

1 Fast and robust method for drug response biomarker 2 identification and sample stratification.

3 *Simanti Bhattacharya** and Amit Das** (*contributed equally)*

4 *Excelra Knowledge Solution Pvt Ltd., NSL SEZ ARENA, IDA Uppal, Hyderabad-500039, India.*

5 **#co-corresponding authors:**

6 *simanti.bhattacharya@gmail.com and amit.das.ku@gmail.com*

7

8 **Highlights:**

- 9 1. Supervised machine learning approach to analyze gene expression data.
- 10 2. Drug response biomarker identification.
- 11 3. Categorization of samples for their drug response with the help of identified
12 biomarkers.
- 13 4. Functional enrichment to understand the biomarkers association with biological
14 processes.
- 15 5. Bayesian network analysis to develop causal structure among identified
16 biomarkers and drug targets.
- 17 6. Time and cost-effective pipeline for fast and robust prediction of drug response
18 biomarkers.

19

20 **Abstract:**

21 With unprecedented progress of cancer research, the world is now prepared with
22 versatile arsenal of drugs to combat cancer. However, individual's response to any
23 drug or combination treatment stands as a major challenge and hence there exists the
24 sheer need for personalized medication. Identification of drug response biomarkers
25 from a wholistic tumor microenvironment analysis would guide researchers to develop
26 custom-tailored treatment regimen.

27 In this study, a fast and robust method has been developed to identify drug
28 response biomarkers from entire transcriptomics data analysis in a data-driven
29 manner. The biomarkers which were identified by the method, were able to stratify
30 patients between responders vs non-responders population. Furthermore, bayesian
31 network (BN) analysis, done on the data, brought forth a mechanistic insight into the
32 role of identified biomarkers in regulating drug's efficacy.

33 The importance of this work lies with the protocol that is time saving and requires
34 less computation power, yet analyzes a whole system data and helps the researchers

35 to take a step forward towards the development of personalized care in effective
36 cancer treatment.

37

38 **Keywords:** cancer; biomarker; machine learning; bayesian network; transcriptomics

39

40 **1. Introduction:**

41 The development of the tumor and its microenvironment is dependent on multi-
42 level systemic regulations [1]. Tumor developmental stages, benign to malignant
43 transition [2], tumor immune escape mechanism [3], epigenetic modulation [4] as well
44 as disease specific mutational switches [5] etc., all regulate the disease prognosis. In
45 addition to these intrinsic modulators, external risk exposure, such as smoking habit,
46 life style, eating habit etc adds another level of complexity [6,7]. Interestingly, with the
47 unmatched speed of cutting-edge technologies and researches, present world is much
48 prepared with essential therapies and medications [8–11], monotherapy or
49 combination, to satisfy the need (1) to fight against cancer progression, (2) to increase
50 the patient's life expectancy and (3) to prohibit the relapse of the disease. Effective
51 cancer therapies can be of several types [12]: surgery (surgical removal of affected
52 tissue), radiation (high doses of radiation applied to kill cancer cells and shrink tumors),
53 chemotherapy (use of chemotherapeutic agents to kill cancer cells), immune therapy
54 (to elicit effective immune system to inhibit tumor cell's immune escape and to destroy
55 tumor cells precisely) and targeted therapy (use of specific drugs that targets cancer
56 specific antigens, receptors etc). For certain cancers, hormone therapy and stem cell
57 transplantation have shown very promising results. These medications, as
58 monotherapy or as a combination, have been very effective to pull the chain of rapid
59 spread of cancer.

60 However, there exists a gridlock that obstructs these cancer therapies to be
61 used effectively in an individual patient. Owing to the complexity of tumor
62 microenvironment, an individual patient responds differently to the given treatment
63 regimen. Hence arises the need for personalized medication [13–15] wherein the
64 treatment regimen will be custom-tailored for the patient to have the best outcome of
65 the therapy. In order to fulfill this efficient treatment, the identification of drug response
66 biomarkers [16] becomes essential. Identification of drug response biomarkers has
67 become an integral part of modern-day personalized therapy [17,18]. With the

68 advancement of data-driven approaches [19,20], the concept behind ‘one-drug-one-
69 target’ has changed. The response outcome of a drug depends on many factors, such
70 as mutations, compensatory mechanisms, epigenetics, dysregulated genes and many
71 more, making a prediction of drug’s efficacy a real challenge for decades. These
72 regulatory factors either could be in the immediate vicinity of drug’s mechanism of
73 action (MoA) or could be trans-regulatory that acts by modulating relevant genes’
74 expression. These down-streams effects (i.e. gene’s up or down-regulations) are easy
75 to assess and can directly be linked to pathways or cellular processes, thereby, giving
76 an answer to how and why a patient could be sensitive to a given therapy [21]. The
77 advantage of such a data-driven approach is that it is free from any pre-conceptualized
78 bias like drug targets, disease genes etc.

79 However, dealing with the enormous amount of gene expression data is both
80 time and computation expensive. There is a sheer need for a rapid and robust method
81 for biomarker identification which can lead to stratify patients as per their response to
82 the drug. In this study, transcriptomic data has been analyzed through a machine
83 learning pipeline which has been developed for rapid identification of drug response
84 biomarkers and thereafter patient stratification.

85 The work has used the data from the experiment submitted to NCBI GEO as
86 GSE2535 [22] where chronic myeloid leukemia (CML) is the disease system and
87 imatinib is the drug. Briefly, CML is a myeloproliferative disorder caused by constitutive
88 tyrosine kinase activity of Bcr-Abl oncoprotein. Although Bcr-Abl-inhibitor imatinib
89 stands the first-line therapy against CML, almost 20-30% of patients develop drug
90 resistance [23]. In general, mutations in the Bcr-Abl domain increase the formation of
91 imatinib resistance. However, not all resistant patients are Bcr-Abl mutation dependent
92 which triggers the thought that there exist non-mutant mechanisms leading to imatinib
93 resistance.

94 After analyzing the transcriptomic data, 12 genes have come up as potential
95 biomarkers which have efficiently categorized unlinked patient samples by their drug
96 response. Further to this analysis, potential links between this drug and the identified
97 biomarkers have been established through function enrichment. However,
98 identification of biomarkers, coupled with their functional analysis may not satisfy the
99 query regarding the causality behind these potential biomarkers and drug’s activity.

100 To understand the system-level regulatory mechanism, here, bayesian network (BN)
101 analysis [24,25] has been employed which developed a probabilistic model of the
102 network with structural characteristics and directed acyclic networks in two different
103 classes: responders vs non-responders so as to enlighten the regulatory map behind
104 the scene.

105 **2. Materials and methods:**

106 **2.1. Patients description:** Patient description was explicitly described in [22].

107 Authors collected two sets of patients: (1) chronic myeloid leukemia (CML)
108 patients at chronic phase 1 from Novartis-sponsored trials at the Department
109 of Hematology of the University of Leipzig, Germany (n=15) and (2) German
110 CML Study Group at the Faculty of Clinical Medicine Mannheim of the
111 University of Heidelberg, Germany (n=14). In both cases, the patient's
112 response to imatinib was captured and cellular lysates for RNA extraction were
113 stored prior to imatinib therapy. Later on, one patient's data was ignored due
114 to defective chip data. Hence total 28 patient samples were analyzed in the
115 work (GSE2535; [22]) which was also utilized in current research. There was
116 no significance difference in average age as well as the duration of the disease
117 between two sets of patients. There were 15 female and 13 male patients.

118 Responders and non-responders were defined as follow: Responders
119 (R) to imatinib who had achieved complete cytogenetic response (CCR) <9
120 months (n=16), while non-responders (NR) to imatinib who had failed to
121 achieve a major cytogenetic response (MCR) within one year of treatment
122 (n=12).

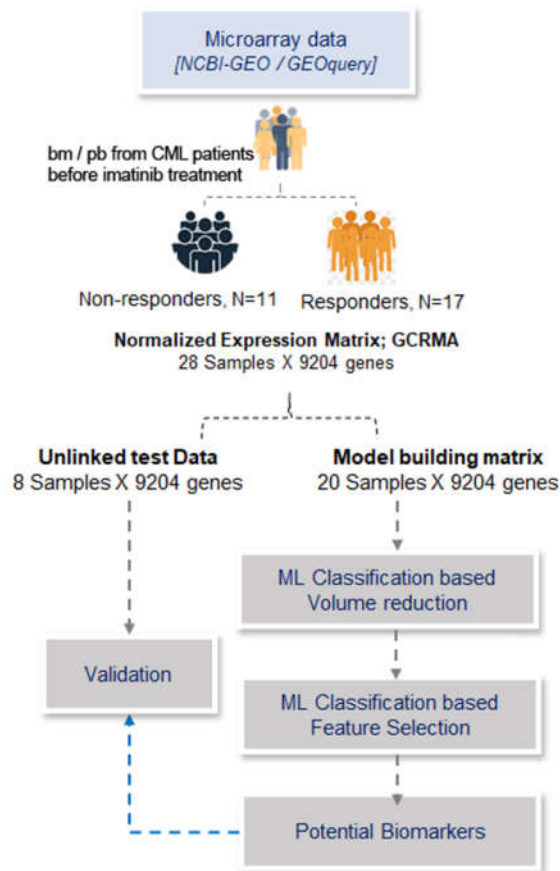
123 **2.2. RNA extraction and microarray:** Details of the sample collection methods
124 and equipment were explained in [22]. Briefly, bone marrow or peripheral blood
125 samples were collected from these patients, before Imatinib treatment. RNA
126 was extracted either with RNeasy kit or cesium-chloride gradient purification
127 method. Quality of RNA extracted was assured with the presence of discrete
128 18S and 28S ribosomal RNA peaks and the absence of irregularly-sized low
129 molecular weight RNA species in the electropherogram. The microarray was
130 performed on Affymetrix chip following standard protocol. Raw samples were
131 downloaded from NCBI GEO and then were normalized with GCRMA.

132 **2.3. Data processing:**

133 The normalized matrix had 12625 genes. Median over genes were done to
134 obtain 9204 unique genes. Samples were labeled according to the response
135 (coded as “1”) o resistance (coded as “-1”) towards the drug. Samples were
136 then randomly shuffled.

137 2.4. Machine learning:

138 **2.4.1. Model building:** 20 samples were used to build the model and to train
139 it. Remaining 8 samples were kept aside from the beginning that would be
140 used as an external or unlined test sample to validate the output. Details
141 of the workflow have been given in Fig. 1.



142

143 **Fig. 1. Detailed workflow for biomarker selection.** This is a schematic diagram
144 of the pipeline used in this study. The work was started with a microarray data 9204
145 genes and 28 CML patients whose drug response was known. Bone marrow (BM)
146 or Peripheral blood (pb) samples for microarray analysis were collected before
147 imatinib treatment. For the main model building, 20 samples were used while 8
148 samples were kept aside for validation. Different ML algorithms were used
149 stepwise to obtain drug response genes.

150

151 The matrix had samples in rows and genes in columns. These genes were
152 considered as variables and genes with the expression pattern were
153 considered as features. In the machine learning algorithm, this 20
154 samples, were further subdivided into training samples (75% of 20
155 samples) and testing samples (25% of 20 samples). Parameters were
156 optimized to improve the prediction of this test set response.

157 **2.4.2. Feature selection:** Feature selection step was preceded by volume
158 reduction with **V**ariable **S**election **U**sing **R**andom**F**orest (VSURF [26])
159 which uses the randomForest method to select significant variables in case
160 of high dimensional data. The reduced matrix had all insignificant gene
161 pruned out and only relevant ones were then given as input in feature
162 selection step where randomForest [27] classification was used.
163 Classification based analysis was done with 500 trees and 34 splitting at
164 each tree node. Out-of-bag error estimation was also calculated. Class
165 error probability is given in the table below

		Predicted		
		non-responder(-1)	responders (1)	class.error
Observed	non-responder (-1)	6	2	0.25
	responder (1)	1	11	0.083333

166 **Table 1. Class error probability for randomForest prediction.** “-1”
167 stands for non-responder and “1” stands for responder class. For non-
168 responder and responder classes, there are 0.25% and 0.08% error
169 probability, respectively for the randomForest training model.

171 Variables were selected based on their importance weight which was
172 equivalent to their contribution towards the sample’s drug response.

173 **2.4.3. Validation:** The knowledge of final features was then applied to the 8
174 samples, kept aside as external test cases to predict their response
175 pattern. Finally, observed and predicted responses were compared to
176 understand the prediction accuracy of the selected feature genes. For
177 validation, 3 different prediction algorithms (**S**upport **V**ector **M**achine: SVM
178 [28], **r**andom**F**orest: RF and **k**-**N**earest **N**eighbor: k-NN [21]) were used to
179 increase the confidence and robustness of prediction.

180 **2.5. Enrichment analysis**

181 Gene ontology (GO) [29,30] to gene association data for Homo sapiens was
182 downloaded from gene ontology consortium (<http://www.geneontology.org/>).
183 GO namespace and category were retrieved from .obo file. These two files
184 were used to create GO to gene dataset which was further used to enrich GO
185 terms (only biological processes, BP) for the selected genes. Hypergeometric
186 p-value was generated to obtain relevant GO-BP terms.

187 **2.6. Bayesian network**

188 Bayesian networks (BN) [24,25,31] are a type of probabilistic graphical model
189 that can be used to build models from data and/or expert opinion. BN is also
190 called a directed acyclic graph or DAG. It constitutes of nodes (each node
191 represents a variable such as genes. A variable might be discrete, such as
192 drug response = {Yes, No} or might be continuous such as gene expression
193 values) and links/edges (connections with directions across node pairs). The
194 absence of link between a pair of nodes does not mean that they are
195 completely independent; rather they may be connected via other nodes.
196 However, such node pairs may become dependent or independent depending
197 on the evidence of other nodes. Each gene pair has a source node/gene and
198 a target node/gene which is associated with a strength value (probability of
199 having this gene pair) and a direction value (probability of having the said
200 direction).

201 In this study, expression data-driven BN was constructed using bnlearn
202 package [32] in R with an aim decipher potential gene regulatory networks or
203 protein interaction networks. Separate networks were created for the
204 responders and the non-responders group. For each group, the input dataset
205 was a matrix with samples in rows and genes in columns. Both datasets had
206 20 genes (Machine learning derived genes paired with known drug target
207 genes). While the responders dataset had 17 samples, the non-responders
208 dataset had 11 samples. Each BN was validated through k-fold cross validation
209 and post boot-strapping, arcs or edges with a minimum of 75% direction and
210 strength value were selected and compared through the R compareDF
211 package.

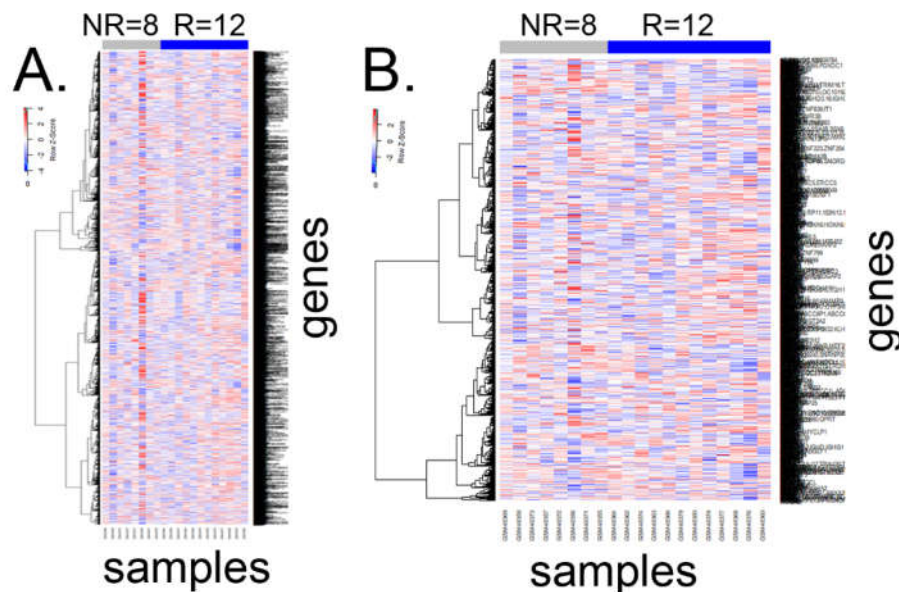
212 **3. Results and discussions:**

213 **3.1. Feature selection**

214 Biological data is unique in nature where variable size (number of genes) is
215 much greater than sample (number of observations) size. Moreover, variables
216 i.e. genes have an influence on each other either directly or through some other
217 genes [33]. The enormous variable size creates computation issues while
218 performing machine learning on such data in the local system. The high
219 dimension of data must be reduced so that machine learning can be performed
220 on less number but significant genes.

221 In this work, the entire gene expression matrix (with 9204 genes) was
222 used (Fig. 2A). From this enormous dataset, significant genes were selected
223 using VSURF algorithm [26] which selects important variables by recursive
224 removal of insignificant or low weight genes. Hence, the matrix dimension in
225 our case, reduced to 1184 genes and 20 samples (Fig. 2B). Performing all
226 three steps of VSURF is a time-consuming process. That is why only VSURF
227 threshold step was performed.

228

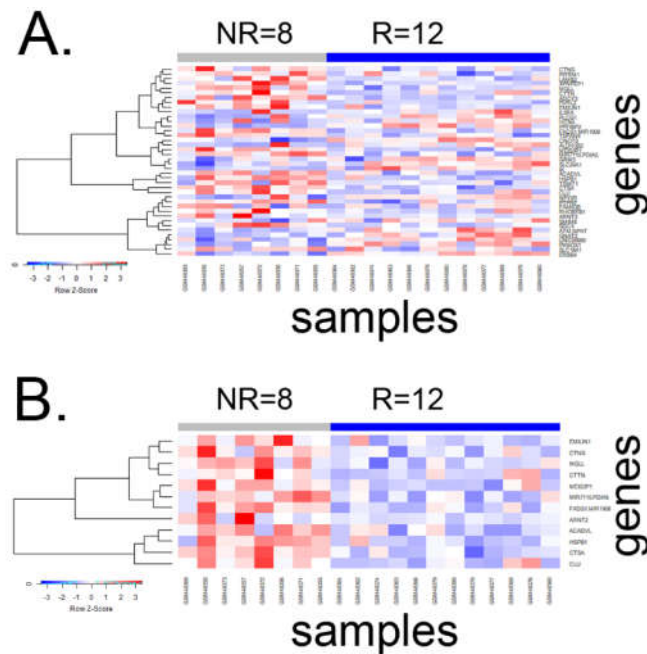


229

230 **Fig. 2. Heatmap for gene expression pattern.** Heatmaps for entire gene set,
231 n=9204 (A) and after VSURF mediated volume reduction, n=1184 (B) are shown
232 here. NR stands for non-responders, 8 samples and R stands for responders, 12
233 samples. This large set of genes are not distinctly different in both classes of
234 samples. In both cases, samples are shown in the x axis and genes are shown in
235 the y axis. During heatmap generation, expression values were scaled. Given color-
236 scale indicates that comparative lower expressions are shown in shades of blue
237 while comparative higher expressions are shown in shades of red.

238

239 Thereafter RF was applied to this model system which estimated the
 240 importance or weights of each predictor variable (here, gene) in modelling the
 241 classes ("1" & "-1"). These weights were equivalent to genes importance in
 242 classifying samples into two distinct classes. 40 genes were selected with
 243 weight ≥ 0.15 criteria (Fig. 3A) and with subsequent steps finally 12 genes
 244 were selected (Fig. 3B) that showed distinct expression pattern between two
 245 classes.
 246



247
 248 **Fig. 3. Feature selection using a data-driven approach.** (A) With randomForest
 249 cut-off >0.15 , 40 genes were selected. The heatmap shows there exist patches of
 250 difference in genes expression between responders(R) and non-responders(NR).
 251 (B) with further shrinking, 12 genes were selected. Their expressions are strikingly
 252 different between two classes, i.e. R vs NR. In both cases, again, samples are
 253 shown in the x axis and genes are shown in the y axis. During heatmap generation,
 254 expression values were scaled. Given color-scale indicates that comparative lower
 255 expressions are shown in shades of blue while comparative higher expressions
 256 are shown in shades of red.

257

258 The identified potential biomarkers with their weights have been given in table 2.

GENE SYMBOL	WEIGHT	GENE NAME	GENE ID
CLU	-0.5344	Clusterin	1191
HSPB1	-0.2965	Heat shock protein family B small member 1	3315
ARNT2	-0.25353	Aryl hydrocarbon receptor nuclear translocator 2	9915
MGLL	-0.24054	Monoglyceride lipase	11343

CTSA	-0.21896	Cathepsin A	5476
CTTN	-0.20909	Cortactin	2017
MIR7110.PDIA5	-0.20887	Protein disulfide isomerase family A member 5	10954
ACADVL	-0.18557	Acyl-CoA dehydrogenase very long chain	37
EMILIN1	-0.16605	Elastin microfibril interfacier 1	11117
CTNS	-0.16163	Cystinosin, lysosomal cystine transporter	1497
FADS1.MIR1908	-0.16033	Fatty acid desaturase 1	3992
MEIS3P1	-0.15428	Meis homeobox 3 pseudogene 1	4213

259 **Table 2. Identified biomarkers with their weights.** Identified biomarkers (i.e.
 260 genes) with their ML weights are given in the table. A negative weight corresponds
 261 to comparative lower expression in drug-responding samples. Gene names and
 262 Gene IDs are also mentioned.

263

264 3.2. Feature selection-based sample stratification

265 A sub matrix was then formed with these 12 genes and 20 samples which was
 266 further used to train model of three different algorithms: SVM, RF and k-NN.
 267 All these three algorithms are distinct from each other. This trained model was
 268 used to predict the response of external test data (8 samples) with 75%
 269 accuracy (table 3).

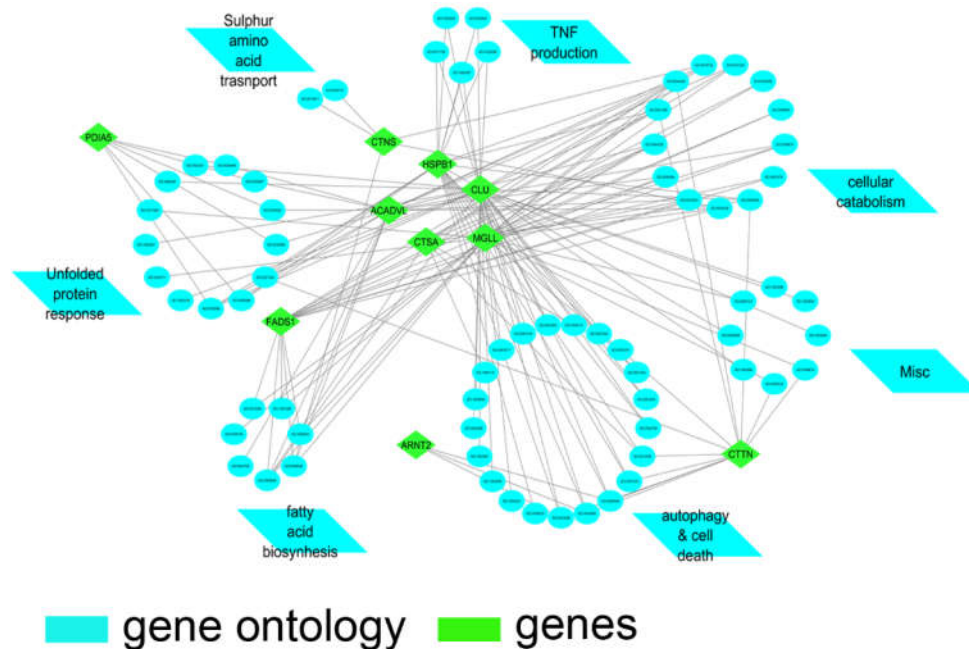
		Different predictive algorithms			
Unlinked Sample N=8	Actual Response	k-NN	randomForest	SVM	Inference
GSM48381	Responder	Non-Responder	Responder	Responder	Responder
GSM48370	Non-Responder	Non-Responder	Non-Responder	Non-Responder	Non-Responder
GSM48354	Non-Responder	Responder	Responder	Responder	Responder
GSM48361	Responder	Responder	Responder	Non-Responder	Responder
GSM48375	Responder	Responder	Responder	Responder	Responder
GSM48365	Responder	Responder	Responder	Responder	Responder
GSM48367	Non-Responder	Responder	Responder	Responder	Responder
GSM48359	Responder	Responder	Responder	Responder	Responder

270 **Table 3. Feature selection-based sample stratification.** The identified biomarkers
 271 and their weights were used to determine the response outcome of remaining 8
 272 samples, unlinked to the main training data. Three different prediction algorithms were
 273 used to predict the outcome. The “Inference” column in the table, holds the cumulative
 274 prediction from all three prediction algorithms. A comparison between “Inference” and
 275 “Actual Response” declares that the pipeline has predicted the outcome of 6 out of 8
 276 samples correctly, i.e. 75% accuracy.

277

278 **3.3. Functional analysis:** Identified genes that could predict the response of
 279 samples to the drug, could be used as potential drug response biomarkers.
 280 Hence it was required to understand their link to drug’s mechanism of action.

281 Most often the not, it has been found that the identified genes might regulate
282 the system distantly, instead of direct regulation, culminating to alternation in
283 response to the drug. In order to decipher the processes, GO functional
284 enrichment was carried out with the selected genes. This analysis enriched 70
285 terms (FDR<0.05). These terms were very specific which were required to be
286 clustered to have a broad category. Revigo webtool (<http://revigo.irb.hr/>) was
287 used to categorize the specific GO terms into six distinct classes and one
288 generic class, grouped as 'other' (Fig. 4).



289

290 **Fig. 4. Functional enrichment analysis.** Gene ontology enrichment was
291 carried out with hypergeometric p-value calculation using identified biomarkers
292 or feature genes (green, rhombus). Selected significant gene ontology (cyan,
293 circle) terms (FDR <0.05) were further categorized into broad clusters (cyan,
294 parallelogram) using Revigo webtool.

295

296 Published experiments were reviewed to establish the link between drug's
297 efficacy and the enriched processes.

298 **3.3.1. Regulation of autophagy:** Autophagy enables recycling of intracellular
299 ingredients as an alternative source during metabolic stress or starvation,
300 especially in cancer cells, to maintain cellular homeostasis and survival.
301 Helgason *et al.*,(2011) experimentally proved that autophagy inhibition in
302 combination with TKIs (imatinib, dasatinib, or nilotinib) resulted in almost
303 elimination of CML stem cells [34]. Moreover, overexpression and

304 knockdown of WNT2 have demonstrated perturbations in the autophagy
305 process and thereby CML drug sensitivity to Imatinib.

306 **3.3.2. Response to topologically incorrect protein:** Topologically incorrect
307 or unfolded protein response (UPR) enables neoplastic cells to resist TKI
308 therapeutics, such as imatinib [35].

309 **3.3.3. Fatty acid biosynthesis:** Certain fatty acid metabolic markers guide
310 resistance towards imatinib [36].

311 **3.3.4. Cellular catabolism:** This representative class majorly includes
312 monocarboxylic acid biosynthesis and fatty acid metabolism processes.
313 Activation fatty-acid oxidation (FAO) (which is compensatory to imatinib
314 mediated suppression of BCR-ABL and glycolysis), contribute to glucose-
315 independent cancer cell survival and thereby resistance to Imatinib therapy
316 [37]. Therefore, patients with activated FAO may exhibit Imatinib
317 resistance.

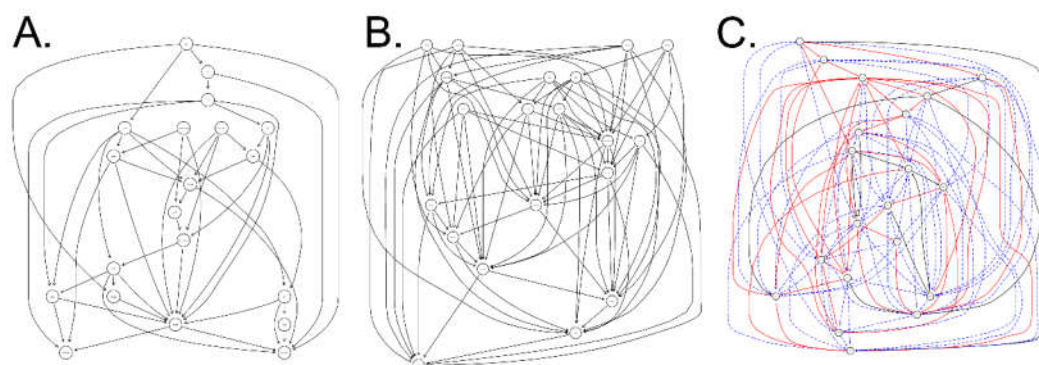
318 **3.3.5. Tumor necrosis factor (TNF) production:** This representative class
319 includes processes like tumor necrosis factor (TNF) superfamily cytokine
320 productions. Experimental evidence showed that FOXO3a activation could
321 trigger increased expression of tumor necrosis factor-related apoptosis-
322 inducing ligand (TRAIL) which could contribute to overcoming imatinib
323 resistance. In general, FOXO3a functions downstream of Bcr–Abl tyrosine
324 kinase in its phosphorylated inactive form in CML, which is reversed by
325 Imatinib activity bringing FOXO3a in its activated form which further
326 increases production of TNF family proteins. Hence TNF superfamily
327 proteins' production is an indication of imatinib sensitivity [38,39].

328 **3.3.6. Sulfur amino acid transport:** Major sulfur containing amino acids in
329 cells are methionine, cysteine, homocysteine, and taurine. Sulphur
330 containing amino acids have great implications in the tumor
331 microenvironment. However, evidence linking sulfur amino acid transport
332 to Imatinib response is yet to be validated.

333 **3.3.7. Others:** This group has different biological processes, such as axon
334 extension, neurofibrillary tangle assembly, endocannabinoid signaling
335 pathway, amyloid-beta formation and cell migration by vascular endothelial
336 growth factor signaling pathway.

337 3.4. Network analysis

338 As was mentioned earlier, it is essential to know the regulatory link between
339 identified genes and drug's MoA or drug's direct targets. To address this, BN
340 was created for responders and non-responders, followed by comparison
341 between two networks (Fig. 5A-C). The network architecture for responder
342 samples (Fig. 5A) differed significantly from that of non-responder samples
343 (Fig. 5B). The comparative figure (Fig. 5C) showed new connection that
344 appeared (indicated with blue dashed line), disappeared (red solid line) or
345 remained unchanged (black line) in non-responders.



346

347 **Fig. 5. Bayesian network analysis.** Bayesian network was created for (A)
348 responders and (B) non-responders. (C) is the comparison between two networks.
349 Blue dashed lines appear in non-responders while red lines disappear in them.
350 Black lines remain unchanged in both cases.

351

352 BN analysis of the non-responder population identified one unique
353 connection to them – ACADVL to PDIA5 (MIR7110) (Table 1). Both of these
354 two proteins are well established for their roles in unfolded protein response
355 (UPR) in the endoplasmic reticulum (ER). Recent studies suggested that these
356 two proteins play important roles in cancer drug-resistance and closely
357 interacts with GPR78 chaperon and transcription regulator TRF6 α [40–43]. In
358 our dataset also both ACADVL and PDIA5 were over-expressed in the non-
359 responders group when compared to the expression patterns of the
360 responders group. Further studies on the not so well understood UPR – ER
361 Stress – Immune response mechanism could provide vital information on the
362 development of novel therapeutics in several cancers [44].

363 Cortactin (CTTN) and Clusterin (CLU) are well established regulators of cancer
364 metastasis and drug resistance [45–47] as well as are associated with the gene

365 ontology biological process - negative regulation of apoptotic signaling
366 pathway (GO:2001234). While CLU has a chaperone role, CTTN works as an
367 actin cytoskeleton regulator downstream to Src kinase, another protein dis-
368 regulated in different cancers [48]. While the direct relationship of CTTN and
369 CLU are not well reported, over-expression of these two proteins is associated
370 with poor cancer prognosis. In our dataset also, these two proteins exhibited
371 over-expression in the non-responders group while responders exhibited lower
372 expressions of both the proteins. CLU-CTTN connection being unique to the
373 responders group (Table 4), it can be guessed that their lower expression is
374 far more important for imatinib sensitivity than over expression mediated
375 imatinib resistance.

376 Putative homeobox protein Meis3-like 1 (MEIS3P1) protein is poorly
377 understood but has been found to be dis-regulated in cancer related contexts
378 [49–52]. Our data also suggest that its lower expression is potentially
379 associated with better imatinib sensitivity. Similarly, ABL1 and ACADVL
380 relationship under cancer context is not well established but both could
381 potentially play important roles in SOCS3 mediated cancer prognosis. CSF1R
382 is well known for having a beneficial effect on several cancers when inhibited
383 by drugs [53] but its relation with BN predicted interactors like KIT, CTNS,
384 CTSA and EMILIN1 are not yet well established. CTNS and CTSA are
385 regulators of lysosome while EMILIN1 is known to have a positive impact on
386 cancers [54–56]. Further studies would be required to decipher the interactions
387 of these predicted interacting pairs.

388 BN helps to identify unique genetic or protein interactions through a completely
389 data driven approach. Several gene connections identified in this study could
390 be rationalized based on existing information from published articles while
391 some were entirely new. While the individual roles of ER stress response
392 proteins or lysosomal proteins are well established, their interplay with other
393 cancer regulators is poorly understood. Few studies have highlighted the
394 potential beneficial impact of these proteins or their in-situ dysregulation in
395 several cancers. Further focused and consolidated experimental studies would
396 be required to validate such findings as well to further understand the
397 therapeutic potential of these genes when targeted by drugs.

From	To	Strength	Direction	Remark
ABL1	ACADVL	0.88	0.76	
CSF1R	KIT	0.8	0.81	
CSF1R	CTNS	0.76	0.76	
CSF1R	CTSA	0.82	0.78	Connections present only in responding group
CSF1R	EMILIN1	0.86	0.8	
CLU	MEIS3P1	0.86	0.81	
CLU	CTTN	0.99	0.9	
ACADVL	MIR7110.PDIA5	0.8	0.79	Connections present only in non-responding group

398 **Table 4: BN Comparative analysis.** Comparative analysis of strong connections
 399 mapped from separate BN analysis of responding and non-responding
 400 population. While no common connections were identified across the non-
 401 responders and responders, seven unique connections (green text) were
 402 mapped to drug responding population compared to one unique connection (red
 403 text) coming from the non-responding group.
 404

405 4. Conclusion

406 In this work, supervised machine learning was employed on patients'
 407 transcriptomics data and 12 potential biomarkers were identified. These
 408 biomarkers were found to be associated with certain biological phenomena, such
 409 as, regulation of autophagy, topologically incorrect proteins, fatty acid biosynthesis,
 410 TNF production etc, which have shown predominant effect regulating imatinib
 411 resistance. A deep dive with BN revealed exciting findings of 7 gene-gene causal
 412 network in responders and 1 gene-gene connection in non-responders that might
 413 play an instrumental role in drug response.

414 Interestingly, this method is disease or drug system independent and can be used
 415 in several kinds of biomarkers prediction and thereby stratification, either patient
 416 level or finding new indications/ disease subtype sensitive to therapy. In this current
 417 study, the aim was to identify drug response biomarkers for which drug linked
 418 response classes have been fed into the pipeline. However, if the aim is to find
 419 disease metastatic biomarkers then only benign and metastatic level gene
 420 information will enable the pipeline to the identification of metastatic regulating
 421 biomarkers. Moreover, Bayesian network itself is an effective way to find the causal
 422 regulatory network and thereby brings forth a mechanistic insight. Together this
 423 machine learning and bn will be very powerful in analyzing the drug response
 424 nature in light of personalized medicine.

425 In this work, a very small data set has been presented as an example case, yet
426 the pipeline has identified significant genes which can be used as potential drug
427 response biomarkers. The pipeline can be easily scaled up for a large set of data
428 and is definite to produce relevant insight.

429 **5. Conflict of interest**

430 The authors declare no conflicts of interest. The work is carried out in personal laptop
431 with following specifications: Intel® Core™ i3-5005U CPU, 64-bit Windows 10 OS,
432 8GB RAM.

433 **6. Acknowledgement**

434 AD and SB want to thank Dr Saurabh Bundela for his guidance and assessment on
435 this work.

436 **7. About authors:** *SB and AD have equally contributed to this work, from hypothesis generation
437 to experiment design and execution. SB and AD are senior scientific managers at Excelra
438 knowledge solution, India, a contract research organization working in the pharma analytics. Both
439 of them have Ph.D. in bioinformatics (2013-2016) and are trained in molecular biology techniques
440 (2009-2013).*

441 **8. References:**

- 442 [1] Gambara G, Gaebler M, Keilholz U, Regenbrecht CRA, Silvestri A. From
443 Chemotherapy to Combined Targeted Therapeutics: In Vitro and in Vivo
444 Models to Decipher Intra-tumor Heterogeneity. *Front Pharmacol* 2018;9:77.
445 doi:10.3389/fphar.2018.00077.
- 446 [2] Abrosimov AY, Dvinskikh NY, Sidorin A V. Cells of Benign and Borderline
447 Thyroid Tumor Express Malignancy Markers. *Bull Exp Biol Med*
448 2016;160:698–701. doi:10.1007/s10517-016-3253-1.
- 449 [3] Deng G. Tumor-infiltrating regulatory T cells: origins and features. *Am J Clin
450 Exp Immunol* 2018;7:81–7.
- 451 [4] Deshmukh A, Binju M, Arfuso F, Newsholme P, Dharmarajan A. Role of
452 epigenetic modulation in cancer stem cell fate. *Int J Biochem Cell Biol*
453 2017;90:9–16. doi:10.1016/j.biocel.2017.07.003.
- 454 [5] Couture F, Sabbagh R, Kwiatkowska A, Desjardins R, Guay S-P, Bouchard L,
455 et al. PACE4 Undergoes an Oncogenic Alternative Splicing Switch in Cancer.
456 *Cancer Res* 2017;77:6863–79. doi:10.1158/0008-5472.CAN-17-1397.
- 457 [6] Vineis P, Fecht D. Environment, cancer and inequalities—The urgent need for
458 prevention. *Eur J Cancer* 2018;103:317–26. doi:10.1016/j.ejca.2018.04.018.

- 459 [7] Salem AA, Mackenzie GG. Pancreatic cancer: A critical review of dietary risk.
460 Nutr Res 2018;52:1–13. doi:10.1016/j.nutres.2017.12.001.
- 461 [8] Qin S, Schulte BA, Wang GY. Role of senescence induction in cancer
462 treatment. World J Clin Oncol 2018;9:180–7. doi:10.5306/wjco.v9.i8.180.
- 463 [9] Urruticoechea A, Alemany R, Balart J, Villanueva A, Viñals F, Capellá G.
464 Recent advances in cancer therapy: an overview. Curr Pharm Des 2010;16:3–
465 10.
- 466 [10] Halbur C, Choudhury N, Chen M, Kim JH, Chung EJ. siRNA-Conjugated
467 Nanoparticles to Treat Ovarian Cancer. SLAS Technol Transl Life Sci Innov
468 2019;247263031881666. doi:10.1177/2472630318816668.
- 469 [11] Wolf Y, Samuels Y. Cancer research in the era of immunogenomics. ESMO
470 Open 2018;3:e000475. doi:10.1136/esmooopen-2018-000475.
- 471 [12] Arruebo M, Vilaboa N, Sáez-Gutierrez B, Lambea J, Tres A, Valladares M, et
472 al. Assessment of the Evolution of Cancer Treatment Therapies. Cancers
473 (Basel) 2011;3:3279–330. doi:10.3390/cancers3033279.
- 474 [13] Vogenberg FR, Isaacson Barash C, Pursel M. Personalized medicine: part 1:
475 evolution and development into theranostics. P T 2010;35:560–76.
- 476 [14] Kumar A, Singh UK, Chaudhary A. Targeting autophagy to overcome drug
477 resistance in cancer therapy. Future Med Chem 2015;7:1535–42.
478 doi:10.4155/fmc.15.88.
- 479 [15] Ramos P, Bentires-Alj M. Mechanism-based cancer therapy: resistance to
480 therapy, therapy for resistance. Oncogene 2015;34:3617–26.
481 doi:10.1038/onc.2014.314.
- 482 [16] Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al.
483 Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic
484 biomarker discovery in cancer cells. Nucleic Acids Res 2012;41:D955–61.
485 doi:10.1093/nar/gks1111.
- 486 [17] Heras SC Ias, Martínez-Balibrea E. CXC family of chemokines as prognostic
487 or predictive biomarkers and possible drug targets in colorectal cancer. World
488 J Gastroenterol 2018;24:4738–49. doi:10.3748/wjg.v24.i42.4738.
- 489 [18] Kamel HFM, Al-Amodi HSAB. Exploitation of Gene Expression and Cancer
490 Biomarkers in Paving the Path to Era of Personalized Medicine. Genomics
491 Proteomics Bioinformatics 2017;15:220–35. doi:10.1016/j.gpb.2016.11.005.
- 492 [19] Biccler JL, Eloranta S, de Nully Brown P, Frederiksen H, Jerkeman M,

- 493 Jørgensen J, et al. Optimizing Outcome Prediction in Diffuse Large B-Cell
494 Lymphoma by Use of Machine Learning and Nationwide Lymphoma
495 Registries: A Nordic Lymphoma Group Study. *JCO Clin Cancer Informatics*
496 2018;1–13. doi:10.1200/CCI.18.00025.
- 497 [20] Kinnersley B, Sud A, Coker EA, Tym JE, Di Micco P, Al-Lazikani B, et al.
498 Leveraging Human Genetics to Guide Cancer Drug Development. *JCO Clin*
499 *Cancer Informatics* 2018;1–11. doi:10.1200/CCI.18.00077.
- 500 [21] Ayyad SM, Saleh AI, Labib LM. Gene expression cancer classification using
501 modified K-Nearest Neighbors technique. *Biosystems* 2019;176:41–51.
502 doi:10.1016/J.BIOSYSTEMS.2018.12.009.
- 503 [22] Crossman LC, Mori M, Hsieh YC, Lange T, Paschka P, Harrington CA, et al. In
504 chronic myeloid leukemia white cells from cytogenetic responders and non-
505 responders to imatinib have very similar gene expression signatures.
506 *Haematologica* 2005;90:459–64.
- 507 [23] Salizzato V, Borgo C, Cesaro L, Pinna LA, Donella-Deana A. Inhibition of
508 protein kinase CK2 by CX-5011 counteracts imatinib-resistance preventing
509 rpS6 phosphorylation in chronic myeloid leukaemia cells: new combined
510 therapeutic strategies. *Oncotarget* 2016;7:18204–18.
511 doi:10.18632/oncotarget.7569.
- 512 [24] Witteveen A, Nane GF, Vliegen IMH, Siesling S, IJzerman MJ. Comparison of
513 Logistic Regression and Bayesian Networks for Risk Prediction of Breast
514 Cancer Recurrence. *Med Decis Mak* 2018;38:822–33.
515 doi:10.1177/0272989X18790963.
- 516 [25] Luo Y, El Naqa I, McShan DL, Ray D, Lohse I, Matuszak MM, et al. Unraveling
517 biophysical interactions of radiation pneumonitis in non-small-cell lung cancer
518 via Bayesian network analysis. *Radiother Oncol* 2017;123:85–92.
519 doi:10.1016/j.radonc.2017.02.004.
- 520 [26] Genuer R, Poggi J-M, Tuleau-Malot C. VSURF: An R Package for Variable
521 Selection Using Random Forests. *R J* 2015;7:19–33.
- 522 [27] Breiman L. Random Forests. *Mach Learn* 2001;45:5–32.
523 doi:10.1023/A:1010933404324.
- 524 [28] Yang ZR. Biological applications of support vector machines. *Brief Bioinform*
525 2004;5:328–38.
- 526 [29] Liu Y, Hua T, Chi S, Wang H. Identification of key pathways and genes in

- 527 endometrial cancer using bioinformatics analyses. *Oncol Lett* 2018;17:897–
528 906. doi:10.3892/ol.2018.9667.
- 529 [30] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene
530 Ontology: tool for the unification of biology. *Nat Genet* 2000;25:25–9.
531 doi:10.1038/75556.
- 532 [31] Machado D, Costa RS, Rocha M, Ferreira EC, Tidor B, Rocha I. Modeling
533 formalisms in Systems Biology. *AMB Express* 2011;1:45. doi:10.1186/2191-
534 0855-1-45.
- 535 [32] Scutari M. Learning Bayesian Networks with the **bnlearn** R Package. *J Stat*
536 *Softw* 2010;35:1–22. doi:10.18637/jss.v035.i03.
- 537 [33] Clarke R, Resson HW, Wang A, Xuan J, Liu MC, Gehan EA, et al. The
538 properties of high-dimensional data spaces: implications for exploring gene
539 and protein expression data. *Nat Rev Cancer* 2008;8:37–49.
540 doi:10.1038/nrc2294.
- 541 [34] Helgason G V., Karvela M, Holyoake TL. Kill one bird with two stones:
542 potential efficacy of BCR-ABL and autophagy inhibition in CML. *Blood*
543 2011;118:2035–43. doi:10.1182/blood-2011-01-330621.
- 544 [35] Bazi A, Keramati MR, Gholamin M. Role of Oxidative Stress in Modulating
545 Unfolded Protein Response Activity in Chronic Myeloid Leukemia Cell Line.
546 *Iran Biomed J* 2016;20:63–7.
- 547 [36] Klawitter J, Kominsky DJ, Brown JL, Klawitter J, Christians U, Leibfritz D, et al.
548 Metabolic characteristics of imatinib resistance in chronic myeloid leukaemia
549 cells. *Br J Pharmacol* 2009;158:588–600. doi:10.1111/j.1476-
550 5381.2009.00345.x.
- 551 [37] Shinohara H, Kumazaki M, Minami Y, Ito Y, Sugito N, Kuranaga Y, et al.
552 Perturbation of energy metabolism by fatty-acid derivative AIC-47 and imatinib
553 in BCR-ABL-harboring leukemic cells. *Cancer Lett* 2016;371:1–11.
554 doi:10.1016/j.canlet.2015.11.020.
- 555 [38] Kikuchi S, Nagai T, Kunitama M, Kirito K, Ozawa K, Komatsu N. Active
556 FKHL1 overcomes imatinib resistance in chronic myelogenous leukemia-
557 derived cell lines via the production of tumor necrosis factor-related apoptosis-
558 inducing ligand. *Cancer Sci* 2007;98:1949–58. doi:10.1111/j.1349-
559 7006.2007.00623.x.
- 560 [39] Nestal de Moraes G, Souza PS, Costas FC de F, Vasconcelos FC, Reis FRS,

- 561 Maia RC. The Interface between BCR-ABL-Dependent and -Independent
562 Resistance Signaling Pathways in Chronic Myeloid Leukemia. *Leuk Res*
563 *Treatment* 2012;2012:1–19. doi:10.1155/2012/671702.
- 564 [40] Cook KL, Soto-Pantoja DR, Clarke PAG, Cruz MI, Zwart A, Wärrri A, et al.
565 Endoplasmic Reticulum Stress Protein GRP78 Modulates Lipid Metabolism to
566 Control Drug Sensitivity and Antitumor Immunity in Breast Cancer. *Cancer Res*
567 2016;76:5657–70. doi:10.1158/0008-5472.CAN-15-2616.
- 568 [41] Higa A, Taouji S, Lhomond S, Jensen D, Fernandez-Zapico ME, Simpson JC,
569 et al. Endoplasmic Reticulum Stress-Activated Transcription Factor
570 ATF6 Requires the Disulfide Isomerase PDIA5 To Modulate
571 Chemoresistance. *Mol Cell Biol* 2014;34:1839–49. doi:10.1128/MCB.01484-
572 13.
- 573 [42] Phan ANH, Vo VTA, Hua TNM, Kim M-K, Jo S-Y, Choi J-W, et al. PPAR γ
574 sumoylation-mediated lipid accumulation in lung cancer. *Oncotarget*
575 2017;8:82491–505. doi:10.18632/oncotarget.19700.
- 576 [43] Lee E, Lee DH. Emerging roles of protein disulfide isomerase in cancer. *BMB*
577 *Rep* 2017;50:401–10. doi:10.5483/BMBREP.2017.50.8.107.
- 578 [44] So J-S. Roles of Endoplasmic Reticulum Stress in Immune Responses. *Mol*
579 *Cells* 2018;41:705–16. doi:10.14348/molcells.2018.0241.
- 580 [45] Zhang X, Liu K, Zhang T, Wang Z, Qin X, Jing X, et al. Cortactin promotes
581 colorectal cancer cell proliferation by activating the EGFR-MAPK pathway.
582 *Oncotarget* 2017;8:1541–54. doi:10.18632/oncotarget.13652.
- 583 [46] Koltai T. Clusterin: a key player in cancer chemoresistance and its inhibition.
584 *Onco Targets Ther* 2014;7:447–56. doi:10.2147/OTT.S58622.
- 585 [47] Jin R, Chen X, Han D, Luo X, Li H. Clusterin modulates transdifferentiation of
586 non-small-cell lung cancer. *BMC Cancer* 2017;17:661. doi:10.1186/s12885-
587 017-3649-y.
- 588 [48] Guarino M. Src signaling in cancer invasion. *J Cell Physiol* 2009;223:n/a-n/a.
589 doi:10.1002/jcp.22011.
- 590 [49] Stewart PA, Luks J, Roycik MD, Sang Q-XA, Zhang J. Differentially Expressed
591 Transcripts and Dysregulated Signaling Pathways and Networks in African
592 American Breast Cancer. *PLoS One* 2013;8:e82460.
593 doi:10.1371/journal.pone.0082460.
- 594 [50] Tang S, Gao L, Bi Q, Xu G, Wang S, Zhao G, et al. SDR9C7 Promotes Lymph

- 595 Node Metastases in Patients with Esophageal Squamous Cell Carcinoma.
596 PLoS One 2013;8:e52184. doi:10.1371/journal.pone.0052184.
- 597 [51] Giulietti M, Occhipinti G, Righetti A, Bracci M, Conti A, Ruzzo A, et al.
598 Emerging Biomarkers in Bladder Cancer Identified by Network Analysis of
599 Transcriptomic Data. Front Oncol 2018;8:450. doi:10.3389/fonc.2018.00450.
- 600 [52] Champion M, Chiquet J, Neuvial P, Elati M, Birmelé E. Identification of
601 deregulated transcription factors involved in subtypes of cancers 2017.
- 602 [53] Cannarile MA, Weisser M, Jacob W, Jegg A-M, Ries CH, Rüttinger D. Colony-
603 stimulating factor 1 receptor (CSF1R) inhibitors in cancer therapy. J
604 Immunother Cancer 2017;5:53. doi:10.1186/s40425-017-0257-y.
- 605 [54] Circu M, Cardelli J, Barr M, O'Byrne K, Mills G, El-Osta H. Modulating
606 lysosomal function through lysosome membrane permeabilization or
607 autophagy suppression restores sensitivity to cisplatin in refractory non-small-
608 cell lung cancer cells. PLoS One 2017;12:e0184922.
609 doi:10.1371/journal.pone.0184922.
- 610 [55] Modica TME, Maiorani O, Sartori G, Pivetta E, Doliana R, Capuano A, et al.
611 The extracellular matrix protein EMILIN1 silences the RAS-ERK pathway via
612 $\alpha 5 \beta 1$ integrin and decreases tumor cell growth.
613 Oncotarget 2017;8:27034–46. doi:10.18632/oncotarget.15067.
- 614 [56] Danussi C, Petrucco A, Wassermann B, Modica TME, Pivetta E, Belluz LDB,
615 et al. An EMILIN1-Negative Microenvironment Promotes Tumor Cell
616 Proliferation and Lymph Node Invasion. Cancer Prev Res 2012;5:1131–43.
617 doi:10.1158/1940-6207.CAPR-12-0076-T.
618