjier *et al.*[47]

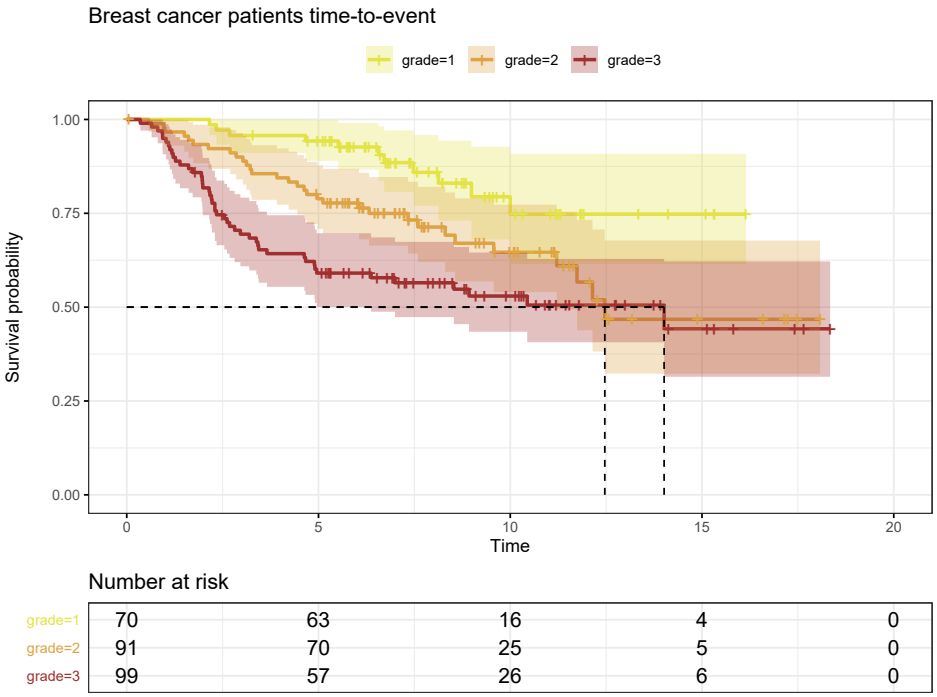

FIGURE 4 Kaplan-Meier curves of Breast cancer patients, stratified on the grade. Dotted lines represent the time to 50 % survival.

Selected pathways		Total pathways	
Pathway	p-value	Pathway	p-value
hsa04115 p53 signaling pathway	$4.3 \times 10{-}6$	hsa04110 cell cycle	$1.6 \times 10{-}9$
hsa00970 aminoacyl trna biosynthesis	$4.3 \times 10{-}6$	g1 to s cell cycle reactome	$1.6 \times 10{-}9$
hsa03030 dna polymerase	$1.4 \times 10{-}5$	pyrimidine metabolism	$9.1 \times 10{-}9$
hsa04010 mapk signaling pathway	$2.5 \times 10{-}6$	cell cycle kegg	$2.1 \times 10{-}8$
breast cancer estrogen signaling	$2.5 \times 10{-}5$	hsa00240 pyrimidine metabolism	$3.1 \times 10{-}8$
hsa00790 folate biosynthesis	$2.5 \times 10{-}5$	dna replication reactome	$5.0 \times 10{-}8$
tnfr1pathway	$4.5 \times 10{-}5$	glycolysis and gluconeogenesis	$2.8 \times 10{-}7$
mapkpathway	$4.5 \times 10{-}5$	inositol metabolism	$6.9 \times 10{-}7$
p53pathway	$4.5 \times 10{-}5$	hsa00010 glycolysis and gluconeogenesis	$1.1 \times 10{-}6$
erkpathway	$9.5 \times 10{-}5$	aminoacyl trna biosynthesis	$1.1 \times 10{-}6$

TABLE 4 Top ten pathways for the total set of 639 pathways and the selected 70 pathways, with their p-values after correction.

5 | DISCUSSION

Only few gene set analysis methods are currently suited to analyze time-to-event data. The current state-of-the-art methods rely on strong assumptions and do not account for the correlation structure between genes, that might be strong in some cases. In this work, we extend the GBJ test for gene set analysis in survival analysis. `sGBJ` relies on Cox model estimations, thus it relies on the proportional hazards hypothesis like the three other competing methods. Yet, `sGBJ` does not require any additional hypothesis on the gene expression distribution, or the significant genes estimated coefficients, or on the correlation between genes within the gene set (as this correlation structure is estimated and taken into account in the `sGBJ` statistic). In numerical simulations, we demonstrated the good performance of `sGBJ` in all situation, making it a versatile alternative to either the `global test`, the `Wald test` or the `global boost test` which can be underperforming as soon as their assumptions are not met or there is a strong correlation between genes. Indeed, `sGBJ` seems more robust to different correlation structures and effect heterogeneity than existing methods. On the contrary, our numerical simulation study shows that neither the `global test`, nor the `Wald test`, nor the `global boost test` have consistent and reliable performance, exhibiting dramatic decrease in statistical power when some of their specific distributional assumptions are violated.

Another advantage of `sGBJ` is the absence of distributional assumptions on the gene expression measurements. Thus `sGBJ` is suited for the analysis of RNAseq data as well as microarray data. `sGBJ` can even be used for other omics data, such as proteomics data, as long as there is a known grouping structure. Notably, as `sGBJ` relies on estimations from a Cox model, it shares its assumptions such as the proportional hazards hypothesis – and so do the three competing methods.

In two real data applications – on glioma and on breast cancer – we showed that `sGBJ` offered greater accuracy compared to competing methods for estimating the lowest p-values, an important feature in multiple testing context[36]. The accuracy of the competing methods can be improved at the price of significantly longer computation times. In the breast cancer study, `sGBJ` identified more significant pathways among those known to have a link with breast cancer than the previous kernel methods applied to analyze this particular dataset[19,20]. Pathways identified in breast cancer and in glioma were consistent with other studies findings (e.g "G1 to S cell cycle reactome" or "glycosis and glucogneogenesis"[50,51] in breast cancer and E2F1 or P53[37,38] in glioma).

The adaptation of `sGBJ` method to the time-to-event context is implemented in a R package available on github at https://github.com/lauravillain/sGBJ

6 | ACKNOWLEDGMENTS

Laura Villain postoctoral contract is funded by the ERA-NET on Translational Cancer Research (TRANSCAN-2) grant, under the GliomaPRD project.

7 | AUTHOR'S CONTRIBUTIONS

BPH, LV, RT & TF designed the study, performed the analyses and wrote the manuscript.

References

1. Wang Zhong, Gerstein Mark, Snyder Michael. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics.* 2009;10(1):57–63.
2. Robinson Mark D, McCarthy Davis J, Smyth Gordon K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26(1):139–140.
3. Love Michael I, Huber Wolfgang, Anders Simon. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology.* 2014;15(12):550.
4. Law Charity W, Chen Yunshun, Shi Wei, Smyth Gordon K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology.* 2014;15(2):R29.
5. Gauthier Marine, Agniel Denis, Thiébaut Rodolphe, Hejblum Boris P. dearseq: a variance component score test for RNA-Seq differential analysis that effectively controls the false discovery rate. *NAR Genomics and Bioinformatics.* 2020;2(4):lqaa093.
6. Chung Woosung, Eum Hye Hyeon, Lee Hae-Ock, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications.* 2017;8(1):1–12.
7. Raeven René HM, Riet Elly, Meiring Hugo D, Metz Bernard, Kersten Gideon FA. Systems vaccinology and big data in the vaccine development chain. *Immunology.* 2019;156(1):33–46.
8. Liberzon Arthur, Subramanian Aravind, Pinchback Reid, Thorvaldsdóttir Helga, Tamayo Pablo, Mesirov Jill P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011;27(12):1739–1740.
9. Ogata Hiroyuki, Goto Susumu, Sato Kazushige, Fujibuchi Wataru, Bono Hidemasa, Kanehisa Minoru. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research.* 1999;27(1):29–34.
10. Ashburner Michael, Ball Catherine A, Blake Judith A, et al. Gene ontology: tool for the unification of biology. *Nature genetics.* 2000;25(1):25–29.
11. Maciejewski Henryk. Gene set analysis methods: statistical models and methodological differences. *Briefings in bioinformatics.* 2014;15(4):504–518.
12. Efron Bradley, Tibshirani Robert. On testing the significance of sets of genes. *The annals of applied statistics.* 2007;1(1):107–129.
13. Subramanian Aravind, Tamayo Pablo, Mootha Vamsi K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences.* 2005;102(43):15545–15550.
14. Hejblum Boris P, Skinner Jason, Thiébaut Rodolphe. Time-course gene set analysis for longitudinal gene expression data. *PLoS computational biology.* 2015;11(6):e1004310.
15. Barnett Ian, Mukherjee Rajarshi, Lin Xihong. The generalized higher criticism for testing SNP-set effects in genetic association studies. *Journal of the American Statistical Association.* 2017;112(517):64–76.
16. Gaynor Sheila M, Sun Ryan, Lin Xihong, Quackenbush John. Identification of differentially expressed gene sets using the Generalized Berk–Jones statistic. *Bioinformatics.* 2019;35(22):4568–4576.
17. Ishwaran Hemant, Kogalur Udaya B, Blackstone Eugene H, Lauer Michael S. Random survival forests. *The annals of applied statistics.* 2008;2(3):841–860.
18. Tibshirani Robert. The lasso method for variable selection in the Cox model. *Statistics in medicine.* 1997;16(4):385–395.
19. Cai Tianxi, Tonini Giulia, Lin Xihong. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics.* 2011;67(3):975–986.

20. Neykov Matey, Hejblum Boris P, Sinnott Jennifer A. Kernel machine score test for pathway analysis in the presence of semi-competing risks. *Statistical methods in medical research.* 2018;27(4):1099–1114.

21. Goeman Jelle J, Oosting Jan, Cleton-Jansen Anne-Marie, Anninga Jakob K, Van Houwelingen Hans C. Testing association of a pathway with survival using gene expression data. *Bioinformatics.* 2005;21(9):1950–1957.

22. Adewale Adeniyi J, Dinu Irina, Potter John D, Liu Qi, Yasui Yutaka. Pathway analysis of microarray data via regression. *Journal of Computational Biology.* 2008;15(3):269–277.

23. Boulesteix Anne-Laure, Hothorn Torsten. Testing the additional predictive value of high-dimensional molecular data. *BMC bioinformatics.* 2010;11(1):1–11.

24. Lee Seungyeoun, Kim Jinheum, Lee Sunho. A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. *BMC bioinformatics.* 2011;12(1):377.

25. Sun Ryan, Lin Xihong. Set-based tests for genetic association using the generalized Berk-Jones statistic. *arXiv preprint arXiv:1710.02469.* 2017;.

26. Sun Ryan, Hui Shirley, Bader Gary D, Lin Xihong, Kraft Peter. Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. *PLoS genetics.* 2019;15(3):e1007530.

27. Engle Robert F. Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of econometrics.* 1984;2:775–826.

28. Ostrom Quinn T, Bauchet Luc, Davis Faith G, et al. The epidemiology of glioma in adults: a "state of the science" review. *Neuro-oncology.* 2014;16(7):896–913.

29. Chen Ricky, Smith-Cohn Matthew, Cohen Adam L, Colman Howard. Glioma subclassifications and their clinical significance. *Neurotherapeutics.* 2017;14(2):284–297.

30. Sathornsumetee Sith, Reardon David A, Desjardins Annick, Quinn Jennifer A, Vredenburgh James J, Rich Jeremy N. Molecularly targeted therapy for malignant glioma. *Cancer.* 2007;110(1):13–24.

31. Norden Andrew D, Wen Patrick Y. Glioma therapy in adults. *The neurologist.* 2006;12(6):279–292.

32. Madhavan Subha, Gusev Yuriy, Harris Michael, et al. G-DOC: a systems medicine platform for personalized oncology. *Neoplasia.* 2011;13(9):771–783.

33. Georgetown Database of Cancer (G-DOC) platform https://gdoc.georgetown.edu/gdoc/.

34. Madhavan Subha, Zenklusen Jean-Claude, Kotliarov Yuri, Sahni Himanso, Fine Howard A, Buetow Kenneth. Rembrandt: helping personalized medicine become a reality through integrative translational research. *Molecular cancer research.* 2009;7(2):157–167.

35. Benjamini Yoav, Hochberg Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological).* 1995;57(1):289–300.

36. Phipson B., Smyth G.. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology.* 2010;9.

37. Xiao Bingxiang, Tan Li, He Benfu, Liu Zhiliang, Xu Ruxiang. MiRNA-329 targeting E2F1 inhibits cell proliferation in glioma cells. *Journal of translational medicine.* 2013;11(1):1–10.

38. Squatrito Massimo, Brennan Cameron W, Helmy Karim, Huse Jason T, Petrini John H, Holland Eric C. Loss of ATM/Chk2/p53 pathway components accelerates tumor development and contributes to radiation resistance in gliomas. *Cancer cell.* 2010;18(6):619–629.

39. Alvarez-Breckenridge Christopher A, Yu Jianhua, Price Richard, et al. NK cells impede glioblastoma virotherapy through NKp30 and NKp46 natural cytotoxicity receptors. *Nature medicine.* 2012;18(12):1827–1834.

40. Hambardzumyan Dolores, Gutmann David H, Kettenmann Helmut. The role of microglia and macrophages in glioma maintenance and progression. *Nature neuroscience.* 2016;19(1):20.

41. Guan Qian, Yuan Li, Lin Ao, et al. KRAS gene polymorphisms are associated with the risk of glioma: a two-center case-control study. *Translational Pediatrics.* 2021;10(3):579.

42. Friedmann-Morvinski Dinorah, Bushong Eric A, Ke Eugene, et al. Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. *Science.* 2012;338(6110):1080–1084.

43. Kaul Aparna, Toonen Joseph A, Cimino Patrick J, Gianino Scott M, Gutmann David H. Akt-or MEK-mediated mTOR inhibition suppresses Nf1 optic glioma growth. *Neuro-oncology.* 2015;17(6):843–853.

44. Scuderi Sarah A, Lanza Marika, Casili Giovanna, et al. TBK1 inhibitor exerts anti-proliferative effect on glioblastoma multiforme cells. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics.* 2021;.

45. Benson John R, Jatoi Ismail. The global breast cancer burden. *Future oncology.* 2012;8(6):697–702.

46. Scully Olivia Jane, Bay Boon-Huat, Yip George, Yu Yingnan. Breast cancer metastasis. *Cancer Genomics-Proteomics.* 2012;9(5):311–320.

47. Van De Vijver Marc J, He Yudong D, Van't Veer Laura J, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine.* 2002;347(25):1999–2009.

48. McPherson Klim, Steel CaMa, Dixon JM. Breast cancer—epidemiology, risk factors, and genetics. *Bmj.* 2000;321(7261):624–628.

49. Kanehisa Minoru, Araki Michihiro, Goto Susumu, et al. KEGG for linking genomes to life and the environment. *Nucleic acids research.* 2007;36(suppl_1):D480–D484.

50. Bassi C, Fortin J, Snow B E, et al. The PTEN and ATM axis controls the G1/S cell cycle checkpoint and tumorigenesis in HER2-positive breast cancer. *Cell Death & Differentiation.* 2021;:1–16.

51. Li Wenhui, Xu Ming, Li Yu, et al. Comprehensive analysis of the association between tumor glycolysis and immune/inflammation function in breast cancer. *Journal of translational medicine.* 2020;18(1):1–12.

52. Tan Juan, Yu Chen-Yang, Wang Zhen-Hua, et al. Genetic variants in the inositol phosphate metabolism pathway and risk of different types of cancer. *Scientific reports.* 2015;5(1):1–8.

53. Siddiqui Aarif, Ceppi Paolo. A non-proliferative role of pyrimidine metabolism in cancer. *Molecular metabolism.* 2020;35:100962.

8 | APPENDIX A: TYPE-I ERROR

We simulated the Case (I) and Case (III) presented Section3 but with no significant genes to evaluate the type-I error on the four methods. Figure 1 shows that all four methods control the Type-I error.

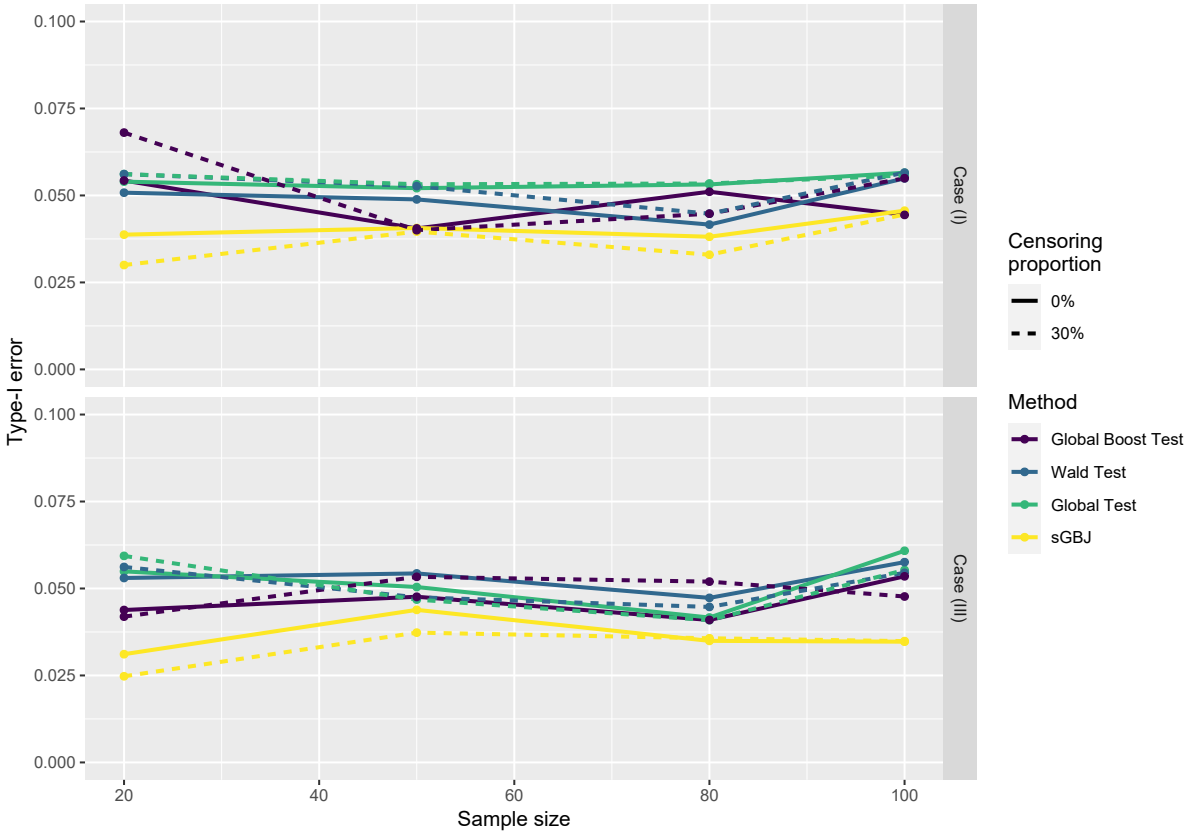

FIGURE 1 Type-I error computed for each method (`sGBJ`, `Global Boost Test`, `Global test`, `Wald test`) with increasing sample size, on the Cases (I) and (III) described Section 3, the Case (II) being identical to the Case (I) when there is no significant genes. Dotted lines represent a censoring fraction of 0.3; full lines, a censoring fraction of 0. The type-I error is the proportion of significant genes among the true negative.

9 | APPENDIX B: COMPARISON OF sGBJ WITH OTHER METHODS FOR THE BREAST CANCER STUDY

As we did for the Rembrandt study, we evaluated the sGBJ method among `global test`, `Wald test` and `global boost test`. Figure 2 presents the raw p-values in function of the ranks of the p-values computed for sGBJ. The Benjamini Hochberg[35] limit shows the value a p-value must cross to be significant after multiple test correction, while the 5% threshold shows where the 0.05 p-value limit is. As we can see, a high number of pathways are found significant, and the four methods performs similarly, with a higher degree of precision for the low p-values with the sGBJ method.

How to cite this article: Villain L, Ferté T, Thiébaut R, and Hejblum BP (2021), Gene Set Analysis for time-to-event outcome with the Generalized Berk–Jones statistic, *Statistics in Medicine*, *vol*.

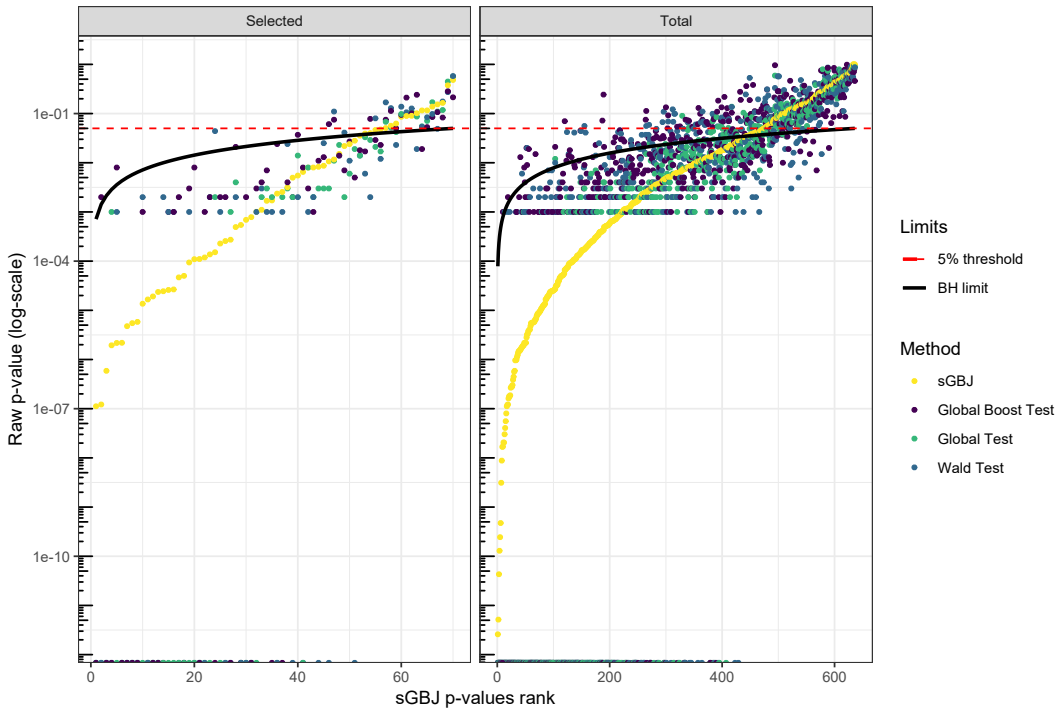

FIGURE 2 Raw p-values for the 4 methods, in function of the ranks of the p-values of sGBJ: sGBJ, `global boost test`, `Wald test` and `global test`, with the 5% threshold and the Benjamini Hochberg limit. *Nota Bene*: The Benjamini Hochberg limit only applies for the sGBJ method, as the ranks are computed for sGBJ only.