1 **Dissecting tumor cell programs through group biology estimation in clinical single-cell**

2 **transcriptomics**

3 **Authors/Affiliations**

Shreya Johri	1,2
Kevin Bi	1,2
Breanna M. Titchen	1,2,3
Jingxin Fu, PhD.	1,2
Jake Conway	1,2,4
Jett P. Crowdis	1,2
Natalie I. Volkes, M.D.	5
Zenghua Fan, PhD.	6
Lawrence Fong, M.D.	6
Jihye Park, PhD.	1,2
David Liu, M.D.	1,2
Meng Xiao He, PhD.	1,2
Eliezer M. Van Allen, M.D.*	1,2

4

5 1. Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA.

6 2. Broad Institute of Harvard and MIT, Cambridge, MA, USA.

7 3. Harvard Graduate Program in Biological and Biomedical Sciences, Boston, MA, USA.

8 4. Harvard Graduate Program in Bioinformatics and Integrative Genomics, Boston, MA,
9 USA.

10 5. Department of Thoracic and Head and Neck Oncology, MD Anderson Cancer Center,
11 Department of Genomic Medicine, MD Anderson Cancer Center.

12 6. Division of Hematology/Oncology, Department of Medicine, University of California, San
13 Francisco, San Francisco, CA 94143, USA; Parker Institute for Cancer Immunotherapy,
14 University of California, San Francisco, San Francisco, CA 94143, USA.

15

16 *corresponding author.

17 Eliezer M. Van Allen

18 Dana-Farber Cancer Institute

19 450 Brookline Ave

20 Boston MA 02215

21 Eliezerm_vanallen@dfci.harvard.edu

22

Abstract

Given the growing number of clinically integrated cancer single-cell transcriptomic studies, robust differential enrichment methods for gene signatures to dissect tumor cellular states for discovery and translation are critical. Current analysis strategies neither adequately represent the hierarchical structure of clinical single-cell transcriptomic datasets nor account for the variability in the number of recovered cells per sample, leading to results potentially confounded by sample-driven biology with high false positives instead of accurately representing true differential enrichment of group-level biology (e.g., treatment responders vs. non-responders). This problem is especially prominent for single-cell analyses of the tumor compartment, because high intra-patient similarity (as opposed to inter-patient similarity) results in stricter hierarchical structured data that confounds enrichment analysis. Furthermore, to identify signatures which are truly representative of the entire group, there is a need to quantify the robustness of otherwise statistically significant signatures to sample exclusion. Here, we present a new nonparametric statistical method, BEANIE, to account for these issues, and demonstrate its utility in two cancer cohorts stratified by clinical groups to reduce biological hypotheses and guide translational investigations. Using BEANIE, we show how the consideration of sample-specific versus group biology greatly decreases the false positive rate and guides identification of robust signatures that can also be corroborated across different cell type compartments.

Introduction

Single-cell transcriptomic profiling of patient tumors has enabled high-resolution dissections of disease progression and treatment response. Building on seminal cellular atlases for specific cancer types, many studies are increasingly focused on deriving hypotheses by evaluating groups of patients (e.g., treated vs. untreated, responders vs. non-responders, and early- vs. late-stage)

47 for differences in gene signatures (which may be an experimentally and/or computationally
48 derived aggregation of related genes or pathways) between the two groups. For this purpose,
49 methods such as the Mann-Whitney U (MWU) tests and Generalised Linear Models (GLMs) have
50 been conventionally used in bulk RNA-sequencing (bulk RNA-seq) studies as well as single-cell
51 transcriptomic analyses;[1–4] however, they may have a number of limitations for the latter
52 application. First, these methods assume mutual independence of samples, and although this is
53 not problematic for bulk RNA-seq analyses, cells derived from the same patient in single-cell
54 analyses do not satisfy this criteria. Second, these methods fall short of representing the
55 hierarchical structure of tumor single-cell transcriptomic data, as tumor cells tend to exhibit more
56 intra-patient similarity as compared to inter-patient similarity due to the expression of patient-
57 specific transcriptional programs driven by DNA-level alterations and epigenetics.[5–8] This
58 challenge, in turn, may lead to differential enrichment results being skewed by patient-specific
59 biology, instead of representing genuine group biology. Finally, the number of cells (and hence,
60 data points) sequenced in these single-cell transcriptomic datasets are typically large compared
61 to bulk RNA-seq datasets, thereby potentially increasing the power of statistical tests to detect
62 differences (by rejecting the null hypothesis) between the groups under consideration, which may
63 not reflect biologically or clinically relevant observations. These challenges exist for other cell
64 types as well, including the immune and stromal cells, albeit to a lesser extent. As a result of these
65 significant methodological challenges, single-cell transcriptomic case/control analyses of cancer
66 samples have thus far often not involved detailed assessments of the tumor compartments, which
67 has restricted the capability to learn from tumor cellular programs in increasingly complex clinical
68 contexts.

69

70 To maximize the utility of single-cell transcriptomic analyses between clinically relevant patient
71 populations and determine how tumor cell programs differ between groups of patients, we

72 developed a nonparametric statistical group biology estimation method (group **B**iology **E**stim**A**tion
73 i**N** s**I**ngle c**E**ll, "BEANIE") inspired from *He et al.*,[9] addressing the above-mentioned issues (Fig.
74 1, see Methods). This method first estimates the statistical significance (empirical p-value) of the
75 test signatures through a Monte Carlo approximation of the test signatures' p-value distribution
76 (test distribution) and that of the random signatures' p-value distribution (background distribution),
77 followed by contextualisation of the former with respect to the latter. It then uses the leave-one-
78 out cross-validation approach (sample exclusion) to infer robustness of the gene signatures (see
79 Methods). We used publicly available datasets to demonstrate the utility of this method, and
80 present suggested guidelines for the design of clinically embedded single-cell transcriptomic
81 studies in oncology.

82

83 **Results**

84 We evaluated single-cell transcriptomic data in two cancer types (melanoma and lung cancer)
85 that have the following clinical groups for comparison: (i) response to treatment; and (ii) disease
86 progression.[1,2,10–12] We contextualised their tumor compartments with signatures from the
87 Molecular Signatures Database (MSigDB),[13,14] including Hallmark (n = 50) and Oncogenic (n =
88 189) gene sets. We compared results obtained from MWU tests followed by Benjamini-Hochberg
89 (BH) corrections and GLMs with results obtained from BEANIE, and characterised our approach
90 relative to these methods (Table 1). Details regarding the implementation and comparisons are
91 available in the Methods section.

92

93 **Group biology analysis of Immune Checkpoint Blockade-naive vs. -exposed melanoma**

94 We first evaluated a melanoma dataset,[1] which included data for both immune checkpoint
95 blockade (ICB)-naive and ICB-exposed patients, to assess methodologies for comparing clinical
96 treatment states in tumor cells. All of the ICB-exposed samples were resistant to treatment and
97 were biopsied from the metastatic sites. We excluded samples having less than 50 tumor cells,
98 and, in total, there were 1891 tumor cells across 14 patients, with 7 patients per group (Fig. S1,
99 see Methods). We first assessed the data with Hallmark and Oncogenic gene set signatures from
100 MSigDb, to characterise treatment-driven biology within the tumor compartment (Fig. 2, Table 1).

101

102 We observed that the MWU test followed by a BH correction and GLMs predicted a large number
103 of differentially enriched signatures (p-value ≤ 0.05), whereas BEANIE was more conservative,
104 detecting fewer signatures as differentially enriched (Fig. 2b, Table S1). Notably, a majority of the
105 signatures identified as significant by MWU test and GLMs were labelled as non-significant and
106 non-robust to sample exclusion by BEANIE.

107

108 Among signatures that were identified as statistically significant and robust to sample exclusion
109 by BEANIE (see Methods), signatures upregulated in the ICB-naive group include those for genes
110 upregulated by *STAT5* in response to *IL2* stimulation (HALLMARK_IL2_STAT5_SIGNALING),
111 genes regulated by NF-κB in response to TNF (HALLMARK_TNFA_SIGNALING_VIA_NFKB),
112 and genes defining inflammatory response (HALLMARK_INFLAMMATORY_RESPONSE). We
113 also verified a previously identified T cell exclusion signature[1] upregulated in the ICB-exposed
114 group as statistically significant and robust to sample exclusion with the BEANIE method. To verify
115 gene-level differential expression, we used a MWU test and observed differential *IL2* gene
116 expression for the ICB-naive group in the T cell compartment (p-value = 0.0078), corroborating

117 our finding in the tumor compartment (differential HALLMARK_IL2_STAT5_SIGNALING). We
118 also identified the top constituent genes (ranked according to log2 fold change and robustness to
119 sample exclusion, see Methods) for these three signatures, and found that these genes were
120 differential in the tumor compartment uniformly across samples of a given group (Fig. 2e, Table
121 S2). Together, these results describe the tumor microenvironment of the ICB-exposed group
122 (consisting of treatment-resistant patients) as one depleted of T cells, with reduced IL2-STAT5
123 signaling, TNFA-NFKB signaling, and inflammatory response relative to the ICB-naive group.

124

125 Additionally, we found that the signature for genes upregulated by *IL6* via STAT3
126 (HALLMARK_IL6_JAK_STAT3_SIGNALING) was upregulated in the ICB-naive group and
127 statistically significant and robust to sample exclusion. Using a MWU test, we found differential
128 *STAT3* expression in the tumor compartment for the ICB-naive group (p-value = 3.28e-36).
129 Furthermore, we also found a positive correlation between the *STAT3* expression and
130 HALLMARK_IL6_JAK_STAT3_SIGNALING signature score in the tumor cells on an individual
131 cell basis (Fig. 2f). This observation supports the finding that *IL6* could potentially induce
132 downstream signaling via *STAT3* in the tumor cells of the ICB-naive group.[15,16]

133

134 We further examined the cause for non-robustness of the signatures that were identified as
135 statistically significant but not robust to sample exclusion by BEANIE (Fig. 2c). We found that the
136 exclusion of one or more samples led to statistically non-significant results, in contrast to when
137 the sample was included, by shifting the empirical p-value to greater than 0.05 as a result of an
138 overlap between the test distribution and the background distribution as shown in Fig. 2d (see
139 Methods). For example, the signature ONCOGENIC_RAF_UP.V1_UP was not robust to the
140 exclusion of sample Mel106, and this particular sample was also the cause of non-robustness for

141 21 other signatures. This variability due to sample exclusion was also not explained by any of the
142 other available clinical variables (e.g., age, sex). Therefore, these signatures were driven by
143 sample-specific biology, and were consequently not representative of the group-level biology, but
144 would have otherwise been considered differentially enriched with statistical significance using
145 either of the conventional MWU test or GLM approaches.

146

147 We next investigated the methodological stability with respect to subsample size (Fig. 2g), and
148 accordingly repeated BEANIE's workflow using smaller subsample sizes. We found that a smaller
149 subsampling of cells led to fewer signatures that were identified as statistically significant and
150 non-robust to sample exclusion, and even fewer that were identified as both statistically significant
151 and robust to sample exclusion. However, the number of statistically significant and robust
152 signatures identified by BEANIE reached saturation around the subsample size of 30 cells per
153 sample, indicating that the subsample size of 60, which had been used for all of the
154 aforementioned results, could successfully capture all statistically significant signatures from the
155 test signature sets that were also robust to sample exclusion.

156

157 To assess the ability to detect noise from a true signal, we additionally used a curated set of
158 immune cell surface marker signatures[17] (including signatures for T cells, NKT cells, NK cells, B
159 cells, mast cells, and a joint dendritic cell/macrophage signature), that should not be relevant to
160 tumor cells, to test the performance of the three methods (MWU test with a BH correction, GLMs,
161 and BEANIE). We observed that a MWU test with a BH correction led to a p-value ≤ 0.05 for all
162 signatures except the B cell signature and GLMs led to a p-value ≤ 0.05 for NKT cell, B cell, NK
163 cell, and the joint dendritic cell/macrophage signatures. By contrast, BEANIE correctly predicted

164 all of the immune cell surface marker signatures as both statistically non-significant and non-
165 robust to sample exclusion (Table S3).

166

167 Finally, we evaluated BEANIE's performance on the previously reported 18 T cell exclusion
168 signatures[1] for the tumor cell compartments of a reduced set of patients from the original set used
169 to derive these (only patient samples with greater than 50 tumor cells were retained, as described
170 above). We observed that while a MWU test with a BH correction had p-values ≤ 0.05 for 18/18
171 signatures and GLMs had p-values ≤ 0.05 for 11/18 signatures, BEANIE had an empirical p-value
172 ≤ 0.05 for 17/18 signatures and additionally found 10/18 of them to be robust to sample exclusion.

173

174 **Group biology analysis of distinct clinical states in non-small cell lung carcinoma**

175 In an effort to demonstrate the applicability of BEANIE for a meta-analysis composed of multiple
176 single-cell transcriptomic datasets, we next analyzed the tumor compartments from four published
177 lung cancer studies[2,10–12] to evaluate potential differentially enriched signatures between early-
178 vs. late-stage samples. We selected patient samples which satisfied the following criteria: (i) had
179 more than 50 tumor cells; (ii) were classified as adenocarcinoma; (iii) were staged as either I, II,
180 or IV (early-stage = I and II; late-stage = IV); and (iv) had received no prior treatment at the time
181 of sample collection. Filtering according to these criteria yielded a total of 18251 malignant cells
182 across 17 patients (11 early-stage, 6 late-stage) (Fig. 3a).

183

184 We sought to characterise the tumor compartment with Hallmark and Oncogenic gene sets from
185 MSigDb (Fig. 3b, Table 1). Again, a large number of gene sets predicted as differentially enriched
186 with statistical significance by a MWU test with a BH correction and GLMs were identified as

187 statistically non-significant and non-robust to sample exclusion by BEANIE (Fig. 3c, Table S1).
188 We found a signature composed of genes important for mitotic spindle assembly
189 (HALLMARK_MITOTIC_SPINDLE) to be statistically significant and robust to sample exclusion
190 for early-stage lung tumors with BEANIE, consistent with prior studies.[18] Another signature
191 comprised of genes encoding proteins involved in glycolysis and gluconeogenesis
192 (HALLMARK_GLYCOLYSIS) was also found to be statistically significant and robust to sample
193 exclusion for the early-stage tumors with BEANIE, which is in agreement with a prior study[19]
194 describing an association between TKI treatment and its effect on decreased activity of glycolysis.
195 Furthermore, the top constituent genes for both of these signatures were consistently upregulated
196 across all samples (Fig. S2, Table S2). Thus, BEANIE was able to detect both statistically
197 significant and robust signatures in the meta-analysis of multiple single-cell transcriptomic
198 datasets.

199

200 We next sought to evaluate the tumor compartment from two of the lung cancer datasets (Kim et
201 al.,[10] Maynard et al.[11]) for treatment responses to tyrosine kinase inhibitors (TKIs). We selected
202 patient samples which satisfied the following criteria: (i) had more than 50 tumor cells; and (ii) the
203 biopsy was derived from the primary tumor. These filtering criteria led to a total of 7576 malignant
204 cells across 10 patients (6 TKI-naive, 4 TKI-exposed) (Fig. S3).

205

206 We again used the Hallmark and Oncogenic gene sets to characterise the tumor compartment
207 (Fig. 4a, 4b, Fig. S3; Table 1, Table S1). Among the signatures that were found to be statistically
208 significant and robust to sample exclusion with BEANIE, signatures upregulated in the TKI-
209 exposed group included a signature for genes upregulated in response to *IFNG*
210 (HALLMARK_INTERFERON_GAMMA_RESPONSE) and a signature for genes upregulated by

211 the overexpression of *WNT1* (ONCOGENIC_WNT_UP.V1_UP). We identified the top constituent
212 genes of both signatures, and found them to be consistently upregulated across all samples in
213 the TKI-exposed group (Fig. 4c, Table S2). Interferon gamma response has been described to be
214 associated with response to TKI treatment in non-small cell lung cancers.[20] Using a MWU test,
215 we observed that genes encoding the *IFNG* receptors (*IFNGR1, IFNGR2*) were differentially
216 expressed with statistical significance in the tumor cells of the TKI-exposed group (p-value
217 [*IFNGR1*] = 3.22e-104, p-value [*IFNGR2*] = 1.81e-13, Fig. S3). WNT signaling has also been
218 extensively studied in the context of cancer development, and increased WNT signaling has been
219 associated with tumor progression and metastasis in many different cancers.[21] We assessed
220 potential intratumoral differential gene expression of *WNT1* in the tumor cells and found an
221 absence of intratumoral *WNT1* expression altogether. We then assessed potential *WNT1*
222 differential gene expression in specific immune cell compartments (NK cells, macrophages, and
223 T cells) and found a statistically significant differential expression of *WNT1* for the TKI-exposed
224 group within the T cell compartment (MWU test, p-value = 8.77e-18), indicating putative cross-
225 compartment communication between the T cells and tumor cells via *WNT1* signaling. To further
226 validate this, we used a MWU test to investigate possible differential gene expression of *WNT1*
227 receptors (*FZD1, FZD2*) in the tumor compartment and found both of the receptors to be
228 upregulated in the TKI-exposed group (p-value [*FZD1*] = 6.35e-21, p-value [*FZD2*] = 1.28e-27).
229 Of note, patients who were treated with TKI were classified with RECIST as having either PD
230 (Progressive Disease) or RD (Residual Disease), which raises the hypothesis that these patients
231 may have developed therapeutic resistance through the WNT/beta-catenin signaling pathway in
232 alignment with prior preclinical studies.[22]

233

234 In addition, we estimated the stability of BEANIE to subsample size for the test signatures used
235 (Fig. S3), and found that the number of robust signatures identified persistently increased at the

236 maximum sample size, indicating the possibility that some of the signatures classified as robust
237 could have been instead classified as non-robust. This may be a result of unbalanced samples
238 per group being tested or may demonstrate the necessity of additional biological samples for the
239 clinical context being evaluated.

240

241 **Estimation of the False Positive Rate**

242 In order to estimate the chance of occurrence of incorrectly identified statistically significant and
243 robust signatures with BEANIE, we calculated the false positive rate (type I error) (see Methods,
244 Fig. 5) for all three methods (MWU test with a BH correction, GLMs, BEANIE) and clinical contexts
245 (ICB-naive vs. -exposed melanoma, early- vs. late-stage lung cancer, and TKI-naive vs. -exposed
246 lung cancer) for the signatures that had been classified by BEANIE as both statistically significant
247 and robust to sample exclusion.

248

249 We observed that across all datasets, the MWU test with a BH correction had a high average
250 false positive rate, followed by GLMs which exhibited a moderately high average false positive
251 rate. By contrast, BEANIE had the lowest average false positive rate, that in some cases also
252 approached the significance level (alpha) of 5% (Table 2). Individual false positive rates calculated
253 for all robust and statistically significant signatures can be found in Table S4.

254

255 To evaluate how a smaller number of cells being tested, and thereby reduced statistical power,
256 would impact the false positive rate for a MWU test with a BH correction and GLMs, we
257 subsampled cells from each sample being tested to a number equivalent to BEANIE's subsample
258 size and repeated the type I error estimation. We found that subsampling decreased the false

259 positive rates for both a MWU test with a BH correction and GLMs, but that their false positive
260 rates were still relatively higher than those calculated with BEANIE, corroborating BEANIE's
261 aptitude for detecting robust and true signals as compared to the other two methods, and also
262 reinforcing the need to incorporate robustness estimation into differential enrichment testing.

263

264 **Group biology estimation in immune cells**

265 While this strategy was primarily developed to overcome challenges for differential enrichment
266 testing specifically within the tumor cell compartment, we also evaluated whether the subsampling
267 and sample exclusion approach implemented within BEANIE would likewise yield biological
268 insights in immune cell compartments, as well as to further validate some of the initial hypotheses
269 from the tumor compartment analyses. As a test case, in continuation of the preliminary evaluation
270 of the CD8+ T cell compartment as described in the earlier tumor compartment analysis, we more
271 comprehensively dissected the CD8+ T cell compartment in ICB-naive vs. -exposed melanoma
272 patients in an isolated context here (Fig. S1).

273

274 We filtered out samples which had fewer than 50 CD8+ T cells, yielding a total of 1292 cells across
275 11 patients (5 ICB-naive, 6 ICB-exposed). We evaluated the potential statistically significant
276 differential enrichment and robustness to sample exclusion of various signatures representing a
277 range of CD8+ T cell subtypes and states from Oliveira et al.[23] between the ICB-naive and -
278 exposed patient groups. We found that previously reported signatures for early activated CD8+ T
279 cells (Sade-Feldman_5[24]) and memory precursor effector CD8+ T cells (Joshi_MPEC,[25] murine-
280 derived) were statistically significant and robust to sample exclusion, and upregulated for the ICB-
281 naive patient group. This result substantiates the finding from the tumor compartment of the same
282 dataset, where we had identified higher IL2-STAT5 signaling, TNF activation, and inflammatory

283 response in the tumor cells from the ICB-naive group as described, which may be a result of
284 CD8+ T cell activation in the naive condition.

285

286 Therefore, in addition to the demonstrated utility for tumor cell compartments, BEANIE likewise
287 exhibited capacity for cross-compartment validations of group biology in single-cell non-tumor
288 populations as well.

289

290 **Conclusions**

291 Conventional differential enrichment methods, such as a MWU test with a BH correction and
292 GLMs, are limited in correctly estimating differential biology in clinical tumor single-cell
293 transcriptomic datasets in two aspects. First, they have an appreciably high false positive rate,
294 which can be attributed in part to an increased power of statistical tests (due to high cell numbers)
295 to detect differences between groups. However, increased power does not necessarily signify a
296 biologically relevant difference. Consequently, interpretation of these differences in a group
297 biology context is requisite to correctly distinguish genuine group biological differences from
298 technical artifacts (such as variation in cell numbers). We also observed that subsampling alone
299 is insufficient to tackle this problem, and it is important to use a background distribution for
300 contextualisation. Second, conventional differential enrichment methods do not assess the
301 robustness of a signature to sample exclusion, and as a consequence, these methods may lead
302 to results being sample-driven and of uncertain translational importance. This issue is particularly
303 relevant in clinical contexts, and especially for tumor cell compartments which demonstrate higher
304 intra-patient similarity than inter-patient similarity, as hypotheses based on group comparison
305 (about treatment effects, disease progression, etc.) may impact future clinical trials.

306

307 To address the shortcomings of conventional differential enrichment methods, we developed
308 BEANIE, a nonparametric statistical method for estimating group biology in clinical single-cell
309 transcriptomic datasets. We demonstrated its application on publicly available datasets from six
310 clinical single-cell transcriptomic studies, and illustrate its aptitude to detect statistically significant
311 and robust gene signatures as compared to conventional methods, through its low false positive
312 rate as compared to its counterparts (MWU test followed by a BH correction and GLMs). We
313 illustrated its extensive application in the tumor compartment, and its potential utility for the
314 immune compartment as well. It may likewise be used to identify differential enrichment of gene
315 signatures in the stromal compartment. Finally, we demonstrated that BEANIE is adept at
316 distinguishing sample-driven signatures from group-driven signatures, whereas conventional
317 differential enrichment methods fail to do so. Alternate models for representation of tumor single-
318 cell data include hierarchical linear models; however, unlike BEANIE, they are parametric and
319 therefore assume normality and homogeneity of variance for the data.

320

321 Despite its potential to estimate group biology and pinpoint both statistically significant as well as
322 robust and therefore prospective biologically relevant signatures in single-cell transcriptomic
323 dissections, BEANIE also possesses a few limitations. First, in spite of its demonstrated value in
324 single-cell transcriptomic tumor compartment analyses, BEANIE's widespread applicability in the
325 immune compartment may be limited, in part due to an absence of comprehensive databases
326 with precise and rigorous signatures representing discrete cell types, states, and pathways. In
327 fact, the ultimate utility of BEANIE's or any group biology analysis tool's framework is in part
328 contingent on the quality of the gene signatures being tested, including for the tumor
329 compartment. Moreover, there also exists scope to further improve the false positive rate within
330 the BEANIE method. In addition, we do not currently have an understanding of why some patient

331 samples are more prone to contribute to the non-robustness of certain signatures as compared
332 to other patient samples, and having additional clinical information (e.g., mutational status) could
333 potentially help delineate some of the biology behind this. Lastly, despite the ability to estimate
334 group biology and identify statistically significant and robust signatures between patient groups
335 with BEANIE, current clinical single-cell transcriptomic datasets have an overall small sample
336 size, which indicates that they are likely not an adequate representation of the broader population
337 and hence could lead to introduction of false negatives (type II error). Therefore, in general, larger
338 datasets, such as those generated via consortium efforts, are needed to improve our ability to
339 draw robust conclusions, and minimise putative false negatives. Broadly, dedicated efforts to
340 analyze larger clinically integrated single-cell cohorts that reflect the diverse clinical and
341 therapeutic contexts across cancer types will accelerate our understanding of the cell states that
342 promote treatment resistance for translational discovery.

343

344 **Methods**

345 **Data Preprocessing**

346 **Melanoma Dataset**

347 We selected cells which were labelled as malignant (authors made use of inferCNV[28] to identify
348 malignant cells).

349

350 **Lung Cancer Datasets**

351 Owing to the variability in collected datasets from the four studies (Kim et al.[10], Maynard et al.[11],
352 Qian et al.[12], and Lambrechts et al.[2]), we carefully assessed the metadata files available. For

353 Lambrechts et al., we reached out to the authors to acquire their Seurat object containing patient
354 ID and cell ID labelling. We used the following criteria for the selection of cells for analysis: (i)
355 must be of epithelial origin; (ii) must be identified as malignant by the authors (all studies made
356 use of inferCNV[28] to identify malignant cells); and (iii) must be isolated from the primary site (i.e.,
357 lung). We also removed cells from patients that had locally advanced lung cancer (stage III
358 tumors), as they are more difficult to classify into early- versus late-stage.[26]

359

360 **BEANIE's Workflow**

361 **Preprocessing and Normalisation**

362 The raw counts matrix is normalised by the library size and converted to counts per million (CPM
363 normalization) to account for differences in library sizes of different cells. Pre-normalised matrices
364 may also be used, in which case this step is ignored. Genes with no expression across all cells
365 are excluded.

366

367 **Signature Scoring**

368 For each cell c_i in the normalised counts matrix, signature scoring is performed for the set of gene
369 signatures provided as input by the user (test signatures). The default signature scoring method
370 is adapted from AUCell.[27]

371

372 (i) For each gene g_k, the cells are ranked by calculating the percentile of each cell across the gene
373 g_k in terms of normalised expression of the gene, i.e., cells with higher expression values of that
374 particular gene will have a higher percentile. The ties are randomly broken (i.e. if two cells have

375 the exact same expression of the gene, which is common in single-cell datasets, those cells are
376 randomly assigned a percentile value).

377
$$P_{c_i, g_k} = \frac{100 \times n_{c_i, g_k}}{n_c}$$

378 where n_{c_i, g_k} = ordinal rank of c_i for expression of g_k (sorted from smallest to largest),

379 P_{c_i, g_k} = percentile of c_i for expression of g_k,

380 and n_c is the total number of cells

381

382 (ii) Next, for every cell c_i, genes are ranked based on their calculated percentile values across
383 that cell. Genes which have a higher percentile across the cell are given lower ranks. This scoring
384 system takes into account the importance of each gene in a given cell relative to that gene's
385 importance in other cells, i.e., genes which have a lower rank are more important for the cell in
386 question as compared to genes with a higher rank.

387
$$P'_{g_k, c_i} = \frac{100 \times n'_{g_k, c_i}}{n_g}$$

388 where n'_{g_k, c_i} = ordinal rank of g_k for P_{c_i, g_k} values (sorted from largest to smallest),

389 P'_{g_k, c_i} = percentile of g_k for P_{c_i, g_k} values,

390 and n_g is the total number of genes for the cell c_i

391

392 (iii) For each gene signature S_j, a recovery curve per cell c_i is generated by calculating the
393 enrichment of the top constituent genes ranked from S_j, followed by a calculation of the Area

Under the Curve (AUC), which measures the expression of c_i's top constituent genes ranked from S_j. The AUC is therefore the score of the cell for S_j.

$$\mathrm{AUC}_{j,i} = \int f_i(x)\,dx\,,\ f_i(x) = \mathrm{n}(\mathrm{S}_j \cap \mathrm{R}_{x,i})$$

where S_j = set of genes comprising a gene signature

and $R_{x,i}$ = set of top constituent x genes based on P'_{g_k, c_i}

Other signature scoring methods available in BEANIE include weighted mean and z-scoring.

Background Distribution Generation

A background distribution is generated for the biological interpretability of the results as follows: (i) Bins are created based on the gene set size of each signature S_j (default bin size = 10, tunable parameter). (ii) Random signatures (r_signatures) (R_k, k = 1, 2, … , n_b, where n_b = total number of bins) for each of the bin sizes are generated such that they are representative of both lowly expressed and highly expressed genes. For this step, the normalised matrix is used and the genes are sorted based on their expression values across all samples. Equal numbers of genes from every 20th percentile are then randomly subsampled such that the sum of all genes equals the bin size. This random sampling is repeated multiple times to generate different random signatures (R_{kl}, l = 1, 2, … , n_r, where n_r = the total number of times subsampling is repeated). The rationale for generating the random signatures is that they should not represent any biologically meaningful gene signature, and as a consequence, their differential expression can be used as a null distribution (background distribution) for interpretation of the results in a biological context. (iii) Each cell c_i is scored for R_{kl}'s using the aforementioned signature scoring method.

416

417 **Folds and Subsampling**

418 To accomplish BEANIE's two-fold aim of having equal sample representation and quantifying
419 robustness for S_js, two statistical techniques, Monte Carlo approximations (subsampling) and
420 leave-one-out cross-validation (sample exclusion), are coupled. First, the data is divided into folds
421 (f_q, q = 1, 2, ..., n_p, where n_p = number of samples), with each fold f_q representing the exclusion of
422 one sample from either group. For each fold f_q, cells are subsampled such that each sample is
423 represented by an approximately equal number of cells. This is done by first subsampling an
424 equal number of cells from all samples, followed by additional subsampling in the sample-
425 excluded group to compensate for the cells that would have otherwise been subsampled from the
426 excluded sample. The additional subsampling ensures that the total number of cells subsampled
427 from the two groups being tested always remains constant regardless of which group the excluded
428 sample belongs to, which is necessary to ensure that the folds are comparable with each other.
429 The subsampling is then repeated multiple times to establish adequate representation of each
430 patient sample.

431

432 **Identification of Differentially Enriched Signatures**

433 A multi-step strategy is adopted to identify differentially enriched signatures. First, for each
434 subsample belonging to the fold f_q, a MWU test is performed between the two groups for every
435 S_j. Additionally, for each fold f_q, a null p-value distribution is generated by a MWU test between
436 the two groups for every R_{kl}. The null distribution generated is fold-specific to ensure that the
437 sample excluded from the fold is also excluded for the generation of the null distribution. The
438 percentile of the subsample's p-value against the null p-value distribution is then calculated,
439 hereafter referred to as the empirical p-value. A median empirical p-value is calculated for these

440 subsamples to represent the p-value for a given fold, followed by a median across all folds to
441 represent the cell's p-value. To quantify the robustness of S_j to sample exclusion, a ratio
442 (henceforth referred to as the Fold Rejection Ratio (FRR)) is defined, and calculated for every fold
443 f_q.

$$\mathrm{FRR}_q = \begin{cases} \dfrac{\mathrm{n}(\mathrm{F}_{q,\, p \leq 0.05})}{\mathrm{n}(\mathrm{F}_{0,\, p \leq 0.05})} & \text{if, } \mathrm{n}(\mathrm{F}_{0,\, p \leq 0.05}) \neq 0 \\ 0 & \text{if, } \mathrm{n}(\mathrm{F}_{0,\, p \leq 0.05}) = 0 \end{cases}$$

444

445 where F_q = set of subsamples for the fold f_q which have an empirical p-value ≤ 0.05

446 and F_0 = set of subsamples for the fold f_0 (when no sample is excluded) which have an
447 empirical p-value ≤ 0.05

448

449 A FRR value closer to 1 indicates that exclusion of the sample has no effect on the empirical
450 significance of the gene signature S_j, and a lower value indicates the opposite. We use a threshold
451 of 0.9 (hyperparameter) to call signatures as robust or not, i.e., if the FRR for a particular S_j is
452 greater than 0.9 for *all* folds, then the gene signature is considered robust to sample exclusion.

453

454 **Gene Ranking**

455 For every gene signature S_j, the genes are then ranked for the robustness of their log2 fold change
456 between the two groups. This is particularly useful for larger gene sets. For every sample, a mean
457 gene expression (MGE) is calculated for every gene using the normalised counts. A similar
458 strategy of subsampling coupled with sample exclusion is used for ranking. The MGE matrix is
459 then divided into folds, with each fold representing the exclusion of one sample. A log2 fold
460 change is then calculated for each fold, and the standard deviation, along with the mean across

461 folds, is also calculated. Genes with both outlier MGE values and outlier log2 fold changes (i.e.,
462 MGE values and log2 fold changes more than 1.5 times the interquartile range above the third
463 quartile or below the first quartile) are classified as non-robust to sample exclusion. The final
464 ranking of genes is performed based on decreasing log2 fold change, increasing standard
465 deviation, and robustness status.

466

467 **Mann-Whitney U tests**

468 Mann-Whitney U (MWU) tests followed by Benjamini-Hochberg (BH) correction are performed for
469 the calculation of p-values. The Python package *scipy* is used for the MWU p-value calculation
470 and the function *multipletests* from the Python package *statsmodels* is used for the BH correction.

471 **Generalised Linear Models**

472 Generalised linear models (GLMs) with a binomial distribution link function are used for calculation
473 of p-values. The Python package *statsmodels* is used to implement this method. The signature
474 scores are used as covariates (exog variable), and the group labels (e.g., treatment-naive or -
475 exposed and early-stage or late-stage) as the response variable to be modelled (exog variable).

476 **Calculation of False Positive Rate (Type I Error)**

477 False positive rate (type I error) refers to the probability of detecting a result by chance. To
478 calculate this, we permute the patient ID and group label in such a way that roughly equal numbers
479 of samples from the original group labels are placed in both comparison groups. We then repeat
480 the BEANIE workflow on these permuted datasets for signatures which are classified by BEANIE
481 as statistically significant and robust to sample exclusion in the original dataset to evaluate the
482 type I error rates for our predictions. In addition, we also run a MWU test followed by a BH

correction and GLMs for these signatures to compare the type I error rates across the three methods. Finally, to investigate whether Monte Carlo subsampling (with equivalent statistical power to that of BEANIE's workflow) would affect the false positive rate. For this, we subsample a random set of cells equal to the number of cells subsampled in the BEANIE workflow and repeat the MWU test and GLM methods.

For the ICB-naive vs. ICB-exposed melanoma dataset (14 samples, 7 in each group) and early-stage vs. late-stage lung cancer dataset (17 samples, 11 early-stage and 6 late-stage), we ran 1000 simulations per gene signature with the above workflow to estimate the false positive rate. For the TKI-naive vs. TKI-exposed lung cancer dataset (10 samples, 6 TKI-naive and 4 TKI-exposed), we ran 100 simulations per gene signature (due to limited combinations of equidistributed samples per group possible).

Data Availability

All datasets used in the study are publicly available. Hallmark and Oncogenic gene sets are available for download from MSigDb.

Code Availability

Code is publicly available as a downloadable Python package from:
https://github.com/sjohri20/beanie.

503 **Acknowledgements**

504 This project was funded by NIH U01 CA233100 (E.M.V.A., L.F.), R01CA227388 (E.M.V.A.),
505 U2CCA233195 (E.M.V.A.), PCF-Movember Challenge Award (E.M.V.A.)

506

507 **Disclosures**

508 E.M.V.A. has received research support (to institution) from Novartis and BMS. E.M.V.A. serves
509 as a consultant or on scientific advisory boards of Tango Therapeutics, Genome Medical,
510 Invitae, Enara Bio, Janssen, Manifold Bio, Monte Rosa. E.M.V.A. has equity in Tango
511 Therapeutics, Genome Medical, Syapse, Enara Bio, Manifold Bio, Microsoft, Monte Rosa.
512 E.M.V.A. receives travel reimbursements from Roche/Genentech. E.M.V.A. has filed institutional
513 patents on chromatin mutations and immunotherapy response, and methods for clinical
514 interpretation, and has intermittent legal consulting on patents for Foaley & Hoag.

515

516 L.F. has received research support (to institution) from Roche/Genentech, Abbvie, Bavarian
517 Nordic, Bristol Myers Squibb, Dendreon, Janssen, Merck, and Partner Therapeutics. L.F. has
518 served on the scientific advisory boards of Actym, Allector, Astra Zeneca, Atreca, Bioalta, Bolt,
519 Bristol Myer Squibb, Immunogenesis, Merck, Merck KGA, Nutcracker, RAPT, Scribe, Senti,
520 Soteria, TeneoBio, and Roche/Genentech.

521

522 M.X.H. has been a consultant to Amplify Medicines, Ikena Oncology, and Janssen. He is also
523 currently an employee of Genentech/Roche.

524

525 J.C. has been a consultant to Tango Therapeutics. He is also currently an employee of PathAI.

526

527 N.I.V. serves on advisory boards of Sanofi and Oncocyte.

528

529 **Author Contributions**

530 M.X.H. and E.M.V.A. conceived the original idea. S.J. developed the idea further, designed
531 experiments, performed the analyses, and developed the Python package for the presented
532 method. K.B., B.M.T., J.F. and M.X.H. helped in development of the method from a single-cell
533 perspective. M.X.H., J.C. and D.L. provided input from a statistical perspective. J.P.C. provided
534 input for the development of visualisation modules in the package. N.I.V. and D.L. helped in
535 clinical interpretation of the results. Z.F., J.P., L.F., D.L., and E.M.V.A. contributed to the overall
536 analyses. S.J., B.M.T., K.B., and E.M.V.A. wrote the manuscript. All authors reviewed and
537 approved the final manuscript.

538

539 **Competing Interests statement**

540 The authors declare no competing interests.

541

542 **References**

543 1. Jerby-Arnon, L. *et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to
544 Checkpoint Blockade. *Cell* **175**, 984–997.e24 (2018).

545 2. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor
546 microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).

547 3. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with
548 COVID-19. *Nat. Med.* **26**, 842–844 (2020).

549 4. Habib, N. *et al.* Disease-associated astrocytes in Alzheimer's disease and aging. *Nat.*
550 *Neurosci.* **23**, 701–706 (2020).

551 5. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with
552 Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

553 6. Lütge, A. *et al.* CellMixS: quantifying and visualizing batch effects in single-cell RNA-seq
554 data. *Life Sci Alliance* **4**, (2021).

555 7. Chen, W. *et al.* A multicenter study benchmarking single-cell RNA sequencing technologies
556 using reference samples. *Nat. Biotechnol.* **39**, 1103–1114 (2020).

557 8. Richards, L. M. *et al.* A comparison of data integration methods for single-cell RNA
558 sequencing of cancer samples. *bioRxiv* 2021.08.04.453579 (2021)
559 doi:10.1101/2021.08.04.453579.

560 9. He, M. X. *et al.* Transcriptional mediators of treatment resistance in lethal prostate cancer.
561 *Nat. Med.* **27**, 426–433 (2021).

562 10. Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular
563 reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11**, 1–15 (2020).

564 11. Maynard, A. *et al.* Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-
565 Cell RNA Sequencing. *Cell* **182**, 1232–1251.e22 (2020).

566 12. Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment

567 revealed by single-cell profiling. *Cell Res.* **30**, 745–762 (2020).

568 13. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for
569 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–
570 15550 (2005).

571 14. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set
572 collection. *Cell Syst* **1**, 417–425 (2015).

573 15. Sansone, P. & Bromberg, J. Targeting the interleukin-6/Jak/stat pathway in human
574 malignancies. *J. Clin. Oncol.* **30**, 1005–1014 (2012).

575 16. Block, K. M., Hanke, N. T., Maine, E. A. & Baker, A. F. IL-6 stimulates STAT3 and Pim-1
576 kinase in pancreatic cancer cell lines. *Pancreas* **41**, 773–781 (2012).

577 17. Zhu, Y., Yao, S. & Chen, L. Cell surface signaling molecules in the control of immune
578 responses: a tide model. *Immunity* **34**, 466–478 (2011).

579 18. Curtis, N. L., Ruda, G. F., Brennan, P. & Bolanos-Garcia, V. M. Deregulation of
580 Chromosome Segregation and Cancer. *Annu. Rev. Cancer Biol.* **4**, 257–278 (2020).

581 19. Poliaková, M., Aebersold, D. M., Zimmer, Y. & Medová, M. The relevance of tyrosine
582 kinase inhibitors for global metabolic pathways in cancer. *Mol. Cancer* **17**, 27 (2018).

583 20. Gurule, N. J. *et al.* A tyrosine kinase inhibitor-induced interferon response positively
584 associates with clinical response in EGFR-mutant lung cancer. *npj Precision Oncology* **5**,
585 1–11 (2021).

586 21. Zhan, T., Rindtorff, N. & Boutros, M. Wnt signaling in cancer. *Oncogene* **36**, 1461–1473
587 (2016).

588 22. Song, Z., Wang, H. & Zhang, S. Negative regulators of Wnt signaling in non-small cell lung

589 cancer: Theoretical basis and therapeutic potency. *Biomed. Pharmacother.* **118**, 109336
590 (2019).

591 23. Oliveira, G. *et al.* Phenotype, specificity and avidity of antitumour CD8+ T cells in
592 melanoma. *Nature* **596**, 119–125 (2021).

593 24. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint
594 Immunotherapy in Melanoma. *Cell* **175**, 998–1013.e20 (2018).

595 25. Joshi, N. S. *et al.* Inflammation Directs Memory Precursor and Short-Lived Effector CD8+ T
596 Cell Fates via the Graded Expression of T-bet Transcription Factor. *Immunity* **27**, 281–295
597 (2007).

598 26. Goldstraw, P. *et al.* The IASLC Lung Cancer Staging Project: Proposals for Revision of the
599 TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for
600 Lung Cancer. *J. Thorac. Oncol.* **11**, 39–51 (2016).

601 27. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat.*
602 *Methods* **14**, 1083–1086 (2017).

603 28. inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV

604

605

606 **Figure Legends**

607 **Main Figures**

608 Figure 1. Overview of the BEANIE method.

609 A. Overall workflow: A counts matrix, sample IDs, group IDs, and a list of signatures for which
610 differential enrichment will be tested (test signatures, t_signatures) are provided as user input.
611 Based on the gene set size, test signatures are first divided into bins. For each bin size, a list of
612 random signatures (r_signatures) of the same gene set size is generated, to be later used for p-
613 value calculation and biological interpretation. Signature scoring per cell is performed for both
614 random signatures and test signatures, followed by differential enrichment testing.

615 B. Differential enrichment testing workflow: The differential enrichment testing algorithm is
616 based on a combination of Monte Carlo approximation of empirical p-value through
617 subsampling, and leave-one-out cross validation through sample exclusion. The data is first
618 divided into folds, where each fold f_q represents the exclusion of a sample from either of the
619 comparison groups. This is followed by the subsampling step, where an equal number of cells
620 are subsampled from every sample to ensure equal patient representation. Next, a Mann-
621 Whitney U test is performed per subsample for all folds, for both the test signatures and the
622 background distribution (generated from the random signatures). The test signatures are then
623 matched to their corresponding background distribution based on bin size, and an empirical p-
624 value (percentile of the test distribution's median with respect to the background distribution) is
625 calculated per test signature for every fold f_q. Additionally, a Fold Rejection Ratio (FRR) (see
626 Methods) is calculated per test signature for every fold, and is used to determine the overall
627 robustness of the test signature to sample exclusion.

628

629 Figure 2. Group biology analysis of the tumor compartment from ICB-naive vs. ICB-exposed
630 melanoma patient samples.

631 A. Bar plot displaying the log(empirical p-value) for all of the signatures identified as statistically
632 significant (empirical p-value ≤ 0.05), along with their robustness status. A '***' above a bar
633 indicates that the empirical p-value for that test signature was below the minimum empirical p-
634 value measured.

635 B. Venn diagram quantifying the intersection of signatures identified as differentially enriched with
636 statistical significance by the three methods (MWU test with a BH correction, GLMs, and BEANIE).

637 C. Plot depicting the signatures identified as statistically significant but non-robust to sample
638 exclusion by BEANIE, the distribution of their Fold Rejection Ratios (FRRs), and the sample IDs
639 having FRRs less than the threshold used to determine robustness, along with a horizontal bar
640 plot of the number of statistically significant but non-robust signatures (dropout signatures) per
641 sample.

642 D. Histogram illustrating the sample exclusion procedure implemented within BEANIE shifting the
643 test distribution to the right such that it overlaps with the background distribution, leading to the
644 fold's empirical p-value being greater than 0.05.

645 E. Heatmap revealing the differential top constituent genes (ranked according to log2 fold change
646 and robustness) from three of the statistically significant and robust signatures identified by
647 BEANIE across all patients from the ICB-naive group.

648 F. Joint scatter and density plot demonstrating a positive correlation between the
649 HALLMARK_IL6_JAK_STAT3_SIGNALING signature score and STAT3 gene expression in
650 individual cells.

651 G. Plot illustrating BEANIE's stability to subsample size for the test signatures used. The curve
652 plateaued as the number of statistically significant test signatures, irrespective of robustness
653 status, reached saturation as the subsample size approached 60 (the max subsample size
654 possible within the constraints of this dataset [see Methods]), whereas the curve plateaued
655 around the subsample size of 30 as the number of signatures identified as both statistically
656 significant and robust reached saturation.

657

658 Figure 3. Group biology analysis for early- vs. late-stage non-small cell lung cancer.

659 A. Overview of the integrated dataset from four studies and a bar plot depicting the number of
660 tumor cells per patient sample.

661 B. Bar plot displaying the log(empirical p-value) for all of the signatures identified as statistically
662 significant (empirical p-value ≤ 0.05), along with their robustness status.

663 C. Venn diagram quantifying the intersection of signatures identified as differentially enriched with
664 statistical significance by the three methods (MWU test with a BH correction, GLMs, and BEANIE).

665 D. Plot depicting the signatures identified as statistically significant but non-robust to sample
666 exclusion by BEANIE, the distribution of their FRRs, and the patient IDs having FRRs less than
667 the threshold used to determine robustness, along with a horizontal bar plot of the number of
668 statistically significant but non-robust signatures (dropout signatures) per sample.

669

670 Figure 4. Group biology analysis for TKI-naive vs. TKI-exposed non-small cell lung cancer.

671 A. Venn diagram quantifying the intersection of signatures identified as differentially enriched with
672 statistical significance by the three methods (MWU test with a BH correction, GLMs, and BEANIE).

673 B. Bar plot displaying the log(empirical p-value) for all of the signatures identified as statistically
674 significant (empirical p-value ≤ 0.05) along with their robustness status. A '***' above a bar
675 indicates that the empirical p-value for that test signature was below the minimum empirical p-
676 value measured.

677 C. Heatmap showing the differential top constituent genes (ranked according to log2 fold change
678 and robustness) from the HALLMARK_INTERFERON_GAMMA_RESPONSE and
679 ONCOGENIC_WNT_UP.V1_UP signatures across all samples from the TKI-exposed group.

680

681 Figure 5. False positive rate (type I error) for the three methods: BEANIE, GLMs, and a MWU
682 test followed by a BH correction.

683 The false positive rate for signatures (from Hallmark and Oncogenic gene sets) which were
684 classified as statistically significant and robust to sample exclusion by BEANIE for:

685 A. ICB-naive vs. ICB-exposed melanoma dataset,

686 B. Early-stage vs. Late-stage lung cancer dataset, and

687 C. TKI-naive vs. TKI-exposed lung cancer dataset

688 The dashed pink line denotes the 5% error mark.

689

690 **Supplementary figures**

691 Figure S1. Distribution of cells from ICB-naive vs. ICB-exposed melanoma patient samples.

692 A. Distribution of tumor cells.

693 B. Distribution of CD8+ T cells.

694

695 Figure S2. Heatmap displaying the differential top constituent genes (ranked according to log2
696 fold change and robustness) from the HALLMARK_MITOTIC_SPINDLE and
697 HALLMARK_GLYCOLYSIS gene signatures in the early- vs. late-stage lung cancer dataset.

698

699 Figure S3. Extended group biology analysis of the tumor compartment from TKI-naive vs. TKI-
700 exposed lung cancer patient samples.

701 A. Distribution of tumor cells in the TKI-naive vs. -exposed lung cancer samples.

702 B. Boxplot illustrating statistically significant differential expression of the genes IFNGR1 and
703 IFNGR2 between TKI-naive vs. TKI-exposed samples.

704 C. Line plot illustrating BEANIE's stability to subsample size for TKI-naive vs. -exposed lung
705 cancer samples. The curve plateaued as the number of signatures identified as statistically
706 significant, irrespective of robustness status, reached saturation as the subsample size
707 approached 100.

708 D. Plot depicting the test signatures identified as statistically significant but non-robust to sample
709 exclusion by BEANIE for TKI-naive vs. TKI-exposed samples, the distribution of their FRRs, and
710 the patient IDs having FRRs less than threshold used to determine robustness, along with a
711 horizontal bar plot of the number of statistically significant but non-robust signatures (dropout
712 signatures) per sample.

713

714 **Supplementary Tables**

715 Table S1: Hallmark and Oncogenic gene set results for MWU test + BH correction, GLMs, and
716 BEANIE, for all datasets (ICB-naive vs. -exposed melanoma, early- vs. late-stage lung cancer,
717 and TKI-naive vs. -exposed lung cancer).

718

719 Table S2: Top genes for Hallmark and Oncogenic gene sets for all datasets (ICB-naive vs. -
720 exposed melanoma, early- vs. late-stage lung cancer, and TKI-naive vs. -exposed lung cancer).

721

722 Table S3: Noise estimation p-values for the three methods (MWU test with a BH correction, GLMs,
723 and BEANIE).

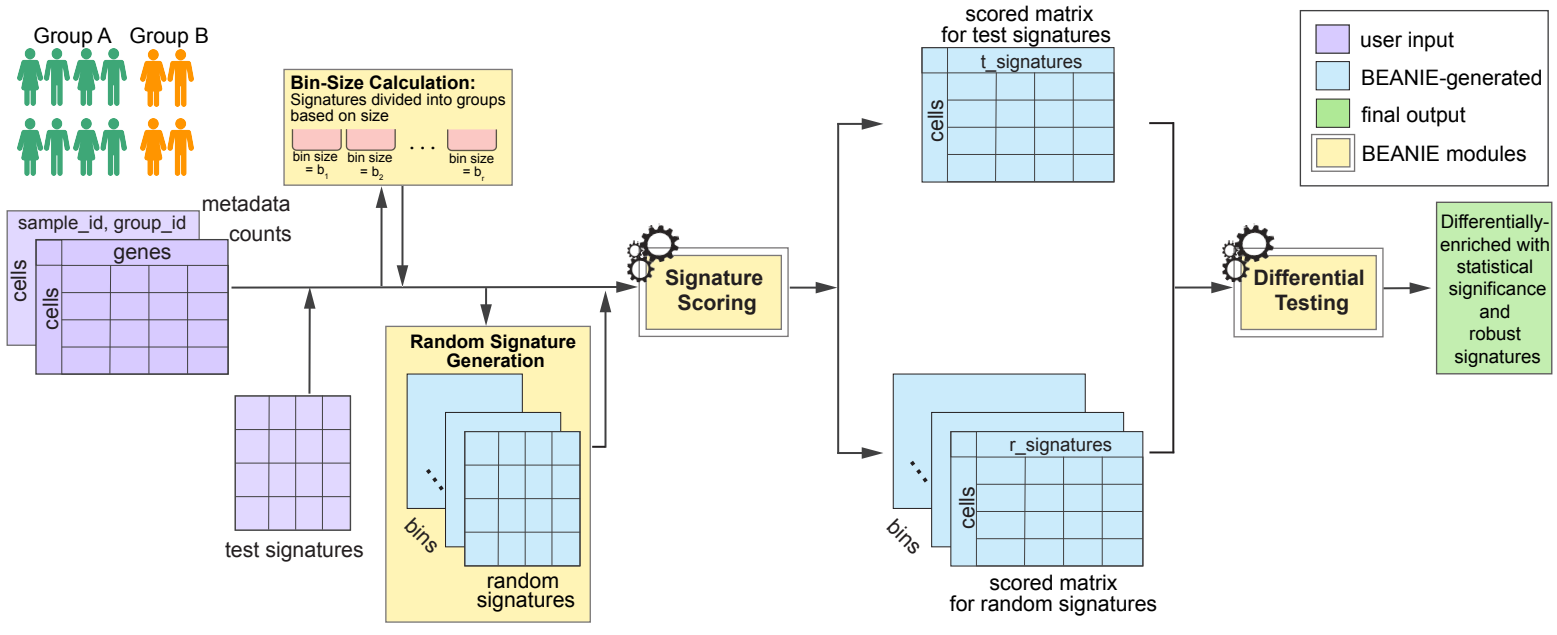
724

725 Table S4: False positive rate (in percentage) for statistically significant and robust signatures
726 identified by BEANIE for the Hallmark and Oncogenic gene sets for all datasets (ICB-naive vs. -
727 exposed melanoma, early- vs. late-stage lung cancer, and TKI-naive vs. -exposed lung cancer).

728

Figure 1.

A.


Group A
Group B
Bin-Size Calculation:
Signatures divided into groups based on size
bin size = b1
bin size = b2
bin size = br
metadata
counts
sample_id, group_id
genes
cells
cells
test signatures
Random Signature Generation
bins
random signatures
Signature Scoring
scored matrix for test signatures
t_signatures
cells
r_signatures
cells
bins
scored matrix for random signatures
Differential Testing
Differentially-enriched with statistical significance and robust signatures
user input
BEANIE-generated
final output
BEANIE modules

B.

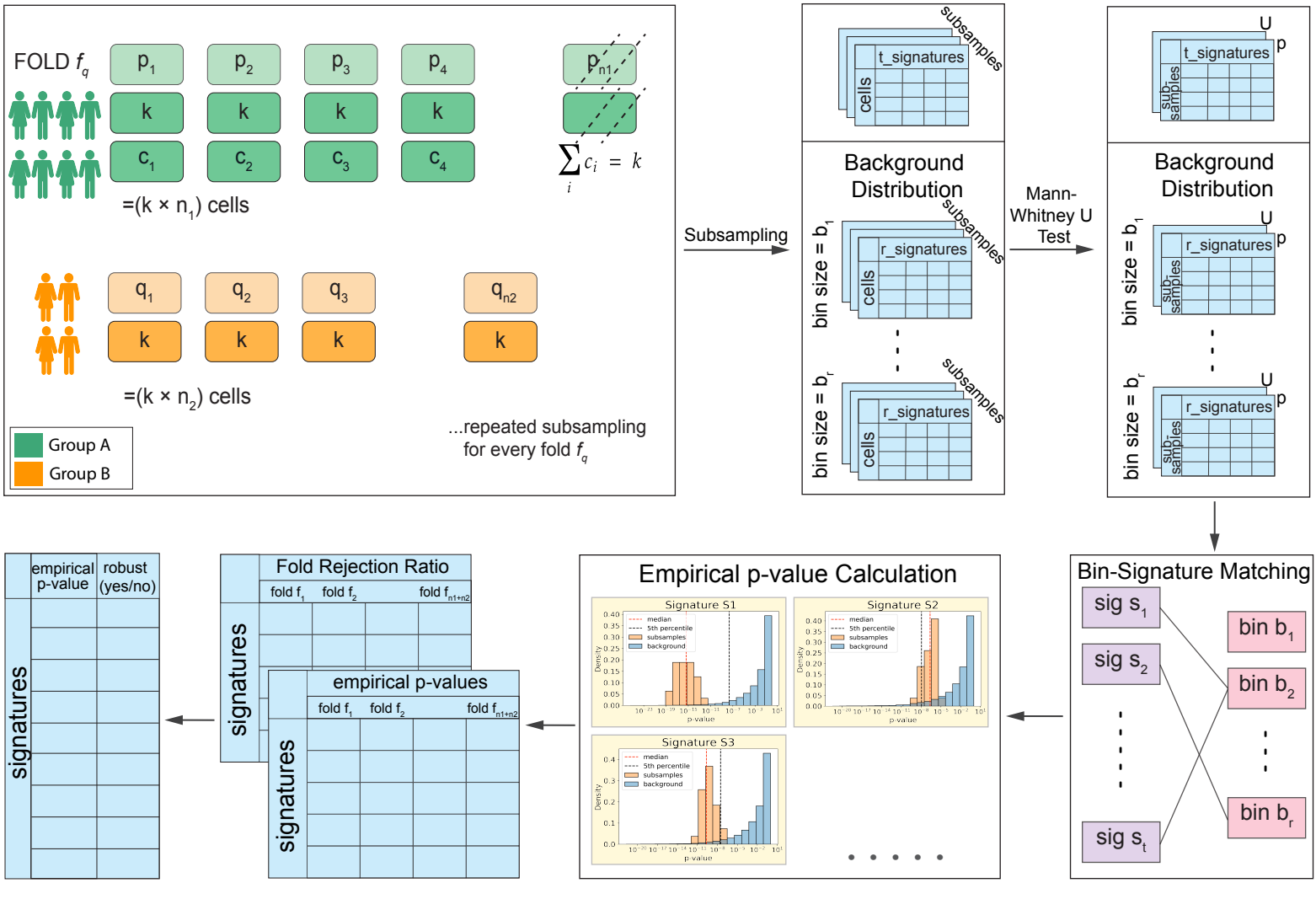

FOLD fq
p1
p2
p3
p4
pn1
k
c1
c2
c3
c4
=(k × n1) cells
q1
q2
q3
qn2
=(k × n2) cells
Group A
Group B
...repeated subsampling for every fold fq
Subsampling
subsamples
t_signatures
cells
Background Distribution
bin size = b1
bin size = br
r_signatures
Mann-Whitney U Test
U
p
sub samples
Bin-Signature Matching
sig s1
sig s2
sig st
bin b1
bin b2
bin br
Empirical p-value Calculation
Signature S1
Signature S2
Signature S3
median
5th percentile
subsamples
background
Density
p-value
empirical p-values
fold f1
fold f2
fold fn1+n2
Fold Rejection Ratio
signatures
empirical p-value
robust (yes/no)

Figure 2.

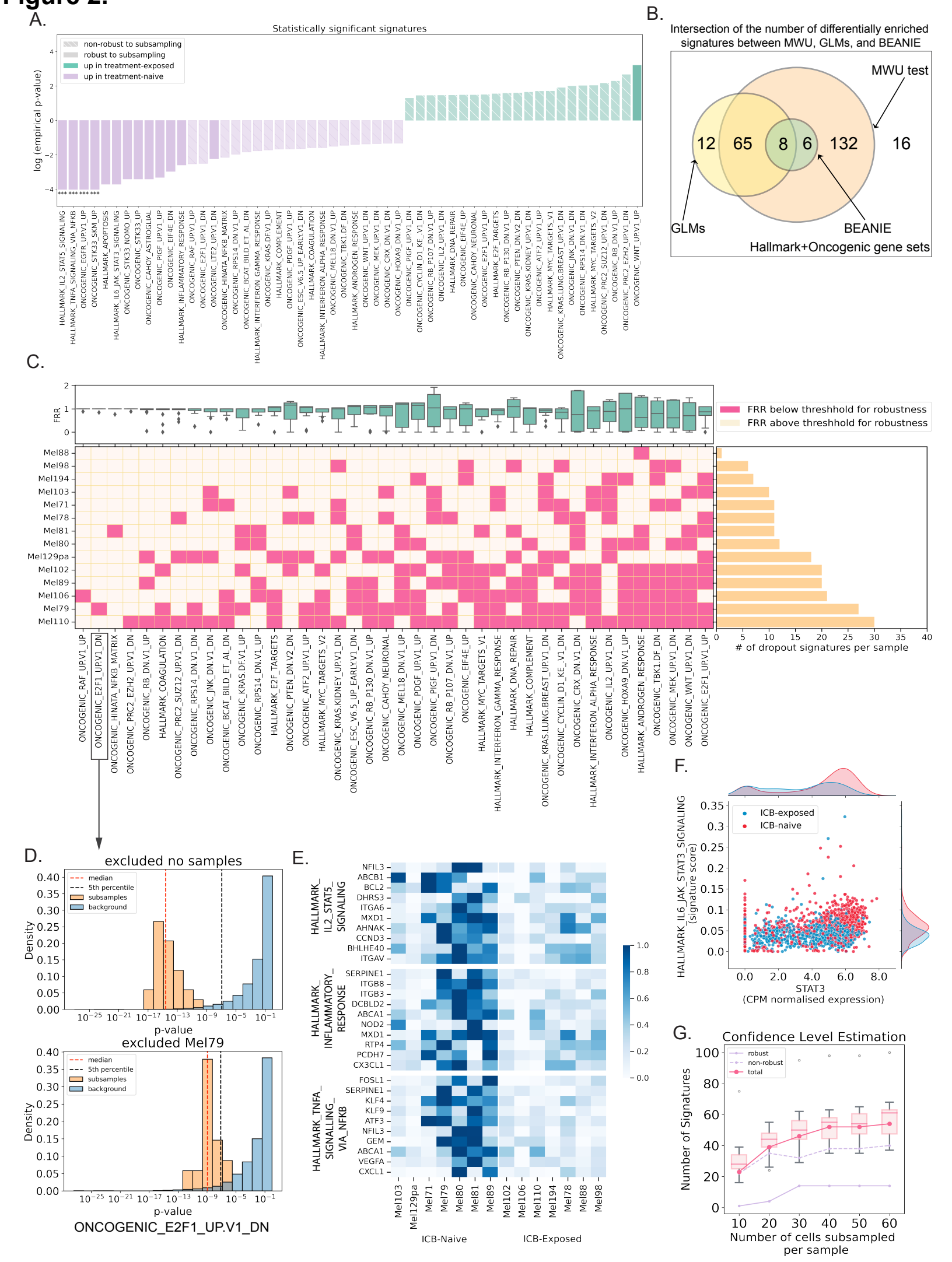

A.
Statistically significant signatures
non-robust to subsampling
robust to subsampling
up in treatment-exposed
up in treatment-naive
log (empirical p-value)
B.
Intersection of the number of differentially enriched signatures between MWU, GLMs, and BEANIE
MWU test
12
65
8
6
132
16
GLMs
BEANIE
Hallmark+Oncogenic gene sets
C.
FRR
FRR below threshhold for robustness
FRR above threshhold for robustness
# of dropout signatures per sample
D.
excluded no samples
median
5th percentile
subsamples
background
Density
p-value
excluded Mel79
ONCOGENIC_E2F1_UP.V1_DN
E.
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_TNFA_SIGNALLING_VIA_NFKB
ICB-Naive
ICB-Exposed
F.
ICB-exposed
ICB-naive
HALLMARK_IL6_JAK_STAT3_SIGNALING (signature score)
STAT3 (CPM normalised expression)
G.
Confidence Level Estimation
robust
non-robust
total
Number of Signatures
Number of cells subsampled per sample

Figure 3.

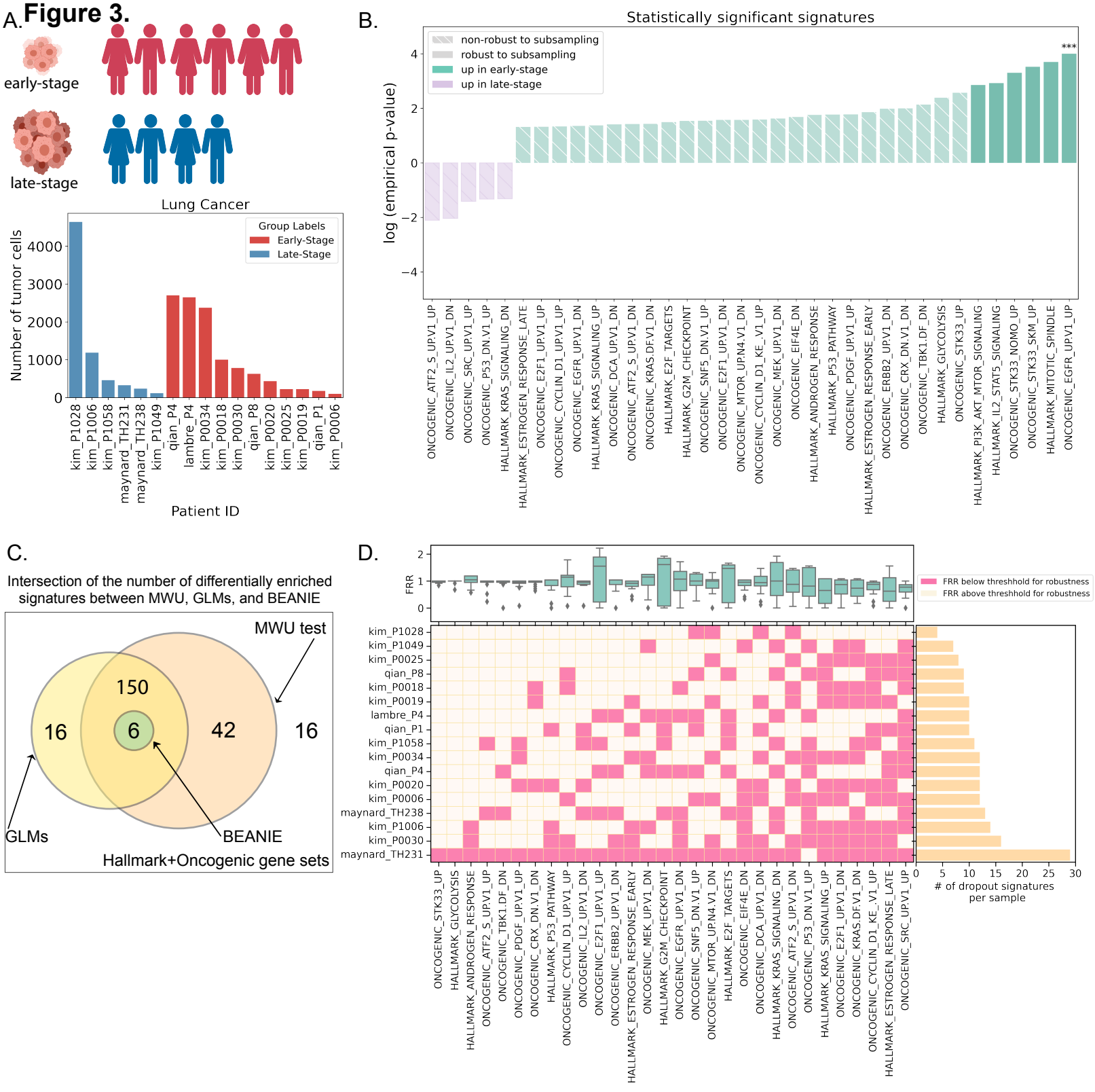
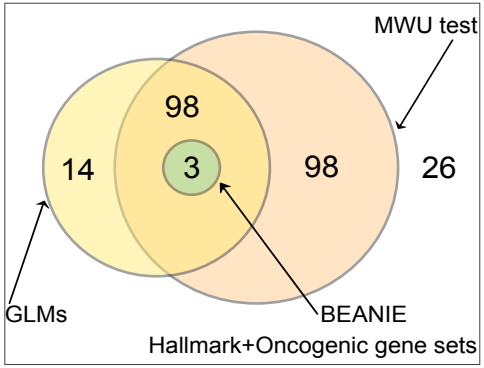
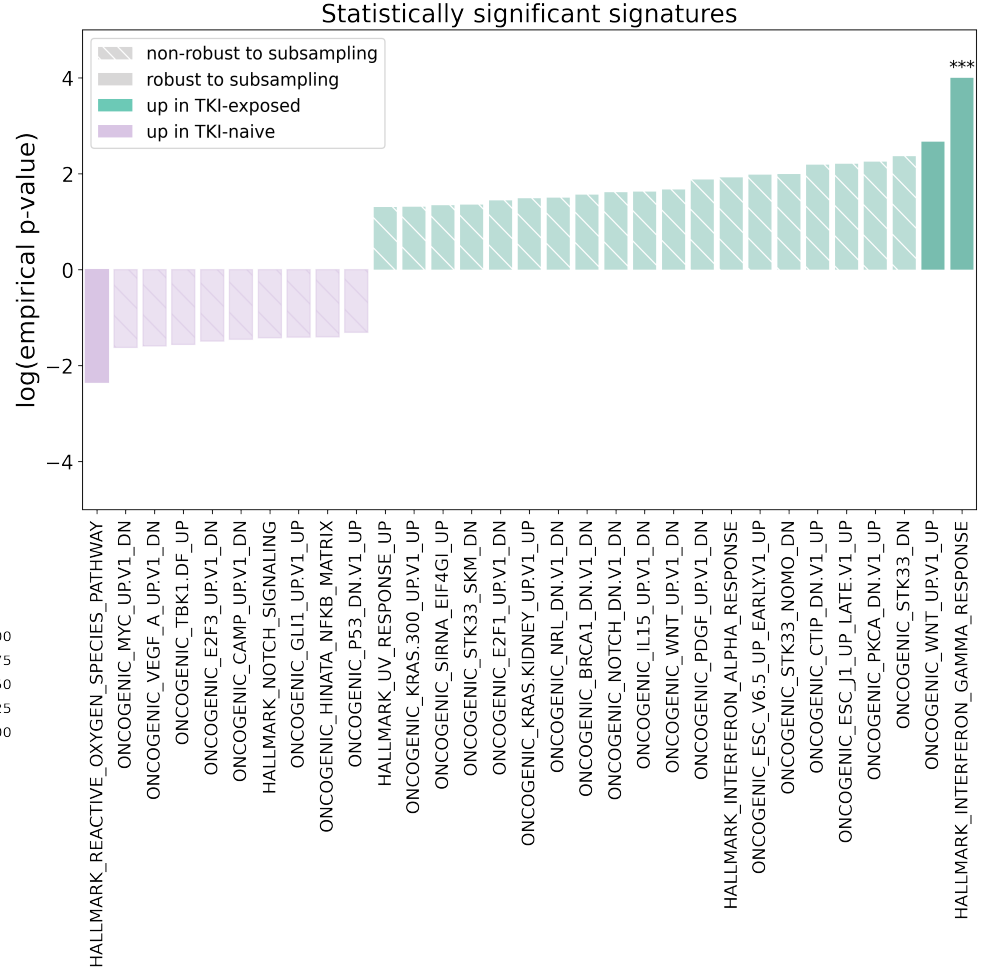

A.
early-stage
late-stage
Lung Cancer
Group Labels
Early-Stage
Late-Stage
Number of tumor cells
Patient ID
kim_P1028
kim_P1006
kim_P1058
maynard_TH231
maynard_TH238
kim_P1049
qian_P4
lambre_P4
kim_P0034
kim_P0018
kim_P0030
qian_P8
kim_P0020
kim_P0025
kim_P0019
qian_P1
kim_P0006
B.
Statistically significant signatures
non-robust to subsampling
robust to subsampling
up in early-stage
up in late-stage
log (empirical p-value)
***
ONCOGENIC_ATF2_S_UP.V1_UP
ONCOGENIC_IL2_UP.V1_DN
ONCOGENIC_SRC_UP.V1_UP
ONCOGENIC_P53_DN.V1_UP
HALLMARK_KRAS_SIGNALING_DN
HALLMARK_ESTROGEN_RESPONSE_LATE
ONCOGENIC_E2F1_UP.V1_UP
ONCOGENIC_CYCLIN_D1_UP.V1_UP
ONCOGENIC_EGFR_UP.V1_DN
HALLMARK_KRAS_SIGNALING_UP
ONCOGENIC_DCA_UP.V1_DN
ONCOGENIC_ATF2_S_UP.V1_DN
ONCOGENIC_KRAS.DF.V1_DN
HALLMARK_E2F_TARGETS
HALLMARK_G2M_CHECKPOINT
ONCOGENIC_SNF5_DN.V1_UP
ONCOGENIC_E2F1_UP.V1_DN
ONCOGENIC_MTOR_UP.N4.V1_DN
ONCOGENIC_CYCLIN_D1_KE_V1_UP
ONCOGENIC_MEK_UP.V1_DN
ONCOGENIC_EIF4E_DN
HALLMARK_ANDROGEN_RESPONSE
HALLMARK_P53_PATHWAY
ONCOGENIC_PDGF_UP.V1_UP
HALLMARK_ESTROGEN_RESPONSE_EARLY
ONCOGENIC_ERBB2_UP.V1_DN
ONCOGENIC_CRX_DN.V1_DN
ONCOGENIC_TBK1.DF_DN
HALLMARK_GLYCOLYSIS
ONCOGENIC_STK33_UP
HALLMARK_PI3K_AKT_MTOR_SIGNALING
HALLMARK_IL2_STAT5_SIGNALING
ONCOGENIC_STK33_NOMO_UP
ONCOGENIC_STK33_SKM_UP
HALLMARK_MITOTIC_SPINDLE
ONCOGENIC_EGFR_UP.V1_UP
C.
Intersection of the number of differentially enriched signatures between MWU, GLMs, and BEANIE
MWU test
150
16
6
42
16
GLMs
BEANIE
Hallmark+Oncogenic gene sets
D.
FRR
FRR below threshhold for robustness
FRR above threshhold for robustness
# of dropout signatures per sample

Figure 4.

A. Intersection of the number of differentially enriched signatures between MWU, GLMs, and BEANIE



B.



C.

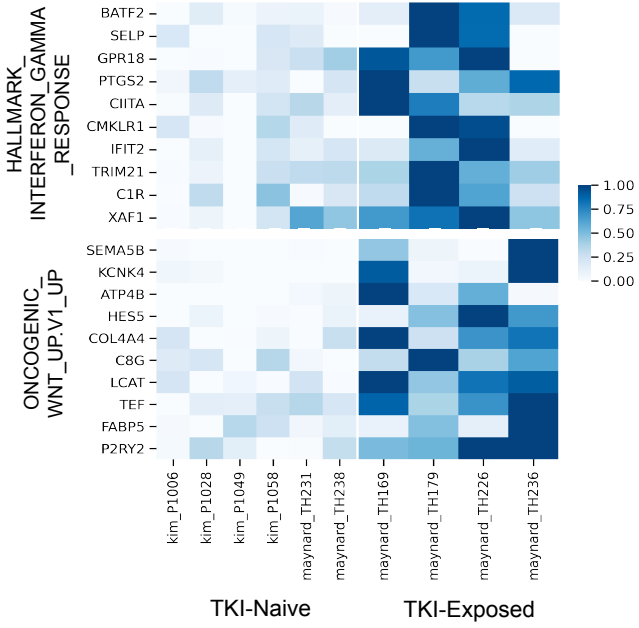


Figure 5.

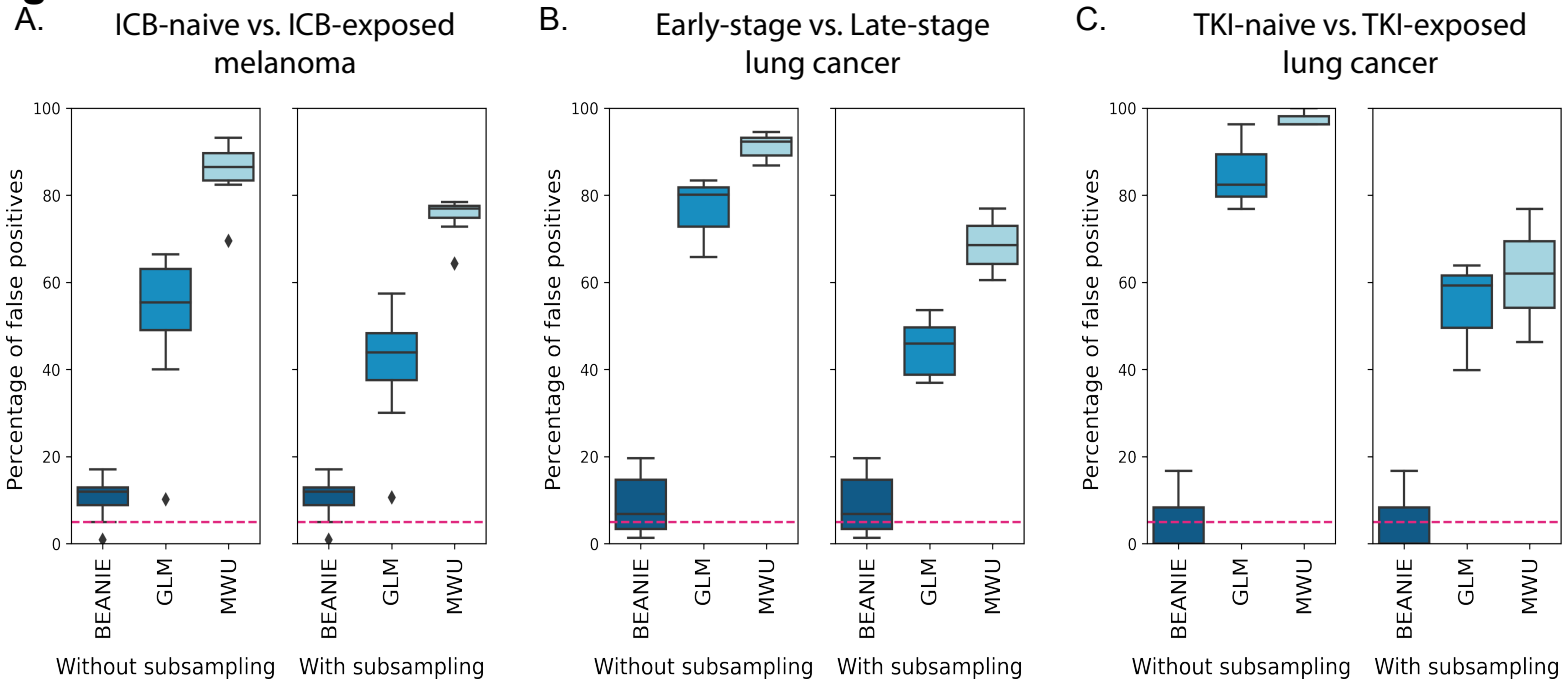
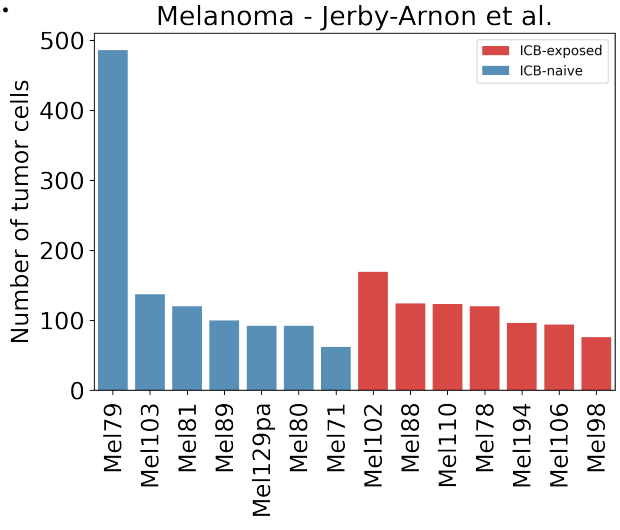

A.
ICB-naive vs. ICB-exposed melanoma
B.
Early-stage vs. Late-stage lung cancer
C.
TKI-naive vs. TKI-exposed lung cancer
Percentage of false positives
0
20
40
60
80
100
BEANIE
GLM
MWU
Without subsampling
With subsampling

Figure S1.

A.


Melanoma - Jerby-Arnon et al.
ICB-exposed
ICB-naive
Number of tumor cells
0
100
200
300
400
500
Mel79
Mel103
Mel81
Mel89
Mel129pa
Mel80
Mel71
Mel102
Mel88
Mel110
Mel78
Mel194
Mel106
Mel98

B.

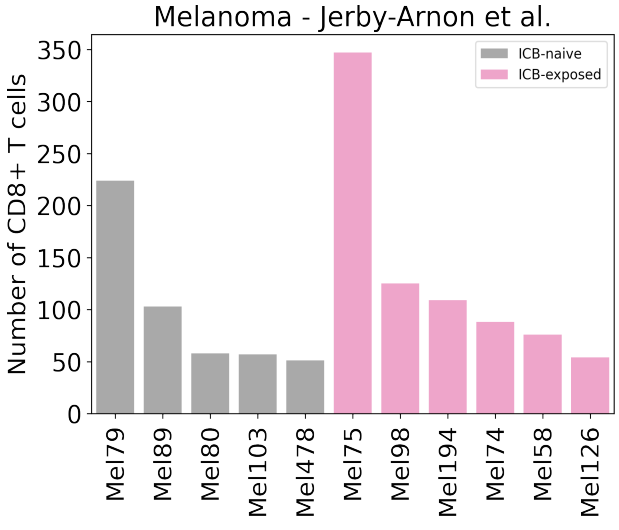

Melanoma - Jerby-Arnon et al.
ICB-naive
ICB-exposed
Number of CD8+ T cells
0
50
100
150
200
250
300
350
Mel79
Mel89
Mel80
Mel103
Mel478
Mel75
Mel98
Mel194
Mel74
Mel58
Mel126

Figure S2.

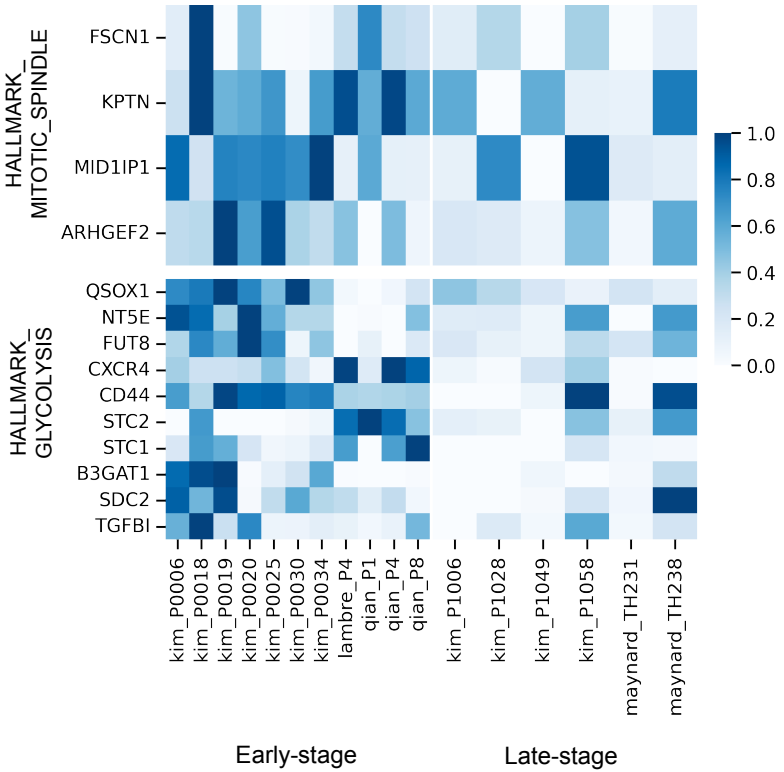
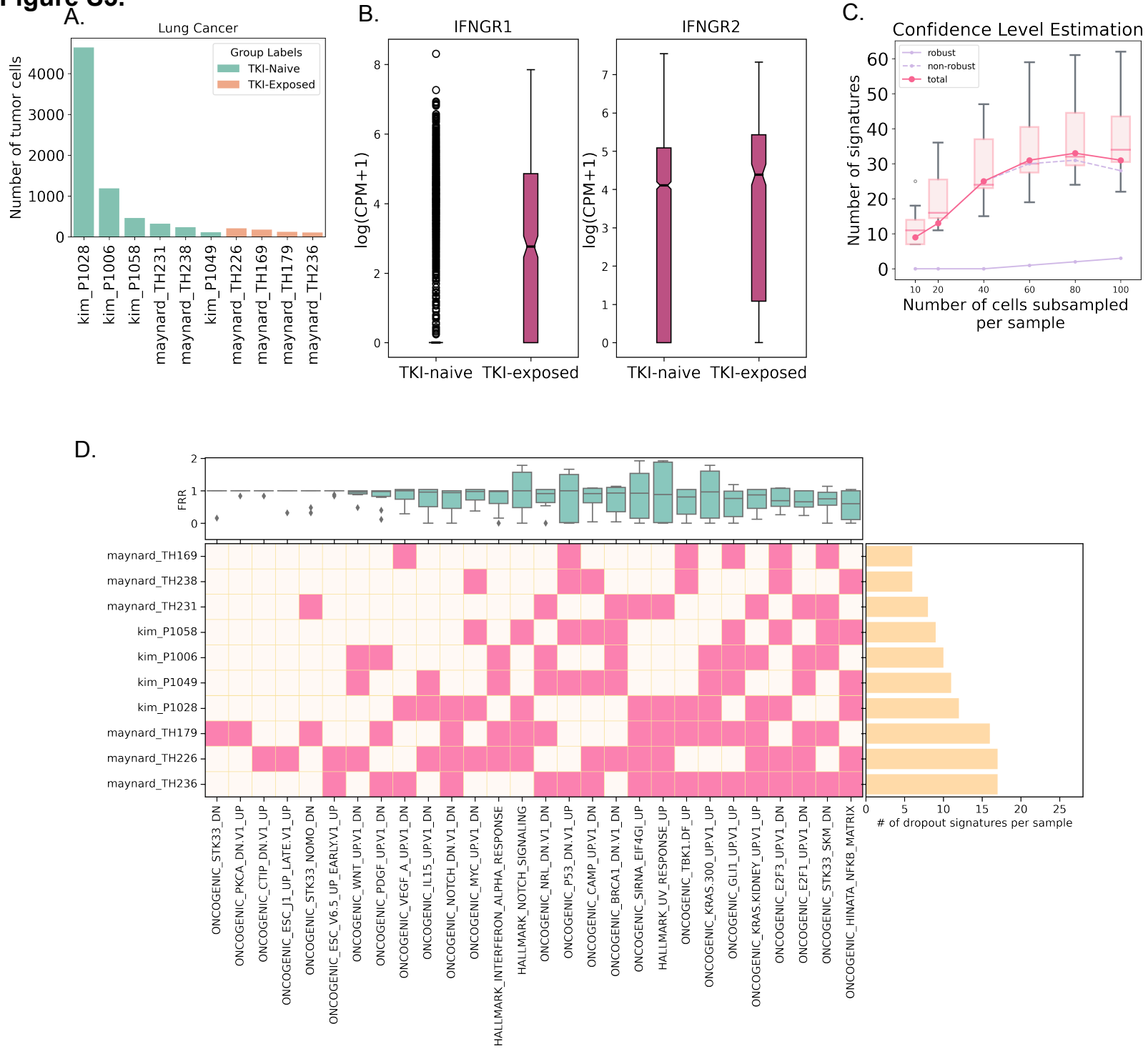

HALLMARK_
MITOTIC_SPINDLE
FSCN1
KPTN
MID1IP1
ARHGEF2
HALLMARK_
GLYCOLYSIS
QSOX1
NT5E
FUT8
CXCR4
CD44
STC2
STC1
B3GAT1
SDC2
TGFBI
kim_P0006
kim_P0018
kim_P0019
kim_P0020
kim_P0025
kim_P0030
kim_P0034
lambre_P4
qian_P1
qian_P4
qian_P8
kim_P1006
kim_P1028
kim_P1049
kim_P1058
maynard_TH231
maynard_TH238
Early-stage
Late-stage
1.0
0.8
0.6
0.4
0.2
0.0

Figure S3.


A.
Lung Cancer
Group Labels
TKI-Naive
TKI-Exposed
Number of tumor cells
kim_P1028
kim_P1006
kim_P1058
maynard_TH231
maynard_TH238
kim_P1049
maynard_TH226
maynard_TH169
maynard_TH179
maynard_TH236
B.
IFNGR1
IFNGR2
log(CPM+1)
TKI-naive
TKI-exposed
C.
Confidence Level Estimation
robust
non-robust
total
Number of signatures
Number of cells subsampled per sample
D.
FRR
# of dropout signatures per sample
ONCOGENIC_STK33_DN
ONCOGENIC_PKCA_DN.V1_UP
ONCOGENIC_CTIP_DN.V1_UP
ONCOGENIC_ESC_J1_UP_LATE.V1_UP
ONCOGENIC_STK33_NOMO_DN
ONCOGENIC_ESC_V6.5_UP_EARLY.V1_UP
ONCOGENIC_WNT_UP.V1_DN
ONCOGENIC_PDGF_UP.V1_DN
ONCOGENIC_VEGF_A_UP.V1_DN
ONCOGENIC_IL15_UP.V1_DN
ONCOGENIC_NOTCH_DN.V1_DN
ONCOGENIC_MYC_UP.V1_DN
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_NOTCH_SIGNALING
ONCOGENIC_NRL_DN.V1_DN
ONCOGENIC_P53_DN.V1_UP
ONCOGENIC_CAMP_UP.V1_DN
ONCOGENIC_BRCA1_DN.V1_DN
ONCOGENIC_SIRNA_EIF4GI_UP
HALLMARK_UV_RESPONSE_UP
ONCOGENIC_TBK1.DF_UP
ONCOGENIC_KRAS.300_UP.V1_UP
ONCOGENIC_GLI1_UP.V1_UP
ONCOGENIC_KRAS.KIDNEY_UP.V1_UP
ONCOGENIC_E2F3_UP.V1_DN
ONCOGENIC_E2F1_UP.V1_DN
ONCOGENIC_STK33_SKM_DN
ONCOGENIC_HINATA_NFKB_MATRIX

729 **Tables**

		MWU test	GLM	BEANIE
Melanoma (Immune Checkpoint Blockade (ICB)-naive vs. ICB-exposed)	Hallmark gene sets	45/50	27/50	5/50
	Oncogenic gene sets	166/189	58/189	9/189
Lung (Early-stage vs. Late-stage)	Hallmark gene sets	46/50	44/50	3/50
	Oncogenic gene sets	168/189	128/189	3/189
Lung (Tyrosine Kinase Inhibitor (TKI) - naive vs. TKI-exposed)	Hallmark gene sets	47/50	33/50	2/50
	Oncogenic gene sets	152/189	82/189	1/189

730 ***Table 1. Number of differentially enriched signatures identified with the three methods***

731 ***(MWU test, GLMs, and BEANIE) using Hallmark (n = 50) and Oncogenic (n = 189) gene sets.***

732

733

	MWU test + BH correction		GLMs		BEANIE
	Without subsampling	With subsampling	Without subsampling	With subsampling	
ICB-naive vs. ICB-exposed	85.74%	75.57%	52.58%	41.72%	10.63%
Early-stage vs. Late-stage	91.23%	68.6%	77.03%	44.93%	9.01%
TKI-naive vs. TKI-exposed	97.53%	61.72%	85.18%	54.32%	5.55%

734

735 ***Table 2. Average false positive rate for the three datasets (ICB-naive vs. ICB-exposed***
736 ***melanoma, early-stage vs. late-stage lung cancer, and TKI-naive vs. TKI-exposed lung***
737 ***cancer) across the three methods (MWU test with a BH correction, GLMs, and BEANIE).***

738