Chakravarthy, Ankur, Furness, Andrew, Joshi, Kroopa, Ghorani, Ehsan, Ford, Kirsty, Ward, Matthew J., King, Emma V., Lechner, Matt, Marafioti, Teresa, Quezada, Sergio A. and others (2018) *Pan-cancer deconvolution of tumour composition using DNA methylation.* Nature Communications, 9 (1). ISSN 2041-1723.

Kent Academic Repository
Full text document (pdf)

Link to record in KAR

http://kar.kent.ac.uk/68522/

Document Version

Author's Accepted Manuscript

1 Pan-cancer deconvolution of tumour composition using DNA methylation

2 Ankur Chakravarthy[1,§], Andrew Furness[2], Kroopa Joshi[2], Ehsan Ghorani[2], Kirsty Ford[3], Matthew J

3 Ward[3], Emma V King[3], Matt Lechner[1], Teresa Marafioti[4], Sergio A. Quezada[2], Gareth J Thomas[3],

4 Andrew Feber[5], Tim R Fenton[6*]

5

6 [1] Department of Oncology, UCL Cancer Institute, University College London, WC1E 6BT, UK

7 [2] Department of Haematology, UCL Cancer Institute, University College London, WC1E 6BT, UK

8 [3] Cancer Sciences Unit, University of Southampton, Tremona Road, Southampton SO16 6YD, UK

9 [4] Department of Pathology, UCL Cancer Institute, University College London, WC1E 6BT, UK

10 [5] Division of Surgery and Interventional Science, University College London, WC1E 6BT, UK

11 [6] School of Biosciences, University of Kent, Canterbury, CT2 7NJ, UK

12

13 [§] Present address: Princess Margaret Cancer Centre, Toronto, Ontario, M5G 2C4, Canada

14

15 * To whom correspondence should be addressed: t.fenton@kent.ac.uk

16

17

18

19

20

21

Abstract

The nature and extent of immune cell infiltration into solid tumours are key determinants of therapeutic response. Here, using a DNA methylation-based approach to tumour cell fraction deconvolution, we report the integrated analysis of tumour composition and genomics across a wide spectrum of solid cancers. Initially studying head and neck squamous cell carcinoma, we identify two distinct tumour subgroups: 'immune hot' and 'immune cold', which display differing prognosis, mutation burden, cytokine signalling, cytolytic activity, and oncogenic driver events. We demonstrate the existence of such tumour subgroups pan-cancer, link clonal-neoantigen burden to cytotoxic T-lymphocyte infiltration, and show that transcriptional signatures of hot tumours are selectively engaged in immunotherapy responders. We also find that treatment-naive hot tumours are markedly enriched for known immune-resistance genomic alterations, potentially explaining the heterogeneity of immunotherapy response and prognosis seen within this group. Finally, we define a catalogue of mediators of active antitumour immunity, deriving candidate biomarkers and potential targets for precision immunotherapy.

Introduction

The tumour microenvironment plays key roles in shaping tumour evolution and in determining treatment responses; prominent intratumoural lymphocyte infiltration is a favourable prognostic marker in multiple tumour types, while a high stromal content of extracellular matrix-producing cancer-associated fibroblasts (CAF), is associated with poor outcomes[1]. The recent clinical success of immunotherapy in subpopulations of patients with previously intractable malignancies has also highlighted the importance of understanding the tumour microenvironment in order to identify those

46   patients who will derive most benefit from targeted therapies[1,2]. Although responses to immune

47   checkpoint blockade (ICB), e.g. antibodies against PD-1 (programmed cell death protein 1), PD-L1

48   (programmed death-ligand 1) and CTLA-4 (cytotoxic T-lymphocyte-associated protein 4), are seen

49   across many solid tumours, the proportion of patients that benefit varies widely by cancer type and

50   we currently lack biomarkers with which to reliably predict immunotherapy response[3]. Emerging

51   evidence from clinical trials indicates higher response rates in those cancer types that typically display

52   greater lymphocyte infiltration (e.g. melanoma, lung cancer, head and neck cancer) and that the

53   tumour neoantigen repertoire (a function of somatic mutation load) is a key determinant[4-7]. These

54   observations point to a model in which, within any given cancer type, there are immune hot and

55   immune cold tumours. Immune hot tumours display greater cytotoxic T-lymphocyte (CTL) infiltration,

56   and reactivation of these tumour-resident CTLs by checkpoint inhibition can result in dramatic tumour

57   regression. Conversely, immune cold tumours display minimal CTL infiltrates and typically fail to

58   respond to checkpoint modulation. If one could accurately identify likely responders for patient

59   stratification and devise strategies by which to convert cold tumours to hot tumours, these would be

60   major steps forward in realising the full clinical potential of cancer immunotherapy.

61

62   Although flow cytometry of disaggregated tumour biopsies is commonly used for investigating cellular

63   composition, this is often unfeasible for several reasons; difficulty in obtaining fresh tumour tissue;

64   lack of defined markers for poorly characterised cell types (e.g. CAFs); and high cost of labour, reagents

65   and equipment required for such analyses. Cellular disaggregation of collagen-rich tumours is also

66   problematic, where cells are embedded in a dense extracellular matrix. To overcome these difficulties,

67   multiple reference-free or reference-based methods have recently been developed to permit the *in-*

68   *silico* deconvolution of complex cellular mixtures or to estimate tumour purity[8-15]. For example,

69   accurate deconvolution of complex cellular mixtures, including tumours, has recently been achieved

70   by application of support vector regression modelling (CIBERSORT) to gene expression microarray

71   data[14,16]. Notably, DNA methylation data are also suitable for deconvolution of tissue mixtures,

72   although studies so far have focussed primarily on simple tissues such as blood, where cell type

73   differences are a major confounder in Epigenome Wide Association Studies[10].

74

75   Here we present CIBERSORT-based deconvolution to genome-wide DNA methylation data from whole

76   tumour tissue (hereafter referred to as 'MethylCIBERSORT'). We provide accurate estimates of tumour

77   purity and cellular composition, and identify immune hot and cold tumours across a broad spectrum

78   of cancer types profiled by The Cancer Genome Atlas Project (TCGA). Using matched genomic and

79   transcriptomic data, we identify multiple copy number alterations enriched in cold tumours, including

80   deletions in *PTEN* and amplifications in *MYC* and *EGFR.* We show that responses to PD1-blockade are

81   associated with a transcriptional signature for hot tumours post-treatment, while the cold signature,

82   and specifically a gene expression module we previously linked to increased aerobic glycolysis

83   downstream of EGFR in HNSCC[17], is enriched in non-responders. Importantly however, defining

84   whether a tumour is hot or cold is not sufficient to accurately predict response to immune checkpoint

85   blockade (ICB), and by interrogating matched genomic data, we show that treatment-naive hot

86   tumours frequently display genomic alterations known to confer immunotherapy resistance.

87

Results

DNA methylation-based tumour deconvolution using CIBERSORT

90   To develop a DNA methylation based deconvolution pipeline for application in tumours, we created a

91   custom R interface to produce basis matrices for use with CIBERSORT and generated a reference using

92   fibroblasts and seven different immune cell types (see methods for details). We then evaluated the

93   ability of our feature selection heuristic to accurately deconvolute mixtures of leukocytes using

94   publicly available methylation data from mixtures of peripheral blood mononuclear cells (PBMCs) with

95   composition verified by flow-cytometry (gold standard). This showed extremely high correlation

96   between the estimated and gold-standard fractions (Pearson's R = 0.986, p < 2.2e-16, **Figure 1a**). We

97   also carried out benchmarking against the performance of RNA-based CIBERSORT using the LM22

98   basis matrix against leukocyte mixtures of similar resolution originally profiled in Newman et al[14]. This

99   revealed that MethylCIBERSORT estimates demonstrate higher correlations, both at the cell-type and

100 the sample level (**Figure 1b, c**) and significantly lower absolute error (**Figure 1d**). Thus, methylation

101 data coupled to CIBERSORT is highly accurate and may offer distinct advantages relative to expression-

102 based CIBERSORT.

103

104 To validate the method on real tumour samples, especially the tumour content in order to permit

105 absolute quantification of tumour composition, we focused initially on head and neck squamous cell

106 carcinoma (HNSCC), a tumour type in which we've previously demonstrated the prognostic

107 significance of tumour-infiltrating lymphocytes (TILs), particularly in those cancers driven by human

108 papillomavirus (HPV)[18,19]. We applied our pipeline to generate an HNSCC specific basis matrix and

109 applied it to the set of 464 HNSCCs that have both RNA-sequencing and DNA methylation profiles

110 available from TCGA[20]. Upon comparing cancer cell proportion (purity) estimates derived using

111 MethylCIBERSORT with estimates derived from ABSOLUTE[21] (which jointly estimates purity and ploidy

112 using mutation and copy number data) relative to other previously published methods of estimating

113 purity (LUMP[22] and ESTIMATE[23]) with data aggregated in[22], MethylCIBERSORT displayed the highest

114 correlation (R=0.82) and better concordance with ABSOLUTE than other methods (**Figure 1e**). Analysis

115 of residuals (method estimate – ABSOLUTE estimate) suggested close concordance with ABSOLUTE

116 estimates for MethylCIBERSORT, with larger deviations only seen when samples were of very high

117 purity (>80%), while other methods tended to overestimate tumour cell content in samples of low

118 purity (**Supplementary Figure 1a**), resulting in statistically significant differences in distributions (FDR

119 < 2.2e-16).

120

121   We also compared the mRNA expression of a panel of cellular lineage markers with MethylCIBERSORT

122   estimates and found significant associations for multiple cell types (**Supplementary Figure 1b**) even

123   though they are derived from different samplings of the same tumour. Many of these marker genes

124   demonstrated more variable expression in tumours with lower estimates of infiltrating cell fraction,

125   suggesting that low coverage on either or both platforms (RNA-seq and methylation array) at the

126   lower end of cellular abundance may result in poorer concordance. Taken together, these

127   observations confirm that MethylCIBERSORT can accurately deconvolute the mixed cell populations

128   in tumour samples using DNA methylation data.

129

130   Having established the potential of MethylCIBERSORT to identify patterns of cellular infiltration in

131   solid tumours, we tested ts ability to detect the elevated TIL levels previously documented in HPV-

132   driven (HPV+) HNSCC[19]. MethylCIBERSORT detected not only the increased TIL levels in HPV+ HNSCC

133   compared with HPV- HNSCC (p= 2.167e-05 , Wilcoxon's Rank Sum Test) but more specifically

134   attributed this to increased numbers of B (CD19+) and cytotoxic T (CD8+) lymphocytes (CTLs, **Figure**

135   **1f**), in agreement with observations made using other methods, including immunohistochemistry and

136   gene expression analysis[24], potentially also helping to explain favourable prognosis displayed by this

137   subgroup, independent of treatment modality[25-27].

138

139   ***Cellular infiltration patterns in HPV-negative HNSCC***

140   Next, we extended our analysis to HPV-negative (HPV-) HNSCC, a heterogeneous, anatomically-diverse

141   group of tumours in which prognosis is typically much poorer than in HPV+ disease. Again, using TCGA

142   data (available for 398 HPV- HNSCCs) we observed interesting relationships between multiple cell

143   types, with 24/36 pairs of cell types showing significant correlations (Spearman's rank correlations,

144 FDR<0.1; **Figure 1g**). CTLs are associated with both CD14+ (monocytes / macrophages / myeloid-

145 derived suppressor cells) and B-lymphocytes (Rho =0.2 and 0.51). CD4+/FoxP3- T-lymphocytes (CD4 +

146 effector T-lymphocytes), meanwhile display inverse correlations with CTLs (R= -0.41) and Tregs (R=-

147 0.38). CD56+ Natural Killer (NK) cell abundance is also inversely correlated with CTLs (R=-0.45). Of

148 note, CTLs are inversely correlated with fibroblast abundance (R=-0.12) and to validate this latter

149 finding, we analysed data from two large studies in which these parameters had been quantified in

150 HNSCC[19,28]. In a pooled analysis of these data, TIL content and SMA expression (a CAF marker) are

151 inversely correlated (r=-0.322 and -0.344 for CD8 and CD3 IHC in the Ward (oropharyngeal SCC) cohort

152 (**Figure 1h, top panels**); -0.4 and -0.424 for TIL scoring of H&E sections in the Ward (oropharyngeal

153 SCC) and Marsh (oral SCC) cohorts respectively). They are also strongly prognostic (**Figure 1h, lower**

154 **panels**; p<0.001, Log Rank Test).

155 Given the complex nature of associations between different cell types, we performed consensus PAM

156 clustering on the estimated cellular fractions to define subgroups by infiltration patterns. We derived

157 two clusters (immune cold and immune hot, hereafter referred to as cold and hot respectively) that

158 show markedly different distributions of multiple cell types, most notably CTLs, Tregs, CD4+ effector

159 T-lymphocytes, CD19+ B-lymphocytes and NK cells, all of which are implicated in antitumour immunity

160 (**Figure 2a**). Consistent with our previous observations, estimates of fibroblast content are higher in

161 the cold group (mean fold change 1.35, FDR < 5e-7, Wilcoxon's Rank Sum Test). To explore the

162 functional significance of our observations, we tested for associations between individual cellular

163 fractions or immune cluster and a recently defined measure of local cytolytic activity based on the

164 expression of Granzyme A and Perforin 1 (*GZMA* and *PRF1*; markers of activated T-cells)[29]. Most

165 infiltrating cell fractions display significant correlations with cytolytic activity, with CD8+ cells showing

166 the maximum positive correlation (**Figure 2b**, FDR < 0.05, Spearman's Rank Correlation). Accordingly,

167 the hot cluster displays significantly higher cytolytic activity (**Figure 2c**, p = 2e-16, Wilcoxon's Rank Sum

168 test), and increased ratios of CTLs to Tregs (**Figure 2d**, p < 2e-16, Wilcoxon's Rank Sum test); a metric

169 that is prognostic in multiple settings[30-32].

170

171   Together, our analyses suggest a tilting of the balance towards CTL activity in the microenvironment

172   of tumours from the hot cluster. Integrated analysis of the impact of different cell populations on

173   cytolytic activity using linear modelling identified CD8+ (coef = 0.13, p < 2e-16), CD14+ (coef = 0.10, p

174   <7e-7), and CD56+ (coef = 0.26, p < 5e-15) cell abundance as positive predictors and fibroblast

175   abundance as a negative predictor (coef = -0.08, p <1e-10).

176

177   ***Transcriptional and proteomic differences between hot and cold HNSCCs***

178   Having observed differential abundance of various leukocytes in hot versus cold HNSCC using

179   MethylCIBERSORT, we used gene expression data to further validate these estimates. Using limma-

180   trend analysis, we identified 458 genes differentially expressed (DEGs) between the hot and cold

181   clusters at a fold-change of greater than 2 (FDR<0.01**, Supplementary Data 1**, genes highlighted in

182   bold). Multiple DEGs are consistent with the MethylCIBERSORT-derived estimates of lymphocyte

183   infiltration; *CD8A, ZAP70* and *CD3D* (CD8 lymphocyte markers), *CD79A* and *CD19* (B-lymphocyte

184   markers), are all upregulated in the hot tumours, as are multiple chemokines and their receptors

185   (*CCL5, CCR5, CXCR5, CXCR6, CCL19, CXCL11)*), immune checkpoint gene transcripts (*LAG3, PD1, IDO1*),

186   and as expected, the cytolytic markers *PRF1* and *GZMA*. In extended analyses of all genes at FDR <

187   0.01 **(Supplementary Data 1)**, multiple other genes, including the Class 1 MHC gene *B2M* (FC = 1.39),

188   PD-1 ligand *CD274* (FC = 1.62) and *ACTA2*, which encodes SMA *(*FC = 0.68), are also differentially

189   expressed between the two clusters, the latter validating fibroblast estimates from MethylCIBERSORT

190   (**Figure 2a**).

191

192   Ingenuity Pathway Analysis further confirmed observations made using MethylCIBERSORT estimates,

193   identifying differential regulation of multiple canonical pathways associated with immune function

194 and inflammatory conditions **(Supplementary Data 2)**, consistent with differential lymphocyte

195 infiltration and activity. Diseases and functions ontology (**Supplementary Data 3)** indicated that the

196 top few pathways activated in hot tumours were associated with leukocyte and lymphocyte migration.

197 Upstream regulatory analysis implicated increased activation of STAT1 and IRF1, and inhibition of

198 Interferon-stimulated transcription mediated by IRF4, in hot tumours **(Supplementary Data 4)**. Finally,

199 analysis of RPPA data identified 10 differentially abundant (FDR<0.1) proteins or phospho-proteins

200 **(Supplementary Data 5)**. Higher levels of cleaved Caspase 7 (FC=1.52) in the hot subgroup indicates

201 increased apoptosis**,** whereas Fibronectin and PAI1 upregulation in immune cold tumours suggest a

202 distinct pattern of TGFβ-driven extracellular matrix remodelling in what may be a CAF-linked

203 phenomenon.

204
205 ***Distinct mutations are associated with HNSCC immune cluster***
206
207 Having established that the two immune clusters display distinct transcriptional patterns, we then

208 sought to identify individual mutations in driver genes (MutSig CV[33] q.value < 0.01) associated with

209 immune cluster using Negative Binomial regression. This identified enrichment of *CASP8, PIK3CA,*

210 *CREBBP, EP300 and HLA-A* mutations in hot HNSCCs, while *TP53* and *KDM6A* mutations are enriched

211 in cold HNSCCs (**Figure 2e**). *CASP8* mutations are implicated in subverting apoptosis induced by

212 lymphocytes; they are enriched in tumours with high immune cytolytic activity and likely reflect an

213 increased selective pressure exerted by the presence of adaptive immune cells[29,34]. Fas-ligand *(FASLG)*,

214 an upstream activator of pro-apoptotic signalling through Caspase 8 is also upregulated in the immune

215 hot tumours, further highlighting the importance of this pathway (**Supplementary Data 1**).

216 Identification of this lymphocyte-rich, good p group displaying *CASP8* mutations and a relative lack of

217 *TP53* mutations is striking, since TCGA previously identified a subset of good-prognosis oral cavity

218 tumours bearing the same genomic hallmarks, which were reported to co-occur with *HRAS*

219 mutations[20].

220

221   Neoantigen burden has previously been identified as a predictor of anti-tumour immune

222   responses[29,35,36] and consistent with this, we identified significantly higher predicted MHC class I

223   neoantigen burdens in the hot tumours (OR = 1.56, p < 5.5e-8, Negative Binomial GLM) and a smaller

224   increase in overall mutational burden (OR = 1.46, p = 6e-7). Moreover, in 15 tumours from the hot

225   cluster versus 5 in the cold cluster, *CASP8* mutations themselves encoded at least one neoantigenic

226   peptide (**Figure 2e**), demonstrating the existence of mutations that could both contribute to the

227   development of a potential selective constraint, and serve as an adaptive mechanism to evade it.

228

229   ***Deconvolution and immune clustering across tumour types***

230   To examine whether the relationships between tumour composition, genomic alterations and clinical

231   behaviour we observed in HNSCC are generally applicable, we derived cancer-type specific basis

232   matrices and conducted deconvolution on 18 further tumour types for which cancer cell line

233   methylation data have recently been published[37]. For 9 of these we were able to compare our

234   predictions of tumour purity with ABSOLUTE estimates and observed strong correlations and

235   significantly lower error margins compared to LUMP and ESTIMATE (**Supplementary Figure 2a, b**).

236   Further, we observed a robust preservation of positive correlations between MethylCIBERSORT and

237   marker expression pan-cancer (**Supplementary Figure 2c**), again with the caveat that the samples

238   were taken from different aliquots of the tumour. Taken together, these findings attest to the general

239   pan-cancer applicability of MethylCIBERSORT. An important potential advantage of DNA-methylation

240   over gene expression-based deconvolution methods is the ease with which accurate DNA methylation

241   profiles can be obtained from formalin-fixed, paraffin-embedded (FFPE) samples[38]. We therefore

242   compared estimates pertaining to fresh frozen and matched FFPE samples (n = 21 from 3 tumours)[39]

243   and recorded very high correlations, indicating our method is applicable also to archival material

244   (**Supplementary Figure 2d**).

245

246 We then trained an elastic-net classifier using 5-fold cross-validation for tuning on the HNSCC cellular

247 abundance data, returning highly accurate recapitulation of clustering (Kappa = 0.9), and predicted

248 immune cluster membership for the validation set of 7596 samples representing 21 further tumour

249 types from TCGA (**Figure 3a**). As expected, we observed strong enrichment for CTLs, Tregs, and B-

250 lymphocytes in hot tumours pan-cancer, while CD4-effectors, NK cells, eosinophils and CAFs were

251 enriched in immune cold tumours (**Figure 3b**). Different tumour types also display markedly varying

252 degrees of lymphocyte infiltration, with the majority of pancreatic ductal adenocarcinomas,

253 colorectal, thyroid, uterine corpus endometrial, kidney, prostate, hepatocellular cancers and

254 sarcomas belonging to the cold cluster (**Figure 3a**). We again observed increased CTL:Treg ratios in hot

255 tumours (**Figure 3c**) and similar relationships between tumour composition and CYT to those seen in

256 HNSCC.

257

258 ***Increased immunoediting and Th1/M1 responses in hot tumours***

259 To further determine if the immune infiltrate was active in these tumours, we assayed immunoediting

260 by testing for reductions from the expected ratio (as previously defined by Rooney et al[29]) of observed

261 neoantigens to total nonsilent mutations per tumour and adapted this approach to derive the

262 estimated number of neoepitopes lost through immune editing while controlling for tumour type.

263 Accordingly, we found significant enrichment for editing in hot tumours compared to cold tumours

264 (OR = 1.28, p = 0.001, Negative Binomial GLM). Additionally, upon integration with T-cell receptor

265 (TCR) repertoire data from Li et al[40], we found more diversity (Number of TCR clones / Total number

266 of TCR reads) in the immune hot tumours (**Figure 3d,** p < 2.2e-16, Wilcoxon's Rank Sum Test),

267 suggesting that broader immune responses may underlie the greater depletion of neoantigens in this

268 group.

269

270   Given the evidence for divergent infiltration patterns and activity between the immune clusters across

271   cancer types, we then investigated the determinants of this response by identifying differentially

272   expressed genes after adjusting for tumour type. We identified 365 genes at FDR < 0.01, FC>2 and in

273   pathway analysis, the top pathways were significantly associated with T-helper 1 (Th1) vs T-helper 2

274   Th2) lymphocyte responses (**Figure 3e, Supplementary Data 6**). Multiple Th1 cytokines and

275   downstream targets were overexpressed in hot tumours (*IFNG, CCL4, CCL5, CXCL9, CXCL10*), along

276   with costimulatory and coinhibitory receptors, suggesting these tumours were marked by a state of

277   lymphocyte activation and counter-responses thereto. We next scored proinflammatory (Th1, Th17)

278   and suppressive (Th2)) CD4+ cell populations using RNA-seq reference profiles from purified cells to

279   derive relative estimates using CIBERSORT[14]. Consistent with our inferences from pathway analysis,

280   we found enrichment for Th1 cells in hot, and Th2 and Th17 cells in cold tumours (**Figure 3f**). Th2 cells

281   have been linked to poor prognosis in multiple studies, while Th1 cells are associated with good

282   prognosis and aiding CTL responses[1].

283

284   We also used expression-based CIBERSORT to derive estimates for different myeloid cell populations

285   (n = 2346 tumours at deconvolution $p < 0.05$, Permutation Test, 1000 replicates), and identified

286   substantially higher fractions of M1 relative to M2 macrophages in hot tumours (p = 2.2e-16,

287   Wilcoxon's Rank Sum Test, **Figure 3g**). Notably, M2-like polarisation is associated not only with Th2

288   immune responses but also with immune-suppressive myeloid derived suppressor cells (MDSCs)[41].

289   Taken together, our analyses implicate Th1 cytokine signalling programmes as responsible for

290   establishing an immune-hot state and suggest MDSC and Th2/Th17 programmes as targets for efforts

291   to switch cold tumours to an immune hot state.

292

293   ***Relation of immune cluster to ICB response***

294   We reasoned that if the signature for immune hot tumours represented active immunity, it could be

295   applicable to the prediction of immunotherapy responses and evaluated this hypothesis using tumour

296   gene expression data from three melanoma cohorts: post-sequential aCTLA4 and aPD1 treatment[42,43];

297   pre-aCTLA4 treatment[7] and post-aPD1 (Nivolumab) treatment[44]. Analysis of these transcriptional

298   patterns indicated differential expression between responders and non-responders (**Figure 4a**), and

299   accordingly, ssGSEA scores for the hot transcriptional signature showed significant enrichment in

300   responders for the latter two datasets (**Figure 4b, c**). Moreover, a similar association emerged from

301   comparing the probability of response to hot/cold class prediction, inferred using a logistic regression

302   fit on TCGA hot/cold transcriptional signature ssGSEA scores (**Figure 4d**). Finally, we evaluated the

303   ability of the hot-signature to stratify patients by response relative to mutational load and Class I

304   neoepitope burden using elastic nets coupled to cross-validation for each dataset (**Figure 4e**). While

305   larger patient cohorts will be required to search for and validate predictive ICB biomarkers, from this

306   limited analysis it is clear that neither cellular composition as described by hot/cold classification, nor

307   total mutation or predicted class 1 neoantigen loads (which are also associated with response and

308   have been proposed as ICB response biomarkers) are reliable predictors.

309

310   Our analysis clearly indicates there is a heterogeneity of ICB response within both hot and cold

311   tumours similar to that described for tumours with high or low mutational loads, underlining the

312   importance of additional factors not captured by these metrics, such as immune cell phenotype and

313   the spatial distribution of immune cells within the tumours[1]. We reasoned that in addition to these

314   factors, the heterogeneity in responses to ICB among hot tumours might be driven by intrinsic

315   resistance to T-cell mediated destruction due to pre-existing genomic alterations within the tumour

316   cells. We set out to test this hypothesis by constructing a pan-cancer catalogue of genomic alterations

317   enriched in hot tumours, with the additional aim of finding those alterations enriched in cold tumours

318   which might drive lymphocyte exclusion or reduce tumour immunogenicity.

319

Genomic features of hot and cold tumours

321 Consistent with our observations in HNSCC, and with estimates from gene expression-based

322 deconvolution[13], immune hot tumours harboured higher overall mutation loads (OR = 1.33, p = 3e-25,

323 negative binomial GLM controlling for cancer type) and more predicted neoantigens than immune

324 cold tumours (**Figure 5a**). Given the recent finding that in addition to the presence of neoantigens,

325 their clonality (i.e. presence in all tumour cells as opposed to minor subclones) is associated with

326 prognosis and response to Pembrolizumab in lung adenocarcinoma[5], we analysed immune

327 microenvironment composition as a function of neoantigen clonality (as denoted by The Cancer

328 Immunome Atlas[45]). We found that the abundance of both CTLs and Tregs is correlated with clonal

329 neoantigen load pan-cancer (**Figure 5b**), while the relationship is much weaker when subclonal

330 neoantigens are considered. CD4+/FOXP3- effector lymphocytes display a striking inverse correlation

331 with clonal neoantigens (**Figure 5b**). Consistent with our earlier observation that they are enriched in

332 CTL / Treg low cold tumours, CAFs are inversely correlated with both clonal and subclonal neoantigen

333 loads. Hot tumours display a significantly higher clonal neoantigen burden (OR = 1.28, p < 4e-7,

334 Negative Binomial GLM) as well as a skew in the neoantigen burden towards clonal neoantigens after

335 adjusting for tumour type (OR = 1.01, p = 0.006 Negative Binomial GLM). These findings provide

336 evidence for a direct link between Class I MHC clonal neoantigen burden and patterns of TIL

337 abundance and may help to explain the observations of McGranahan and colleagues, that high clonal

338 neoantigen burden predicts favourable response to immune checkpoint modulation using

339 Pembrolizumab[5].

340

341 Next, we examined if the genomic features associated with immune cluster were also reproducible

342 across cancer types, performing adjusted binomial regressions to estimate cluster association after

343 controlling for tumour type for genes previously implicated as pan-cancer drivers based on signatures

344   of positive selection[46] and recorded 56 hits at FDR < 0.05 (**Supplementary Data 7, Figure 5c**).

345   Interestingly, these putative drivers of hot tumours were significantly enriched (OR = 11.75, p < 0.0004,

346   Fisher's Exact Test) in a list of genes demonstrated to confer resistance to CD8 T-cell mediated killing

347   in a recent CRISPR-Cas9 screen[47]. The functionally-verified immune-resistance genes that were

348   disproportionally mutated in hot tumours included those involved in antigen presentation (*B2M* and

349   *HLA-A*), apoptosis (*CASP8*) and interferon signalling (*JAK1*). *ARID2*, which encodes a component of the

350   SWI/SNF chromatin remodelling complex also implicated in modulating sensitivity to T-cell mediated

351   killing downstream of JAK-STAT signalling, was also disproportionately mutated in hot tumours[48,49].

352   Mutant *KRAS*, recently implicated in generating a fibrotic tumour microenvironment by activating

353   paracrine signalling with CAFs in pancreatic cancer[50], is enriched in cold tumours. Taken together,

354   these findings support a model wherein mutations in certain genes render tumours hot as a

355   consequence (and therefore susceptible to checkpoint blockade), or may enable tumours to survive

356   in a hot tumour microenvironment, potentially also bestowing resistance to checkpoint blockade. We

357   sought to test this model by linking our candidate immune-resistance mutations to lack of ICB

358   response in pre-treatment hot tumours and although we observed a trend, (OR = 0.13, logistic

359   regression, p = 0.09), the number of treated tumours with sequence data available is currently too

360   low (54 hot tumours across four studies) to gain a definitive answer.

361

362   To complement these analyses, we called copy number alterations across 11,000 tumours and tested

363   for differential association of peaks with immune cluster after adjusting for tumour type for the subset

364   with immune cluster assignment available. This led to the identification of multiple events that

365   occurred at different frequencies between cold and hot tumours (FDR < 0.1, **Figure 5d**). Of these,

366   prominent examples included amplifications targeting the epidermal growth factor receptor (*EGFR*)

367   (7p11.2) and *MYC* (8q24.3) and deletions at 10q23.31, encompassing the *PTEN* tumour suppressor

368   gene in cold tumours and *JAK2* (9p24.1) amplifications in hot tumours. Some of these candidates

369   already have known associations with immune evasion; *MYC* has been linked to an immune evasion

370   phenotype that is amenable to targeting through gene-body demethylation[51] and *PTEN* deletion has

371   recently been linked to failure of immunotherapy and decreased cytotoxic T-lymphocyte infiltration

372   in patients and in a mouse model of melanoma[43,52,53]. Among the genomic alterations we identify (for

373   full list of predicted driver events see **Supplementary Data 8**), it is likely that some establish, while

374   others are selected for, in different immune microenvironments. In either case, alongside *PTEN*

375   deletion, these alterations warrant further investigation as candidate genomic markers for response

376   to ICB. The enrichment of *EGFR* and *MYC* amplification, together with *PTEN* deletion in cold tumours

377   pan-cancer was striking given the co-expression module linked to increased tumour cell glycolysis and

378   immune evasion in HNSCC, which includes *EGFR* and in which pathway analysis also predicts increased

379   c-MYC and mTORC1 activity[17]. A similar relationship has been observed in triple negative breast

380   cancer[54] and we therefore investigated this relationship further, initially interrogating the link

381   between *EGFR* protein levels and TILs in two HNSCC cohorts (n=518)[19,28] we found that samples

382   classified as EGFR high and moderate were significantly more likely to be TIL low than EGFR low

383   cancers after accounting for anatomic site and HPV status (**Figure 5e**, $p < 0.05$ and 0.01 for EGFR

384   moderate and high cancers, logistic regression). The positive correlation between EGFR levels (which

385   are themselves correlated with EGFR phosphorylation (activation), **Supplementary Figure 3**) and the

386   glycolytic signature is maintained across TCGA when matched RPPA profiles and RNA-seq data are

387   compared (**Figure 5f**). Notably, the glycolytic signature is enriched in progressing melanomas after PD-

388   1 blockade (**Figure 5g**, $p = 0.06$, t-test, $p = 0.02$ when excluding stable disease) and is inversely

389   associated with expression of the hot transcriptional signature (Rho = -0.44) that is associated with

390   ICB response (**Figure 4**, **Figure 5h**).

391

392 **Discussion**

393 Of the methods developed to deconvolve cell mixtures into multiple cell types from methylation data,

394 there have not been, to the best of our knowledge, exhaustive studies employed across cancer types.

395 Methods such as LUMP and the Leukocyte Methylation Score estimate only the overall leukocyte

396 fraction, while methods based on expression data either produce relative estimates of abundance

397 within the immune fraction or enrichment scores (CIBERSORT, TIMER) or perform low resolution

398 deconvolution (ESTIMATE)[13,14,22,55,56]. Combining methylation-based feature selection from both

399 stromal and cancer cells with the robust performance of CIBERSORT previously displayed on gene

400 expression microarray data[14] allowed us to derive estimates for different infiltrating cell populations

401 as a fraction of the overall sample. We see MethylCIBERSORT and CIBERSORT as complementary tools

402 for studying the tumour microenvironment in cases where both DNA methylation and gene expression

403 data can be obtained, in addition to serving as alternatives in cases where only one data type is

404 available.

405

406 While approaches using RNA-sequencing or other transcriptional profiling, such as the construction of

407 an index of cytolytic activity, have been useful in predicting immunotherapy response[7] and in

408 identifying the role of mutations in genes like *CASP8* in immune evasion[29], the deeper level of

409 deconvolution made feasible using DNA methylation data allows the roles of distinct cellular subsets

410 and their interdependencies to be dissected. Here, by applying the method to HNSCC, which is marked

411 by a great degree of clinical heterogeneity, we identified lymphocyte-rich and stromal-rich prognostic

412 subgroups consistent with those discovered previously using a variety of independent methods

413 [19,20,24,28,57-59] and derived insights into the microenvironmental alterations that might be relevant for

414 prognosis. In the process, we showed that our scheme for classifying cancers correlates with well-

415 established immune metrics such as cytolytic activity, neoantigen/mutational load and CTL:Treg

416 ratios. We demonstrated that tumours similar to the HNSCC immune-hot subgroup exist in varying

417   fractions across the vast majority of cancer types. The congruence of our classification with the

418   aforementioned metrics is maintained throughout and translates to broader TCR responses which

419   presumably drive the greater depletion of neoantigens seen in the hot tumours. Combining the

420   accuracy and ability to determine absolute cell fractions of MethylCIBERSORT with the high resolution

421   afforded by CIBERSORT enabled detection of a skew towards antitumour T-helper and macrophage

422   phenotypes in hot tumours, reinforcing our hypothesis that this cluster is enriched for active

423   antitumour immune responses. Consistent with this, we also showed in the limited post-treatment

424   cohorts with gene expression data, that melanomas responding to ICB (particularly anti-CTLA4) display

425   a gene expression signature derived from our hot cluster. Our ICB response analysis also indicates

426   that hot/cold classification based on cellular composition is a superior predictor of response than total

427   mutation or predicted class I neoantigen loads (both previously proposed ICB response biomarkers),

428   however none of these metrics displayed sufficient accuracy to be of clinical use. Given the complex

429   interplay between tumour genomics, epigenomics and anti-tumour immune responses, it is likely that

430   prediction algorithms incorporating multiple biomarkers such as these, in addition to checkpoint

431   protein expression and spatial information on immune infiltrates (e.g. Immunoscore[60]) will be required

432   for patient stratification.

433

434   Genomic analysis identified significant enrichment for events that confer resistance to T-cell mediated

435   destruction in hot tumours as well as potential sensitisers. Our copy number analysis revealed that

436   *PTEN* deletion, *MYC* amplification and *EGFR* amplification are associated with immune depletion. All

437   three events have been associated with increased glycolysis, which we have previously linked to

438   immune evasion[17]. Our finding that *PTEN* deletion is associated with poor CTL infiltration in this pan-

439   cancer cohort adds substantial support and mechanistic rationale for its proposed role as a

440   determinant of response to ICB. Taken together with the identification of *EGFR* and *MYC* amplification

441   in cold tumours, our analysis suggests that pharmacological inhibition of EGFR/mTORC1/MYC-driven

442   glycolysis could be an effective means by which to 'warm-up' these tumours and potentially enhance

443   responses to ICB. The finding that hot tumours frequently harbour functionally-validated immune

444   resistance mutations offers a potential explanation for the heterogeneity in ICB response even

445   amongst hot tumours (or equally, those with high mutation loads or high cytolytic activity[7]). Secondly,

446   the relative paucity of these mutations in cold tumours (presumably due to the absence of a selection

447   pressure for them) suggests that if we could induce lymphocyte infiltration (e.g. by targeting glycolysis

448   or CAFs[61]), we may improve the effectiveness of checkpoint blockade across a broader range of

449   patients.

450

451   Finally, our analysis of neoantigen clonality and immune infiltration patterns adds mechanistic insight

452   to the value of clonal neoantigen burden in predicting response to ICB[5]. In particular, we show that

453   clonal neoantigens are associated with infiltration of CTLs and Tregs, while Th2 cells and CAFs are

454   enriched in tumours with lower clonal neoantigen loads. Why these relationships between neoantigen

455   loads and T-lymphocytes are apparent only when one considers clonal neoantigens is an intriguing

456   question. It could be that since many subclonal neoantigens are expressed by a small minority of cells

457   within the tumour, these evade effective presentation to the immune system. Indeed, in a previous

458   study by several of the authors, it was possible to isolate T-lymphocytes reactive against clonal but

459   not subclonal neoantigens from lung cancer patients[5]. Our data suggest that this is due to a relative

460   paucity of CTLs in tumours with low clonal neoantigen loads and that this is true across a wide range

461   of cancer types.

462

463   In summary, the development of a stand-alone method to estimate both tumour purity and stromal

464   composition from DNA methylation data has provided a number of insights that shed light on

465   potential biomarkers for immunotherapy response and the way in which tumour genomes influence,

466   and are shaped by, the immune microenvironment. Beyond analyses of publicly available data such as

467 those we present here, the applicability of the method to both fresh and archival samples should

468 readily allow researchers to explore questions related to the tumour microenvironment and potential

469 therapeutic response across a diverse range of experimental settings.

470

471 **Methods**

472

473 ***Development of methylation signatures for deconvolution***

474 Raw data were obtained in the form of IDAT files from the following sources (the number of samples

475 from which each profile was derived is shown in parentheses): Granulocytes (12), CD8+ (cytotoxic T-

476 lymphocytes) (6), CD19+ (B-lymphocytes) (6), CD56+ (Natural Killer cells) (6), CD14+ (monocyte

477 lineage) (6), Eosinophils (6) were from the *FlowSorted.Blood.450k* Bioconductor package[62]. CD4+ cells

478 were removed from the Blood.450k dataset and CD4+ T-cells from the Zhang dataset[63] (data kindly

479 provided by Dr Alicia Oshlack) were further divided into FOXP3+ (Tregs) (4) or FOXP3- (6) groups.

480 Fibroblast profiles (4) were from the Gene Expression Omnibus (GSE74877

481 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74877]).   Neutrophils are the most

482 abundant subset of granulocytes and these samples were therefore aggregated into a single category

483 for further analysis. To generate a DNA methylation signature for cancer cells, we used 450k

484 methylation profiles we previously obtained from a series of 6 HNSCC cell lines: UM-SCC47; 93VU147T;

485 UPCI:SCC090;   PCI-30;   UPCI:SCC036   and   UPCI:SCC003   (GSE38270

486 [[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38270], described in[64]) and additionally

487 those   from   Iorio   et   al   (GSE68379

488 [[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68379])[65]. The files were parsed into R

489 using the *minfi*[66] Bioconductor package and were normalised using single sample Noob as

490 implemented in *minfi*. To derive signature features, a custom limma based wrapper function was used

491 to fit a series of linear models for all pairwise comparisons between candidate cell types. Features

492 from this set of analyses were then restricted to MVPs that showed a median beta-value difference of

493 0.25 at an FDR of 0.01 for that fit or less, with a maximum of 100 MVPs per pairwise comparison.

494 Finally, for use with CIBERSORT, data were transformed from beta values (bound between 0 and 1) to

495 percentages (0 – 100). Type-wise means were estimated for each probe and cell type and the matrix

496 exported for upload to CIBERSORT.

497

498 **Benchmarking using PBMC mixtures**

499 We applied the feature selection pipeline to the matrix of stromal cells that we assembled and then

500 tested performance against 450k profiles of PBMC mixtures with flow-cytometry gold standards. We

501 also applied *LM22* (Expression-based CIBERSORT) to datasets consisting of PBMC samples and

502 Follicular Lymphoma biopsies originally evaluated in CIBERSORT[14]. Wilcoxon's rank sum tests were

503 used to test for differences in correlations with flow-cytometry for cell types and samples, and

504 absolute errors between flow-cytometry and deconvolution estimate. For the expression-based

505 CIBERSORT estimates, we performed comparisons against both calibrated (i.e, enforcing a sum-to-one

506 constraint as reflected in the flow cytometry) and uncalibrated (straight estimates of cell fractions

507 from CIBERSORT) estimates.

508

509 **Running deconvolution experiments on HNSCC using CIBERSORT**

510 Data for 464 methylation profiled TCGA HNSCC samples were downloaded in the form of raw IDAT

511 files for the 450k array from the TCGA data. Data were normalised using functional normalisation[67] in

512 the minfi[66] package and BMIQ[68], with 10,000 reference probes for Expectation Maximisation fitting.

513 HPV status was determined using VirusSeq[69] based on detection of viral gene transcripts.

514

515   Beta values for deconvolution associated features, and the signature matrix derived in the previous

516   step, were uploaded to CIBERSORT at https://cibersort.stanford.edu. The data were not quantile

517   normalised due to the potential for global methylation shifts in cancers, and CIBERSORT was run using

518   1000 permutations. Output files were downloaded as tab-delimited text files and custom parsers were

519   used to import results into R for downstream analysis. FFPE methylation profiles for 42 HNSCC were

520   obtained   from   Gene   Expression   Omnibus   (Accession   GSE38266

521   [[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38266]) using the GEOquery R package,

522   and beta values were BMIQ normalised and analysed using CIBERSORT as described for the TCGA

523   cohort. Wilcoxon's Rank Sum Tests were used to test for differences in total TIL abundance and TIL

524   subsets.

525

526   ***Estimating accuracy in tumour deconvolution***

527   In the absence of flow-cytometry based estimates for the different cell types in the analyzed tumours,

528   the estimated fraction of cancer cells from MethylCIBERSORT was compared to sequencing-data

529   based estimates from ABSOLUTE available for 466 HNSCCs from previously published work[22] using

530   Spearman's Rank Correlation. Correlations were between ABSOLUTE and other methods of estimating

531   purity/immune cell fraction in this subset of tumours; (LUMP, ESTIMATE[56] and H&E staining

532   assessment of tumour purity (data available in[22]). Residuals were computed by subtracting the

533   method estimate from the ABSOLUTE value. Distributions were compared using Wilcoxon's Rank Sum

534   Test. Spearman's Rank Correlation was used to estimate correlations between expression of marker

535   transcripts and MethylCIBERSORT estimates for multiple cellular populations. Where applicable,

536   multiple testing correction was performed using the Benjamini Hochberg approach.

537

538   ***Clustering and correlation analyses***

539   Estimates of immune cell fractions in HPV- HNSCC (HPV-transcript negative) were examined for

540   correlations with other infiltrating cell types using Spearman's Rank Correlation with BH correction for

541   multiple testing. Clustering was carried out using the clusterCons package with 100 iterations using a

542   manhattan distance metric. The most robust number of clusters was then selected.

543   Differences in the distribution of infiltrating cell types by immune cluster were summarised using

544   mean fold changes and tested using Wilcoxon's Rank Sum Test with BH-correction for multiple testing.

545   Differentially expressed genes were identified using Limma-trend and were defined at a threshold of

546   a 2-fold change and BHFDR < 0.01. Pathway analysis was carried out using Ingenuity Pathway Analysis,

547   with findings restricted to experimentally confirmed direct interactions in human cells/tissues.

548   Cytolytic activity (CYT) was calculated as the geometric mean of *GZMA* and *PRF1* expression as defined

549   previously[29]. To estimate the contributions of cell population abundances to this, a linear model was

550   fit against log2(CYT) with the different populations as predictors. Wilcoxon's Rank Sum tests were used

551   to test differences in CYT and CD8:Treg ratios between the immune clusters.

552

553   ***Survival Analyses***

554   Multivariate Cox Regression was used to estimate the prognostic utility of clusters derived using

555   infiltration patterns with age and stage as covariates. The survival effect of estimated purity was

556   regressed with the same covariates using a Cox regression with coefficients defined per percent

557   increase in purity.

558

559   ***Genomic Correlates***

560   We obtained a list of driver genes inferred by MutSigCV[33] in TCGA HNSCC cohort from the Broad

561   Institute's GDAC. GISTIC Copy number estimates thresholded by genes were also obtained from this

562   source. MAF files were obtained from the TCGA data portal. MutSigCV drivers were filtered at a

563   q.value threshold of 0.01 and mutations in this set were tested for differences in frequencies of

564   occurrence using a chi-squared test for differences in proportion. Multiple testing correction was

565   carried out using the Benjamini-Hochberg method. Tables of predicted neoantigens were downloaded

566   from The Cancer Immunome Atlas (http://tcia.at).

567

568   ***Benchmarking performance across other tumour types***

569   Signature features were derived from 450k profiles using the aforementioned heuristic (delta-Beta

570   and FDR cutoffs) with a maximum of 100 features per cell type for a wide range of tumour types, using

571   cell   lines   allocated   to   the   corresponding   tissue   in   GSE68379

572   [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68379][37]   (**Table   S10**)   and   the

573   aforementioned infiltrating cell types. These signatures were applied to deconvolve methylation

574   profiles and estimates of purity were derived using TCGA samples for which ABSOLUTE, ESTIMATE and

575   LUMP purity estimates were available[22].

576   The cell line data were functionally normalized with the infiltrating cell types described earlier before

577   signature extraction was carried out. 450k data for the aforementioned tumour types were loaded

578   from a pan-cancer freeze derived from SAGE synapse for TCGA pan-cancer (syn2812961

579   [https://www.synapse.org/#!Synapse:syn2812961]) and a custom function was used to extract

580   signature probes and generate methylation percentage matrices for deconvolution with CIBERSORT

581   CIBERSORT was run using the graphical user interface [https://cibersort.stanford.edu]. Correlation

582   and residuals analysis were carried out as described above with MethylCIBERSORT purity estimates vs

583   ABSOLUTE, and between previously published methods and ABSOLUTE. Wilcoxon's Rank Sum Test

584   with Benjamini-Hochberg correction for multiple testing were used to compare distributions, with

585   these estimates sourced from[22].

586

587 ***Pan-cancer analyses of immune cluster assignment***

588 An elastic net model was fit using cellular abundance estimates for HPV- HNSCC using 3 iterations of

589 5-fold cross-validation to identify the optimal values of lambda and alpha with Kappa values being the

590 selection criterion. The classifier was then applied to MethylCIBERSORT estimates from 18 further

591 tumour types for which corresponding cancer cell line methylation profiles were available[37] to allocate

592 immune cluster. Deconvolution was performed as described above and class allocations were made

593 using the elastic net classifier derived from HNSCC.

594 For immunoediting analyses, we estimated the number of nonsynonymous mutations encoding at

595 least one immunogenic peptide empirically by summing coefficients across each of six base change

596 contexts as well as the number of non-neoepitope nonsynonymous mutations. Together , these were

597 applied to silent mutation counts in each cancer to derive an expected fraction of neoantigens to

598 nonneoantigens. Comparing the observed fraction to the expected fraction yielded the percentage of

599 neoantigens depleted, and using this in combination with the number of observed neoantigens

600 yielded the count of neoantigen-encoding mutations lost specifically to immunoediting. This was then

601 modelled using a negative binomial framework to estimate the influence of immune cluster on

602 immunoediting.

603 MAF files for mutations were again downloaded from SAGE synapse for all tumours from the MC3

604 calling effort (syn7214402 [https://www.synapse.org/#!Synapse:syn7214402]). Driver mutations

605 were defined based on pan-cancer MutSig analyses previously published[33] and logistic regression

606 GLMs were used to estimate coefficients for mutation frequencies for immune cluster with tumour

607 type as a covariate. Significant genes were defined at BHFDR < 0.05. Survival analyses were performed

608 using data downloaded from Synapse (syn7343873

609 [https://www.synapse.org/#!Synapse:syn7343873]) using Cox proportional hazards regression with

610 stage and cancer type as covariates. Substages were aggregated into stages and only Stages I-IV were

611 considered. Neoantigen abundance and clonality data were downloaded from The Cancer Immunome

612 Atlas[45].

613 Negative binomial modelling was used to model all count data, cytolytic activity was modelled using

614 linear models, and binomial GLMs were used to model proportions. Details of covariates, hypotheses

615 and response variables are presented inline. For copy number analyses, SNP6 data were downloaded

616 from the GDC data portal and processed using GISTIC 2.0[70] on the GenePattern Public Server (arm-

617 level peel off, noise threshold 0.3, FDR < 0.01, driver-gene confidence > 95%) and modelled similarly

618 to mutation data.

619

620 ***Further resolution of cell types using expression-based CIBERSORT***

621 RNA-seq data were downloaded from the European Nucleotide Archive for the following datasets:

622 PRJEB11844 [https://www.ebi.ac.uk/ena/data/view/PRJEB11844][71]; PRJNA258216

623 [https://www.ebi.ac.uk/ena/data/view/PRJNA258216][72], and PRJEB5468

624 [https://www.ebi.ac.uk/ena/data/view/PRJEB5468][73]. Kallisto[74] was used to quantify gene expression

625 with a reference transcriptome consisting of Gencode Grch37 assembly of protein coding and lincRNA

626 transcripts. Data were then modelled using limma trend and the top 50 markers by t-statistics were

627 selected for each cell subset from one versus all comparisons after thresholding with a 2-fold change

628 and FDR < 0.05. These cell types were used to generate a reference profile and CIBERSORT was run

629 to deconvolute samples. For M1/M2 macrophage analyses we used LM22 from the CIBERSORT server

630 as the reference. In both cases, Wilcoxon's Rank Sum Test was used to estimate differences in

631 distributions.

632

633 ***Analysis of immunotherapy response***

634   Nanostring data for a panel of immune genes and and exome sequencing data were obtained from

635   Chen et al[42] and Roh et al[43] respectively for patients treated using sequential anti-CTLA4 and anti-PD1

636   checkpoint blockade.

637   Clustering and machine learning were carried out using the subset of genes intersecting with the Hot-

638   vs-Cold pancancer signature. .632 bootstrapping was used for hyperparameter tuning and ROC

639   estimation. Negative binomial regression was used to model neoantigen, mutation and subclone

640   numbers, and logistic regression to estimate predictive performance of count data on response.

641   The number of subclones present in each tumour from the Roh cohort, derived from the EXPANDS

642   algorithm, were obtained from the associated publication[75]. RNAseq data for aCTLA4 pretreatment

643   biopsies were kindly provided by Eliezer Van Allen and genomic data were from the associated

644   publication[7]. Data for post-treatment Nivolumab treated melanomas were obtained from[44].

645

Valiation of EGFR association with cold tumours

647   RPPA data were downloaded for TCGA cancers from the TCGA portal. IHC data were derived from[1,7]

648   for comparison of EGFR protein levels vs TIL levels, previously defined in[1,7]. ssGSEA scores were used

649   to summarise the activity of the glycolytic gene signature (described in[17]) and standard statistical

650   procedures were used to assess interrelationships.

651

Data Availability

653   The data analysed in this study are available either from the Gene Expression Omnibus (accession

654   numbers: GSE35069 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35069]; GSE74877

655   [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74877]; GSE38270

656   [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38270]; GSE68379

657   [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68379]), the European Nucleotide Archive

658   (accession   numbers: PRJEB11844   [https://www.ebi.ac.uk/ena/data/view/PRJEB11844];

659   PRJNA258216   [https://www.ebi.ac.uk/ena/data/view/PRJNA258216];   PRJEB5468

660   [https://www.ebi.ac.uk/ena/data/view/PRJEB5468]), SAGE Synapse (accession numbers: syn7214402

661   [https://www.synapse.org/#!Synapse:syn7214402];   syn7343873

662   [https://www.synapse.org/#!Synapse:syn7343873];   syn2812961

663   [https://www.synapse.org/#!Synapse:syn2812961]), The Cancer Genome Atlas Project or from the

664   authors upon reasonable request.

665

666   ***Code Availability.***

667   Knit R-markdowns of the code used for analysis have been deposited on Zenodo at
668   https://doi.org/10.5281/zenodo.1304766. The MethylCIBERSORT R-package and the signatures we
669   generated are on Zenodo at

670   https://doi.org/10.5281/zenodo.1298968. A Google Group for users of the package can be found

671   here: https://groups.google.com/d/forum/methyldeconvolution.

672

Acknowledgments

683   Information about TCGA and the investigators and institutions who constitute TCGA research network

684   can be found at http://cancergenome.nih.gov.

685

Author Contributions

687   A.C., A.Feber. and T.R.F. conceived the study, designed research and wrote the paper; A.C., A.Furness,

688   K.J., E.G., T.M., M.J.W., G.J.T. and S.A.Q. conducted and analysed experiments; K.F., E.V.K and M.L.

689   provided samples; T.R.F. supervised the research.

690

Conflict of Interest Statement

692   The authors declare no conflict of interest.

693

References

695   1.   Fridman, W.H., Zitvogel, L., Sautès–Fridman, C. & Kroemer, G. The immune contexture in
696   cancer prognosis and treatment. *Nature Reviews Clinical Oncology* **14**, 717 (2017).
697   2.   Furness, A.J., Vargas, F.A., Peggs, K.S. & Quezada, S.A. Impact of tumour microenvironment
698   and Fc receptors on the activity of immunomodulatory antibodies. *Trends Immunol* **35**, 290-
699   298 (2014).
700   3.   Patel, S.P. & Kurzrock, R. PD-L1 Expression as a Predictive Biomarker in Cancer
701   Immunotherapy. *Mol Cancer Ther* **14**, 847-856 (2015).
702   4.   Le, D.T.*, et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* **372**,
703   2509-2520 (2015).
704   5.   McGranahan, N.*, et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to
705   immune checkpoint blockade. *Science* **351**, 1463-1469 (2016).
706   6.   Rizvi, N.A.*, et al.* Mutational landscape determines sensitivity to PD-1 blockade in non–small
707   cell lung cancer. *Science* **348**, 124 (2015).
708   7.   Van Allen, E.M.*, et al.* Genomic correlates of response to CTLA-4 blockade in metastatic
709   melanoma. *Science (New York, N.Y.)* **350**, 207-211 (2015).
710   8.   Houseman, E.A.*, et al.* Reference-free deconvolution of DNA methylation data and
711   mediation by cell composition effects. *BMC Bioinformatics* **17**, 1-15 (2016).
712   9.   Houseman, E.A., Molitor, J. & Marsit, C.J. Reference-free cell mixture adjustments in analysis
713   of DNA methylation data. *Bioinformatics* **30**, 1431-1439 (2014).
714   10.   Jaffe, A.E. & Irizarry, R.A. Accounting for cellular heterogeneity is critical in epigenome-wide
715   association studies. *Genome Biol* **15**, R31 (2014).
716   11.   Koestler, D.C.*, et al.* Blood-based profiles of DNA methylation predict the underlying
717   distribution of cell types: a validation analysis. *Epigenetics* **8**, 816-826 (2013).

718   12.   Li, B. & Li, J.Z. A general framework for analyzing tumor subclonality using SNP array and
719   DNA sequencing data. *Genome Biol* **15**, 473 (2014).
720   13.   Li, B.*, et al.* Comprehensive analyses of tumor immunity: implications for cancer
721   immunotherapy. *Genome Biol* **17**, 174 (2016).
722   14.   Newman, A.M.*, et al.* Robust enumeration of cell subsets from tissue expression profiles.
723   *Nature methods* **12**, 453-457 (2015).
724   15.   Teschendorff, A.E., Breeze, C.E., Zheng, S.C. & Beck, S. A comparison of reference-based
725   algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies.
726   *BMC Bioinformatics* **18**, 105 (2017).
727   16.   Gentles, A.J.*, et al.* The prognostic landscape of genes and infiltrating immune cells across
728   human cancers. *Nat Med* **21**, 938-945 (2015).
729   17.   Ottensmeier, C.H.*, et al.* Upregulated Glucose Metabolism Correlates Inversely with CD8+ T-
730   cell Infiltration and Survival in Squamous Cell Carcinoma. *Cancer Res* **76**, 4136-4148 (2016).
731   18.   Chakravarthy, A.*, et al.* Human Papillomavirus Drives Tumor Development Throughout the
732   Head and Neck: Improved Prognosis Is Associated With an Immune Response Largely
733   Restricted to the Oropharynx. *J Clin Oncol* **34**, 4132-4141 (2016).
734   19.   Ward, M.J.*, et al.* Tumour-infiltrating lymphocytes predict for outcome in HPV-positive
735   oropharyngeal cancer. *Br J Cancer* **110**, 489-500 (2014).
736   20.   The Cancer Genome Atlas Research, N. Comprehensive genomic characterization of head
737   and neck squamous cell carcinomas. *Nature* **517**, 576-582 (2015).
738   21.   Carter, S.L.*, et al.* Absolute quantification of somatic DNA alterations in human cancer.
739   *Nature biotechnology* **30**, 413-421 (2012).
740   22.   Aran, D., Sirota, M. & Butte, A.J. Systematic pan-cancer analysis of tumour purity. *Nat
741   Commun* **6**, 8971 (2015).
742   23.   Yoshihara, K.*, et al.* Inferring tumour purity and stromal and immune cell admixture from
743   expression data. *Nature Communications* **4**, 2612 (2013).
744   24.   Wood, O.*, et al.* Gene expression analysis of TIL rich HPV-driven head and neck tumors
745   reveals a distinct B-cell signature when compared to HPV independent tumors. *Oncotarget*
746   (2016).
747   25.   Fischer, C.A.*, et al.* Is the improved prognosis of p16 positive oropharyngeal squamous cell
748   carcinoma dependent of the treatment modality? *International Journal of Cancer* **126**, 1256-
749   1262 (2010).
750   26.   Hong, A.M.*, et al.* Human papillomavirus predicts outcome in oropharyngeal cancer in
751   patients treated primarily with surgery or radiation therapy. *Br J Cancer* **103**, 1510-1517
752   (2010).
753   27.   Ward, M.J.*, et al.* Staging and treatment of oropharyngeal cancer in the human
754   papillomavirus era. *Head & Neck* **37**, 1002-1013 (2015).
755   28.   Marsh, D.*, et al.* Stromal features are predictive of disease mortality in oral cancer patients. *J
756   Pathol* **223**, 470-481 (2011).
757   29.   Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G. & Hacohen, N. Molecular and genetic
758   properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48-61 (2015).
759   30.   deLeeuw, R.J., Kost, S.E., Kakal, J.A. & Nelson, B.H. The Prognostic Value of FoxP3+ Tumor-
760   Infiltrating Lymphocytes in Cancer: A Critical Review of the Literature. *Clinical Cancer
761   Research* **18**, 3022 (2012).
762   31.   Sato, E.*, et al.* Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high
763   CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer.
764   *Proceedings of the National Academy of Sciences of the United States of America* **102**,
765   18538-18543 (2005).
766   32.   Suzuki, H.*, et al.* Intratumoral CD8+ T/FOXP3+ cell ratio is a predictive marker for survival in
767   patients with colorectal cancer. *Cancer Immunology, Immunotherapy* **59**, 653-661 (2010).

768   33.   Lawrence, M.S.*, et al.* Discovery and saturation analysis of cancer genes across 21 tumour
769   types. *Nature* **505**, 495-501 (2014).
770   34.   Li, C., Egloff, A.M., Sen, M., Grandis, J.R. & Johnson, D.E. Caspase-8 mutations in head and
771   neck cancer confer resistance to death receptor-mediated apoptosis and enhance migration,
772   invasion, and tumor growth. *Mol Oncol* **8**, 1220-1230 (2014).
773   35.   Brown, S.D.*, et al.* Neo-antigens predicted by tumor genome meta-analysis correlate with
774   increased patient survival. *Genome Research* **24**, 743-750 (2014).
775   36.   Matsushita, H.*, et al.* Cancer exome analysis reveals a T-cell-dependent mechanism of cancer
776   immunoediting. *Nature* **482**, 400-404 (2012).
777   37.   Iorio, F.*, et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754
778   (2016).
779   38.   Thirlwell, C.*, et al.* Genome-wide DNA methylation analysis of archival formalin-fixed
780   paraffin-embedded tissue using the Illumina Infinium HumanMethylation27 BeadChip.
781   *Methods (San Diego, Calif.)* **52**, 248-254 (2010).
782   39.   Siegel, E.M.*, et al.* Expanding epigenomics to archived FFPE tissues: an evaluation of DNA
783   repair methodologies. *Cancer epidemiology, biomarkers & prevention : a publication of the
784   American Association for Cancer Research, cosponsored by the American Society of
785   Preventive Oncology* **23**, 2622-2631 (2014).
786   40.   Li, B.*, et al.* Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat Genet* **48**,
787   725-732 (2016).
788   41.   Elliott, L.A., Doherty, G.A., Sheahan, K. & Ryan, E.J. Human Tumor-Infiltrating Myeloid Cells:
789   Phenotypic and Functional Diversity. *Frontiers in Immunology* **8**, 86 (2017).
790   42.   Chen, P.-L.*, et al.* Analysis of immune signatures in longitudinal tumor samples yields insight
791   into biomarkers of response and mechanisms of resistance to immune checkpoint blockade.
792   *Cancer Discovery* (2016).
793   43.   Roh, W.*, et al.* Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-
794   1 blockade reveals markers of response and resistance. *Science Translational Medicine*
795   **9**(2017).
796   44.   Riaz, N.*, et al.* Tumor and Microenvironment Evolution during Immunotherapy with
797   Nivolumab. *Cell* **171**, 934-949.e915 (2017).
798   45.   Charoentong, P.*, et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-
799   Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell
800   Rep* **18**, 248-262 (2017).
801   46.   Martincorena, I.*, et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*
802   **171**, 1029-1041.e1021 (2017).
803   47.   Patel, S.J.*, et al.* Identification of essential genes for cancer immunotherapy. *Nature* **548**,
804   537-542 (2017).
805   48.   Miao, D.*, et al.* Genomic correlates of response to immune checkpoint therapies in clear cell
806   renal cell carcinoma. *Science (New York, N.Y.)* **359**, 801 (2018).
807   49.   Pan, D.*, et al.* A major chromatin regulator determines resistance of tumor cells to T cell–
808   mediated killing. *Science (New York, N.Y.)* **359**, 770 (2018).
809   50.   Tape, C.J.*, et al.* Oncogenic KRAS Regulates Tumor Cell Signaling via Stromal Reciprocation.
810   *Cell* **165**, 910-920 (2016).
811   51.   Topper, M.J.*, et al.* Epigenetic Therapy Ties MYC Depletion to Reversing Immune Evasion and
812   Treating Lung Cancer. *Cell* **171**, 1284-1300.e1221 (2017).
813   52.   George, S.*, et al.* Loss of PTEN Is Associated with Resistance to Anti-PD-1 Checkpoint
814   Blockade Therapy in Metastatic Uterine Leiomyosarcoma. *Immunity* **46**, 197-204 (2017).
815   53.   Peng, W.*, et al.* Loss of PTEN Promotes Resistance to T Cell–Mediated Immunotherapy.
816   *Cancer Discovery* **6**, 202 (2016).
817   54.   Lim, S.O.*, et al.* EGFR Signaling Enhances Aerobic Glycolysis in Triple-Negative Breast Cancer
818   Cells to Promote Tumor Growth and Immune Escape. *Cancer Res* **76**, 1284-1296 (2016).

819 55. Carter, S.L.*, et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat*
820   *Biotech* **30**, 413-421 (2012).
821 56. Yoshihara, K.*, et al.* Inferring tumour purity and stromal and immune cell admixture from
822   expression data. *Nat Commun* **4**, 2612 (2013).
823 57. Balermpas, P.*, et al.* Tumour-infiltrating lymphocytes predict response to definitive
824   chemoradiotherapy in head and neck cancer. *Br J Cancer* **110**, 501-509 (2014).
825 58. Keck, M.K.*, et al.* Integrative Analysis of Head and Neck Cancer Identifies Two Biologically
826   Distinct HPV and Three Non-HPV Subtypes. *Clinical Cancer Research* **21**, 870-881 (2015).
827 59. Nguyen, N.*, et al.* Tumor infiltrating lymphocytes and survival in patients with head and neck
828   squamous cell carcinoma. *Head & Neck* **38**, 1074-1084 (2016).
829 60. Angell, H. & Galon, J. From the immune contexture to the Immunoscore: the role of
830   prognostic and predictive immune markers in cancer. *Current Opinion in Immunology* **25**,
831   261-267 (2013).
832 61. Hanley, C.J.*, et al.* Targeting the Myofibroblastic Cancer-Associated Fibroblast Phenotype
833   Through Inhibition of NOX4. *Journal of the National Cancer Institute* **110**(2018).
834 62. Reinius, L.E.*, et al.* Differential DNA methylation in purified human blood cells: implications
835   for cell lineage and studies on disease susceptibility. *PLoS One* **7**, e41361 (2012).
836 63. Zhang, Y.*, et al.* Genome-wide DNA methylation analysis identifies hypomethylated genes
837   regulated by FOXP3 in human regulatory T cells. *Blood* **122**, 2823-2836 (2013).
838 64. Lechner, M.*, et al.* Identification and functional validation of HPV-mediated
839   hypermethylation in head and neck squamous cell carcinoma. *Genome medicine* **5**, 15
840   (2013).
841 65. Iorio, F.*, et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754
842   (2016).
843 66. Aryee, M.J.*, et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis
844   of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369 (2014).
845 67. Fortin, J.P.*, et al.* Functional normalization of 450k methylation array data improves
846   replication in large cancer studies. *Genome biology* **15**, 503 (2014).
847 68. Teschendorff, A.E.*, et al.* A beta-mixture quantile normalization method for correcting probe
848   design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189-196
849   (2013).
850 69. Chen, Y.*, et al.* VirusSeq: software to identify viruses and their integration sites using next-
851   generation sequencing of human cancer tissue. *Bioinformatics* **29**, 266-267 (2013).
852 70. Mermel, C.H.*, et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of
853   focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
854 71. De Simone, M.*, et al.* Transcriptional Landscape of Human Tissue Lymphocytes Unveils
855   Uniqueness of Tumor-Infiltrating T Regulatory Cells. *Immunity* **45**, 1135-1147.
856 72. Linsley, P.S., Speake, C., Whalen, E. & Chaussabel, D. Copy number loss of the interferon
857   gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis.
858   *PLoS One* **9**, e109760 (2014).
859 73. Bonnal, R.J.P.*, et al.* De novo transcriptome profiling of highly purified human lymphocytes
860   primary cells. *Scientific Data* **2**, 150051 (2015).
861 74. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
862   quantification. *Nature biotechnology* **34**, 525-527 (2016).
863 75. Andor, N., Harness, J.V., Müller, S., Mewes, H.W. & Petritsch, C. EXPANDS: expanding ploidy
864   and allele frequency on nested subpopulations. *Bioinformatics* **30**, 50-60 (2014).

865

866

867 **Figure Legends**

868

869 **Figure 1. Validation of DNA methylation-based deconvolution for the analysis of tumour**

870 **composition (a)** Correlation between MethylCIBERSORT fractions and flow cytometry for PBMC

871 mixtures in independent data. **(b-d)** Boxplots showing comparisons between MethylCIBERSORT and

872 flow cytometry versus Expression-CIBERSORT and flow cytometry in mixtures of similar complexity for

873 correlations by cell type, correlations within samples, and finally absolute error. **(e)** Spearman's

874 Correlations between ABSOLUTE and MethylCIBERSORT versus other previously published purity

875 estimation methods. **(f)** Validation of previously reported associations between CD8 T-cells and B-cells

876 and HPV status by HPV status. **(g)** Correlation plot showing Spearman's Rho between cell-types in HPV-

877 HNSCC, red boxes indicate nonsignificance at q < 0.1. **(h)** IHC showing representative image of CD8

878 and SMA and Kaplan-Meier curves confirming the prognostic impact of TILs and fibroblasts in HPV-

879 negative HNSCC.

880

881 **Figure 2: Classification of HNSCC into hot and cold subgroups on the basis of immune cell infiltration**

882 **patterns (a)** Boxplot of cell-types based on clustering HNSCC. **(b)** Bar-graph showing associations

883 between cytolytic activity and cell types. **(c)** Cytolytic activity is elevated in immune-hot HNSCC. **(d)**

884 CD8/Treg ratios by HNSCC Immune Cluster. **(e)** Mutations significantly associated with HNSCC immune

885 cluster. In boxplots, the ends of the boxes and the middle line represent the lower and upper quartiles,

886 and medians, respectively. Whiskers represent 1.5 times the interquartile range (IQR).

887

888

889 **Figure 3. Identification and characterization of hot and cold tumours pan-cancer (a)** Barplot of

890 distribution of Immune-hot and cold tumours across TCGA. Cancers known to respond favourably to
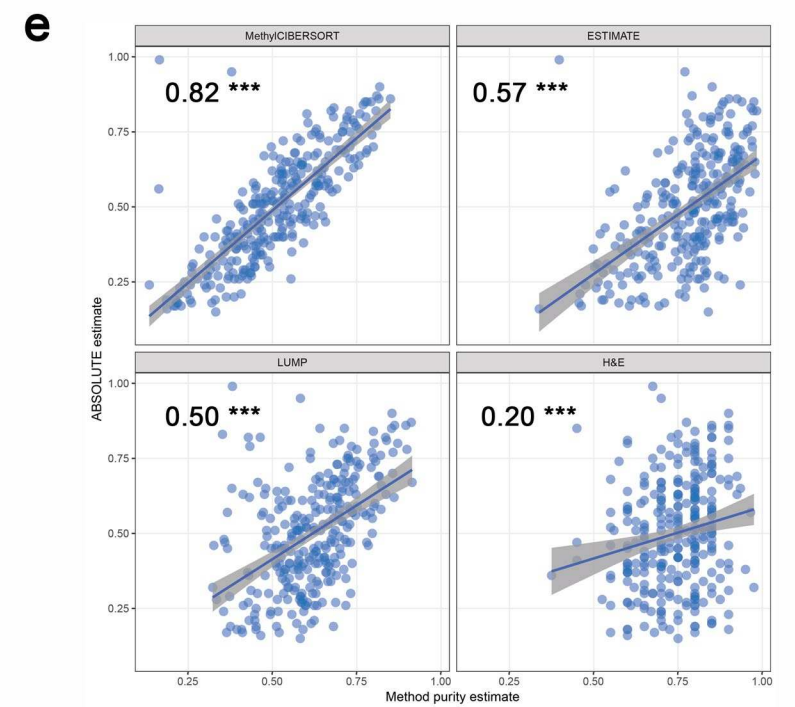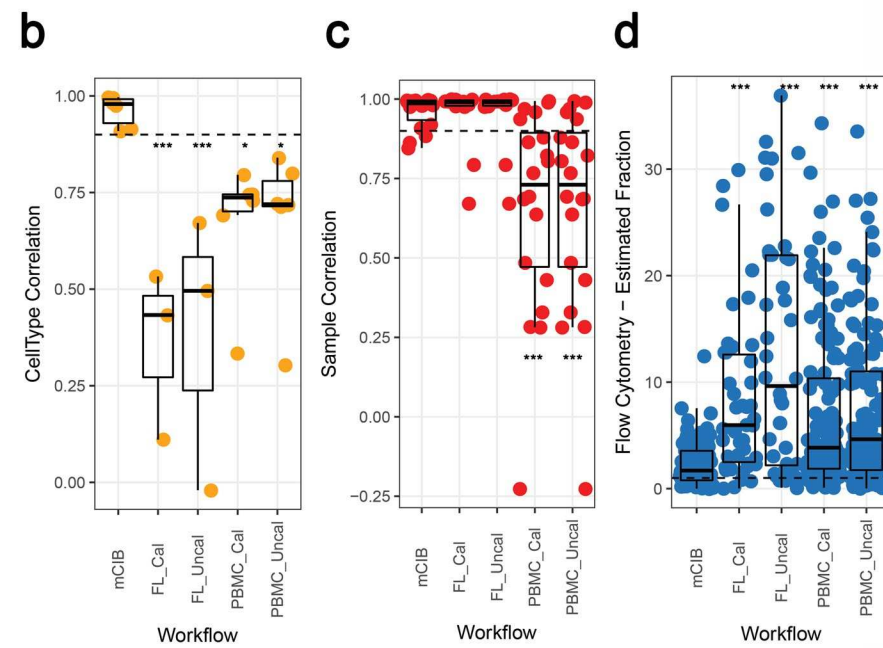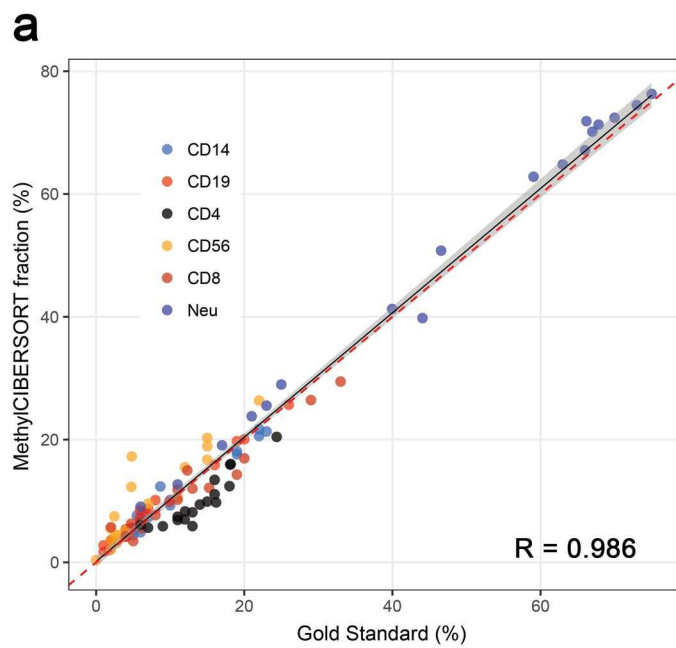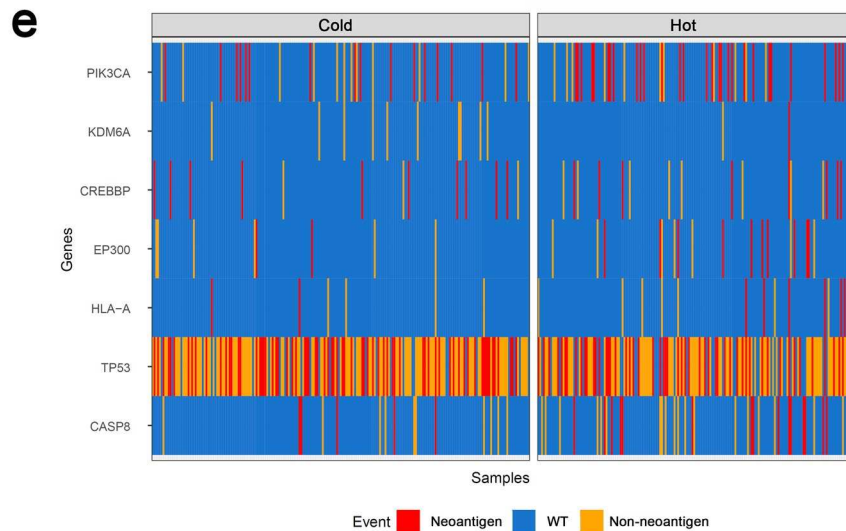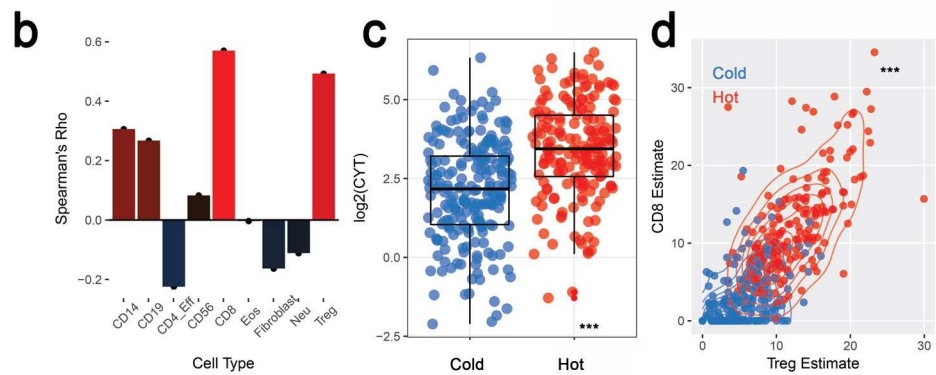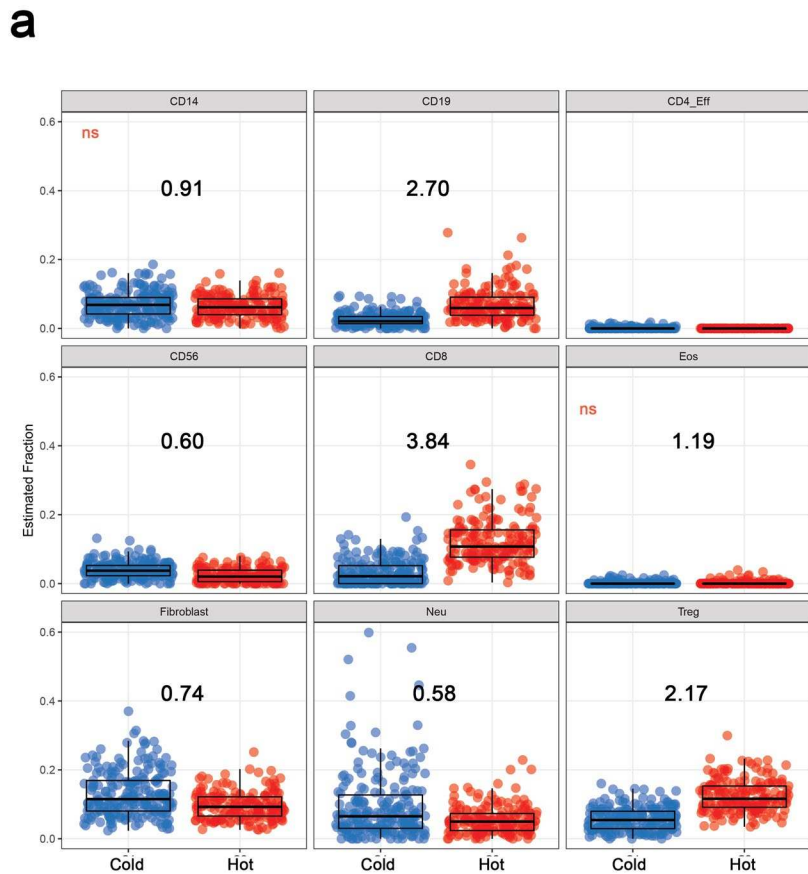
891   checkpoint blockade, such as lung cancer and melanomas, show high fractions of hot tumours. **(b)**

892   Boxplot of cell-type estimates by immune cluster. All at q < 0.05. Numbers represent mean fold

893   changes. **(c)** CD8:Treg ratio is elevated in hot tumours pan-cancer. **(d)** Increased breadth of TCR

894   sequences in Immune Hot tumours (Wilcoxon's Rank Sum Tests). **(e)** Results of IPA canonical pathway

895   analysis comparing hot and cold tumours pan-cancer after adjusting for tumour type. **(f)**

896   Transcriptional deconvolution by Expression-based CIBERSORT shows immune cluster (x-axis) is

897   associated with distinct CD4 polarisation and **(g)** macrophage polarisation (CIBERSORT fractions on y-

898   axis). P.values are from Wilcoxon's Rank Sum Tests. In boxplots, the ends of the boxes and the middle

899   line represent the lower and upper quartiles, and medians, respectively. Whiskers represent 1.5 times
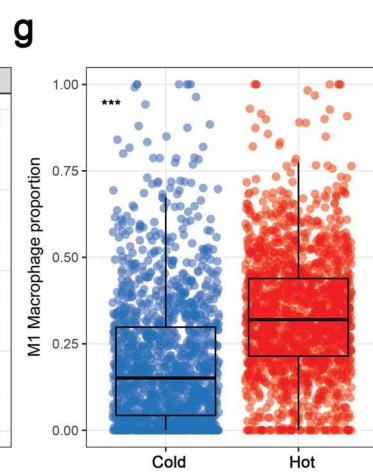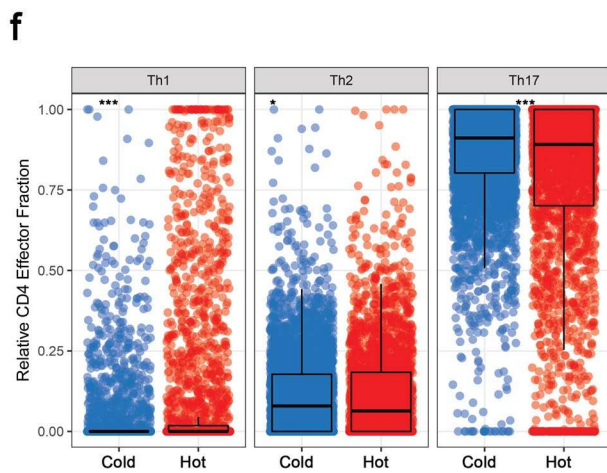
900   the interquartile range (IQR).

901

902   **Figure 4. The immune-hot signature is associated with but not predictive of response to ICB in**

903   **melanoma (a)** heatmap showing expression of the hot-tumour transcriptional signature in Nanostring

904   data from posttreatment biopsies of immunotherapy patients. **(b)** heatmaps showing the same

905   signature in RNAseq data of aCTLA4 (pre-treatment) and aPD1 (post-treatment) respectively. **(c)**

906   boxplots highlighting significant differences in ssGSEA scores for the hot-tumour transcriptional

907   signature in the datasets featured in (b) (p.values from logistic regression). **(d)** barplots display

908   similarity to TCGA hot and cold tumours based on logistic regression class probabilities from a model

909   fit to TCGA data, which are associated with response. **(e)** boxplots showing Kappa values from cross-

910   validation for models examining the performance of the Immune-hot signature, Class I neoepitope

911   burden, and finally mutational load on immunotherapy response classification (p.values from

912   Wilcoxon's Rank Sum Test). In boxplots, the ends of the boxes and the middle line represent the lower

913   and upper quartiles, and medians, respectively. Whiskers represent 1.5 times the interquartile range
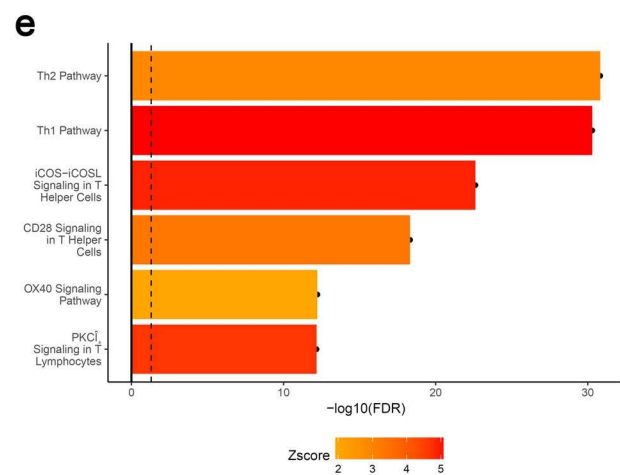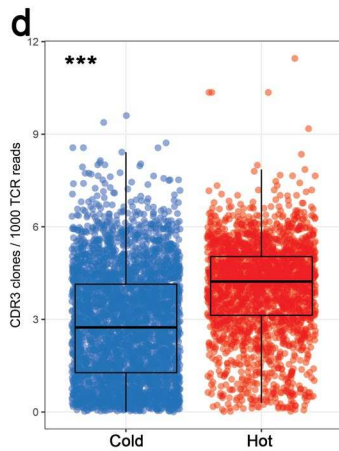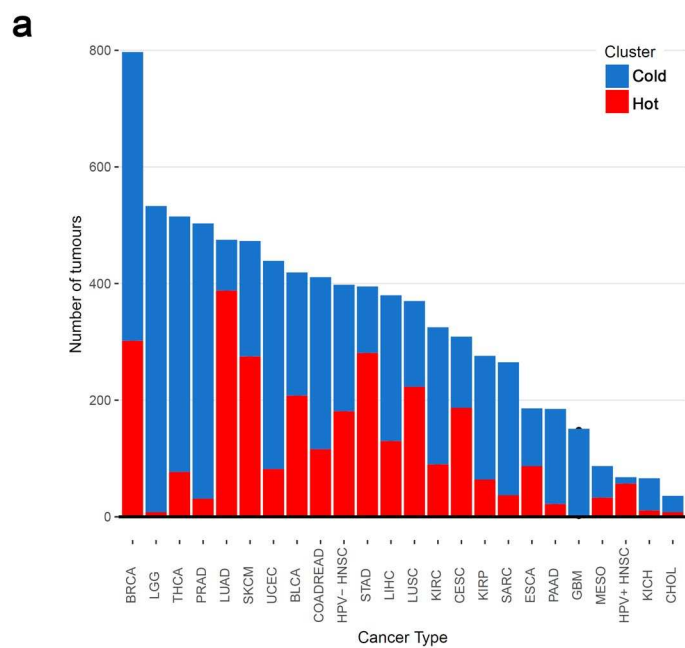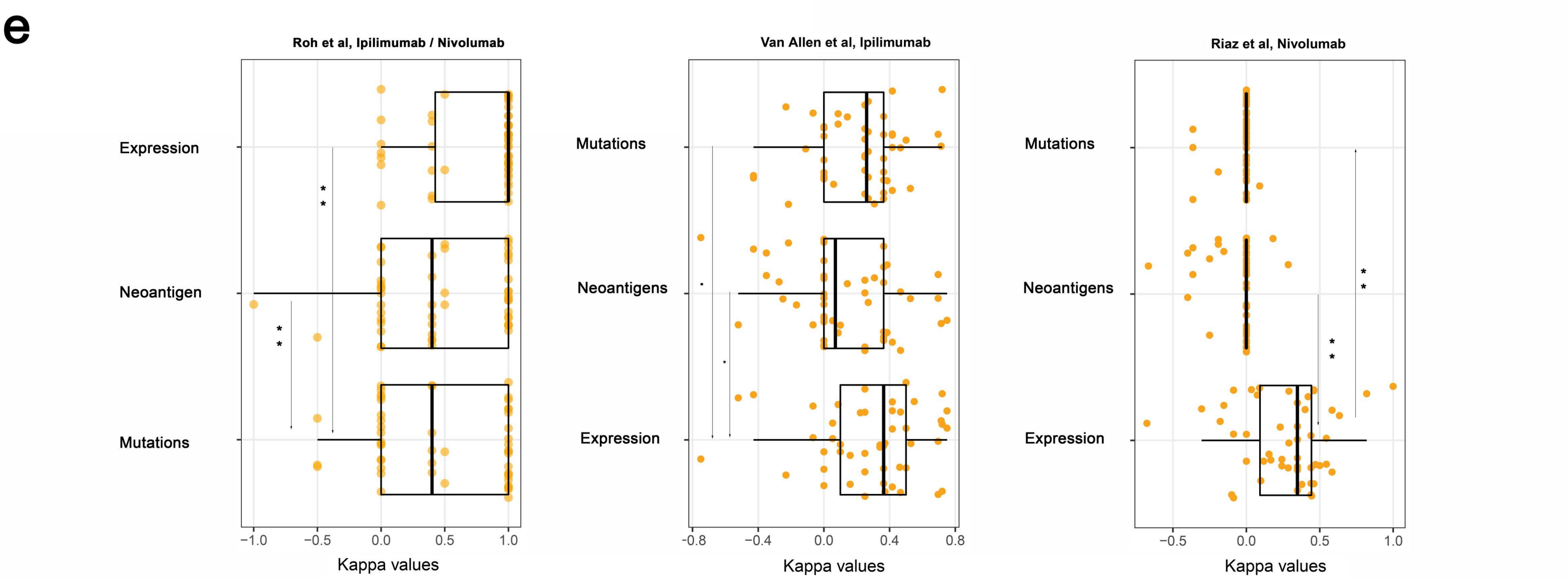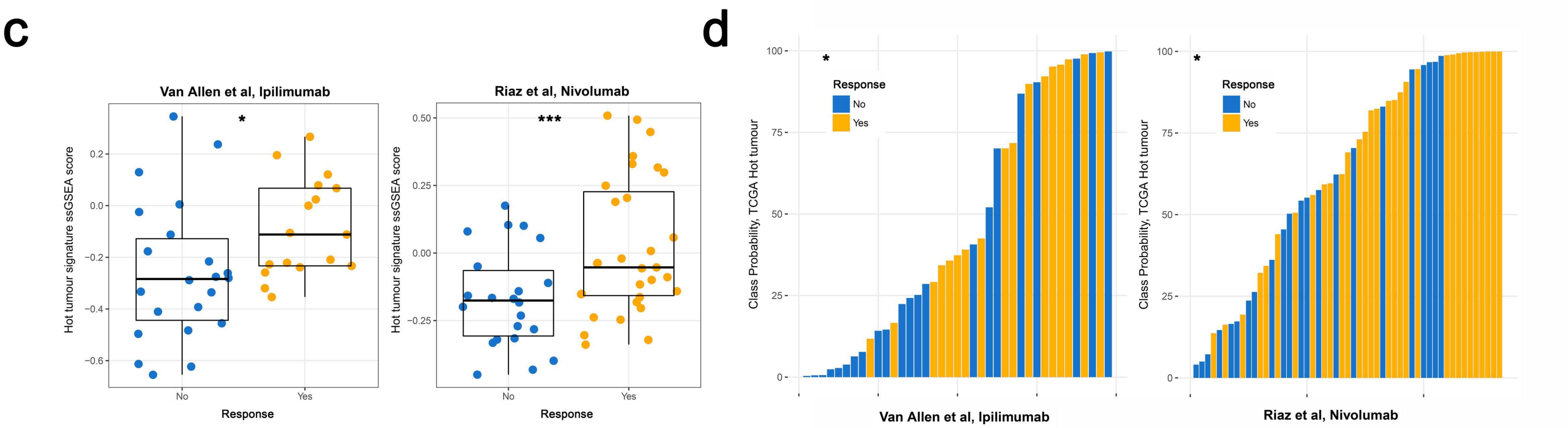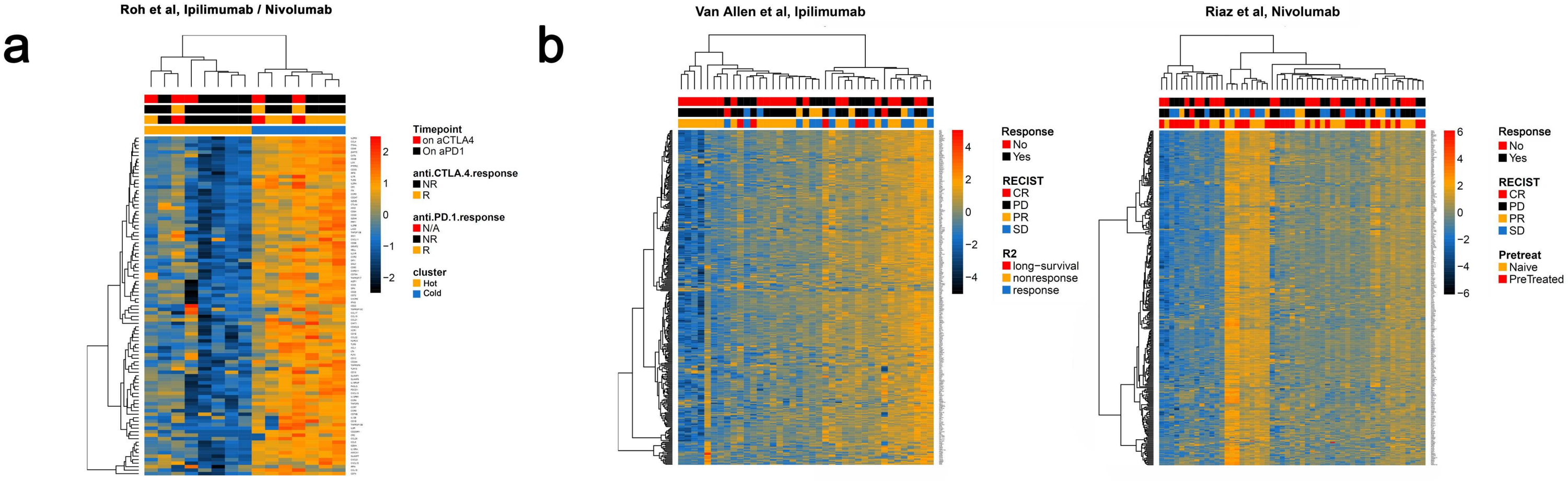
914   (IQR).

915

916 **Figure 5. Genomic features of hot and cold tumours (a)** Density plots showing differences in

917 neoantigen burden by immune cluster pan-cancer. P.value from negative binomial regression that

918 accounts for tumour type. **(b)** Clonal neoantigens and subclonal neoantigens are correlated with

919 different infiltration profiles. Volcanoplot shows Spearman's Rho on the X-axis and –log10(FDR) on the

920 y-axis. **(c)** Volcanoplot showing results of binomial regression testing for associations between

921 Immune-hot cancers and mutation frequencies in candidate cancer driver genes. Those genes

922 implicated in resistance to T-cell mediated destruction are highlighted in orange. **(d)** Volcanoplot

923 showing associations between GISTIC candidate driver copy number peaks and immune cluster. **(e)**

924 Plot showing results of logistic regression in a cohort of HNSCCs where the probability of being

925 classified TIL-high was regressed against anatomic subsite, EGFR IHC (low/moderate/high) and HPV

926 status. **(f)** Correlation between glycolytic coexpression signature ssGSEA scores and EGFR levels by

927 RPPA. **(g)** Association of glycolytic signature post-Nivolumab with response and **(h)** Inverse correlation

928 between the glycolytic signature and the immune-hot expression signature, Spearman's correlation
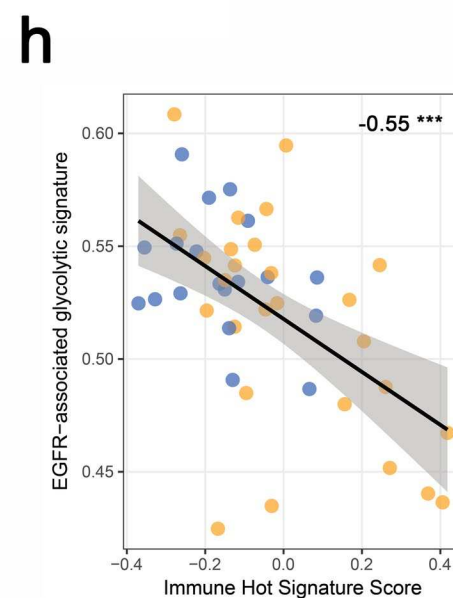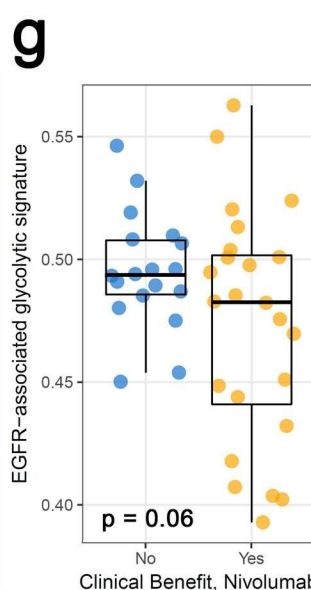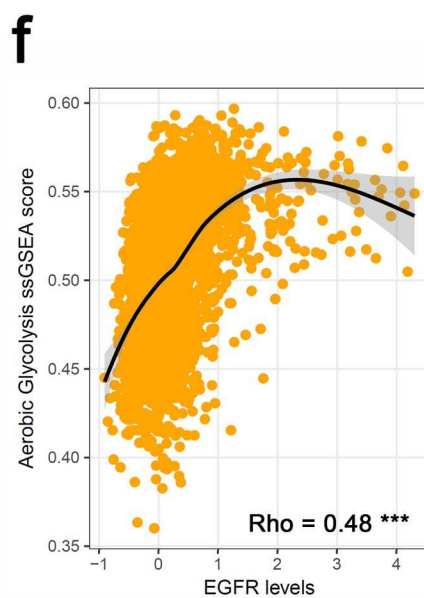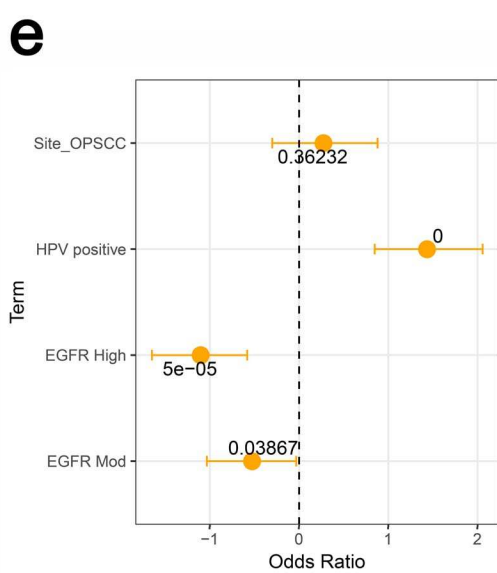
929 has been plotted.

930

a Scatter plot with MethylCIBERSORT fraction (%) on y-axis and Gold Standard (%) on x-axis. Legend: CD14, CD19, CD4, CD56, CD8, Neu. R = 0.986

b CellType Correlation vs Workflow (mCIB, FL_Cal, FL_Uncal, PBMC_Cal, PBMC_Uncal). *** *** * *

c Sample Correlation vs Workflow (mCIB, FL_Cal, FL_Uncal, PBMC_Cal, PBMC_Uncal). *** ***

d Flow Cytometry – Estimated Fraction vs Workflow (mCIB, FL_Cal, FL_Uncal, PBMC_Cal, PBMC_Uncal). *** *** *** ***

e ABSOLUTE estimate vs Method purity estimate.
MethylCIBERSORT 0.82 ***
ESTIMATE 0.57 ***
LUMP 0.50 ***
H&E 0.20 ***

f Estimated Abundance vs Class (Neg, Pos) panels: CD19 ***, CD4_Eff, Treg *, CD8 ***

g Correlation matrix:

CD14 = Monocytes
CD19 = B.Lymphocytes
CD56 = Natural Killer Cells
CD8 = Cytotoxic T Lymphocytes
Eos = Eosinophils
Neu = Neutrophils
CD4_Eff = Helper/Effector Lymphocytes
CD4_Treg = Regulatory T Lymphocytes

	CD14	CD19	CD4_Eff	CD56	CD8	Eos	Fibroblast	Neu	Treg
CD14	1								
CD19	-0.02	1							
CD4_Eff	-0.13	-0.29	1						
CD56	0.24	-0.23	0.22	1					
CD8	0.2	0.51	-0.41	-0.45	1				
Eos	0	-0.03	0.01	0.02	0	1			
Fibroblast	0.35	-0.12	0.02	0.3	-0.12	0.21	1		
Neu	0.07	-0.14	0.02	-0.06	-0.1	-0.5	-0.31	1	
Treg	0.3	0.63	-0.38	-0.17	0.74	0.05	0.05	-0.22	1

h

CD8-high / SMA-low
CD8-low / SMA-high
CD8 SMA

Survival (%) vs Years. N=387. TIL_high, TIL_mod, TIL_low. Log rank: *p<0.001*

Survival (%) vs Years. N=380. SMA_low, SMA_mod, SMA_high. Log rank: *p<0.001*

TIL_high	72	30	7	1
TIL_mod	126	45	6	1
TIL_low	187	48	6	

SMA_low	168	69	10	2
SMA_mod	119	36	6	
SMA_high	93	17	3	

a

CD14 — ns — 0.91

CD19 — 2.70

CD4_Eff

CD56 — 0.60

CD8 — 3.84

Eos — ns — 1.19

Fibroblast — 0.74

Neu — 0.58

Treg — 2.17

Estimated Fraction

Cold   Hot

b

Spearman's Rho

Cell Type: CD14, CD19, CD4_Eff, CD56, CD8, Eos, Fibroblast, Neu, Treg

c

log2(CYT)

Cold   Hot

***

d

Cold
Hot

***

CD8 Estimate

Treg Estimate

e

Cold   Hot

Genes: PIK3CA, KDM6A, CREBBP, EP300, HLA-A, TP53, CASP8

Samples

Event: Neoantigen, WT, Non-neoantigen

Pan-cancer deconvolution of tumour composition using DNA methylation

Chakravarthy et al

Supplementary Information
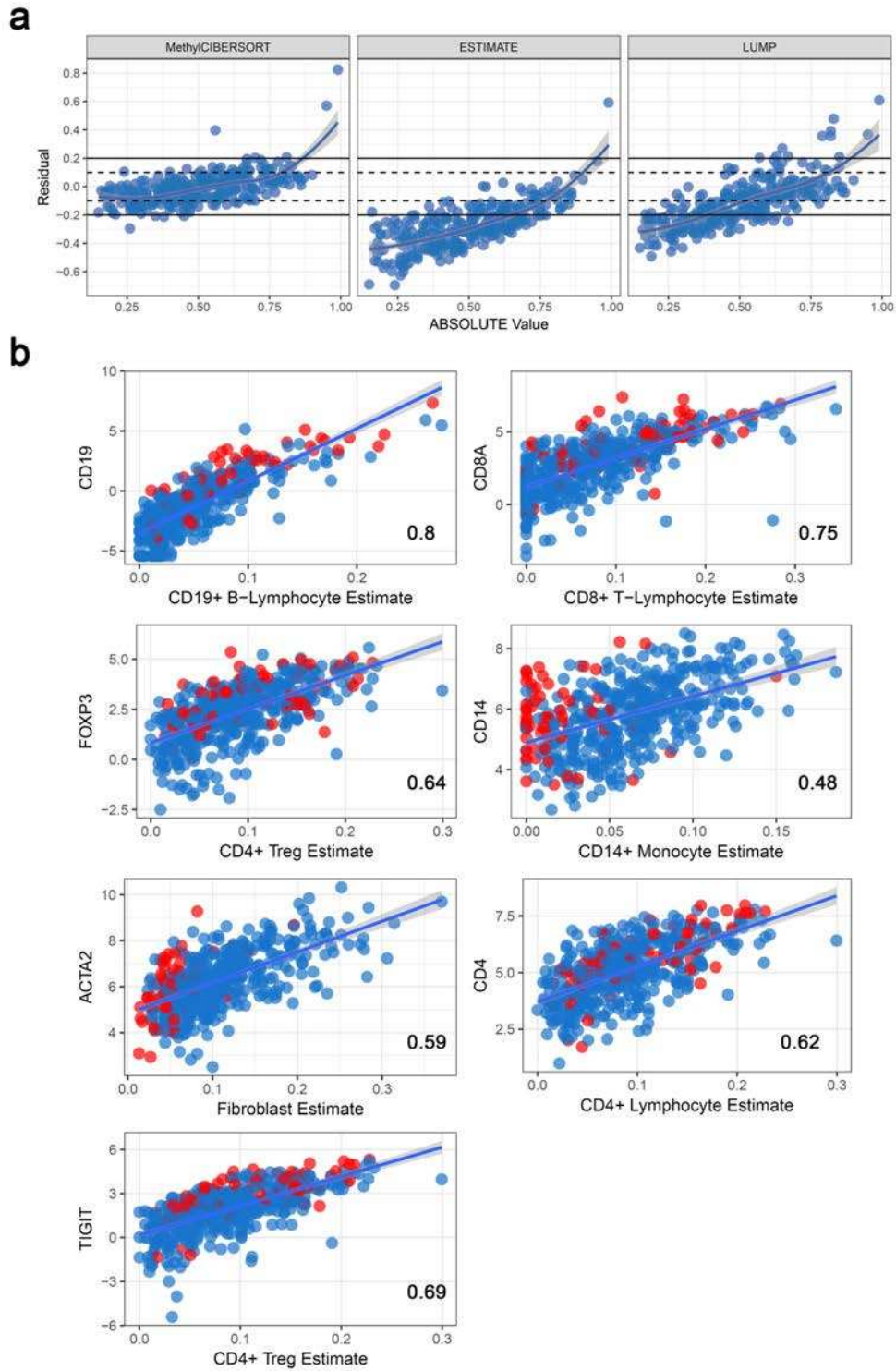
Supplementary Figure 1: (a) Analysis of ABSOLUTE estimate (x-axis) and error from MethylCIBERSORT, ESTIMATE and LUMP in estimating purity in relation (y-axis). **(b)** Correlations (Spearman's Rho) between MethylCIBERSORT estimates and marker gene expression in TCGA HNSCC. Grey areas either side of fitted lines represent 95% confidence intervals.
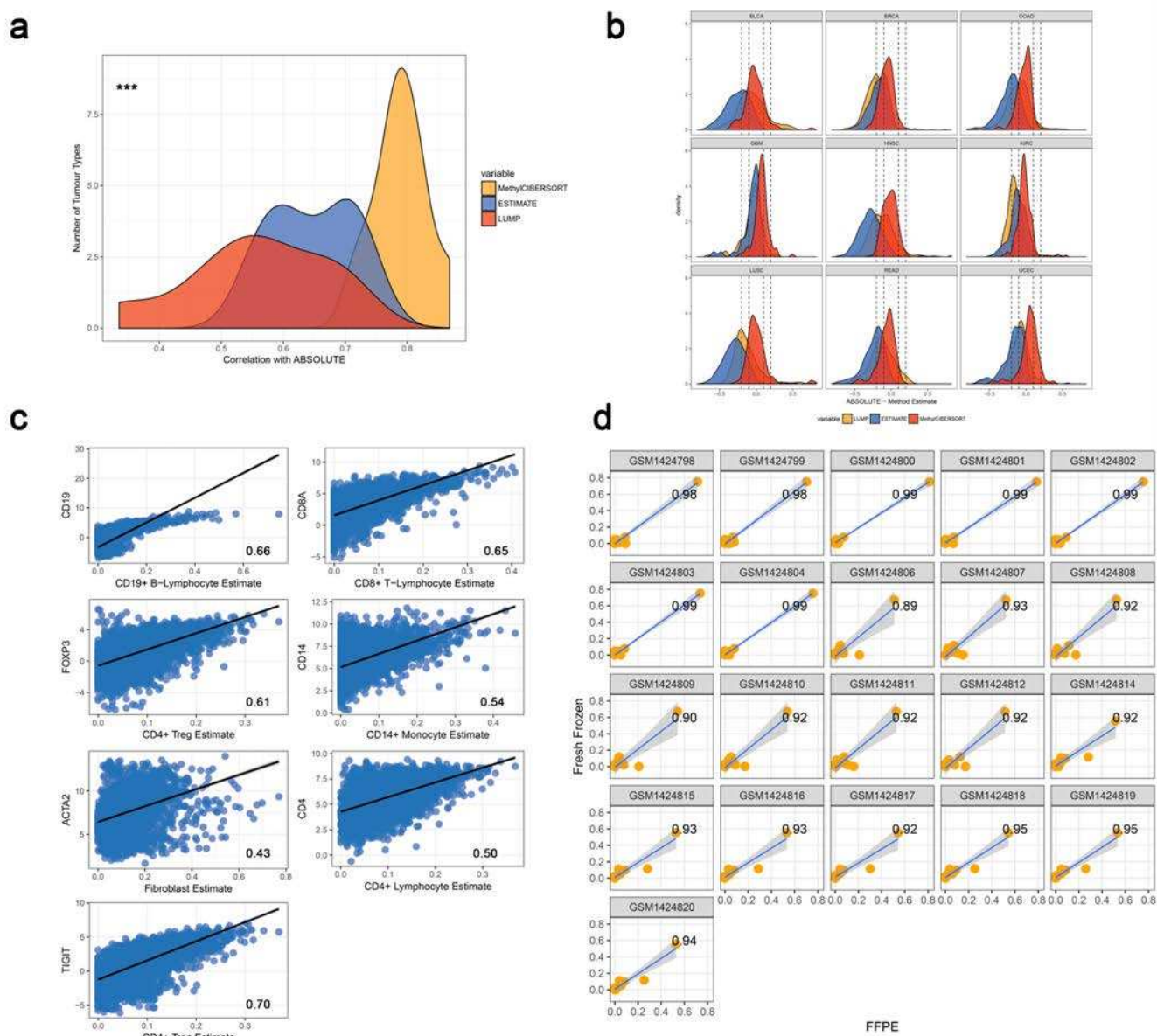
Supplementary Figure 2: (a) Correlation densities for associations with ABSOLUTE purity for MethylCIBERSORT, ESTIMATE and LUMP across tumour types. **(b)** Density plots showing error relative to ABSOLUTE in individual tumour types for MethylCIBERSORT, ESTIMATE and LUMP. **(c)** Marker correlation plots between MethylCIBERSORT estimates and expression of marker genes. **(d)** Correlation plots for 21 450k methylomes relative from FFPE samples relative to their fresh frozen counterparts. Grey areas either side of fitted lines represent 95% confidence intervals.
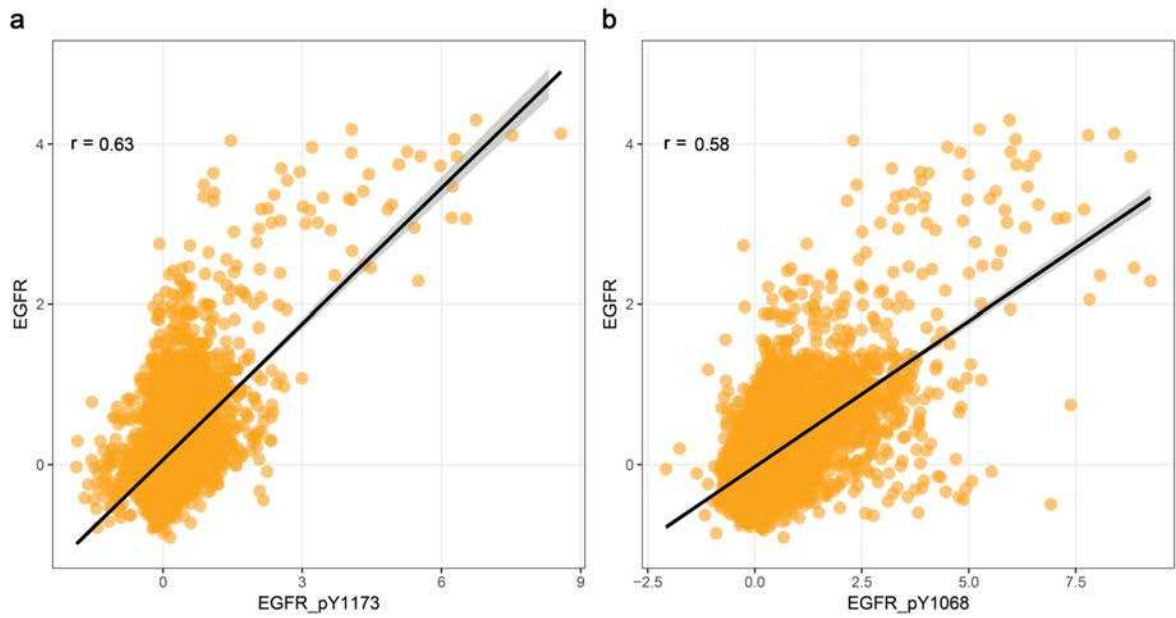
Supplementary Figure 3: Scatterplots showing association between EGFR levels by RPPA and phosphorylation at key activating residues; tyrosine 1173 (a) and tyrosine 1068 (b). Grey areas either side of fitted lines represent 95% confidence intervals. Pearson's correlation coefficients (r) are stated in the panels.