Using Empirical Biological Knowledge to Infer Regulatory Networks From Multi-omics Data

Anna Pačínková (✉ anna.pacinkova@recetox.muni.cz)

Research Centre for Toxic Compounds in the Environment

Vlad Popovici

Research Centre for Toxic Compounds in the Environment

Research Article

Keywords: integrative analysis, multimodal omics, Bayesian networks, regulatory networks, knowledge discovery

Posted Date: February 1st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1291540/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Using empirical biological knowledge to infer regulatory networks from multi-omics data

Anna Pačínková^{1,2*} and Vlad Popovici¹

*Correspondence:

anna.pacinkova@recetox.muni.cz

¹RECETOX, Faculty of Science,
Masaryk University, Kotlarska 2,
Brno, Czech Republic

Full list of author information is
available at the end of the article

Abstract

Background: Integration of multi-omics data can provide a more complex view of the biological system consisting of different interconnected molecular components, the crucial aspect for developing novel personalised therapeutic strategies for complex diseases. Various tools have been developed to integrate multi-omics data. However, an efficient multi-omics framework for regulatory network inference at the genome level that incorporates prior knowledge is still to emerge.

Results: We present IntOMICS, an efficient integrative framework based on Bayesian networks. IntOMICS systematically analyses gene expression, DNA methylation, copy number variation and biological prior knowledge to infer regulatory networks. IntOMICS complements the missing biological prior knowledge by so-called *empirical* biological knowledge, estimated from the available experimental data. Regulatory networks derived from IntOMICS provide deeper insights into the complex flow of genetic information on top of the increasing accuracy trend compared to a published algorithm designed exclusively for gene expression data. The ability to capture relevant crosstalks between multi-omics modalities is verified using known associations in microsatellite stable/instable colon cancer samples. IntOMICS is also applied to detect potential predictive biomarkers in microsatellite stable stage III colon cancer samples.

Conclusions: We provide IntOMICS, a framework for multi-omics data integration using a novel approach to biological knowledge discovery. IntOMICS is a powerful resource for exploratory systems biology and can provide valuable insights into the complex mechanisms of biological processes that has a vital role in personalised medicine.

Availability: Source code and data used for producing the presented results are available at: <https://gitlab.ics.muni.cz/bias/intomics>.

Keywords: integrative analysis; multimodal omics; Bayesian networks; regulatory networks; knowledge discovery

Introduction

The rapid development of high-throughput technologies has led to large production and availability of *omics* data. Single-omics technologies measure simultaneously molecules of the same type from biological samples. On the contrary, multi-omics data collect multiple modalities from the same set of samples and describe different aspects of cellular functioning. Therefore, multi-omics data contain complementary information and provide a holistic view of the biological system consisting of different interconnected molecular components.

Integration of multi-*omics* data can enhance our understanding of biological systems, crucial for developing novel personalised therapeutic strategies for complex diseases. Hence, developing a computational framework to infer regulatory relationships by integrating multiple modalities is one of the most relevant and challenging problems in systems biology.

The regulatory network inference from high-throughput data is limited by noise or measurement errors. However, it could be significantly improved by incorporating a wealth of biological prior knowledge from the scientific literature [1, 2]. Gene regulatory networks inferred from single-*omics* data (gene expression) and prior knowledge were frequently used to model gene-gene interactions [3, 4, 1, 5]. However, the flow of genetic information in biological systems is very complex, and gene expression is a product of multiple biological processes and control mechanisms, such as copy number variations (CNVs), transcription factors, non-coding RNAs, DNA methylations, or histone modifications. Therefore, it is crucial to progress from single-*omics* data analysis and derive causal relationships between features from multi-*omics* data.

There are several tools estimating the dependence structure among multi-*omics* data [6, 7], estimating a large number of networks for each gene where different modalities are treated as nodes in a graphical model. Although this approach provides valuable insights, they are limited and should be complemented with gene-gene interactions. Several other tools for gene regulatory network inference from multi-*omics* data based on correlation or regression were proposed [8, 9, 10]. However, these tools do not integrate biological prior knowledge from the databases, and their main limitation is missing implementation.

[11] propose a framework to identify disease-specific pathways by integrating gene expression, mutation information and prior knowledge through a Bayesian network. One of its drawbacks is data discretisation, which implies substantial information loss. Recently, [12] published one of the most promising tools, integrating the prior knowledge, transcriptomics and metabolomics data into an interaction network. However, several diseases such as cancer are driven by changes at the genome level, such as somatic mutations, copy number variations or methylation. Another limitation of these two methods is the data availability for differential analysis between experimental groups (usually normal versus diseased phenotype).

Motivated mainly by [3, 4], we present IntOMICS, a novel Bayesian framework for multi-*omics* data integration using prior knowledge to infer regulatory networks. Even if the intensive research has deposited a wealth of biological prior knowledge into public databases, some regulatory events between genes are still missing. Although databases such as DbVar [13] or iMETHYL [14] exist, a database with known CNV/METH and gene expression interactions is missing. Therefore, IntOMICS incorporates a novel approach to biological knowledge discovery - estimates the *empirical* biological knowledge to complement the available data from public databases. IntOMICS is designed to capture relevant crosstalks within and between copy number variation, DNA methylation and gene expression. The model parameters tuning guarantees accurate model design and robust results inference.

[4] algorithm (W&H) represents one of the most relevant gene regulatory network reconstruction tools based on Bayesian networks. Therefore, the W&H algorithm is

selected for performance comparison with IntOMICS at the gene expression level. Both algorithms resemble the core formulation of prior distribution and integration of biological prior knowledge. On the contrary, they differ in two key aspects: *(i)* IntOMICS combines prior knowledge with data-derived evidence - the *empirical* biological knowledge *(ii)* IntOMICS is designed to infer not only dependencies among gene expression but also between gene expression, DNA methylation and copy number variation. W&H algorithm relies on conventional MCMC sampling, which tends to be slow in convergence and mixing and can often be stuck in low-probability regions. The inclusion of adaptive MCMC simulation and Markov blanket resampling (MBR) [15] minimise the weaknesses of the W&H algorithm.

IntOMICS can theoretically be extended with any additional modality if the proposed model assumptions are satisfied. We assume that variables come from the multivariate Gaussian distribution, so no discretisation is needed. In addition, only biologically relevant dependencies respecting the central dogma of molecular biology must be considered.

Bayesian Networks

A Bayesian network (BN) is a graphical model representing probabilistic relationships among random variables. BN is defined by the joint probability distribution over the variables specified by *(i)* a network structure G represented by a directed acyclic graph (DAG) with a set of nodes (indicating a set of random variables) and a set of directed edges (indicating conditional dependence relations among random variables), and *(ii)* a family of conditional probability distributions with corresponding parameters characterising the dependencies represented by the set of edges.

Due to the Markov condition (each variable is conditionally independent of the set of all its non-descendants given the set of all its parents), the joint probability distribution factorises as follows:

$$P(X_1, \dots, X_N | G) = \prod_{j=1}^N P(X_j | X_{pa_j(G)}), \quad (1)$$

where $X_1, \dots, X_N$ are random variables, $X_{pa_j(G)}$ are parents of X_j implied by the network structure G .

Learning the Bayesian network structure from the data is one of the most challenging tasks. It is guided by a scoring metric S , which assess the agreement between a given network structure and the available data D . The aim is to identify the highest-scoring network structure.

S is proportional to the posterior probability of the network structure given the data D and factorises into a product where each term depends only on a given node and its parents:

$$P(G|D) \propto P(D|G)P(G) = \prod_{j=1}^N S(X_j, X_{pa_j}|D), \quad (2)$$

The choice of scoring metric either requires the data to be discretised (BDe score, [16]) or can only capture linear regulatory relationships (BGe score, [17]). The

BGe score is developed for continuous data sampled from a multivariate normal distribution. Since gene expression, copy number variation, and DNA methylation data are continuous, and we want to avoid data discretisation because of information loss, we consider BGe scoring metric. Furthermore, it has been shown that BGe remains as powerful as BDe, even in the case of slight departures from the linearity assumption [18].

Methods

We present IntOMICS, a powerful Bayesian framework for multi-*omics* data integration to capture dependencies among different molecular features. Fig. 1. summarises the key steps of our novel framework.

IntOMICS integrates the biological knowledge from databases and is designed to capture relevant crosstalks within and between copy number variation, DNA methylation and gene expression. The missing biological prior knowledge is complemented by so-called *empirical* biological knowledge, estimated from the available experimental data. The *empirical* biological knowledge reflects hidden patterns in the available datasets derived from past iterations of the Markov chain.

The proposed framework avoids data discretisation, which implies substantial information loss. IntOMICS parameters tuning guarantees accurate model design and robust results inference. The inclusion of adaptive MCMC simulation and Markov blanket resampling (MBR) [15] improves convergence and mixing of the Markov chain.

Biological Prior Knowledge Integration

Biological prior knowledge with information on known interactions among molecular features is integrated into the regulatory network inference employing a prior probability of the network structure G . As in [3] and [4], we define the prior distribution over network structures as:

$$P(G|\beta) = \frac{e^{-\beta E(G)}}{Z(\beta)} = \frac{e^{-\beta E(G)}}{\sum_{G \in \mathcal{G}} e^{-\beta E(G)}}, \quad (3)$$

where $\mathcal{G}$ indicates a set of all possible network structures, and the parameter β controls the strength of the influence of the biological prior knowledge. The energy function $E(G)$ measures the agreement between the biological prior knowledge and the current network structure:

$$E(G) = \sum_{j=1}^N \varepsilon(X_j, X_{pa_j(G)}), \quad (4)$$

$$\varepsilon(X_j, X_{pa_j(G)}) = \sum_{i \in X_{pa_j}} (1 - B_{ij}) + \sum_{i \notin X_{pa_j}} B_{ij}, \quad (5)$$

where B is adjacency matrix with $B_{ij} \in [0, 1]$. The matrix B represents the biological prior knowledge with information on known interactions among molecular features. If there is the prior knowledge about the direct interaction from node i to

node j , $B_{ij} = 1$. If there is no knowledge about the direct interaction from node i to node j , $B_{ij} = 0.5$. If we have prior knowledge that there is no edge from node i to node j , $B_{ij} = 0$.

Then, from (4) we obtain the upper bound of the partition function $Z(\beta)$ (see [4]):

$$Z(\beta) = \prod_j \sum_{X_{pa_j}} e^{-\beta \varepsilon(X_j, X_{pa_j(G)})}. \quad (6)$$

In the current experiment, curated regulatory relationships of publicly available network database KEGG [20] are used as prior knowledge. Nevertheless, our approach is not limited to KEGG, any other available network database can be used.

MCMC Sampling Scheme for Bayesian Network Structure Learning

A complete comparison of posterior probabilities is intractable since the search space of all possible network structures grows super-exponentially with the number of random variables. Hence, a Markov Chain Monte Carlo (MCMC) sampling scheme [21] is adopted to generate a sample of network structures from the posterior distribution.

First, we sample a network structure G_c while keeping the β parameter fixed. We need to define the proposal distribution $Q(G_c|G_s)$ to draw candidate network structures from an intractable posterior distribution. The candidate network structure G_c is proposed by either adding, deleting or reversing a particular directed edge from the current network structure G_s . Besides a conventional single edge proposal move, the Markov blanket resampling step [15] is introduced with a fixed probability $p_{MBR} = 1/15$ to achieve faster mixing and convergence (p_{MBR} suggested by the authors). The acceptance probability A indicates how probable the candidate network structure is with respect to the current network structure, according to the posterior distribution. A candidate network structure proposed from the proposal distribution $Q(G_c|G_s)$ is accepted according to the Metropolis-Hastings rule [22] with the acceptance probability given by:

$$A = \min \left\{ \frac{P(D|G_c)P(G_c|\beta)Q(G_s|G_c)}{P(D|G_s)P(G_s|\beta)Q(G_c|G_s)}, 1 \right\}. \quad (7)$$

After each iteration except the sampling phase (see section technical details), a new parameter β_c for fixed network structure G is proposed and accepted according to the following acceptance probability:

$$A_\beta = \min \left\{ \frac{P(G|\beta_c)}{P(G|\beta_s)}, 1 \right\}, \quad (8)$$

where β_s refers to the current parameter value.

Technical Details

The main steps of IntOMICS are summarised in Algorithm 1.

IntOMICS framework takes as input (i) gene expression matrix GE ($m \times n_1$), (ii) the associated copy number variation matrix CNV ($m \times n_2$), (iii) the associated DNA methylation matrix of beta-values $METH$ ($m \times n_3$) sampled from the same individuals, and (iv) the biological prior knowledge matrix B ($n_1 \times n_1$)

with information on known interactions among molecular features. The ordered quantile normalisation [23] is used to transform DNA methylation beta-values to Gaussian distribution. Because multiple DNA methylation loci are associated with a single gene, we can use linear regression to filter irrelevant DNA methylation loci. In this study, we considered DNA methylation probes with statistically significant coefficient (p-value < 0.5) and $R^2 > 0.3$.

Adaptive MCMC algorithms use information from past iterations of the chain to improve computational efficiency. We use an automatically tuned MCMC algorithm [19] with default hyper-parameters to estimate parameter β and *empirical* biological knowledge matrix $\mathcal{B}$ through multiple phases. The automatically tuned MCMC algorithm consists of several distinct phases.

The first adaptive phase is used to roughly tune the parameter β , more precisely the variance of its proposal distribution σ_s^2 . The proposal distribution is $\beta_c \sim N(\beta_s, \sigma_s^2)$, where β_s refers to the current parameter value, and β_c refers to the candidate parameter value. The initial value of the parameter β_s is randomly drawn from $U[0, 10]$, and then we require $\beta \geq 0.5$. The initial value $\sigma_s = 5$.

The transient phase is applied to diagnose whether the chain has reached the mode of the target distribution.

The second adaptive phase is used to fine-tune the variance σ_s^2 , β parameter estimation and compute the *empirical* biological prior matrix $\mathcal{B}$. Assuming $B_{ij} = 0.5$, the prior knowledge about interaction from node i to node j is updated during the second adaptive phase after every conventional single edge proposal move. The B_{ij} value corresponds to the ratio of acceptance (number of iterations with accepted candidate edge from node i to node j) and frequency (number of iterations with proposed candidate edge from node i to node j) (for details, see Table 1). Reversing an edge is equivalent to deleting the edge and adding the edge in the opposite direction.

Table 1 Empirical biological knowledge estimation. Assuming there is no prior knowledge about the direct interaction from node i to node j , the *empirical* biological matrix $\mathcal{B}$ is estimated, with $B_{ij} = \frac{a_{ij}}{f_{ij}} \in [0, 1]$.

edge	operation	frequency	candidate	acceptance
$G_i \rightarrow G_j$	add	$f_{ij} = f_{ij} + 1$	accepted	$a_{ij} = a_{ij} + 1$
			rejected	$a_{ij} = a_{ij}$
$G_i \rightarrow G_j$	delete	$f_{ij} = f_{ij} + 1$	accepted	$a_{ij} = a_{ij}$
			rejected	$a_{ij} = a_{ij} + 1$
$G_i \rightarrow G_j$	reverse	$f_{ij} = f_{ij} + 1$	accepted	$a_{ij} = a_{ij}$
		$f_{ji} = f_{ji} + 1$	rejected	$a_{ji} = a_{ji} + 1$
$G_i \rightarrow G_j$	reverse	$f_{ij} = f_{ij} + 1$	accepted	$a_{ji} = a_{ji} + 1$
		$f_{ji} = f_{ji} + 1$	rejected	$a_{ij} = a_{ij} + 1$
$G_i \rightarrow G_j$	reverse	$f_{ij} = f_{ij} + 1$	accepted	$a_{ji} = a_{ji} + 1$
		$f_{ji} = f_{ji} + 1$	rejected	$a_{ij} = a_{ij} + 1$

The *empirical* biological matrix $\mathcal{B}$ and the parameter β determined by the second adaptive phase are used in the last sampling phase. In our simulation, two independent samples of network structures are produced. Each sample consists of 200,000 DAGs (with a burn-in period of 100,000 iterations). The resulting samples of DAGs are thinned - discarded all but every 500th DAG. The burn-in period of 100,000 iterations and thinning frequency of 500 are arbitrary choices. We tested different settings of these parameters but they did not influence the resulting accuracy (see Supplementary Figures 1-4).

Distinct DAGs can describe the same set of independence relations and have the same likelihood score. Such DAGs are from the same equivalence class. The

equivalence class can be represented by completed partially directed acyclic graphs (CPDAGs). Therefore, we convert DAGs into corresponding CPDAGs and discard duplicated CPDAGs.

The convergence of resulting Markov chains is examined using the c_{rms} measure [24]. The c_{rms} threshold is given by the third quartile of $|c_{rms_k} - c_{rms_{k-1}}|$ for each iteration k . If the c_{rms} value of the last 100 iterations (after thinning) is smaller than the c_{rms} threshold, the MCMC simulation stops. Otherwise, the simulation proceeds until this condition is met. Subsequently, we discard the CPDAGs from the burn-in period.

The strength of the probabilistic relationships expressed by the edges in the resulting network structure is measured as their empirical frequency over both independent sets of CPDAGs.

In the context of gene expression, a gene is usually controlled by a small number of regulator genes. In contrast, a regulator-gene can control an almost unlimited number of genes. Therefore, we apply the upper bound of the number of parents for each GE node: each GE node can have at most three GE parent nodes and one corresponding CNV node. There is no restriction on the number of METH parents for given GE node. CNV and METH nodes cannot have any parents. This restriction is supported by the biological literature [25].

Analyses were carried out using the free R software (www.r-project.org) version 4.0.0.

Data and Test Procedures

Objective assessment of regulatory network inference is challenging, especially if we are interested in multiple molecular features, such as gene expression, copy number variation, and DNA methylation. Simulated datasets cannot fully reflect underlying biological processes. There are multiple publicly available databases with well-known interaction pathways. However, well-known interaction pathways capture only simplified representations of mechanisms ongoing in a most population under given conditions. We can use them as the ground truth. Nevertheless, a certain level of disagreement is likely to emerge due to biological diversity.

As a test case for IntOMICS, we focus mainly on the Wnt/Wingless and MAPK signalling pathways that have a prominent role in cancer development. Hence, they have been extensively studied in the context of colon cancer, and most of the key gene-gene interactions have been revealed [26, 27, 28, 29]. Besides wild-type/normal tissue experiments, we assess the IntOMICS performance in samples affected by some perturbation (interventional gene expression data from knock-out experiments, tumour tissues).

Datasets

A summary of all data sets used in this study can be found in Table 2.

The Cancer Genome Atlas (TCGA) [30] provides publicly available multi-*omics* datasets for human cancers, including colon cancer and its histologically normal tissue adjacent to the tumour (NAT). The copy number variation of the associated genes from TCGA-COAD samples were downloaded from the Broad Institute GDAC Firehose (<https://gdac.broadinstitute.org/>; accessed 30 December 2020). We

Algorithm 1 IntOMICS

Initialisation: the random network structure G_s ; $\beta_s \sim U[0, 10]$; $\sigma_s = 5$; $d = 0$.

The first adaptive phase:

while β acceptance rate $\notin (0.28; 0.60)$ **do**

$\sigma_s = \ln(\sigma_s) \pm 0.05$

Propose a candidate network structure G_c .

Accept the candidate network structure G_c with the acceptance probability A given by eq. (7), otherwise leave the current network structure G_s .

Propose a candidate $\beta_c \sim N(\beta_s, \sigma_s^2)$; $\beta_c \geq 0.5$.

Accept the candidate β_c with the acceptance probability A_β given by eq. (8), otherwise leave the current β_s .

end while

The transient phase:

For every 200 iterations average the β values, and fit the linear model using 5 last averages to check if the chain values show a linear trend.

if p-value of the slope < 0.1 **then**

run next 200 iterations and fit new linear model

end if

The second adaptive phase:

while β acceptance rate < 0.02 **do**

$d = d + 1$

$\sigma_s = SD(\beta_{transient}) * 2.38 / (1.5^d)$

Estimate the *empirical* biological knowledge matrix B .

end while

default hyper-parameters setting from the automatically tuned MCMC algorithm [19] used in the first, transient, and second phase

The sampling phase:

β_s from the second adaptive phase is fixed.

for $i \leftarrow 1$ to $2 * burn-in$ (or until the c_{rms} stopping criterion is met) **do**

Run two independent MCMC simulations: G_{s_1} from the second adaptive phase; G_{s_2} empty network structure.

Propose a candidate network structure G_c .

Accept the candidate network structure G_c with the acceptance probability A given by eq. (7), otherwise leave the current network structure G_s .

end for

Discarding *burn-in* DAGs and thinning (save every 500th DAG from a sequence of resulting DAGs).

Conversion of DAGs into CPDAGs and saving only unique CPDAGs.

Edge weights calculation.

Table 2 Summary of all used data sets. MSS microsatellite stability; MSI microsatellite instability; NAT histologically normal tissue adjacent to the tumour; GE gene expression; CNV copy number variation; METH methylation probe

dataset	investigated gene set	samples	details
GSE127960 ZIC5 WT	16GE / 0CNV / 0METH	2	HCT116 colon cancer cell line
GSE127960 ZIC5 KO	16GE / 0CNV / 0METH	4	HCT116 colon cancer cell line, ZIC5 knockout
TCGA-COAD MSS	24GE / 24CNV / 4METH	27	primary tumour stage II/III with MSS phenotype
TCGA-COAD MSI	24GE / 24CNV / 3METH	69	primary tumour with MSI phenotype
TCGA-COAD NAT	24GE / 24CNV / 16METH	19	histologically normal tissue adjacent to the tumour
PETACC-3 MSS	39GE / 23CNV / 0METH	176	primary tumour stage III with MSS phenotype

use only samples with DNA methylation data from Illumina Infinium HumanMethylation450 (450K) BeadChip array available. The resulting subset of the TCGA-COAD consists of $n = 115$ samples (27 microsatellite stable phenotype (MSS) stage II/III, 69 microsatellite instable phenotype (MSI), 19 NAT). Several TCGA-COAD samples lack information on microsatellite status. These samples were classified into MSI/MSS groups using MSI gene expression signature [31].

GSE127960 includes the gene expression profiling of the human colon cancer cell lines HCT116 with ZIC5 wild-type (ZIC5 WT) and ZIC5 knockout (ZIC5 KO) replicates.

PETACC-3 clinical trial [32, 33] investigates the benefit of irinotecan in the adjuvant treatment of stage III colon cancer patients. We use gene expression and copy number variation data of 176 MSS stage III colon cancer samples (PETACC-3 MSS). Missing MSS/MSI phenotype was determined by the MSI gene expression signature mentioned above.

Evaluation criteria

Regulatory networks derived using TCGA-COAD MSI/NAT, and GSE127960 ZIC5 WT/KO are compared with the Wnt signalling pathway from the KEGG database, which is used as the ground truth. One of the main performance indexes is used the accuracy and 95% confidence intervals. The IntOMICS performance is also compared with the W&H algorithm - one of the most relevant gene regulatory network reconstruction tools based on Bayesian networks [4]. The W&H algorithm was designed to infer only dependencies among gene expression data. Thus, we exclude CNV-GE and METH-GE edges identified by IntOMICS in the performance comparison of these two algorithms. The criterion of the 83% confidence intervals overlap [34] is used to test if there is a difference in the accuracy of both algorithms.

To measure the agreement between IntOMICS and W&H algorithm, we add another performance metric – Cohen’s κ coefficient.

Edge weight in the resulting network structure reflects its empirical frequency over the final set of CPDAGs. From a practical point of view, the edge weight expresses confidence in the real edge presence. If not stated otherwise, we define a threshold for a presence of an edge as 0.75 quantile of all present edges in a given network structure.

β -catenin regulation in MSI colon cancer: Activating mutations in β -catenin (CTNNB1) result in decreased phosphorylation by GSK3 β and increased Wnt signalling through the Tcf/Lef transcription factors. These mutations are more frequent in microsatellite instable (MSI) colon cancer [35, 36, 37]. KEGG Colorectal cancer pathway has MSI specific information on missing interaction between GSK3 β and CTNNB1 genes. Therefore, we compare regulatory networks inferred by IntOMICS using both TCGA-COAD MSI samples and TCGA-COAD NAT samples.

SLC2A1 regulation in colon cancer cell lines: [38] found that the effectors of Wnt signalling β -catenin (CTNNB1) and TCF7L2 in collaboration with ZIC proteins directly regulate SLC2A1 gene expression. We choose gene expression of 16 genes from the KEGG Wnt signalling pathway together with ZIC5 and SLC2A1 genes and observe any difference in SLC2A1 regulation between ZIC5 WT and ZIC5 KO samples. GSE127960 data set was used to assess the IntOMICS performance only at the gene expression level.

CNVs specific for MSS colon cancer: The TCGA colon cancer (TCGA-COAD) microsatellite stable (MSS) samples were used to evaluate IntOMICS ability to infer dependencies among different molecular features - gene expression and copy number variation. In this part, we utilise gene expression, DNA methylation, and CNV data of TCGA-COAD MSS stage II/III samples ($n = 27$). We choose a subset of 24 genes from the KEGG Colorectal cancer pathway concerning CNVs identified by [33] in MSS stage II/III primary tumours. [33] identified several MSS specific aberrations. We focus on amplification of KRAS, MYC, BIRC5, CCND1, RAC3 and deletion of SMAD4.

DNA methylation specific for MSI colon cancer: The TCGA colon cancer (TCGA-COAD) microsatellite instable (MSI) samples have characteristic molecular biomarkers such as gene expression silencing through the MLH1 promoter hypermethylation [39, 40]. TCGA-COAD MSI samples were used to evaluate IntOMICS ability to infer dependencies among other molecular features - gene expression and DNA methylation. In this part, we utilise gene expression, DNA methylation, and CNV data of TCGA-COAD MSI samples ($n = 69$). We choose a subset of 24 genes from the KEGG Colorectal cancer pathway and MLH1. Because there are many methylation probes with significant regression coefficient and $R^2 > 0.3$, we perform differential methylation analysis using ChAMP R-package [41]. We select CpG island methylation probes from the promoter region with differential methylation between NAT and MSI samples (p -value < 0.05), and the absolute value of delta beta was greater than its 0.75 quantile.

To assess which of the tested algorithms estimate more biologically relevant gene-gene interactions, we define a set of edges identified exclusively by IntOMICS or W&H, respectively, and compare them against other publicly available interaction databases [42, 43, 44].

Real Application of InfOMICS: ABCG2-related chemoresistance in MSS stage III colon cancer

Resistance to chemotherapy is a major clinical problem in colon cancer treatment. Mechanisms associated with chemoresistance and novel biomarkers can identify patients with benefit from irinotecan-based regimens that could substantially improve

the selection of cancer therapy for the individual patient. PETACC-3 clinical trial [32] randomised colon cancer patients to fluorouracil/leucovorin (5FU/FA) +/– irinotecan treatment groups. The combination of ABCG2 and TOP1 gene expression significantly divided the stage III colon cancer patients enrolled in PETACC-3 into two groups regarding benefit from adjuvant treatment with FOLFIRI [45]. ABCG2 plays an essential role as an efflux transporter with various substrates, including chemotherapy drugs. Thus it is responsible for chemotherapy failure [46]. MYCN (by analogy, c-MYC) can contribute to chemoresistance by regulating a specific set of ABC transporter genes, including ABCG2. Direct interaction was determined by chromatin immunoprecipitation (ChIP) assays and luciferase-reporter assays [47]. The gene expression of ABCG2 could be increased by activation of mitogen-activated protein kinase cascade via phosphorylation of extracellular regulated kinase ERK1/2 and c-jun NH-terminal kinase/stress-activated protein kinase (JNK/SAPK) [48].

We use IntOMICS to investigate mechanisms associated with chemoresistance using 176 MSS stage III colon cancer samples from the PETACC-3 (PETACC-3 MSS) clinical trial [32, 33]. We select 37 genes from the KEGG MAPK signalling pathway together with MYCN and ABCG2 genes. Copy number variation data are available for 23 out of 39 selected genes. Both PETACC-3 MSS samples treated by 5FU/FA (PETACC-3 MSS 5FU/FA) and PETACC-3 MSS samples treated by 5FU/FA + irinotecan (PETACC-3 MSS FOLFIRI) are dichotomized by 5-year relapse-free survival (RFS) into high/low-RFS groups (Table 3).

Table 3 Number of samples in PETACC-3 clinical trial according to the treatment and relapse-free survival. 5FU/FA fluorouracil/leucovorin; FOLFIRI fluorouracil/leucovorin with irinotecan; RFS relapse-free survival

	5FU/FA	FOLFIRI
low-RFS	31	34
high-RFS	53	58

Results

The accuracy of IntOMICS and W&H algorithm in TCGA-COAD and GSE127960 datasets is shown in Fig. 2. In TCGA-COAD NAT data, IntOMICS accuracy is significantly higher than the W&H. In TCGA-COAD MSI data, the difference between IntOMICS accuracy and W&H accuracy is at the margin of statistical significance (the 83% confidence intervals share the endpoints). Even if the difference in the accuracy of both algorithms is not statistically significant in all test datasets, the results show IntOMICS achieves higher accuracy than the W&H algorithm.

Cohen’s κ ranges between 0.19 to 0.34. In the GSE127960 ZIC5 WT dataset, Cohen’s $\kappa = 0.19$ is the lowest, probably due to the lack of convergence in W&H MCMC simulation. As a result of the Markov chain poor convergence in the W&H algorithm, we can detect apparent anomalies in the trace plots of β values (Supplementary Figure 5a,b) and inconsistency of posterior probabilities of the edges (Supplementary Figure 6a,b). [4] decided to run a fixed number of MCMC simulations for performance evaluation. We followed the same procedure in the W&H application and found out that this setting is insufficient in some datasets. Consequently, using a measure of convergence, such as the c_{rms} [24], is necessary for every

MCMC simulation. IntOMICS trace plots of β values and the edge posterior probabilities of two independent MCMC simulations can be found in Supplementary Figures 7 and 8.

β -catenin regulation in MSI colon cancer: As expected, the interaction from GSK3 β to CTNNB1 is present in the IntOMICS regulatory network of TCGA-COAD NAT samples. The edge from GSK3 β to CTNNB1 is missing in TCGA-COAD MSI samples, even if $B_{ij} = 1$ for $i = \text{GSK3}\beta$ and $j = \text{CTNNB1}$. This result suggests IntOMICS can discover relevant data-derived interactions despite distinct prior knowledge. The W&H algorithm identified the above-mentioned interaction in both TCGA-COAD MSI and TCGA-COAD NAT datasets. If we examine interaction network derived from TCGA-COAD NAT, IntOMICS identified the interactions from PSEN1 to CTNNB1 and from CSNK1A1 to AXIN1 in NAT samples (the direct interactions present in KEGG). The W&H algorithm did not detect the edge. We can conclude that IntOMICS performs better in TCGA-COAD MSI/NAT datasets than the W&H.

In the TCGA-COAD NAT dataset, 22% and 33% of edges identified exclusively by the W&H and IntOMICS algorithm were found in other databases, respectively.

SLC2A1 regulation in colon cancer cell lines: Based on the [38] experiment, we assume SLC2A1 is directly regulated by ZIC5, CTNNB1, and TCF7L2 in ZIC5 WT samples ($B_{ij} = 1$). In ZIC5 WT samples, W&H revealed all three interactions. IntOMICS revealed TCF7L2 and CTNNB1 as direct regulators of SLC2A1, and ZIC5 was identified as directly regulated by SLC2A1. We expect a substantial difference in SLC2A1 regulation in ZIC5 KO samples. In ZIC5 KO samples, W&H revealed CTNNB1 as a direct regulator of SLC2A1 and ZIC5 as directly regulated by SLC2A1. IntOMICS identified a direct regulation from SLC2A1 to CTNNB1.

In the GSE127960 WT dataset, 20% and 29% of edges identified exclusively by the W&H and IntOMICS algorithm were found in other databases, respectively.

CNVs specific for MSS colon cancer: IntOMICS identified edges from CNV to associated GE in 5 out of 6 genes of interest: KRAS, MYC, BIRC5, CCND1, and SMAD4. Except five genes mentioned above, we should highlight other CNVs with direct interaction to gene expression identified by IntOMICS: deletion of tumour suppressor SMAD2 and amplification of proto-oncogenes MYC, KRAS and BRAF. The resulting regulatory network inferred by the IntOMICS algorithm using TCGA-COAD MSS stageII/III samples can be seen in Supplementary Figure 9.

DNA methylation specific for MSI colon cancer: IntOMICS identified 6 DNA methylation probes as MLH1 direct regulators together with associated CNV. There are no descendant nodes of the MLH1 gene. This is in concordance with the hypothesis that DNA methylation of the MLH1 promoter region influences its gene expression. The resulting regulatory network inferred by the IntOMICS algorithm using TCGA-COAD MSI samples can be seen in Supplementary Figure 10.

Presented results indicate a difference in favour of IntOMICS, and we can conclude that the proposed framework can compete with the W&H algorithm at the gene expression level.

Real Application of InfOMICS: ABCG2-related chemoresistance in MSS stage III colon cancer

There are several interesting differences in ABCG2 regulation in the PETACC-3 MSS samples. Genes with ABCG2 direct interaction determined by IntOMICS are listed in Table 4. In all PETACC-3 MSS samples, IntOMICS identified direct interaction from ABCG2 CNV to ABCG2 GE.

MAPK1/ERK2 and HRAS genes are identified as the direct regulators of the ABCG2 gene in the FOLFIRI high-RFS samples. In contrast, the ABCG2 regulation confidence is very low in other groups. In our study, these genes have the highest predictive potential. These results support the findings of [48] about the regulation of ABCG2 by MAPK1/ERK2, which can have therapeutic consequences.

The direct interaction from MAPK3/ERK1 to ABCG2 is detected in 5FU/FA low-RFS samples. In the context of ABCG2 regulation, we can speculate about the functional redundancy of MAPK1/ERK2 and MAPK3/ERK1.

Regardless of the treatment arm, RPS6KA1 directly regulates ABCG2 in high-RFS samples. In contrast, IntOMICS identified direct interaction between ABCG2 and FGFR4, MAPK3, RPS6KA3, and KRAS genes in all low-RFS samples.

The role of MYCN in ABCG2-related chemoresistance is not deciphered. In the FOLFIRI treatment arm, MYCN directly regulates ABCG2. In the 5FU/FA treatment arm, MYCN regulates ABCG2 with the edge weight ≥ 0.5 quantile of all present edges.

Table 4 Genes with ABCG2 direct interaction and the confidence of the regulation (w) determined by IntOMICS. Genes with the highest predictive potential are highlighted in bold. low $w < 0.5$ quantile of all edge weights; med $0.5 \geq w < 0.75$ quantile of all edge weights; high empirical frequency $w \geq 0.75$ quantile of all edge weights; NA the edge was not identified

	5FU/FA		FOLFIRI	
	low-RFS	high-RFS	low-RFS	high-RFS
FGFR4	high	low	low	NA
HRAS	low	low	low	high
KRAS	high	NA	low	NA
MAPK1/ERK2	low	low	low	high
MAPK3/ERK1	high	low	low	low
MAPK8	low	high	low	low
RPS6KA3	high	NA	NA	low
RPS6KA1	NA	high	low	high
ELK1	high	low	NA	med
GRB2	med	med	high	med
MAP3K1	high	low	low	med
MKNK1	med	high	low	med
MRAS	low	low	high	med
MYC	high	low	low	med
MYCN	med	med	high	high
RAF1	high	low	med	high
ABCG2 CNV	high	high	high	high

Discussion

We present IntOMICS, a Bayesian framework for multi-*omics* data and prior knowledge integration to infer regulatory networks using a novel approach to biological knowledge discovery. Besides the integration of known interactions as prior knowledge, IntOMICS complements the prior knowledge using *empirical* biological matrix, which reflects hidden patterns in the available datasets. IntOMICS is designed to infer not only dependencies among gene expression but also between gene expression, DNA methylation and copy number variation. Pathogenic copy number variations and epigenetic changes (such as DNA methylation) can affect gene expression, contribute to increased DNA instability and play an essential role in the initiation and progression of complex diseases such as cancer. The great benefit of IntOMICS is the use of continuous data. Because frequently used data discretisation in multi-*omics* data analysis implies substantial information loss. The proposed framework minimises the weaknesses of MCMC-based algorithms utilising state-of-the-art approaches such as Markov blanket resampling or adaptive MCMC algorithm. IntOMICS can be extended with any additional modality if the proposed model assumptions are satisfied (variables come from the multivariate Gaussian distribution).

Although IntOMICS evaluation and application is demonstrated using *multi-omics* data of colon cancer samples, it is not limited to any particular phenotype. IntOMICS is also designed to infer regulatory network, even if copy number variation or DNA methylation data (or both) are not available.

We have to mention that IntOMICS is not designed to infer genome-wide regulatory networks because of time complexity. At present, IntOMICS is restricted to infer regulatory networks with up to tens of features (nodes). Our next objective is to upgrade the proposed workflow to infer regulatory networks with an extensive set of features.

At the gene expression level, the performance of IntOMICS shows an increasing trend compared to a published algorithm based on Bayesian networks. The ability to capture relevant crosstalks between copy number variation and gene expression is

verified using known associations between copy number variation and gene expression in microsatellite stable stage II/III colon cancer samples. IntOMICS identified 5 out of 6 known associations. Microsatellite unstable samples were used to verify crosstalks between gene expression and methylation. IntOMICS identified 6 DNA methylation probes as MLH1 direct regulators together with associated CNV. Using our novel framework, several ABCG2 regulator genes are discovered as potential predictive markers in microsatellite stable stage III colon cancer samples. However, all regulatory relationships discovered by IntOMICS need to be verified using more refined approaches.

IntOMICS is a powerful resource for exploratory systems biology and can provide valuable insights into the complex mechanisms of biological processes that has a vital role in personalised medicine.

Declaration section

Availability of data and materials:

Source code and data used for producing the presented results are available at:

<https://gitlab.ics.muni.cz/bias/intomics>.

Competing interests:

The authors declare that they have no competing interests.

Author's contributions:

AP and VP designed the research. The code was developed and written by AP. The evaluation was performed by AP and VP. All authors have read and approved the final manuscript.

Acknowledgements:

Authors thank to Research Infrastructure RECETOX RI [LM2018121] financed by the Ministry of Education, Youth and Sports, and Operational Programme Research, Development and Innovation - project CETOCOEN EXCELLENCE [CZ.02.1.01/0.0/0.0/17.043/0009632] for supportive background.

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" CESNET [LM2015042] is greatly appreciated.

The published results here are partly based upon data generated by the TCGA Research Network:

<https://www.cancer.gov/tcga>.

Funding:

This work was supported by Czech Science Foundation (GACR) through grant no. 19-08646S.

Ethics approval and consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Author details

¹RECETOX, Faculty of Science, Masaryk University, Kotlarska 2, Brno, Czech Republic. ²Faculty of Informatics, Masaryk University, Botanická 68a, Brno, Czech Republic.

References

1. Djebbari, A., Quackenbush, J.: Seeded Bayesian Networks: Constructing genetic networks from microarray data. *BMC Systems Biology* **2** (2008). doi:10.1186/1752-0509-2-57
2. Gao, S., Wang, X.: Quantitative utilization of prior biological knowledge in the Bayesian network modeling of gene expression data. *BMC Bioinformatics* **12** (2011). doi:10.1186/1471-2105-12-359
3. Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., Miyano, S.: Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Journal of Bioinformatics and Computational Biology* **2**, 77–98 (2004). doi:10.1142/s021972000400048x
4. Werhli, A.V., Husmeier, D.: Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical applications in genetics and molecular biology* **6** (2007). doi:10.2202/1544-6115.1282
5. de Campos, L.M., Cano, A., Castellano, J.G., Moral, S.: Combining gene expression data and prior knowledge for inferring gene regulatory networks via Bayesian networks using structural restrictions. *Stat Appl Genet Mol Biol.* **18** (2019). doi:10.1515/sagmb-2018-0042

6. Xu, Y., Zhang, J., Yuan, Y., Mitra, R., Müller, P., Ji, Y.: A Bayesian graphical model for integrative analysis of TCGA data. *IEEE Int Workshop Genomic Signal Process Stat*, 135–138 (2012). doi:10.1109/GENSIPS.2012.6507747
7. Ni, Y., Stingo, F.C., Baladandayuthapani, V.: Integrative Bayesian Network Analysis of Genomic Data. *Cancer informatics* **13**, 39–48 (2014). doi:10.4137/CIN.S13786
8. Kim, D.-C., Kang, M., Zhang, B., Wu, X., Liu, C., Gao, J.: Integration of DNA Methylation, Copy Number Variation, and Gene Expression for Gene Regulatory Network Inference and Application to Psychiatric Disorders. In: *2014 IEEE International Conference on Bioinformatics and Bioengineering*, pp. 238–242 (2014). doi:10.1109/BIBE.2014.71
9. Zarayeneh, N., Ko, E., Oh, J.H., Suh, S., Liu, C., Gao, J., Kim, D., Kang, M.: Integration of multi-omics data for integrative gene regulatory network inference. *International journal of data mining and bioinformatics* **18**, 223–239 (2017). doi:10.1504/IJDMB.2017.10008266
10. Yuan, L., Guo, L.-H., Yuan, C.-A., Zhang, Y.-H., Han, K., Nandi, A., Honig, B., Huang, D.-S.: Integration of Multi-omics Data for Gene Regulatory Network Inference and Application to Breast Cancer. In: *EEE/ACM Transaction on Computational Biology and Bioinformatics*, vol. 16, pp. 782–791 (2018). doi:10.1109/TCBB.2018.2866836
11. Zhao, Y., Hoang, T.H., Joshi, P., Hong, S.-H., Giardina, C., Shin, D.-G.: A route-based pathway analysis framework integrating mutation information and gene expression data. *Methods* **124**, 3–12 (2017). doi:10.1016/j.ymeth.2017.06.016
12. Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K.B., Vieira, V., Bekker-Jensen, D.B., Kranz, J., Bindels, E.M.J., Costa, A.S.H., Sousa, A., Beltrao, P., Rocha, M., Olsen, J.V., Frezza, C., Kramann, R., Saez-Rodriguez, J.: Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol Syst Biol* **17** (2021). doi:10.15252/msb.20209730
13. Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J.D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P., Church, D.M.: DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res* **41**, 936–941 (2013). doi:10.1093/nar/gks1213
14. Komaki, S., Shiwa, Y., Furukawa, R., Hachiya, T., Ohmomo, H., Otomo, R., Satoh, M., Hitomi, J., Sobue, K., Sasaki, M., Shimizu, A.: iMETHYL: an integrative database of human DNA methylation, gene expression, and genomic variation. *Human Genome Variation* **5** (2018). doi:10.1038/hgv.2018.8
15. Su, C., Borsuk, M.E.: Improving Structure MCMC for Bayesian Networks through Markov Blanket Resampling. *Journal of Machine Learning Research* **17**, 1–20 (2016)
16. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* **9**, 309–347 (1992). doi:10.1007/BF00994110
17. Geiger, D., Heckerman, D.: Learning Gaussian Networks. In: *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pp. 235–243 (1994)
18. Grzegorzczak, M.: Comparative evaluation of different graphical models for the analysis of gene expression data. PhD thesis, The Department of Statistics of the University Dortmund. (2006). <http://dx.doi.org/10.17877/DE290R-607>
19. Yang, J., Rosenthal, J.S.: Automatically tuned general-purpose MCMC via new adaptive diagnostics. *Computational Statistics* **32**, 315–348 (2017). doi:10.1007/s00180-016-0682-2
20. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**, 29–34 (1999). doi:10.1093/nar/27.1.29
21. Madigan, D., York, J., Allard, D.: Bayesian Graphical Models for Discrete Data. *International Statistical Review / Revue Internationale De Statistique* **63**, 215–232 (1995). doi:10.2307/1403615
22. Hastings, W.K.: Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **51**, 97–109 (1970). doi:10.2307/2334940
23. Peterson, R.A., Cavanaugh, J.E.: Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics* **47**, 2312–2327 (2020). doi:10.1080/02664763.2019.1630372
24. Agostinho, N.B., Machado, K.S., Werhli, A.V.: Inference of regulatory networks with a convergence improved MCMC sampler. *BMC Bioinformatics* **16** (2015). doi:10.1186/s12859-015-0734-6
25. Leclerc, R.D.: Survival of the sparsest: robust gene networks are parsimonious. *Molecular systems biology* **4** (2008). doi:10.1038/msb.2008.52
26. Barker, N., Clevers, H.: Mining the Wnt pathway for cancer therapeutics. *Nature Reviews Drug Discovery* **5**, 997–1014 (2006). doi:10.1038/nrd2154
27. Sebio, A., Kahn, M., Lenz, H.-J.: The potential of targeting Wnt/ β -catenin in colon cancer. *Expert Opin Ther Targets* **18**, 611–615 (2014). doi:10.1517/14728222.2014.906580
28. Novellademunt, L., Antas, P., Li, V.S.W.: Targeting Wnt signaling in colorectal cancer. A Review in the Theme: Cell Signaling: Proteins, Pathways and Mechanisms. *Am J Physiol Cell Physiol* **309**, 511–521 (2015). doi:10.1152/ajpcell.00117.2015
29. Duchartre, Y., Kim, Y.-M., Kahn, M.: The Wnt signaling pathway in cancer. *Crit Rev Oncol Hematol* **99**, 141–149 (2016). doi:10.1016/j.critrevonc.2015.12.005
30. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013). doi:10.1038/ng.2764
31. Pačínková, A., Popovici, V.: Cross-platform Data Analysis Reveals a Generic Gene Expression Signature for Microsatellite Instability in Colorectal Cancer. *Biomed Res Int* (2019). doi:10.1155/2019/6763596
32. Cutsem, E.V., Labianca, R., Bodoky, G., Barone, C., Aranda, E., Nordlinger, B., Topham, C., Tabernero, J., André, T., Sobrero, A.F., Mini, E., Greil, R., Costanzo, F.D., Collette, L., Cisar, L., Zhang, X., Khayat, D., Bokemeyer, C., Roth, A.D., Cunningham, D.: Randomized Phase III Trial Comparing Biweekly Infusional Fluorouracil/Leucovorin Alone or With Irinotecan in the Adjuvant Treatment of Stage III Colon Cancer: PETACC-3. *Journal of Clinical Oncology* **27**, 3117–3125 (2009). doi:10.1200/JCO.2008.21.6663
33. Xie, T., d'Ario, G., Lamb, J.R., Martin, E., Wang, K., Tejpar, S., Delorenzi, M., Bosman, F.T., Roth, A.D.,

- Yan, P., Bougel, S., Narzo, A.F.D., Popovici, V., Budinská, E., Mao, M., Weinrich, S.L., Rejto, P.A., Hodgson, J.G.: A comprehensive characterization of genome-wide copy number aberrations in colorectal cancer reveals novel oncogenes and patterns of alterations. *PLoS one* **7** (2012). doi:10.1371/journal.pone.0042001
34. Goldstein, H., Healy, M.J.R.: The graphical presentation of a collection of means. *Journal of the Royal Statistical Society* **158**, 175–177 (1995)
 35. Mirabeli-Primdahl, L., Gryfe, R., Kim, H., Millar, A., Luceri, C., Dale, D., Holowaty, E., Bapat, B., Gallinger, S., Redston, M.: β -Catenin Mutations Are Specific for Colorectal Carcinomas with Microsatellite Instability but Occur in Endometrial Carcinomas Irrespective of Mutator Pathway. *Cancer Research* **59**, 3346–3351 (1999)
 36. Fukushima, H., Yamamoto, H., Itoh, F., Horiuchi, S., Min, Y., Iku, S., Imai, K.: Frequent alterations of the beta-catenin and TCF-4 genes, but not of the APC gene, in colon cancers with high-frequency microsatellite instability. *J Exp Clin Cancer Res* **20**, 553–559 (2001)
 37. Kim, S., Jeong, S.: Mutation Hotspots in the β -Catenin Gene: Lessons from the Human Cancer Genome Databases. *Mol Cells* **42**, 8–16 (2019). doi:10.14348/molcells.2018.0436
 38. Zhao, Z., Wang, L., Bartom, E., Marshall, S., Rendleman, E., Ryan, C., Shilati, A., Savas, J., Chandel, N., Shilatifard, A.: β -Catenin/Tcf7l2-dependent transcriptional regulation of GLUT1 gene expression by Zic family proteins in colon cancer. *Science Advances* **5** (2019). doi:10.1126/sciadv.aax0698
 39. Herman, J.G., Umar, A., Polyak, K., Graff, J.R., Ahuja, N., Issa, J.-P.J., Markowitz, S., Willson, J.K.V., Hamilton, S.R., Kinzler, K.W., Kane, M.F., Kolodner, R.D., Vogelstein, B., Kunkel, T.A., Baylin, S.B.: Incidence and functional consequences of hmlh1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci USA* **95**, 6870–6875 (1998). doi:10.1073/pnas.95.12.6870
 40. Kuusmanen, S.A., Holmberg, M.T., Salovaara, R., de la Chapelle, A., Peltomäki, P.: Genetic and epigenetic modification of mlh1 accounts for a major share of microsatellite-unstable colorectal cancers. *The American Journal of Pathology* **156**, 1773–1779 (2000). doi:10.1016/S0002-9440(10)65048-1
 41. Tian, Y., Morris, T.J., Webster, A.P., Yang, Z., Beck, S., Feber, A., Teschendorff, A.E.: ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* **33**, 3982–3984 (2017). doi:10.1093/bioinformatics/btx513
 42. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R.: IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, 452–455 (2004). doi:10.1093/nar/gkh052
 43. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, 535–539 (2006). doi:10.1093/nar/gkj109
 44. Wu, G., Haw, R.: Functional Interaction Network Construction and Analysis for Disease Discovery. *Methods Mol Biol.* **1558**, 235–253 (2017). doi:10.1007/978-1-4939-6783-4_11
 45. Stenvang, J., Budinská, E., van Cutsem, E., Bosman, F., Brünner, V.P.N.: An Explorative Analysis of ABCG2/TOP-1 mRNA Expression as a Biomarker Test for FOLFIRI Treatment in Stage III Colon Cancer Patients: Results from Retrospective Analyses of the PETACC-3 Trial. *Cancers* **12** (2020). doi:10.3390/cancers12040977
 46. Doyle, L.A., Douglas, R.D.: Multidrug resistance mediated by the breast cancer resistance protein BCRP (ABCG2). *Oncogene* **22**, 7340–7358 (2003). doi:10.1038/sj.onc.1206938
 47. Porro, A., Haber, M., Diolaiti, D., Iraci, N., Henderson, M., Gherardi, S., Valli, E., Munoz, M.A., Xue, C., Flemming, C., Schwab, M., Wong, J.H., Marshall, G.M., Valle, G.D., Norris, M.D., Perini, G.: Direct and coordinate regulation of ATP-binding cassette transporter genes by Myc factors generates specific transcription signatures that significantly affect the chemoresistance phenotype of cancer cells. *The Journal of Biological Chemistry* **285**, 19532–19543 (2010). doi:10.1074/jbc.M109.078584
 48. zu Schwabedissen, H.E.M., Grube, M., Dreisbach, A., Jedlitschky, G., Meissner, K., Linnemann, K., Fusch, C., Ritter, C.A., Völker, U., Kroemer, H.K.: Epidermal growth factor-mediated activation of the map kinase cascade results in altered expression and function of ABCG2 (BCRP). *Drug metabolism and disposition* **34**, 524–533 (2006). doi:10.1124/dmd.105.007591

Additional Files

Additional file 1 — Supplementary_Figures_IntOMICS.pdf

Supplementary Files

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

- [SupplementaryFiguresPacinkovaPopovici.pdf](#)