1-1-2017

ProcessDriver: A Computational Pipeline to Identify Copy Number Drivers and Associated Disrupted Biological Processes in Cancer

Brittany Baur
Marquette University

Serdar Bozdag
Marquette University, serdar.bozdag@marquette.edu

Accepted version. *Genomics,* Vol. 109, Nos. 3-4 (July 2017): 233-240. DOI.

Genomics, Vol. 109, No. 3-4 (July 2017): 233-240. DOI.

ProcessDriver: A computational pipeline to identify copy number drivers and associated disrupted biological processes in cancer

Brittany Baur
Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI
Serdar Bozdag
Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI

Abstract

Copy number amplifications and deletions that are recurrent in cancer samples harbor genes that confer a fitness advantage to cancer tumor proliferation and survival. One important challenge in computational biology is to separate the causal (i.e., driver) genes from passenger genes in large, aberrated regions. Many previous studies focus on the genes within the aberration (i.e., *cis genes*), but do not utilize the genes that are outside of the aberrated region and dysregulated as a result of the aberration (i.e., *trans genes*). We propose a computational pipeline, called ProcessDriver, that prioritizes candidate drivers by relating *cis* genes to dysregulated trans genes and biological processes. ProcessDriver is based on the assumption that a driver *cis* gene should be closely associated with the dysregulated *trans* genes and biological processes, as

opposed to previous studies that assume a driver *cis* gene should be the most correlated gene to the copy number of an aberrated region. We applied our method on breast, bladder and ovarian cancer data from the Cancer Genome Atlas database. Our results included previously known driver genes and cancer genes, as well as potentially novel driver genes. Additionally, many genes in the final set of drivers were linked to new tumor events after initial treatment using survival analysis. Our results highlight the importance of selecting driver genes based on their widespread downstream effects in *trans*.

Keywords

Copy number, Copy number driver, Biological process, Gene expression

1. Introduction

Copy number amplifications and deletions that are recurrent in cancer samples harbor driver genes that confer a fitness advantage to cancer tumor proliferation and survival [1]. Passenger genes that do not have a selective advantage are amplified or deleted along with the drivers due to their proximity to the driver and as a result, have similar changes in expression with respect to copy number. Due to their similar copy number and expression profiles, separating drivers from passengers is an important and difficult challenge.

One of the tools to compute significant recurrent copy number alterations in a given set of samples is GISTIC. GISTIC relies on copy number data to detect regions of the genome that harbor likely drivers [2], [3]. GISTIC leveraged the notion that a region containing a driver gene should be altered significantly more than expected by chance. This method has proven useful in identifying regions that likely harbor candidate driver genes. However, it is difficult to distinguish passengers from drivers in large regions based on copy number data alone.

Some studies have integrated copy number and gene expression data to determine the effects of copy number on gene expression for genes within a copy number aberration, known as cis genes [4], [5], [6], [7]. The underlying assumption is that driver genes will have a more altered expression due to a copy number aberration than passenger genes. For example, Oncodrive-CIS is a method to score the cis genes as drivers by comparing the gene expression of samples with the aberration to the gene expression of samples without the aberration [4]. The strength of the correlation between copy number and gene expression is also used to detect drivers [6], [7].

Some studies have identified drivers by taking into account the wider impact of a driver on downstream target genes located outside of the aberration, known as trans genes. For instance, Akavia et al. had the underlying assumption that copy number influences the driver gene expression, which in turn alters the expression of a group of downstream trans genes [8]. Aure et al. determined which cis genes were highly correlated to their own copy number [9]. The authors then determined which of these cis genes played a network perturbing role in cancer through expression correlation to all other genes.

Certain biological processes are known to be disrupted in cancer such as apoptosis and cell cycle [10]. Therefore, identifying modules of cis and trans genes based on biological processes would allow for additional insight into the specific biological processes that the driver disrupts. Additionally, a driver cis gene changes the pathology of the cell and therefore influences the expression of many other genes in trans. Therefore, the cis genes in the module can also be narrowed down to a set of likely drivers based on the strength of the association of the cis genes with the downstream trans genes, as opposed to the strength of a cis gene's association with its own copy number.

In this study, we proposed a pipelined called ProcessDriver that detects driver cis genes, associated trans genes and disturbed biological processes. We first find all of the differentially expressed cis and trans genes with respect to an aberration. For a given aberration, the pipeline creates modules of differentially expressed cis

genes and differentially expressed trans genes based on biological processes. The module is subject to further refinements to determine likely drivers from the cis genes based on the relationship between cis gene expression and trans gene expression. The pipeline is therefore able to determine which biological processes and trans genes are dysregulated by the driver gene. We found that our selected drivers were more enriched in cancer genes and were associated with a higher risk of new tumor events after initial treatment. Additionally, consistent with previous studies [4], [5], [6], [7], we found that the selected drivers were more correlated with their own copy number.

2. Materials and methods

2.1. ProcessDriver

We implemented a computational pipeline called ProcessDriver in R to compute candidate copy-number driven driver genes by relating cis genes to dysregulated trans genes and biological processes. ProcessDriver utilizes gene expression, copy number alteration data and GO database. ProcessDriver consists of two main steps, namely GO term enrichment step and driver selection step. The entire pipeline of ProcessDriver is illustrated in Fig. 1. In what follows, we describe each main step of ProcessDriver. The source code for ProcessDriver is freely available at www.github.com/brittanybaur/ProcessDriver.

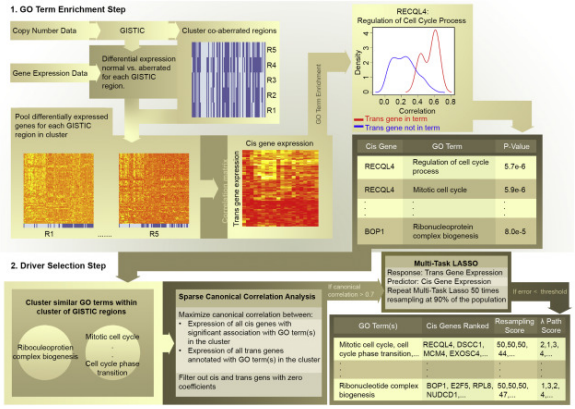

Fig. 1. Flowchart of ProcessDriver. In the GO term association step, cis and trans genes that were differentially expressed with respect to a copy number aberration were computed. Each cis gene was associated with up to ten biological processes by performing a Kolmogorov-Smirnov test using the correlation between the expression of the cis gene and every trans gene as a score. In the driver selection step, a GO term module containing similar GO terms and associated cis and trans genes was formed. The sparse canonical correlation analysis and multi-task LASSO were performed to narrow down potential drivers of the biological processes in the module from the cis genes.

2.1.1. GO term enrichment step

The GO term enrichment step first identifies differentially expressed cis and trans genes for a given aberration. Next, cis genes are associated with biological processes through the trans genes.

2.1.1.1. Computing GISTIC regions and differential expressed genes on GISTIC regions

GISTIC 2.0 was used to detect significant recurrent somatic copy number alterations (GISTIC regions hereafter) [2]. A GISTIC region with a log2 ratio above 0.1 was considered amplified, and a GISTIC region with a log2 ratio below − 0.1 was considered deleted. A confidence level of 0.75 was used to calculate the GISTIC region. The differential expression analysis was performed using DESeq2 for each GISTIC region between samples with no significant deletions or amplifications versus amplified or deleted samples [11] (p-value < 0.001).

Genes were considered differentially expressed with respect to an aberration if their adjusted *p*-value was < 0.001 in DESeq2 in one or more of the GISTIC regions within an aberration. Aberrations with > 50 differentially expressed genes were considered. These are aberrations of interest suitable for our algorithm because of the widespread effects of the aberration in trans, as well as the need to determine which cis genes are drivers. Batch effects were taken into account using the TCGA batch IDs as a covariate in DESeq2.

2.1.1.2. Clustering GISTIC regions into aberrations

To account for co-occurring aberrations, GISTIC regions were clustered together such that more similar regions were considered as a single aberration containing the individual GISTIC regions. Throughout the rest of the manuscript, a cluster of GISTIC regions will be referred to as an aberration. To cluster GISTIC regions into aberrations, a distance matrix was calculated where each entry was 1 minus the Pearson correlation of the copy number of two different GISTIC regions across all samples. Hierarchical clustering was performed on the distance matrix using average linkage using the stats package in R and the resulting dendrogram was cut at half of the maximum distance between the inter-cluster pairs.

The set of differentially expressed genes as determined by DESeq2 for each GISTIC region within the aberration were pooled together. Aberrations with > 50 differentially expressed genes were considered. These are aberrations of interest suitable for our algorithm because of the widespread effects of the aberration in trans, as well as the need to determine which cis genes are drivers. For each aberration, a differentially expressed gene is hereafter called *cis gene* if its chromosomal position was within a GISTIC region of that aberration, or called *trans gene* otherwise.

2.1.1.3. Computing aberration-adjusted expression

In the remaining steps of ProcessDriver algorithm, we related expression changes between cis genes and trans genes beyond the effects of copy number aberration. Both cis and trans genes expression are potentially under the influence of the copy number aberration of interest to varying degrees, and possibly other copy number aberrations in cis and trans. Due to the confounding effects of copy number aberration on gene expression, correlation between all gene expression will be high, making it difficult to establish relationships based solely on gene expression. To alleviate these copy number effects on gene expression, we computed *aberration-adjusted expression.* First we computed the variance stabilizing rlog transformation of the RNA-seq data. Then we applied principal component regression (PCR) between a gene's expression as a response and the copy number of all the GISTIC regions as predictors. The aberration-adjusted expression was the residual expression after PCR. We chose the PCR method as it is a suitable model to address the multicollinearity issue between the copy numbers of the GISTIC regions. All the remaining steps in ProcessDriver used the aberration-adjusted expression data.

2.1.1.4. GO term association

To link cis genes in aberrations to possible dysregulated biological processes in trans, each cis gene was associated with up to ten GO biological process terms through the trans genes. For a given aberration, the correlation between each cis gene's expression and each of the trans gene's expression in that aberration was calculated. A cis gene's correlation to all trans genes was used as a score in a Kolmogorov-Smirnov (KS) test to determine significant GO terms using the TopGO package in R [12]. The KS test examined whether trans genes annotated with a particular GO term were more correlated to the cis gene than trans genes not related to that GO term. KS test repeated for each cis gene in each aberration and up to ten GO terms with p-value < 0.05 were chosen to be associated with each cis gene.

2.1.2. Driver selection step

The driver selection step clusters cis and trans genes to form modules based on associated biological processes. Next, expression data are utilized in a sparse canonical correlation analysis to filter cis and trans genes with canonical correlation > 0.7. Finally, cis genes are ranked as drivers using two multi-task LASSO-based methods.

2.1.2.1. Clustering of significant GO terms into GO modules

Since some of the GO terms are semantically similar to each other and closely related in the GO term hierarchy, for each aberration, the set of GO terms associated with the cis genes were clustered using the getTermSim function with the relevance measure in the GOSim package in R [13]. For each GO term cluster, we defined *GO term module* as the collection of cis genes that were significantly associated with at least one GO term in that GO term cluster, and the trans genes that were annotated with at least one GO term in that GO term cluster.

2.1.2.2. Applying sparse canonical correlation analysis to refine GO term modules

To further refine a GO term module to determine likely drivers, we performed sparse canonical correlation analysis (SCCA) between the expression of p cis genes and the expression of K trans genes [14]. Let $X_{ij}$and Y_{ij} be the expression for patient i for cis and trans gene j, respectively. The goal of CCA is to maximize the canonical correlation, ρ, between two groups of variables X and Y, by finding a linear combination Yu and Xvcalled canonical variates, where $u = (u_1, \dots, u_K)$, $v = (v_1, \dots, v_p)$, are weight vectors [15]:

$$\rho = \frac{v'X'Yu}{\sqrt{v'X'Xv}\sqrt{u'Y'Yu}} \quad (1)$$

SCCA maximizes the correlation [Eq. 1] while also applying penalties to $u$and v such that some of the weights become zero resulting in $q < p$ cis genes and $M < K$ trans genes [14].

If the canonical correlation was > 0.7, cis and trans genes that had non-zero coefficients were left in the GO term module while those with zero coefficients were removed. If the canonical correlation was < 0.7, the module was no longer considered.

2.1.2.3. Applying multi-task LASSO to compute driver cis genes

Multi-task LASSO was performed with the expression of the remaining trans genes as a response and the expression of the remaining cis genes as the predictors to rank the cis genes based on their influence on trans gene expression. Let X and Y now represent the remaining q cis and M trans gene expression, respectively. Multi-task LASSO is the multi-response version of LASSO [16]. Friedman et al., defines the multi-task LASSO model [Eq. 2] for q cis genes, M trans genes and N patients as:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{(q+1)\times M}} \frac{1}{2N}\sum_{i=1}^{N}\left\|Y_{i,1:M} - \beta_0 - \beta^T X_{i,1:q}\right\|_F^2 + \lambda\sum_{j=1}^{q}\left\|\beta_j\right\|_2 \quad (2)$$

In Eq. 2, $Y_{i,1:M}$ is a vector corresponding to the expression values of the trans genes in patient i and $X_{i,1:q}$ is the covariate vector of cis genes. β_j is the jth row of the $q \times M$ coefficient matrix corresponding to jth cis gene and λ is the tuning parameter controlling the strength of the penalty.

We ranked cis genes as drivers based on the order of appearance of each of the cis gene predictors in the model as λ goes from largest to smallest. As λgets smaller, more cis genes will be non-zero and included in the model. The multi-task sharing portion involves which variables are selected. For each variable, a separate coefficient is

fit for each response, resulting in the $q + 1 \times M$ coefficient matrix [16]. Therefore, for all the trans genes, the coefficient for a given cis gene is either zero or non-zero, although the value of the non-zero coefficients will vary between trans genes. Therefore, this ranking will be the same for every trans gene, regardless of the non-zero coefficient value for the included cis genes

As an additional ranking system, the multi-task LASSO was rerun fifty times, each time resampling 90% of the samples without replacement. For a single resample, the value of λ used was the simplest model where the cross-validation error was within one standard error of the minimum cross-validation error. The number of times a cis gene was selected out of fifty resamples was used as a system to rank cis genes within the module. This ranking system would identify potential drivers that are robust to sample variation.

2.2. Datasets to assess ProcessDriver

To assess the performance of ProcessDriver, we used Illumina HiSeq 2000 RNA sequencing and level 3 segmented copy number inferred from Affymetrix Genome-Wide Human SNP 6.0 copy number data were downloaded for 92 luminal A breast cancer samples, 120 ovarian cancer samples and 120 bladder cancer samples from the Cancer Genome Atlas (TCGA) repository [17], [18], [19]. The TCGA IDs used in this study are provided in Supplemental Table 1.

3. Results

We downloaded RNA-seq and segmented copy number data from the TCGA repository for 92 luminal A breast cancer, 120 bladder cancer and 120 ovarian cancer samples. A summary of the thresholds used for the parameters described throughout Section 2 is provided in Supplemental Table 2. We used GISTIC 2.0 to identify recurrent copy number aberrated GISTIC regions using segmented copy number data from each cancer type and clustered them into aberrations (see Materials and Methods section). For breast, ovarian and bladder cancer, 175, 116 and 156 GISTIC regions were clustered into 66, 82 and 79 aberrations, respectively. DESeq2 was used to compute differentially expressed cis and trans genes for each aberration. Supplemental Tables 3–5 contain information about the cytoband locations and number of differentially expressed cis and trans genes for the aberrations considered in each cancer type.

For each cis gene in each aberration, associated dysregulated GO biological process terms were computed (Section 2.1.1). For each aberration, GO term modules were formed (Section 2.1.2.1 and then the cis and trans genes were filtered with SCCA (Section 2.1.2.2). Finally, the cis genes were ranked as likely drivers with two multi-task LASSO-based ranking methods (Section 2.1.2.3). The number of GO terms, and the average number of cis and trans genes per module before and after SCCA are summarized in Supplemental Table 6.

In the following sections, to evaluate the performance of ProcessDriver, we categorize cis genes into various groups namely, multiple driver, driver, semi-driver, last in λ path, and filtered. A *driver gene* is a cis gene that was selected 50 out of 50 times during resampling of multi-task LASSO and appears as the first gene in the λ path in at least one GO term module. A *multiple driver gene* is a gene that was selected as a driver in more than one GO term module. A cis gene that is *last in the λ path* is a gene that was selected last in λ path in every GO term module it appeared in. A *semi-driver*was never selected as a driver gene, but was not last in λ path in at least one module. A cis gene that in the *filtered* group was filtered because the canonical correlation of the GO term module was < 0.7 (Supplemental Fig. 1) or its coefficient was 0 in a GO term module with canonical correlation > 0.7, and otherwise never appeared in the multi-task LASSO phase (Section 2.1.2.3).

For comparison purposes, we imitated some of the existing methods and selected drivers based solely on the magnitude of correlation between their gene expression and their copy number. For each GO module, cis genes

with highest correlation between their expression and copy number were selected as *top correlated* group. This group served to highlight the differences between methods that take into account the relationship between trans gene expression and cis gene expression to select drivers and existing methods that selected drivers based on gene expression correlation to cis copy number.

3.1. Multiple drivers are enriched in known cancer genes

Table 1 lists the entire multiple driver genes computed by ProcessDriver using breast cancer data and Supplemental Tables 7 and 8 lists the multiple driver genes in ovarian and bladder cancer, respectively. For breast cancer, 19 out of 44 of the multiple driver genes were associated with cancer in the literature using the tool OncoSearch [20] as one or more publications describe their involvement in a cancer. Additionally, we found articles associating five more genes with cancer [21], [22], [23], [24], [25]. Seven multiple drivers were known cancer genes in the AGCOH or intOgen database [26], [27]. Additionally, we used the BioGRID database to find genes that the multiple driver interacts with and then determined which of the interacting genes are cancer genes in the AGCOH or intOgen database [28]. Overall, our results indicate that 27 out of the 44 breast cancer multiple drivers are a likely cancer gene or interact with a known cancer gene.

Table 1. Multiple driver genes in breast cancer. The GO terms column indicates the GO terms that the multiple driver is associated with through the trans genes in ProcessDriver. The number of articles column lists the number of articles found with OncoSearch tool indicating the multiple driver's involvement in cancer as a biomarker, tumor suppressor or oncogene [20]. Some additional literature references were found manually [21], [22], [23], [24], [25]. For cancer type (CT) column, BC – breast cancer, C – cancer based on the supporting literature, * indicates the multiple driver is cancer gene in AGCOH [26] or intOgen databases [27]. The number of cancer gene interactions column indicates the number of cancer genes in AGCOH or intOgen databases that interact with the multiple driver gene.

Multiple driver gene	GO terms	# articles	CT	# cancer gene interactions
AURKA	Mitotic cell cycle, cell cycle	71	BC*	8
SMARCB1	Macromolecule metabolic process, RNA biosynthetic process	59	C*	24
ADAM17	Positive regulation of cellular process, positive regulation of nucleobase-containing compound	32	BC	–
TRADD	Purine nucleoside metabolic process	10	C	4
CUL5	Carbohydrate metabolic process, nucleobase-containing compound metabolic process	6	BC	3
ELAC2	Cellular component organization, macromolecular complex assembly	5	C*	1
PSMA7	Cellular protein metabolic process, cellular macromolecule metabolic process, mitotic cell cycle process	5	C	4
RBM5	Cellular response to endogenous stimulus	5	BC*	–
COPS3	Cellular component organization, cellular component biogenesis	4	C	7
TBX21	T cell receptor signaling pathway, immune system process	2	C	3
APPBP2	Cell cycle process, cellular protein localization	1	C	2
ARFGAP1	Cellular protein metabolic process, gene expression, mitotic cell cycle process	1	C*	–

BOP1	Ribosome biogenesis, ribonucleoprotein complex biogenesis	1	C*	–
DDT	Macromolecule metabolic process, RNA biosynthetic process	1	C	–
HAGH	Organelle organization, regulation of RNA metabolic process	1	C	–
MED17	Carbohydrate metabolic process, cellular response to stress, cellular response to DNA damage stimulus	1	C*	7
PTDSS1	G2/M transition of mitotic cell cycle	1	C	–
RBM38	Hemostasis, wound healing, regulation of protein metabolic process	1	C	–
RRS1	Mitotic cell cycle, regulation of protein complex assembly	1	C	–
DIDO1	Phosphorus metabolic process, phosphorylation	Ref [21]	C	2
EIF4ENIF1	Macromolecule metabolic process, RNA biosynthetic process	Ref [22]	C	–
DSCC1	Mitotic cell cycle, cell cycle phase transition	Ref [23]	C	–
AZIN1	Cellular cation homeostasis, cellular ion homeostasis	Ref [24]	C	2
BCL2L13	Gene expression, macromolecule localization	Ref [25]	C	–
COG4	Organic substance metabolic process, nucleobase-containing compound metabolic process	–	–	1
PSMD7	Cellular response to stress, cellular response to DNA damage stimulus	–	–	1
AMDHD2	Transcription, DNA-templated, RNA biosynthetic process	–	–	–
C8orf55	Regulation of apoptotic process	–	–	–
C8orfk29	Nucleotide metabolic process	–	–	–
CCDC64B	Negative regulation of macromolecule biological process, regulation of macromolecule biosynthetic process	–	–	–
COG6	Regulation of cellular metabolic process, regulation of nitrogen compound metabolic process	–	–	–
DCUN1D5	Protein transport, macromolecule localization, establishment of protein localization	–	–	–
DDTL	Organic substance metabolic process, macromolecule catabolic process	–	–	–
DNTTIP1	Cellular protein catabolic process, proteolysis involved in cellular protein…	–	–	–
DUS2L	Organic substance metabolic process, RNA processing	–	–	–
DYNC1LI2	Organonitrogen compound catabolic process, macromolecule biosynthetic process	–	–	–
GPR172A	Organelle assembly, organelle organization, mitotic nuclear division	–	–	–
KARS	Metabolic process, viral process, symbiosis	–	–	–
KIAA1731	Mitotic cell cycle process, DNA metabolic process	–	–	–

KIFC2	Carbohydrate derivative catabolic process, nucleoside catabolic process	–	–	–
NAT15	Gene expression, regulation of gene expression	–	–	–
OSBPL2	Phosphate-containing compound metabolic …, negative regulation of biological process	–	–	–
RHOT2	Organelle organization, protein complex assembly, transcription, DNA-templated	–	–	–
STX8	Intracellular protein transport, intracellular transport	–	–	–

For ovarian cancer, 18 out of 33 multiple driver genes were associated with cancer through the literature or an interactor with a known cancer gene (Supplemental Table 7). Articles for nine genes were found with OncoSearch and supporting literature was found for seven more. The remaining two were found to have interactions with known cancer genes in the OCGene ovarian cancer database [29]. For bladder cancer, 17 out of 26 multiple driver genes were a likely cancer gene or an interactor with one (Supplemental Table 8). Eight drivers had articles found by OncoSearch and supporting literature was found for seven more. The remaining two had interactions with known cancer genes in the AGCOH or intOgen databases [26], [27].

Our methods associated cis genes with disrupted biological process in trans. Many of the multiple driver genes in all three datasets were appropriately associated with biological processes that they are known to be involved in. For example, in breast cancer, BOP1 is required for the maturation of ribosomal RNAs [30] and was associated in our algorithm with "ribosome biogenesis" (Table 1). In ovarian cancer, candidate GSDMD is involved in the release of Interleukin 1-Beta, and was associated with out methods with "lymphocyte activation" and "response to cytokines" (Supplemental Table 7). HSPA9 in bladder cancer is a heat shock proteinand was associated "cellular response to stress" (Supplemental Table 8). These genes and others are all involved in cancer, and are candidate copy number drivers and respective candidate disrupted processes.

In order to compute the enrichment of cis gene categories in known cancer gene lists, we created a list of cancer genes by combining 727 known cancer genes from the AGCOH database [26] and 475 known cancer genes from the intOgen database [27]. The overlap between all cis genes in ovarian cancer and the cancer gene list was poor (hypergeometric *p*-value = 0.28). Thus, for ovarian cancer, we used a more specific cancer list from the OCGene ovarian cancer database [29]. The OCGene ovarian cancer database had a stronger, but marginal overlap with the cis genes (hypergeometric *p*-value = 0.09). Cis genes in breast and bladder cancer had sufficient overlap with the intOgen and AGCOH database (*p*-value = 0.0025 for bladder and 0.11 for breast cancer). We found that drivers and multiple drivers had lower *p*-values than genes that were filtered out by ProcessDriver and cis genes that were the most correlated with their own copy number (Table 2). Although some of the *p*-values were marginal, the enrichment for drivers and/or multiple drivers was higher than for cis genes that were filtered out. The marginal *p*-values could be due the incompleteness of the databases. As shown in Table 1 and Supplemental Tables 7 and 8, additional literature was found via a manual search for some multiple drivers supporting their involvement in cancer, despite not being present in the databases, yet.

Table 2. Enrichment of cis genes with known cancer genes. Hypergeometric p-values for the enrichment of known cancer genes in selected drivers, cis genes that were filtered out by ProcessDriver, and cis genes that were the most correlated with their own copy number. Number of genes indicates the number of cis genes in each of the groups defined by ProcessDriver. The p-values were computed using AGCOH and intOgen databases for bladder and breast cancer [27], [28], and using the OCGene ovarian cancer database for ovarian cancer [29].

		Multiple driver	Driver	Semi-driver	Last in λpath	Filtered	Top Cor
Bladder cancer	**# genes**	26	89	197	43	197	120
	p-Value	**0.06**	0.12	0.86	0.52	0.77	0.6
Breast cancer	**# genes**	44	116	266	51	259	128
	p-Value	**0.01**	0.15	0.96	0.19	0.27	0.52
Ovarian cancer	**# genes**	33	82	184	45	389	138
	p-Value	0.25	**0.07**	0.35	0.71	0.85	0.8

Bold indicates the minimum enrichment p -value for each cancer type.

We tested the effect of adjusting the threshold for the canonical correlation in SCCA step on the enrichment of cancer genes in multiple drivers for breast cancer (Supplemental Table 9). We observed that a threshold of 0.7 had the highest enrichment. Higher thresholds such as 0.8 and 0.9 had insignificant enrichments, suggesting that higher thresholds exclude too many cancer genes. Lower thresholds such as 0.5 and 0.6 had significant enrichment, but reduced specificity by including more multiple drivers.

3.2. SCCA filters cis genes with a lower correlation of expression to their own copy number

The underlying assumption in many previous studies on cancer drivers is that driver gene expression has a higher correlation to their own copy number than passenger genes [4], [5], [6], [7]. Although we did not use correlation of cis gene expression to its copy number to narrow down likely drivers, we expect that our drivers would have a higher correlation between their gene expression and copy number than the correlation of other genes' expression to their own copy number. Fig. 2 illustrates the distribution of the correlation of cis copy number to gene expression in the different groups of cis genes for bladder and breast cancer data and Supplemental Fig. 2 shows the same distribution for the ovarian cancer data. Cis genes that were filtered by SCCA had a significantly lower average correlation of expression with copy number than driver genes in all three cancers (Wilcoxon rank-sum *p*-value < 0.001 for ovarian and breast cancer and < 0.05 for bladder cancer). We also observed that for cis genes filtered by SCCA, there were still genes with extremely high correlation between expression and copy number. These results suggest that utilizing correlation between gene expression and copy number to select potential driver genes could make false positive selections.

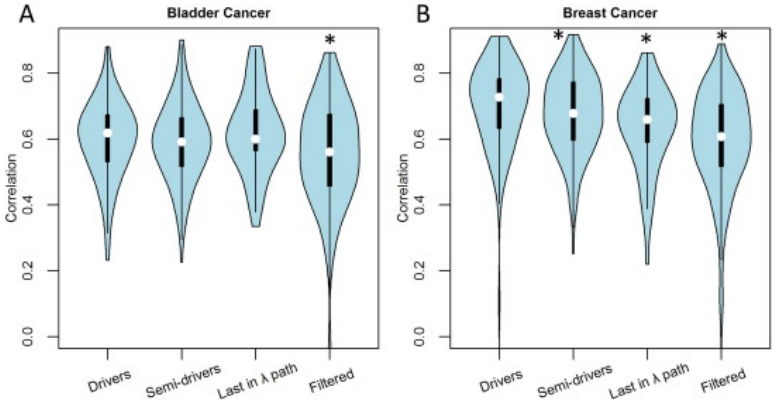

Fig. 2. Correlation of copy number to cis gene expression. Violin plots representing the correlation of cis genes to their own copy number for selected drivers and cis genes filtered-out by ProcessDriver for (A) bladder cancer and (B) breast cancer. Definition of each group is in the results section. Asterisk indicates $p < 0.05$ in a Wilcoxon rank-sum test compared to the drivers group.

3.3. Driver genes are associated with a higher risk of new tumor events after initial treatment

In order to evaluate if the driver genes could predictive new tumor events after initial treatment, we performed survival analysis on cis genes. We fit a univariate Cox proportional hazard model for each cis gene for the number of days to a new tumor event after the initial treatment and used the cis gene expression as a covariate. If a patient did not experience a new tumor event after the initial treatment, the days until the last follow-up were used and the patient was censored. In the bladder cancer cohort, 97 out of 120 patients have had new tumor events after the initial treatment and in the ovarian cancer cohort 86 out of 120 patients have had new tumor events. Only two out of 92 of the luminal A patients had new tumor events after initial treatment, therefore luminal A was not included in this analysis.

A hazard ratio > 1 implies that an increase of expression of the cis gene increases the risk of a new tumor event, while a hazard ratio < 1 implies that an increase of the cis gene expression decreases the risk of a new tumor event. Overall in bladder cancer, drivers had hazard ratios greater than one (Fig. 3A and C). We compared the mean of the hazard ratios of each group using the Wilcoxon rank-sum test. We observed that the mean of the hazard ratios was significantly higher in the driver group compared to the top correlated, filtered and last in λ path groups with $p < 0.05$.

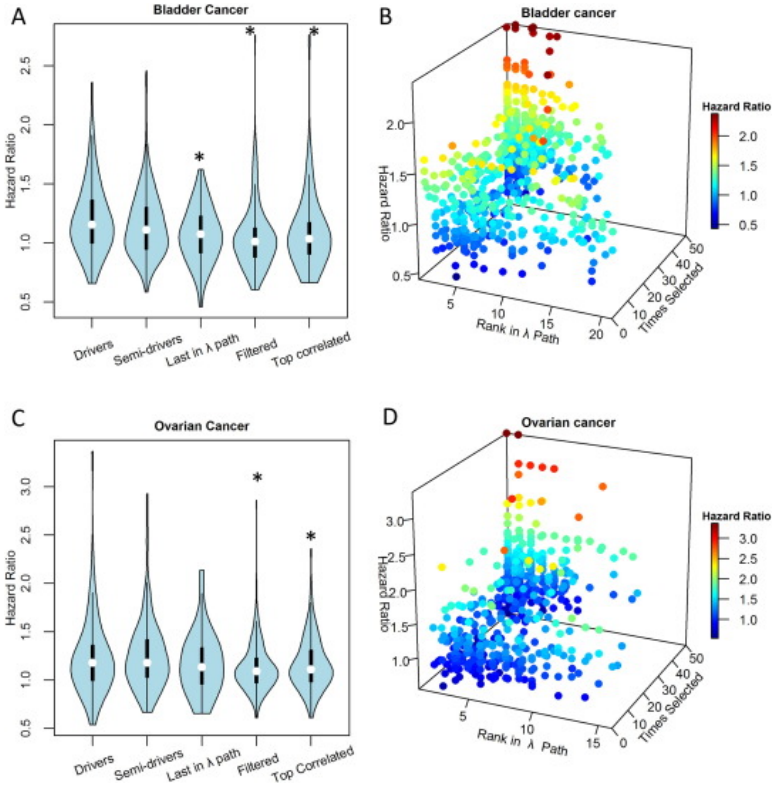

Fig. 3. Hazard ratios for new tumor events in a univariate Cox proportional hazards model. Violin plots of hazard ratios for genes filtered out or selected at various stages of the driver selection step for (A) bladder cancer and (C) ovarian cancer. Asterisk indicates $p < 0.05$ in a Wilcoxon rank-sum test in bladder cancer and F-test of variances in ovarian cancer compared

to the driver group. Hazard ratios were plotted for genes in the multi-task LASSO stage against the number of times they were selected by resampling and the rank in the λ path for (B) bladder cancer and (D) ovarian cancer.

In ovarian cancer, multiple driver RAF1, a putative oncogene, had the highest hazard ratio of any cis gene of 3.2. However, multiple driver CASP3, which promotes apoptosis and is in a deleted region, had the lowest hazard ratio of any cis gene of 0.55. This highlights that the hazard ratio could be dependent on the drivers oncogenic or tumor suppressor activities since a lower hazard ratio implies lower risk with increased expression. We found that the driver group ($\sigma^2 = 0.16$) had a significantly higher variance than the top correlated ($\sigma^2 = 0.07$), and filtered ($\sigma^2 = 0.055$) groups (Levenne's test p-value < 0.05). Although not significant, drivers also had a larger variance than the last in λ path group. This suggests that drivers of ovarian cancer have a higher or lower hazard ratio due to tumor suppressor and oncogenic activities.

Bladder cancer also contains drivers with low hazard ratios. For example, multiple driver FEM1B has a hazard ratio of 0.8 and is a pro-apoptotic protein [31]. Fig. 3B and D illustrates the hazard ratio for new tumor events after initial treatment for cis genes that appeared in the multi-task LASSO phase in bladder and ovarian data sets, respectively. The results show that that cis genes with the highest hazard ratios were selected close to 50 out of 50 times during resampling and had a relatively low rank in the λ path.

4. Conclusions

We designed and implemented ProcessDriver in three different cancer sets and found consistently that the most likely candidate drivers are more enriched in known cancer genes. For each dataset, more than half of the multiple drivers are known to be involved in cancer. Biological processes are associated with each driver through the trans genes, and all the trans genes are differentially expressed as a result of the aberration. Therefore, the processes associated with a driver are the ones that are likely disrupted.

We also found that the selected drivers have more extreme hazard ratios for new tumor events after initial treatment with respect to new tumor events compared to cis genes filtered out by ProcessDriver and cis genes selected on the basis of their correlation of expression to their own copy number. Since drivers promote tumorigenesis, it is expected that drivers would be linked to new tumor events.

Aside from ensuring that all cis genes and trans genes are differentially expressed with respect to an aberrated region, we do not use the correlation of copy number to cis gene expression in our filtering of drivers. However, as expected, the cis genes that were selected as drivers had expression that was more correlated to their own copy number compared to cis genes filtered by SCCA. This result suggests that drivers tend to have higher correlation to copy number. However, when we selected the cis genes that are most correlated to their own copy number for each GO term module, it results in a lower enrichment of known cancer genes and lower hazard ratios with respect to new tumor events compared to drivers selected by ProcessDriver. These results highlight the importance of selecting drivers based on the relationship between cis gene expression and trans gene expression, as opposed to selecting the cis genes based on correlation to their own copy number as in previous studies [4], [5], [6], [7].

While a couple of studies relate cis genes to other genes in trans, our approach differs from previous approaches in a number of ways. The statistical approaches outlined in this pipeline strongly emphasize a close relationship between a potential driver and downstream target trans genes and also provide insight into disrupted biological processes. Akavia et al. relates the expression of cis genes to downstream targets, but does not integrate information about biological processes [8]. Aure et al. associates cis genes with biological processes in trans. However, all other genes are used as trans genes [9]. In this study, all trans genes must be differentially

expressed with respect to the aberration. In [9] the correlation with cis genes to their own copy number is to first narrow down cis genes. Here, we demonstrate that the relationship between cis gene expression and trans gene expression is more valuable in selecting drivers than the correlation of cis genes to their own copy number.

ProcessDriver will narrow down a list of driver genes from many genes that are cis-affected by copy number. This could help find drivers which could be therapeutic targets of drugs. Additionally, the algorithm associates drivers with biological processes through the trans genes, which could aid in gaining insight into the widespread, downstream effects of the driver.

Acknowledgements

This work was supported by the Richard W. Jobling research assistantship of BB from Marquette University. We would also like to thank the Marquette High Performance Computing center.

References

[1] D. Hanahan, R.A. Weinberg. **The hallmarks of cancer.** *Cell*, 100 (1) (2000), p. 57

[2] C.H. Mermel, S.E. Schumacher, B. Hill, M.L. Meyerson, R.Beroukhim, G. Getz. **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.** *Genome Biol*., 12 (4) (2011), pp. 1-14

[3] R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, *et al.* **Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.** *Proc. Natl. Acad. Sci. U. S. A*., 104 (50) (2007 Dec 11), pp. 20007-20012

[4] D. Tamborero, N. Lopez-Bigas, A. Gonzalez-Perez. **Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression.** *PLoS One*, 8 (2) (2013 02/08), Article e55489

[5] S. Ambatipudi, M. Gerstung, M. Pandey, T. Samant, A. Patil, S.Kane, *et al.* **Genome-wide expression and copy number analysis identifies driver genes in gingivobuccal cancers.** *Genes Chromosomes Cancer*, 51 (2) (2011 11/10), pp. 161-173

[6] B. Fan, S. Dachrut, H. Coral, S.T. Yuen, K.M. Chu, S. Law, *et al.*. **Integration of DNA copy number alterations and transcriptional expression analysis in human gastric cancer.** *PLoS One*, 7 (4) (2012 04/23), Article e29824

[7] C.R. Pickering, J. Zhang, S.Y. Yoo, L. Bengtsson, S. Moorthy, D.M. Neskey, *et al.* **Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers.** *Cancer*

[8] U.D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H.C. Causton, *et al.* **An integrated approach to uncover drivers of cancer.** *Cell*, 143 (6) (2010 12/10), pp. 1005-1017

[9] M.R. Aure, I. Steinfeld, L.O. Baumbusch, K. Liestøl, D. Lipson, S. Nyberg, *et al.* **Identifying In-*Trans* process associated genes in breast cancer by integrated analysis of copy number and expression data.** *PLoS One*, 8 (1) (2013 01/30), Article e53014

[10] G.I. Evan, K.H. Vousden. **Proliferation, cell cycle and apoptosis in cancer.** *Nature*, 411 (6835) (2001), pp. 342-348

[11] M.I. Love, W. Huber, S. Anders. **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** Genome Bio*l*., 15 (12) (2014), pp. 1-21

[12] A. Alexa, J. Rahnenfuhrer. **topGO: topGO: enrichment analysis for Gene Ontology.** *R package version 2.18.0 (2010)*

[13] H. Frohlich, N. Speer, A. Poustka, T. Beissbarth. **GOSim–an R-package for computation of information theoretic GO similarities between terms and gene products.** *BMC Bioinformatics*, 8 (2007 May 22), p. 166

[14] D.M. Witten, Robert Tibshirani, Trevor Hastie. **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.** *Biostatistics*, 10 (3) (2009 7), pp. 515-534

[15] H. Hotelling. **Relations between two sets of variates.** *Biometrika*, 28 (3–4) (1936), pp. 321-377

[16] J. Friedman, T. Hastie, R. Tibshirani. **Regularization paths for generalized linear models via coordinate descent.** *J. Stat. Softw*., 33 (1) (2010), pp. 1-22

[17] Comprehensive molecular portraits of human breast tumours, Nature, 490 (7418) (2012 10/04), pp. 61-70

[18] Integrated genomic analyses of ovarian carcinoma, Nature, 474(7353) (2011 06/30), pp. 609-615

[19] The Cancer Genome Atlas Research Network. **Comprehensive molecular characterization of urothelial bladder carcinoma.** *Nature*, 507 (7492) (2014 03/20), pp. 315-322

[20] H. Lee, T.C. Dang, H. Lee, J.C. Park. **OncoSearch: cancer gene search engine with literature evidence.** *Nucleic Acids Res*., 42 (2014 02/22), pp. W416-W421

[21] S. Braig, A. Bosserhoff. **Death inducer-obliterator 1 (Dido1) is a BMP target gene and promotes BMP-induced melanoma progression.** *Oncogene*, 32 (7) (2013 02/14), pp. 837-848

[22] L. Furic, L. Rong, O. Larsson, I.H. Koumakpayi, K. Yoshida, A.Brueschke, *et al.* **eIF4E phosphorylation promotes tumorigenesis and is associated with prostate cancer progression.** *Proc. Natl. Acad. Sci. U. S. A.,* 107 (32) (2010 08/02), pp. 14134-14139

[23] K. Yamaguchi, R. Yamaguchi, N. Takahashi, T. Ikenoue, T.Fujii, M. Shinozaki, *et al.* **Overexpression of cohesion establishment factor DSCC1 through E2F in colorectal cancer.** *PLoS One,* 9 (1) (2014 01/17), Article e85750

[24] L. Chen, Y. Li, C.H. Lin, T.H.M. Chan, R.K.K. Chow, Y. Song, *et al.* **Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma.** *Nat. Med.,* 19 (2) (2013), pp. 209-216. print

[25] S.A. Jensen, A.E. Calvert, G. Volpert, F.M. Kouri, L.A. Hurley, J.P. Luciano, *et al.* **Bcl2L13 is a ceramide synthase inhibitor in glioblastoma.** *Proc. Natl. Acad. Sci.,* 111 (15) (2014), pp. 5682-5687

[26] J.L. Huret, M. Ahmad, M. Arsaban, A. Bernheim, J. Cigna, F.Desangles, *et al.* **Atlas of genetics and cytogenetics in oncology and haematology in 2013.** *Nucleic Acids Res.,* 41 (Database issue) (2013 Jan), pp. D920-D924

[27] G. Gundem, C. Perez-Llamas, A. Jene-Sanz, A. Kedzierska, A.Islam, J. Deu-Pons, *et al.* **IntOGen: integration and data mining of multidimensional oncogenomic data.** *Nat. Methods,* 7 (2) (2010), pp. 92-93 print

[28] A. Chatr-aryamontri, B. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, *et al.* **The BioGRID interaction database: 2015 update.** *Nucleic Acids Res.* (2014)

[29] Y. Liu, J. Xia, J. Sun, M. Zhao. **OCGene: a database of experimentally verified ovarian cancer-related genes with precomputed regulation information.** *Cell Death Dis*., 6 (12) (2015 12/31), Article e2036

[30] Y.R. Lapik, C.J. Fernandes, L.F. Lau, D.G. Pestov. **Physical and functional interaction between Pes1 and Bop1 in mammalian ribosome biogenesis.** *Mol. Cell*, 15 (1) (2004), p. 17

[31] M.C. Subauste, O.J. Sansom, N. Porecha, N. Raich, L. Du, J.F.Maher. **Fem1b, a proapoptotic protein, mediates proteasome inhibitor-induced apoptosis of human colon cancer cells.** *Mol. Carcinog*., 49 (2) (2010 02/01), pp. 105-113