ELIMINATOR: Essentiality anaLysIs using MultIsystem Networks And inTeger prOgRamming

Asier Antoranz
Intelligent Biodata Ltd
Maria Ortiz
A Bristol Myers Squibb Company
Jon Pey (jonpey@gmail.com)
Intelligent Biodata Ltd

Research Article

Keywords: gene essentiality analysis, constrain-based modelling, multisystem networks, in-silico methods.

Posted Date: March 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1347932/v1

ELIMINATOR: Essentiality anaLysIs using MultIsystem Networks And inTeger prOgRamming

Asier Antoranz[(1)], Maria Ortiz[(2)], and Jon Pey[(1)]

(1) Intelligent Biodata Ltd, San Sebastian, Spain

(2) BMS Center for Innovation and Translational Research Europe (CITRE), A Bristol Myers Squibb Company, Sevilla, Spain

Corresponding authozr: Jon Pey, jonpey@gmail.com

Abstract

A gene is considered as essential when it is indispensable for cells to grow and replicate under a certain environment. However, gene essentiality is not a structural property but rather a contextual one, which depends on the specific biological conditions affecting the cell. This circumstantial essentiality of genes is what brings the attention of scientist since we can identify genes essential for cancer cells but not essential for healthy cells. This same contextuality makes their identification extremely challenging. Huge experimental efforts such as Project Achilles where the essentiality of thousands of genes is measured together with a plethora of molecular data (transcriptomics, copy number, mutations, etc.) in over one thousand cell lines can shed light on the causality behind the essentiality of a gene in a given environment.

Here, we present an in-silico method for the identification of patient-specific essential genes using constraint-based modelling (CBM). Our method expands the ideas behind traditional CBM to accommodate multisystem networks. In essence, it first calculates the minimum number of non-expressed genes required to be active by the cell to sustain life as defined by a set of requirements; and second, it performs an exhaustive in-silico gene knockout to find those that lead to the need of activating additional non-expressed genes.

We validated the proposed methodology using a set of 452 cancer cell lines derived from the Cancer Cell Line Encyclopedia where an exhaustive experimental large-scale gene knockout study using CRISPR (Achilles Project) evaluates the impact of each removal. We also show that the integration of different essentiality predictions per gene, what we called Essentiality Congruity Score, (derived from multiple pathways) reduces the number of false positives. Finally, we explored the gene essentiality predictions for a breast cancer patient dataset, and our results showed high concordance with previous publications.

These findings suggest that identifying genes whose activity is fundamental to sustain cellular life in a patient-specific manner is feasible using in-silico methods. The patient-level gene essentiality predictions can pave the way for precision medicine by identifying potential drug targets whose deletion can induce death in tumour cells.

Keywords: gene essentiality analysis, constrain-based modelling, multisystem networks, in-silico methods.

Introduction

We can define an essential gene as a gene whose activity is fundamental to sustain life (Bartha et al., 2018). It is precisely the critical importance of these genes that brings the attention of scientists. For instance, in cancer research, specific essential genes of this condition are considered as promising drug targets as their deletion can induce death in tumour cells (Tsherniak et al., 2017).

The essentiality of a gene is not a structural property, it depends on the biological scenario under consideration (Zhang and Lin 2009), including the cellular environmental conditions, disease phenotypes, etc. The contextual dependency of essential genes makes their experimental identification an extremely difficult task. The huge effort of experimental initiatives such as Project Achilles in creating an archive of essential genes is of utmost interest to the scientific community (Tsherniak et al., 2017). However, the biological context in which a particular gene turns out to be essential is exceptionally critical in cancer, where the essentiality of a gene could emerge at patient level (Fernald et al., 2011). This highlights the role of in-silico gene essentiality identification approaches that effectively integrate -omics datasets to contextualize a given biological scenario.

During the last decade, many successful examples have been presented on integrating omics datasets with biological networks in the context of efficient mathematical models to address an assortment of biomedical problems (Yan et al., 2018), including the identification of essential genes (Li et al., 2020). We can find relevant insights provided by these algorithms in different fields, ranging from microbiology (Plata et al., 2010) to cancer research (Frezza et al., 2011), among others.

Despite the recent advent of machine-learning based gene essentiality analyses (Kuang et al., 2020), (Schapke et al., 2020), traditionally, approaches referred to as Constraint-Based Modelling (CBM) leaded the field setting the foundations for the development of different methodologies to predict essential genes (Apaolaza et al., 2017), (Pey et al., 2017), (Tobalina et al., 2016). In essence, CBM integrates omics data in the context of genome-scale metabolic networks resulting in a linear system of inequalities. The arising system of inequations is usually solved using linear optimization techniques (Martin 2012). Here, essential genes emerge from their indispensability when ensuring the activity of an artificial metabolic reaction,

referred to as biomass reaction, which involves the metabolic requirements of the cell for its replication (Agren et al., 2014).

In this work, we extend the ideas in traditional CBM by going beyond metabolism considering multisystem networks (Schaefer et al., 2009). In addition, and in analogy with CBM, here we identify genes whose activity is essential for a relevant biological task. Thus, the emerging set of essential genes will be richer and more diverse than in traditional CBM, capturing a variety of biological processes (Vaske et al., 2010).

Overall, this article introduces a new methodology for the in-silico identification of essential genes. This approach combines three main inputs: (i) An indispensable biological entity/process required to sustain cellular life, (ii) a set of interaction networks including the molecular requirements to activate the aforementioned indispensable entity and (iii) an experimental dataset that reflects, at least qualitatively, the genetic landscape of the sample/patient, e.g., gene expression data.

These inputs are subsequently encoded into a mathematical model (Integer Linear Program, ILP) (Schrijver, 1998) that finds the minimum number of non-expressed genes required to activate the given relevant function. Then, a systematic approach identifies artificial gene knockouts that lead to require additional unexpressed genes to activate the critical biological entity/process. These knockouts are precisely considered as **essential genes**. This is further illustrated in the manuscript through a series of toy examples. In addition, we successfully validated a continuous score representing the degree of essentiality of a given gene, referred to as the *Essentiality Congruity Score*. We also show the relevance of each of these inputs by evaluating the performance of the method in a different set of scenarios. Finally, we apply the methodology to a group of breast cancer patients and subsequently support the relevance of the emerging essential genes based on a literature review.

Methods

In the following section, we introduce the *in-silico* gene-essentiality framework presented in this article. In the first subsection, referred to as **Pathways**, we describe the biological pathway compendium used for the model; in the second subsection, mentioned as **Datasets**, we describe all the experimental data used throughout the study; the third subsection, **Mathematical Model**, describes the mathematical equations modelling the pathways and integrating the experimental data; and, in the fourth subsection, called **Gene Essentiality Analysis**, we present the pipeline that systematically find essential genes. Moreover, we present the *Essentiality Congruity Score*, which assigns a quantitative value to an otherwise binary score to represent the essentiality of a gene.

Pathways

As in (Vaske et al., 2010), we consider a set of well-curated pathways from the (NCI-PID) (Schaefer et al., 2009) database, which are represented in the UCSC Pathway Tab Format. Vaske and co-workers provided further details about the characteristics of these pathways, including their consistency when capturing cancer related knowledge. In essence, these pathways comprise vertices and edges representing various types of biological entities and their interactions respectively. For instance, vertices could denote a gene/protein, gene complex or biological abstracts like “mitosis” or “cell motility”, among others, whilst edges represent activations/inhibitions or member/component associations (Vaske et al., 2010).

Following the UCSC Pathway Tab Format (Vaske et al., 2010), we will consider the following interactions: member (member>), component (component>), activation (-a>, -t>, -ap>) and inhibition (-t|,-ap|, -a|). As will be introduced in the next subsection, each one of these interactions is modelled by a specific set of mathematical equations.

Datasets

We initially show the behaviour of the algorithm in a toy-example simulated dataset using the Wnt receptor signaling pathway, planar cell polarity pathway, from the NCI-PID which is shown in **Figure 3** (Schaefer et al., 2009).

Secondly, we applied the methodology in the set of pathways from the NCI-PID (Schaefer et al., 2009) using the gene expression data from the Cancer Cell Line Encyclopedia (CCLE)

(Barretina et al., 2012) and validated the biological relevance of the predictions using the essentiality scores from the Achilles project (Dempster et al., 2019). Project Achilles is a systematic effort aimed at identifying and cataloguing gene essentiality across hundreds of genomically characterized cancer cell lines. These gene essentiality scores are obtained from CRISPR knockouts (CERES method) (Meyers et al., 2017) on several of the cell lines included in the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012), a compilation of gene expression, chromosomal copy number and massively parallel sequencing data from nearly 1,000 human cancer cell lines. The gene expression data for the cell lines was obtained from the Gene Expression Omnibus (GSE36133) which includes 917 cell-lines annotated with 23,521 gene identifiers (HGNC format). Gene expression data was binarized (1 expressed, 0 not expressed) using The Gene Expression Barcode 3.0 (McCall et al., 2012), (McCall et al., 2014). Probes were mapped to HGNC identifiers (GPL570, Affymetrix Human Genome U133 Plus 2.0 Array). The Achilles Essentiality Scores were downloaded from the DepMap portal (https://depmap.org/portal/download/, version 20Q1) which contained essentiality scores for 18,333 genes in 739 cell lines, 478 of which were in common with the CCLE. In total, 1,660 HGNC IDs were in common between Achilles, CCLE and the genes/proteins present in the NCI-PID pathways. The Achilles database included in DepMap contains missing values for several of these genes (NA values). After removing NA values, 26 cell lines didn't contain any data for the 1,660 genes in common, thus reducing the number of included cell lines to 452. Achilles scores represent gene essentiality, the more negative the score, the more essential the knockout of the gene is for a given cell-line.

Finally, we applied the gene essentiality method to Breast Cancer patient samples (Maubant S, 2012), (Maire V, 2013). This dataset includes transcriptome analysis of 130 breast cancer samples (41 TNBC; 30 Her2; 30 Luminal B and 29 Luminal A), 11 normal breast tissue samples and 14 TNBC cell lines. This dataset contains 178 array samples. 153 arrays were used to analyse 130 unique breast cancer samples from as many patients and 23 technical duplicates. In addition, 11 "Normal" samples from healthy breast tissue obtained from mammoplasty are included, as well as a collection of 14 breast cancer cell lines. Data production involved different array batches and hybridation series which were accounted for in the pre-processing of the data. Processed gene expression data and sample meta-data was obtained from the Gene Expression Omnibus (GSE65194). Samples belonging to cell lines were removed from further analysis. Gene expression data was discretized using The Gene Expression Barcode 3.0

(McCall et al, 2012), (McCall et al, 2014). Probes were mapped to HGNC identifiers (GPL570, Affymetrix Human Genome U133 Plus 2.0 Array).

Mathematical model

The network format described above can be translated into a series of Boolean rules. However, the inherent complexity of these rules grows exponentially when regular-sized pathways are considered. In this subsection, we present the Integer Linear Programming framework (ILP) able to contend with complex networks and capturing all the essence of the Boolean rules. The ILP constitutes the core of the methodology. In essence, it calculates the minimum number of non-expressed genes required to activate a given biological function necessary to sustain cellular life. The mathematical equations in the ILP arise precisely from the structure of the pathway and its different interactions **(Figure 1)**.

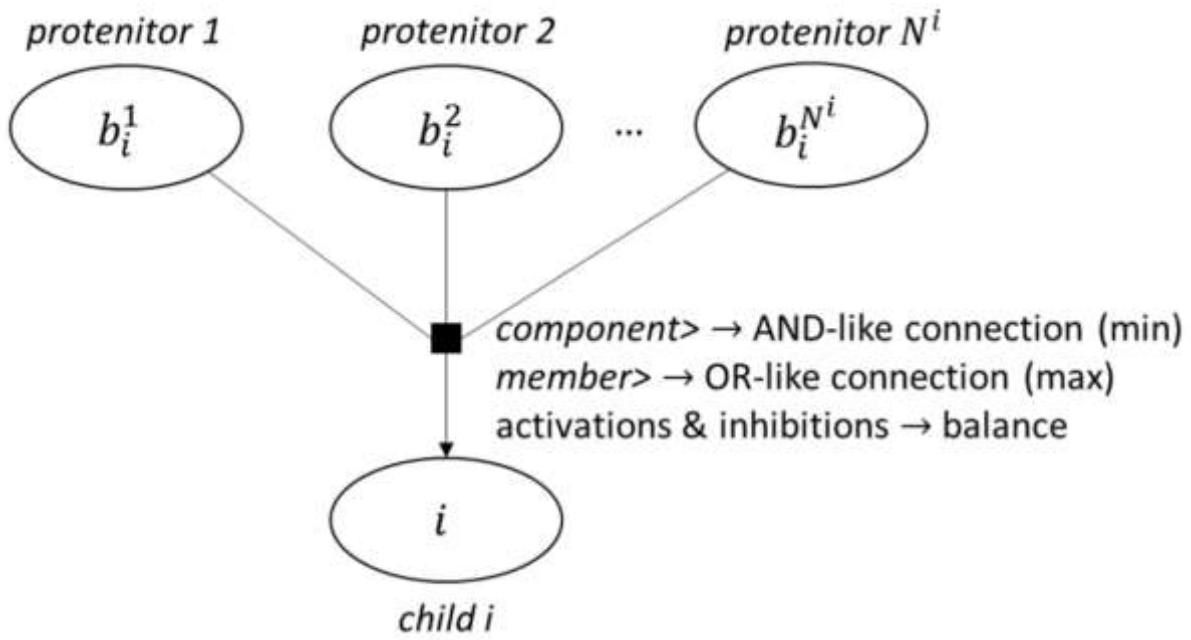

Figure 1 - Conversion of UCSC Pathway Tab Format to valid pathways for the mathematical model. *These pathways model the relationship between the i-th child and its progenitors using various types of interactions including component>, member>, and activations & inhibitions.*

Let $\boldsymbol{B^i}$ represent the set of all the parents for a given entity i. Let N^i be the cardinality of $\boldsymbol{B^i}$, namely $N^i = |\boldsymbol{B^i}|$. We can define E_i as a binary variable $\{0,1\}$ that represents the final binary activation status of i in a given simulation ($E_i = 0$ if inactive, $E_i = 1$ if active). The reader should note that E_i is not the same as the experimental expression of the genes, the experimental data will be included later in the model. In particular, the method focuses on minimizing the number of non-expressed genes being active but does not apply direct restrictions to expressed genes.

Now we will proceed to mathematically define the constraints based on the nature of the interaction between the *i*-th child and its progenitors.

component>

In analogy to the AND-like connection considered for components of a complex in Vaske *et al.,* 2010, the final activation status of the child (E_i) is determined by the minimum value from all its components. So $E_i = 1$ if, and only if, $E_j = 1\ \forall b \in \boldsymbol{B_i}$. Otherwise, we impose that $E_i = 0$.

$$E_i \geq \sum_{\forall b \in \boldsymbol{B^i}} E_b - \left(N^i - 1\right) \quad (1)$$

$$N^i \cdot E_i \leq \sum_{\forall b \in \boldsymbol{B^i}} E_b \quad (2)$$

member>

The final status of node *i* is determined by the maximum value from all its members. So $E_i = 0$ if, and only if, $E_b = 0\ \forall b \in \boldsymbol{B_i}$. Otherwise, we impose that $E_i = 1$. Note the similarity with (Vaske *et al.,* 2010) where members are modelled in a OR-like fashion.

$$N^i \cdot E_i \geq \sum_{\forall b \in \boldsymbol{B^i}} E_b \quad (3)$$

$$E_i \leq \sum_{\forall b \in \boldsymbol{B^i}} E_b \quad (4)$$

activations & inhibitions

The final status of the target is determined by a balance between all its activators and inhibitors. For simplicity, we can define an intermediate variable $F_i \in \mathbb{Z}$ that expresses the activation/inhibition state of *i*,

$$F_i = \sum_{\forall b \in \boldsymbol{J_i}} E_b - \sum_{\forall b \in \boldsymbol{I_i}} E_b \quad (5)$$

where $\boldsymbol{J_i}$ and $\boldsymbol{I_i}$ represent the set of activators and inhibitors of *i*, respectively. The activation status of the child *i* is then determined by its activation state,

$$M \cdot (E_i - 1) \leq F_i - w \quad (6)$$

where *M* is an auxiliary positive large integer ($M = 1{,}000$) and *w* the relative weight between activators and inhibitors that modules the sign of F_i. Here, we considered an arbitrary value of $w > 0{,}5$. The reader should note how equation 6 forces $E_i = 0$ when $F_i \prec w$ and does not constrain E_i when $F_i \geq w$. That is, an inhibitory state of *i* is sufficient for the inhibition of the child, while an activation state of *i* is necessary for the activation of the child. Note that equation

Error! Reference source not found. is only imposed when the target *i* is an abstract or a complex because the genes and proteins generally represent the entries of the network, and their global activators-inhibitors scenario are often not properly captured in individual pathways.

Artificially activating an abstract/complex

We will impose the activation of relevant biological functions. To that end, we define the set of all entities required to sustain cellular life as $\boldsymbol{A}$, from now on defined as actives, and an independent problem is defined for each of them.

$$E_a = 1, \forall a \in \boldsymbol{A} \quad (7)$$

Where E_a represents the activity of the entity a. In practice, for a given pathway, the set **A** consists of all its abstracts and complexes.

Minimizing the number of lowly expressed

Let $\boldsymbol{L}$ represent the set of non-expressed genes. $\forall a \in \boldsymbol{A}$, we define the optimal solution as the one that directly minimizes the number of non-expressed genes active in the final solution whilst $E_a = 1$. Note that the model will provide a specific value of the objective function for each $a \in \boldsymbol{A}$. We will refer to this solution as S_a^{wild}.

$$S_a^{wild} = min \sum_{\forall i \in \boldsymbol{L}} E_i \,, \forall a \in \boldsymbol{A} \quad (8)$$

Gene Essentiality Analysis

Let $\boldsymbol{G}$ represent the set of expressed genes. $\forall a \in \boldsymbol{A}$, we model each gene removal ($g \in \boldsymbol{G}$) sequentially to quantify the biological impact of its knockout for a given abstract and experimental picture. The gene removal is basically imposed by forcing E_g to be equal to zero (E_g=*0*) with g representing the gene that is being knocked out. Note that genes that appear in the pathway models and are not experimentally measured are considered as expressed and therefore we include them in the knock-out process.

Afterwards the problem is solved (equation 8) and the minimum number of non-expressed genes active is calculated (S_a^g). If $S_a^g > S_a^{wild}$ the gene is considered as essential for the cell to carry out that biological process (a) in the given pathway. Else, the gene is considered as not essential. In other words, if the new solution S_a^g modelling the knockout of gene g requires the presence of more lowly expressed genes that the wild type $Sa\ wild$, we assume that the removal

of g is causing a significant impact to the phenotype represented by the gene expression dataset. The flow diagram corresponding to the methodology is summarised in **Figure 2**.

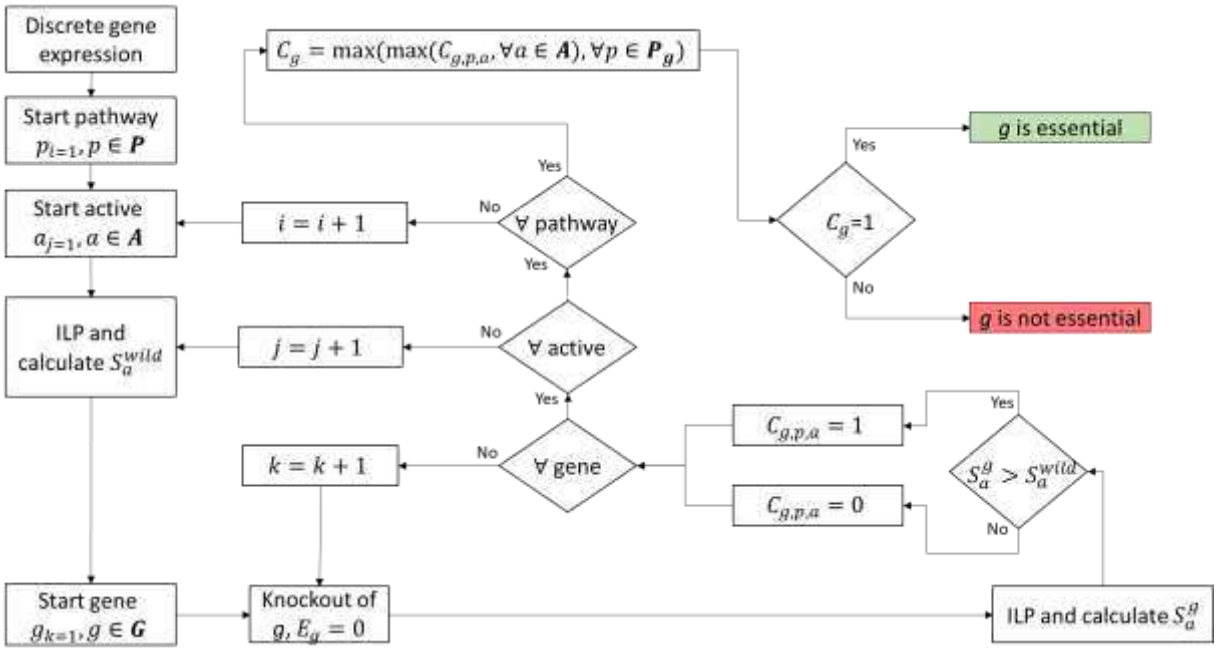

***Figure 2* – Flow diagram of the methodology.** Starting from a specific experimental picture (discrete gene expression), we calculate the minimum number of non-expressed genes required to be active for the cell to sustain cellular life (S_a^{wild}). Then, we systematically knock-out one by one all the expressed genes g present in the pathway **P** (E_g=0) and recalculate the minimum number of non-expressed genes required to be active for the cell to sustain cellular life (S_a^g). We define a gene as essential for a given active if $S_a^g > S_a^{wild}$. We repeat this process for all the genes, actives, and pathways included in the database. The essentiality of a gene g is finally defined as the maximum of all its essentiality predictions.

Note that the proposed methodology, analyses every pathway independently and thus produces a prediction of essentiality for every gene and every active (gene complex or biological abstract) in the pathway. This means, that for a given gene, multiple predictions of essentiality can be produced in the same pathway (as many as there are elements in $\boldsymbol{A}$ for that pathway). Conceptually, our method assumes that if the gene is essential for at least one entity required to sustain cellular life (active), then its knockout would be fatal for the cell overall. Therefore, a gene is essential for a pathway, if it is essential for any of its actives.

$$C_{g,p} = \max(C_{g,p,a}, \forall a \in \boldsymbol{A}) \quad (9)$$

Where $C_{g,p,a}$ is a binary variable ($C_{g,p,a} \in \{0,1\}$) that represents the essentiality of the gene g in the pathway p for the entity a.

Moreover, different pathways are not completely disjoint sets and often have common genes. This means, that we can have more than one prediction of essentiality for a gene in different pathways. Similarly, we assume that if the gene is essential for at least one pathway, then its

knockout would be fatal for the cell overall. Therefore, a gene is essential, if it is essential for any of its pathways.

$$C_g = \max(C_{g,p}, \forall p \in \boldsymbol{P_g}) \quad (10)$$

Where $\boldsymbol{P_g}$ represents the set of pathways where the gene *g* is present.

Globally essential and globally not essential genes

If the knockout of a gene *g* leads to $C_g = 1$ for each of the experimental datasets, the gene is considered globally essential. Similarly, if the knockout of a gene leads to $C_g = 0$ for every experimental dataset, the gene is considered globally not essential. Both globally essential and globally not essential genes are excluded from downstream analysis. Given the ulterior motives of the method, we are particularly interested in genes whose essentiality depends on the experimental dataset. Therefore, if a particular gene turns out to be essential in a cancer phenotype but not in the corresponding healthy tissue, we can identify it as a potential drug target.

Essential Congruity Score (ECS)

The proposed methodology assumes that predictions of essentiality ($C_{g,p,a} = 1$) are more impactful than predictions of no essentiality ($C_{g,p,a} = 0$) and the essentiality of a gene *g* is defined as the maximum of all its predictions (equations 9 and 10). This assumption, however, is very susceptible to false positive predictions (not essential genes predicted as essential) that can have a huge influence in the obtained results. To address this issue, we defined the *Essential Congruity Score* (ECS) as:

$$ECS_g = \frac{\sum_{\nabla p \in \boldsymbol{P_g}} (\sum_{\nabla a \in \overline{\boldsymbol{A_p}}} C_{g,p,a})}{\sum_{\nabla p \in \boldsymbol{P_g}} (|\boldsymbol{A_p}|)} \quad (11)$$

Where ECS_g is the Essential Congruity Score for the gene *g*, $\boldsymbol{P_g}$ represents the set of pathways in which the gene *g* is present, $\overline{\boldsymbol{A_p}}$ is the set of actives for the pathway *p* with at least one prediction of essentiality, and $C_{g,p,a}$ is the prediction of essentiality for the gene *g*, in the pathway *p*, and for the active *a*. $ECS_g = 0$ means that in none of the instances the gene *g* was predicted as essential while $ECS_g = 1$ means that in 100% of the predictions the gene *g* was essential.

Results

In this section, we show the results obtained with the proposed methodology in a different set of scenarios: 1) a simple toy example showing the key conceptual aspects of the methodology and the functioning of the equations; 2) a case study using the gene essentiality data from the Achilles project illustrating the biological validity of the obtained results; 3) a breast cancer dataset which results are validated in the literature.

Toy Example

First, we considered a simplification of the *Wnt receptor signaling pathway, planar cell polarity pathway*, which is shown in **Figure 3** (Schaefer *et al.,* 2009). The simplified subnetwork comprises four genes (WNT5A, FZD7, WNT3A and FCD1), two complexes (WNT5A/FZD7 and WNT3A/FZD1) and one abstract (Wnt receptor signaling pathway, planar cell polarity pathway). As mentioned earlier, the methodology comprises two main steps: (i) calculating the minimum number of non-expressed genes that we need to activate in order to trigger a given active a (S_a^{wild}) and (ii) performing an exhaustive *in-silico* gene knockout to find deletions that unavoidably lead to the need of activating extra non-expressed genes in order to trigger the given entity ($S_a^g > S_a^{wild}$).

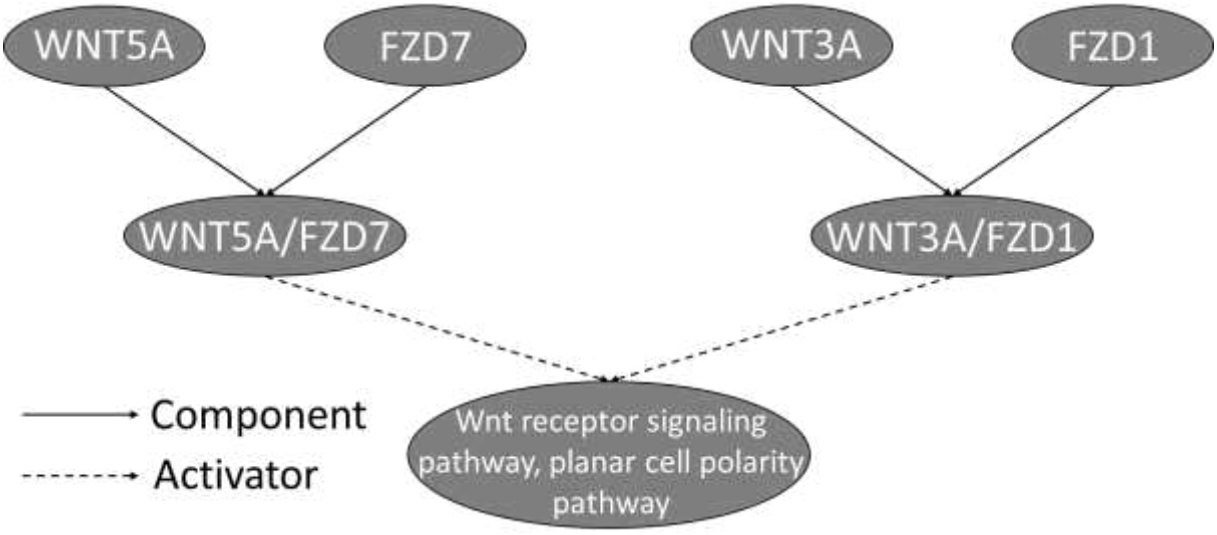

Figure 3 - Toy Example. *Graphical representation of the pathway activating the Wnt receptor signaling pathway, planar cell polarity pathway. Component-type interactions are represented with solid arrows whilst activation-type interactions are illustrated with dashed lines.*

In the forthcoming lines we will define three scenarios based on simulated data. These scenarios show different solutions based on whether WNT5A and WNT3A are expressed or not while FZD7 and FZD1 are always expressed i.e., FZD7 & FZD1 ∈ **G**. Table 1 summarises the solution

of the different proposed scenarios. The complete solution of the mathematical model for each scenario is included in Supplementary Results 1.

Table 1- Toy example solution. *Possible scenarios when FZD7 and FZD1 are expressed. For each scenario, the expression values of each gene and the activity values of each entity are included. W5F7C represents the WNT5A/FZD7 complex, W3F1C represents the WNT3A/FZD1 complex, and ABSTR represents the Wnt receptor signaling pathway, planar cell polarity pathway. A) WNT5A and WNT3A are not expressed. For the abstract to be active we need to activate one non-expressed gene (WNT3A in the example). B) WNT5A is not expressed and WNT3A is expressed. For the abstract to be active we do not need to activate any non-expressed gene. C) WNT5A is expressed and WNT3A is not expressed. For the abstract to be active we do not need to activate any non-expressed gene. B') Scenario B after FZD7 is knocked-out. For the abstract to be active we do not need to activate any non-expressed gene. C') Scenario C after FZD7 is knocked-out. For the abstract to be active we need to activate one non-expressed gene (WNT3A).*

Scenario	Gene Expression				Entity Activity						
	WNT5A	FZD7	WNT3A	FZD1	WNT5A	FZD7	WNT3A	FZD1	W5F7C	W3F1C	ABSTR
A	0	1	0	1	0	0	1	1	0	1	1
B	0	1	1	1	0	0	1	1	0	1	1
C	1	1	0	1	1	1	0	0	1	0	1
B'	0	1	1	1	0	0	1	1	0	1	1
C'	1	1	0	1	0	0	1	1	0	1	1

(i) Minimum number of non-expressed genes required to activate an entity

The scenario shown in **Figure 4.A** has two expressed genes (FZD7 and FZD1) and two non-expressed genes (WNT5A and WNT3A). For the abstract to be active, one of the two complexes needs to be active. The condition for either complex is that both of its gene components need to be active. Thus, in scenario A we need to activate one non-expressed gene (WNT5A or WNT3A) for the abstract to be active ($S^{wild}_{Abstract} = 1$). Scenarios B and C do not require the activation of any non-expressed gene to activate the abstract and therefore $S^{wild}_{Abstract} = 0$ (for the complete solution, please refer to Supplementary Results 1).

(ii) in-silico exhaustive gene knockout

In the scenario shown in **Figure 4.B**, a knock-out in FZD7 ($E_{FZD7} = 0$) does not require the activation of any non-expressed gene because the abstract can be activated through the WNT3A/FZD1 complex and both its components are expressed, that is $S^{wild}_{Abstract} = S^{g}_{Abstract} = 0$. Therefore, in this scenario, FZD7 is not an essential gene. On the other hand, if we consider the scenario described in **Figure 4.C**, a knockout in FZD7 means that the WNT5A/FZD7 complex cannot be active, and thus the abstract needs to be activated via the WNT3A/FZD1 complex which requires the activation of one non-expressed gene (WNT3A). In this scenario, $S^{g}_{Abstract} = 1$, while $S^{wild}_{Abstract} = 0$ and therefore FZD7 is considered an essential gene.

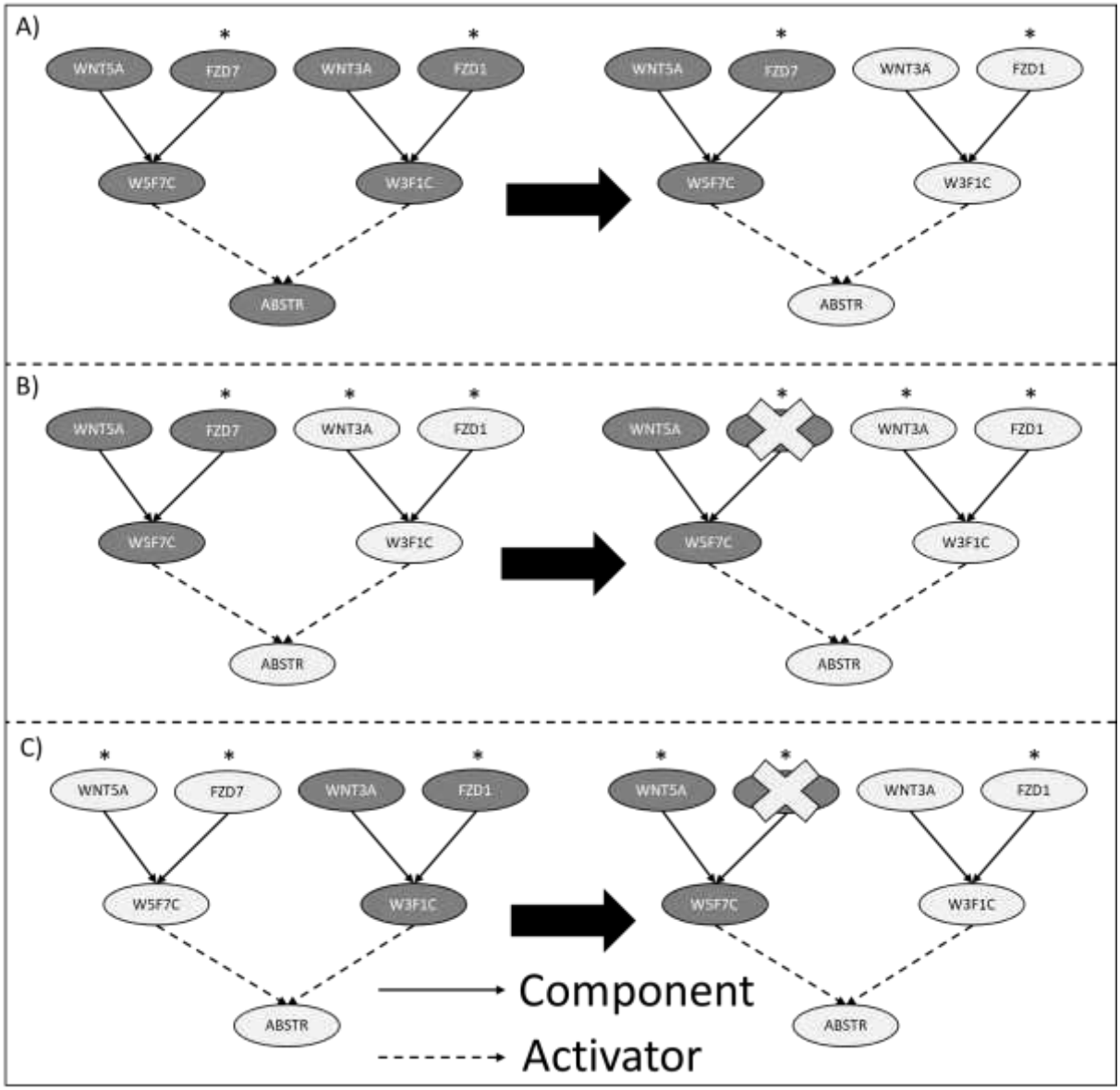

Figure 4 - Toy example solution*. Possible scenarios when FZD7 and FZD1 are expressed. W5F7C represents the WNT5A/FZD7 complex, W3F1C represents the WNT3A/FZD1 complex, and ABSTR represents the Wnt receptor signaling pathway, planar cell polarity pathway.* Dark and light nodes represent inactive and active nodes in the final solution respectively, namely $E_i = 0$ and $E_i = 1$. The star on top of a gene g represents the expression state, in particular if g has a star it is expressed, namely $g \in G$, and not expressed otherwise, i.e. $g \in L$. A) WNT5A and WNT3A are not expressed. For the abstract to be active we need to activate one non-expressed gene (WNT3A in the example). B) WNT5A is not expressed and WNT3A is expressed. For the abstract to be active we do not need to activate any non-expressed gene. A knock-out of FZD7 does not require the activation of any non-expressed gene for the abstract to be active. C) WNT5A is expressed and WNT3A is not expressed. For the abstract to be active we do not need to activate any non-expressed gene. A knock-out of FZD7 requires the activation of one non-expressed gene (WNT3A) for the abstract to be active.

Method Validation

To validate the biological relevance of the gene essentiality predictions of our method, for a given cell line, we compared the Achilles scores of the genes g predicted as essential (C_g=1) versus the scores of the genes predicted as not essential (Cg=0) (**Figure 5.A**). For this analysis, globally essential (genes predicted as essential in all cell lines) and globally not essential genes (genes not

essential in all the cell lines) were not included in the analysis (Methods section). This reduced the number of genes included in the comparison to 159.

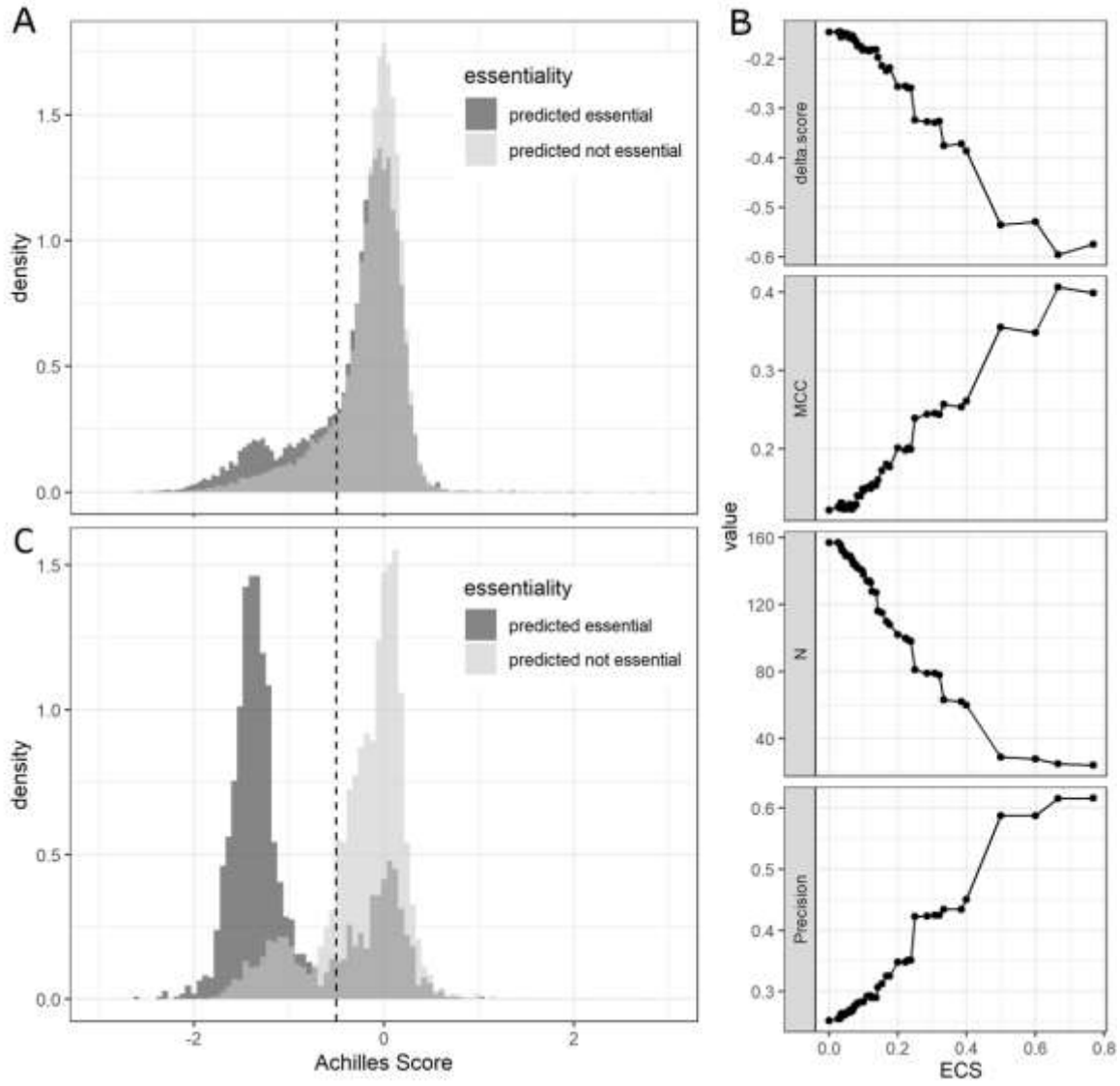

***Figure 5* - Method validation.** A) Histogram showing the results from the validation of the method. The dark distribution shows the Achilles scores of those pair gene & cell-lines predicted as essential; the light distribution shows the Achilles scores of those predicted as not essential. Genes predicted as essential have significantly lower Achilles score than genes predicted as not essential (p-value = 6.4032^{-246}). The average difference between both distributions is defined by the parameter delta.score =-0.1463. B) Impact of ECS in the performance of the method. Evolution of the results when different thresholds of ECS are used to define a gene as essential. delta.score: average difference in Achilles score between the genes predicted as essential and the genes predicted as not essential; MCC: Matthew's Correlation Coefficient; N: number of genes included in the comparison; Precision: obtained precision assuming as real essential genes those with an Achilles score < -0.5. C) Histogram when MCC finds its maximum (ECS = 0.6667). The average difference in Achilles Score between genes predicted as essential and genes predicted as not essential becomes bigger (delta.score = -0.5954) and so does their significance (p-value = 0).

Figure 5.A shows how the genes predicted as essential have a significantly lower Achilles score than the genes predicted as not essential (p-value = 6.4032^{-246}). The results illustrated in **Figure 5.A** follow the definition of essentiality represented in equations 9 and 10 of the methods where a gene is considered essential if is predicted as essential for any active in any of the pathways where it

appears. However, the ECS defined in equation 11 is a continuous score ($ECS \in [0,1]$) and allows to describe flexible threshold when defining the essentiality of a gene. For example, we can define genes as essential if their ECS is larger than a given threshold th ($C_g = 1, if\ ECS_g > th$). We studied the impact of applying different thresholds to the ECS by evaluating the evolution of the obtained results (**Figure 5.B**). For this analysis, we defined a gene as essential for a given cell-line if its Achilles score was below -0.5.

Figure 5.B shows how as the minimum ECS required to consider a gene as essential increases, so does the quality of the predictions. Most of the statistics shown in the different subfigures improve their performance when more demanding values of ECS are needed to define a gene as essential. We defined as the optimal cut-off the ECS where the MCC parameter finds its maximum (ECS = 0.67, MCC = 0.41, **Figure 5.C).** We selected the MCC because it has been proven to be the most robust metric for imbalanced data (Boughorbel et al, 2017). However, as the minimum required threshold increases, so does the number of genes considered globally not essential which decreases the number of genes included in the analysis (represented by N). **Figure 5.B** also shows the monotonically increasing behaviour of the Precision curve when the minimum required ECS to define essentiality increases. This is particularly interesting for reducing experimental validation costs, as we want to make sure that genes predicted as essential are indeed essential while genes predicted as not essential are not as relevant.

Synergistic behaviour of the method

This gene essentiality method finds its success on the synergy between three different factors: 1) biologically relevant gene expression data, 2) a robust prior-knowledge-network (PKN), and 3) the mathematical formulation described in the methods section. Alterations in each of these fundamental pillars affect downstream results increasing the number of false positive predictions. To test the first pillar, biologically relevant gene expression data, we fed the method with "nonsense" expression data by inverting the binary scores obtained from The Gene Expression Barcode 3.0 (McCall et al., 2012), (McCall et al., 2014). This reduced the maximum MCC (starting from a baseline of 0.41 using a ECS of 0.67) to 0.1 (using a ECS of 0.5). To validate the need of a representative prior-knowledge network we repeated the analysis using only the subset of 50 NCI-PID pathways that were labelled as tumorigenic which increased the maximum MCC to 0.53 (using a ECS of 0.67). Finally, we evaluated that this improvement in MCC was only present when the gene expression data was biologically meaningful. To that end, we repeated the analysis using the subset of NCI-PID pathways and the "nonsense" expression data obtaining a MCC of 0.12 (using a ECS of 0.9). The reader should refer to Supplementary Results 2 for the complete evaluation. When

compared with other state of the art methods (Cubuk et al., 2018) our method produces less false positives (Supplementary Results 2).

Case Study – Breast Cancer

Finally, we applied the gene essentiality method to Breast Cancer patient samples (Maubant S, 2012), (Maire V, 2013) and looked for genes significantly predicted as essential in cancer patients using hypergeometric tests. For this purpose, technical duplicates were considered as independent samples. A gene was considered essential for a given patient if ECS > 0. The same procedure was repeated for the different cancer subtypes.

Table 2 shows the top 10 results for the Healthy vs BRCA case while Supplementary Table 2 shows the top 10 results for the group-specific comparison. The complete lists can be found in Supplementary Tables 3 and 4. In the following lines, we will highlight the relevance of the top 4 (elbow criterion) genes reported in **Table 2** with a higher coverage of patients by relying on existing knowledge in the literature.

Table 2 **- BRCA essentiality results.** geneID: gene identifier in HGNC nomenclature; p.hyper: p-value from a hypergeometric test; p.adj: adjusted p-value after multiple-testing comparison, # cancer essential: number of cancer samples predicted as essential; # total essential: number of total samples predicted as essential; Achilles score: average Achilles score across all the BRCA cell lines from the CCLE. Number of cancer samples = 153, Total number healthy samples = 11.

geneID	p.hyper	p.adj	# cancer essential	# total essential	Achilles score
RACGAP1	3.25E-07	4.91E-05	118	118	-1.45764
MIB1	2.33E-06	1.76e-04	110	110	-0.34808
EZR	1.02E-05	5.13e-04	103	103	0.094074
PCNA	1.94E-05	7.34e-04	126	128	-1.88689
TUBG1	1.40e-04	4.21e-03	88	88	-1.31599
CASP3	5.20e-04	9.81e-03	79	79	0.073582
PKCDELTA	5.20e-04	9.81e-03	79	79	NA
SDC2	5.20e-04	9.81e-03	79	79	-0.06373
BIRC3	1.68e-03	2.31e-02	89	90	0.149018
GNAI1	1.68e-03	2.31e-02	70	70	-0.10703

RACGAP1: Rac GTPase-activating protein 1. RACGAP1 is a protein involved in several biological processes including cell cycle, cell division, and differentiation and with a key role in various cellular phenomena including cytokinesis, invasive migration and metastasis. Increased expression of RACGAP1 protein has been previously associated with poor survival as well as significantly associated with increased tumour malignancy in colorectal cancer (Imaoka et al, 2015).

It has been shown that its knockdown – in combination with radiotherapy – is associated with a decrease of tumour viability and invasiveness in 4T1 mouse models (Wu et al, 2019).

PCNA: Proliferating cell nuclear antigen. PNCA is a protein involved in DNA replication by increasing the processivity of DNA polymerase delta. Immunohistochemical staining of PCNA has been used extensively in breast cancer diagnosis and prognosis (Malkas et al, 2006). It has been shown that targeting the EGFR/PCNA signalling suppresses tumour growth of triple-negative breast cancer cells (Yu et al, 2013) and inhibit cancer growth in neuroblastoma and breast cancer mouse xenograft models (Choe et al, 2017).

MIB1: Mindbomb E3 ubiquitin protein ligase 1. MIB1 is a protein that positively regulates Notch signaling by ubiquitinating the Notch receptors, thereby facilitating their endocytosis. It has been shown that MicroRNA-198 suppresses prostate tumorigenesis by targeting MIB1 (Ray et al, 2019). MicroRNA-198 also represses cell proliferation and migration and promotes cell adhesion in breast cancer cells (Hu et al 2017).

EZR: Ezrin. EZR is protein that plays a key role in cell surface structure adhesion, migration and organization. Its inhibition synergizes with lapatinib in a PKC-dependent fashion to inhibit proliferation and promote apoptosis in HER2-positive breast cancer cells (Jeong et al, 2019). EZR inhibition in hepatocellular carcinoma (HCC) cells decreases their migratory and invasive potential (Zhang et al, 2006).

Discussion

This article introduces a new methodology for the *in-silico* identification of essential genes which integrates high-throughput gene expression data with predefined biological pathways to provide patient-specific gene essentiality predictions. This method uses a mathematical formulation that identifies the number of non-expressed genes required to be active for the cell to sustain life, here modelled by the activation of a relevant biological task. This work expands the ideas behind existing CBM-based methodologies going beyond metabolism by considering multisystem networks (Schaefer et al., 2009).

We have validated the proposed methodology using a set of 452 cancer cell lines derived from the Cancer Cell Line Encyclopedia where the essential genes had been previously identified using CRISPR knockouts (Achilles Project). When compared to competing methods, our approach identifies essential genes with fewer false positives. Because cell-lines do not represent the entire complexity of cancer, we have further supported the obtained essential genes in an independent breast cancer dataset using existing literature.

The mathematical formulation presented in the methods section makes it possible to have several predictions of essentiality for the same gene. Due to the nature of the problem, initially a single prediction of essentiality was a sufficient condition to consider the gene as essential thus these multiple predictions were summarized into their maximum for each gene. This summarization is very susceptible to false positive results which can have a huge impact in downstream results. We have shown how the integration of multiple predictions into the *Essentiality Congruity Score* (ECS) improves our ability to identify essential genes.

The presented methodology finds its success on the synergy between its three core constituents: biologically relevant gene expression data, a robust prior-knowledge-network that effectively captures cancer biological events and the constraint-based mathematical model described in the methods section.

We have demonstrated that all three elements are necessary by modifying individual constituents. We have proven that missense input data (produced by inverting the discrete expression values) does not yield to valid results. We have also shown that including pathways that do not represent tumorigenic events worsen the essentiality predictions. Finally, we have proven how diluting the impact of false positive predictions derived from the methodology using the ECS further improves the precision when identifying essential genes.

The mathematical formulation described in the methods section distinguishes between expressed genes and non-expressed genes. This discrimination, however, is derived from continuous gene expression data, which was previously discretized using The Gene Expression Barcode 3.0 (McCall et al, 2012), (McCall et al, 2014). This work does not directly tackle this issue but the selection of discretization strategy can have a tremendous impact on downstream results.

The present methodology assumes that all the actives (abstracts + complexes) included in the PKN are equally relevant for the cell to sustain life. This represents an oversimplification of the reality as not all the actives will affect the cell in the same way. We have shown that removing pathways that do not capture tumorigenic events improve the obtained results demonstrating that there needs to exist harmony between the biological network and the mathematical model.

The advent of *in-silico* approaches predicting essential genes will pave the way for precision medicine by identifying potential drug targets whose deletion can induce death in tumour cells (Tsherniak *et al.,* 2017). The work presented here contributes in this direction. However, further efforts are required to develop disruptive *in-silico* methodologies that accounts for further biophysical knowledge, such as dynamic models or multi-omics data. Overcoming this ambitious challenge will set the foundations for addressing biological questions that were unreachable before.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The gene essentiality scores from the Achilles project were downloaded from the DepMap (CRISPR knockouts, CERES method, version 20Q1): https://depmap.org/portal/download/.

The gene expression data for the cell-lines included in the Cell Line Encyclopedia (CCLE) was obtained from the Gene Expression Omnibus (GEO), accession number: GSE36133

The breast cancer dataset used as case study was obtained from the Gene Expression Omnibus under accession number GSE65194.

Conflict of Interest

The authors declare no competing interest with the presented work.

Funding

Not applicable

Authors' Contributions

AA: conceptualization, formal analysis, interpretation of the results, writing of the manuscript, preparation of the figures, revision.

JP: conceptualization, method implementation, interpretation of the results, writing of the manuscript, revision.

MO: conceptualization, interpretation of the results, revision.

Acknowledgements

The authors would like to thank PhD Matthew Trotter and PhD Manuel Sanchez Castillo for the fruitful discussions and the useful suggestions and comments.

References

Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M., & Nielsen, J. (2014). Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Molecular systems biology, 10(3), 721

Apaolaza, I., San José-Eneriz, E., Tobalina, L., Miranda, E., Garate, L., Agirre, X., ... & Planes, F. J. (2017). An in-silico approach to predict and exploit synthetic lethality in cancer metabolism. Nature communications, 8(1), 1-9

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D. and Reddy, A., 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483(7391), pp.603-607.

Bartha, I., di Iulio, J., Venter, J. C., & Telenti, A. (2018). Human gene essentiality. Nature Reviews Genetics, 19(1), 51.

Boughorbel, S., Jarray, F. and El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS one, 12(6), p.e0177678.

Choe, K.N. and Moldovan, G.L., 2017. Forging ahead through darkness: PCNA, still the principal conductor at the replication fork. Molecular cell, 65(3), pp.380-392.

Cubuk, C., Hidalgo, M.R., Amadoz, A., Pujana, M.A., Mateo, F., Herranz, C., Carbonell-Caballero, J. and Dopazo, J., 2018. Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape. Cancer research, 78(21), pp.6059-6072.

Dempster, J.M., Rossen, J., Kazachkova, M., Pan, J., Kugener, G., Root, D.E. and Tsherniak, A., 2019. Extracting biological insights from the Project Achilles genome-scale CRISPR screens in cancer cell lines. BioRxiv, p.720243.

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., & Altman, R. B. (2011). Bioinformatics challenges for personalized medicine. Bioinformatics, 27(13), 1741-1748

Frezza, C., Zheng, L., Folger, O., Rajagopalan, K. N., MacKenzie, E. D., Jerby, L., ... & Kalna, G. (2011). Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. Nature, 477(7363), 225-228.

Hu, Y., Tang, Z., Jiang, B., Chen, J. and Fu, Z., 2017. miR-198 functions as a tumor suppressor in breast cancer by targeting CUB domain-containing protein 1. Oncology Letters, 13(3), pp.1753-1760.

Imaoka, H., Toiyama, Y., Saigusa, S., Kawamura, M., Kawamoto, A., Okugawa, Y., Hiro, J., Tanaka, K., Inoue, Y., Mohri, Y. and Kusunoki, M., 2015. RacGAP1 expression, increasing tumor malignant potential, as a predictive biomarker for lymph node metastasis and poor prognosis in colorectal cancer. Carcinogenesis, 36(3), pp.346-354.

Jeong, J., Choi, J., Kim, W., Dann, P., Takyar, F., Gefter, J.V., Friedman, P.A. and Wysolmerski, J.J., 2019. Inhibition of ezrin causes PKCα-mediated internalization of erbb2/HER2 tyrosine kinase in breast cancer cells. Journal of Biological Chemistry, 294(3), pp.887-901.

Kuang, S., Wei, Y., & Wang, L. (2020). Expression-based prediction of human essential genes and candidate lncRNAs in cancer cells. Bioinformatics

Li, X., Li, W., Zeng, M., Zheng, R., & Li, M. (2020). Network-based methods for predicting essential genes or proteins: a survey. Briefings in bioinformatics, 21(2), 566-583.

Maire, V., Baldeyron, C., Richardson, M., Tesson, B., Vincent-Salomon, A., Gravier, E., Marty-Prouvost, B., De Koning, L., Rigaill, G., Dumont, A. and Gentien, D., 2013. TTK/hMPS1 is an attractive therapeutic target for triple-negative breast cancer. PloS one, 8(5), p.e63712.

Malkas, L.H., Herbert, B.S., Abdel-Aziz, W., Dobrolecki, L.E., Liu, Y., Agarwal, B., Hoelz, D., Badve, S., Schnaper, L., Arnold, R.J. and Mechref, Y., 2006. A cancer-associated PCNA expressed in breast cancer has implications as a potential biomarker. Proceedings of the National Academy of Sciences, 103(51), pp.19472-19477.

Maubant, S., Tesson, B., Maire, V., Ye, M., Rigaill, G., Gentien, D., Cruzalegui, F., Tucker, G.C., Roman-Roman, S. and Dubois, T., 2015. Transcriptome analysis of Wnt3a-treated triple-negative breast cancer cells. PloS one, 10(4), p.e0122333.

Martin, R. K. (2012). Large scale linear and integer optimization: a unified approach. Springer Science & Business Media.

McCall, M.N., Jaffee, H.A. and Irizarry, R.A., 2012. fRMA ST: frozen robust multiarray analysis for Affymetrix Exon and Gene ST arrays. Bioinformatics, 28(23), pp.3153-3154.

McCall, M.N., Jaffee, H.A., Zelisko, S.J., Sinha, N., Hooiveld, G., Irizarry, R.A. and Zilliox, M.J., 2014. The Gene Expression Barcode 3.0: improved data processing and mining tools. Nucleic acids research, 42(D1), pp.D938-D943.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S. and Goodale, A., 2017. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nature genetics, 49(12), pp.1779-1784.

Pey, J., San José-Eneriz, E., Ochoa, M. C., Apaolaza, I., de Atauri, P., Rubio, A., ... & Carracedo, A. (2017). In-silico gene essentiality analysis of polyamine biosynthesis reveals APRT as a potential target in cancer. Scientific reports, 7(1), 1-10

Plata, G., Hsiao, T. L., Olszewski, K. L., Llinás, M., & Vitkup, D. (2010). Reconstruction and flux-balance analysis of the Plasmodium falciparum metabolic network. Molecular systems biology, 6(1), 408

Ray, J., Hoey, C., Huang, X., Jeon, J., Taeb, S., Downes, M.R., Boutros, P.C. and Liu, S.K., 2019. MicroRNA-198 suppresses prostate tumorigenesis by targeting MIB1. Oncology reports, 42(3), pp.1047-1056.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., & Buetow, K. H. (2009). PID: the pathway interaction database. Nucleic acids research, 37, D674-D679.

Schapke, J., Tavares, A., & Recamonde-Mendoza, M. (2020). EPGAT: Gene Essentiality Prediction With Graph Attention Networks. arXiv preprint arXiv:2007.09671

571 Schrijver, A. (1998). Theory of linear and integer programming. John Wiley & Sons

572 Tobalina, L., Pey, J., Rezola, A., & Planes, F. J. (2016). Assessment of FBA based gene essentiality
573 analysis in cancer with a fast context-specific network reconstruction method. PloS one, 11(5),
574 e0154583

575 Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., ... &
576 Meyers, R. M. (2017). Defining a cancer dependency map. Cell, 170(3), 564-576.

577 Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., ... & Stuart, J. M. (2010).
578 Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using
579 PARADIGM. Bioinformatics, 26(12), i237-i245.

580 Wu, P.H., Onodera, Y., Recuenco, F.C., Giaccia, A.J., Le, Q.T., Shimizu, S., Shirato, H. and Nam,
581 J.M., 2019. Lambda-Carrageenan Enhances the Effects of Radiation Therapy in Cancer Treatment
582 by Suppressing Cancer Cell Invasion and Metastasis through Racgap1 Inhibition. Cancers, 11(8),
583 p.1192.

584 Yan, J., Risacher, S. L., Shen, L., & Saykin, A. J. (2018). Network approaches to systems biology
585 analysis of complex disease: integrative methods for multi-omics data. Briefings in bioinformatics,
586 19(6), 1370-1381

587 Yu, Y.L., Chou, R.H., Liang, J.H., Chang, W.J., Su, K.J., Tseng, Y.J., Huang, W.C., Wang, S.C.
588 and Hung, M.C., 2013. Targeting the EGFR/PCNA signaling suppresses tumor growth of triple-
589 negative breast cancer cells with cell-penetrating PCNA peptides. PLoS One, 8(4), p.e61362.

590 Zhang, R., & Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and
591 eukaryotes. Nucleic acids research, 37(suppl_1), D455-D58

592 Zhang, Y., Hu, M.Y., Wu, W.Z., Wang, Z.J., Zhou, K., Zha, X.L. and Liu, K.D., 2006. The
593 membrane-cytoskeleton organizer ezrin is necessary for hepatocellular carcinoma cell growth and
594 invasiveness. Journal of cancer research and clinical oncology, 132(11), pp.685-697.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AntoranzetalELIMINATORSupplementaryResults.docx
- SupplementaryTable1Pathwayannotations.xlsx
- SupplementaryTable2Top10groupspecificcomparison.xlsx
- SupplementaryTable3CancervsHealthy.xlsx
- SupplementaryTable4Cancergroups.xlsx