Robust metabolic transcriptional components in 34,494 patient-derived samples and cell lines

V.C. Leeuwenburgh[1,2,#], C.G. Urzúa-Traslaviña[1,#], A. Bhattacharya[1], M.T.C. Walvoort[2], M. Jalving[1], S. de Jong[1], R.S.N. Fehrmann[1]

[1] Department of Medical Oncology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands

[2] Department of Chemical Biology, Stratingh Institute for Chemistry, University of Groningen, the Netherlands

[#] Contributed equally

Corresponding author / Lead contact:

Rudolf S.N. Fehrmann, MD, PhD

Department of Medical Oncology, University Medical Center Groningen

P.O. Box 30.001, 9700 RB Groningen, the Netherlands

Tel. +31503612821, Fax +31503614862

E-mail: r.s.n.fehrmann@umcg.nl

ABSTRACT

Patient-derived expression profiles of cancers can provide insight into transcriptional changes that underlie reprogrammed metabolism in cancer. These profiles represent the average expression pattern of all heterogeneous tumor and non-tumor cells present in biopsies of tumor lesions. Therefore, subtle transcriptional footprints of metabolic processes can be concealed by other biological processes and experimental artifacts. We, therefore, performed consensus Independent Component Analyses (c-ICA) with 34,494 bulk expression profiles of patient-derived tumor biopsies, non-cancer tissues, and cell lines. c-ICA enabled us to create a transcriptional metabolic landscape in which many robust metabolic transcriptional components and their activation score in individual samples were defined. Here we demonstrate how this landscape can be used to explore associations between the metabolic transcriptome and drug sensitivities, patient outcomes, and the composition of the immune tumor microenvironment. The metabolic landscape can be explored at http://www.themetaboliclandscapeofcancer.com.

32 **INTRODUCTION**

33 Reprogrammed energy metabolism is a hallmark of cancer (Hanahan and Weinberg, 2011).
34 Metabolic reprogramming supports the survival, proliferation, and maintenance of cancer cells
35 by ensuring sufficient biosynthetic capacity, redox potential, and energy (Pavlova and Thompson,
36 2016; Vazquez et al., 2016). Additionally, metabolic reprogramming enables tumor cells to adapt
37 to challenging microenvironmental conditions, such as hypoxia and low nutrient availability, and
38 become resistant to cancer treatment (Huang et al., 2014; Viale and Draetta, 2016). Moreover,
39 metabolic reprogramming of cancer cells influences the composition and function of immune
40 cells present in the tumor microenvironment (TME), affecting the anti-cancer immune response
41 to immunotherapy (Le Bourgeois et al., 2018; Quail and Joyce, 2013).

42 Metabolic dependencies have been successfully exploited to treat cancer, as illustrated by
43 the efficacy of antifolate drugs such as methotrexate (Walling, 2006). More recent knowledge
44 about cancer cell metabolism has resulted in novel therapeutic targets, such as glutaminase and
45 mutant forms of IDH1/2, currently being evaluated in pre-clinical models and phase I/II clinical
46 trials (Shah and Chen, 2020; Tang et al., 2021). However, adverse effects or lack of effectiveness
47 still hamper the clinical development of most metabolic therapies. A potential reason is that many
48 metabolic targeting drugs are developed based on insights derived from model systems of human
49 cancer, which do not fully reflect the complexities of cancer in humans (Ghaffari et al., 2015). In
50 particular, cell line models lack the immune cells present in the TME (Hynds et al., 2018; Jiang et
51 al., 2016; Vincent and Postovit, 2017) and often require specific metabolic conditions to grow
52 (Ben-David et al., 2018; Hynds et al., 2018).

53 Evidence is emerging that transcriptional changes play an important role in the metabolic
54 plasticity of cancer cells: gene expression can influence metabolite levels, and metabolic changes
55 can result in altered gene expression (Desvergne et al., 2006; Martin-Martin et al., 2018; Peng et
56 al., 2018). The availability of large numbers of gene expression profiles — from a broad spectrum
57 of cancer types — in the public domain provides a unique opportunity to study metabolic
58 reprogramming in patient-derived cancer tissue.

59 Almost without exception, these gene expression profiles were generated from complex
60 biopsies that contain tumor cells and cells present in the TME (e.g., immune cells). Accordingly,
61 these profiles represent the average gene expression pattern of all cells present in the biopsy.
62 Therefore, detecting metabolic processes relevant to cancer biology in expression profiles from
63 complex biopsies can be challenging, especially when their transcriptional footprints (TFs) are
64 subtle and concealed by more pronounced TFs from other biological processes or experimental
65 artifacts.

66 In the present study, we used consensus Independent Component Analyses (c-ICA), a
67 statistical method capable of separating the average gene expression profiles generated from
68 complex biopsies into additive transcriptional components (TCs). This enabled us to detect both
69 the pronounced and more subtle transcriptional footprints of metabolic processes. We
70 performed c-ICA with 32,409 gene expression profiles obtained from the Gene Expression
71 Omnibus (GEO) and The Cancer Genome Atlas (TCGA), as well as 2,085 gene expression profiles
72 obtained from the Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in
73 Cancer portal (GDSC) (Barret et al., 2013; Barretina et al., 2012; Yang et al., 2013). Comprehensive
74 characterization of the TCs with gene set enrichment analysis (GSEA) identified TCs associated

with metabolic processes, i.e., metabolic TCs (mTCs). This enabled us to create a metabolic landscape showing the activity of these mTCs in all 34,494 samples. We demonstrate how this landscape (www.themetaboliclandscapeofcancer.com) can be used to explore associations between the metabolic transcriptome and drug sensitivities, patient outcomes, and the composition of immune cells in the TME.

RESULTS

A subset of transcriptional components is associated with metabolic processes

Previously, we collected gene expression data from four databases: the Gene Expression Omnibus (GEO dataset, *n* = 21,592), The Cancer Genome Atlas (TCGA dataset, *n* = 10,817), the Cancer Cell Line Encyclopedia (CCLE dataset, *n* = 1,067), and the Genomics of Drug Sensitivity in Cancer (GDSC dataset, *n* = 1,018) (**Figure 1A**), totaling 34,494 samples (Bhattacharya et al., 2020). Overall, 28,200 expression profiles originated from patient-derived complex tissue cancer biopsies, 4,209 from complex tissue biopsies of non-cancerous tissue, and 2,085 from cell lines. The samples in these four databases encompass 89 cancer tissue types and subtypes and 19 non-cancerous tissue types. For GEO and CCLE data sets, the expression profiles were generated with Affymetrix HG-U133 Plus 2.0. Expression profiles within the GDSC dataset were generated with Affymetrix Human Genome U219, and TCGA profiles were generated with RNA sequencing.

Gene expression profiling measures the net expression level of individual genes, thus reflecting the integrated activity of underlying regulatory factors, including experimental, genetic, and non-genetic factors. To gain insight into the number and nature of these regulatory factors and their effects on gene expression levels, i.e., their transcriptional footprints, we previously

97 performed consensus-independent component analysis (c-ICA) on each of the abovementioned
98 four datasets separately (Bhattacharya et al., 2020), resulting in four sets of transcriptional
99 components (TCs). In every TC, each gene has a specific weight. This weight describes how
100 strongly and in which direction the underlying transcriptional regulatory factor influences the
101 expression level of that gene. c-ICA also provides a 'mixing-matrix' per dataset, in which each
102 column corresponds to a TC and each row corresponds to a sample. Values in the mixing matrix
103 are interpreted as measurements of the activity of the TCs in an individual sample; we refer to
104 these as 'activity scores'. Ultimately, the analysis yielded 855, 1383, 466, and 467 TCs for GEO,
105 TCGA, CCLE, and GDSC datasets, respectively (**Figure 1A**).

106 Gene set enrichment analysis (GSEA) with 608 gene sets that describe metabolic processes
107 was performed to identify TCs enriched for metabolic processes. The gene sets were selected
108 from the gene set collections Biocarta (*n* = 7), the Kyoto Encyclopedia of Genes and Genomes
109 (KEGG, *n* = 64), the Gene Ontology Consortium (GO, *n* = 508), and Reactome (*n* = 29) within the
110 Molecular Signatures DataBase (MSigDB, v6.1; for the systematic selection strategy see
111 **Methods**). We performed consensus clustering on the enrichment scores of the 608 metabolic
112 gene sets to identify potential biological redundancy in the metabolic gene set definitions (**Figure**
113 **S1**). This resulted in 50 clusters of gene sets, which can be ascribed to different metabolic themes
114 (**Table S1**). Based on these 50 enrichment clusters, 132 (GEO), 151 (TCGA), 136 (CCLE), and 136
115 (GDSC) mTCs were defined (**Figure 1A** and **B**); see **Methods** for the systematic selection strategy).
116 These mTCs represent the metabolic transcriptional footprints present in our broad set of
117 samples, i.e., patient-derived samples, cancer cell line samples, and non-cancer samples.
118 However, some of the identified mTCs may capture the transcriptional footprints of experimental

119 factors. Therefore, we investigated how much of the variance in activity scores of each mTC could
120 be explained by experimental batches. For GEO mTCs, experimental batches were determined by
121 the provided GSE identifiers (i.e., experiment series identifiers). For TCGA mTCs, experimental
122 batches were determined by the tissue source site of samples (e.g., 2H, Erasmus MC, esophageal
123 carcinoma). We observed that 12/132 GEO mTCs showed a potential putative batch effect with
124 more than 10% explained variance (**Figure S2A**). However, six of the 12 GEO mTCs with a putative
125 batch effect also explained more than 10% of the variance in gene expression of samples
126 belonging to a single tissue subtype (**Figure S2A**). One of the 151 TCGA mTCs showed a putative
127 batch effect with 20.5% explained variance (**Figure S2B**). This mTC, TCGA mTC 43, also showed
128 tissue-specificity for thymoma, a tissue type that is not present in the GEO dataset. These
129 observations might indicate that the mTCs showing a putative batch effect in fact describe tissue-
130 specific biology of tissues that are only present in a single experiment in our dataset.
131
132 **Metabolic TCs are robust across different datasets and platforms**
133 Pair-wise comparison of mTCs between datasets, based on gene weights, showed that 91-99% of
134 mTCs per dataset were highly correlated ($|r_s| \geq 0.5$, P-value < 0.05 as a threshold) with at least
135 one mTC identified in another dataset (**Figure 1C**, **D** and **Figure S3A-G**). This indicates that most
136 of the mTCs were cross-platform and cross-dataset robust.
137 Given the selected correlation threshold ($|r_s| \geq 0.5$, P-value < 0.05), 72 mTCs could be
138 identified with a highly similar gene weight pattern in all four datasets (**Figure 1D**). Thus, these
139 mTCs capture a transcriptional footprint that is very similar in both patient-derived complex
140 biopsies and cell lines. As cell lines lack a TME, these 72 mTCs were considered to capture

141 metabolic processes that reflect tumor cell characteristics. Six GEO mTCs were identified that
142 were highly correlated with TCGA mTCs, but not highly correlated with any CCLE or GDSC mTC
143 (**Figure 1D**). These mTCs, therefore, might capture transcriptional footprints that are specific for
144 complex biopsies obtained from patient-derived cancer tissue and may originate from the TME
145 or capture a transcriptional footprint from tissue only present in the GEO and TCGA datasets. One
146 pair of mTCs was identified with a gene weight pattern that was highly similar in CCLE and GDSC
147 datasets only, capturing a metabolic transcriptional footprint that could only be found in cell line
148 models (**Figure 1D**).
149

150 **Metabolic TCs identify new genes potentially involved in metabolic processes**

151 Among the 'top' genes in every mTC —defined as the genes with an absolute weight > 3 in an mTC
152 — many genes were member of the 608 metabolic gene sets (**Figure 2A** and **2B, Figure S4**).
153 However, even for the mTCs with the absolute highest gene set enrichment scores for a metabolic
154 gene set, at least 20% of top genes were not members of any of the metabolic gene sets. Because
155 these genes were nevertheless part of an mTC, they may be potentially involved in the metabolic
156 processes that showed enrichment.

157 For example, two strongly correlated mTCs, GEO mTC 54 and TCGA mTC 127 ($|r_s| = 0.77$),
158 both showed enrichment for glycolysis and the metabolic process of ADP (**Figure 2C** and **2D, Table**
159 **S1**). GEO mTC 54 contained 262 top genes, of which 155 (59.1%) were also among the top genes
160 in TCGA mTC 127. Both mTCs contained multiple top genes that are known targets of the HIF-1
161 complex, and genes previously found to be part of a hypoxic signature (Benita et al., 2009; Ye et
162 al., 2018). Several top genes of both GEO mTC 54 and TCGA mTC 127 (e.g., *FAM162A*, *C4orf3*,

163 *C4orf47*, and *ANKRD37*) are currently not a member of any of the 608 metabolic gene sets.
164 However, these data suggest that these four genes are involved in glycolysis and are possibly
165 hypoxia related. Indeed, several studies have indicated that at least *FAM162A* and *ANKRD37* are
166 regulated by the transcription factor HIF-1α (Copple et al., 2012; Sørensen et al., 2015).

167 As a second example, we investigated two highly correlated mTCs, GEO mTC 11 and TCGA
168 mTC 141 (r_s = 0.68), which showed enrichment for mitochondrial metabolic processes such as
169 oxidative phosphorylation and the TCA cycle (**Figure 2E** and **2F, Table S1**). GEO mTC 11 contained
170 427 top genes, of which 270 (63.2%) were among the top genes in TCGA mTC 141. In these two
171 mTCs, *C6orf136* and *IMMT* are top genes currently not assigned to any of the 608 metabolic gene
172 sets. *C6orf136* and *IMMT* were previously identified in functional mitochondria' proteome
173 profiles (Lefort et al., 2009). These results suggest that mTCs could assign metabolic functions to
174 genes currently not members of known gene sets describing metabolic processes.

175

176 **Clustering sample activity scores of mTCs reveal multiple metabolic subtypes**

177 To investigate the heterogeneity of the metabolic transcriptome in a broad range of cancer
178 subtypes, we hierarchically clustered the mixing matrix provided by consensus-ICA that contains
179 the activity score of mTCs in every sample (**Figure 3A**, **3B** and **S5A, S5B**). We selected the cutoff
180 heights of the resulting dendrograms so that every cluster – referred to as metabolic subtype –
181 contained at least 50 samples (**Figure S5C** and **S5D**). This clustering analysis divided the 21,592
182 GEO samples into 67 metabolic subtypes with a median of 276 samples per subtype (range 54-
183 1,252) and the 10,817 TCGA samples into 58 metabolic subtypes with a median of 167 samples

184 per subtype (range 52-536). For an overview of the metabolic subtypes and their sample
185 composition, see **Figure S6**, **S7,** and **Table S2.** Two types of patterns emerged.

186 The first pattern consisted of tumor types with samples that belong to one dominant
187 metabolic subtype. For example, 102/133 (76.7%) of thyroid cancer samples in the GEO dataset
188 fell into one metabolic subtype (subtype 27, **Figure S6, Table S2**). Similarly, 446/509 (87.6%) of
189 thyroid cancer samples in the TCGA dataset fell into metabolic subtype 43 (**Figure S7, Table S2**).
190 In line with the biology of thyroid tissue, both GEO metabolic subtype 27 and TCGA metabolic
191 subtype 43 were characterized by high activity scores of mTCs enriched for thyroid hormone
192 metabolism (GEO mTC 64 and TCGA mTC 87; **Table S1**).

193 The second pattern consisted of several tumor types that were not characterized by a few
194 dominant metabolic subtypes. Instead, their samples were divided across multiple metabolic
195 subtypes. For example, the 3,512 breast cancer samples in the GEO dataset were divided across
196 33 metabolic subtypes (**Figure 3C**). These metabolic subtypes did not follow the breast cancer
197 classification based on ER and HER2 receptor status (**Figure 3C** and **Table S2**). In line with this
198 observation in the GEO dataset, the 1,100 breast cancer samples in the TCGA dataset were also
199 scattered across 29 metabolic subtypes.

200 Several metabolic subtypes likewise contained samples from multiple tumor types. For
201 example, GEO metabolic subtype 22 contained samples from 25 tumor types, including 42 ovarian
202 cancer samples (22% of all ovarian cancer), 33 synovial sarcoma samples (97% of all synovial
203 sarcoma), and 15 Ewing's sarcoma samples (58% of all Ewing's sarcoma; **Figure S6** and **Table S2**).
204 GEO mTC 111 had the highest absolute median activity score in GEO metabolic subtype 22 (**Table**

205 S2). This mTC showed enrichment for the metabolism of nicotinamide adenine dinucleotide
206 phosphate (NADP) and genes involved in the activation of an innate immune response (Table S1).
207 These results show that the classification of samples based on metabolic subtype yields
208 different patterns than current classification systems, such as histotype or receptor status in
209 breast cancer.
210
211 **Metabolic subtypes are associated with distant relapse-free survival in breast cancer**
212 We then investigated if metabolic subtypes could have clinical relevance. We had previously
213 collected distant relapse-free survival (DRFS) data for 1,207 breast cancer samples (Bense et al.,
214 2017). As mentioned earlier, breast cancer samples in the GEO dataset were divided across 33 of
215 the 67 metabolic subtypes. Of these 33 subtypes, eight contained > 50 breast cancer samples with
216 data available for DRFS: subtypes 15, 16, 20, 31, 32, 33, 34, and 35. We found that patients from
217 breast cancer samples assigned to metabolic subtypes 16 and 33 showed the best and worst
218 DRFS, respectively (P-value = $1.08 \cdot 10^{-23}$, Log-Rank test; Figure 3**D**). Distributions of standard
219 prognostic factors within these eight metabolic subtypes are presented in Table S3. These results
220 show that metabolic subtypes are associated with disease outcomes in breast cancer.
221
222 **The activity of mTCs is associated with drug sensitivity**
223 The CCLE and GDSC databases contain the sensitivities of cell lines to a large panel of drugs
224 expressed as IC_{50} values. With a threshold of $|r_s|>0.2$, we observed associations between the
225 activity scores of 61 CCLE mTCs, 90 GDSC mTCs, and the IC50 values of 238 drugs (Table S4).

226 For example, in the GDSC dataset, an increase in activity score of GDSC mTC 3 was
227 associated with a decrease in IC_{50} value of (i.e., increased sensitivity to) nutlin-3a ($|r_s|$= 0.42;
228 **Figure 4A** and **4B**). Nutlin-3a targets the p53 pathway through inhibition of MDM2. In line with
229 this, GDSC mTC 3 showed strong enrichment for genes involved in the p53 pathway, with *MDM2*
230 ranked as the second gene (**Table S1**). GDSC mTC 3 was strongly correlated with CCLE mTC 4 ($|r_s|$=
231 0.84), GEO mTC 57 ($|r_s|$= 0.79), and TCGA mTC 110 ($|r_s|$= 0.74) (**Figure 4D**), suggesting that this
232 mTC was captured in cell line datasets as well as in the two patient-derived datasets. Indeed, an
233 increase in activity score of CCLE mTC 4 was associated with a decrease in IC_{50} value of nutlin-3a
234 as well ($|r_s|$= 0.25; **Figure 4E**). Cell lines with wildtype *TP53* had a higher activity score of GDSC
235 mTC 3 (**Figure 4C**). Also, cell lines with wildtype *TP53* had a higher activity score of CCLE mTC 4
236 (**Figure 4F**).
237 In another example, the activity score of GDSC mTC 18 was found to be associated with the
238 IC50 values of 142 drugs ($|r_s|$ range 0.20 – 0.44; **Figure 4G**). An increase in activity score of GDSC
239 mTC 18 in a sample was associated with a higher IC_{50} value (i.e., increased resistance) for 135 of
240 these drugs, including the widely used DNA synthesis-inhibiting antimetabolites 5-fluorouracil
241 ($|r_s|$ = 0.41) and methotrexate ($|r_s|$ = 0.38). GDSC mTC 18 was strongly correlated with CCLE mTC
242 28 ($|r_s|$= 0.84), GEO mTC 35 ($|r_s|$= 0.59), and TCGA mTC 58 ($|r_s|$= 0.55), indicating that this mTC
243 is also captured in both cell line datasets and the two patient-derived datasets. In line with this,
244 CCLE mTC 28 was associated with a higher IC_{50} value (i.e., increased resistance) for 7 drugs
245 including topoisomerase inhibitors topotecan ($|r_s|$ = 0.35) and irinotecan ($|r_s|$ = 0.34) (**Figure 4H**).
246 All four of the highly correlated mTCs were enriched for genes involved in glutathione
247 metabolism, cellular ketones and xenobiotics, and drug detoxification (**Table S1**). Specifically,

248 genes belonging to the aldo-keto reductase family 1 (AKR1) were among the top genes in these
249 mTCs. Previous studies have reported a role for the glutathione system in resistance to irinotecan
250 and 5-fluorouracil (Goto et al., 2002), and specifically the AKR1 family in resistance to e.g.
251 methotrexate and irinotecan (Heibein et al., 2012; Matsunaga et al., 2020; Selga et al., 2008). In
252 contrast, we observed that an increased activity score of GDSC mTC 18 was associated with a
253 decrease in IC_{50} value (i.e., increased sensitivity) for only seven drugs ($|r_s|$ range 0.20-0.41; **Figure**
254 **4G**). The drug with the highest negative correlation was tanespimycin (17-AAG), an Hsp90
255 inhibitor ($|r_s| = 0.41$). An increased activity score of CCLE mTC 28 was associated with decrease
256 in IC_{50} value for tanespimycin as well ($|r_s| = 0.26$; **Figure 4H**). A direct link between the functions
257 of glutathione and Hsp90 in oxidative stress has been suggested, as well as a relationship between
258 tanespimycin sensitivity and *NQO1* expression, a gene coding for an enzyme reducing quinones
259 to hydroquinones that is involved in detoxification pathways (Gaspar et al., 2009; Kim et al.,
260 2015). In line with these findings, we found that the *NQO1* gene is present near the top of GDSC
261 mTC 18, CCLE mTC 28, GEO mTC 35, and TCGA mTC 58.

262 As a final example, increased activity of GDSC mTC 108 was associated with a lower IC_{50}
263 value (i.e., increased sensitivity) to the MEK inhibitor trametinib ($|r_s| = 0.48$) and a higher IC_{50}
264 value (i.e., increased resistance) to the histone deacetylase inhibitor vorinostat ($|r_s| = 0.46$; **Figure**
265 **4I** and **Table S4**). GDSC mTC 108 was correlated with CCLE mTC 97 ($|r_s| = 0.32$). Consistent with
266 the observation for GDSC mTC 108, we found that increased activity of CCLE mTC 97 was
267 associated with a lower IC_{50} value (i.e., increased sensitivity) to the MEK inhibitor mirdametinib
268 ($|r_s| = 0.24$) and a higher IC_{50} value (i.e., increased resistance) to the histone deacetylase inhibitor
269 panobinostat ($|r_s| = 0.43$; **Figure 4J** and **Table S4**). This contrasting sensitivity for MEK and histone

270 deacetylase inhibition is in line with data from a study that used *BRAF*-mutated melanoma cell
271 lines. The authors showed that cell lines with acquired resistance to MEK inhibitors subsequently
272 became sensitive to treatment with the histone deacetylase inhibitor vorinostat (Wang et al.,
273 2018). They concluded that the MEK-inhibitor resistance mechanism results from the activation
274 (or reactivation) of MAPK cascades (Wagle et al., 2014). These findings are in line with our
275 observation that both GDSC mTC 108 and CCLE mTC 97 were enriched for genes involved in the
276 negative regulation of the MAPK cascade (**Table S1**). These examples demonstrate how mTCs can
277 capture cross-dataset robust metabolic transcriptional footprints relevant for drug response.
278

279 **The activity of mTCs is associated with the immune composition of the tumor**
280 **microenvironment**

281 We determined the association between the activity of mTCs and the immune composition of the
282 TME (**Table S5;** see **Methods** for details). The immune composition for all samples in the GEO and
283 TCGA dataset was determined by inferring fractions of 22 immune cell types using the CIBERSORT
284 algorithm (Chen et al., 2018). We observed that the mTCs that were correlated with immune cell
285 fractions could be divided into two groups. The first group included mTCs that were only identified
286 in the patient-derived datasets. The second group contained mTCs that were identified in both
287 the patient-derived and the cell line datasets.

288 For example, the activity score of GEO mTC 123 was associated with estimated fractions of
289 CD8+ T cells ($|r_s|$ = 0.40), γδ T cells ($|r_s|$ = 0.36), activated CD4 memory T cells ($|r_s|$ = 0.34), and
290 regulatory T cells ($|r_s|$ = 0.32, **Figure 5A**). Belonging to the group of mTCs only identified in the
291 patient-derived datasets, GEO mTC 123 was correlated highly with only TCGA mTC 34 ($|r_s|$= 0.28).

292 In line with this, the activity score of TCGA mTC 34 was also associated with CD8+ T cell fractions
293 ($|r_s|$ = 0.58, **Figure 5B**). Both GEO mTC 123 and TCGA mTC 34 showed enrichment for genes
294 involved in immunological processes such as leukocyte activation and cytokine metabolism and
295 metabolic processes such as phosphatidylinositol and phospholipid metabolism (**Table S1**). The
296 fact that both GEO mTC 123 and TCGA mTC 34 have no high correlation to mTCs in the cell line
297 datasets suggests that they indeed capture transcriptional activity from non-cancerous cells in
298 the immune TME.

299 GEO mTC 14 is illustrative of the second group of mTCs correlated with immune cell
300 fractions and identified in both the patient-derived and the cell line datasets. The activity scores
301 of GEO mTC 14 were correlated with the fractions of M1 macrophages ($|r_s|$= 0.65) and M2
302 macrophages ($|r_s|$= 0.59; **Figure 5C**). GEO mTC 14 was correlated with TCGA mTC 70 ($|r_s|$= 0.44)
303 and with CCLE mTC 124 ($|r_s|$= 0.47), and GDSC mTC 33 ($|r_s|$= 0.33). All four mTCs were enriched
304 for genes involved in the metabolism of extracellular macromolecules (**Tables S1**). Genes coding
305 for several types of collagens were among the top-ranked in these mTCs. This is in line with
306 previous reports indicating that macrophages can function as collagen-producing cells in the TME
307 (Schnoor et al., 2008; Vaage and Harlos, 1991). GEO mTC 14 and TCGA mTC 70 showed a high
308 activity score in subsets of breast cancers, lung cancers, and sarcomas (**Figure S8A** and **S8B**). A
309 negative activity score of GEO mTC 14 and TCGA mTC 70 was observed in a subset of
310 hematological cancers and hematological cancer cell lines in both GDSC and CCLE mTCs. Because
311 these mTCs were present in both patient data sets and cell line datasets, this indicates that the
312 captured metabolic processes reflect tumor cell characteristics, which are associated with the
313 fraction of macrophages present in the immune TME.

By correlating inferred immune cell fractions of samples with the activity scores of mTCs in samples, the relationship between the metabolic transcriptome and the various components of the immune TME could be assessed.

DISCUSSION

In the present study, we used consensus-Independent Component Analysis (c-ICA) in combination with Gene Set Enrichment Analysis (GSEA) to identify a broad set of robust metabolic Transcriptional Components (mTCs). With these mTCs, the transcriptional metabolic landscape was defined in patient-derived cancer tissue, cancer cell lines, and non-cancer samples. We also showed how this metabolic landscape could be used to generate hypotheses by exploring associations between metabolic processes and drug sensitivities, patient outcomes, and the composition of the immune tumor microenvironment.

We used the wealth of publicly available pan-cancer transcriptomic data to study human metabolism on a large scale. Previous work used either single-cell sequencing or bulk cell transcriptomic profiles to study metabolism in specific cancer types (Hakimi et al., 2016; Xiao et al., 2019), or pan-cancer, but based on a single platform (Cubuk et al., 2018; Rosario et al., 2018). Our present study differs from this previous work in two essential aspects. Firstly, we used c-ICA to segregate the average expression patterns of complex biopsies into statistically independent components (Biton et al., 2014; Kong et al., 2008). Previous studies investigated average gene expression profiles of complex biopsies and can therefore only distinguish the gene expression signature and regulation of more pronounced metabolic processes. With c-ICA it is possible to identify statistically independent regulatory factors and their transcriptional footprints and

336 distinguish both pronounced from more subtle metabolic processes. This enabled us to determine
337 the association of both pronounced and subtle metabolic processes with, e.g., patient outcome
338 and the composition of the TME in a complex biopsy. Secondly, the present study is the most
339 extensive transcriptional analysis of metabolism and the first that integrated patient-derived data
340 from GEO and TCGA with cell line data from CCLE and GDSC. The samples in these four datasets
341 were obtained from a multitude of independently constructed, publicly available cohorts, and the
342 expression profiles were generated using different technologies (microarray or RNA-sequencing).
343 This integrated dataset enabled us to demonstrate that most of the identified mTCs were robust
344 and independent from dataset-specific and platform-specific characteristics. The observed
345 overlap, or lack of overlap, between patient-derived and cell line-derived mTCs can help
346 researchers understand how metabolic genes and pathways identified in cell lines can be
347 translated to a patient tissue context and vice versa.

348 Furthermore, we hypothesize that metabolic processes identified only in patient-derived
349 samples and not in cell line samples are more likely to originate from cells in the tumor
350 microenvironment. These microenvironment-specific metabolic processes will not be captured
351 by mTCs in cell line datasets. This is because bulk expression profiles of cancer cell line samples
352 do not harbor transcriptional footprints associated with non-cancerous cells.

353 The metabolic landscape enabled us to classify samples based on the transcriptional activity
354 of metabolic processes, resulting in metabolic subtypes. However, this metabolic classification
355 was often not in full alignment with current classification systems based on aspects such as
356 histotype. We demonstrated that metabolic subtypes were associated with disease outcomes for
357 breast cancer, emphasizing the relevance of metabolic pathway-based classification in cancer.

358 The heterogeneity (metabolic and otherwise) within and between cancer types is well recognized,
359 and alternative subtyping based on metabolite profiling and the metabolic transcriptome have
360 been proposed before (Reznik et al., 2018; Rosario et al., 2018; Tang et al., 2014). More
361 specifically, clinically significant metabolism-based classifications have been proposed in breast
362 cancer (Cappelletti et al., 2017; Serrano-Carbajal et al., 2020; Wang et al., 2019). The most active
363 mTCs in a metabolic subtype relevant to disease outcome could thus be used to generate new
364 hypotheses for treatment targets. Additionally, the association between the activity of mTCs and
365 drug sensitivity could help to design these future therapeutic strategies.

366 Metabolic heterogeneity and plasticity are not limited to cancer cells but are also applicable
367 to the immune cells present in the tumor micro-environment. Immune cells undergo metabolic
368 changes when activated, and their metabolic status can overlap with the metabolic state of cancer
369 cells (Andrejeva and Rathmell, 2017). For example, the Warburg effect is classically seen as an
370 example of a metabolic transformation in cancer cells. However, it is also observed in activated T
371 cells (Bantug et al., 2018; Patel and Powell, 2017; Wang and Green, 2012). In the context of
372 metabolism, this complex interplay between cancer cells and immune cells present in the micro-
373 environment gives a new dimension to the use of drugs that target metabolic processes
374 (O'Sullivan et al., 2019; Patel et al., 2019). For instance, inhibiting glutamine metabolism has been
375 shown to inhibit tumor growth and increase the sensitivity of triple-negative breast cancers to
376 immune checkpoint blockade (Oh et al., 2020), and reducing oxidative stress has been shown to
377 prevent the generation of tumor-associated macrophages (Zhang et al., 2013). Furthermore,
378 modulating metabolism in T cells from glycolytic to an OXPHOS-weighted profile has been shown
379 to improve CAR T cell immunotherapy (Fraietta et al., 2018; O'Sullivan and Pearce, 2015; Sukumar

380 et al., 2017). Our transcriptional metabolic landscape can contribute to knowledge on
381 immunometabolism and, combined with the association of mTCs with drug sensitivity, also
382 contribute to the formulation of new hypotheses on how to metabolically engage the tumor and
383 its immune microenvironment, thus improving the response to immunotherapy.

384 Further research to gain an even more comprehensive understanding of metabolism in
385 patient-derived cancer samples should ideally integrate genomics, transcriptomics, proteomics,
386 and metabolomics to capture the complexity of metabolic processes within cancer cells (Buescher
387 and Driggers, 2016). Recent initiatives are the Recon1, Edinburgh Human Metabolic Network
388 (EHMN), and Human1 projects (Brunk et al., 2018; Ma et al., 2007; Robinson et al., 2020).
389 However, challenges for these initiatives lie in the limited set of samples for which these high-
390 dimensional multi-omics features are available and the use of predominantly cell line samples.
391 Paired datasets on a large scale are needed to unleash the full potential of such an integrated
392 approach.

393 To facilitate the use of our transcriptional metabolic landscape, we have provided access
394 to all data via a web portal (www.themetaboliclandscapeofcancer.com). In this portal, users can
395 explore genes, metabolic processes, and tissue types of interest. We invite researchers and
396 clinicians to use this portal as a guide to the metabolic transcriptome in cancer or as a starting
397 point for further research into cancer metabolism.

398

399 **MATERIALS AND METHODS**

400 **Resource availability**

401 Further information and requests for resources should be directed to the Lead Contact, Rudolf
402 S.N. Fehrmann (r.s.n.fehrmann@umcg.nl).
403
404 **Data and code availability**
405 Data can be explored at http://themetaboliclandscapeofcancer.com. Code is available at
406 github.com/MetabolicLandscape/
407
408 **Data acquisition**
409 A detailed description of the data acquisition of the four datasets has been described previously
410 (Bhattacharya et al., 2020). In short, the GEO dataset contained microarray expression data
411 generated with Affymetrix HG-U133 Plus 2.0 (accession number GPL570). A two-step search
412 strategy was applied to select healthy or cancer tissue samples – automatic filtering on keywords
413 followed by manual curation. Samples from cell lines, cultured human biopsies, and animal-
414 derived tissue were excluded. The TCGA dataset contained the preprocessed and normalized level
415 3 RNA-seq (version 2) data for 34 cancer datasets available at the Broad GDAC Firehose portal
416 (https://gdac.broadinstitute.org/). The profiles in the CCLE dataset were generated with
417 Affymetrix HG-U133 Plus 2.0. The CCLE project conducted a detailed genetic characterization of
418 a large panel of human cancer cell lines. Expression data within the CCLE project was generated
419 with Affymetrix HG-U133 Plus 2.0. The GDSC dataset contained expression data generated with
420 Affymetrix HG-U219. The GDSC project aims to identify molecular features of cancer that predict
421 response to anti-cancer drugs.
422

423 **Preprocessing, normalization, and quality control**

424 A more detailed description has been provided previously (Bhattacharya et al., 2020). In short,
425 preprocessing and aggregation of raw expression data (CEL files) within the GEO dataset, CCLE
426 dataset, and GDSC dataset was performed according to the robust multi-array average algorithm
427 RMAExpress (version 1.1.0). Quality control was performed on the GEO dataset, CCLE dataset,
428 and GDSC dataset separately with principal component analysis (PCA). Duplicate CEL files were
429 removed by generating a message-digest algorithm 5 (MD5) hash for each CEL file. The expression
430 levels for each probeset (in the GEO dataset, CCLE dataset, and GDSC dataset) or gene (in the
431 TCGA dataset) were standardized to a mean of zero and variance of one to remove probeset-
432 specific or gene-specific variability in the datasets.

433

434 **Consensus independent component analysis**

435 We used consensus independent component analysis (c-ICA) to segregate the average gene
436 expression patterns of complex biopsies into statistically independent transcriptomic
437 components. The input gene expression dataset was preprocessed using whitening
438 transformation, making all profiles uncorrelated and giving them a variance of one. Next, ICA was
439 performed on the whitened dataset using the FastICA algorithm, resulting in the extraction of
440 estimated sources (ESs) and a mixing matrix (MM). The number of principal components which
441 captured 90% of the variance seen in the whitened dataset was chosen as the number of ESs to
442 extract. Each ES contains all genes with a specific weight. This weight represents the direction and
443 magnitude of the influence of an underlying transcriptional regulatory process on that gene
444 expression level. The MM contains the coefficients of ESs in each sample, representing the activity

445 of an ES in the corresponding sample. We performed 25 ICA runs with different random
446 initialization weight factors to assess the robustness of the ESs and exclude ICA results derived
447 from convergence at local solutions. ESs extracted from these runs were clustered together if the
448 absolute value of the Pearson correlation between them was > 0.9. We calculated consensus
449 transcriptional components (TCs) by taking the mean vector of weights in the co-clustering ESs.
450 We considered a consensus TC robust when clustering included individual TCs from > 50% of the
451 runs. The consensus TCs, in combination with the original input expression profiles, were used to
452 obtain the consensus mixing matrix (MM) with the individual activity scores of the consensus TCs
453 in each sample via matrix inversion.
454
455 **Identification of transcriptional components enriched for metabolic processes**
456 First, we selected gene sets defining metabolic processes from five gene set collections obtained
457 from the Molecular Signatures Database (MSigDb version 6.1); BioCarta, Gene Ontology –
458 Biological Process (GO-BP), Gene Ontology – Molecular Function (GO-MF), KEGG, and Reactome.
459 From BioCarta, gene sets were selected manually based on their title. Selected gene sets
460 described metabolic pathways or regulatory pathways regulated by metabolic processes. From
461 GO-BP, all gene sets were selected that contained the motif 'METABOLIC_PROCESS' in the title.
462 Also, gene sets that included the name of a metabolite or class of metabolites combined with the
463 motif'_TRANSPORT' in the title were selected.
464 Furthermore, gene sets not containing these title motifs but associated with (cancer)
465 metabolism were manually selected based on metabolic pathway names. From GO-MF, all gene
466 sets were selected that included the name of a metabolite or class of metabolites combined with

467 the motif '_ACTIVITY' or '_BINDING' in the title. From KEGG, all gene set containing the motif
468 "METABOLISM" or "BIOSYNTHESIS" in the title in combination with the name of a known
469 metabolic route was selected. Furthermore, gene sets concerning metabolism-related regulatory
470 pathways were chosen based on their titles. From Reactome, all gene set that falls within the
471 hierarchy of the "Metabolism"-pathways were selected (see reactome.org/PathwayBrowser).
472 The metabolism of Abacavir was not included. A complete list of all metabolic gene sets selected
473 is presented in **Table S1**.

474 To identify transcriptional components enriched for metabolic processes, gene set
475 enrichment analysis (GSEA) was performed using the selected metabolic gene sets. Enrichment
476 of each metabolic gene set was tested according to the two-sample Welch's t-test for unequal
477 variance between the metabolic set of genes under investigation versus the set of genes that was
478 not under investigation. To compare gene sets of different sizes, we transformed Welch's t
479 statistic to a Z-score.

480 A biological process can be captured by multiple gene sets in several gene set collections.
481 Therefore, it is possible that within the selection of 608 gene sets, multiple gene sets describe the
482 same metabolic process. These will then show a similar pattern in gene set enrichment scores of
483 transcriptional components. To reduce this redundancy, consensus clustering was performed
484 gene set-wise on the GSEA data for the GEO, TCGA, CCLE, and GDSC datasets. Consensus
485 clustering was performed using the ConsensusClusterPlus-package (v1.51.1) within R, using the
486 default hierarchical clustering algorithm and Pearson correlation distance, a maximum amount of
487 clusters (maxK) of 150, 2000 resamplings (reps), with 80% row and 80% column resampling
488 (pFeature and pItem, respectively). The optimal number of clusters (*k)* was determined as the *k*

489 at which the relative change in area under the CDF curve was minimized (<0.01). This resulted in
490 a *k* of 50 clusters (**Figure S1**).
491 The 50 clusters of gene sets were subsequently used to select transcriptional components
492 based on their enrichment for metabolic processes. Per gene set cluster, the three TCs with the
493 highest absolute enrichment score for any gene set in that cluster were selected. In addition to
494 this, the three TCs with the highest absolute mean enrichment score for all gene sets in that
495 cluster were selected. The selected TCs were then referred to as metabolic Transcriptional
496 Components (mTCs). In the end, four different sets of mTCs were identified (GEO mTCs, TCGA
497 mTCs, CCLE mTCs, GDSC mTCs)
498
499 **Approximation of batch effects and tissue specificity of mTCs**
500 First, the explained variance of every component from the perspective of a sample (as a
501 percentage) was estimated using the squares of the mixing matrix weights of a sample divided by
502 the sum of the squares. This percentage explained variance matrix for samples was then
503 summarized into a mean explained variance for studies by summarizing samples belonging to the
504 same study (through the annotated GEO series accession number or TCGA tissue source site
505 code). In the figures, only the highest explained variance available for any study is given. Similarly,
506 tissue specificity was approximated by calculating the mean explained variance for tissue types
507 by summarizing samples belonging to the same tissue subtype.
508
509 **Pair-wise gene-level correlations of mTCs between datasets**

510 To correlate two mTCs of different datasets, the subset of genes with an absolute weight
511 higher than 3 in two mTCs was selected. Then, the overlap between these two sets of top genes
512 was determined. Using the gene weights of the overlapping genes in both mTCs, pair-wise
513 correlations were calculated. Specifically, Spearman correlations were performed in R using the
514 *pspearman*-package (v0.3-0) in R, with a t-distribution approximation to determine the P-value.
515 As the number of genes with an absolute weight above 3 was different for every mTC, the size of
516 the overlap in genes between two mTCs changed. The significance of the Spearman correlation
517 found between two mTCs, therefore, was dependent on the number of overlapping genes. Hence,
518 the significance of the found size of the overlap in genes between mTCs should be determined.
519 To this end, for a pair of mTCs, two sets of random gene identifiers were selected from all possible
520 gene identifiers. The amount of randomly selected genes per set corresponded to the number of
521 genes with a weight >3 in both mTCs. Subsequently, the overlap in gene identifiers between the
522 two random sets of gene identifiers was determined. By repeating this 10,000 times, the chance
523 of finding a given overlap between two sets of genes could be determined.
524 Ultimately, mTCs were said to be concordant when their correlation was > 0.5, with a P-
525 value < 0.05, given that there was a significant overlap in genes (P-value of overlap <0.05).
526
527 **Clustering of Metabolic Transcriptional Components, Genes and Samples**
528 For each of the four datasets, the mixing matrix (MM) containing activity scores was clustered
529 both on samples and mTCs. To this end, hierarchical clustering was performed using ward-D2 as
530 the method and 1-cor(data) as the distance function. Heatmaps were created using R's *gplots*
531 package (v3.0.1). Based on the MM clustering for every dataset, metabolic subtypes were

532 defined. To determine the sizes of clusters of samples that would make up a metabolic subtype,
533 the dendrograms resulting from hierarchical clustering of the samples were systematically cut at
534 dissimilarity values ranging from 0.0 to 8.0 with increments of 0.2. For each of the four datasets
535 GEO, TCGA, CCLE, and GDSC, the cutoff was chosen at the dendrogram height at which the
536 smallest cluster reached a size of 50 samples (**Figure S6**).
537

538 **CIBERSORT**

539 Relative and absolute immune fractions for 22 immune cell types were estimated for all samples
540 in GEO and TCGA datasets using the CIBERSORT algorithm running with default parameters, 1000
541 permutations, and selecting 'absolute nosumto1' as output. This output was then associated with
542 the activity of the mTCs through spearman correlation.
543

544 **Statistical Analyses**

545 Univariate OS on breast cancer samples from GEO and univariate DRFS analyses on melanoma
546 samples from TCGA were performed using a cox regression model through *survminer* (v0.4.3) and
547 *survival* (v2.43-3) packages in R. Confidence intervals were set at 0.95. Significance was tested
548 through the Log Rank test. Scripts are available at github.com/**MetabolicLandscape/.** Pearson
549 correlations were performed in R using the cor.test()-function from the *stats* package (v.3.5.1).
550 Spearman correlations and the corresponding exact P-values were calculated using the
551 *pspearman*-package (v0.3-0) in R, with a t-distribution as an approximation.
552

553 **ACKNOWLEDGMENTS**

554 This research was supported by grants awarded by the Young Academy Groningen (to R.S.N.F.,
555 M.T.C.W., and V.C.L.), the Netherlands Organization for Scientific Research (NWO-VENI grant 916-
556 16025 to R.S.N.F), the Dutch Cancer Society (RUG 2013-5960 to R.S.N.F, RUG 2014-6691 to S.J.,
557 Young Investigator Grant 10913/2017-1 to M.J.), the European Union through the Rosalind
558 Franklin Fellowship (COFUND project 600211 to M.T.C.W.), and a grant from the Hanarth Fonds,
559 The Netherlands (2019N1552 to R.S.N.F).

560

561 **AUTHOR CONTRIBUTIONS**

562 R.S.N.F, V.C.L., C.G.U, and A.B collected and compiled the data. R.S.N.F., V.C.L., and C.G.U
563 performed data analyses. All authors contributed to the data interpretation, writing of the
564 manuscript, and the final decision to submit the manuscript.

565

566 **COMPETING INTERESTS**

567 All authors declare no competing interests.

568 **REFERENCES**

569 • Andrejeva, G., and Rathmell, J.C. (2017). Similarities and Distinctions of Cancer and Immune
570 Metabolism in Inflammation and Tumors. Cell Metab *26*, 49-70.

571 • Bantug, G.R., Galluzzi, L., Kroemer, G., and Hess, C. (2018). The spectrum of T cell metabolism
572 in health and disease. Nat Rev Immunol *18*, 19-34.

573 • Barret, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A.,
574 Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional
575 genomics data sets - update. Nucleic Acids Research *41*, D991-D995.

576 • Barretina, J., Caponigro, G., Stransky, N., K., V., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J.,
577 Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive
578 modelling of anticancer drug sensitivity. Nature *483*, 603-607.

579 • Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., Strathdee, C.A.,
580 Dempster, J., Lyons, N.J., Burns, R., et al. (2018). Genetic and transcriptional evolution alters
581 cancer cell line drug response. Nature *560*, 325-330.

582 • Benita, Y., Kikuchi, H., Smith, A.D., Zhang, M.Q., Chung, D.C., and Xavier, R.J. (2009). An
583 integrative genomics approach identifies Hypoxia Inducible Factor-1 (HIF-1)-target genes that
584 form the core response to hypoxia. Nucleic Acids Research *37*, 4587-4602.

585 • Bense, R.D., Sotiriou, C., Piccart-Gebhart, M.J., Haanen, J., van Vugt, M., de Vries, E.G.E.,
586 Schroder, C.P., and Fehrmann, R.S.N. (2017). Relevance of Tumor-Infiltrating Immune Cell
587 Composition and Functionality for Disease Outcome in Breast Cancer. J Natl Cancer Inst *109*.

588 • Bhattacharya, A., Bense, R.D., Urzua-Traslavina, C.G., de Vries, E.G.E., van Vugt, M., and
589 Fehrmann, R.S.N. (2020). Transcriptional effects of copy number alterations in a large set of
590 human cancers. Nat Commun *11*, 715.
591 • Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Perez, C., Lopez-
592 Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., et al. (2014). Independent component
593 analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into
594 luminal and basal subtypes. Cell Rep *9*, 1235-1245.
595 • Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Drager, A., Mih, N., Gatto, F., Nilsson, A.,
596 Gonzalez, G.A.P., Aurich, M.K., et al. (2018). Recon3D enables a three-dimensional view of
597 gene variation in human metabolism. Nat Biotechnol *36*, 272-+.
598 • Buescher, J.M., and Driggers, E.M. (2016). Integration of omics: more than the sum of its parts.
599 Cancer Metab *4*, 4.
600 • Cappelletti, V., Iorio, E., Miodini, P., Silvestri, M., Dugo, M., and Daidone, M.G. (2017).
601 Metabolic Footprints and Molecular Subtypes in Breast Cancer. Dis Markers *2017*, 7687851.
602 • Chen, B., Khodadoust, M.S., Liu, C.L., Newman, A.M., and Alizadeh, A.A. (2018). Profiling
603 Tumor Infiltrating Immune Cells with CIBERSORT. Methods Mol Biol *1711*, 243-259.
604 • Copple, B.L., Bai, S., Burgoon, L.D., and Moon, J.O.K. (2012). Hypoxia-inducible Factor-1α
605 Regulates Expression of Genes in Hypoxic Hepatic Stellate Cells Important for Collagen
606 Deposition and Angiogenesis. Liver int. *31*, 230-244.
607 • Cubuk, C., Hidalgo, M.R., Amadoz, A., Pujana, M.A., Mateo, F., Herranz, C., Carbonell-
608 Caballero, J., and Dopazo, J. (2018). Gene Expression Integration into Pathway Modules
609 Reveals a Pan-Cancer Metabolic Landscape. Cancer Res *78*, 6059-6072.

610 • Desvergne, B., Michalik, L., and Wahli, W. (2006). Transcriptional Regulation of Metabolism.
611 Physiol Rev *86*, 465-514.
612 • Fraietta, J.A., Lacey, S.F., Orlando, E.J., Pruteanu-Malinici, I., Gohil, M., Lundh, S., Boesteanu,
613 A.C., Wang, Y., O'Connor, R.S., Hwang, W.T., et al. (2018). Determinants of response and
614 resistance to CD19 chimeric antigen receptor (CAR) T cell therapy of chronic lymphocytic
615 leukemia. Nat Med *24*, 563-571.
616 • Gaspar, N., Sharp, S.Y., Pacey, S., Jones, C., Walton, M., Vassal, G., Eccles, S., Pearson, A., and
617 Workman, P. (2009). Acquired resistance to 17-allylamino-17-demethoxygeldanamycin (17-
618 AAG, tanespimycin) in glioblastoma cells. Cancer Res *69*, 1966-1975.
619 • Ghaffari, P., Mardinoglu, A., and Nielsen, J. (2015). Cancer Metabolism: A Modeling
620 Perspective. Front Physiol *6*, 382.
621 • Goto, S., Kamada, K., Soh, Y., Ihara, Y., and Kondo, T. (2002). Significance of nuclear
622 glutathione S-transferase pi in resistance to anti-cancer drugs. Jpn J Cancer Res *93*, 1047-1056.
623 • Hakimi, A.A., Reznik, E., Lee, C.H., Creighton, C.J., Brannon, A.R., Luna, A., Aksoy, B.A., Liu,
624 E.M., Shen, R., Lee, W., et al. (2016). An Integrated Metabolic Atlas of Clear Cell Renal Cell
625 Carcinoma. Cancer Cell *29*, 104-116.
626 • Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*,
627 646-674.
628 • Heibein, A.D., Guo, B., Sprowl, J.A., Maclean, D.A., and Parissenti, A.M. (2012). Role of aldo-
629 keto reductases and other doxorubicin pharmacokinetic genes in doxorubicin resistance, DNA
630 binding, and subcellular localization. BMC Cancer *12*, 381.

631 • Huang, M., Shen, A.J., Ding, J., and Geng, M.Y. (2014). Molecularly targeted cancer therapy:
632 some lessons from the past decade. Trends Pharmacol Sci *35*, 41-50.
633 • Hynds, R.E., Vladimirou, E., and Janes, S.M. (2018). The Secret Lives of Cancer Cell Lines. Dis
634 Model Mech *11*, 1-5.
635 • Jiang, G.L., Zhang, S.J., Yazdanparast, A., Li, M., Pawar, A.V., Liu, Y.L., Inavolu, S.M., and Cheng,
636 L.J. (2016). Comprehensive comparison of molecular portraits between cell lines and tumors
637 in breast cancer. Bmc Genomics *17*.
638 • Kim, Y.S., Seo, H.W., and Jung, G. (2015). Reactive oxygen species promote heat shock protein
639 90-mediated HBV capsid assembly. Biochem Biophys Res Commun *457*, 328-333.
640 • Kong, W., Vanderburg, C.R., Gunshin, H., Rogers, J.T., and Huang, X. (2008). A review of
641 independent component analysis application to microarray gene expression data.
642 Biotechniques *45*, 501-520.
643 • Le Bourgeois, T., Strauss, L., Aksoylar, H.-I., Daneshmandi, S., Seth, P., Patsoukis, N., and
644 Boussiotis, V.A. (2018). Targeting T Cell Metabolism for Improvement of Cancer
645 Immunotherapy. Frontiers in Oncology *8*, 1-17.
646 • Lefort, N., Yi, Z., Bowen, B., Glancy, B., De Filippis, E.A., Mapes, R., Hwang, H., Flynn, C.R.,
647 Willis, W.T., Civitarese, A., et al. (2009). Proteome profile of functional mitochondria from
648 human skeletal muscle using one-dimensional gel electrophoresis and HPLC-ESI-MS/MS. J
649 Proteomics *72*, 1046-1060.
650 • Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. (2007). The
651 Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol
652 *3*, 135.

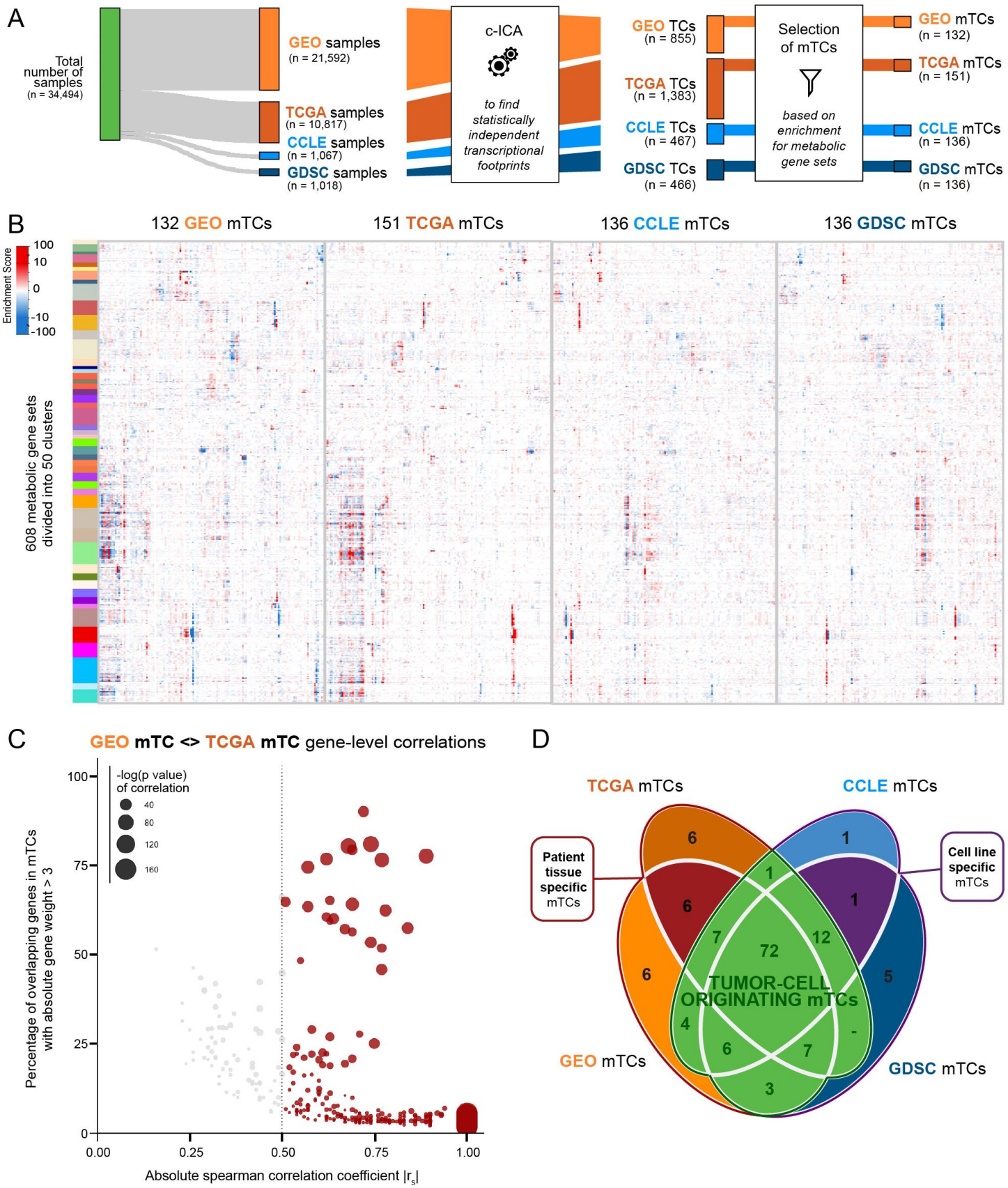
653 • Martin-Martin, N., Carracedo, A., and Torrano, V. (2018). Metabolism and Transcription in
654 Cancer: Merging Two Classic Tales. Frontiers in Cell and Developmental Biology *5*, 1-8.
655 • Matsunaga, T., Okumura, N., Saito, H., Morikawa, Y., Suenami, K., Hisamatsu, A., Endo, S., and
656 Ikari, A. (2020). Significance of aldo-keto reductase 1C3 and ATP-binding cassette transporter
657 B1 in gain of irinotecan resistance in colon cancer cells. Chem Biol Interact *332*, 109295.
658 • O'Sullivan, D., and Pearce, E.L. (2015). Targeting T cell metabolism for therapy. Trends
659 Immunol *36*, 71-80.
660 • O'Sullivan, D., Sanin, D.E., Pearce, E.J., and Pearce, E.L. (2019). Metabolic interventions in the
661 immune response to cancer. Nat Rev Immunol *19*, 324-335.
662 • Oh, M.H., Sun, I.H., Zhao, L., Leone, R.D., Sun, I.M., Xu, W., Collins, S.L., Tam, A.J., Blosser, R.L.,
663 Patel, C.H., et al. (2020). Targeting glutamine metabolism enhances tumor specific immunity
664 by modulating suppressive myeloid cells. J Clin Invest.
665 • Patel, C.H., Leone, R.D., Horton, M.R., and Powell, J.D. (2019). Targeting metabolism to
666 regulate immune responses in autoimmunity and cancer. Nat Rev Drug Discov *18*, 669-688.
667 • Patel, C.H., and Powell, J.D. (2017). Targeting T cell metabolism to regulate T cell activation,
668 differentiation and function in disease. Curr Opin Immunol *46*, 82-88.
669 • Pavlova, N.N., and Thompson, C.B. (2016). The Emerging Hallmarks of Cancer Metabolism.
670 Cell Metab *23*, 27-47.
671 • Peng, X., Chen, Z., Farshidfar, F., Xu, X., Lorenzi, P.L., Wang, Y., Cheng, F., Tan, L., Mojumdar,
672 K., Du, D., et al. (2018). Molecular Characterization and Clinical Relevance of Metabolic
673 Expression Subtypes in Human Cancers. Cell Reports *23*, 255-269.

674 • Quail, D.F., and Joyce, J.A. (2013). Microenvironmental regulation of tumor progression and
675 metastasis. Nat Med *19*, 1423-1437.
676 • Reznik, E., Luna, A., Aksoy, B.A., Liu, E.M., La, K., Ostrovnaya, I., Creighton, C.J., Hakimi, A.A.,
677 and Sander, C. (2018). A Landscape of Metabolic Variation across Tumor Types. Cell Syst *6*,
678 301-313 e303.
679 • Robinson, J.L., Kocabas, P., Wang, H., Cholley, P.E., Cook, D., Nilsson, A., Anton, M., Ferreira,
680 R., Domenzain, I., Billa, V., et al. (2020). An atlas of human metabolism. Science Signaling *13*,
681 eeaz1482.
682 • Rosario, S.R., Long, M.D., Affronti, H.C., Rowsam, A.M., Eng, K.H., and Smiraglia, D.J. (2018).
683 Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome
684 Atlas. Nat Commun *9*, 5330.
685 • Schnoor, M., Cullen, P., Lorkowski, J., Stolle, K., Robenek, H., Troyer, D., Rauterberg, J., and
686 Lorkowski, S. (2008). Production of type VI collagen by human macrophages: a new dimension
687 in macrophage functional heterogeneity. J Immunol *180*, 5707-5719.
688 • Selga, E., Noe, V., and Ciudad, C.J. (2008). Transcriptional regulation of aldo-keto reductase
689 1C1 in HT29 human colon cancer cells resistant to methotrexate: role in the cell cycle and
690 apoptosis. Biochem Pharmacol *75*, 414-426.
691 • Serrano-Carbajal, E.A., Espinal-Enriquez, J., and Hernandez-Lemus, E. (2020). Targeting
692 Metabolic Deregulation Landscapes in Breast Cancer Subtypes. Front Oncol *10*, 97.
693 • Shah, R., and Chen, S. (2020). Metabolic Signaling Cascades Prompted by Glutaminolysis in
694 Cancer. Cancers (Basel) *12*.

695 • Sørensen, B.S., Knudsen, A., Wittrup, C.F., Nielsen, S., Aggerholm-Pedersen, N., Busk, M.,
696 Horsman, M., Høyer, M., Bouchelouche, P.N., Overgaard, J., et al. (2015). The usability of a
697 15-gene hypoxia classifier as a universal hypoxia profile in various cancer cell types.
698 Radiotherapy and Oncology *116*, 346-351.
699 • Sukumar, M., Kishton, R.J., and Restifo, N.P. (2017). Metabolic reprograming of anti-tumor
700 immunity. Curr Opin Immunol *46*, 14-22.
701 • Tang, X., Lin, C.C., Spasojevic, I., Iversen, E.S., Chi, J.T., and Marks, J.R. (2014). A joint analysis
702 of metabolomics and genetics of breast cancer. Breast Cancer Res *16*, 415.
703 • Tang, Z., Xu, Z., Zhu, X., and Zhang, J. (2021). New insights into molecules and pathways of
704 cancer metabolism and therapeutic implications. Cancer Commun (Lond) *41*, 16-36.
705 • Vaage, J., and Harlos, J.P. (1991). Collagen production by macrophages in tumour
706 encapsulation and dormancy. Br J Cancer *63*, 758-762.
707 • Vazquez, A., Kamphorst, J.J., Markert, E., Schug, Z.T., Tardito, S., and Gottlieb, E. (2016).
708 Cancer metabolism at a glance. J Cell Sci *129*, 3367-3373.
709 • Viale, A., and Draetta, G.F. (2016). Metabolic features of cancer treatment resistance. Resent
710 Results in Cancer Res *207*, 135-156.
711 • Vincent, K.M., and Postovit, L.M. (2017). Investigating the utility of human melanoma cell lines
712 as tumour models. Oncotarget *8*, 10498-10509.
713 • Wagle, N., Van Allen, E.M., Treacy, D.J., Frederick, D.T., Cooper, Z.A., Taylor-Weiner, A.,
714 Rosenberg, M., Goetz, E.M., Sullivan, R.J., Farlow, D.N., et al. (2014). MAP kinase pathway
715 alterations in BRAF-mutant melanoma patients with acquired resistance to combined
716 RAF/MEK inhibition. Cancer Discov *4*, 61-68.

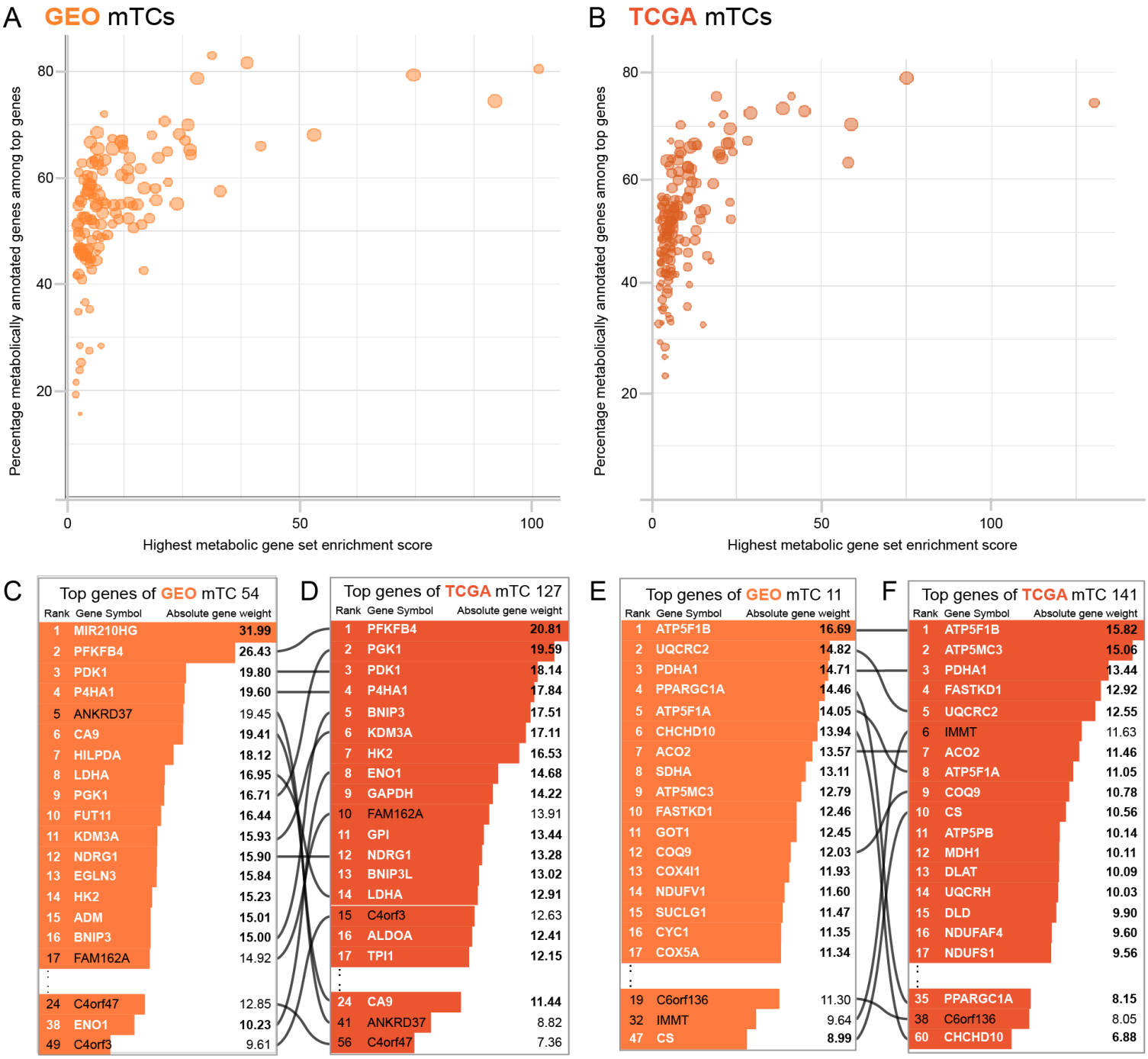
717 • Walling, J. (2006). From methotrexate to pemetrexed and beyond. A review of the
718 pharmacodynamic and clinical properties of antifolates. Invest New Drug *24*, 37-77.
719 • Wang, L., Leite de Oliveira, R., Huijberts, S., Bosdriesz, E., Pencheva, N., Brunen, D., Bosma, A.,
720 Song, J.Y., Zevenhoven, J., Los-de Vries, G.T., et al. (2018). An Acquired Vulnerability of Drug-
721 Resistant Melanoma with Therapeutic Potential. Cell *173*, 1413-1425 e1414.
722 • Wang, R., and Green, D.R. (2012). Metabolic checkpoints in activated T cells. Nat Immunol *13*,
723 907-915.
724 • Wang, R., Zhao, H., Zhang, X., Zhao, X., Song, Z., and Ouyang, J. (2019). Metabolic
725 Discrimination of Breast Cancer Subtypes at the Single-Cell Level by Multiple Microextraction
726 Coupled with Mass Spectrometry. Anal Chem *91*, 3667-3674.
727 • Xiao, Z., Dai, Z., and Locasale, J.W. (2019). Metabolic landscape of the tumor
728 microenvironment at single cell resolution. Nat Commun *10*, 3763.
729 • Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D.,
730 Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a
731 resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research *41*, D955-
732 961.
733 • Ye, I.C., Fertig, E.J., DiGiacomo, J.W., Considine, M., Godet, I., and Gilkes, D.M. (2018).
734 Molecular Portrait of Hypoxia in Breast Cancer: A Prognostic Signature and Novel HIF-
735 Regulated Genes. Mol Cancer Res *16*, 1889-1901.
736 • Zhang, Y., Choksi, S., Chen, K., Pobezinskaya, Y., Linnoila, I., and Liu, Z.G. (2013). ROS play a
737 critical role in the differentiation of alternatively activated macrophages and the occurrence
738 of tumor-associated macrophages. Cell Res *23*, 898-914.

739 **FIGURES**


A
Total number of samples (n = 34,494)
GEO samples (n = 21,592)
TCGA samples (n = 10,817)
CCLE samples (n = 1,067)
GDSC samples (n = 1,018)
c-ICA
to find statistically independent transcriptional footprints
GEO TCs (n = 855)
TCGA TCs (n = 1,383)
CCLE TCs (n = 467)
GDSC TCs (n = 466)
Selection of mTCs
based on enrichment for metabolic gene sets
GEO mTCs (n = 132)
TCGA mTCs (n = 151)
CCLE mTCs (n = 136)
GDSC mTCs (n = 136)
B
132 GEO mTCs
151 TCGA mTCs
136 CCLE mTCs
136 GDSC mTCs
Enrichment Score
100
10
0
-10
-100
608 metabolic gene sets divided into 50 clusters
C
GEO mTC <> TCGA mTC gene-level correlations
-log(p value) of correlation
40
80
120
160
Percentage of overlapping genes in mTCs with absolute gene weight > 3
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
Absolute spearman correlation coefficient |r_s|
D
TCGA mTCs
CCLE mTCs
Patient tissue specific mTCs
Cell line specific mTCs
6
1
1
6
1
7
72
12
6
5
TUMOR-CELL ORIGINATING mTCs
4
6
7
-
3
GEO mTCs
GDSC mTCs

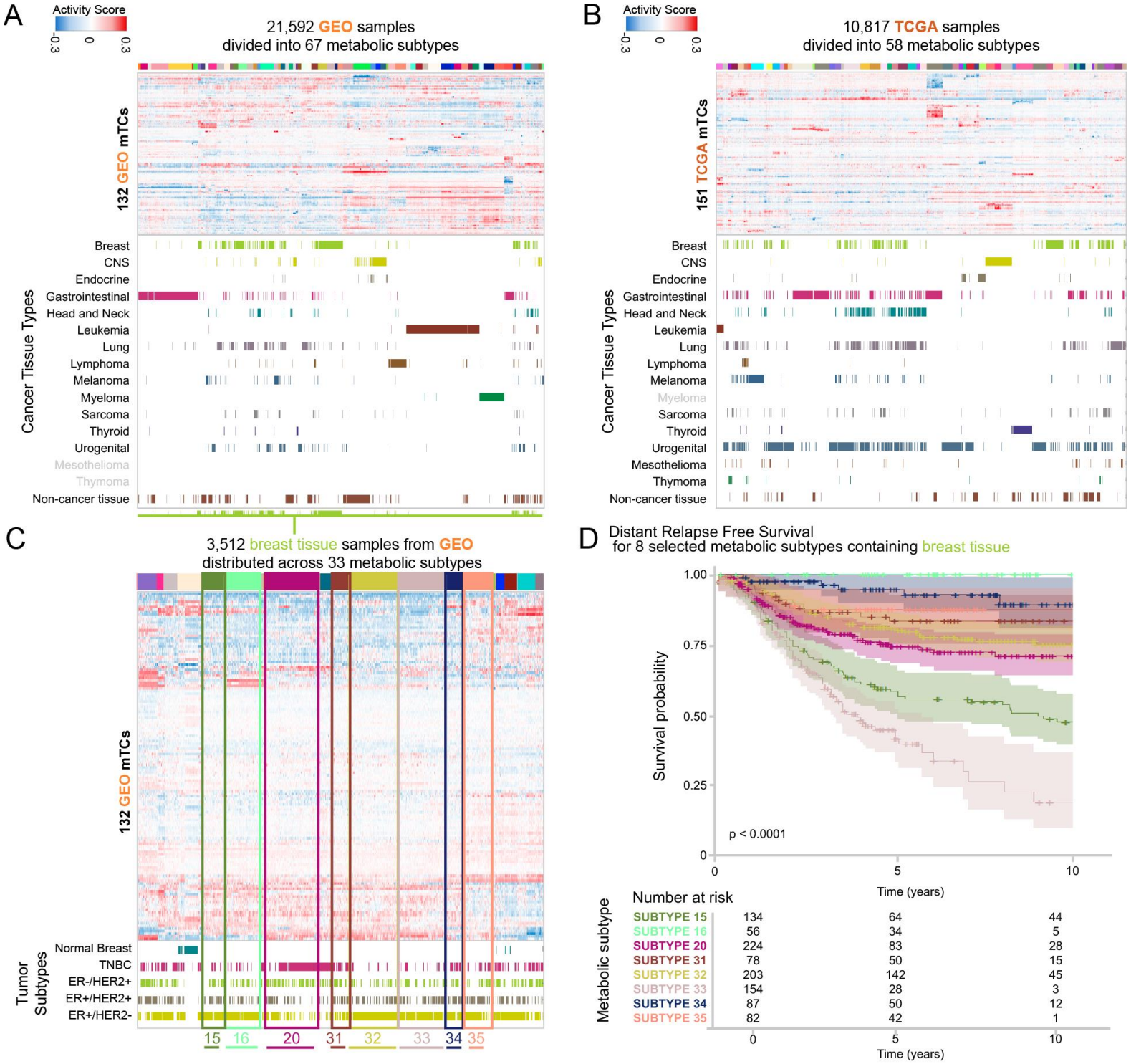
740 **Figure 1 – Identification of metabolic transcriptional components (mTCs). (A)** Workflow for
741 identification of mTCs. Consensus-Independent Component Analysis (c-ICA) applied to identify
742 transcriptional components (TCs). Subsequent systematic selection of TCs enriched for metabolic
743 processes resulted in in 132, 151, 136, and 136 mTCs for the GEO, TCGA, CCLE, and GDSC datasets,
744 respectively. **(B)** Hierarchically clustered heatmaps showing the enrichment of the 608 metabolic
745 gene sets of mTCs identified in GEO, TCGA, CCLE, and GDSC datasets. **(C)** Scatter plot showing
746 absolute spearman correlation coefficients (x-axis), versus the percentage of overlapping top
747 genes (genes with absolute weight >3) between GEO mTCs and TCGA mTCs (y-axis). Only
748 significant pair-wise correlations (with P-values <0.05) are shown. Colored dots show correlations
749 > 0.5, the size of the dots represent the P-value of these spearman correlations. **(D)** Venn diagram
750 quantifying overlap of mTCs between each dataset based on their pair-wise correlations. Two
751 mTCs are counted as shared between datasets, when they have a high absolute spearman
752 correlation ($|r_s|>0.5$). Three groups of (shared) mTCs, mentioned in the text, are designated.

753


A GEO mTCs
B TCGA mTCs
Percentage metabolically annotated genes among top genes
Highest metabolic gene set enrichment score
80
60
40
20
0
50
100
C
Top genes of GEO mTC 54
Rank Gene Symbol Absolute gene weight
1 MIR210HG 31.99
2 PFKFB4 26.43
3 PDK1 19.80
4 P4HA1 19.60
5 ANKRD37 19.45
6 CA9 19.41
7 HILPDA 18.12
8 LDHA 16.95
9 PGK1 16.71
10 FUT11 16.44
11 KDM3A 15.93
12 NDRG1 15.90
13 EGLN3 15.84
14 HK2 15.23
15 ADM 15.01
16 BNIP3 15.00
17 FAM162A 14.92
24 C4orf47 12.85
38 ENO1 10.23
49 C4orf3 9.61
D
Top genes of TCGA mTC 127
Rank Gene Symbol Absolute gene weight
1 PFKFB4 20.81
2 PGK1 19.59
3 PDK1 18.14
4 P4HA1 17.84
5 BNIP3 17.51
6 KDM3A 17.11
7 HK2 16.53
8 ENO1 14.68
9 GAPDH 14.22
10 FAM162A 13.91
11 GPI 13.44
12 NDRG1 13.28
13 BNIP3L 13.02
14 LDHA 12.91
15 C4orf3 12.63
16 ALDOA 12.41
17 TPI1 12.15
24 CA9 11.44
41 ANKRD37 8.82
56 C4orf47 7.36
E
Top genes of GEO mTC 11
Rank Gene Symbol Absolute gene weight
1 ATP5F1B 16.69
2 UQCRC2 14.82
3 PDHA1 14.71
4 PPARGC1A 14.46
5 ATP5F1A 14.05
6 CHCHD10 13.94
7 ACO2 13.57
8 SDHA 13.11
9 ATP5MC3 12.79
10 FASTKD1 12.46
11 GOT1 12.45
12 COQ9 12.03
13 COX4I1 11.93
14 NDUFV1 11.60
15 SUCLG1 11.47
16 CYC1 11.35
17 COX5A 11.34
19 C6orf136 11.30
32 IMMT 9.64
47 CS 8.99
F
Top genes of TCGA mTC 141
Rank Gene Symbol Absolute gene weight
1 ATP5F1B 15.82
2 ATP5MC3 15.06
3 PDHA1 13.44
4 FASTKD1 12.92
5 UQCRC2 12.55
6 IMMT 11.63
7 ACO2 11.46
8 ATP5F1A 11.05
9 COQ9 10.78
10 CS 10.56
11 ATP5PB 10.14
12 MDH1 10.11
13 DLAT 10.09
14 UQCRH 10.03
15 DLD 9.90
16 NDUFAF4 9.60
17 NDUFS1 9.56
35 PPARGC1A 8.15
38 C6orf136 8.05
60 CHCHD10 6.88

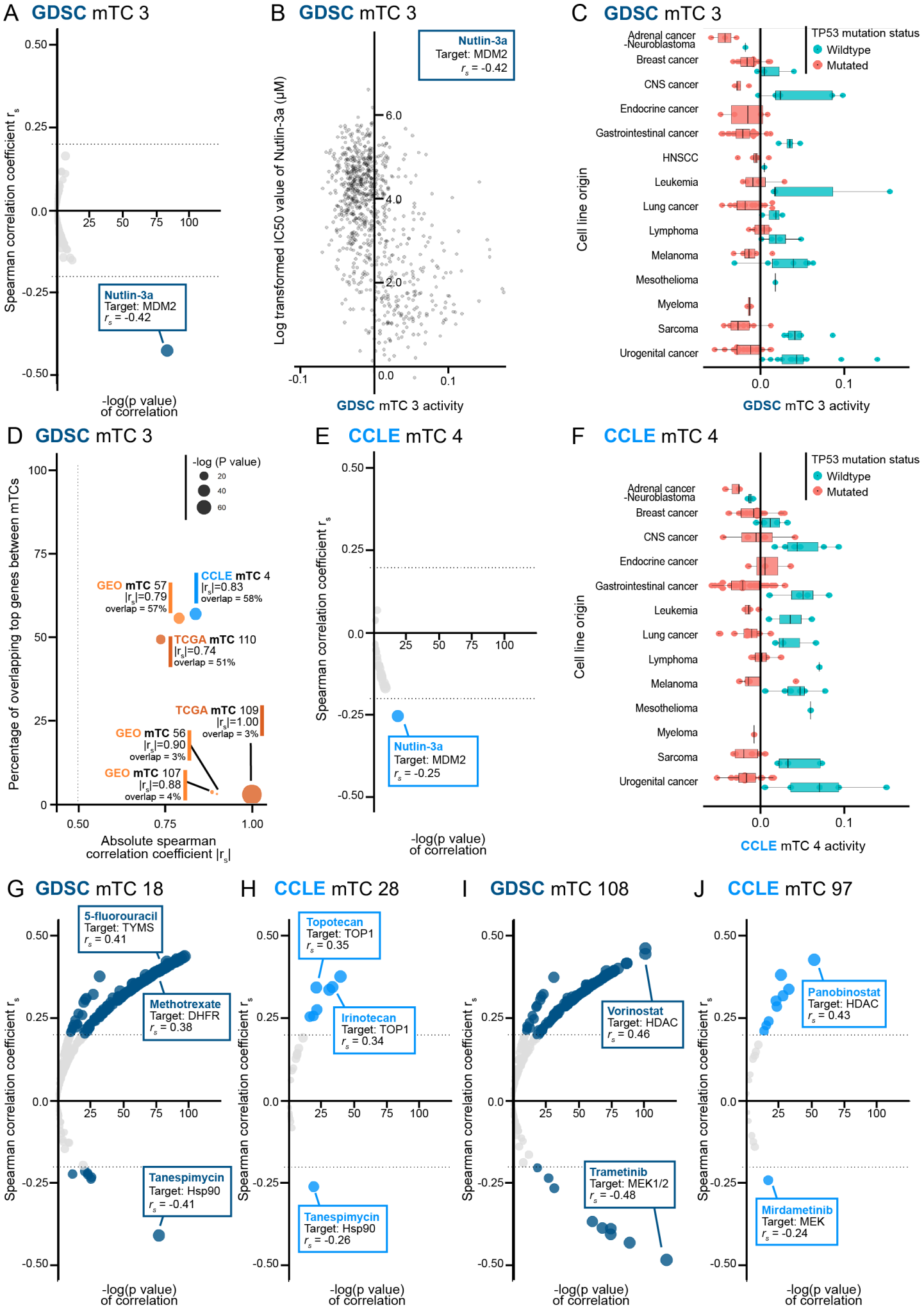
754 **Figure 2** - **Metabolic TCs identify new genes potentially involved in metabolic processes**. **(A-B)**
755 Scatterplots showing the highest metabolic gene set enrichment score for every GEO (A) and
756 TCGA (B) mTC (x-axis) versus the percentage of metabolically annotated genes among the top
757 genes (genes with absolute weight >3) in those mTCs. Size of dots correspond to the absolute
758 amount of metabolically annotated genes in the corresponding mTC. **(C-D)** Top genes in GEO mTC
759 54 and TCGA mTC 127. Text colored white shows genes that are a member of at least one of the
760 608 defined metabolic gene sets. Lines signify genes that are top genes in both GEO and TCGA
761 mTCs. **(E-F)** Top genes in GEO mTC 11 and TCGA mTC 141.

762

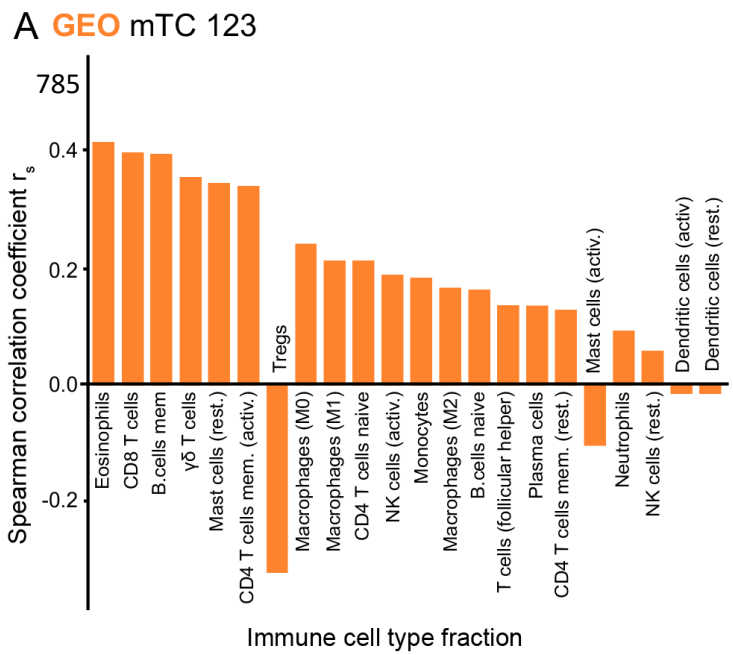

A
Activity Score
-0.3
0
0.3
21,592 GEO samples
divided into 67 metabolic subtypes
132 GEO mTCs
Cancer Tissue Types
Breast
CNS
Endocrine
Gastrointestinal
Head and Neck
Leukemia
Lung
Lymphoma
Melanoma
Myeloma
Sarcoma
Thyroid
Urogenital
Mesothelioma
Thymoma
Non-cancer tissue
B
Activity Score
-0.3
0
0.3
10,817 TCGA samples
divided into 58 metabolic subtypes
151 TCGA mTCs
Cancer Tissue Types
Breast
CNS
Endocrine
Gastrointestinal
Head and Neck
Leukemia
Lung
Lymphoma
Melanoma
Myeloma
Sarcoma
Thyroid
Urogenital
Mesothelioma
Thymoma
Non-cancer tissue
C
3,512 breast tissue samples from GEO
distributed across 33 metabolic subtypes
132 GEO mTCs
Tumor Subtypes
Normal Breast
TNBC
ER-/HER2+
ER+/HER2+
ER+/HER2-
15
16
20
31
32
33
34
35
D
Distant Relapse Free Survival
for 8 selected metabolic subtypes containing breast tissue
Survival probability
1.00
0.75
0.50
0.25
0
p < 0.0001
0
5
10
Time (years)
Number at risk
Metabolic subtype
SUBTYPE 15 134 64 44
SUBTYPE 16 56 34 5
SUBTYPE 20 224 83 28
SUBTYPE 31 78 50 15
SUBTYPE 32 203 142 45
SUBTYPE 33 154 28 3
SUBTYPE 34 87 50 12
SUBTYPE 35 82 42 1
0
5
10
Time (years)

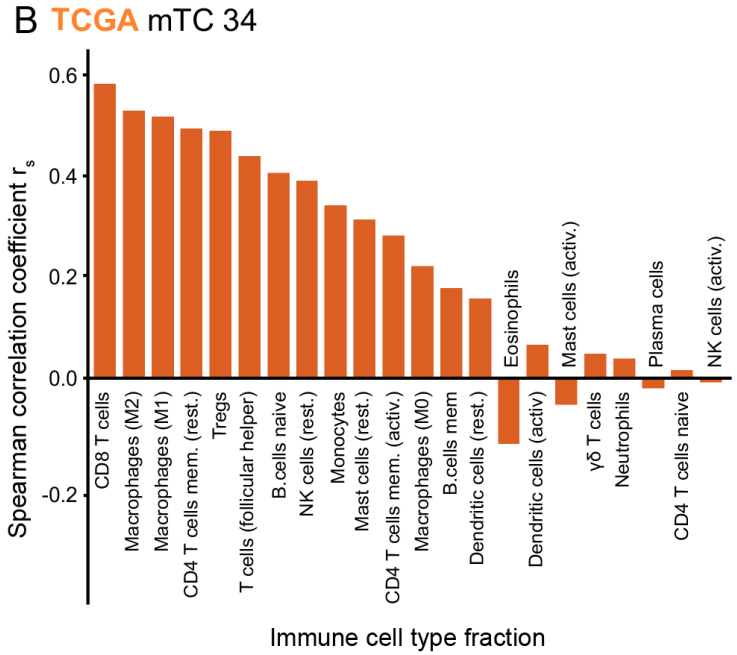
763 **Figure 3 – Clustering activity scores of mTCs reveal multiple metabolic subtypes**

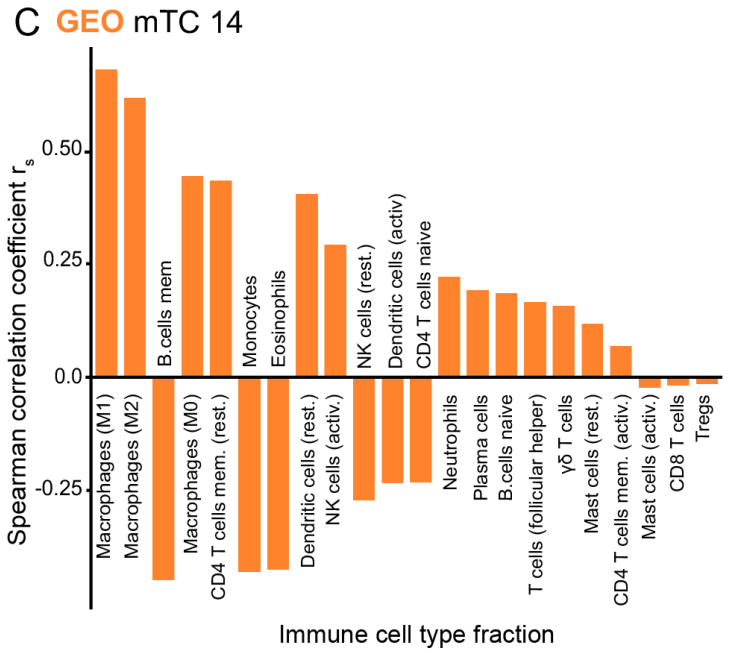
764 **(A)** 21,592 GEO samples were hierarchically clustered based on mTC activity scores and divided
765 into 67 metabolic subtypes. **(B)** 10,817 TCGA samples were hierarchically clustered based on mTC
766 activity scores and divided into 58 metabolic subtypes. **(C)** Metabolic landscape of the subset of
767 breast tissue samples in the GEO dataset. Subtypes with DFS data were selected for survival
768 analysis are highlighted. Grey labels designate tissue types that are present in other datasets, but
769 are not present in the given dataset. **(D)** Distant relapse-free survival of breast cancer patients in
770 the GEO dataset. Patient-derived samples were stratified per metabolic subtype. Kaplan Meier
771 curves are shown with a confidence interval of 0.95.

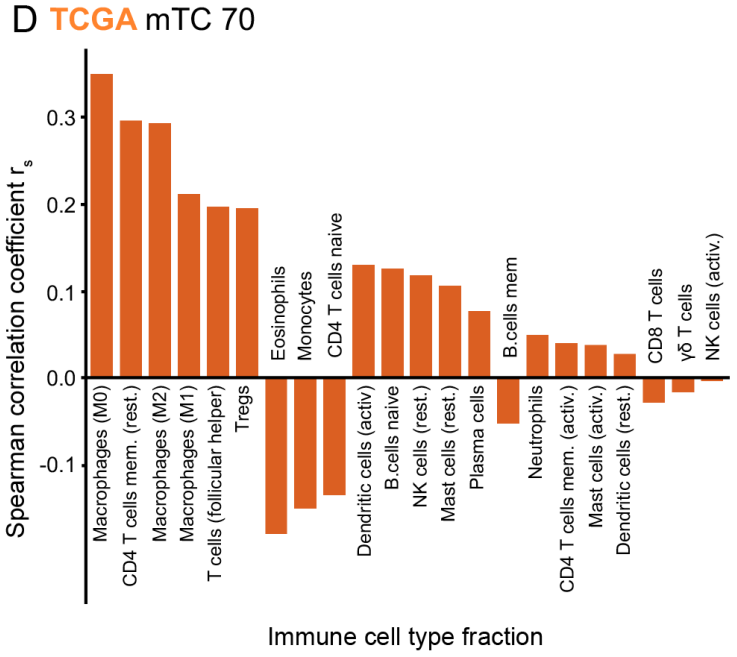

A GDSC mTC 3
Spearman correlation coefficient r_s
-log(p value) of correlation
Nutlin-3a
Target: MDM2
r_s = -0.42
B GDSC mTC 3
Log transformed IC50 value of Nutlin-3a (μM)
GDSC mTC 3 activity
Nutlin-3a
Target: MDM2
r_s = -0.42
C GDSC mTC 3
TP53 mutation status
Wildtype
Mutated
Cell line origin
Adrenal cancer -Neuroblastoma
Breast cancer
CNS cancer
Endocrine cancer
Gastrointestinal cancer
HNSCC
Leukemia
Lung cancer
Lymphoma
Melanoma
Mesothelioma
Myeloma
Sarcoma
Urogenital cancer
GDSC mTC 3 activity
D GDSC mTC 3
-log (P value)
Percentage of overlapping top genes between mTCs
Absolute spearman correlation coefficient $|r_s|$
GEO mTC 57 $|r_s|$=0.79 overlap = 57%
CCLE mTC 4 $|r_s|$=0.83 overlap = 58%
TCGA mTC 110 $|r_s|$=0.74 overlap = 51%
TCGA mTC 109 $|r_s|$=1.00 overlap = 3%
GEO mTC 56 $|r_s|$=0.90 overlap = 3%
GEO mTC 107 $|r_s|$=0.88 overlap = 4%
E CCLE mTC 4
Spearman correlation coefficient r_s
-log(p value) of correlation
Nutlin-3a
Target: MDM2
r_s = -0.25
F CCLE mTC 4
TP53 mutation status
Wildtype
Mutated
Cell line origin
CCLE mTC 4 activity
G GDSC mTC 18
5-fluorouracil
Target: TYMS
r_s = 0.41
Methotrexate
Target: DHFR
r_s = 0.38
Tanespimycin
Target: Hsp90
r_s = -0.41
H CCLE mTC 28
Topotecan
Target: TOP1
r_s = 0.35
Irinotecan
Target: TOP1
r_s = 0.34
Tanespimycin
Target: Hsp90
r_s = -0.26
I GDSC mTC 108
Vorinostat
Target: HDAC
r_s = 0.46
Trametinib
Target: MEK1/2
r_s = -0.48
J CCLE mTC 97
Panobinostat
Target: HDAC
r_s = 0.43
Mirdametinib
Target: MEK
r_s = -0.24

773 **Figure 4 – Associations between mTCs and drug sensitivity for selected examples. (A)** Spearman
774 correlations between drug IC50 values and the activity of GDSC mTC 3 **(B)** Scatter plot showing
775 the association between the (log-transformed) IC50 value of Nutlin-3a and activity of GDSC mTC
776 3 in samples. **(C)** Box plot of activity of GDSC mTC 3 across cell lines, colored for their *TP53*
777 mutation status. **(D)** Pair-wise correlations between GDSC mTC 3 and mTCs from GEO, TCGA and
778 CCLE datasets. Every dot corresponds to an mTC with a correlation to GDSC mTC 3 ≥ 0.5. Dot sizes
779 correspond to the P-value of the spearman correlation coefficient; the y-axis gives the percentage
780 of overlapping top genes between the two mTCs involved in the correlation. **(E)** Spearman
781 correlations between drug IC50 values and the activity of CCLE mTC 4 **(F)** Box plot of activity of
782 CCLE mTC 4 across cell lines, colored for their *TP53* mutation status. **(G-J)** Spearman correlations
783 between drug IC50 values and the activity of GDSC mTC 18, CCLE mTC 28, GDSC mTC 108 and
784 CCLE mTC 97.


A GEO mTC 123
785
Spearman correlation coefficient r_s
0.4
0.2
0.0
-0.2
Eosinophils
CD8 T cells
B.cells mem
γδ T cells
Mast cells (rest.)
CD4 T cells mem. (activ.)
Tregs
Macrophages (M0)
Macrophages (M1)
CD4 T cells naive
NK cells (activ.)
Monocytes
Macrophages (M2)
B.cells naive
T cells (follicular helper)
Plasma cells
CD4 T cells mem. (rest.)
Mast cells (activ.)
Neutrophils
NK cells (rest.)
Dendritic cells (activ)
Dendritic cells (rest.)
Immune cell type fraction


B TCGA mTC 34
Spearman correlation coefficient r_s
0.6
0.4
0.2
0.0
-0.2
CD8 T cells
Macrophages (M2)
Macrophages (M1)
CD4 T cells mem. (rest.)
Tregs
T cells (follicular helper)
B.cells naive
NK cells (rest.)
Monocytes
Mast cells (rest.)
CD4 T cells mem. (activ.)
Macrophages (M0)
B.cells mem
Dendritic cells (rest.)
Eosinophils
Dendritic cells (activ)
Mast cells (activ.)
γδ T cells
Neutrophils
Plasma cells
CD4 T cells naive
NK cells (activ.)
Immune cell type fraction


C GEO mTC 14
Spearman correlation coefficient r_s
0.50
0.25
0.0
-0.25
Macrophages (M1)
Macrophages (M2)
B.cells mem
Macrophages (M0)
CD4 T cells mem. (rest.)
Monocytes
Eosinophils
Dendritic cells (rest.)
NK cells (activ.)
NK cells (rest.)
Dendritic cells (activ)
CD4 T cells naive
Neutrophils
Plasma cells
B.cells naive
T cells (follicular helper)
γδ T cells
Mast cells (rest.)
CD4 T cells mem. (activ.)
Mast cells (activ.)
CD8 T cells
Tregs
Immune cell type fraction


D TCGA mTC 70
Spearman correlation coefficient r_s
0.3
0.2
0.1
0.0
-0.1
Macrophages (M0)
CD4 T cells mem. (rest.)
Macrophages (M2)
Macrophages (M1)
T cells (follicular helper)
Tregs
Eosinophils
Monocytes
CD4 T cells naive
Dendritic cells (activ)
B.cells naive
NK cells (rest.)
Mast cells (rest.)
Plasma cells
B.cells mem
Neutrophils
CD4 T cells mem. (activ.)
Mast cells (activ.)
Dendritic cells (rest.)
CD8 T cells
γδ T cells
NK cells (activ.)
Immune cell type fraction

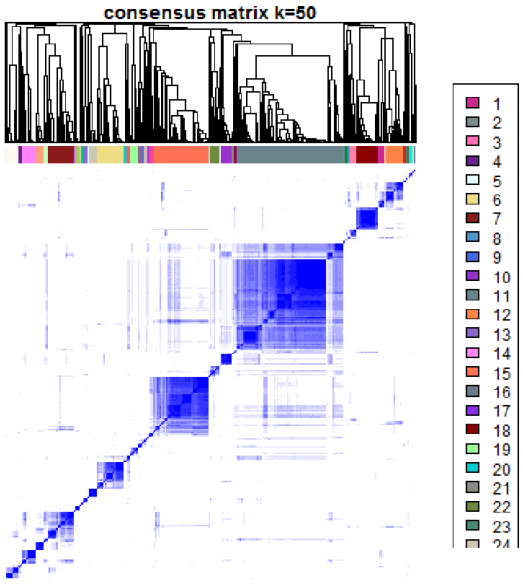
786 **Figure 5 – Associations between mTCs and the composition of the immune tumor**
787 **microenvironment for selected examples. (A-B)** Spearman correlations between CIBERSORT
788 estimated immune cell fractions and the activity of GEO mTC 123 and TCGA mTC 34. **(C-D)**
789 Spearman correlations between CIBERSORT estimated immune cell fractions and the activity of
790 GEO mTC 14 and TCGA mTC 70.

SUPPLEMENTARY INFORMATION

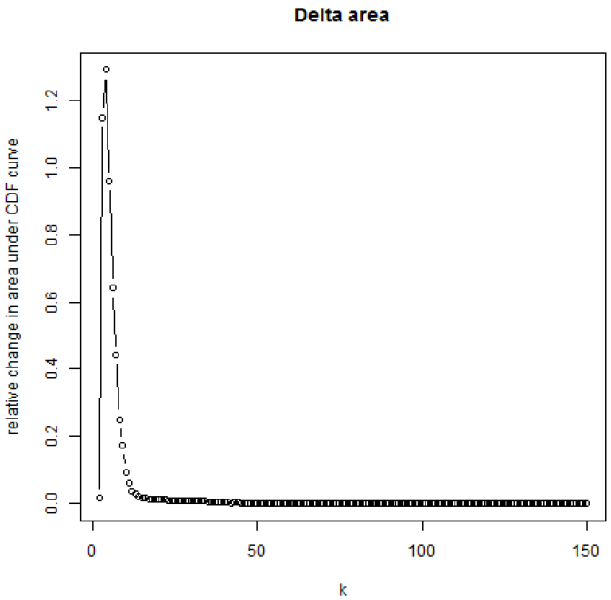
Contains supplementary Figures S1 – S8 and their legends.

To accommodate the editorial process, and due to file constraints, Supplementary Tables S1 – S5 are available as excel files but omitted from the generated composite pdf file.

A


consensus matrix k=50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

B

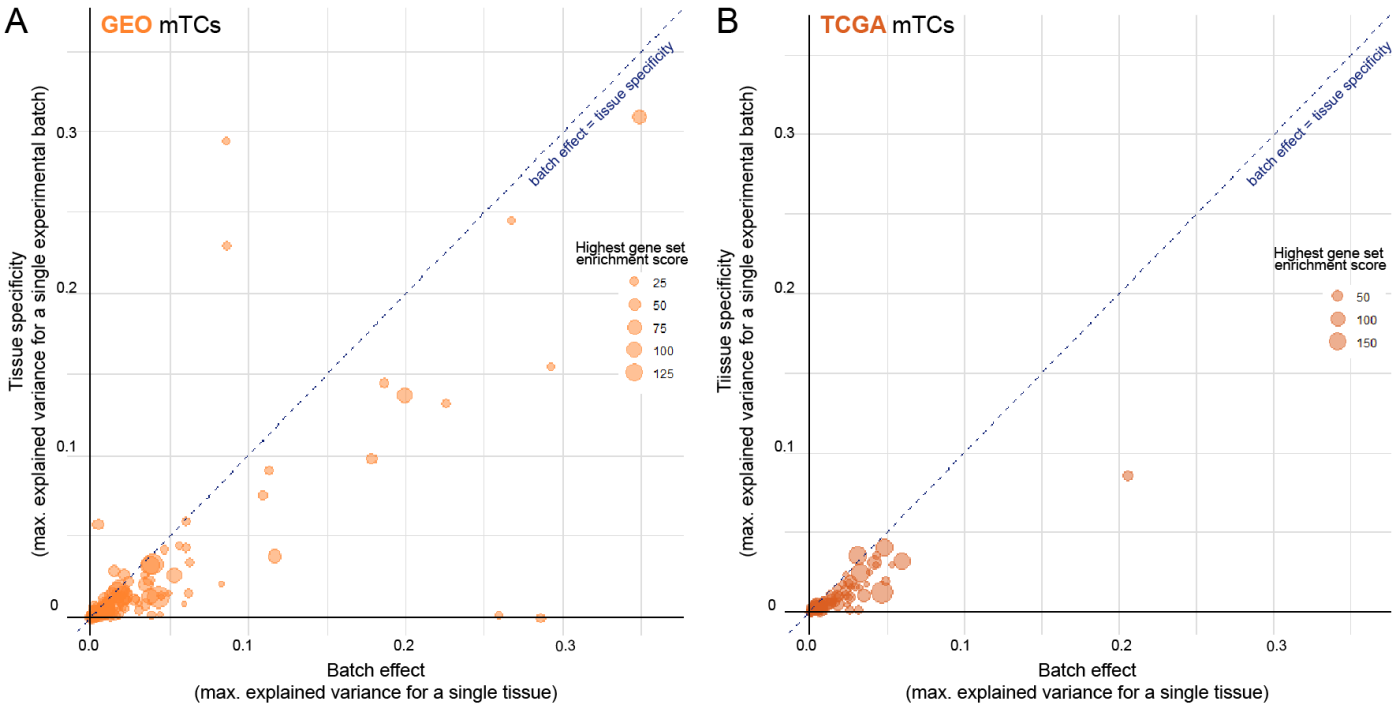

Delta area
relative change in area under CDF curve
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
50
100
150
k

796

797 **Figure S1 – Related to Figure 1;**

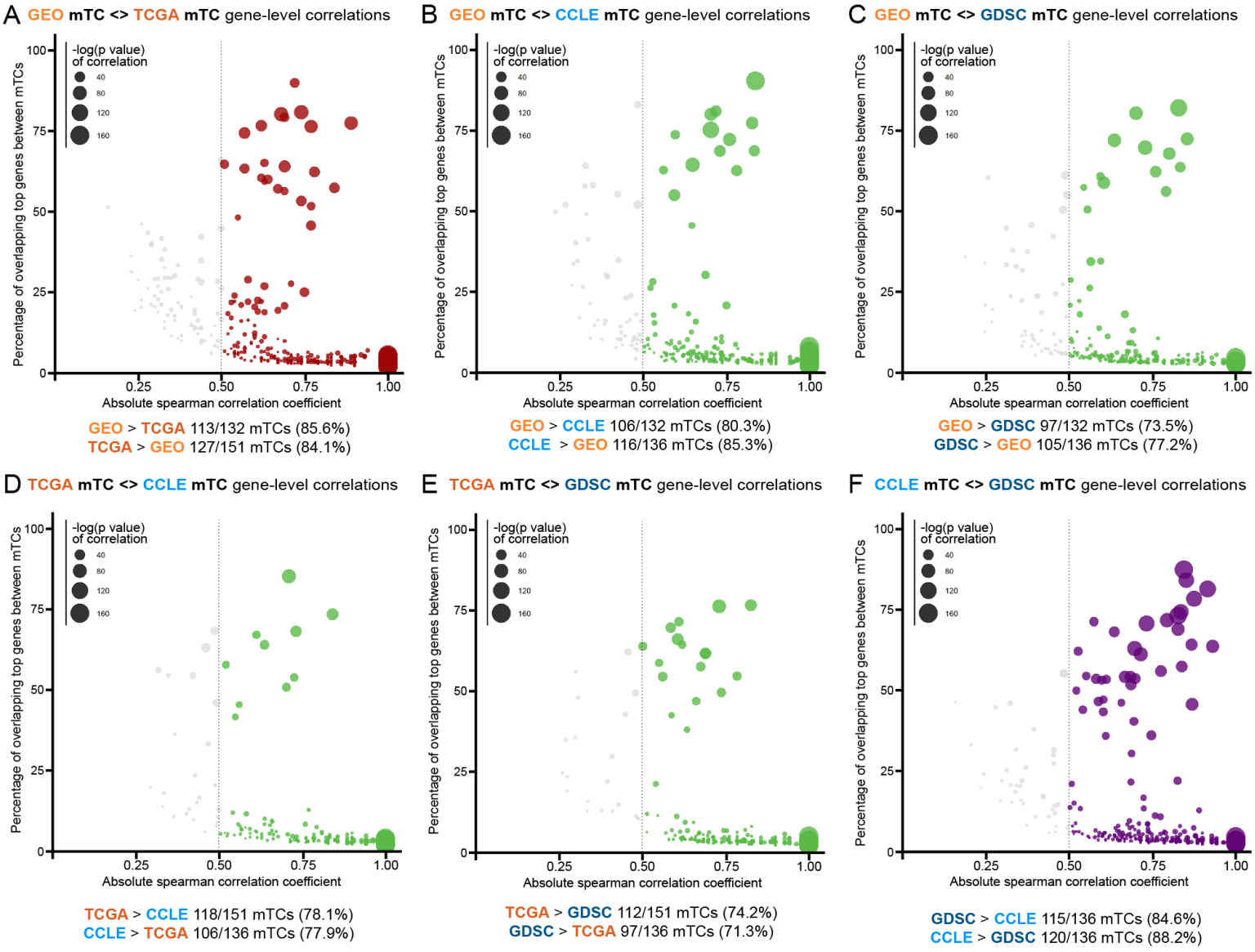
798 **(A)** Consensus clustering gene set enrichment scores of all TCs in the GEO dataset. Consensus
799 matrix for a *k* of 50 gene set clusters. **(B)** Consensus clustering gene set enrichment scores of all
800 TCs in the GEO dataset. Relative change in area under the consensus CDF curve with increasing *k*.

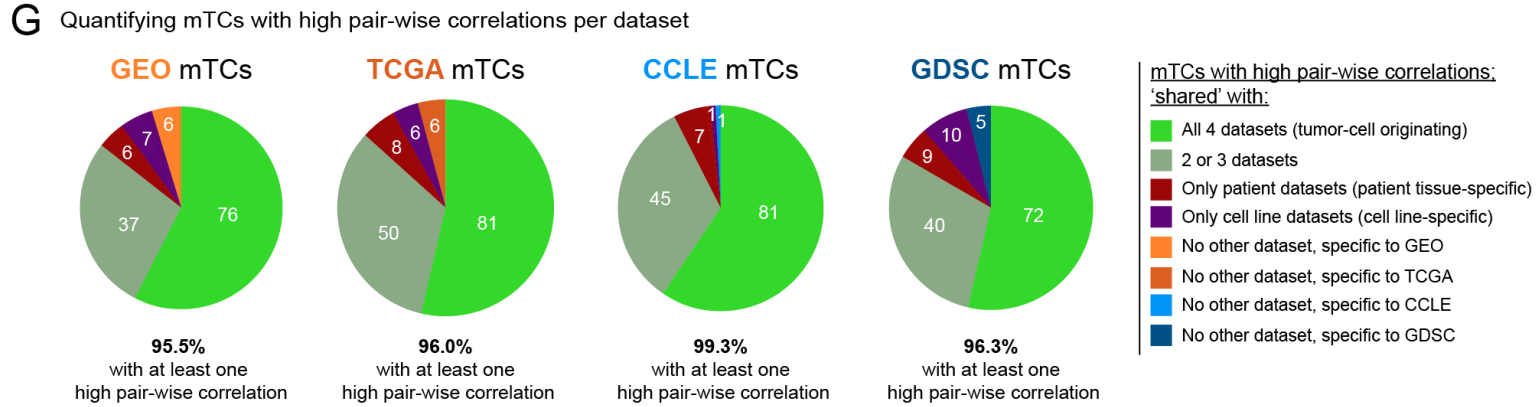
801


A
GEO mTCs
Tissue specificity
(max. explained variance for a single experimental batch)
Batch effect
(max. explained variance for a single tissue)
batch effect = tissue specificity
Highest gene set enrichment score
25
50
75
100
125
0
0.1
0.2
0.3
0.0
B
TCGA mTCs
Tiissue specificity
(max. explained variance for a single experimental batch)
Batch effect
(max. explained variance for a single tissue)
batch effect = tissue specificity
Highest gene set enrichment score
50
100
150

803 **Figure S2 – Related to Figure 1;**

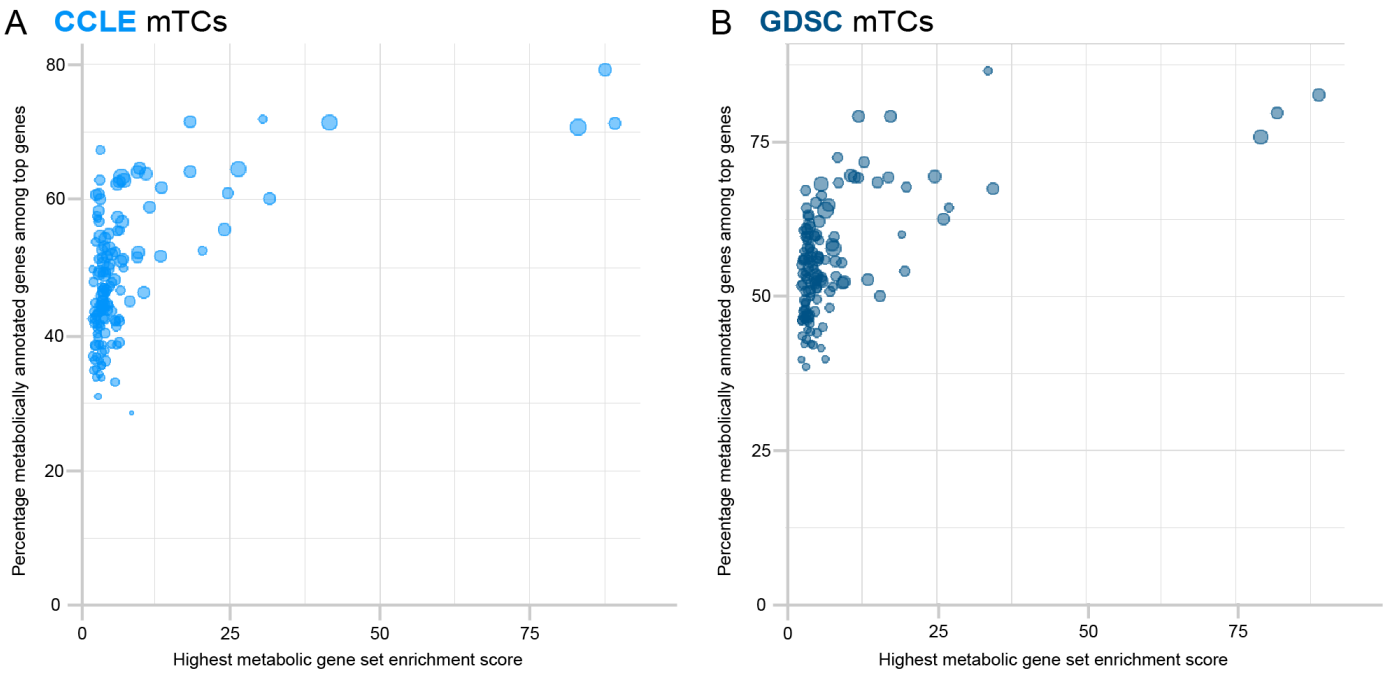
804 Scatter plots showing the maximum batch effect and tissue specificity for GEO **(A)** and TCGA **(B)**
805 mTCs. Size of the dots correspond to the highest gene set enrichment score of that mTC. The
806 magnitude of the batch effect in an mTC is estimated by the maximum fraction of the sample
807 variance in an experimental batch that is explained by that mTC. Similarly, the tissue specificity
808 of an mTC is estimated by the maximum fraction of the sample variance in a tissue type that is
809 explained by that mTC.


A GEO mTC <> TCGA mTC gene-level correlations
B GEO mTC <> CCLE mTC gene-level correlations
C GEO mTC <> GDSC mTC gene-level correlations
D TCGA mTC <> CCLE mTC gene-level correlations
E TCGA mTC <> GDSC mTC gene-level correlations
F CCLE mTC <> GDSC mTC gene-level correlations
-log(p value) of correlation
40
80
120
160
Percentage of overlapping top genes between mTCs
Absolute spearman correlation coefficient
0
25
50
75
100
0.25
0.50
0.75
1.00
GEO > TCGA 113/132 mTCs (85.6%)
TCGA > GEO 127/151 mTCs (84.1%)
GEO > CCLE 106/132 mTCs (80.3%)
CCLE > GEO 116/136 mTCs (85.3%)
GEO > GDSC 97/132 mTCs (73.5%)
GDSC > GEO 105/136 mTCs (77.2%)
TCGA > CCLE 118/151 mTCs (78.1%)
CCLE > TCGA 106/136 mTCs (77.9%)
TCGA > GDSC 112/151 mTCs (74.2%)
GDSC > TCGA 97/136 mTCs (71.3%)
GDSC > CCLE 115/136 mTCs (84.6%)
CCLE > GDSC 120/136 mTCs (88.2%)


G Quantifying mTCs with high pair-wise correlations per dataset
GEO mTCs
76
37
6
6
7
95.5%
with at least one
high pair-wise correlation
TCGA mTCs
81
50
8
6
6
96.0%
with at least one
high pair-wise correlation
CCLE mTCs
81
45
7
1
1
99.3%
with at least one
high pair-wise correlation
GDSC mTCs
72
40
9
10
5
96.3%
with at least one
high pair-wise correlation
mTCs with high pair-wise correlations; 'shared' with:
All 4 datasets (tumor-cell originating)
2 or 3 datasets
Only patient datasets (patient tissue-specific)
Only cell line datasets (cell line-specific)
No other dataset, specific to GEO
No other dataset, specific to TCGA
No other dataset, specific to CCLE
No other dataset, specific to GDSC

811 **Figure S3 – Related to Figure 1;**

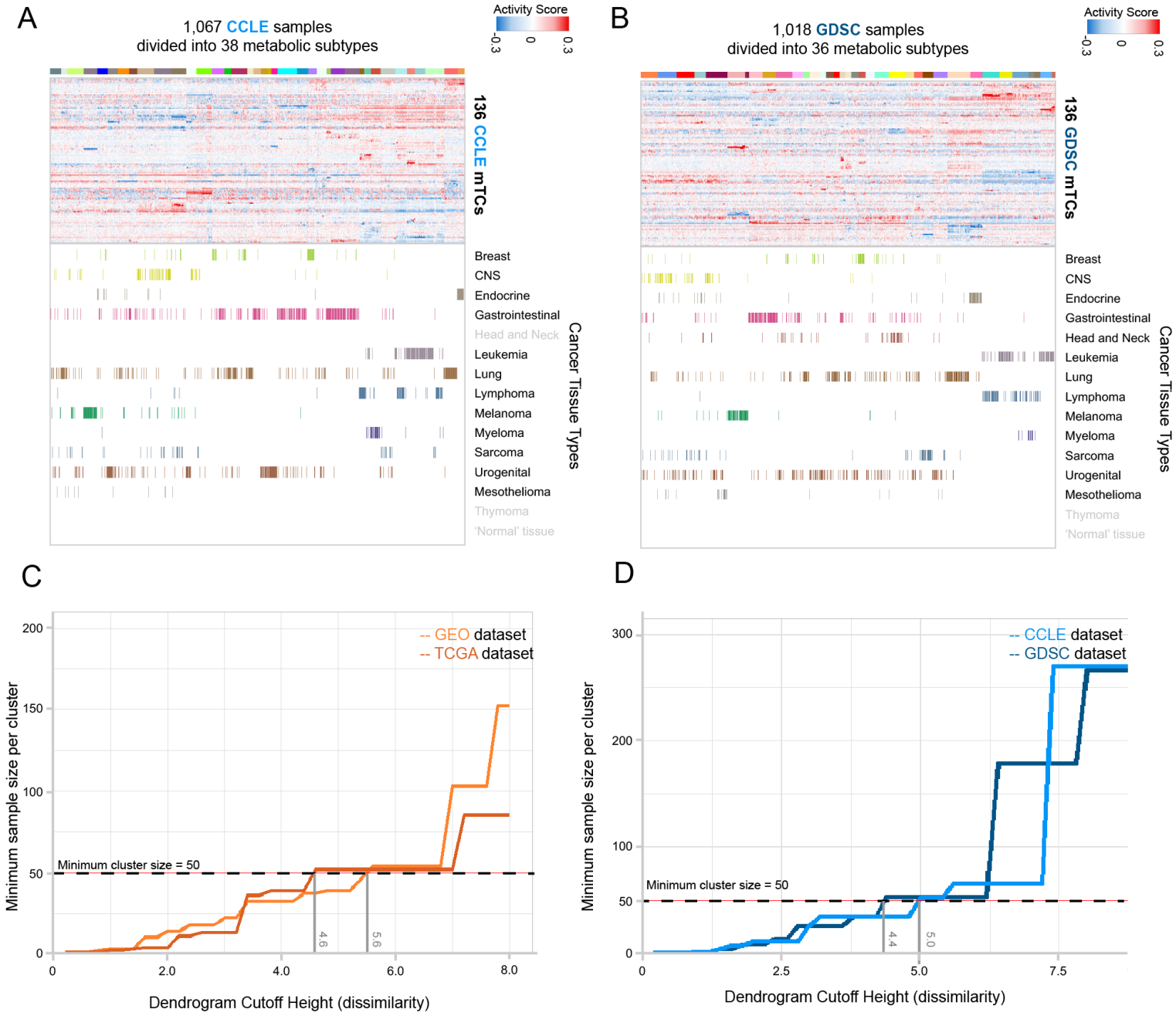
812 Scatter plot showing absolute spearman correlation coefficients (x-axis), versus the percentage
813 of overlapping top genes (genes with absolute weight >3) between mTCs from different datasets
814 (y-axis). Only significant pair-wise correlations (with P-value <0.05 and top gene overlap
815 significance <0.05) are shown. Colored dots show absolute correlations > 0.5, the size of the dots
816 represent the P-value of these spearman correlations. Scatter plots are shown for correlations
817 between **(A)** GEO and TCGA mTCs, **(B)** GDSC and CCLE mTCs, **(C)** GEO and GDSC mTCs, **(D)** TCGA
818 and GDSC mTCs, **(E)** GEO and CCLE mTCs, **(F)** TCGA and CCLE mTCs. **(G)** Pie graphs quantifying the
819 amount of mTCs with high correlations for every dataset.


A CCLE mTCs
Percentage metabolically annotated genes among top genes
Highest metabolic gene set enrichment score
0
20
40
60
80
0
25
50
75
B GDSC mTCs
Percentage metabolically annotated genes among top genes
Highest metabolic gene set enrichment score
0
25
50
75
0
25
50
75

820

821 **Figure S4 – Related to Figure 2;** Dot plots showing the highest metabolic gene set enrichment
822 score for every CCLE (A) and GDSC (B) mTC (x-axis) versus the percentage of metabolically
823 annotated genes in the top genes (genes with absolute weight >3) in those mTCs (y-axis).

824

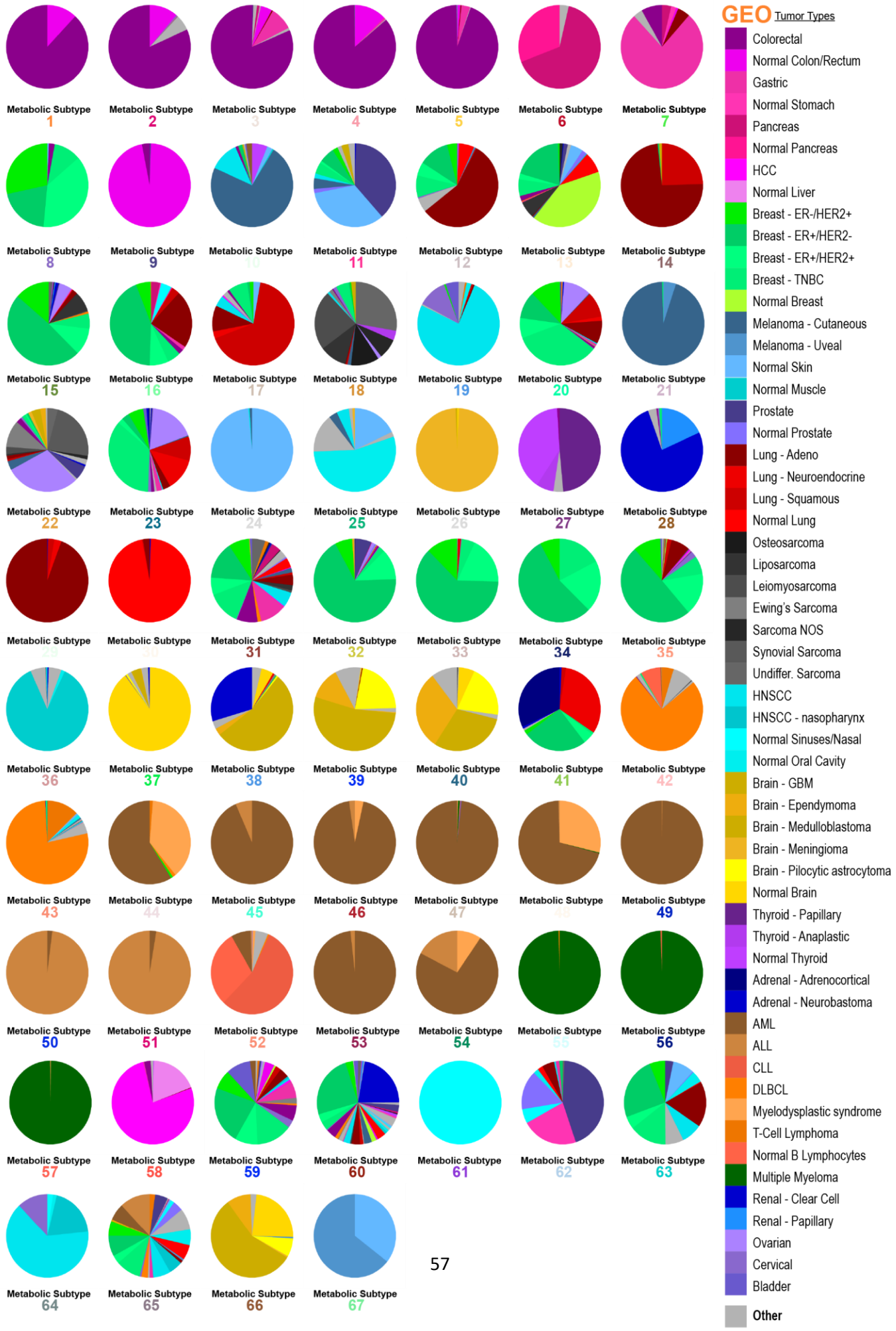

A
1,067 CCLE samples
divided into 38 metabolic subtypes
Activity Score
-0.3
0
0.3
136 CCLE mTCs
Breast
CNS
Endocrine
Gastrointestinal
Head and Neck
Leukemia
Lung
Lymphoma
Melanoma
Myeloma
Sarcoma
Urogenital
Mesothelioma
Thymoma
'Normal' tissue
Cancer Tissue Types
B
1,018 GDSC samples
divided into 36 metabolic subtypes
Activity Score
-0.3
0
0.3
136 GDSC mTCs
Breast
CNS
Endocrine
Gastrointestinal
Head and Neck
Leukemia
Lung
Lymphoma
Melanoma
Myeloma
Sarcoma
Urogenital
Mesothelioma
Thymoma
'Normal' tissue
Cancer Tissue Types
C
-- GEO dataset
-- TCGA dataset
Minimum sample size per cluster
Minimum cluster size = 50
4.6
5.6
Dendrogram Cutoff Height (dissimilarity)
D
-- CCLE dataset
-- GDSC dataset
Minimum sample size per cluster
Minimum cluster size = 50
4.4
5.0
Dendrogram Cutoff Height (dissimilarity)

825 **Figure S5 – Related to Figure 3;**

826 **(A)** Metabolic landscape for CCLE samples. The 1,067 samples were hierarchically clustered and
827 divided into 38 metabolic subtypes. **(B)** Metabolic landscape for GDSC samples. The 1,018
828 samples were hierarchically clustered and divided into 36 clusters metabolic subtypes. Grey labels
829 designate tissue types that are present in other datasets, but are not present in the given dataset.
830 **(C)** Hierarchical clustering of activity scores of mTCs in samples from GEO and TCGA datasets used
831 in order to define metabolic subtypes. The plot shows the minimum sample size of a cluster
832 depending on the chosen cutoff height of the dendrogram resulting from hierarchical clustering.
833 The heights at which the minimum cluster size reaches 50 is given for both GEO and TCGA
834 datasets. **(D)** Hierarchical clustering of activity scores of mTCs in samples from CCLE and GDSC
835 datasets used in order to define metabolic subtypes. The plot shows the minimum sample size of
836 a cluster depending on the chosen cutoff height of the dendrogram resulting from hierarchical
837 clustering. The heights at which the minimum cluster size reaches 50 is given for both CCLE and
838 GDSC datasets.

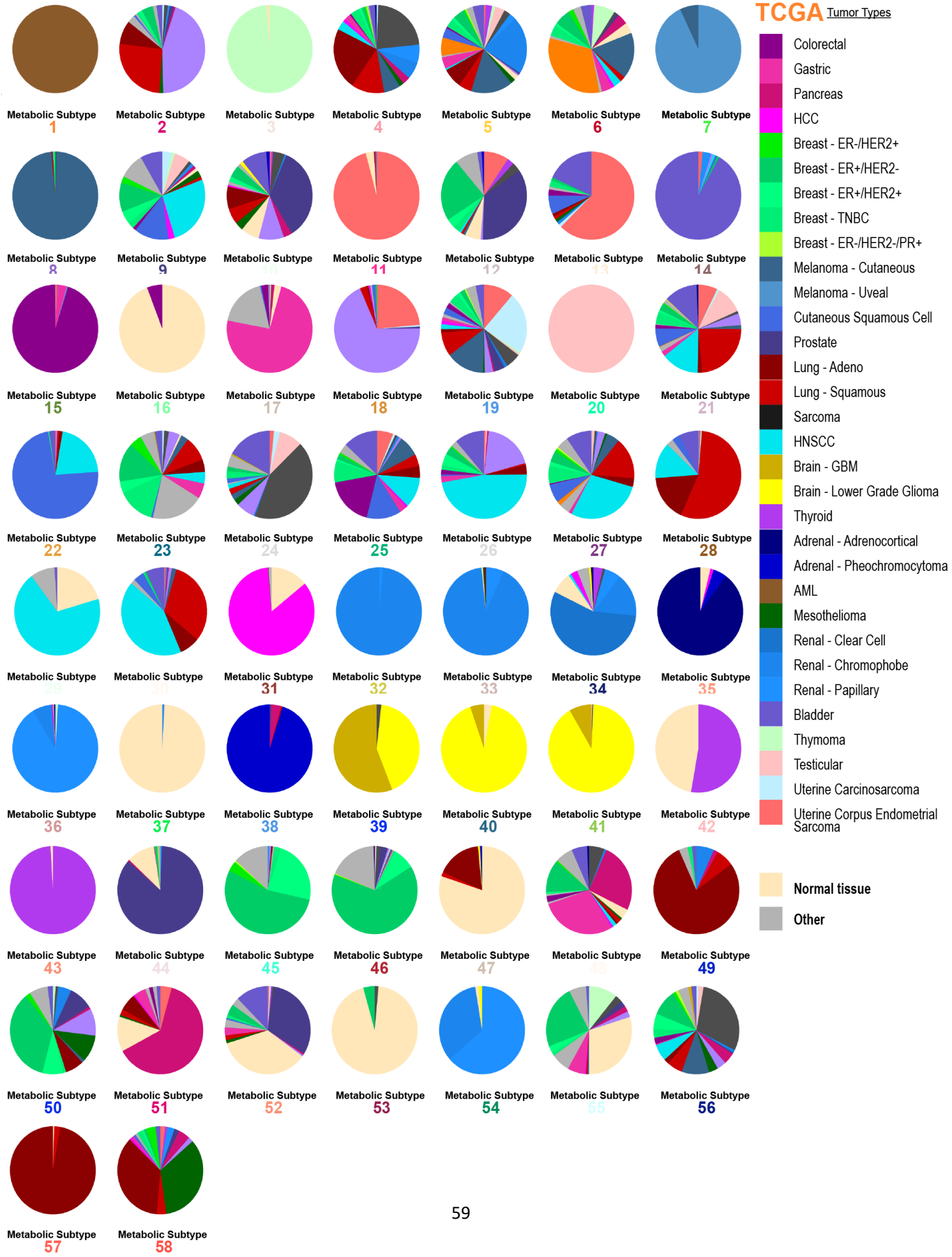
839

840


GEO Tumor Types
Colorectal
Normal Colon/Rectum
Gastric
Normal Stomach
Pancreas
Normal Pancreas
HCC
Normal Liver
Breast - ER-/HER2+
Breast - ER+/HER2-
Breast - ER+/HER2+
Breast - TNBC
Normal Breast
Melanoma - Cutaneous
Melanoma - Uveal
Normal Skin
Normal Muscle
Prostate
Normal Prostate
Lung - Adeno
Lung - Neuroendocrine
Lung - Squamous
Normal Lung
Osteosarcoma
Liposarcoma
Leiomyosarcoma
Ewing's Sarcoma
Sarcoma NOS
Synovial Sarcoma
Undiffer. Sarcoma
HNSCC
HNSCC - nasopharynx
Normal Sinuses/Nasal
Normal Oral Cavity
Brain - GBM
Brain - Ependymoma
Brain - Medulloblastoma
Brain - Meningioma
Brain - Pilocytic astrocytoma
Normal Brain
Thyroid - Papillary
Thyroid - Anaplastic
Normal Thyroid
Adrenal - Adrenocortical
Adrenal - Neurobastoma
AML
ALL
CLL
DLBCL
Myelodysplastic syndrome
T-Cell Lymphoma
Normal B Lymphocytes
Multiple Myeloma
Renal - Clear Cell
Renal - Papillary
Ovarian
Cervical
Bladder
Other

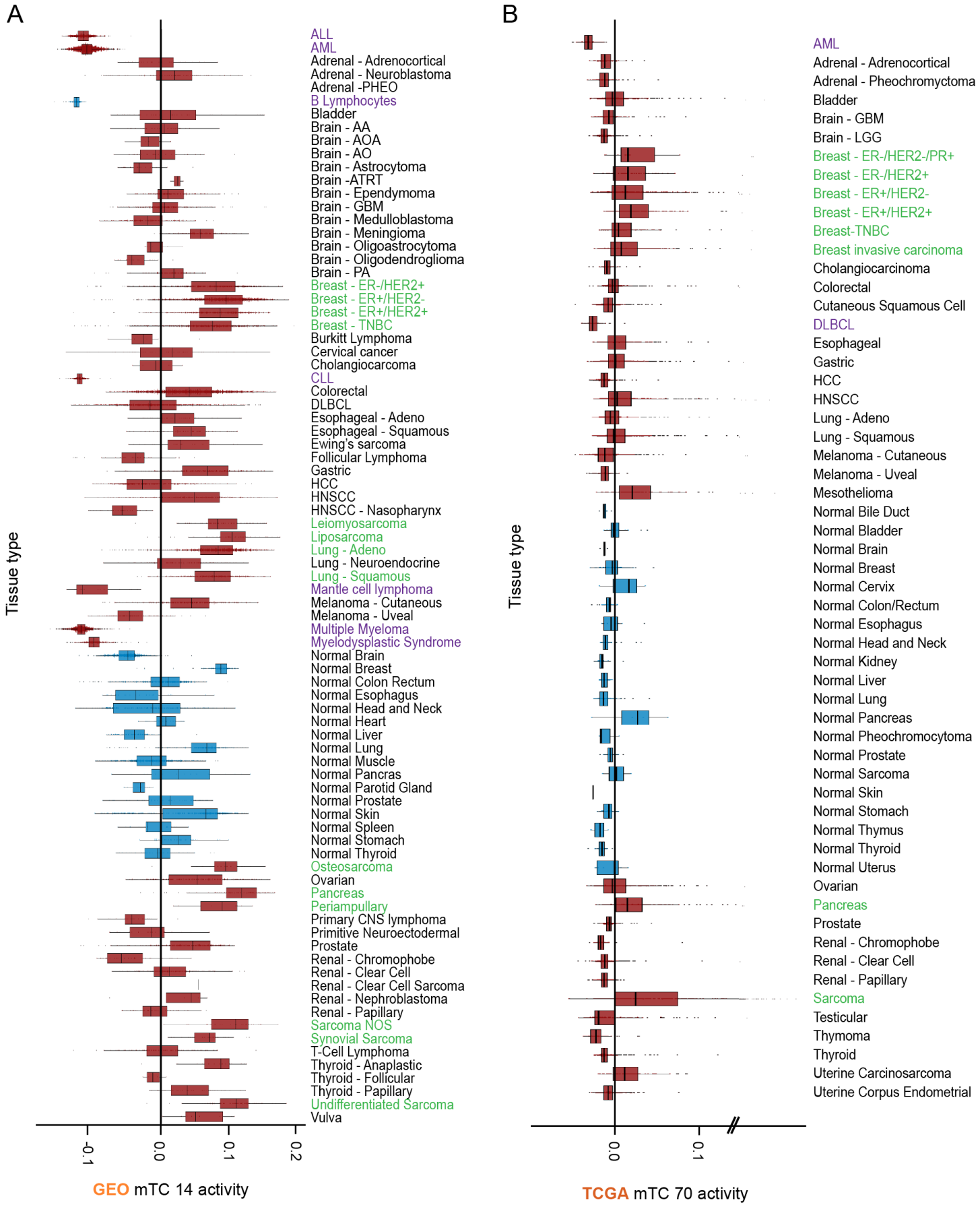
841 **Figure S6 – Related to Figure 3;** Pie graphs depicting the tissue type composition of the 67

842 metabolic subtypes defined for the GEO dataset.


TCGA Tumor Types
Colorectal
Gastric
Pancreas
HCC
Breast - ER-/HER2+
Breast - ER+/HER2-
Breast - ER+/HER2+
Breast - TNBC
Breast - ER-/HER2-/PR+
Melanoma - Cutaneous
Melanoma - Uveal
Cutaneous Squamous Cell
Prostate
Lung - Adeno
Lung - Squamous
Sarcoma
HNSCC
Brain - GBM
Brain - Lower Grade Glioma
Thyroid
Adrenal - Adrenocortical
Adrenal - Pheochromocytoma
AML
Mesothelioma
Renal - Clear Cell
Renal - Chromophobe
Renal - Papillary
Bladder
Thymoma
Testicular
Uterine Carcinosarcoma
Uterine Corpus Endometrial Sarcoma
Normal tissue
Other
Metabolic Subtype 1
Metabolic Subtype 2
Metabolic Subtype 3
Metabolic Subtype 4
Metabolic Subtype 5
Metabolic Subtype 6
Metabolic Subtype 7
Metabolic Subtype 8
Metabolic Subtype 9
Metabolic Subtype 10
Metabolic Subtype 11
Metabolic Subtype 12
Metabolic Subtype 13
Metabolic Subtype 14
Metabolic Subtype 15
Metabolic Subtype 16
Metabolic Subtype 17
Metabolic Subtype 18
Metabolic Subtype 19
Metabolic Subtype 20
Metabolic Subtype 21
Metabolic Subtype 22
Metabolic Subtype 23
Metabolic Subtype 24
Metabolic Subtype 25
Metabolic Subtype 26
Metabolic Subtype 27
Metabolic Subtype 28
Metabolic Subtype 29
Metabolic Subtype 30
Metabolic Subtype 31
Metabolic Subtype 32
Metabolic Subtype 33
Metabolic Subtype 34
Metabolic Subtype 35
Metabolic Subtype 36
Metabolic Subtype 37
Metabolic Subtype 38
Metabolic Subtype 39
Metabolic Subtype 40
Metabolic Subtype 41
Metabolic Subtype 42
Metabolic Subtype 43
Metabolic Subtype 44
Metabolic Subtype 45
Metabolic Subtype 46
Metabolic Subtype 47
Metabolic Subtype 48
Metabolic Subtype 49
Metabolic Subtype 50
Metabolic Subtype 51
Metabolic Subtype 52
Metabolic Subtype 53
Metabolic Subtype 54
Metabolic Subtype 55
Metabolic Subtype 56
Metabolic Subtype 57
Metabolic Subtype 58

844 **Figure S7 – Related to Figure 3;** Pie graphs depicting the tissue type composition of the 58

845 metabolic subtypes defined for the TCGA dataset.


A
Tissue type
ALL
AML
Adrenal - Adrenocortical
Adrenal - Neuroblastoma
Adrenal -PHEO
B Lymphocytes
Bladder
Brain - AA
Brain - AOA
Brain - AO
Brain - Astrocytoma
Brain -ATRT
Brain - Ependymoma
Brain - GBM
Brain - Medulloblastoma
Brain - Meningioma
Brain - Oligoastrocytoma
Brain - Oligodendroglioma
Brain - PA
Breast - ER-/HER2+
Breast - ER+/HER2-
Breast - ER+/HER2+
Breast - TNBC
Burkitt Lymphoma
Cervical cancer
Cholangiocarcoma
CLL
Colorectal
DLBCL
Esophageal - Adeno
Esophageal - Squamous
Ewing's sarcoma
Follicular Lymphoma
Gastric
HCC
HNSCC
HNSCC - Nasopharynx
Leiomyosarcoma
Liposarcoma
Lung - Adeno
Lung - Neuroendocrine
Lung - Squamous
Mantle cell lymphoma
Melanoma - Cutaneous
Melanoma - Uveal
Multiple Myeloma
Myelodysplastic Syndrome
Normal Brain
Normal Breast
Normal Colon Rectum
Normal Esophagus
Normal Head and Neck
Normal Heart
Normal Liver
Normal Lung
Normal Muscle
Normal Pancras
Normal Parotid Gland
Normal Prostate
Normal Skin
Normal Spleen
Normal Stomach
Normal Thyroid
Osteosarcoma
Ovarian
Pancreas
Periampullary
Primary CNS lymphoma
Primitive Neuroectodermal
Prostate
Renal - Chromophobe
Renal - Clear Cell
Renal - Clear Cell Sarcoma
Renal - Nephroblastoma
Renal - Papillary
Sarcoma NOS
Synovial Sarcoma
T-Cell Lymphoma
Thyroid - Anaplastic
Thyroid - Follicular
Thyroid - Papillary
Undifferentiated Sarcoma
Vulva
-0.1
0.0
0.1
0.2
GEO mTC 14 activity
B
Tissue type
AML
Adrenal - Adrenocortical
Adrenal - Pheochromyctoma
Bladder
Brain - GBM
Brain - LGG
Breast - ER-/HER2-/PR+
Breast - ER-/HER2+
Breast - ER+/HER2-
Breast - ER+/HER2+
Breast-TNBC
Breast invasive carcinoma
Cholangiocarcinoma
Colorectal
Cutaneous Squamous Cell
DLBCL
Esophageal
Gastric
HCC
HNSCC
Lung - Adeno
Lung - Squamous
Melanoma - Cutaneous
Melanoma - Uveal
Mesothelioma
Normal Bile Duct
Normal Bladder
Normal Brain
Normal Breast
Normal Cervix
Normal Colon/Rectum
Normal Esophagus
Normal Head and Neck
Normal Kidney
Normal Liver
Normal Lung
Normal Pancreas
Normal Pheochromocytoma
Normal Prostate
Normal Sarcoma
Normal Skin
Normal Stomach
Normal Thymus
Normal Thyroid
Normal Uterus
Ovarian
Pancreas
Prostate
Renal - Chromophobe
Renal - Clear Cell
Renal - Papillary
Sarcoma
Testicular
Thymoma
Thyroid
Uterine Carcinosarcoma
Uterine Corpus Endometrial
0.0
0.1
TCGA mTC 70 activity

847 **Figure S8 - Related to Figure 5; (A)** Activity of GEO mTC 14 in samples, grouped per tissue type.
848 Tissue types with a higher median activity highlighted in the text are given a red axis label, tissue
849 types with a lower median activity highlighted in the text are given a blue axis label. **(B)** Activity
850 of TCGA mTC 70 in samples, grouped per tissue type. Tissue types with a higher median activity
851 highlighted in the text are given a red axis label, tissue types with a lower median activity
852 highlighted in the text are given a blue axis label.