1 # Subnetwork-based prognostic biomarkers exhibit

2 # performance and robustness superior to gene-based

3 # biomarkers in breast cancer

4 Michal R. Grzadkowski[1], Syed Haider[1,2], Dorota H. Sendorek[1] and Paul C. Boutros[1,3,4,§]

5 [1] Ontario Institute for Cancer Research, Toronto, Ontario, Canada M5G 0A3

6 [2] Centre for Molecular Oncology, Barts Cancer Institute, London EC1M 6BQ, United Kingdom

7 [3] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada M5G

8 2M9

9 [4] Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario,

10 Canada M5S 1A8

11 [§]Corresponding author: paul.boutros@oicr.on.ca; 661 University Avenue, Suite 510; Toronto,

12 Ontario, Canada; M5G 0A3; 416-673-8564

13 **Author Emails:**

14 Michal R. Grzadkowski   mgrzad@gmail.com

15 Syed Haider   syed.haider@oicr.on.ca

16 Dorota H. Sendorek   dorota.sendorek@oicr.on.ca

17 Paul C. Boutros   paul.boutros@oicr.on.ca

18 **Keywords:** prognostic signatures, subnetwork biomarker, gene biomarker, breast cancer

Abstract

Background

Effective classification of cancer patients into groups with differential survival remains an important and unsolved challenge. Biomarkers have been developed based on mRNA abundance data, but their replicability and clinical utility is modest. Integrating functional information, such as pathway data, has been suggested to improve biomarker performance. To date, however, the advantages of subnetwork-based biomarkers have not been quantified.

Results

We deeply sampled the population of prognostic gene-based and subnetwork-based biomarkers in a breast cancer meta-dataset of 4,960 patients. Analysing the performance and robustness of 22,000,000 gene biomarkers and 6,250,000 subnetwork biomarkers across twenty different training:testing cohort partitions of the meta-dataset revealed that subnetwork biomarkers exhibit superior overall performance and higher concordance across partitions. We find evidence of an upper bound for optimal biomarker size of ~200 genes or ~100 subnetworks. Additionally, with both biomarker feature types, larger biomarkers tend to show less consistency in performance across partitions, suggestive of over-fitting. Finally, an evaluation of varying training cohort sizes quantifies the effects of training cohort size.

Conclusions

Many groups are developing techniques for exploiting network-based representations of biological pathways to characterize cancer and other diseases. By considering the distribution of gene- and subnetwork-based biomarkers, we show that pathway data improves

40 performance and replicability, and that smaller biomarkers are more robust across patient

41 cohorts. These insights may facilitate development of clinically useful biomarkers.

42

43 # Background

44 Breast cancer continues to be the leading cause of neoplastic incidence and mortality in

45 women worldwide [1]□. A major obstacle to the effective diagnosis and treatment of breast

46 tumours is their molecular heterogeneity, which contributes to widely divergent outcomes [2–

47 4]□. Techniques for accurate classification of breast cancer patients by risk profile or

48 therapeutic response are therefore urgently needed as they enable more efficient allocation of

49 clinical resources. A wide variety of models utilizing the mRNA abundance scores of a select

50 set of genes, commonly referred to as "prognostic biomarkers", have been developed for this

51 purpose in breast cancer [5–11]□ as well as other tumour types and diseases [12–16]□.

52 Although several prognostic biomarkers have proven sufficiently effective to be successfully

53 commercialized [5,6]□, there is no consensus on the best procedure for identifying useful and

54 robust biomarkers. Furthermore, the persistent difficulty in replicating the reported

55 performance of published prognostic biomarkers has led some to question the value of this

56 approach [17–20]□. A proposed method for improving the efficacy and robustness of

57 biomarkers has been to leverage regulatory pathway information. Traditionally, a prognostic

58 biomarker comprises a set of genes whose individual expression levels serve as input for a

59 scoring algorithm that computes the individual risk levels for a patient sample cohort.

60 However, more recently introduced biomarkers consist of *gene subnetworks*, with each

61 subnetwork being a set of genes identified *a priori* as sharing a common functionality via a

62 separate dataset [21–26].

63 Subnetworks are most commonly derived from protein-protein interaction (PPI) networks

64 since protein network data are relatively easy to obtain and contain information on gene-

65 product interactions that it is believed cannot be inferred from mRNA abundance levels alone

66 [27]. A s*ubnetwork biomarker* thus differs from a *gene biomarker* in that the expression

67 levels of the genes it uses are first aggregated into a set of subnetwork expression scores

68 before being fed into the risk scoring algorithm. We recently introduced a novel scoring

69 system for subnetwork biomarkers constructed using PPI data [28]. This "SIMMS node-only"

70 model (hereafter referred to simply as *SIMMS*) first calculates the dependence of patient

71 survival on the mRNA abundance of each gene to get a gene risk factor. It then aggregates

72 the gene-level mRNA abundance values within a subnetwork for each patient by taking their

73 sum weighed by the gene risk factors. Several other subnetwork-based schemes have been

74 proposed, and research in this area is active [29–31].

75 Several studies have cast doubt on the value of using multi-gene models compared to simpler

76 univariate models by citing the performance of subnetwork biomarkers relative to gene

77 biomarkers [32, 33]. This raises the question of whether time and resources increasingly

78 designated for the development of novel feature selection algorithms as well as the

79 implementation of increasingly composite data for biomarker discovery would be best

80 reserved for research using a more traditional approach of simply identifying patient outcome-

81 related gene sets (*i.e.* gene biomarkers). If building increasingly complex models using

82 advanced machine learning methods and interaction network data yields only marginal

83 improvements in biomarker performance, then identifying informative groups of genes is

84 clearly a more efficient approach to generating clinically relevant prognostic tools.

85 To resolve this issue, we examined whether using subnetworks as the building blocks of

86 prognostic biomarkers confers any advantage over orthodox gene-based approaches. The

87 SIMMS model was generalized to score gene biomarkers as well as subnetwork biomarkers

88 and then used to evaluate and compare the performance of 22,000,000 gene biomarkers and

89 6,250,000 subnetwork biomarkers. This experiment was performed with the same set of

90 biomarkers on twenty different permutations of training and testing patient cohorts to test

91 biomarker robustness. Our systematic evaluation of the behaviour of subnetwork biomarkers

92 and gene biomarkers is an important contribution to determining the worth of network-based

93 approaches in biomarker discovery.

94

Results and discussion

Using SIMMS to measure the performance of gene and subnetwork biomarkers

97 The prognostic performance of subnetwork biomarkers was evaluated on a 4,960 breast

98 cancer patient meta-dataset using the SIMMS model as fully detailed in Methods [28]□.

99 Briefly, a univariate Cox proportional hazards model is used to estimate the association

100 between mRNA abundance levels and survival information of the training cohort for each

101 gene separately. Next, for each patient, the per-gene coefficients from the model are

102 aggregated into subnetwork risk scores. Then, for each subnetwork biomarker, a multivariate

103 Cox model is fit to the patient risk scores of the subnetworks comprising the biomarker.

104 Finally, the subnetwork coefficients of this model are used to generate patient risk scores for

105 an independent testing cohort by combining them with the mRNA abundance values. A

106 univariate Cox model is fit to the testing cohort (dichotomized by the training cohort's median

107 risk score) and the resulting hazard ratio is used to assess the biomarker's prognostic

108 capability.

109 In order to evaluate the prognostic capability of a given gene biomarker in a manner that is

110 directly comparable to subnetwork biomarker performance, we modified the SIMMS model so

111 that it handles individual genes as subnetworks (*i.e.* each gene is treated as a subnetwork of

112 size one). We refer to this approach as *geneSIMMS*.

Jackknifing reveals patterns in gene and subnetwork biomarker performance

114 To study the characteristics of the population of gene biomarkers and subnetwork biomarkers,

115 we randomly sampled 22,000,000 gene biomarkers and 6,250,000 subnetwork biomarkers by

116 utilizing *jackknifing*, a resampling method in which samples are drawn from a population

117 without replacement [34]□. For this purpose, we only used genes for which mRNA abundance

118 scores were available in all eighteen datasets comprising our breast cancer meta-dataset

119 (**Table 1**). To ensure that our set of gene biomarkers contained the same underlying mRNA

120 abundance data as our set of subnetwork biomarkers, all gene biomarkers were drawn from a

121 pool of the 1,500 genes which were included in at least one of the 500 subnetworks from

122 which the subnetwork biomarkers were created. Since gene biomarkers were drawn from a

123 set of 1,500 features while subnetwork biomarkers were drawn from a set of 500 features, the

124 population of all possible gene biomarkers is substantially larger than that of all possible

125 subnetwork biomarkers. This necessitated a larger test set of gene biomarkers to fully capture

126 the null distribution of gene biomarker performance. The prognostic capability of each gene

127 and subnetwork biomarker was estimated using SIMMS on twenty different partitions of the

128 breast cancer meta-dataset, each of which was composed of varying training and testing

129 cohort sizes (**Tables 2-3**). For full details on how the meta-dataset was partitioned, see

130 Methods.

131 Considering only a single meta-dataset partition of equally-sized training and testing cohorts

132 (the 'default' partition in **Table 2**), we found that the performance of gene and subnetwork

133 biomarkers is highly sensitive to biomarker size (**Figure 1**). Adding more features to gene

134 biomarkers results in substantial improvements in performance up to ~100 genes, at which

135 point improvement in performance starts levelling off before peaking at ~175 genes. After this,

136 larger gene biomarkers show a sharp and consistent decline in prognostic capability,

137 particularly after biomarker size exceeds ~200 features. On the other hand, the performance

138 of subnetwork biomarkers improves much more gradually with increasing biomarker size,

139 plateauing at around 100 subnetworks.

140 **Comparison of gene and subnetwork biomarkers across gene counts**

141 To better contrast the performance of gene and subnetwork biomarkers, we first normalized

142 the sizes of our subnetwork biomarkers to make them more directly comparable to our gene

143 biomarkers. Since a subnetwork biomarker is essentially a set of gene lists, the most

144 straightforward way to perform this normalization is to simply sum the sizes of all these gene

145 lists to get the total number of genes in the biomarker. However, there is gene content overlap

146 in our set of 500 subnetworks so a single subnetwork biomarker may include multiple

147 subnetworks that contain the same gene. Therefore, we additionally considered the number of

148 unique genes in each biomarker. At each gene count, we compared the 95th to 99th

149 percentiles of subnetwork biomarker performance against the 95th to 99th percentiles of gene

150 biomarker performance (**Figure 2**).

151 The difference in performance between the two gene count normalization techniques is

152 negligible as both gene and subnetwork biomarkers exhibit similar peak performance.

153 Although gene biomarkers require fewer genes to reach optimal prognostic efficiency, they

154 also show a higher propensity for over-fitting at higher gene counts.

Randomly sampling training:testing meta-dataset partitions reveals effect of training cohort size on biomarker performance and stability

157 To test the robustness of our finding that biomarker size affects performance, we re-evaluated

158 our biomarker sets on a partition of the meta-dataset with a substantially smaller training

159 cohort at only 36.5% of the total patient number. As seen in **Additional file 1:**

160 **Supplementary Figure 1**, the general characteristics of biomarker performance with regard

161 to biomarker size remain the same even though the performance of both gene and

162 subnetwork biomarkers decreases slightly overall in this setting. However, subnetwork

163 biomarkers do reach higher peak performance than gene biomarkers, suggesting that they

164 are less sensitive to a suboptimal training environment. To verify these observations, we went

165 on to create eighteen more meta-dataset partitions using random sampling for a total of ten

166 partitions with larger training cohorts and ten partitions with smaller training cohorts.

167 As first evidenced by our initial finding, both gene and subnetwork biomarkers perform

168 routinely better when trained on the larger patient cohorts (**Figure 3**). In only one dataset

169 partition containing a smaller training cohort do biomarkers consistently outperform those

170 trained on the larger cohorts. For most of the additional eighteen dataset partitions, the

171 performance patterns resemble our initial two partitions. This is particularly true for gene

172 biomarker performance which peaks rapidly at smaller biomarker sizes then drops with

173 increasing biomarker size. Subnetwork biomarker performance peaks slowly, as seen

174 previously, and then either plateaus or undergoes a more gradual decline.

175 To quantify biomarker performance stability across partitions we used Lin's concordance

176 correlation coefficient (CCC) [35, 36]□. The output of this metric ranges from -1, indicating

177 complete discordance in biomarker performance across all partitions, to 1, indicating identical

178 biomarker performance across all partitions. The stability of biomarker performance at

179 individual biomarker sizes is only slightly greater in partitions with larger training cohorts

180 compared to partitions with smaller training cohorts in gene and subnetwork biomarkers.

181 However, when stability is measured using all biomarkers simultaneously this effect gets

182 amplified (**Additional file 2: Supplementary Figure 2**). More significantly, both classes of

183 biomarkers show the highest performance stability at smaller biomarker sizes, with steep

184 declines in stability as more features are included. This demonstrates that although larger

185 biomarkers may exhibit better prognostic performance, the superiority of individual large

186 biomarkers is much harder to replicate. Therefore, partitioning a dataset to have fewer

187 patients in the training cohort than in the testing cohort has a clear negative effect on both

188 biomarker performance and stability. However, when only small biomarkers are considered,

189 the difference in stability is quite trivial. Since the stability of small gene and subnetwork

190 biomarkers is very high, these biomarkers can be relied upon to have consistent performance

191 when trained on patient cohorts of varying sizes.

192 Comparing subnetwork biomarkers to gene biomarkers reveals advantages of

193 subnetworks as features

194 Finding a link between biomarker performance and feature size prompted us to explore the

195 performance of gene and subnetwork biomarkers in more detail. A clearly defined optimal size

196 was observed for gene biomarkers in all twenty meta-dataset permutations and we used this

197 size as the standard against which subnetwork biomarkers would be judged. Subnetwork

198 biomarker performance was normalized against peak gene biomarker performance in all

199 twenty dataset permutations (**Figure 4**). This reveals that optimal subnetwork biomarker

200 performance tends to be greater than optimal gene biomarker performance, an effect that is

201 particularly pronounced in partitions with small training cohorts. Furthermore, in the majority of

202 partitions where the performance of subnetwork biomarkers exceeds that of optimally-sized

203 gene biomarkers, it does so at a wide range of feature and gene sizes (**Figure 4** and

204 **Additional file 3: Supplementary Figure 3**). In particular, when considering only these

205 partitions, the approximate biomarker sizes at which subnetwork biomarkers start to

206 outperform optimal gene biomarkers is never greater than 50 features, 300 unique genes or

207 400 total genes.

208 Although the stability of gene biomarker performance is significantly higher than the stability of

209 subnetwork biomarker performance when considering biomarkers of all feature sizes, at

210 comparable gene counts this relationship is inverted (**Figure 5**). The instability of subnetwork

211 biomarkers that contain many large subnetworks obscures the fact that subnetwork

212 biomarkers consisting of a small or moderate number of modules perform relatively

213 consistently across different training:testing cohort partitions. As discussed above, subnetwork

214 biomarkers tend to outperform even optimal gene biomarkers at fairly small sizes.

215 Subnetwork biomarkers without an exorbitant number of features are, therefore, not only

216 more stable than gene biomarkers but also have greater prognostic capabilities.

217

Conclusions

219 In this study, we applied jackknifing to a 4,960 breast cancer patient meta-dataset and

220 identified two key advantages that subnetwork biomarkers have over gene biomarkers. In the

221 majority of our randomly sampled partitions of training and testing patient cohorts, subnetwork

222 biomarkers of various sizes outperform optimally-sized gene biomarkers. Additionally,

223 subnetwork biomarkers show more consistency in their performance across different meta-

224 dataset partitions than gene biomarkers of equal biomarker size (*i.e.* biomarkers containing

225 the same number of features). In other words, a randomly selected subnetwork biomarker can

226 be expected to do a better job of predicting high-risk patients repeatedly across numerous

227 patient cohorts than a randomly selected gene biomarker of comparable size, provided that

228 the number of genes comprising the biomarker does not exceed a relatively high threshold.

229 The breadth of our analysis also allows us to make several other important observations

230 about prognostic biomarker performance and robustness. Small training cohort sizes have a

231 greater effect on overall biomarker performance than on the consistency of said performance.

232 Furthermore, the performance of large biomarkers is substantially more unstable than that of

233 small biomarkers. Therefore, to optimize biomarker development, it is critical to maximize

234 training cohort size and to minimize biomarker size.

235 Nevertheless, we recognize there are several caveats to be made regarding our findings that

236 subnetwork biomarkers are superior to gene biomarkers. Because generally gene biomarkers

237 will contain fewer genes than subnetwork biomarkers, and biomarkers with fewer genes are

238 more robust, it is easier to find gene biomarkers with consistently high performance especially

239 since gene biomarker performance peaks at a smaller number of features. Moreover, our

240 study considers only the null distribution of biomarker performance. It is thus possible, albeit

241 highly unlikely, that among the "rare" biomarkers, whose performance falls well above that of

242 those that are randomly sampled, gene biomarkers are preferable to subnetwork biomarkers.

243 Finally, there is a potential limitation in considering only 500 subnetworks and, in particular,

244 only 1500 genes as these may not necessarily be representative of the actual human genome

245 and the approximately 20,000 genes it contains. As information on PPI and gene pathways

246 continues to accumulate, this matter can be thoroughly investigated in the future with access

247 to a more complete and biologically accurate set of data.

248 Finally, deeply sampling the total biomarker space allows for a comprehensive evaluation of

249 biomarker characteristics that is not attainable through analyses on a smaller number of

250 biomarkers acquired using unsubstantiated feature selection strategies. Since experiment

251 replicability has proven to be a great challenge in the development of clinically useful

252 biomarkers, it is vital to study the performance of biomarkers in a variety of settings,

253 especially using different training and testing cohorts. Jackknifing and meta-dataset

254 permutation are, therefore, two key tools for prognostic biomarker research and can be

255 utilized to great effect in studies hindered by limited sample sizes and data access.

256

Methods

257

Preparation of mRNA abundance and pathway data

258

259 Datasets and subnetworks were prepared as described by Haider *et al* [28]*.* Raw breast

260 cancer mRNA abundance datasets (**Table 1**) were normalized using the RMA algorithm [37]☐

261 (R package affy v1.28.0) and probes were mapped to Entrez genes using custom CDFs (R

262 packages hgu133ahsentrezgcdf v12.1.0, hgu133bhsentrezgcdf v12.1.0,

263 hgu133plus2hsentrezgcdf v12.1.0, hthgu133ahsentrezgcdf v12.1.0 and

264 hgu95av2hsentrezgcdf v12.1.0). Subnetworks were drawn from the NCI-Nature and

265 BioCarta/Reactome databases and screened such that there was a maximum overlap of 80%

266 between any two subnetworks and each subnetwork consisted of at least three genes. Genes

267 without mRNA abundance values available in all eighteen datasets were removed along with

268 genes that did not appear in any of the subnetworks. This resulted in a pool of 500

269 subnetworks and 1500 genes.

SIMMS and geneSIMMS

270

271 The original SIMMS node-only model trains a set of subnetworks on a cohort of patients to

272 obtain a "module dysregulation score" (MDS) for each subnetwork. For each gene g, the

273   training cohort is median dichotomized by the mRNA abundance value of *g*. A univariate Cox

274   proportional hazards model is then fit to the dichotomization by gene *g* using training cohort

275   survival information and a hazard ratio HR_g is obtained (R package survival v2.36.14). A risk

276   score for each combination of patient *P* and subnetwork *S* is computed using the formula

$$risk_{S,P} = \sum_{g \in S} \log_2(HR_g) \times mRNA_{g,P} \tag{1}$$

277   where $mRNA_{g,P}$ is the mRNA abundance in patient *P* of gene *g* contained within subnetwork

278   *S*. Next, for each subnetwork biomarker *B*, a multivariate Cox proportional hazards model is fit

279   to its subnetworks' training cohort's risk scores, as calculated in **(1)**, and the training cohort

280   survival data to get a Cox beta β_S for each subnetwork *S* in the biomarker. For each patient *P*,

281   we thus calculate a risk score as measured by *B* using the formula

$$\sum_{s \in B} \beta_s \times risk_{s,P} \tag{2}$$

282   The median risk score is calculated for the training cohort and is then used as the break-point

283   at which to dichotomize the testing cohort. A univariate Cox model is fit to this testing cohort

284   and the risk classification is assessed by the resulting hazard ratio.

285   The SIMMS model was extended to gene biomarkers by treating each gene as a subnetwork

286   of size one. We refer to this as the geneSIMMS model. The risk score for patient *P* and gene

287   *g* is thus given by

$$risk_{g,P} = \log_2(HR_g) \times mRNA_{g,P} \tag{3}$$

288   with HR_g and $mRNA_{g,P}$ defined as in **(1)**. For a gene biomarker *B*, a Cox beta β_g is calculated

289   using a multivariate model for each gene *g* in the biomarker (as is done with the subnetworks

290   in a subnetwork biomarker). For each patient *P*, we thus calculate a risk score as measured

291   by *B* using the formula

$$\sum_{g\_B} \beta_g \times risk_{g,P} \tag{4}$$

292

293   This is followed by patient dichotomization of the testing cohort and then biomarker

294   performance evaluation, same as described above for SIMMS.

Biomarker jackknifing

296   Jackknifing is a resampling method in which samples are drawn from a population without

297   replacement [34]☐. This technique was applied to estimate the characteristics of the gene and

298   subnetwork biomarker populations. Twenty-five subnetwork biomarker sizes were chosen (all

299   multiples of five between 5 and 100 inclusive and all multiples of ten between 110 and 150

300   inclusive) along with forty-four gene biomarker sizes (all multiples of five between 5 and 150

301   inclusive, all multiples of twenty-five between 175 and 400 inclusive and all multiples of one

302   hundred between 500 and 800 inclusive). 250,000 biomarkers were jackknifed at each

303   subnetwork biomarker size for a total of 6,250,000 biomarkers and 500,000 biomarkers were

304   jackknifed at each gene biomarker size for a total of 22,000,000 biomarkers.

305   The jackknifing was performed to sample more gene biomarkers than subnetwork biomarkers

306   since the pool of potential gene biomarkers is much larger due to there being a greater

307   number of gene features (n = 1500) than subnetwork features (n = 500). Because of the large

308   computational expense of using signatures with a high feature number, gene biomarkers

309   comprising 500 or more genes were only tested on the two initial (default) meta-dataset

310   partitions.

Normalizing subnetwork biomarker size

A subnetwork biomarker S of size n consists of a set of modules $M_1, M_2, ... M_n$, each of which is a list of genes. The total number of genes in S is therefore

$$\sum_{i=1}^{n} |M_i| \tag{5}$$

i.e. the sum of the sizes of the modules comprising S. The number of unique genes in S is

$$\left| \bigcup_{i=1}^{n} M_i \right| \tag{6}$$

i.e. the size of the set of genes appearing in at least one of the modules comprising S. A gene biomarker G of size n consists of a set of n unique genes, so the total number of genes and the number of unique genes in G both equal n.

Meta-dataset partitioning and concordance correlation calculations

As seen in **Table 1**, our breast cancer meta-dataset consists of eighteen different datasets. To create a partition of this meta-dataset into training and testing patient cohorts, we randomly chose a subset of the eighteen datasets to use for training such that the total number of patients in this subset satisfied some bound and the remaining datasets served as the testing cohort. For partitions with large training cohorts, this bound was at least 47% of the patients in the meta-dataset (**Table 2**). For partitions with small training cohorts this bound was between 33% and 40% of the patients in the meta-dataset (**Table 3**). To ensure random sampling over all partitions satisfying a given bound, we first exhaustively identified all possible partitions and then sampled from that set.

The concordance correlation coefficient for multiple classes was used as first defined and

329 later amended by Lin [35, 36]. Each dataset partition was treated as a class (or observer) in

330 our analysis. Therefore, all gene biomarkers containing 400 or fewer genes and all

331 subnetwork biomarkers had twenty matched observations of their performance as measured

332 by the raw hazard ratio returned by SIMMS or geneSIMMS.

333 **Comparing subnetwork biomarker performance to peak gene biomarker performance**

334 To determine peak gene biomarker performance on a given meta-dataset partition, we

335 considered the 99th percentile of biomarker performance at each biomarker size as measured

336 by the log-transformed Cox model hazard ratio. The biomarker size which had the highest

337 such percentile of performance defined the optimal size and the corresponding 99th

338 percentile was set as the peak gene biomarker performance for the partition. To compare

339 subnetwork biomarker performance at a given size to this peak, we again took the 99th

340 percentile of performance as measured by the log-transformed hazard ratio and calculated

341 the difference of subnetwork biomarker performance and peak gene biomarker performance

342 to obtain a "relative" performance.

343 **Visualizations**

344 All figures were created in the R statistical environment (v2.15.3) using the lattice (v0.20-16),

345 latticeExtra (v0.6-24), RColorBrewer (v1.0-5), BPG (v5.6.8) [38] and cluster (v1.14.3)

346 packages.

347 **List of abbreviations used**

348 CCC: Lin's concordance correlation coefficient

349 PPI: protein-protein interaction

350 SIMMS: subnetwork integration for multi-modal signatures

351

Declarations

353 **Ethics approval and consent to participate**

354 Not applicable

355 **Consent for publication**

356 Not applicable

357 **Availability of data and materials**

358 The datasets supporting the conclusions of this article are included within the article and its

359 additional files, as well as obtained from the ArrayExpress, Gene Expression Omnibus (GEO)

360 and European Genome-phenome Archive (EGA) repositories:

361 - Bild: https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-3143

362 - Chin: https://www.ebi.ac.uk/arrayexpress/experiments/E-TABM-158

363 - Desmedt-1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7390

364 - Desmedt-2: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16446

365 - Hatzis-1: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25055

366 - Hatzis-2: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25065

367 - Ivshina: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4922

368 - Miller: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3494

369 - Pawitan: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1456

370 • Sabatier: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21653

371 • Schmidt: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11121

372 • Sotiriou: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6532

373 • Symmans-JBI: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17700

374 • Symmans-MDA: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17705

375 • Wang: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2034

376 • Zhang: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12093

377 • Metabric-Training: https://www.ebi.ac.uk/ega/datasets/EGAD00010000210

378 • Metabric-Validation: https://www.ebi.ac.uk/ega/datasets/EGAD00010000211

379 **Competing interests**

380 All authors declare that they have no competing interests.

381 **Funding**

385 **Authors' contributions**

386 Initiated the project: MRG, PCB

387 Developed the SIMMS model: SH

388 Extended the SIMMS model to gene biomarkers: MRG

389 Performed biomarker jackknifing and meta-dataset permutation analysis: MRG

390 Wrote the first draft of the manuscript: MRG, DHS

Supervised research: PCB

Acknowledgements

References

1. Stewart BW, Wild CP. World Cancer Report 2014. Lyon: 2014.

2. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747–52.

3. Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, Richesson DA, Bojesen SE, Nordestgaard BG, Axelsson CK, Arias JI, Milne RL, Ribas G, González-Neira A, Benítez J, Zamora P, Brauch H, Justenhoven C, Hamann U, Ko YD, Bruening T, Haas S, Dörk T, Schürmann P, Hillemanns P, Bogdanova N, Bremer M, Karstens JH, Fagerholm R, Aaltonen K, Aittomäki K, von Smitten K, Blomqvist C, Mannermaa A, Uusitupa M, Eskelinen M, Tengström M, Kosma VM, Kataja V, Chenevix-Trench G, Spurdle AB, Beesley J, Chen X; Australian Ovarian Cancer Management Group; Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer, Devilee P, van Asperen CJ, Jacobi CE, Tollenaar RA, Huijts PE, Klijn JG, Chang-Claude J, Kropp S, Slanger T, Flesch-Janys D, Mutschelknauss E, Salazar R, Wang-

Gohrke S, Couch F, Goode EL, Olson JE, Vachon C, Fredericksen ZS, Giles GG, Baglietto L, Severi G, Hopper JL, English DR, Southey MC, Haiman CA, Henderson BE, Kolonel LN, Le Marchand L, Stram DO, Hunter DJ, Hankinson SE, Cox DG, Tamimi R, Kraft P, Sherman ME, Chanock SJ, Lissowska J, Brinton LA, Peplonska B, Klijn JG, Hooning MJ, Meijers-Heijboer H, Collee JM, van den Ouweland A, Uitterlinden AG, Liu J, Lin LY, Yuqing L, Humphreys K, Czene K, Cox A, Balasubramanian SP, Cross SS, Reed MW, Blows F, Driver K, Dunning A, Tyrer J, Ponder BA, Sangrajrang S, Brennan P, McKay J, Odefrey F, Gabrieau V, Sigurdson A, Doody M, Struewing JP, Alexander B, Easton DF, Pharoah PD. Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS Genet. 2008;4(4):e1000054.

4. Voduc KD, Cheang MC, Tyldesley S, Gelmon K, Nielsen TO, Kennecke H. Breast cancer subtypes and the risk of local and regional relapse. J Clin Oncol. 2010;28(10):1684–91.

5. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;15:530–6.

6. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med.

432   2004;351(27):2817–26.

433   7. Ma Y, Qian Y, Wei L, Abraham J, Shi X, Castranova V, Harner EJ, Flynn DC, Guo L.
434   Population-based molecular prognosis of breast cancer by transcriptional profiling. Clin
435   Cancer Res. 2007;13(7):2014–22.

436   8. Naderi A, Teschendorff A, Barbosa-Morais NL, Pinder SE, Green AR, Powe DG,
437   Robertson JF, Aparicio S, Ellis IO, Brenton JD, Caldas C. A gene-expression signature
438   to predict survival in breast cancer across independent data sets. Oncogene.
439   2007;26(10):1507–16.

440   9. Wan Y-W, Qian Y, Rathnagiriswaran S, Castranova V, Guo NL. A breast cancer
441   prognostic signature predicts clinical outcomes in multiple tumor types. Oncol Rep.
442   2010;24(2):489–94.

443   10. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N Engl J Med.
444   2009;360(8):790–800.

445   11. Duan Q, Kou Y, Clark NR, Gordonov S, Ma'ayan A. Metasignatures identify two major
446   subtypes of breast cancer. CPT Pharmacometrics Syst Pharmacol. 2013;2:e35.

447   12. Sun Z, Wigle DA, Yang P. Non-overlapping and non-cell-type-specific gene expression
448   signatures predict lung cancer survival. J Clin Oncol. 2008;26(6):877–83.

449   13. Boutros PC, Lau SK, Pintilie M, Liu N, Shepherd FA, Der SD, Tsao M-S, Penn LZ,
450   Jurisica I. Prognostic gene signatures for non-small-cell lung cancer. Proc Natl Acad
451   Sci U S A. 2009;106(8):2824–8.

14. Sabatier R, Finetti P, Bonensea J, Jacquemier J, Adelaide J, Lambaudie E, Viens P, Birnbaum D, Bertucci F. A seven-gene prognostic model for platinum-treated ovarian carcinomas. Br J Cancer. 2011;105(2):304–11.

15. Hoshida Y, Villanueva A, Sangiovanni A, Sole M, Hur C, Andersson KL, Chung RT, Gould J, Kojima K, Gupta S, Taylor B, Crenshaw A, Gabriel S, Minguez B, Iavarone M, Friedman SL, Colombo M, Llovet JM, Golub TR. Prognostic gene expression signature for patients with hepatitis C-related early-stage cirrhosis. Gastroenterology. 2013;144(5):1024–30.

16. Glatt SJ, Stone WS, Nossova N, Liew C-C, Seidman LJ, Tsuang MT. Similarities and differences in peripheral blood gene expression signatures of individuals with schizophrenia and their first-degree biological relatives. Am J Med Genet B Neuropsychiatr Genet. 2011;156B(8):869–87.

17. Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A. Biomarkers and surrogate end points--the challenge of statistical validation. Nat Rev Clin Oncol. 2010;7(6):309–17.

18. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst. 2010;102(7):464–74.

19. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A. 2006;103(15):5923–8.

472   20. Chibon F. Cancer gene expression signatures - the rise and fall? Eur J Cancer.
473   2013;49(8):2000–9.

474   21. Dao P, Colak R, Salari R, Moser F, Davicioni E, Schönhuth A, Ester M. Inferring cancer
475   subnetwork   markers   using   density-constrained   biclustering.   Bioinformatics.
476   2010;26(18):i625–31.

477   22. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T. Network-based classification of breast
478   cancer metastasis. Mol Syst Biol. 2007;3:140.

479   23. Abraham G, Kowalczyk A, Loi S, Haviv I, Zobel J. Prediction of breast cancer prognosis
480   using gene set statistics provides signature stability and biological context. BMC
481   Bioinformatics. 2010;11:277.

482   24. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM.
483   Inference of patient-specific pathway activities from multi-dimensional cancer genomics
484   data using PARADIGM. Bioinformatics. 2010;26(12):i237–45.

485   25. Zhang W, Ota T, Shridhar V, Chien J, Wu B, Kuang R. Network-based survival analysis
486   reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment.
487   PLoS Comput Biol. 2013;9(3):e1002975.

488   26. Liu K-Q, Liu Z-P, Hao J-K, Chen L, Zhao X-M. Identifying dysregulated pathways in
489   cancers from pathway interaction networks. BMC Bioinformatics. 2012;13:126.

490   27. Chowdhury SA, Nibbe RK, Chance MR, Koyutürk M. Subnetwork state functions define
491   dysregulated subnetworks in cancer. J Comput Biol. 2011;18(3):263–81.

28. Haider S, Yao C, Sabine V, Grzadkowski M, Stimper V, Starmans MH, Wang J, Nguyen F, Moon NC, Lin X, Drake C, Crozier CA, Brookes C, van de Velde CJH, Hasenburg A, Kieback D, Markopoulos CJ, Dirix LY, Seynaeve C, Rea DW, Kasprzyk A, Lio P, Lambin P, Bartlett JMS, Boutros PC. Network-based biomarkers enable cross-disease biomarker discovery. BioRxiv 2018; doi: https://doi.org/10.1101/289934.

29. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.

30. Wu G. Stein L. A network module-based method for identifying cancer prognostic signatures. Genome Biol. 2012;13(12):R112.

31. Li W, Bai X, Hu E, Huang H, Li Y, He Y, Lv J, Chen L, He W. Identification of breast cancer prognostic modules based on weighted protein-protein interaction networks. Oncol Lett. 2017;13(5):3935–41.

32. Haibe-Kains B, Desmedt C, Sotiriou C, Bontempi G. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? Bioinformatics. 2008;24(1):2200–8.

33. Staiger C, Cadot S, Kooter R, Dittrich MT, Müller T, Klau GW, Wessels LF. A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. PLoS One. 2012;7(4):e34796.

34. Miller RG. The jackknife – a review. Biometrika. 1974;61(1):1–15.

35. Lin LI-K. A concordance correlation coefficient to evaluate reproducibility. Biometrics.

512   1989;45(1):255–68.

513   36. Lin LI-K. A note on the concordance correlation coefficient. Biometrics. 2000;56:324–5.

514   37. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP.

515   Exploration, normalization, and summaries of high density oligonucleotide array probe

516   level data. Biostatistics. 2003;4(2):249–64.

517   38. P'ng C, Green J, Chong LC, Waggott D, Prokopec SD, Shamsi M, Nguyen F, Mak DYF,

518   Lam F, Albuquerque MA, Wu Y, Jung EH, Starmans MHW, Chan-Seng-Yue MA, Yao

519   CQ, Liang B, Lalonde E, Haider S, Simone NA, Sendorek D, Chu KC, Moon NC, Fox

520   NS, Grzadkowski MR, Harding NJ, Fung C, Murdoch AR, Houlahan KE, Wang J,

521   Garcia DR, de Borja R, Sun RX, Lin X, Chen GM, Lu A, Shiah Y-J, Zia A, Kearns R,

522   Boutros P. BPG: Seamless, automated and interactive visualization of scientific data.

523   bioRxiv 2017; doi: https://doi.org/10.1101/156067.

524

Figure Legends

Figure 1: Biomarker performance percentiles on a 50:50 training:testing meta-dataset partition

250,000 subnetwork biomarkers were sampled at 25 biomarker sizes ranging from 5 to 150 for a total of 6,250,000 subnetwork biomarkers. 500,000 gene biomarkers were sampled using the same pool of genes at 44 biomarker sizes ranging from 5 to 800 for a total of

531 22,000,000 gene biomarkers. Each biomarker was trained on an initial cohort of 2,490

532 patients and then applied to dichotomize an independent testing cohort of 2,470 patients

533 using the SIMMS algorithm. The various percentiles of biomarker performance are displayed

534 by biomarker size for **(A)** gene biomarkers and **(B)** subnetwork biomarkers. Performance is

535 defined by the hazard ratio of the Cox model fit to the biomarker dichotomized testing cohort

536 survival data.

537 **Figure 2: Gene versus subnetwork biomarker performance on 50:50 meta-dataset**

538 **partition using two biomarker size normalization techniques**

539 For each subnetwork biomarker, we obtained the total number of genes from across all

540 individual subnetworks as well as the number of unique genes. The range between the 95th

541 and 99th percentiles of biomarker performance was re-calculated by grouping biomarkers

542 according to **(A)** the total number of genes and **(B)** the number of unique genes. Note that

543 because the features of a gene biomarker are by definition always a unique set of genes, the

544 total number of genes and the number of unique genes are the same in this case.

545 **Figure 3: Gene biomarker versus subnetwork biomarker performance across all twenty**

546 **meta-dataset partitions**

547 Twenty unique partitions of the breast cancer meta-dataset into training and testing cohorts

548 were created such that ten partitions would have at least 47% of the total number of patients

549 to train on (*large training set*) and the remaining ten partitions would have at most 40% of the

550 patients to train on (*small training set*). The performance of 20 million gene biomarkers (the

551 original set of 22 million biomarkers minus those of over 400 genes in size) and the original

552 set of 6,250,000 subnetwork biomarkers was re-evaluated on each of these training:testing

553 cohort partitions. The 99th percentile of **(A)** gene biomarker performance and **(B)** subnetwork

554 biomarker performance on each of the meta-dataset partitions by biomarker size, along with

555 the 99th percentile of subnetwork biomarker performance by **(C)** the total number of genes

556 and by **(D)** the number of unique genes.

557 **Figure 4: Subnetwork biomarker performance relative to peak gene biomarker**

558 **performance**

559 On each meta-dataset partition, the peak gene biomarker performance was calculated by

560 taking the maximum of the 99th percentile of jackknifed biomarker performance across all

561 biomarker sizes. The 99th percentile of jackknifed subnetwork biomarker performance at each

562 biomarker size was normalized to this peak on each partition by taking the difference between

563 their Cox model betas. Relative subnetwork biomarker performance on **(A,B)** partitions with

564 large training sets (in blue) and on **(C,D)** partitions with small training sets (in red) by **(A,C)**

565 biomarker size (*i.e.* the number of features) and by **(B,D)** the number of unique biomarker

566 genes. For relative subnetwork performance by total number of biomarker genes, see

567 **Additional file 3: Supplementary Figure 3**.

568 **Figure 5: Subnetwork biomarker concordance versus gene biomarker concordance**

569 Lin's concordance correlation coefficient (CCC) was calculated for gene and subnetwork

570 biomarkers using the twenty meta-dataset partitions as independent observers of biomarker

571 performance. This was done at **(A)** each biomarker size (*i.e.* the number of features) and at

572 **(B)** each total gene count. Dotted lines indicate the CCC calculated for gene and subnetwork

573 biomarkers of all sizes combined.

574

Tables

576 **Table 1: Breast cancer dataset sizes**

Source	Number of Samples
Bild	158
Chin	129
Desmedt-1	198
Desmedt-2	107
Hatzis-1	310
Hatzis-2	198
Ivshina	249
Metabric-Training	996
Metabric-Validation	992
Miller	236
Pawitan	159
Sabatier	252
Schmidt	200
Sotiriou	94
Symmans-JBI	65
Symmans-MDA	195
Wang	286
Zhang	136

577

578 **Table 2: Meta-dataset partitions with larger training cohorts**

Partition	Training datasets	Testing datasets	% of cohort used for training	Training Size	Testing Size
Default (1)	Chin, Hatzis-1, Ivshina, Metabric-Training, Miller, Pawitan, Sabatier, Sotiriou, Symmans-	Bild, Desmedt-1, Desmedt-2, Hatzis-2, Metabric-Validation, Schmidt, Symmans-MDA,	50.20	2490	2470

	JBI	Wang, Zhang			
2	Bild, Desmedt-2, Hatzis-1, Hatzis-2, Metabric-Validation, Miller, Pawitan, Sabatier, Zhang	Chin, Desmedt-1, Ivshina, Metabric-Training, Schmidt, Sotiriou, Symmans-JBI, Symmans-MDA, Wang	51.37	2548	2412
3	Bild, Chin, Desmedt-1, Hatzis-1, Metabric-Training, Pawitan, Wang, Zhang	Desmedt-2, Hatzis-2, Ivshina, Metabric-Validation, Miller, Sabatier, Schmidt, Sotiriou, Symmans-JBI, Symmans-MDA	47.82	2372	2588
4	Bild, Chin, Desmedt-1, Hatzis-1, Ivshina, Metabric-Validation, Miller, Schmidt, Symmans-JBI	Desmedt-2, Hatzis-2, Metabric-Training, Pawitan, Sabatier, Sotiriou, Symmans-MDA, Wang, Zhang	51.15	2537	2423
5	Chin, Hatzis-1, Ivshina, Metabric-Validation, Schmidt, Symmans-JBI, Wang, Zhang	Bild, Desmedt-1, Desmedt-2, Hatzis-2, Metabric-Training, Miller, Pawitan, Sabatier, Sotiriou, Symmans-MDA	47.72	2367	2593
6	Bild, Hatzis-1, Hatzis-2, Metabric-Validation, Pawitan, Sabatier, Schmidt, Symmans-JBI	Chin, Desmedt-1, Desmedt-2, Ivshina, Metabric-Training, Miller, Sotiriou, Symmans-MDA, Wang, Zhang	47.06	2334	2626
7	Desmedt-1, Hatzis-2, Ivshina, Metabric-Validation, Sotiriou, Symmans-MDA, Wang, Zhang	Bild, Chin, Desmedt-2, Hatzis-1, Metabric-Training, Miller, Pawitan, Sabatier, Schmidt, Symmans-JBI	47.34	2348	2612
8	Bild, Hatzis-1, Ivshina, Metabric-Training, Miller, Pawitan, Schmidt,	Chin, Desmedt-1, Desmedt-2, Hatzis-2, Metabric-Validation, Sabatier,	49.74	2467	2493

	Sotiriou, Symmans-JBI	Symmans-MDA, Wang, Zhang			
9	Chin, Desmedt-2, Hatzis-2, Ivshina, Metabric-Validation, Sabatier, Sotiriou, Symmans-JBI, Wang	Bild, Desmedt-1, Hatzis-1, Metabric-Training, Miller, Pawitan, Schmidt, Symmans-MDA, Zhang	47.82	2372	2588
10	Bild, Chin, Desmedt-2, Hatzis-1, Hatzis-2, Ivshina, Metabric-Training, Pawitan, Schmidt, Symmans-MDA, Wang	Desmedt-1, Metabric-Validation, Miller, Sabatier, Sotiriou, Symmans-JBI, Zhang	60.22	2987	1973

579

580 **Table 3: Meta-dataset partitions with smaller training cohorts**

Partition	Training datasets	Testing datasets	% of cohort used for training	Training Size	Testing Size
Default (1)	Bild, Chin, Desmedt-2, Hatzis-1, Hatzis-2, Miller, Pawitan, Sabatier, Symmans-JBI, Symmans-MDA	Desmedt-1, Ivshina, Metabric-Training, Metabric-Validation, Schmidt, Sotiriou, Wang, Zhang	36.47	1809	3151
2	Desmedt-1, Desmedt-2, Hatzis-1, Hatzis-2, Ivshina, Miller, Schmidt, Symmans-MDA, Zhang	Bild, Chin, Metabric-Training, Metabric-Validation, Pawitan, Sabatier, Sotiriou, Symmans-JBI, Wang	36.88	1829	3131
3	Bild, Desmedt-1, Hatzis-2, Ivshina, Metabric-Validation	Chin, Desmedt-2, Hatzis-1, Metabric-Training, Miller, Pawitan, Sabatier, Schmidt, Sotiriou, Symmans-JBI, Symmans-MDA,	36.19	1795	3165

		Wang, Zhang			
4	Bild, Desmedt-1, Metabric-Training, Miller, Pawitan, Symmans-JBI	Chin, Desmedt-2, Hatzis-1, Hatzis-2, Ivshina, Metabric-Validation, Sabatier, Schmidt, Sotiriou, Symmans-MDA, Wang, Zhang	36.53	1812	3148
5	Bild, Desmedt-2, Hatzis-1, Hatzis-2, Ivshina, Schmidt, Sotiriou, Symmans-JBI, Wang, Zhang	Chin, Desmedt-1, Metabric-Training, Metabric-Validation, Miller, Pawitan, Sabatier, Symmans-MDA	36.35	1803	3157
6	Bild, Desmedt-1, Desmedt-2, Hatzis-2, Sabatier, Schmidt, Sotiriou, Symmans-MDA, Wang, Zhang	Chin, Hatzis-1, Ivshina, Metabric-Training, Metabric-Validation, Miller, Pawitan, Symmans-JBI	36.77	1824	3136
7	Bild, Chin, Desmedt-1, Desmedt-2, Hatzis-1, Ivshina, Pawitan, Sabatier, Symmans-MDA	Hatzis-2, Metabric-Training, Metabric-Validation, Miller, Schmidt, Sotiriou, Symmans-JBI, Wang, Zhang	35.42	1757	3203
8	Desmedt-2, Metabric-Training, Miller, Sotiriou, Symmans-JBI, Symmans-MDA	Bild, Chin, Desmedt-1, Hatzis-1, Hatzis-2, Ivshina, Metabric-Validation, Pawitan, Sabatier, Schmidt, Wang, Zhang	34.13	1693	3267
9	Bild, Desmedt-1, Desmedt-2, Ivshina, Pawitan, Sabatier, Schmidt, Sotiriou, Wang	Chin, Hatzis-1, Hatzis-2, Metabric-Training, Metabric-Validation, Miller, Symmans-JBI, Symmans-MDA, Zhang	34.33	1703	3257
10	Bild, Chin, Ivshina,	Desmedt-1,	34.72	1722	3238

| | Pawitan, Sabatier, Schmidt, Sotiriou, Symmans-MDA, Wang | Desmedt-2, Hatzis-1, Hatzis-2, Metabric-Training, Metabric-Validation, Miller, Symmans-JBI, Zhang | | | |

581

Additional files

583 **Additional file 1: Supplementary Figure 1.**

584 TIFF 4.5 Mb

585 **Gene biomarker versus subnetwork biomarker performance using a 37/63 meta-dataset**

586 **partition.** The performances of the 22,000,000 gene biomarkers and 6,250,000 subnetwork

587 biomarkers sampled were re-evaluated using a smaller training set. From the total 4,960

588 patient meta-dataset, 1,809 patients were randomly drawn for training and the remaining

589 3,151 patients were used to test dichotomization efficacy. The various percentiles of

590 biomarker performance using this smaller training cohort by biomarker size for **(A)** gene

591 biomarkers and **(B)** subnetwork biomarkers, along with the range of biomarker performance

592 between the 95th and 99th percentiles by **(C)** the total number of feature genes and by **(D)**

593 the number of unique feature genes for both biomarker types.

594 **Additional file 2: Supplementary Figure 2.**

595 TIFF 4.8 Mb

596 **Biomarker performance concordance across small training cohorts and large training**

597 **cohorts.** Lin's concordance correlation coefficient (CCC) was calculated treating meta-

dataset partitions as independent observers of biomarker performance. Partitions were grouped together using the two training cohort size classes ('small' and 'large') and the CCC for each partition class was calculated for **(A)** gene biomarkers and **(B)** subnetwork biomarkers at each biomarker size and also for subnetwork biomarkers at each **(C)** total gene count and **(D)** unique gene count. Dotted lines indicate the CCC calculated for each partition class using all biomarkers at once.

Additional file 3: Supplementary Figure 3.

TIFF 6.2 Mb

Subnetwork biomarker performance relative to gene biomarker performance by the total number of genes. On each meta-dataset partition, the peak gene biomarker performance was calculated by taking the maximum of the 99th percentile of jackknifed biomarker performance across all biomarker sizes. The 99th percentile of jackknifed subnetwork biomarker performance at each biomarker size was normalized to this peak on each partition by taking the difference between the Cox model betas. Relative subnetwork biomarker performance on **(A)** partitions with large training sets and on **(B)** partitions with small training sets by the total number of biomarker genes.