MDPI

Article

Identification of driver epistatic gene pairs combining germline and somatic mutations in cancer

Jairo Rocha[1,2,*], Jaume Sastre[1], Emilia Amengual-Cladera[2], Jessica Hernandez-Rodriguez[2], Victor Asensio-Landa[2], Damià Heine-Suñer[2] and Emidio Capriotti[3]

[1] Department of Mathematics and Computer Science, University of the Balearic Islands, Palma de Majorca, Spain
[2] Health Research Institute of the Balearic Islands (Idisba), Palma de Majorca, Spain
[3] Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Bologna, Italy
* Correspondence: jairo@uib.es

Abstract: Cancer arises from the complex interplay of various factors. Traditionally, the identification of driver genes focuses primarily on the analysis of somatic mutations. We describe a new method for the detection of driver gene pairs based on an epistasis analysis that considers both germline and somatic variations. Specifically, the identification of significantly mutated gene pairs entails the calculation of a contingency table, wherein one of the co-mutated genes can exhibit a germline variant. By adopting this approach it is possible to select gene pairs in which the individual genes do not exhibit significant association with cancer. Finally, the survival analysis is used to select clinically relevant gene pairs. To test the efficacy of the new algorithm, we analyzed the Colon Adenocarcinoma (COAD) and Lung Adenocarcinoma (LUAD) samples available at The Cancer Genome Atlas. In the analysis of COAD and LUAD samples, we identified epistatic gene pairs significantly mutated in tumor tissue with respect to normal tissue. We believe that further analysis of the gene pairs detected by our method will unveil new biological insights enhancing a better description of the cancer mechanism.

Keywords: gene pairs; epistasis; cancer driver variations; contingency table; survival analysis; lung cancer; colon cancer.

Citation: Rocha, J.; Sastre, J.; Amengual-Cladera, E.; Hernandez-Rodriguez, J.; Asensio-Landa, V.; Heine-Suñer, D.; Capriotti, E. Identification of driver epistatic gene pairs combining germline and somatic mutations in cancer. *Int. J. Mol. Sci.* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

Cancer is a complex disease driven by several factors [1,2]. No single genetic factor alone can explain cancer onset, thus a multigenic mechanism should be taken into consideration [3,4]. In this context, the epistasis analysis allows the identification of gene pair interactions associated with the insurgence and progress of cancer.

One of the main challenges in current genomics is to identify the correlation between genotype and phenotype. GWAS (Genome Wide Association Analysis) experiments have attempted to establish this association between genotype and phenotype by relying on SNP (Single Nucleotide Polymorphism) data. However, there is only a small proportion of the phenotype that can be explained with common SNPs, raising the well-known debate about the "missing heritability" [5].

Usually SNPs are tested for their statistical relationship with a disease considering only additive effects. In other words, each SNP is presumed to independently contribute to the phenotype. Nevertheless, it has been shown that the missing heritability may arise from genetic variants that exhibit effects through interactions with one another [6,7]. Under this condition, the concept of epistasis becomes relevant, referring to the combinatorial effects of one or more genetic variants that can manifest even without any individual effect. We believe that complex traits would be better explained by considering the interaction between pairs of genes and variants. Our work aims to uncover novel gene pairs associated with the onset and progression of cancer.

Research consortiums such as The Cancer Genome Atlas (TCGA) [8], and the International Cancer Genome Consortium (ICGC) [9] have generated a huge amount of cancer genome data, which allowed to discover potential causative variants [10]. Different authors have reviewed the available computational methods for assessing the impact of mutations in the cancer genome [11–13].

Several methods have been used to discriminate between driver mutations, which provide a selective advantage to the cancerous cells and contribute to the onset of the disease [14–17], from passenger mutations, which have no pathogenic effect. The prevalent strategy to identify cancer driver genes works by detecting significantly over-mutated genes in tumors, which are more likely the drivers.

One potential strategy for identifying gene interactions can involve conducting exhaustive tests on all combinations of variants. Many software packages consider statistical interactions between loci, due to this complex mechanism that involves multiple genes. For example, Marchini's algorithm and PLINK [18,19] test for all two-locus interactions in a reasonable timeframe. PLINK has a specific option (*--epistatic*) to discover such pairs. However, the TCGA datasets consist of two samples per subject, encompassing both normal and tumor tissues. Consequently, the experimental design based on paired-data from the same subject challenges the hypothesis of the sample independence assumed by regression methods.

Like PLINK, BOOST [20] uses general linear regression models over all possible gene pairs. However, it adopts a fast approximation strategy that guarantees significant interactions are not filtered out. Nevertheless, interpreting the regression differences is not straightforward. This is because BOOST is specifically designed for the analysis of case/control data, which is not directly applicable to our problem that involves paired-data.

An alternative approach is MOSGWA [21], which addresses the selection of interacting genes as a variable selection problem using a modified version of the Bayesian Information Criterion (mBIC2). When compared with methods employing a similar strategy [22,23], MOSGWA resulted in a lower fraction of false positives.

A specific study of colorectal cancer [24] presented a simple, but powerful method for testing gene-gene interactions, reducing the number of pairs by considering only the variants with a certain marginal association with cancer. The authors discovered two significant interactions involving known loci and variants that exhibited marginal associations.

Other approaches by Vandin and colleagues [25,26] focused on the identification of cancer associated gene networks. GeneralizedHotNet [25] detects subnetworks of mutated genes within established or known interaction networks, correlating them with a given phenotype. The procedure is based on a heat diffusion process to obtain a measure of influence between pairs of genes in a protein-protein interaction (PPI) network. A two-stage statistical test scoring the association between mutated genes and clinical data is adopted for the selection of significant subnetworks. A more recent method, namely NoMAS [26], identifies subnetworks of a large gene-gene interaction network with mutations associated with survival time. NoMAS implements an efficient algorithm that leverages a color-coding technique and a log-rank statistical test to compare the survival of two specific populations.

In this study, we present a new method for the identification of gene pairs associated with cancer by comparing samples of normal and tumor tissues. Unlike previous methods developed by Vandin and colleagues [25,26], which rely on known protein-protein interaction (PPI) network, our approach does not assume any predefined gene pairs. Our algorithm represents paired data in a contingency table and, focusing on specific elements of the table, assesses the statistical significance of cases of mutated gene pairs, where at least one of the genes exhibit a somatic variant. To the best of our knowledge, our method is the first to consider the association between somatic and germline mutations. Furthermore, the analysis proposed in this study enables the identification of cancer-causing gene pairs, wherein each single gene by itself does not exhibit a significant association with cancer.

2. Results

Analyzing the Variant Calling Files, which include the detected variants in normal and tumor tissues, we calculated a 3x3 contingency table to identify patients holding relevant gene-pair mutations (RGPMs) corresponding to the states (s,s), (s,b), and (b,s). Indeed, we assume that epistasis effects in cancer can not only be attributed to pairs of somatic (s) variations, but it can also arise from the combination of a germline (b) and a somatic (s) variant within a given gene pair. This condition is equivalent to a congenital predisposition associated with a germline variant, wherein the interaction with a somatic variant confers to the cell a selective advantage.

To identify relevant epistatic interactions, we employed two criteria when evaluating all potential gene pairs. The first criterion involved selecting gene pairs that exhibited a significant number of patients with a RGPM in the contingency table. The second criterion entailed identifying gene pairs with a significant difference in overall survival (OS) time between subjects with RGPMs and those with a background single gene mutation (BSGM) corresponding to the states (w,b), (w,s), (b,w) and (s,w). The definition of gene pair mutation states and the procedure for calculating the contingency table are described in the Materials and Methods section.

Despite considering all gene pairs candidates, the procedure itself is remarkably fast. The output of the analysis is a comprehensive list of gene pairs along with their corresponding p-values, which show significant association with cancer after correcting for multiple testing hypothesis.

2.1. Colon Cancer

2.1.1. Detection of Epistatic Interactions

We applied our method to a TCGA dataset comprising 422 patients affected by COAD. Our analysis detected 358,774 variants affecting a protein sequence (VAPs) that appear in at least one individual. After the filtering procedure, 11,670 genes were retained and ∼68 million ($11,670 \times 11,669/2 = 68088615$) gene pair contingency table were calculated for further analysis. Among them, we selected ∼51 million (51,098,733) gene pairs for which the number of observed patients with any RGPM is higher than the relative the expected value. To keep the level of false positive rate below 5%, gene pairs with a p-value lower than 9.8×10^{-10} ($0.05/51,098,733$) were selected.

After this filtering procedure, 450 gene pairs passed the test (Supplementary File 1). On this subset, we performed the survival analysis with Bonferroni correction, searching for the gene pairs with an associate p-value lower than 1.1×10^{-4} ($0.05/450$).

Although no pairs pass this test, we found 16 gene pairs with a p-value under 0.05 (Supplementary Materials Section 1). For each selected gene pair, Table 1 presents the p-values and the hazard rate (HR) which measures the relative rate of deaths comparing two groups of patients. On average, for the 16 gene pairs, the group of individuals with RGPMs doubles the death rate of the group of patients with BSGMs.

Furthermore, we study the association between the two groups of patients and the clinical data provided by the TCGA consortium. In detail, we conducted regression analyses to detect the potential associations between tumor stage, gender, and age at initial diagnosis in relation to the patient class. However, no significant associations were found in these analyses (p-values not shown).

All contingency tables are analysed in less than 3 minutes at 2.6GHz with a program written in C. The survival analysis of the pairs that pass the first test is performed in less than a minute with a dedicated Python package [27].

To further characterize the type of diseases associated with the genes that are selected through the epistasis test, we performed an enrichment analysis of the genes within the selected pairs using EnrichR [28]. In Figure 1A we displayed the Jensen's disease entities [29] with the highest level of significance associated with the list of selected genes. It is evident that the most significant entities are related to different types of carcinoma. This observation underscores the effectiveness of our selection procedure in capturing

Table 1. Analysis of COAD samples. List of 16 gene pairs with a survival difference p-value lower than 0.05 sorted according to p-value for epistasis. The final columns indicate the hazard ratio (HR) and its corresponding confidence interval at a 95% level (HR 95% CI). SA p-value: Survival analysis p-value calculated comparing the overall survival rates of the groups of subjects with RGPMs and BSGMs. The contingency tables and survival curves for each gene pair in this table are reported in Supplementary Materials.

Gene1	Gene2	Epistasis p-value	SA p-value	HR	HR 95% CI
CCDC73	HTR2B	$< 1.04 \times 10^{-10}$	0.032	2.2	1.06 - 4.49
DDX4	KCNJ16	$< 1.04 \times 10^{-10}$	0.018	2.1	1.12 - 4.01
DDX4	SNX13	$< 1.04 \times 10^{-10}$	0.045	1.9	1.01 - 3.60
SEMG1	CYP2E1	$< 1.04 \times 10^{-10}$	0.033	1.6	1.04 - 2.43
TRIP12	BTAF1	2.1×10^{-10}	0.0042	2.7	1.34 - 5.45
ZNF99	HECTD2	3.1×10^{-10}	0.0065	2.1	1.21 - 3.54
LGR5	MBD5	4.1×10^{-10}	0.038	1.9	1.03 - 3.68
TOPORS	FLT1	4.1×10^{-10}	0.022	2.4	1.12 - 5.35
ABCA8	C1orf168	6.2×10^{-10}	0.044	1.5	1.01 - 2.22
MAGI3	KIF20A	6.2×10^{-10}	0.046	2.0	1.01 - 4.14
RASAL2	TRIM37	6.2×10^{-10}	0.014	2.8	1.20 - 6.55
PHLPP1	CLASP1	6.2×10^{-10}	0.021	2.2	1.11 - 4.46
ZNF491	BTAF1	6.2×10^{-10}	0.027	2.3	1.09 - 4.81
GTF3C3	CATSPERB	7.3×10^{-10}	0.011	1.9	1.15 - 3.12
MTNR1B	VIT	7.3×10^{-10}	0.046	1.7	1.01 - 2.71
ARHGAP20	ZBTB39	7.3×10^{-10}	0.036	2.0	1.04 - 3.82

crucial biological aspects associated with the mechanisms of cancer. The robustness of our procedure is demonstrated by the consistent results obtained even when the larger list of the top 1000 gene pairs is considered (Figure 1B).

Finally, we analyzed the contingency tables of the first two significant gene pairs (Table 2). Although these gene pairs do not meet the criteria for significance according to the Bonferroni test, their survival analysis p-value is lower than 0.05.

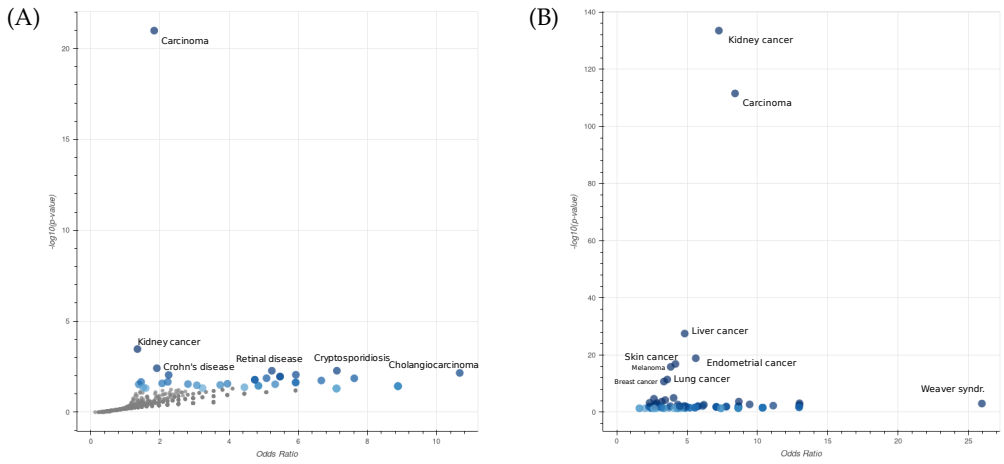

Figure 1. Enrichment analysis of Jensen's disease entities performed on a list of 444 genes involved in 450 significant epistatic interactions (A), and a second list of 782 genes involved in the top 1000 epistatic interactions (B) in COAD. The axes x and y are the odds ratios and p-values of the disease enrichment, respectively. The entity Carcinoma describes a family of diseases characterized by abnormal proliferation of epithelial cells.

Table 2. Contingency tables of the CCDC73-HTR2B (A) and DDX4-KCNJ16 (B) gene pairs. The mutation states of each gene (w, s and b) are defined in Materials and Methods. The number of patients holding relevant gene-pair mutations (RGPMs) correspond to the contingency table elements (s,b), (b,s) and (s,s).

(A) CCDC73-HTR2B

g_1/g_2	w	s	b
w	359	1	18
s	7	**8**	**2**
b	27	**0**	0

(B) DDX4-KCNJ16

g_1/g_2	w	s	b
w	282	3	23
s	0	**6**	**1**
b	100	**1**	6

For the gene pair CCDC73-HTR2B (Table 2A), the number of observed subjects for the entry (s,s), corresponding to the case where both genes hold somatic mutations, is 8 which is significantly higher than the expected value of 0.36. As shown in Figure 2A, the average overall survival time of the 10 patients with relevant gene-pair mutations (RGPM), is 212 days shorter than the value calculated for the subjects with a background single gene mutation (BSGM). For this gene pair, the survival analysis p-value is 0.03. Similarly, for the gene pair DDX4-KCNJ16 (Table 2B), the number of subjects observed for the entry (s,s) is 6, which is significantly above the expected value of 0.17. For the group of 8 patients holding a RGPM, the average overall survival time is 355 days shorter than the value calculated on the set of individuals with a BSGM (Figure 2B). For this gene pair, the survival analysis p-value is 0.05. Figure 3 shows the VAPs that passed the filters and generate the mutations in the four genes.

We aim to discover pairs of loci that are not known to act together in the tumor cells. However, some evidence may already be present elsewhere. We will observe some relationships with gene function in public databases, and with protein interaction networks.

2.1.2. Gene and protein knowledge-based analysis

To generate new hypotheses regarding the cancer mechanism, we performed an analysis of the functions of the identified genes and their corresponding protein interaction network. A previous study has shown that lung cancer patients with a high transcription level of CCDC73 gene tend to have a good prognosis [31]. Another study found a relationship between the expression of the serotonin 2B (HTR2B) receptor and human uveal

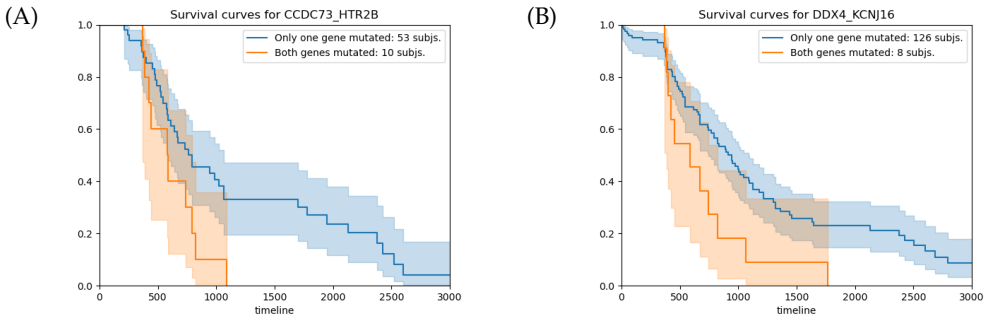

Figure 2. (A) Survival analysis for the CCDC73-HTR2B gene pair. The average overall survival times for the groups with RGPMs and BSGMs are 791 and 579 days, respectively, with a relative p-value of 0.03. (B) Survival analysis for the DDX4-KCNJ16 gene pair. The average overall survival times for the groups with RGPMs and BSGMs are 943 and 588 days, respectively, with a relative p-value of 0.02. The orange and blue curves represent the groups of subjects with RGPMs and BSGMs, respectively. RPGMs: (b,s), (s,b) and (s,s). BSGMs: (w,b), (w,s), (b,w) and (s,w). Mutation states are: w (no mutation), s (somatic) and b (germline).

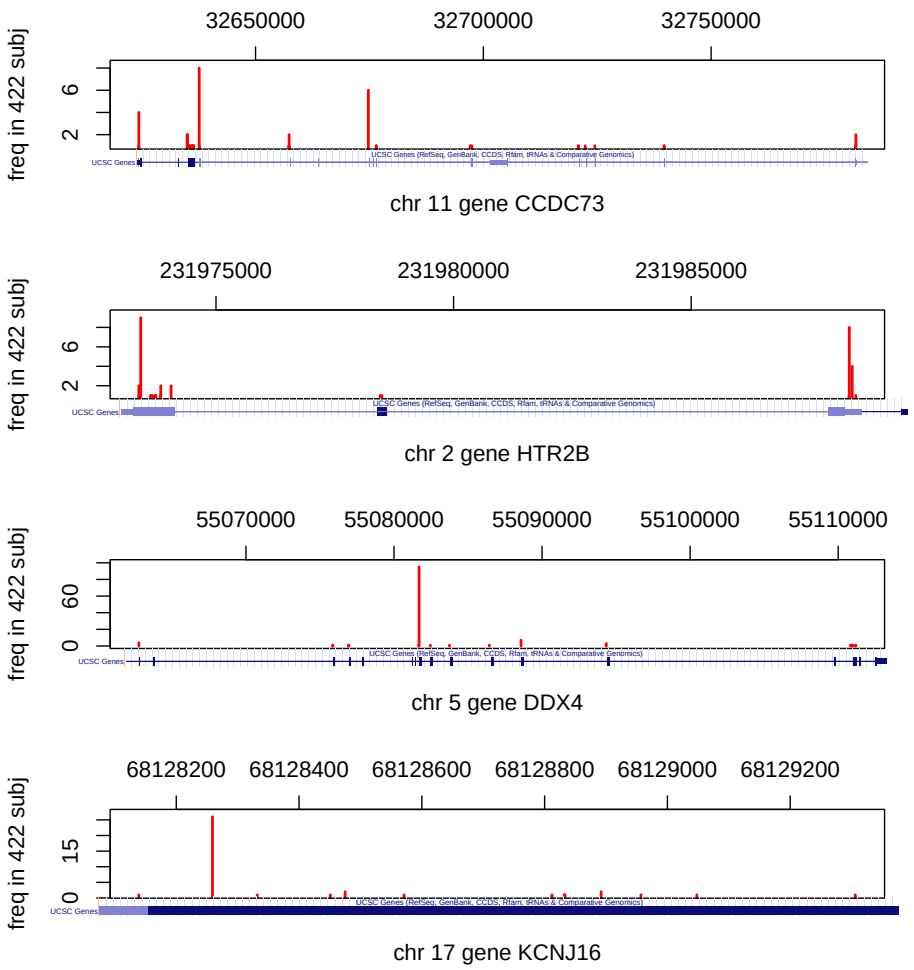

Figure 3. Mutation landscape in genes CCDC73, HTR2B, DDX4 and KCNJ16. Only the VAPs that pass the filters are shown. We can see where the actual VAP positions in each gene are. The gene structures are shown below each figure in a blue axis [30]. Mutations of gene KCNJ16 are in the short range of 1.2 Kb, which explains why the depicted gene structure corresponds exclusively to the exonic region.

melanoma [32]. Nevertheless, no relationship with colon cancer has been found until now. Based on previous studies, DDX4 has already been recognized as a potential molecular target for chemotherapy due to its involvement in regulating cell cycle progression in various somatic-derived-blood cancer cells [33]. Furthermore, KCNJ16 is one of the 154 signature genes that could be used to differentiate carcinoma types [34] although no direct association with colon cancer has been reported to date. Therefore, our approach was able to identify two potential candidate gene pairs that could be proposed as molecular targets for the development of colon cancer therapies.

In the database of known and predicted protein-protein interaction STRING [35], neither the genes CCDC73 and HTR2B nor DDX4 and KCNJ16 interact, not even at a low confidence level. In Wikipathways [36], HTR2B, DDX4 and KCNJ16 do not share pathways, while CCDC73 is not present. HTR2B, DDX4 and KCNJ16 do not share pathways.

Therefore, the interactions between the gene pairs identified in this work are currently unknown and require further investigation.

Table 3. Gene pairs for lung cancer that have a significance of the survival difference with a p-value under 0.05 among the top 100 pairs according to the epistasis significance. The final columns indicate the hazard ratio (HR) and its corresponding confidence interval at a 95% level (HR 95% CI). SA p-value: Survival analysis p-value calculated comparing the overall survival rates of the groups of subjects with RGPMs and BSGMs. The contingency tables and survival curves for each gene pair in this table are reported in Supplementary Materials.

Gene1	Gene2	Epistasis p-value	SA p-value	HR	HR 95% CI
CHRM2	SLC6A15	1.2×10^{-8}	0.042	1.9	1.14 - 3.60
PSD2	SUGP1	3.5×10^{-7}	0.006	2.6	1.21 - 5.66
UBR1	ACAD9	5.2×10^{-7}	0.031	2.2	1.04 - 4.51

2.2. Lung Cancer

2.2.1. Detection of Epistatic Interactions

As a second case study, we applied our approach to the analysis of a dataset of 405 subjects with lung adenocarcinoma (LUAD) identifying 192,304 VAPs. After filtering procedure, 6,854 mutated genes were retained for further analysis. Considering 6,854 mutated genes, we calculated the contingency table for ∼23 million ($6,854 \times 6,853/2 = 23,485,231$) gene pairs. Among them, 10,961,208 have at least one RGPM with a number of observations above the expected number. To keep a false positive rate below 0.05, a p-value threshold of $5 \times 10^{-9} (0.05/10,961,208)$ should be considered.

With this p-value cut-off only four gene pairs pass the filter but none of them pass the survival difference significance test. Given the limited number of significant gene pairs, we relax the p-value cut-off to recover other possible pairs. We consider the top 100 pairs (Supplementary File 2).

For each of the top 100 pairs (epistasis p-value $< 7.32 \times 10^{-7}$), the p-value of the differences in the overall survival time is calculated. With p-value threshold of $5.0 \times 10^{-4} (0.05/100)$ corrected for Bonferroni multiple testing hypothesis, no gene pairs pass this second control. Although no pairs pass this test, we found three gene pairs with a p-value under 0.05 (Table 3 and Supplementary Materials Section 2). For each selected gene pair, Table 3 presents the p-values and the hazard rate (HR) which measures the relative rate of deaths comparing two groups of patients. Similar to the observations in COAD, in

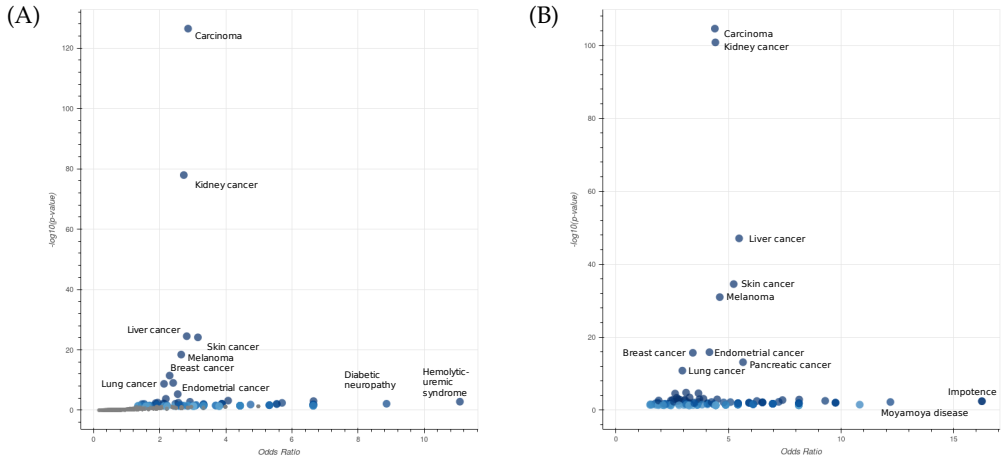

Figure 4. Enrichment analysis of Jensen's disease entities performed on a list of 115 genes involved in 61 interactions with a survival significance under 0.05 among the top 1000 epistatic interactions(A), and a second list of 1220 genes involved in all top 1000 epistatic interactions (B) in LUAD. The axes x and y are the odds ratios and p-values of the disease enrichment, respectively. The entity Carcinoma describes a family of diseases characterized by abnormal proliferation of epithelial cells.

Table 4. Contingency tables of the CHRM2-SLC6A15 (A) and PSD2-SUGP1 (B) gene pairs. The mutation states of each gene (w, s and b) are defined in Materials and Methods. The relevant gene-pair mutations (RGPM) correspond to the contingency table elements (s, b), (b, s) and (s, s).

(A) CHRM2-SLC6A15

g_1/g_2	w	s	b
w	267	0	114
s	6	**5**	**4**
b	7	**0**	2

(B) PSD2-SUGP1

g_1/g_2	w	s	b
w	300	1	80
s	1	**3**	**4**
b	13	**0**	3

the case of LUAD, the group of individuals with RGPMs on average doubles the death rate of the other group of patients.

For characterizing the type of diseases associated with the selected gene pairs, we performed an enrichment analysis by using EnrichR [28]. In Figure 4A we displayed the Jensen's disease entities [29] with the highest level of significance associated with the list of selected genes. Also in the case of lung cancer, the most significant entities are related to different types of carcinoma. The robustness of our algorithm is demonstrated by the consistent results obtained even when the larger list of the top 1000 gene pairs is considered (Figure 4B).

Let's examine the gene pairs CHRM2-SLC6A15 (Table 4A) and PSD2-SUGP1 (Table 4B) which are the top epistatic pairs that pass the survival significance threshold.

The number of observed subjects for the entry (s, s) of the gene pair CHRM2-SLC6A15 is 5, which is significantly above the expected value of 0.19. For this gene pair, the epistatic p-value is 1.2×10^{-8}. Analysing the contingency table of the gene pair PSD2-SUGP1, we found that for (s, s) the observed and expected subjects are 3 and 0.08 respectively, while for (s, b) the observed and expected subjects are 4 and 1.72 respectively. The epistatic p-value for the PSD2-SUGP1 gene pair is 3.52×10^{-7}.

The survival analysis for the CHRM2-SLC6A15 gene pair shows that the subjects with RGPMs have 137 fewer days of survival on average than the subjects with BSGMs (Figure 5A). This difference of the overall survival time correspond to a p-value of 0.042. For the PSD2-SUGP1 gene pair, the average overall survival time difference between the patients with RGPMs and those with BSGM is 421 days (Figure 5B) which corresponds to a p-value

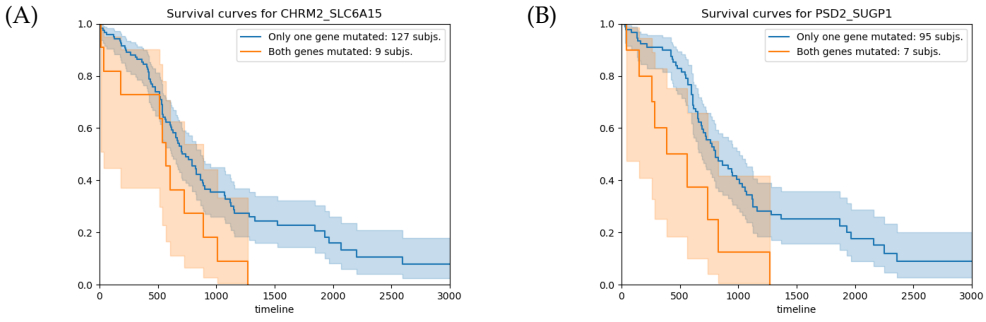

Figure 5. (A) Survival analysis for the CHRM2-SLC6A15 gene pair. The average overall survival times for the groups with RGPMs and BSGMs are 705 and 568 days, respectively, with a relative p-value of 0.042. (B) Survival analysis for the PSD2-SUGP1 gene pair. The average overall survival times for the groups with RGPMs and BSGMs are 806 and 385 days, respectively, with a relative p-value of 0.006. The orange and blue curves represent the groups of subjects with RGPMs and BSGMs, respectively. RPGMs: (b, s), (s, b) and (s, s). BSGMs: (w, b), (w, s), (b, w) and (s, w). Mutation states are: w (no mutation), s (somatic) and b (germline).

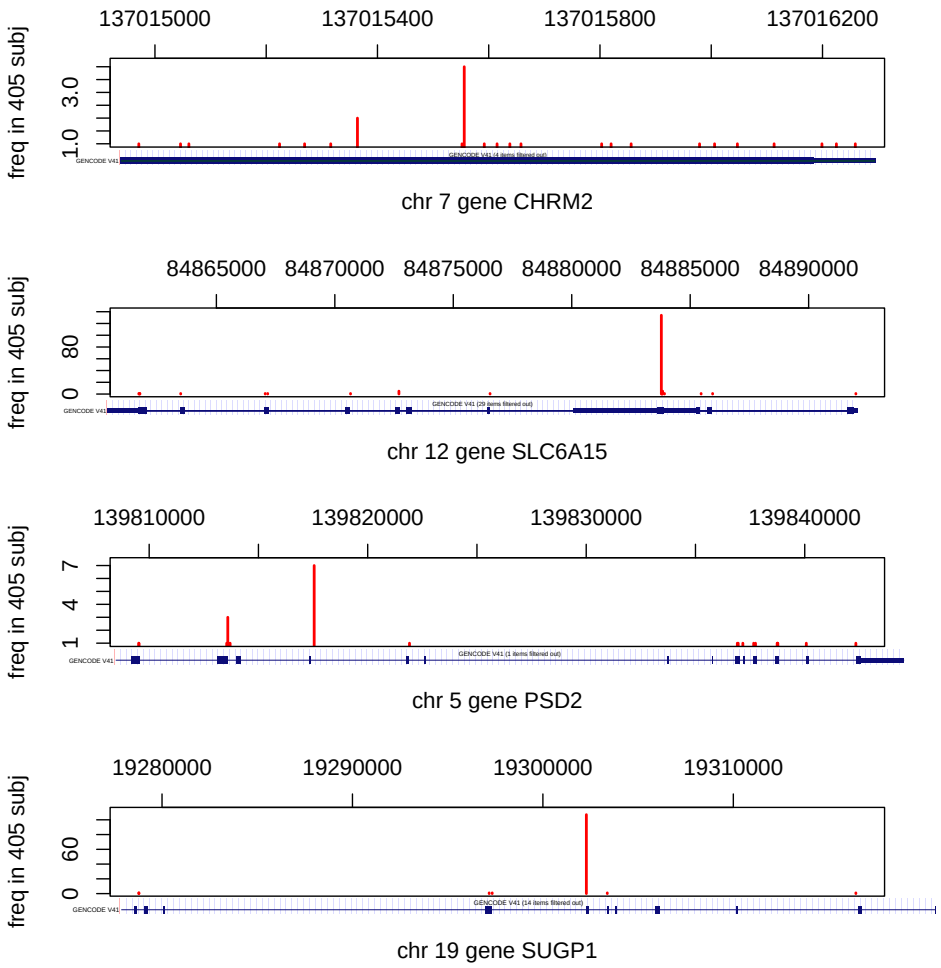

Figure 6. Mutation landscape in genes CHRM2, SLC6A15, PSD2 and SUGP1. The gene structure is shown below each gene figure. Mutations of gene CHRM2 are in the short range of 1.3 Kb, which is why the whole gene structure shown corresponds to an exonic region.

of 0.006. Although these p-values are lower than 0.05, nevertheless they do not pass the control for multiple testing hypothesis (q-value $< 5.0 \times 10^{-4}$).

Figure 6 shows the VAP positions that passed the filters and generate the mutation of the pair of genes (CHRM2, SLC6A15) and (PSD2, SUGP1).

2.2.2. Gene and protein knowledge-based analysis

To gain new insight about the mechanism of lung adenocarcinoma we analyzed the existing information on four genes forming the top two gene pairs. CHRM2 has been associated with nicotine dependence [37] while SLC6A15 with depression disorders [38]. Furthermore, SUGP1 has been recently linked to lung cancer[39] and PSD2, to neurological diseases according to GeneCards. These findings constitute an initial validation of the association between the selected gene pairs and lung cancer, providing early evidence to support their potential significance.

In STRING, none of the pairs have an interaction, not even at the low confidence level, nor a common interacting protein at a low confidence level. Looking in Wikipathways, the two gene pairs have no pathways in common. Thus, the cancer interactions between the two top gene pairs detected with our approach are not yet known.

3. Conclusion and Discussion

This research proposes a new method to identify gene-pairs potentially associated with cancer. In particular, our approach is based on the analysis of paired normal and tumor samples from the same patient and exclude already known driver genes. The procedure developed to detect epistatic gene pairs is effective and fast, as the analysis of the contingency tables that we defined is rigorous and straightforward. The new statistical framework, based on the analysis of specific table cells, correctly characterizes the epistatic gene pairs and can be used in case of dependencies among paired samples from normal and tumor tissues.

In detail, we present an analysis of four epistatic gene pairs detected in colon and lung cancer. Although, no protein-protein interactions between these gene pairs have been identified, the results of the survival and the disease enrichment analysis are promising. The enrichment analysis of diseases suggests that the process presented allows the detection of genes associated to cancer, and candidate epistaic pairs can be validated combining this results with clinical data.

In conclusion, the identification of epistatic gene pairs enables the formulation of new hypotheses on cancer mechanism that should be validated through targeted experimental assays.

4. Materials and Methods

4.1. Statistical Analysis of Genomic Association

The statistical analysis designed for detecting significantly mutated gene pairs relies on the calculation of a contingency table. Given a gene g in a patient, it can assume three possible states:

- w: the gene is mutated neither in normal and tumor tissues ;
- s: the gene is mutated only in the tumor tissue (somatic mutation);
- b: the gene is mutated in both normal and tumor tissues (germline mutation).

In this study, we focus on the patients holding the following relevant gene-pair mutations (RGPMs):

- (s, s): both genes present somatic mutation being mutated only in tumor tissue;
- (s, b), (b, s): one gene carries a germline mutation (both in normal and tumor tissues) and the second gene presents a somatic variant (only in tumor tissue).

Considering the three possible gene states $i, j \in \{w, s, b\}$, the following notations can be adopted for calculating the contingency table:

- o_{ij} is the number of observed subjects in the cell (i, j),
- e_{ij} is the expected number of subjects in the cell (i, j),
- m_i is the row marginal frequency,
- n_j is the column marginal frequency, and
- N is the total number of subjects.

For each pair of genes g_1 and g_2, we focus on the elements of the contingency table accounting for the three case studies mentioned above:

g_1/g_2	w	s	b
w			
s		o_{ss}	o_{sb}
b		o_{bs}	

To identify potential gene interactions, an independence test on the contingency table could be conducted. However, our specific interest lies in testing the dependency exhibited by the three pre-defined elements. For this purpose, we perform a test by considering contingency tables with identical marginal frequencies, while fixing the three cells, aiming to find gene pairs with minimum dependency.

Given that the number of degrees of freedom is 4, fixing the value of the three cells, only one degree of freedom is left. Consequently, our strategy consists in finding the optimal cell value that minimizes the G-test statistic. This is a minimization problem over a single variable, subject to the non-negativity constraints of all the table cells.

If we define x as the new value of o_{bb}, then the other five cell values can be calculated as a function of x, and the statistic of the following table is a function of x:

g_1/g_2	w	s	b
w	$c_3 + x$	o_{ws}	$c_1 - x$
s	o_{sw}	o_{ss}	o_{sb}
b	$c_2 - x$	o_{bs}	x

where $c_1 = n_b - o_{sb}$, $c_2 = m_b - o_{bs}$ and $c_3 = n_w - m_s + o_{ss} + o_{sb} - m_b + o_{bs}$ are the constant values fixed by the marginal constraints.

Notice that under the null hypothesis of independence, the values e_{ij} are defined only on the marginals and therefore do not depend on x. Also, the values of o_{sw} and o_{ws} are constant as the other two cells in the same row or column are constant.

The cells must be positive or zero and satisfy the following positivity constraints:

$$\begin{aligned} o_{bb}(x) &= \quad x \geq 0 \\ o_{wb}(x) &= \quad c_1 - x \geq 0 \\ o_{bw}(x) &= \quad c_2 - x \geq 0 \\ o_{ww}(x) &= \quad c_3 + x \geq 0, \end{aligned}$$

The G statistic can be defined as follows

$$G(x) = 2 \sum_{ij} o_{ij}(x) \ln \frac{o_{ij}(x)}{e_{ij}}$$

After some simplifications, its differential can be expressed as:

$$G'(x) = 2 \ln \frac{(c_1 + x)x}{(c_2 - x)(c_3 - x)}.$$

Thus, $G'(x)$ is null when

$$x^* = \frac{c_1 c_2}{c_1 + c_2 + c_3}.$$

The second differential has four terms all of which positive under the positivity constraints. Consequently, the function $G(x)$ is strictly convex and the x^* value corresponds to a minimum.

The denominator does not present any singularities because

$$\begin{aligned} c_3 &= -o_{sw} + n_w - m_b + o_{bs} \\ &= -o_{sw} + n_w - o_{bb} - o_{bw} \\ &= o_{ww} - o_{bb}, \end{aligned}$$

and,

$$\begin{aligned} c_1 + c_2 + c_3 &= n_b - o_{sb} + m_b - o_{bs} + o_{ww} - o_{bb} \\ &= o_{wb} + m_b - o_{bs} + o_{ww} \\ &= o_{wb} + o_{bw} + o_{bb} + o_{ww}, \end{aligned}$$

If the denominator in the equation for x^* is 0, all the four corner of the table are 0, and the remaining table values are fixed. In such cases, we set $x^* = o_{bb} = 0$.

The positivity constraints are satisfied by x^* which is obtained by is non-negative numerator and denominator. Furthermore, $x^* \leq c_1$ since $c_1 + c_3 = o_{wb} + o_{ww} \geq 0$ and symmetrically, $x^* \leq c_2$.

For c_3 the following conditions are valid: If $c_3 \geq 0$, then $x^* \geq -c_3$.

If $c_3 < 0$, then $x^* \geq -c_3$, but in the latter case the proof is more technical. First, we observe that $o_{bb} = o_{ww} - c_3$, thus, $c_1 = o_{wb} + o_{bb} = o_{wb} + o_{ww} - c_3$, and $c_1 \geq -c_3$. Similarly, $c_2 \geq -c_3$. If we fix $c_3 < 0$ and define the function,

$$f(c_1, c_2) = \frac{c_1 c_2}{c_1 + c_2 + c_3},$$

then

$$\nabla f = \left(\frac{c_2(c_2 + c_3)}{(c_1 + c_2 + c_3)^2}, \frac{c_1(c_1 + c_3)}{(c_1 + c_2 + c_3)^2} \right).$$

The gradient is null when $c_1 = c_2 = -c_3$, and it has only non-negative components in the quadrant $c_1, c_2 \geq -c_3$. Therefore, f has a minimum at $(-c_3, -c_3)$ with value $f(-c_3, -c_3) = -c_3$, and

$$f(c_1, c_2) \geq -c_3.$$

In summary, we have a value x^* that generates a new valid contingency table with a minimum $G(x^*)$ for any table. We take this value as a measure of dependency of the three relevant cells we have focused on. To calculate the p-value, we generate random permutations of subjects for one of the genes and calculate the values of the minimum statistic for ∼10 billion tables. The whole process ∼10 hours on a 2.6GHz computer with 16 processors.

The mutation dependency of two genes can be related to observing more or fewer subjects than expected in a cell. We assume that tumors are caused by the accumulation of mutations. Thus, we retain statistically significant gene pairs where at least one of the three cells has more observations than expected.

We select the gene pairs that pass the Bonferroni correction with the FWER (family-wise error rate) Q of 0.05, and a total number of experiments equal to the number of considered contingency tables.

4.2. Survival Analysis

We analysed the survival data available at the TCGA to select possible clinically relevant epistatic gene pairs. We used overall survival (OS) time and not the disease free interval (DFI) time because the latter one is related to the cancer recurrence and therefore it would be more appropriate for a study on treatment effectiveness. Since our study refers to the overall malignity of the mutations, we considered the OS time.

For each gene pair, we categorized the patients in two groups based on their gene mutation state (w, s, b) and compare their OS times. The first group consists of patients holding one of the three RGPMs (s, b), (b, s) and (s, s). The second group comprises individuals who have either a germline or somatic variant in only one gene of the pair. The mutation states of these individuals, corresponding to the elements (w, b), (w, s), (b, w) and (s, w) of the contingency table, are referred to as background single gene mutations (BSGMs).

To estimate the survival function, we use the Kaplan-Meier Estimate and the Cox Hazard Model available in the `lifelines` [27] package for Python.

From the gene pairs that show significant gene association, we select the ones that show a significant difference in the survival expectancy between double-mutated and non-double-mutated, with a Bonferroni FWER q-value of 0.05.

For each gene pair, we also consider the two subject groups defined above to check for dependency with clinical variables like pathological tumor stage, gender and age at initial of pathological diagnosis. We use logistic regression to find association between the group and the clinical variable. There are 10 levels for the tumor stage that were modelled numerically; the "discrepancy" labels were also changed to "Not Available".

Algorithm Overview

The algorithm can be summarized as follows:

For each pair of genes:

1. Calculate the 3×3 contingency table for the s, b, n values
2. **If** $o_{ij} > e_{ij}$ at some of the three cells $\{(s,s),(b,s),(s,b)\}$:
    2.1. Calculate x^* and the new contingency table
    2.2. Perform the G-test for the new table and calculate the p-value
3. Select the pairs that pass the Bonferroni FWER control
4. Apply the survival analysis and select gene pairs with significantly lower OS time

4.3. Dataset Composition and Processing

We collected two datasets one consisting by samples of Colon Adenocarcinoma (COAD) and the other composed by sample of Lung Adenocarcinoma (LUAD), both of which were released by TCGA consortium. For each patient, only one pair of samples were considered. The two dataset consist of $N = 422$ and $N = 405$ unique paired samples (normal/tumor) for COAD and LUAD respectively. The VCF files associated to each sample were obtained considering GRCh37 and GRCh38 reference human genomes [40] for COAD and LUAD respectively. The annotation of mutations was performed using Annovar [41].

We collected the variants affecting a protein sequence (VAPs) in each sample: nonsynonymous Single Nucleotide Variants, frameshift-deletion, frameshift-insertion, stop-gain, stop-loss, nonframeshift-deletion and nonframeshift-insertion. We assumed that any VAP can impair the gene function.

The considered VAPs should either be labelled as "PASS" in the "Filter" column of the VCF or meet the specific criteria for the read depth (DP) in support of the alternative allele. In detail, we filtered out the VAPs whose DP was lower than 10 reads and with a fraction of reads supporting the alternative allele (alternative alleles supporting read divided by read depth) lower than 10%. Additionally, we excluded from the analysis the VAPs with an excess of mutated cases in tumor tissue with respect to normal because they usually correspond to sequencing artefacts. To reduce the impact of these artefacts, we focused only on somatic mutations that exceed the number of cases with a germline mutation by three subjects or fewer.

A gene is considered mutated with respect to the reference genome if it contains any type of mutation at any locus. To reduce the number of gene tested, we filtered out all the genes with less than 5% of the mutated subjects in a tumor tissue.

In addition, all pseudogenes, all genes associated with olfactory receptors, and the macro-gene TTN (titin) were eliminated.

To identify pairs of genes which are not individual associated with cancer, we excluded for our analysis all the driver genes for the Lung or Colon Adenocarcinoma reported by Intogen [42].

Among the 422 patients affected by COAD, the OS time is available for 403 of them, while the information on 93 individuals is censored. In the case of LUAD cohort, which comprises 405 subjects, the OS time is available for 392 patients, while data for 115 individuals is censored.

Supplementary Materials: All the programs and data generated are available at `https://github.com/jairo-rocha55/epistasis` .

Author Contributions: Conceptualization, J.R., J.S. and E.C.; methodology, J.R., J.S., E.A., D.H. and E.C.; validation, E.A., J.H., V.A. and D.H.; writing, J.R. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been supported by the grant PID2021-126114NB-C44 funded by MCIN/AEI/10.13039/501100011033 and by ERDF, A way of making Europe. Also, by the projects

PI13/02778 and PI18/00847 from Carlos III Health Institute (ISCIII) and by the grant 423/C/2015 from Fundació La Marató TV3. J.H. is funded by a fellowship from IDISBA (Folium Program-INTRES Project, AETIB annual plan 2019). E.A. was funded by IdISBA, program TALENT-Plus TECH funded by the sustainable tourism tax from the Govern de les Illes Balears (AETIB).

This work was also supported by the PRIN project, "Integrative tools for defining the molecular basis of the diseases: Computational and Experimental methods for Protein Variant Interpretation" of the "Ministero Istruzione, Università e Ricerca" [201744NR8S].

Acknowledgments: We thank the reviewers for their helpful comments that allowed us to improve the quality of this paper. We acknowledge The Cancer Genome Atlas (TCGA) consortium for allowing access to the whole-exome sequencing data for colon and lung adenocarcinomas.

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

References

1. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A.; Kinzler, K.W. Cancer genome landscapes. *Science* **2013**, *339*, 1546–1558. https://doi.org/10.1126/science.1235122.
2. Alexandrov, L.B.; Nik-Zainal, S.; Wedge, D.C.; Aparicio, S.A.J.R.; Behjati, S.; Biankin, A.V.; Bignell, G.R.; Bolli, N.; Borg, A.; Børresen-Dale, A.L.; et al. Signatures of mutational processes in human cancer. *Nature*, *500*, 415–421. https://doi.org/10.1038/nature12477.
3. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **2011**, *144*, 646–674. https://doi.org/10.1016/j.cell.2011.02.013.
4. Alexandrov, L.B.; Kim, J.; Haradhvala, N.J.; Huang, M.N.; Tian Ng, A.W.; Wu, Y.; Boot, A.; Covington, K.R.; Gordenin, D.A.; Bergstrom, E.N.; et al. The repertoire of mutational signatures in human cancer. *Nature* **2020**, *578*, 94–101. https://doi.org/10.1038/s41586-020-1943-3.
5. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; et al. Finding the missing heritability of complex diseases. *Nature* **2009**, *461*, 747–753. https://doi.org/10.1038/nature08494.
6. Eichler, E.E.; Flint, J.; Gibson, G.; Kong, A.; Leal, S.M.; Moore, J.H.; Nadeau, J.H. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **2010**, *11*, 446–450. https://doi.org/10.1038/nrg2809.
7. Capriotti, E.; Ozturk, K.; Carter, H. Integrating molecular networks with genetic variant interpretation for precision medicine. *Wiley Interdiscip Rev Syst Biol Med* **2019**, *11*, e1443. https://doi.org/10.1002/wsbm.1443.
8. The Cancer Genome Atlas Program (TCGA). https://www.cancer.gov/tcga, 2022.
9. International Cancer Genome Consortium.; Hudson, T.J.; Anderson, W.; Artez, A.; Barker, A.D.; Bell, C.; Bernabé, R.R.; Bhan, M.K.; Calvo, F.; Eerola, I.; et al. International network of cancer genome projects. *Nature* **2010**, *464*, 993–998. https://doi.org/10.1038/nature08987.
10. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.
11. Tian, R.; Basu, M.K.; Capriotti, E. Computational methods and resources for the interpretation of genomic variants in cancer. *BMC Genomics* **2015**, *16 Suppl 8*, S7. https://doi.org/10.1186/1471-2164-16-S8-S7.
12. Petrosino, M.; Novak, L.; Pasquo, A.; Chiaraluce, R.; Turina, P.; Capriotti, E.; Consalvi, V. Analysis and Interpretation of the Impact of Missense Variants in Cancer. *Int J Mol Sci* **2021**, *22*, 5416. https://doi.org/10.3390/ijms22115416.
13. Cortés-Ciriano, I.; Gulhan, D.C.; Lee, J.J.K.; Melloni, G.E.M.; Park, P.J. Computational analysis of cancer genome sequencing data. *Nat Rev Genet* **2022**, *23*, 298–314. https://doi.org/10.1038/s41576-021-00431-y.
14. Gonzalez-Perez, A.; Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res* **2012**, *40*, e169. https://doi.org/10.1093/nar/gks743.
15. Lawrence, M.S.; Stojanov, P.; Polak, P.; Kryukov, G.V.; Cibulskis, K.; Sivachenko, A.; Carter, S.L.; Stewart, C.; Mermel, C.H.; Roberts, S.A.; et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **2013**, *499*, 214–218. https://doi.org/10.1038/nature12213.
16. Khurana, E.; Fu, Y.; Colonna, V.; Mu, X.J.; Kang, H.M.; Lappalainen, T.; Sboner, A.; Lochovsky, L.; Chen, J.; Harmanci, A.; et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **2013**, *342*, 1235587. https://doi.org/10.1126/science.1235587.

17. Tian, R.; Basu, M.K.; Capriotti, E. ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples. *Bioinformatics* **2014**, *30*, i572–578. https://doi.org/10.1093/bioinformatics/btu466.
18. Marchini, J.; Donnelly, P.; Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* **2005**, *37*, 413–417. https://doi.org/10.1038/ng1537.
19. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **2007**, *81*, 559–575. https://doi.org/10.1086/519795.
20. Wan, X.; Yang, C.; Yang, Q.; Xue, H.; Fan, X.; Tang, N.L.S.; Yu, W. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* **2010**, *87*, 325–340. https://doi.org/10.1016/j.ajhg.2010.07.021.
21. Dolejsi, E.; Bodenstorfer, B.; Frommlet, F. Analyzing genome-wide association studies with an FDR controlling modification of the Bayesian Information Criterion. *PLoS One* **2014**, *9*, e103322. https://doi.org/10.1371/journal.pone.0103322.
22. Hoggart, C.J.; Whittaker, J.C.; De Iorio, M.; Balding, D.J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet* **2008**, *4*, e1000130. https://doi.org/10.1371/journal.pgen.1000130.
23. He, Q.; Lin, D.Y. A variable selection method for genome-wide association studies. *Bioinformatics* **2011**, *27*, 1–8. https://doi.org/10.1093/bioinformatics/btq600.
24. Jiao, S.; Hsu, L.; Berndt, S.; Bézieau, S.; Brenner, H.; Buchanan, D.; Caan, B.J.; Campbell, P.T.; Carlson, C.S.; Casey, G.; et al. Genome-wide search for gene-gene interactions in colorectal cancer. *PLoS One* **2012**, *7*, e52535. https://doi.org/10.1371/journal.pone.0052535.
25. Vandin, F.; Clay, P.; Upfal, E.; Raphael, B.J. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac Symp Biocomput* **2012**, pp. 55–66. https://doi.org/10.1142/9789814366496_0006.
26. Altieri, F.; Hansen, T.V.; Vandin, F. NoMAS: A Computational Approach to Find Mutated Subnetworks Associated With Survival in Genome-Wide Cancer Studies. *Front Genet* **2019**, *10*, 265. https://doi.org/10.3389/fgene.2019.00265.
27. Davidson-Pilon, C. lifelines: survival analysis in Python. *Journal of Open Source Software* **2019**, *4*, 1317. https://doi.org/10.21105/joss.01317.
28. Xie, Z.; Bailey, A.; Kuleshov, M.V.; Clarke, D.J.B.; Evangelista, J.E.; Jenkins, S.L.; Lachmann, A.; Wojciechowicz, M.L.; Kropiwnicki, E.; Jagodnik, K.M.; et al. Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **2021**, *1*, e90. https://doi.org/10.1002/cpz1.90.
29. Pletscher-Frankild, S.; Pallejà, A.; Tsafou, K.; Binder, J.X.; Jensen, L.J. DISEASES: text mining and data integration of disease-gene associations. *Methods* **2015**, *74*, 83–89. https://doi.org/10.1016/j.ymeth.2014.11.020.
30. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. *Genome Res* **2002**, *12*, 996–1006. https://doi.org/10.1101/gr.229102.
31. Zhu, L.; Chen, P.; Wang, H.; Zhao, L.; Guo, H.; Jiang, M.; Zhao, S.; Li, W.; Zhu, J.; Yu, J.; et al. Analysis of prognostic and therapeutic values of drug resistance-related genes in the lung cancer microenvironment. *Transl Cancer Res* **2022**, *11*, 339–357. https://doi.org/10.21037/tcr-21-1841.
32. Benhassine, M.; Le-Bel, G.; Guérin, S.L. Contribution of the STAT Family of Transcription Factors to the Expression of the Serotonin 2B (HTR2B) Receptor in Human Uveal Melanoma. *Int J Mol Sci* **2022**, *23*, 1564. https://doi.org/10.3390/ijms23031564.
33. Schudrowitz, N.; Takagi, S.; Wessel, G.M.; Yajima, M. Germline factor DDX4 functions in blood-derived cancer cell phenotypes. *Cancer Sci* **2017**, *108*, 1612–1619. https://doi.org/10.1111/cas.13299.
34. Xu, Q.; Chen, J.; Ni, S.; Tan, C.; Xu, M.; Dong, L.; Yuan, L.; Wang, Q.; Du, X. Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. *Mod Pathol* **2016**, *29*, 546–556. https://doi.org/10.1038/modpathol.2016.60.
35. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **2019**, *47*, D607–D613. https://doi.org/10.1093/nar/gky1131.
36. Kelder, T.; Pico, A.R.; Hanspers, K.; van Iersel, M.P.; Evelo, C.; Conklin, B.R. Mining Biological Pathways Using WikiPathways Web Services. *PLoS One* **2009**, *4*, e6447. https://doi.org/10.1371/journal.pone.0006447.

37. Mobascher, A.; Rujescu, D.; Mittelstraß, K.; Giegling, I.; Lamina, C.; Nitz, B.; Brenner, H.; Fehr, C.; Breitling, L.P.; Gallinat, J.; et al. Association of a variant in the muscarinic acetylcholine receptor 2 gene (CHRM2) with nicotine addiction. *Am J Med Genet B Neuropsychiatr Genet* **2010**, *153B*, 684–690. https://doi.org/10.1002/ajmg.b.31011.
38. Kohli, M.A.; Lucae, S.; Saemann, P.G.; Schmidt, M.V.; Demirkan, A.; Hek, K.; Czamara, D.; Alexander, M.; Salyakina, D.; Ripke, S.; et al. The neuronal transporter gene SLC6A15 confers risk to major depression. *Neuron* **2011**, *70*, 252–265. https://doi.org/10.1016/j.neuron.2011.04.005.
39. Alsafadi, S.; Dayot, S.; Tarin, M.; Houy, A.; Bellanger, D.; Cornella, M.; Wassef, M.; Waterfall, J.J.; Lehnert, E.; Roman-Roman, S.; et al. Genetic alterations of SUGP1 mimic mutant-SF3B1 splice pattern in lung adenocarcinoma and other cancers. *Oncogene* **2021**, *40*, 85–96. https://doi.org/10.1038/s41388-020-01507-5.
40. 1000 Genomes Project Consortium.; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. https://doi.org/10.1038/nature15393.
41. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **2010**, *38*, e164. https://doi.org/10.1093/nar/gkq603.
42. Gonzalez-Perez, A.; Perez-Llamas, C.; Deu-Pons, J.; Tamborero, D.; Schroeder, M.P.; Jene-Sanz, A.; Santos, A.; Lopez-Bigas, N. IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* **2013**, *10*, 1081–1082. https://doi.org/10.1038/nmeth.2642.