





# Review and meta-analysis of the genetic Minimal Cut Set approach for gene essentiality prediction in cancer metabolism

Danel Olaverri-Mendizabal , Luis V. Valcárcel , Naroa Barrena , Carlos J. Rodríguez and Francisco J. Planes 

Corresponding author: Francisco J. Planes, Instituto de Ciencia de los Datos e Inteligencia Artificial (DATAI), University of Navarra, Campus Universitario, Pamplona 31080, Spain. Tel.: +34-943-219877; Fax: +34-943-311442; E-mail: [fplanes@tecnun.es](mailto:fplanes@tecnun.es)

## Abstract

Cancer metabolism is a marvellously complex topic, in part, due to the reprogramming of its pathways to self-sustain the malignant phenotype in the disease, to the detriment of its healthy counterpart. Understanding these adjustments can provide novel targeted therapies that could disrupt and impair proliferation of cancerous cells. For this very purpose, genome-scale metabolic models (GEMs) have been developed, with Human1 being the most recent reconstruction of the human metabolism. Based on GEMs, we introduced the genetic Minimal Cut Set (gMCS) approach, an uncontextualized methodology that exploits the concepts of synthetic lethality to predict metabolic vulnerabilities in cancer. gMCSs define a set of genes whose knockout would render the cell unviable by disrupting an essential metabolic task in GEMs, thus, making cellular proliferation impossible. Here, we summarize the gMCS approach and review the current state of the methodology by performing a systematic meta-analysis based on two datasets of gene essentiality in cancer. First, we assess several thresholds and distinct methodologies for discerning highly and lowly expressed genes. Then, we address the premise that gMCSs of distinct length should have the same predictive power. Finally, we question the importance of a gene partaking in multiple gMCSs and analyze the importance of all the essential metabolic tasks defined in Human1. Our meta-analysis resulted in parameter evaluation to increase the predictive power for the gMCS approach, as well as a significant reduction of computation times by only selecting the crucial gMCS lengths, proposing the pertinency of particular parameters for the peak processing of gMCS.

**Keywords:** synthetic lethality; genetic Minimal Cut Sets; genome-scale metabolic models; constraint based modelling; gene essentiality analysis

## INTRODUCTION

### Synthetic lethality and genetic minimal cut sets

Precision and targeted medicine have emerged as compelling topics in cancer research. Their main challenge is the development of novel treatments that selectively target malignant cells while sparing healthy ones in order to increase therapy sensitivity and decrease side-effects, leading to a better quality of life for patients. A promising approach to achieving this goal is synthetic lethality (SL), which refers to the interaction between two genes in which a perturbation, such as a mutation, RNA interference knockdown, or inhibition, affecting either gene alone does not imply a loss in cell viability; however, the perturbation of both genes simultaneously results in lethality [1]. SL provides a promising avenue for

developing targeted therapies for cancer, and several approaches have been used to detect SL in different cancer types, including *in vitro* studies [2] and computational methods [3, 4].

Network-based approaches to predict SL in cancer cells have received much attention in the field of systems biology [5]. In particular, constraint-based modelling, a computational framework for the analysis of genome-scale metabolic networks, has experienced great advance in the last decade to study cancer metabolism [6–10]. Among existing approaches, the genetic Minimal Cut Sets (gMCSs) approach constitutes a unique strategy that directly connects with the concept of SL [11]. The gMCS approach extends the concept of SL to two or more genes whose deletion disables a key metabolic task in the reference (uncontextualized)

**Danel Olaverri-Mendizabal** is a Biomedical Engineer with an MSc in Biomedical Engineering (Data Analytics) from the University of Navarra. He is currently pursuing his PhD in Computational Biology at Tecnum, University of Navarra.

**Luis V. Valcárcel** is an Industrial Engineer with an MSc in Biomedical Engineering (Data Analytics) from the University of Navarra. He received his PhD in Computational Biology from the University of Navarra.

**Naroa Barrena** is a Biomedical Engineer with an MSc in Biomedical Data Analytics from the University of Navarra. She is currently pursuing her PhD in Computational Biology at Tecnum, University of Navarra.

**Carlos J. Rodríguez** is a Guatemalan Biomedical Engineer with an MSc in Biomedical Data Analytics from the University of Navarra. He is currently pursuing his PhD in Computational Biology at Tecnum, University of Navarra.

**Francisco J. Planes** PhD, is an Industrial Engineer from the University of Navarra. He received his PhD from the Brunel University (UK). He has been working in Computational Biology since 2009. His research interests are precision medicine and systems biology, focusing on cancer and gut microbiota research.

**Received:** November 22, 2023. **Revised:** February 15, 2024. **Accepted:** February 26, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

network of human cells. This key metabolic task is usually the biomass production, an artificial reaction which defines all the metabolic requirements for cell proliferation. Following the concept of SL, gMCSs can be used to identify cancer-specific essential genes (targeted therapies) based on available omics data. We describe the evolution of the concept of gMCS and more recent advancements.

## The evolution of the gMCS framework

In the initial iteration of the gMCS framework [11], microarray gene expression data were utilized to identify a set of genes with low expression levels within a sample using BARCODE. Subsequently, potential target genes were examined, with emphasis placed on gMCSs that comprised both a highly expressed gene and a collection of genes exhibiting low expression levels. These gMCSs were searched to find those that can be inactivated with a single gene knockout. Target selectivity could be achieved by focusing on gMCSs where only one gene is active in the malignant cell, while the healthy counterpart has multiple activated genes. In their analysis, they established a possible mechanism to explain the essentiality of RRM1 in cancer. The caveat on this study was the limited computational efficiency in the search strategy of gMCSs and essential genes. This issue was addressed in a subsequent work [12], where a more sophisticated algorithm was developed to conduct a global search of gMCSs in metabolic networks.

Afterwards, Valcárcel et al. [13] took the framework one step further and implemented an user-friendly online tool, *gmctool*, that comprises the following advancements: the use of latest human genome-scale metabolic network reference, Human1 [9], novel gene expression thresholding techniques, double knockout lethality analysis, use of RNA-Seq data instead of microarray data, faster computation times and integration with DepMap information. Following the strategy introduced in Apaolaza et al., *gmctool* integrates more than 57,000 unique gMCSs of diverse length associated with biomass production. Furthermore, they computed gMCSs that are necessary for other fundamental metabolic tasks, such as *oxidative phosphorylation*, *uptake of essential amino acids*, *beta oxidation of fatty acids*, *protein turnover*, etc. In total, 57 different tasks, including the production of biomass, are considered that can be blocked in order to define SL groups within the application. The sum of the gMCSs for all tasks tallies to 97,000 distinct gMCSs.

Candidate genes partaking in the distinct biological tasks are selected using a similar approach anew. This involves identifying genes that are the unique expressed gene in a given gMCS for the malignant condition, whilst the corresponding healthy cells has multiples genes expressed for that gMCS. Using the developed *gmctool*, Valcárcel et al. demonstrated the essentiality of CTPS1 in a sub-group of multiple myeloma patients.

The online application, *gmctool*, stores all the computed gMCSs, in which it projects the RNA-Seq gene expression data, uploaded by the user, with customizable parameters, such as the essential metabolic tasks included in the gMCSs or the threshold used for weighing a gene as highly or lowly expressed, for a tailored and thorough gene essentiality analysis.

## Novel gMCS advancements

Altogether, cancer metabolism is complex, as evidenced by the emerging field of research on the tumour microenvironment, which encompasses not only the tumoral but also the surrounding cells. Recent studies have focused on these interactions to gain a deeper understanding of the impact of microenvironment on tumour progression [14]. In this regard, Apaolaza et al. [15],

developed a generalization to the gMCS algorithm to determine an innovative group of metabolic synthetic lethal interactions that integrate nutritional perturbations in the surrounding medium, called *nutrient-gene Minimal Cut Sets* (ngMCSs). Instead of only having genetic interactions, this new paradigm would make possible to have a list of synthetic lethals formed by both gene knockouts and metabolite deprivations from the surrounding medium. In that work, they successfully prove the essentiality of DHFR, subject to the lack of the metabolites thymidine and hypoxanthine in the growth medium, whose presence rescues cellular proliferation.

As promising as gMCSs can be, one major limitation is that they are confined to the metabolic reactions and genes, blind to the impact transcription factors have on reprogramming metabolic networks [16]. On a cutting-edge study, Barrera et al. [17] have integrated regulatory pathways on a metabolic network based on Boolean networks. Using well-known and different regulatory network databases, they assessed gene essentiality prediction in *in vitro* gene silencing data with promising results: an increase in the predicted number of essential genes in integrated metabolic and regulatory to the pure metabolic model, as well as the incorporation of key signalling genes to the study.

Taken together, these three advancements are built upon the same concept first studied by Apaolaza et al. [11]. Examples of the aforementioned algorithms, as well as the basic formulation needed for the computation of gMCSs are thoroughly explained in Supplementary Note 1. These studies show the relevance of gMCSs in SL prediction and gene essentiality analysis.

## Availability of gMCS algorithms and tools

A short description of different gMCS algorithms and tools, as well as their software and code availability, is summarized in Table 1. Different algorithms for calculating gMCSs and ngMCSs have been integrated into The COBRA Toolbox, a MATLAB software suite for the analysis of genome-scale metabolic networks [18]. In fact, a short tutorial can be found in The COBRA Toolbox for gMCSs: <https://opencobra.github.io/cobratoolbox/stable/tutorials/tutorialGMCS.html>. In addition, we recently released an open source Python package, called *gMCSpy*, which overcomes the need of commercial software for computing gMCSs while improving the computational performance of previous algorithms [19].

## Motivation: Fine-tuning the gMCS approach for gene essentiality analysis

To tailor subsequent studies based on gMCS with finely tuned parameters, this work reviews and interrogates the gMCS methodology to make the best predictions on gene essentiality based on several criteria: the optimal threshold that should be considered to classify the genes as highly or lowly expressed; the importance of the length of a gMCS (also called order) for their predictive value, the partaking of target genes in several gMCSs or essential tasks, and, lastly, the relevance of other essential tasks beyond the biomass production.

Our analysis was conducted using a small but curated cell line cohort from Hart et al. [20], hereafter referred to as Hart2015, as well as data from CCLE and the Cancer Dependency Map [21], referred to as DepMap (see Methods section). As a result, we aim to reveal the optimal threshold for gene expression levels. In addition, we investigate how the length of a gMCS and the number of essential tasks performed by a gene impact the accuracy of the predictions of gene essentiality. These findings can guide users in

**Table 1:** Summary of concepts, algorithms and tools in gMCS approach

Name	Description	Software and code availability	Reference
gMCSs	Minimal subsets of gene knockout interventions in metabolic networks.	MATLAB—The COBRA toolbox, function name: calculateGeneMCS	[11, 12]
gmctool	Online tool for gene essentiality analysis based on the gMCS approach, Human1 and RNA-seq data.	The tool is freely available at: <a href="https://biotecnun.unav.es/app/gmctool">https://biotecnun.unav.es/app/gmctool</a> .	[13]
Nutrient gMCSs (ngMCSs)	Minimal subsets of gene knockout and/or nutrient deprivation interventions in metabolic networks.	MATLAB—The COBRA toolbox, function name: calculateGeneMCS, with the option 'only Nutrients' as TRUE.	[15]
gMCSs and regulatory networks	Minimal subsets of gene knockout interventions in integrated metabolic and regulatory (iMR) networks.	MATLAB functions available in <a href="https://github.com/PlanesLab/iMR_gmcs">https://github.com/PlanesLab/iMR_gmcs</a>	[17]
gMCSpy	Computation of gMCSs and ngMCSs in metabolic networks.	Open-source Python package built in COBRAPy available in <a href="https://github.com/PlanesLab/gMCSpy">https://github.com/PlanesLab/gMCSpy</a>	[19]

Note: Short description, necessary software and references for each methodology

the application of gene essentiality analysis in cancer based on the gMCS framework.

## RESULTS

In order to carry out the fine-tuning of the gMCS approach for gene essentiality analysis, we focus on true positives (TPs), false positives (FPs) and the positive predictive value [PPV = TP/(TP + FP)], also called precision. These metrics are important because cancer-specific essential genes are scarce, as they comprise less than 1% of all genes. Thus, accurate prediction of their presence is challenging, and labelling all genes as non-essential would result in an extremely high, yet misleading, accuracy rate, but at the cost of missing the essential genes. With these three metrics (FPs, TPs, PPV), however, we can correctly assess the performance of the different cases considered in this study.

### Optimal gene expression threshold

Once gMCSs are computed, the RNA-Seq expression data are projected to them. A required step to identify cancer-specific essential genes is to define whether a gene is highly or lowly expressed. We considered the two thresholding techniques implemented in *gmctool*: *gmcsTHX* and *localT2* [22].

In the case of *gmcsTHX*, for each sample, the most expressed gene from each gMCS is taken, as it should be the most expressed from the lot. Only unique genes are considered, as repetition will occur due to having more gMCSs than genes; then, using the expression of those unique genes, a distribution is built for each sample, and finally, the threshold is assigned after an arbitrary number. All genes below the coefficient defined by that sample-specific threshold will be deemed as lowly expressed, while those genes above the threshold value are appointed as highly expressed. After some analysis, the authors of *gmctool* decided on setting a threshold of 5% (*gmcsTH5*), which implies that all genes that have a lower expression than the bottom 5% of the most expressed genes of the gMCSs are considered as low expressed. To check whether this threshold conveys the best prediction, we analysed several thresholds: 0, 1, 2, 2.5, 5, 10 and 20%.

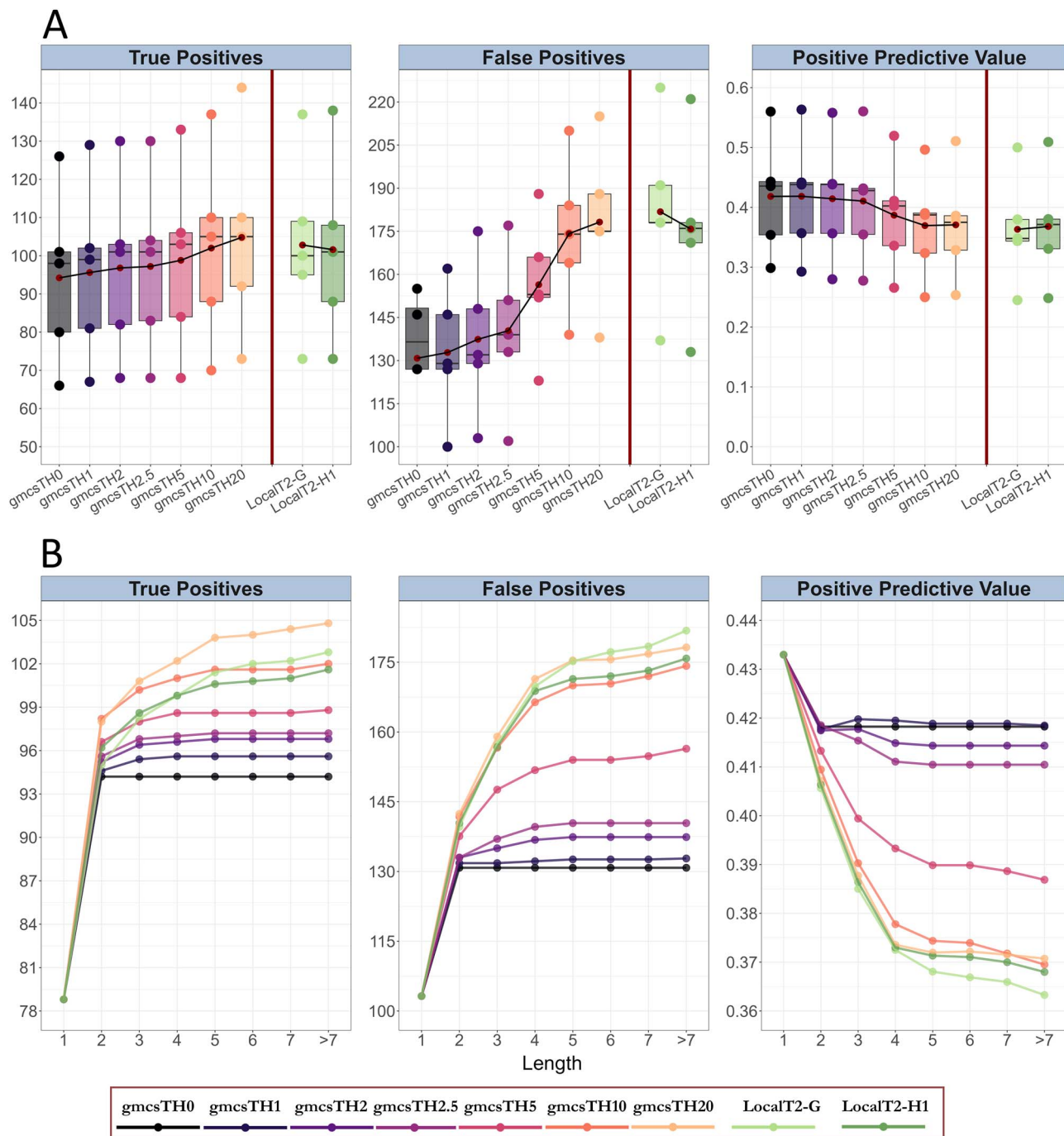
With respect to *localT2*, a cohort-dependent methodology that defines a threshold for each gene based on the observed expression distribution across the samples of the cohort, we considered two cases with different gene universes, one containing the genes

partaking in the gMCS (totalling 1,244), *localT2-G*, and the other has all Human1 genes (3,650), *LocalT2-H1*, and reviewed their predictive capabilities.

We started analysing Hart2015. First, we observed clear differences among the distinct thresholds, as it can be seen in the three facets of Figure 1A. In the case of *gmcsTHX* methodology, we can observe a scarcely escalating trend as the threshold increases in TPs (full details in Supplementary Table 1). Specifically, the mean of the predicted TPs varies from 94.2 for *gmcsTH0* to 104.8 in the case of *gmcsTH20*. In the case of the *localT2* thresholding methodology, considering only the genes partaking in gMCSs (*localT2-G*) yields more TPs than the procedure in which we consider all the universe of metabolic genes (*localT2-H1*). For the FP metrics, the boost as the threshold increases is more pronounced, jumping from 131 FP in *localT2-G* to 178 in *LocalT2-H1*. Concurrently, both *localT2* methods show similar performance to that of the highest thresholds in *gmcsTHX*, and they have the highest FP of the study, with 181.8 for the case of *localT2-G*. Finally, the PPV, subordinated to previous metrics, depicts a decreasing tendency as the threshold increases, having its maximum value of 0.418 for *gmcsTH0*; the lowest PPV (0.363) is held by *localT2* methodologies, due to their high FP values.

We observed similar trends and variances in DepMap to those observed in Hart2015 (Supplementary Figure 1A), finding a significant variation in TPs, FPs and PPVs for the different thresholds and cases tested (Kruskal–Wallis  $P$ -value  $< 2.2 \times 10^{-16}$ ). Precise mean values for the different scenarios are held in Supplementary Table 1. Overall, these results indicate that the higher the threshold, the less we can trust the positive results of the methodology, as the increase in FPs are greater than the increase in TP.

Prior to studying the impact of lengths in the analysis, we contemplated how would the metrics change in each threshold for each length. As the number of gMCSs sharply decreases after length seven in *gmctool*, lengths greater than seven were merged in one single group. Figure 1B shows the cumulative mean for all the thresholds through the eight considered lengths for Hart2015. The most relevant aspect is the importance of gMCSs of length one. Approximately 75% of TPs come from here, finding a small variance among the different thresholds, particularly for highest thresholds, which range from 74% to 79%. Switching to FPs, comparable behaviour is observed, for the lowest thresholds, the first orders are the parameters that provide most FPs. A similar pattern was found in DepMap (Supplementary Figure 1B). Note here that



**Figure 1.** Gene expression threshold analysis and essentiality predictions under the gMCS framework for Hart2015. TPs, FPs and PPVs are shown for the different cases analysed: gmcsthX for six thresholds (1, 2, 2.5, 3.5, 5, 10 and 20%) and localT2 for two gene universes (localT2-G and localT2-H1). Results for gmcsthX and localT2 are separated by a red vertical line. (A) Boxplot of TPs, FPs and PPVs in Hart2015 for the different cases analysed. Average values are linked together for the different cases in gmcsthX and localT2, respectively. (B) Accumulated mean value of TPs, FPs and PPVs across lengths in Hart2015 for the different cases considered.

gMCSs of length one define essential genes for any human cell type under specific growth medium conditions, here, the one used by default in Human1 (Ham's medium). For this reason, this subset of essential genes is conserved across different thresholds and no changes are observed for TPs and FPs that derived from gMCSs of length one in Figure 1B.

Intriguingly, in the case of the lowest thresholds, we can observe that higher lengths do not contribute at all to both TPs and FPs, which then result in a stagnation of the PPV (Figure 1B,

Supplementary Figure 1B). Finally, the higher the threshold, longer gMCSs are considered, as the metrics of TP and FP are seen steadily ramp up.

In summary, when we consider only the effect of changing the threshold, Figure 1 shows that lower thresholds have higher PPVs, although fewer TPs and less contribution based on longer gMCSs. In contrast, the higher thresholds have smaller PPV; however, they show more TPs, due to the gMCSs of higher orders that provide novel candidates for analysis.

## Length of gMCS

When considering candidate genes, shorter gMCS are more suitable for *in vitro* experimental validation, as it is easier to overexpress or inhibit one or two genes than altering, for example, six genes to prove valid a gMCS of length seven. *gmctool* does not use this data for anything else than display, because it assumes that the prediction is not affected by gMCS length. The objective of this section is to further study the impact that length has on TP and FP predictions, based on the previews of the previous section, but, instead of showing cumulative results, we examined each length independently.

As done in the previous section, lengths greater than seven were merged in one single group. For the sake of clarity, we only considered in our study two of the thresholds previously discussed with *gmcsTHX*: *gmcsTH2* and *gmcsTH5*. For completeness, we also included one case considered with *localT2*: *localT2-G*, which is slightly less computationally expensive than *localT2-H1* with similar performance in gene essentiality predictions. Note here both *localT2-H1* and *localT2-G* serve as a proxy for higher threshold values of the *gmcsTHX* technique.

As seen in [Figure 1B](#), a high proportion of TPs are obtained in the first order for all three thresholds. Since order 1 is not affected by the threshold, there were between 60 and 100 TPs depending on the cell line. However, starting from the lowest threshold, *gmcsTH2*, [Figure 2A](#), we found that only the four first orders yield any TP. The mean, depicted by the red dot in the figures, drastically drops from 78.8 for length one to 3 for length four ([Supplementary Table 2](#)). Only one cell line had any TPs predicted for further gMCS orders. The PPV remained consistent across lengths one to four, with high variance for the remaining lengths. The high mean PPV for lengths five and six was due to the unique predicted cell line with essential genes for these orders, which turned to be correct predictions; however, the limited sample size makes them unreliable.

For *gmcsTH5*, [Figure 2B](#), similarities with the previous thresholding method are perceived, but now length five has more TPs and FPs for three out of five cell lines, while lengths six and higher had some TPs and FPs for two cell lines. The PPV shows similar trends as before, but the fifth order had a higher median, despite high variance. Curiously, the sixth order has also a relatively high median, but an even larger variance.

Lastly, we considered *localT2-G*, [Figure 2C](#). As expected, all lengths are now not null. The number of TPs fluctuated between 20 and 30 along lengths two to five and decreased for higher orders. Nevertheless, the number of FPs increases substantially; hence, the PPV is lower than in the last shown cases, even if the mean is non-zero for all the lengths.

We also inquired DepMap data ([Supplementary Figure 2](#)), finding similar patterns to Hart2015. In the three cases considered, we identified significant variations of TPs, FPs and PPV for different length values (Kruskal–Wallis  $P$ -value  $< 2.2 \times 10^{-16}$ , [Supplementary Table 2](#)), which demonstrates that length is a critical parameter.

Summing up, both threshold and length analysis are closely correlated, as we can see with the previous analyses, as higher thresholds (or both *localT2* analyses) entail a greater repercussion of longer lengths. Nonetheless, shorter lengths have higher PPVs, but their lengths have a lower impact.

## gMCS promiscuity

When working with more than 97,000 gMCS, it is expected to have genes that partake in several gMCS, case which we termed

*Multiple-gMCS*; however, there is also the possibility of a gene intervening only in a particular gMCS, named *Single-gMCS*. We studied whether the user should equally rely on both conditions, or essential genes involved in more than one gMCS constitute a more robust strategy.

Taking all lengths into account and the same thresholds as in the previous exercise i.e. *gmcsTH2*, *gmcsTH5* and *localT2-G*, essential genes were separated in two groups for all cell lines, depending on the number of gMCSs they were part of, either having only one gMCS associated or more than one. As seen in previous analyses, gMCSs of length one are the most relevant genes essentiality wise and have the highest predictive values. To avoid introducing bias in the *Single-gMCS* study, these gMCSs have been removed from the analysis, because all these genes are associated with a unique gMCS.

[Figure 3A](#) depicts the result obtained with Hart2015 for TPs, FPs and PPVs. We detected more essential genes that are part of many gMCS, as seen by the sheer number of TPs in both comparisons, the median ranging from three to five in the case of *Single-gMCS* and from 10 to 20 for *Multiple-gMCS*. FPs raise in a similar fashion as seen in previous analyses, increasing as the threshold value increases. Finally, PPVs are spread along 0.2 for the *Single-gMCS* condition; values for *Multiple-gMCS* are slightly higher, but they have more variance. Mean values for the different metrics in [Figure 3A](#) are available at [Supplementary Table 3](#).

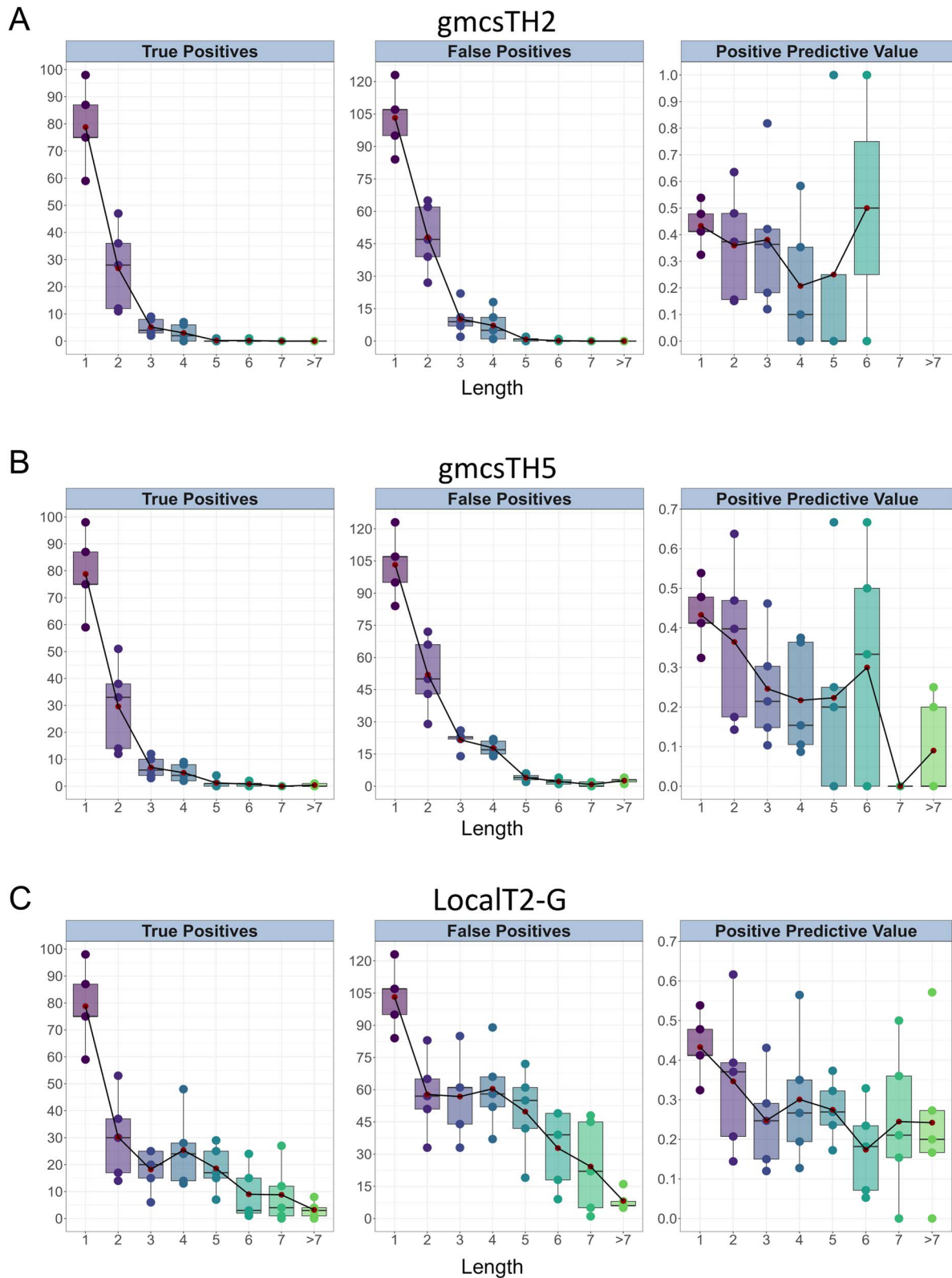
Again, the results for DepMap follow the same pattern than Hart2015 ([Supplementary Figure 3A](#)). We compared TPs, FPs and PPVs between *Single-gMCS* and *Multiple-gMCS* for the all the conditions tested, finding significantly higher values in *Multiple-gMCS* (see two-sided Wilcoxon  $P$ -values in [Supplementary Table 3](#)). However, although the effect size in TPs and FPs is very clear, it is more limited in the case of PPVs ([Supplementary Table 3](#)).

## Task promiscuity

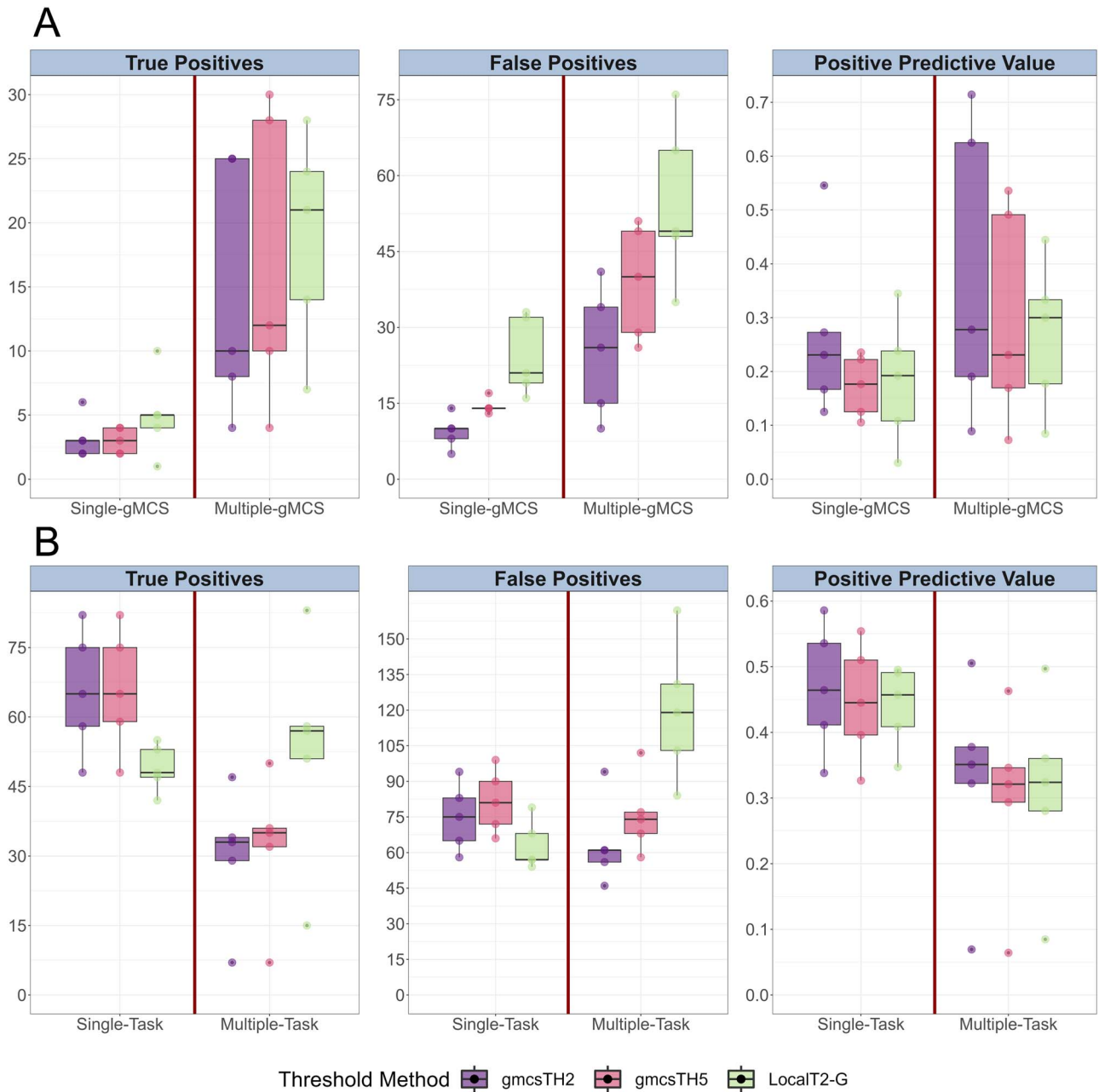
The metabolic model of Human1 defines essential tasks not only regarding proliferation, but also contemplating key cellular metabolic functions, such as ‘*de novo synthesis of nucleotides*’, ‘*beta oxidation of fatty acids*’, ‘*vitamins and co-factors*’, etc. The addition of those tasks in the initial biomass analysis enriches the prediction, although brings up an interesting question closely related to the last one: are genes associated with a unique task, called here *Single-task*, more reliable than genes associated with many tasks, called here *Multiple-task*? In this case, essential genes of the first order are considered, as of the 235 genes, 100 partake only in one task, whereas the rest are part of multiple tasks.

In Hart2015, we note that the TPs are higher in the *Single-task* cases in *gmcsTH2* and *gmcsTH5* ([Figure 3B](#)), where the TP medians are 65.6 and 65.9, respectively, to their *Multiple-Task* counterparts. For *localT2-G*, TPs keep high with more TPs detected in the *Multiple-task* case. In DepMap ([Supplementary Figure 3B](#)), TPs are slightly higher for the *Multiple-task* case in the different thresholds. Once again, the FPs show similar trends; however, the number of FPs is significantly higher in the *Multiple-task* case. Full details can be found in [Supplementary Table 4](#).

In Hart2015, *Single-task* PPVs are higher for *gmcsTH2* and *gmcsTH5*, in which the single task shows a consistently high PPV. Considering DepMap, *Multiple-task* PPVs are significantly lower for all the cases considered (two-sided Wilcoxon  $P$ -value  $< 2.2 \times 10^{-6}$ , [Supplementary Table 4](#)), which suggest that essential genes coming from the *Single-Task* case are more reliable.



**Figure 2.** Length of gMCSs and gene essentiality predictions in Hart2015. (A) Boxplot of TPs, FPs and PPVs in Hart2015 for gmcsTH2 for different lengths considered. (B) Boxplot of TPs, FPs and PPVs in Hart2015 for gmcsTH5 for different lengths considered. (C) Boxplot of TPs, FPs and PPVs in Hart2015 for localT2-G for different lengths considered. Average values are linked together for the different cases analysed.

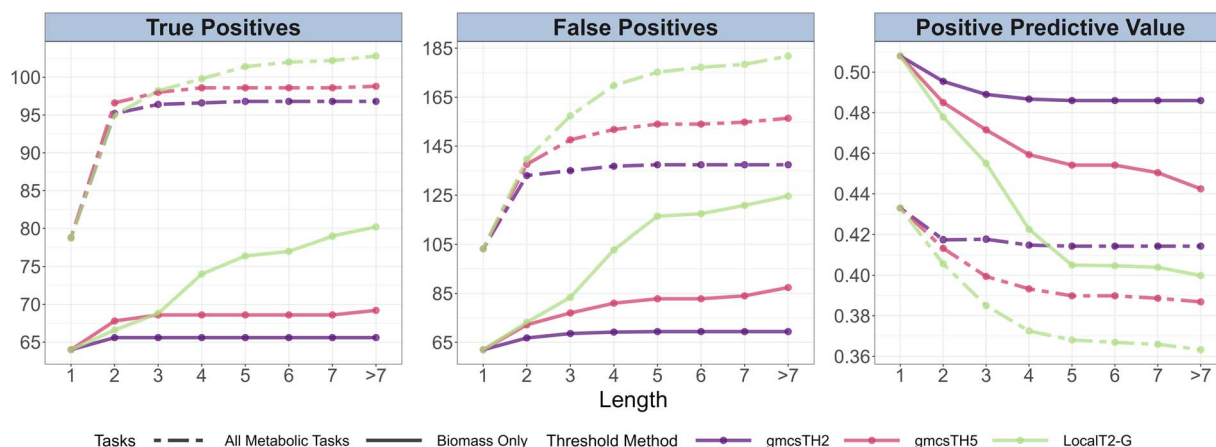


**Figure 3.** gMCS and metabolic task promiscuity and gene essentiality predictions in Hart2015. (A) Boxplot of TPs, FPs and PPVs in Hart2015 for gmcsth2, gmcsth5 and localT2-G considering gMCS promiscuity, which defines two cases: essential genes involved in one gMCS ('Single-gMCS') and more than one ('Multiple-gMCS'). (B) Boxplot of TPs, FPs and PPVs in Hart2015 for gmcsth2, gmcsth5 and localT2-G considering task promiscuity, which defines two cases: essential genes involved in one essential metabolic task ('Single-task') and more than one ('Multiple-task').

### Compare all-tasks to biomass results

Valcárcel *et al.* [13] used all the tasks deemed as essential by Human1 to compute gMCSs. However, the initial approach that was available when Apaolaza *et al.* [11] published their work was to target the biomass function, a reaction which measures the cell's ability to grow. The number of gMCSs obtained for each analysis changes drastically, as the latter is contained in the former, there is a vast difference of 40,000gMCSs between the two analyses. We pose here the question as to whether essential tasks really helps in gene essentiality predictions, or they are not more than a liability for the analysis and only the biomass function should be deliberated.

Figure 4 shows the cumulative median of TPs, FPs and PPVs for different lengths of gMCSs in Hart2015 using the same threshold techniques as in previous analyses. The dashed line represents the All Tasks analysis. It can be observed for all thresholds that TPs are increased when considering all tasks (Figure 4). In Hart2015, the mean of TPs moves from 65.6, 69.2 and 80.2 in the biomass analysis to 96.8, 98.8 and 102.8 when All Tasks were considered in gmcsth2, gmcsth5 and localT2, respectively. Per contra, FPs increase substantially too, from a mean of 69.4, 87.4 and 125 in gMCSs associated with Biomass analysis to 137.4, 156.4 and 182.2 in the case of All Tasks (Figure 4).



**Figure 4.** Effect of metabolic tasks in gene essentiality predictions in Hart2015. TPs, FPs and PPV are shown when considering only the biomass tasks ('Biomass', solid line) or all metabolic tasks ('All Tasks', Two dashed line) in the gene essentiality predictions for the different for three thresholding techniques (gmcsTH2, gmcsTH5 and localT2-G) and different lengths for Hart2015.

The increase in FPs in Hart2015 can be clearly observed in the PPV metrics, which are reduced in the *All Tasks* condition (0.414, 0.387, 0.363) with respect to the *biomass* analysis (0.486, 0.442 and 0.400) (Supplementary Table 5). The same result can be found in DepMap (Supplementary Figure 4), where we found a statistically significant difference for all the metrics and lengths between both conditions (see Kruskal–Wallis P-values in Supplementary Table 5).

## DISCUSSION

The aim of this work has been to perform an *in silico* meta-analysis of the gMCS algorithm developed by Apaolaza *et al.* [11] and honed by Valcárcel *et al.* [13]. Getting to know the parameters that suit best the analysis is of utmost importance in order to find candidate genes that could exploit metabolic vulnerabilities. This thorough analysis has focused on certain aspects of the methodology, in the following order, impact of gene expression thresholding methodology and thresholding value on prediction, impact of gMCS length on prediction, association analysis: whether a gene predicted present in one gMCS is as reliable as a gene predicted present in many gMCS, as well as, a gene predicted present in one essential metabolic task or many; and last, the comparison of whether the biomass function is sufficient for the analysis or adding the rest of deemed essential tasks by the authors of Human1 provides any boost to the analysis.

We now would like to discuss which are the parameters we consider to be the most relevant for the analysis. For it, we will one by one delve into each one in the same structure as has been discussed in the Results section. Before beginning, we would like to note that even though PPVs of less than 0.5 are obtained, which could let the reader think that this prediction capability is irrelevant, we consider this an achievement, because gene essentiality prediction has always been a challenge in the field of Computational Biology, and higher precision values are not easily obtained using the current state-of-the-art metabolic models and network-based strategies.

Insomuch as threshold is concerned, it is clear in both Hart2015 and DepMap that higher threshold values imply more TPs; however, we are preoccupied by the sheer number of FPs predicted in these cases, as they will make the selection of candidates less reliable. Also, high gmcsTHX values and localT2 thresholding techniques behave similarly. Following the same train of thought,

we decant for lower threshold values. Thus, we value more the high results of the PPV rather than the slightly higher TPs; hence, we advise a threshold value of around 2% to maintain a high PPV as well as obtaining additional TPs.

Taking length in mind, we observe length one carries the most importance of all lengths, as half of them really are essential as we predict. We cannot discuss length without considering threshold, as they are closely related because higher lengths are only relevant when examining higher thresholds; nonetheless, longer thresholds have higher FPs due to longer lengths not being as precise as lower lengths. Which reinforces the last point, longer thresholding implies worse predictions due to longer length gMCSs not being as accurate as shorter ones. Keeping in mind that we ascertain lower thresholds to improve PPV, as we consider gmcsTH2 to be the optimal threshold, therefore, only lengths up to five should be considered, as the rest is unreliable and probably null.

When considering genes partaking in more than one gMCS, which we called *gMCS promiscuity*, we have not seen significant visual differences in any of the cases, we therefore think that a gene involved in multiple gMCSs does not imply that the gene is more relevant. In the case of *Task promiscuity*, clear differences appear, and, admitting that the P-values for Hart2015 are not significant, both visually and in DepMap, we can observe significant changes. The PPV is 0.15 higher when a gene partakes in a single task, so, we expect that genes partaking in multiple tasks are going to be less important than those who only partake in one task. This result could imply that the inhibited gene, if partaking in many different tasks, has multiple readjustments that are not considered in the metabolic model, while disrupting a job-specific gene should have a higher impact in cellular metabolism.

Finally, we checked whether all tasks improve biomass predictions. We noticed that adding more tasks to the analysis increased in 47.8% all the detected TPs in lower thresholds, decreasing to 28.1% more for higher thresholds, yet decidedly increasing the FPs 97.9% more for lower thresholds and 45.7% more for longer thresholds. The change resulted in a lower PPV of 14.8% less for gmcsTH2 and 9.25% for localT2. According to these results, we encourage the community to revise and improve further the quality of tasks, as it adds a critical number of TPs that could broaden the number of candidate genes to study, but currently a significantly higher number of FPs.



All in all, the optimal parameter combination for maximizing both TPs and PPVs would involve using gmcsTH2. Therefore, only lengths up to five should be considered, and the promiscuity of gMCSs should not impact the final results. At this stage, the focus is solely on studying the biomass reaction, making an examination of gene promiscuity among multiple tasks unnecessary.

## METHODS

### Gene expression data

Validation analysis has been developed using the data available in Hart *et al.* [20], which includes five cell lines: HCT116, HeLa, GBM, RPE1 and DLD1, and DepMap, release 21Q4 [21]. RNA-seq data for Hart2015 are available in Gene Expression Omnibus database [23] under the accession number GSE75189. In the case of DepMap, the intersection of cell lines which had both data from CRISPR knockout (Achilles-Chronos) and RNA-seq expression consists of 913 cell lines. Processed gene expression data are available at the DepMap (<https://depmap.org/portal/download/all/>) and raw data at the Sequence Read Archive [24] under accession number PRJNA523380 [25].

### gMCS computation

The gMCSs were downloaded from *gmctool* in GitHub: <https://github.com/PlanesLab/gmctool>. Human1, version 1.4.0, has been used as the reference metabolic network, which contains 13,101 reactions, 8,400 metabolites and 3,628 genes [9]. As aforementioned, this network defines 57 essential metabolic tasks for any human cell, which define the output metabolites that must be obtained from a list of input metabolites, subject to reaction constraints. Human1 was downloaded from <https://github.com/SysBioChalmers/Human-GEM>.

### Essentiality analysis

The five cell lines from Hart2015 have an associated Bayes Factor threshold to identify essential genes. Each of them has a particular curated value based on a false discovery threshold (<5%). Genes with a score higher than the defined thresholds will be considered as essential. The intersection of the predicted essential genes with the gMCS approach and the essential genes from Hart2015 leads to TPs and FPs. Bayes Factor data for Hart2015 were obtained from <https://www.cell.com/cms/10.1016/j.cell.2015.11.015/attachment/11268869-c530-4f14-9e70-6a19db4bf8d0/mmc3.xlsx>.

For DepMap analysis, essentiality score below  $-0.6$  has been considered as essential. This threshold is the same that the one used by Robinson *et al.* [9]. The complete list of essential genes for all the cases considered in Hart2015 and DepMap can be found in Supplementary Tables 6 and 7 respectively.

DepMap essentiality score data were downloaded from <https://depmap.org/portal/download/all/> and, specifically, the DepMap21Q4 database, [https://depmap.org/portal/download/all/?releasename=DepMap+Public+21Q4&filename=CRISPR\\_gene\\_effect.csv](https://depmap.org/portal/download/all/?releasename=DepMap+Public+21Q4&filename=CRISPR_gene_effect.csv).

#### Key Points

- The gMCS approach constitutes a promising network-based strategy to predict SL and metabolic vulnerabilities in cancer;
- A review of the gMCS approach, recent extensions and future challenges are described and discussed;

- A systematic analysis and fine-tuning of the gMCS approach for gene essentiality analysis in cancer is carried out with two different datasets of large-scale silencing experiments;
- We show the important and related role of gene expression thresholding techniques and length of gMCSs for the accurate prediction of essential genes;
- We emphasize the importance of correctly defining the essential metabolic tasks to be blocked, beyond biomass production, which is significantly more reliable than the rest of essential tasks deemed in Human1 metabolic reconstruction.

## SUPPLEMENTARY DATA

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

## AUTHORS' CONTRIBUTIONS

Conceptualization: Danel Olaverri-Mendizabal, Luis V. Valcárcel, Francisco J. Planes.

Formal Analysis: Danel Olaverri-Mendizabal.

Funding Acquisition: Francisco J. Planes.

Investigation: Danel Olaverri-Mendizabal, Luis V. Valcárcel, Francisco J. Planes.

Methodology: Danel Olaverri-Mendizabal, Luis V. Valcárcel, Francisco J. Planes.

Project Administration: Danel Olaverri-Mendizabal, Luis V. Valcárcel, Francisco J. Planes.

Resources: Danel Olaverri-Mendizabal, Luis V. Valcárcel, Francisco J. Planes.

Software: Danel Olaverri-Mendizabal, Luis V. Valcárcel.

Supervision: Luis V. Valcárcel, Francisco J. Planes.

Validation: Danel Olaverri-Mendizabal, Luis V. Valcárcel, Francisco J. Planes.

Visualization: Danel Olaverri-Mendizabal.

Writing—Original Draft: Danel Olaverri-Mendizabal, Luis V. Valcárcel.

Writing—Review & Editing: Danel Olaverri-Mendizabal, Luis V. Valcárcel, Naroa Barrena, Carlos J. Rodríguez, Francisco J. Planes.

## FUNDING

This work was supported by the Minister of Economy and Competitiveness of Spain [PID2019-110344RB-I00 and PID2022-143298OB-I00, F.J.P.], PIBA Programme of the Basque Government [PIBA\_2020\_01\_0055, F.J.P.], ERANET program ERAPerMed [MEET-AML, F.J.P.], Elkartek programme of the Basque Government [KK-2022/00045, F.J.P.], Ramon Areces grant [to F.J.P.]. N.B. received his salary from a Basque Government predoctoral grant [PRE\_2021\_2\_0025]. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## CODE AVAILABILITY

Data processing and all the statistical analysis presented in this work have been conducted using R software, version 4.2.1. The code for the tool presented in this article is available in [https://github.com/PlanesLab/gmcs\\_meta\\_analysis](https://github.com/PlanesLab/gmcs_meta_analysis). Within the

repository, all necessary data and code can be found. Due to their size, CCLE, DepMap and Hart Data exceed GitHub's capacity, but there are instructions available for downloading these crucial datasets.

## DATA AVAILABILITY

The authors declare that all data supporting the findings of this study are available within the article, its Supplementary Materials and GitHub repository, or from the corresponding authors upon request.

## REFERENCES

- O'Neil NJ, Bailey ML, Hieter P. Synthetic lethality and cancer. *Nat Rev Genet* 2017;**18**:613–23.
- Thompson NA, Ranzani M, van der Weyden L, et al. Combinatorial CRISPR screen identifies fitness effects of gene paralogues. *Nat Commun* 2021;**12**:1302.
- de Kegel B, Ryan CJ. Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. *PLoS Genet* 2019;**15**:e1008466.
- Ferrer-Bonsoms JA, Jareno L, Rubio A. Rediscover: an R package to identify mutually exclusive mutations. *Bioinformatics* 2022;**38**:844–5.
- Wang J, Zhang Q, Han J, et al. Computational methods, databases and tools for synthetic lethality prediction. *Brief Bioinform* 2022;**23**:1–22.
- Folger O, Jerby L, Frezza C, et al. Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol* 2011;**7**:1–10.
- Agren R, Mardinoglu A, Asplund A, et al. Identification of anti-cancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol Syst Biol* 2014;**10**:1–13.
- Pacheco MP, Bintener T, Ternes D, et al. Identifying and targeting cancer-specific metabolism with network-based drug target prediction. *EBioMedicine* 2019;**43**:98–106.
- Robinson JL, Kocabaş P, Wang H, et al. An atlas of human metabolism. *Sci Signal* 2020;**13**:1–22.
- Bintener T, Pacheco MP, Philippidou D, et al. Metabolic modelling-based in silico drug target prediction identifies six novel repurposable drugs for melanoma. *Cell Death Dis* 2023;**14**:468.
- Apaolaza I, San José-Eneriz E, Tobalina L, et al. An in-silico approach to predict and exploit synthetic lethality in cancer metabolism. *Nat Commun* 2017;**8**:1–9.
- Apaolaza I, Valcarcel LV, Planes FJ. GMCS: fast computation of genetic minimal cut sets in large networks. *Bioinformatics* 2019;**35**:535–7.
- Valcárcel LV, José-Enériz ES, Ordoñez R, et al. gMCStool: automated network-based tool to search for metabolic.  *biorXiv*. 2022; 1–30.
- Viñado AC, Calvo IA, Cenzano I, et al. The bone marrow niche regulates redox and energy balance in MLL::AF9 leukemia stem cells. *Leukemia* 2022;**36**:1969–79.
- Apaolaza I, San José-Enériz E, Valcarcel LV, et al. A network-based approach to integrate nutrient microenvironment in the prediction of synthetic lethality in cancer metabolism. *PLoS Comput Biol* 2022;**18**:1–20.
- Wang R, Dillon CP, Shi LZ, et al. The transcription factor Myc controls metabolic reprogramming upon T lymphocyte activation. *Immunity* 2011;**35**:871–82.
- Barrena N, Valcárcel LV, Olaverri-Mendizabal D, et al. Synthetic lethality in large-scale integrated metabolic and regulatory network models of human cells. 2023;**9**:32.
- Heirendt L, Arreckx S, Pfau T, et al. Creation and analysis of biochemical constraint-based models using the COBRA toolbox v3.0. *Nat Protoc* 2019;**14**:639–702.
- Rodríguez CJ, Barrena N, Olaverri-mendizabal D, et al. gMCSpy: efficient and accurate computation of genetic minimal cut sets in python supplementary information.  *bioRxiv*. 2024.
- Hart T, Chandrashekar M, Aregger M, et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 2015;**163**:1515–26.
- DepMap, Broad. DepMap 21Q4 Public. figshare. Dataset. 2021. <https://doi.org/10.6084/m9.figshare.16924132.v1>.
- Richelle A, Joshi C, Lewis NE. Assessing key decisions for transcriptomic data integration in biochemical networks. *PLoS Comput Biol* 2019;**15**:e1007185.
- Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**:207–10.
- Katz K, Shutov O, Lapoint R, et al. The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res* 2022;**50**:D387–90.
- Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the cancer cell line Encyclopedia. *Nature* 2019;**569**:503–8.