mathematics

MDPI

Article

Gene Set Analysis Using Spatial Statistics

Angela L. Riffo-Campos [1,2], Guillermo Ayala [2,*] and Francisco Montes [2]

[1] Centro de Excelencia de Modelación y Computación Científica, Universidad de La Frontera, Temuco 4780000, Chile; angela.riffo@ufrontera.cl
[2] Departamento de Estadística e Investigación Operativa, Universidad de Valencia, Avda. Vicent Andrés Estellés, 1, 46100 Burjasot, Spain; francisco.montes@uv.es
[*] Correspondence: Guillermo.Ayala@uv.es

Abstract: Gene differential expression consists of the study of the possible association between the gene expression, evaluated using different types of data as DNA microarray or RNA-Seq technologies, and the phenotype. This can be performed marginally for each gene (differential gene expression) or using a gene set collection (gene set analysis). A previous (marginal) per-gene analysis of differential expression is usually performed in order to obtain a set of significant genes or marginal p-values used later in the study of association between phenotype and gene expression. This paper proposes the use of methods of spatial statistics for testing gene set differential expression analysis using paired samples of RNA-Seq counts. This approach is not based on a previous per-gene differential expression analysis. Instead, we compare the paired counts within each sample/control using a binomial test. Each pair per gene will produce a p-value so gene expression profile is transformed into a vector of p-values which will be considered as an event belonging to a point pattern. This would be the first component of a bivariate point pattern. The second component is generated by applying two different randomization distributions to the correspondence between samples and treatment. The self-contained null hypothesis considered in gene set analysis can be formulated in terms of the associated point pattern as a random labeling of the considered bivariate point pattern. The gene sets were defined by the Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The proposed methodology was tested in four RNA-Seq datasets of colorectal cancer (CRC) patients and the results were contrasted with those obtained using the edgeR-GOseq pipeline. The proposed methodology has proved to be consistent at the biological and statistical level, in particular using Cuzick and Edwards test with one realization of the second component and between-pair distribution.

Keywords: colorectal cancer; RNA-Seq; paired samples; spatial point pattern

check for updates

Citation: Riffo-Campos, A.L.; Ayala, G.; Montes, F. Gene Set Analysis Using Spatial Statistics. *Mathematics* **2021**, *9*, 521. https://doi.org/10.3390/math9050521

Academic Editor: Elvira Di Nardo

Received: 26 January 2021
Accepted: 23 February 2021
Published: 3 March 2021

1. Introduction

Global gene expression (transcriptome) can be quantified using DNA microarrays [1] and RNA sequencing (RNA-Seq) technologies [2]. RNA-Seq is widely used to understand and describe the biological mechanisms involved in transcription observed under different experimental conditions. The statistical comparison of the means of the gene expression is known as differential gene expression. This comparison can be performed at the gene level, i.e., a marginal analysis of each gene. There are many pipelines to perform an RNA-Seq data analysis [3]. A list of differentially expressed genes is obtained, the significant genes. Usually, it is expected to find a relationship between these significant genes and the biological mechanisms that underlie the observed phenotype. This biological mechanism is controlled by a gene set. This justifies to analyze the differential expression of gene sets by considering them from the very initial step. This is called gene (enrichment) set analysis [4]. For gene set analysis, the choice of the statistical method, the type of null hypothesis, and the gene-association measure are the most important considerations. In addition, the set of genes can be biologically defined using, among others, GO [5] and/or the KEGG Ontology (KO) [6]. GO include three categories: the first one is the biological objective to which the gene or its product contributes and is called Biological process; the second one is the biochemical activity of a gene product, called Molecular Function; the third is called Cellular Component and it refers to the place in the cell where a gene product is active [5]. On the other hand, KO includes all molecular networks in the categories of Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organisms Systems, Human Diseases, Brite Hierarchies and Not Included in Pathway or Brite [6].

The methods for gene set analysis can be classified, according to the null hypothesis tested, into self-contained or competitive tests [7]. The basic question can be formulated as: Is there any association between a set of genes and the phenotype? This is a very vague question. A more precise formulation is required. Different interpretations of the question are possible. In [8], they formulate the following two null hypotheses that specify the previous question in two different ways. We reproduce the corresponding null hypotheses. The competitive hypothesis is formulated as "The genes in a gene set show the same pattern of associations with the phenotype compared with the rest of the genes". The self-contained hypothesis is formulated as "The gene set does not contain any genes whose expression levels are associated with the phenotype of interest." We are concerned in this paper with the testing of the self-contained null hypothesis using RNA-Seq data observed under two conditions (case-control) using a paired design. A huge literature of gene set analysis exists. A large part of it was developed and implemented for microarray data and later adapted to RNA-Seq data. Good reviews can be found in [9,10]. Gene set over-representation analysis (ORA) consists of evaluating if a previously defined gene set, such as GO terms, is more represented than the others in the list of genes previously selected as differentially expressed, and if this over-representation is unlikely to be due to chance. This is used for instance by GOseq [11]. This approach assumes independent gene expressions, i.e., it starts from a marginal analysis of the gene. Nonetheless, it is well known that, when an alteration occurs, it is not an individual gene, but a gene set that is affected. To take this into account, gene set enrichment analysis (GSEA) [12,13] offers a modification over the previous methods. The GSEA method does not start from a previously selected list of genes differentially expressed, but, instead, it uses the gene set as the initial unit; an example is the SeqGSEA method [14]. However, these tools apply first a gene-level test on the original data.

An alternative is performing the analysis considering first the gene relationships and to focus directly on the gene set differential expression. This idea has been exploited by some methods previously proposed on microarray data [15]. Nevertheless, the gene set differential expression for RNA-Seq data are less studied. Some interesting examples are [16,17]. The method proposed in this paper is designed to test the self-contained hypothesis by using statistical analysis of spatial point patterns.

The study of spatial aggregation or clustering has a long history in the spatial statistics literature [18]. Nonetheless, it is, up to our knowledge, the first time in which spatial statistical methods,, i.e., the Cuzick and Edwards test, the Diggle, Morris, and Morton–Jones test and the Diggle test, are applied as an effective approach for gene set analysis. The methodology is implemented in the R package OMICfpp2 available at http://www.uv.es/ayala/software/OMICfpp2_1.0.tar.gz (accessed on 25 February 2021) and https://github.com/JdMDE/OMICfpp2 (accessed on 25 February 2021).

2. Methodology

2.1. General Notation

Let us introduce the basic notation needed later. This paper is concerned with paired design and the notation is given accordingly.

Let N be the number of genes and n the number of pairs of samples. The value of N is much larger than n, $N \gg n$. Let x_{ijk} denote the count corresponding to the i-th gene ($i = 1, \ldots, N$), j-th pair ($j = 1, \ldots, n$) and k the element within the pair ($k = 1, 2$). The total number of counts for the sample (j,k), its library size, will be denoted by $m_{jk} = \sum_{i=1}^{N} x_{ijk}$. We assume the values x_{ijk} for a given i are in the i-th row of the expression matrix in such a way that each column would be associated with a (j,k) sample, so we have an $N \times 2n$ matrix. If we consider the random expression matrix instead of the observed one, then their rows would be dependent random vectors (the expressions of the genes are dependent), and the columns would be independent random vectors (corresponding to different individuals).

We are interested in a gene set collection $S_1, \ldots, S_T$, where $S_t \subset G$ for $t = 1, \ldots, T$ and $G = \{1, \ldots, N\}$ is the universe of genes. These gene sets do not need to be necessarily disjoint. Given the observed expression matrix $\boldsymbol{x}$ and a gene set S_t, we can extract the matrix corresponding to the rows in S_t, i.e., $\boldsymbol{x}_{S_t}$. Let $\phi(S_t)$ be the set composed by the columns of $\boldsymbol{x}_{S_t}$, $\phi(S_t) \subset \mathbb{R}^{|S_t|}$. This set $\phi(S_t)$ can be considered as a point pattern (it will be called sample point pattern) where the corresponding point process could be denoted as $\Phi(S_t)$. A formal presentation of point process theory can be found in [19].

For the (j,k) sample, we have a phenotypic covariable, $\boldsymbol{y}_{jk}$ ($\in \mathbb{R}$), for instance an experimental factor indicating case or control. The previous point pattern, $\phi(S_t)$, and this covariable can be put together in a so-called marked point pattern.

This simple idea is used later to connect two different topics: statistical analysis of marked point patterns and gene set analysis.

2.2. Paired RNA-Seq Samples

The most common setup consists of two groups of samples to be compared. Our data are pairs of RNA-Seq counts quantifying the gene expression i.e., the samples are grouped in pairs corresponding to two conditions observed on the same individual. We will have a binary phenotypic covariable y_{jk} where $y_{jk} = 0$ (respectively 1) corresponds to a control (respectively case).

A simple procedure was proposed in [20] to test the null hypothesis of no association between condition and expression. This procedure assumes that, under the null hypothesis, the random count X_{ij1} has a binomial distribution $X_{ij1} \sim Bi\left(x_{ij1} + x_{ij2}, \frac{m_{j1}}{m_{j1}+m_{j2}}\right)$, where $x_{ij1} + x_{ij2}$ is the count for the i-the gene and the j-th pair. This count and the sum of library sizes, $m_{j1} + m_{j2}$, are considered given. The p-value will be calculated as the sum of the probabilities lesser or equal than the observed x_{ij1} value and will be denoted as p_{ij}. If we have n pairs of samples, then a p-value will be observed per pair and gene so we will have $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{in})$ for the i-th gene. Our original $N \times 2n$ gene expression matrix is transformed in a $N \times n$ p-value matrix where the (i,j) entry will correspond to the p-value of the j-th pair of the i-th gene. Let the observed matrix of p-values be $\tilde{\boldsymbol{p}}_0$. The corresponding random p-value matrix will be denoted $\tilde{\boldsymbol{P}}_0$. Note that the columns of the random matrix $\tilde{\boldsymbol{P}}_0$ are independent but not their rows. Under the null hypotheses of no expression-phenotype association for all genes, the random p-value $\tilde{P}_0(i,j)$ follows approximately a uniform distribution.

2.3. Gene Set Point Pattern

We are interested in the study of the differential expression between conditions for a given (previously defined) gene set and $G = \{1, \ldots, N\}$ is the universe of genes. The i-th gene will have its expression profile in the i-th row of the expression matrix. Let $S = \{i_1, \ldots, i_{|S|}\}$ a given gene set with $S \subset G$. Our approach will test the self-contained null hypothesis of no differential expression for the gene set. The most common approach consists of two steps. Firstly, the null hypotheses of no differential expression per each gene are tested. Secondly, the statistics (or p-values) of these (marginal) tests are aggregated ignoring later the dependencies between them. This point has to be incorporated in the analysis and we deal with it by using point processes.

Our important gene set is $S = \{i_1, \ldots, i_{|S|}\}$. The vector $\boldsymbol{v}_j = (p_{i_1,j}, \ldots, p_{i_{|S|},j})' \in [0,1]^{|S|}$ contains all the observed p-values for the genes in S corresponding to the j-th pair. We can consider all samples jointly in $\boldsymbol{\phi}_S = \{v_1, \ldots, v_n\}$. It is a finite set of n points contained in the unit hyper-cube $[0,1]^{|S|}$, a point pattern. Each event corresponds to a sample and each dimension of the point corresponds with a gene. As we are working with a paired design and p-values, this point pattern is a natural description of the differential expression of the gene set in both conditions. No independence between genes is assumed. Let $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$ be a random sample of n random vectors, distributed as $\boldsymbol{V}$, whose corresponding observed values are $v_1, \ldots, v_n$. Analogously, $\boldsymbol{\phi}_S = \{v_1, \ldots, v_n\}$ is the point pattern and $\Phi_S = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n\}$ is the point process. In our approach, each event of the point pattern $\phi(S)$ corresponds to a pair of samples, more precisely to the p-values of these pair of samples.

2.4. Testing Differential Expression

We are going to test the gene set differential expression by using a bivariate point process. Each realization will have two components with n points each. The first component is the point process corresponding to the p-values obtained using the original sample classification. This first component will be called cases. The second component is generated by applying a randomization to the original sample classification. The second component will be called controls.

This bivariate point process, under the null hypothesis, would be just a random labeling (with n point per component) of the union of both processes.

This random labeling hypothesis will be tested using different statistical procedures taken from the literature of spatial point processes, and they were designed in order to look for some characteristic of the joint distribution. We provide next a short description of the tests.

We consider a given gene set S. The outline of our procedure is as follows:

1. Using the original pairs, we obtain the first point pattern corresponding to the original pairs or cases, ϕ_0.
2. We choose a randomization distribution, between-pair or complete distribution, and a number of randomizations B to be performed.
3. Using the chosen distribution in the previous step, we generate new pairs and a new sample point pattern associated with these pairs $\phi_1, \ldots, \phi_B$.
4. For the i-th bivariate point pattern (ϕ_0, ϕ_i), it is tested if it can be considered a random labeling of the point pattern $\phi_0 \cup \phi_i$, and the corresponding p-value is obtained.

2.4.1. Generation of Controls

These are the randomization distributions used to generate the random points.

Between-pair (BP) distribution. The first element of each pair is maintained as the original one. The second element of each pair is obtained randomly permuting the second components of all pairs between them. We have $(y_{i,1}, y_{\gamma(i),2})$ for $i = 1, \ldots, n$, where γ is now a permutation of $(1, \ldots, n)$. The number of possible permutations is $n!$.

Complete (C) distribution. Let us choose $I = \{i_1, \ldots, i_n\}$ as a random subset of $\{1, \ldots, 2n\}$. The indices of $\{1, \ldots, 2n\}$ not in $\{i_1, \ldots, i_n\}$ can be denoted $J = \{j_1, \ldots, j_n\}$. A random correspondence between I and J will produce the pairs. The number of possible values is $\frac{(2n)!}{n!}$.

2.4.2. Statistical Tests

Cuzick and Edwards test (CET) was proposed in [21]. We have a bivariate point pattern corresponding with cases and controls. The objective is to detect spatial clustering of cases. It is assumed that their spatial distribution is not homogeneous like in our problem. It is based on distances between nearest neighbor (NN) pairs of points. Let $\{z_j\}_{j=1,\ldots,2n}$ be the locations of the combined sample where the indices have been randomly permuted. We define for $i = 1, .., n$

$$\delta_i = \begin{cases} 1 & \text{if} \quad z_i \text{ is a case} \\ 0 & \text{if} \quad z_i \text{ is a control} \end{cases}$$

and

$$d_i = \begin{cases} 1 & \text{if the NN to } z_i \text{ is a case} \\ 0 & \text{if the NN to } z_i \text{ is a control} \end{cases}$$

The statistic is $T = \sum_{i=1}^{n} \delta_i d_i$, i.e., we are counting the number of cases with a case as the nearest neighbor. It is clear that large values of the statistic are associated with clusters of cases corresponding with the alternative hypothesis of a phenotype-expression association.

Diggle, Morris, and the Morton–Jones test (DMMT). Ref. [22] was proposed within the research of a high risk around a specified point. Consider again the variables $\{\delta_i\}_{i=1,\ldots,2n}$ previously defined. Let $\gamma_i = P(\delta_i = 1)$ and $\gamma_{(i)}$, the corresponding values ordered according to the distance to the origin. Under the null hypothesis of no differential expression, the maximum likelihood estimators of the $\gamma_{(i)}$ are easily obtained by the pool-adjacent violators algorithm: $\hat{\gamma}_{(i)} = \min_{s \le i} \max_{t \ge i} \frac{\sum_{r=s}^{t} \delta_{(r)}}{t-s+1}$. The maximum likelihood ratio test statistic is given by $T_D = 2\sum_{i=1}^{2n}\{\delta_{(i)} \log \frac{\hat{\gamma}_{(i)}}{1/2} + (1-\delta_{(i)}) \log \frac{1-\hat{\gamma}_{(i)}}{1/2}\}$.

Diggle test (DT) [23] studies the possible raised incidence of certain types of cancer near nuclear installations. The test is based on fitting a particular class of a non-homogeneous Poisson point process model to data. Let $\lambda(x)$ be the intensity function of $\Phi_1(S)$ under the alternative hypothesis H_1. We can assume that $\lambda(x)$ has a multiplicative decomposition as $\lambda(\boldsymbol{x};\gamma) = \rho\lambda_0(\boldsymbol{x}) f(\boldsymbol{x}^t\boldsymbol{x};\boldsymbol{\theta})$, where $\boldsymbol{x}^t$ is the transpose of $\boldsymbol{x}$ and $\lambda_0(\boldsymbol{x})$ would represent the spatial variation in intensity under the self-contained null hypothesis. This null intensity function could be estimated using a kernel estimator from the control sample point process $\Phi_2(S)$ of p-values. For a given realization of $\Phi_2(S)$, i.e., $\phi_2(S) = \{\boldsymbol{v}_j\}_{j=1}^{n}$, the kernel estimator using a Gaussian kernel is given by $\hat{\lambda}_0(\boldsymbol{x}) = \frac{\sum_{j=1}^{n} \exp\{\frac{-1}{2h}(\boldsymbol{x}-\boldsymbol{v}_j)^t(\boldsymbol{x}-\boldsymbol{v}_j)\}}{2\pi h^2}$. The function f can be quite general. However, the following function permits an easy computation of the maximum likelihood estimator of the parameters $\boldsymbol{\theta} = (\alpha,\beta)^t$: $f(\boldsymbol{x};\alpha,\beta) = 1 + \alpha\exp(-\beta\boldsymbol{x}^t\boldsymbol{x})$. Using the Gaussian kernel estimate for the function λ_0 and the proposed f, it is easy to obtain the formulas needed to obtain the maximum pseudo likelihood estimator of $\theta = (\alpha,\beta)^t$. Note that the null hypothesis of no clustering around the origin corresponds with $\alpha = \beta = 0$. In order to test this null hypothesis, we compare $D = 2(L(\hat{\alpha},\hat{\beta}) - L(0,0))$ with critical values of χ^2_2. Details can be found in [23]. We have implemented it in the n-dimensional case in our R package OMICfpp2.

2.4.3. Testing the Self-Contained Hypothesis

Under this hypothesis, the original point process, $\Phi_0(S)$, would be a non-homogeneous Poisson point process in the hyper cube $[0,1]^{|S|}$. Note that, under the alternative hypothesis, the point process will tend to produce clustering around the origin. Many other statistical tests could be used and this could be a good line of future research. The three previous tests have been taken from an epidemiological context.

The points correspond to the p-values for the different genes in our important gene set S. If there is no gene set differential expression, then the $2n$ points are independent and identically distributed following a common unknown distribution. We preserve the original label of the point if it corresponds to an original pair or to a randomly generated pair. No differential expression means that the cases are just a random selection of n points from the total $2n$ points, i.e., the bivariate point pattern is just a random labeling of the original point set. We reformulate the testing of no gene set differential expression in that labeling has been tested in a bivariate point process.

This random labeling has been tested using a Monte Carlo test. Let us give a short description. Let (ϕ_0, ϕ_i) be a bivariate point pattern where its first component, ϕ_0, is the original point pattern and its second component, ϕ_i, is a control. Let t_0 be any of the three previous statistics evaluated for this bivariate point pattern. The set $\phi_0 \cup \phi_i$ is randomly partitioned into two new sets of n points. The same statistic is evaluated for this new bivariate point pattern. We repeat the selection process B times independently obtaining the statistics $t_1, \ldots, t_B$. Under the null hypothesis, any order of the vector $(t_0, t_1, \ldots, t_B)$

has the same probability. The Monte Carlo p-value [24] is given by $p = \frac{|\{b:|t_b|>|t_0|, b=1,\ldots,B\}|}{B+1}$. In the experimental study, $B = 100$ will be used. This Monte Carlo p-value will be used to test the self-contained null hypothesis.

The sample point pattern for cases is unique, but many sample point patterns corresponding to controls can be generated. For each bivariate point pattern generated, a Monte Carlo p-value is obtained. It is clear that the computational time is greater when more than one control sample point pattern is generated. More than one realization produces different p-values that will be aggregated using meta-analysis techniques for p-values as the Fisher's method. It is interesting to evaluate if just one realization of controls is enough or if more than realization produces better results. This could be evaluated in Section 3.

2.5. Data

A total of four RNA-Seq data sets with paired (tumor/adjacent normal tissue) samples from CRC patients have been analyzed. The first three data sets correspond to the Bioprojects PRJNA411984 [25], PRJNA413956 [26] and PRJNA218851 [27] with 2, 7, and 18 raw data pairs, respectively. The fourth data set include 50 pairs of preprocessed data (count files) from The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/ (accessed on 26 January 2021)).

We are going to evaluate if just one or multiple realizations are needed to generate the second component of the bivariate point process using the TCGA dataset. The gene sets were defined using GO terms and KEGG ontology. The GO gene set collection uses only the biological process category and gene sets with ten or more genes in the set, resulting in a total of 2815. The KEGG gene set collections have been entirely used and there are a total of 340.

The TCGA dataset has been analyzed using the three tests proposed: CET, DMMT, and DT. One realization (OR) of the between-pair (BP) randomization distribution has been used to generate the second component of bivariate point process. The random labeling hypothesis has been tested using 100 simulations.

The four datasets were analyzed using edgeR-GOseq pipeline. The method edgeR can be found in [28,29]. The `GOseq` R package [11] allows us to analyze gene sets from GO using RNA-Seq data and also KEGG pathways analysis.

The whole code needed to reproduce our paper can be found in the supplementary file `SupplementaryMaterialMethods_pointgene.pdf`.

3. Results

3.1. One or Multiple Realizations?

Out of all GO gene sets, 8% reported as significant (p-value < 0.05) using OR were reported too using MR with all tests (Figure 1A). For KEGG, 143 unique gene sets have been reported in OR, and 52 unique gene sets have been reported in MR, which corresponds to a decreasing of 64% (Figure 1B).

3.2. Analyzing the Tests

A total of 80, 0, 1 GO terms and 63, 5, 6 KEGG pathways were reported (p-value < 0.05) by CET, DMMT, and DT, respectively. The DMMT and DT were more conservative than CET at reporting differentially expressed gene sets. No common gene sets were reported between the spatial tests.

The first five gene sets with the lowest p-values, top genes, obtained by each test are compared in order to identify the most appropriate approach according to the biological relevance. However, the number of papers dealing with CRC, associated with the gene sets reported as significant by the test, could not be sufficient criteria to evaluate them because this number is closely related with the method of detection used and its age.

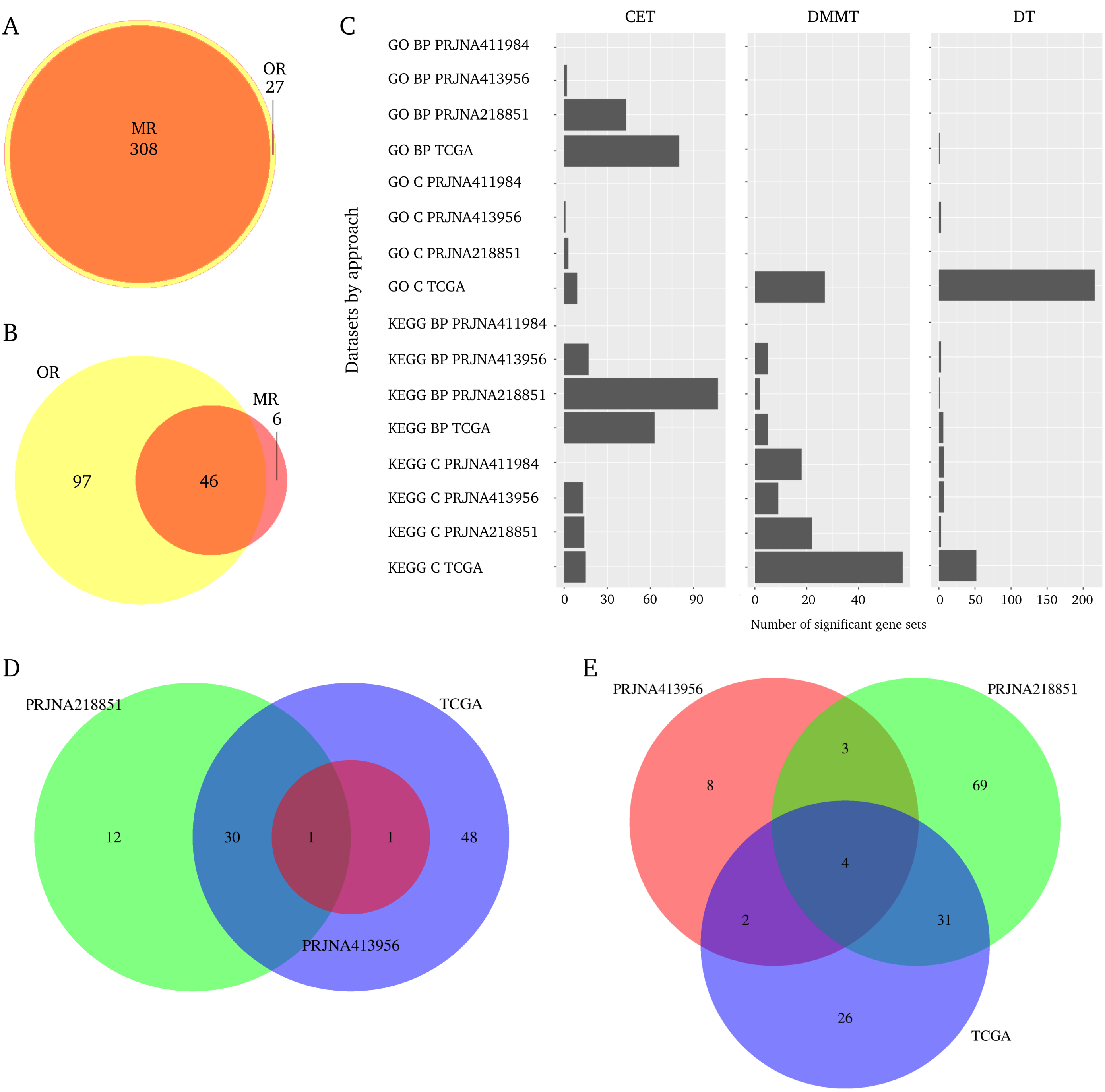

Figure 1. Overlapping between significant gene sets across the methods. (**A**) Venn diagram of significant GO gene sets obtained using TCGA dataset and OR or MR with all tests. (**B**) Venn diagram of significant KEGG gene sets obtained using a TCGA dataset and OR or MR with all tests. (**C**) Number of significant GO and KEGG gene sets (using one realization) for the three tests (CET, DMMT, DT) and the between-pair and complete randomization distributions. (**D**) Venn diagram of significant GO gene sets for all datasets using BP and CET. (**E**) Venn diagram of significant KEGG gene sets for all datasets using BP CET.

Thus, as a complement to these results, we include a list of biological pathways that have been shown to be associated with CRC (see [30–34]) including EGFR, MAPK, Notch, PI3K, P53, Ras, TGF-β, Wnt/β-catenin, JAK-STAT, VEGF, or NF-kappaB signaling pathway, and, therefore, could be (but not necessarily) a gold standard. In this sense, 14 KEGG biological pathways and 156 GO terms that represent the signaling pathway were selected to evaluate which tests reported these gene sets more frequently in their results. Of the 156

GO terms, many were made up of a single gene, being subsets of the signaling pathways, so a short list of 15 GO gene sets was used, which includes only the general signaling pathways (see tables in Supplementary Material).

The top GO and KEGG gene sets reported by CET were highly related to CRC and other cancer types, as reported by Comparative Toxicogenomics Database and bibliographic databases (Table 1). Additionally, canonical pathways involved in CRC as PI3K-Akt, JAK-STAT, and Ras signaling pathways were reported in the first places by CET. The DMMT did not report gene sets differentially expressed in GO terms. For KEGG results, the gene sets reported were less associated with CRC than the CET as for the number of articles, but also includes canonical pathways such as EGFR tyrosine kinase inhibitor resistance. In general, the KEGG results obtained by the DT method were less associated (but also related) with CRC than those obtained by CET and DMMT. This was in concordance with the p-value reported in the gene sets by each test.

Table 1. Five gene sets with lowest p-values reported by each test using the between-pair distribution: CET, DMMT, and DT tests. The column headed "n" refers to the number of genes in the gene set. The number of papers "n rep" reporting the gene set association with Colorectal or Colonic Neoplasms have been obtained from the Comparative Toxicogenomics Database (GO terms) and MEDLINE bibliographic database (KEGG ID AND "colorectal cancer"). The asterisk * indicates that the gene set has been described as related to other cancer types.

ID Gene Set	Name	n	Test	p-Value	n rep
GO:0035195	Gene silencing by miRNA	577	CET	<0.00001	4
GO:0007186	G protein-coupled receptor signaling pathway	868	CET	<0.00001	18
GO:0045944	Positive regulation of transcription by RNA polymerase II	975	CET	<0.00001	111
GO:0006357	Regulation of transcription by RNA polymerase II	917	CET	<0.00001	76
GO:0050911	Detection of chemical stimulus involved in sensory perception of smell	427	CET	<0.00001	0 *
GO:0045190	Isotype switching	17	DT	0.0450	7
hsa05200	Pathways in cancer	530	CET	<0.001	170
hsa04014	Ras signaling pathway	232	CET	<0.001	12
hsa04020	Calcium signaling pathway	193	CET	<0.001	46
hsa04151	PI3K-Akt signaling pathway	354	CET	<0.001	59
hsa04630	JAK-STAT signaling pathway	162	CET	<0.001	35
hsa05340	Primary immunodeficiency	38	DMMT	0.01	12
hsa01212	Fatty acid metabolism	57	DMMT	0.02	7
hsa00071	Fatty acid degradation	44	DMMT	0.03	13
hsa05167	Kaposi sarcoma-associated herpesvirus infection	186	DMMT	0.03	8
hsa01521	EGFR tyrosine kinase inhibitor resistance	79	DMMT	0.04	7
hsa04512	ECM-receptor interaction	88	DT	<0.001	91
hsa04071	Sphingolipid signaling pathway	119	DT	<0.001	5
hsa03410	Base excision repair	33	DT	0.02	9
hsa05033	Nicotine addiction	40	DT	0.03	8
hsa04659	Th17 cell differentiation	107	DT	0.04	7

3.3. Choosing the Randomization Distribution

The between-pair randomization distribution is the most natural one for paired designs, but the complete randomization distribution (forgetting the paired design) could be applied too. Thus, the more appropriate randomization distribution to generate the second component of the point process has been evaluated. A total of 9, 27, and 216 GO terms and 63, 5, and 6 KEGG pathways were reported (p-value < 0.05) by CET, DMMT, and DT, respectively, using complete distribution.

No common gene sets were reported between the spatial tests using either complete distribution or between-pair distribution. The results obtained using complete and between-pair distributions for each test were compared: a total of 9, 0, and 1 GO terms and 11, 1, and 1 KEGG pathways were reported (p-value < 0.05) in common using CET, DMMT, and DT, respectively. These data are not shown and can be found in Supplementary Material. Thus, most of the gene sets reported using CET in complete distribution were also included using the between-pair distribution. The number of GO gene sets reported decreased at least eight times for CET and increases by one hundred percent DMMT and DT. Regarding the biological assertiveness of the results, fewer articles associated with CRC were included in the groups reported using complete distribution, although canonical pathways were included in the results in all spatial tests (Table 2).

Table 2. List of the top five gene sets reported by each statistic test using the complete distribution: CET, DMMT, and DT. The "n" column refers to genes in the set. The number of papers (n rep) reporting the gene set association with Colorectal Neoplasms obtained from the Comparative Toxicogenomics Database (GO terms) and MEDLINE bibliographic database (KEGG ID AND "colorectal cancer"). The * indicates that the gene set was related to other cancer types.

ID Gene Set	Name	n	Test	*p*-Value	n rep
GO:0050911	Detection of chemical stimulus involved in sensory perception of smell	427	CET	0.0001	0*
GO:0035195	Gene silencing by miRNA	577	CET	0.0005	4
GO:0006396	RNA processing	544	CET	0.0036	4
GO:0018149	Peptide cross-linking	26	CET	0.0159	0 *
GO:0070268	Cornification	112	CET	0.0294	0*
GO:0071320	Cellular response to cAMP	52	DMMT	0.0309	6
GO:1990403	Embryonic brain development	13	DMMT	0.0323	14
GO:0071392	Cellular response to estradiol stimulus	34	DMMT	0.0338	13
GO:0051965	Positive regulation of synapse assembly	61	DMMT	0.0364	3
GO:0031145	Anaphase-promoting complex-dependent catabolic process	81	DMMT	0.0368	2
GO:0071363	Cellular response to growth factor stimulus	45	DT	0.0019	14
GO:0050808	Synapse organization	38	DT	0.0019	15
GO:0045190	Isotype switching	17	DT	0.0029	7
GO:0015721	Bile acid and bile salt transport	19	DT	0.0030	2
GO:0002931	Response to ischemia	50	DT	0.0033	33
hsa04630	JAK-STAT signaling pathway	162	CET	<0.001	35
hsa04740	Olfactory transduction	448	CET	<0.001	10
hsa05206	MicroRNAs in cancer	310	CET	<0.001	31
hsa05218	Melanoma	72	CET	<0.001	39
hsa05224	Breast cancer	147	CET	<0.001	5
hsa04215	Apoptosis multiple species	32	DMMT	<0.001	5
hsa04660	T cell receptor signaling pathway	104	DMMT	<0.001	37
hsa04150	mTOR signaling pathway	153	DMMT	<0.001	32
hsa04934	Cushing syndrome	155	DMMT	<0.001	0
hsa04928	Parathyroid hormone synthesis, secretion and action	106	DMMT	<0.001	0
hsa01521	EGFR tyrosine kinase inhibitor resistance	79	DT	<0.001	7
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	70	DT	<0.001	15
hsa03030	DNA replication	36	DT	<0.001	25
hsa04724	Glutamatergic synapse	114	DT	<0.001	14
hsa04012	ErbB signaling pathway	85	DT	<0.001	59

3.4. Effect of Sample Size

The methodology has been evaluated for different sample sizes: an extreme case of two pairs (PRJNA411984); two intermediate cases of 7 (PRJNA413956) and 18 pairs of samples (PRJNA218851) and the TCGA dataset with 50 pairs.

The number of gene sets reported in all methods increases with the number of samples (Figure 1C).

All spatial tests using between-pair or complete distribution reported results from 50 pairs of samples (TCGA dataset), with the exception of DMMT using BP distribution. The DMMT and DT report less gene set that are significant when using between-pair distribution, while, in CET, the opposite occurs in all sample sizes.

In the PRJNA411984 dataset (2 pairs), only DMMT and DT using complete distribution reported 18 and 7 significant KEGG gene sets, respectively. Genes grouping using KEGG ontology seems to be more appropriate for using spatial tests than GO categories.

Regarding the consistency in the results reported by each test across datasets, the BP CET reported more results in common between the datasets than the other tests, for both GO (Figure 1D) and KEGG (Figure 1E).

3.5. Comparison with the GOseq Method

The four datasets were analyzed using edgeR-GOseq pipeline. A total number of 1000 permutations has been used, and we have restricted the analysis to Biological process category in GO. The KEGG pathways analysis was done using the default values for the arguments in the package `GOseq`. Note that the package `GOseq` uses its own GO and KEGG gene set collections.

A total of 1318, 2834, 2605, and 1613 GO terms are differentially regulated in PRJNA411984 (2 pairs), PRJNA413956 (7 pairs), PRJNA218851 (18 pairs) and TCGA (50 pairs) datasets, respectively, and 415 were reported for all datasets (Figure 2A).

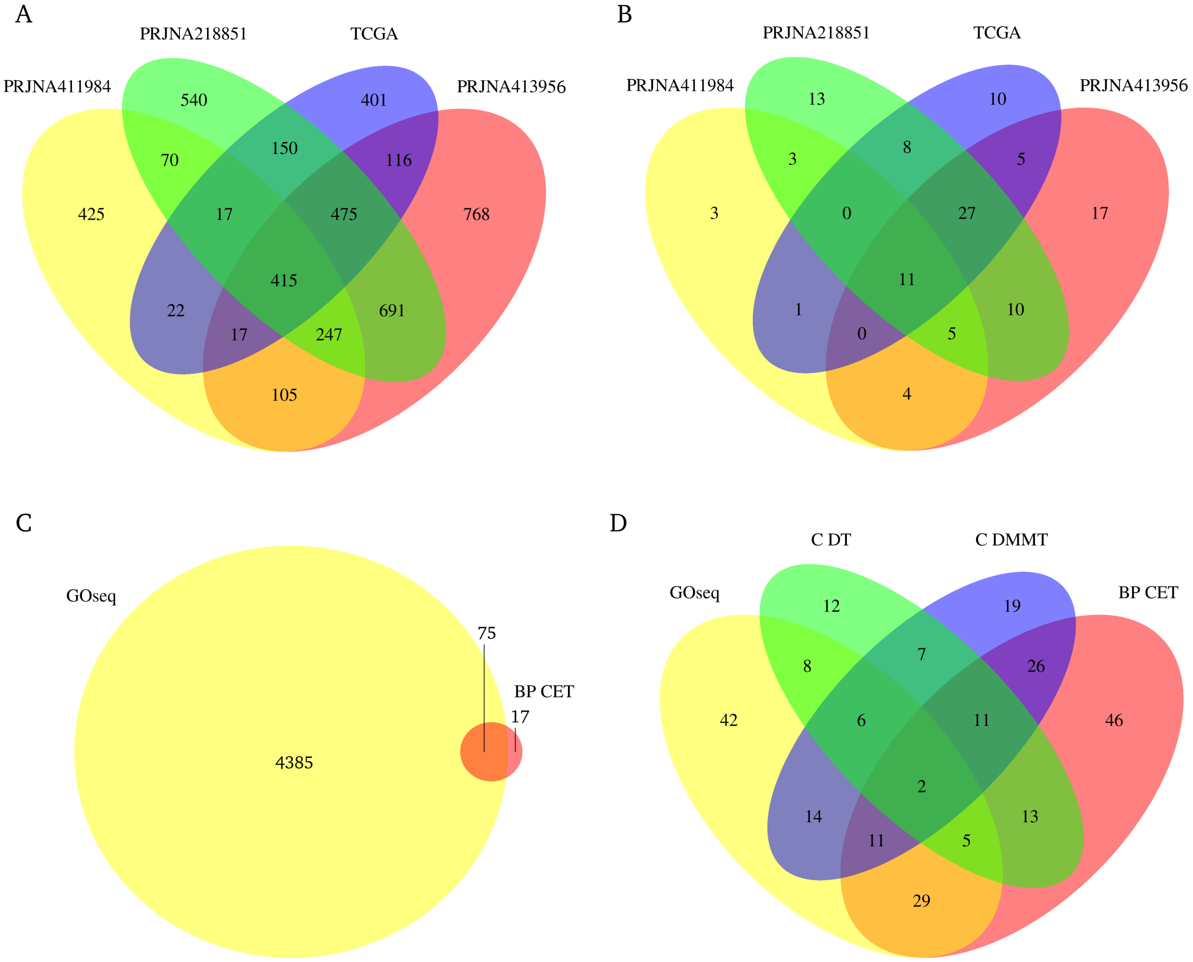

Figure 2. GO terms and KEGG pathways reported in common between datasets using GOseq and spatial tests. (**A**) GO terms reported by GOseq; (**B**) KEGG pathways reported by GOseq; (**C**) GO terms reported in common by GOseq and between-pair CET; (**D**) KEGG pathways reported in common by GOseq and between-pair CET, Complete DT, and Complete DMMT.

If we use the KEGG pathways: 27, 79, 77, and 62 gene sets are reported in PRJNA411984 (2 pairs), PRJNA413956 (7 pairs), PRJNA218851 (18 pairs), and TCGA (50 pairs) datasets, respectively. Of these, 11 gene sets are shared for all datasets (Figure 2B).

When comparing our results with those obtained by `GOseq`, we observe that, using the CET method with BP distribution, a large number of gene sets in common for GO terms (Figure 2C). In KEGG ontology, we also include the results obtained DMMT and DT with complete distribution in the comparison because these tests were appropriate for small sample datasets. The results indicate that there is high agreement between the implemented methodologies (Figure 2D).

4. Discussion

One realization proved to be enough to generate the second component of the bivariate point process because, in the case of GO groups, only 8% of the results differ when using one or multiple realizations (Figure 1A). For KEGG, by increasing the number of realizations, the number of gene sets declared as significant decreased; even so, most of the gene sets reported in MR were reported by OR. Furthermore, the biological results were consistent using only one realization. At the computational level, the use of one realization reduces the computing time. Regarding the randomization distribution, between-pair and complete were tested, and the results indicate that, when applying between-pair or complete distribution, the gene sets reported as significant changes depending on the spatial test used and also on the criteria to group genes (Tables 1 and 2).

If Cuzick and Edwards tests (CET) are used, then a greater agreement has been observed because all GO terms and most of the KEGG gene sets reported by complete distribution were also reported by between-pair distribution. Furthermore, the number of gene sets reported as significant decreases when using complete distribution (Figure 1C). It could be expected because we forget the original design, and the complete distribution produces a greater variability of the realizations. The same signal has been evaluated with a distribution with a higher variability. The biological results were consistent, reporting gene sets highly associated with CRC (Tables 1 and 2). The same results were observed when reducing the sample size to 18 and 7 pairs (Figure 1D,E), showing a high coincidence between the results obtained in each datasets. However, no significant gene sets were reported when using a 2-pair dataset. The power of our test is really small with such a small sample size.

For small samples (as just with two pairs), the DMMT and DT seem appropriate, since the biological results were consistent, particularly if genes are grouped using KEGG ontology (Figure 1C). For instance, significant KEGG pathways as Rap1 signaling pathway (hsa04015), Hepatocellular carcinoma (hsa05225), Thyroid cancer (hsa05216), Bladder cancer (hsa05219), or Acute myeloid leukemia (hsa05221) were reported by DMMT and DT using complete distribution on the PRJNA411984 dataset (see Supplementary Material). Thus, the results obtained through the proposed methodology were consistent at biological level, even though there are only two pairs of samples.

When comparing the results obtained in all datasets using BP-CET for GO terms and including C-DT and C-DMMT for KEGG pathways, with the results obtained by GOseq (Figure 2), we observed that there was a great coincidence between both methods. This indicates that, in biological terms, they are comparable. However, our approach is completely different and has many possible generalizations. This kind of coincidence is not clear for us. It could be a future line of research.

The spatial statistic is a well developed field of research. In this paper, we have tried to show how simple procedures taken from spatial statistics can be used with good results to test null hypotheses of the omics data field. A lot of different possibilities can be explored. No independence between genes needs to be assumed. We think that, except for such a small sample size like two pairs, the results are good for seven pairs. Obviously, they are better for fifty pairs. It seems that the method performs well with small sample sizes.

More complex experimental designs with more than one covariable (categorical or continuous) could be considered and the methodology adapted without great difficulty.

In our opinion, the complexity of the original data makes a valid simulation study difficult. However, a simulation study is included in the Supplementary Material. It shows the good performance of our methods. Additional comments can be found in the file.

5. Conclusions

Our method performs a gene set analysis without a previous marginal (per gene) differential expression analysis by taking into account the dependencies between genes.

The three tests (CET, DMMT and DT) were applied for the first time into the omics data context in order to evaluate the gene set differential expression analysis. The proposed methodology is able to efficiently report the biological processes associated with the pathology studied.

It is important to note that each statistical test shows complementary biological results, i.e., it is convenient to use all of them and to evaluate all results. An important contribution of this paper is to show how these spatial tests can deal with the well known problem of the low sample sizes assuming the interdependence between genes in the context of gene set analysis.

Supplementary Materials: The following are available online at https://www.mdpi.com/2227-7390/9/5/521/s1, the Data, gene set collections, proportion test, global analysis, edgeR-GOseq pipeline and Groups and test names.

Author Contributions: G.A. and F.M. propose the original idea. The theoretical setup was developed by G.A. The biological analysis has been done by A.L.R.-C. The software implementation was realized by G.A. The R package was built by G.A. All authors have read and agreed to the published version of the manuscript.

Funding: This paper has been partially supported by the Spanish grant DPI2017-87333-R (GA) and by the Chilean ANID/FONDECYT-POSTDOCTORADO N0. 3180486 (ALRC).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We are thankful for previous discussions with Amelia Simó (simo@uji.es).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Draghici, S. *Statistics and Data Analysis for Microarrays Using R and BioConductor*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2012.
2. Pevsner, J. *Bioinformatics and Functional Genomics*; Wiley-Blackwell: Hoboken, NJ, USA, 2009.
3. Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szcześniak, M.W.; Gaffney, D.J.; Elo, L.L.; Zhang, X.; et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **2016**, *17*, 1–19. [CrossRef] [PubMed]
4. Maleki, F.; Ovens, K.; Hogan, D.J.; Kusalik, A.J. Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front. Genet.* **2020**, *11*, 654. [CrossRef]
5. Consortium, T.G.O. Gene ontologie: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
6. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44*, D457–D462. [CrossRef] [PubMed]
7. de Leeuw, C.; Neale, B.; Heskes, T.; Posthuma, D. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* **2016**, *17*, 353–364. [CrossRef]
8. Tian, L.; Greenberg, S.A.; Kong, S.W.; Altschuler, J.; Kohane, I.S.; Park, P.J. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13544–13549. [CrossRef]
9. Ackermann, M.; Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC Bioinform.* **2009**, *10*, 47. [CrossRef]
10. Rahmatallah, Y.; Emmert-Streib, F.; Glazko, G. Gene set analysis approaches for RNA-seq data: Performance evaluation and application guideline. *Briefings Bioinform.* **2016**, *17*, 393–407. [CrossRef]
11. Young, M.D.; Wakefield, M.J.; Smyth, G.K.; Oshlack, A. Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.* **2010**, *11*, R14. [CrossRef]
12. Mootha, V.K.; Lindgren, C.; Eriksson, K.F.; Subramanian, A.; Sihag, S.; Lehar, J.; Puigserver, P.; Carlsson, E.; Ridderstråle, M.; Laurila, E.; et al. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **2003**, *34*, 267–273. [CrossRef]

13. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
14. Wang, X.; Cairns, M.J. SeqGSEA: A Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. *Bioinformatics* **2014**, *30*, 1777–1779. [CrossRef] [PubMed]
15. Goeman, J.J.; van de Geer, S.A.; de Kort, F.; van Houwelingen, H.C. A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics* **2004**, *20*, 93–99. [CrossRef]
16. Chen, Y.; Lun, A.T.L.; Smyth, G.K. From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* **2016**, *5*, 1438.
17. Law, C.W.; Alhamdoosh, M.; Su, S.; Dong, X.; Tian, L.; Smyth, G.K.; Ritchie, M.E. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research* **2016**, *5*. [CrossRef]
18. Diggle, P.J. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, 3rd ed.; CRC Press: Boca Raton, FL, USA, 2013.
19. Chiu, S.N.; Stoyan, D.; Kendall, W.S.; Mecke, J. *Stochastic Geometry and Its Applications*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2013.
20. Kal, A.J.; van Zonneveld, A.J.; Benes, V.; van den Berg, M.; Koerkamp, M.G.; Albermann, K.; Strack, N.; Ruijter, J.M.; Richter, A.; Dujon, B.; et al. Dynamics of Gene Expression Revealed by Comparison of Serial Analysis of Gene Expression Transcript Profiles from Yeast Grown on Two Different Carbon Sources. *Mol. Biol. Cell* **1999**, *10*, 1859–1872. [CrossRef]
21. Cuzick, J.; Edwards, R. Spatial Clustering for Inhomogeneus Populations. *J. R. Stat. Soc.* **1990**, *B52*, 73–104.
22. Diggle, P.; Morris, S.; Morton-Jones, T. Case-control isotonic regression for investigation of elevation in risk around a point source. *Stat. Med.* **1999**, *18*, 1605–1613. [CrossRef]
23. Diggle, P.J. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* **1990**, *153*, 349–362. [CrossRef]
24. Barnard, G. Contribution to the discussion of Professor Bartlett's paper. *J. R. Stat. Soc. B* **1963**, *25*, 294.
25. Yamada, A.; Yu, P.; Lin, W.; Okugawa, Y.; Boland, C.R.; Goel, A. A RNA-Sequencing approach for the identification of novel long non-coding RNA biomarkers in colorectal cancer. *Sci. Rep.* **2018**, *8*, 2–11. [CrossRef] [PubMed]
26. Li, M.; Zhao, L.; Li, S.; Li, J.; Gao, B.; Wang, F.; Wang, S.; Hu, X.; Cao, J.; Wang, G. Differentially expressed lncRNAs and mRNAs identified by NGS analysis in colorectal cancer patients. *Cancer Med.* **2018**, *7*, 4650–4664. [CrossRef]
27. Kim, S.K.; Kim, S.Y.; Kim, J.H.; Roh, S.A.; Cho, D.H.; Kim, Y.S.; Kim, J.C. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Mol. Oncol.* **2014**, *8*, 1653–1666. [CrossRef]
28. Robinson, M.D.; Smyth, G.K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **2008**, *9*, 321–332. [CrossRef] [PubMed]
29. Robinson, M.D.; Smyth, G.K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **2007**, *23*, 2881–2887. [CrossRef] [PubMed]
30. Intracrine VEGF signalling mediates colorectal cancer cell migration and invasion. *Br. J. Cancer* **2017**, *117*, 848–855. [CrossRef]
31. Farooqi, A.A.; de la Roche, M.; Djamgoz, M.B.; Siddik, Z.H. Overview of the oncogenic signaling pathways in colorectal cancer: Mechanistic insights. *Semin. Cancer Biol.* **2019**, *58*, 65–79. [CrossRef]
32. Koveitypour, Z.; Panahi, F.; Vakilian, M.; Peymani, M.; Seyed Forootan, F.; Nasr Esfahani, M.H.; Ghaedi, K. Signaling pathways involved in colorectal cancer progression. *Cell Biosci.* **2019**, *9*, 1–14. [CrossRef]
33. Soly, W.; Zhanjie, L.; Lunshan, W.; Xiaoren, Z. NF-κB signaling pathway, inflammation and colorectal cancer. *Chin. J. Cell. Mol. Immunol.* **2009**, *6*, 327–334. [CrossRef]
34. Sanchez-Vega, F.; Mina, M.; Marra, M.A. Pathways, Oncogenic Signaling Cancer, The Atlas, Genome. *Cell* **2019**, *173*, 321–337. [CrossRef]