bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Pan-cancer deconvolution of cellular composition identifies molecular correlates of antitumour immunity and checkpoint blockade response.

Ankur Chakravarthy[1,§], Andrew Furness[2], Kroopa Joshi[2], Ehsan Ghorani[2], Kirsty Ford[3], Matthew J Ward[3], Emma V King[3], Matt Lechner[5], Teresa Marafioti[4], Sergio Quezada[2], Gareth J Thomas[3], Andrew Feber[5], Tim R Fenton[6*]

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

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

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

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

[5] Department of Cancer Biology, UCL Cancer Institute, University College London, WC1E 6BT, UK

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

The nature and extent of immune cell infiltration into solid tumours are key determinants of therapeutic response. Here, using a novel 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 hot tumours, 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 and define a catalogue of novel and known mediators of active antitumour immunity, deriving 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[1-5], while a high stromal content of extracellular matrix-producing cancer-associated fibroblasts (CAF), is associated with poor outcomes[6-8]. 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 patients who will derive most benefit from targeted therapies[9]. Although responses to immune checkpoint modulation (e.g. antibodies against PD-1 (programmed cell death protein 1), PD-

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

L1 (programmed death-ligand 1 and CTLA-4 (cytotoxic T-lymphocyte-associated protein 4)) are seen across many solid tumours, the proportion of patients that benefit varies widely by cancer type and we currently lack biomarkers with which to reliably predict immunotherapy response[10]. Emerging evidence from clinical trials indicates higher response rates in those cancer types that typically display greater lymphocyte infiltration (e.g. melanoma, lung cancer, head and neck cancer) and that the tumour neoantigen repertoire (a function of somatic mutation load) is a key determinant[11-14]. These observations point to a model in which, within any given cancer type, there are 'immune hot' and 'immune cold' tumours. Immune hot tumours display greater cytotoxic T-lymphocyte (CTL) infiltration, and reactivation of these tumour-resident CTLs by checkpoint inhibition can result in dramatic tumour regression. Conversely, immune cold tumours display minimal CTL infiltrates and typically fail to respond to checkpoint modulation. If one could accurately identify likely responders for patient stratification and devise strategies by which to convert cold tumours to hot tumours, these would be major steps forward in realising the full clinical potential of cancer immunotherapy. Applying a novel method to estimate tumour composition from DNA methylation data, we set out to address these questions by identifying immune hot and cold tumours across a broad spectrum of cancer types. We aimed to understand the differences in the cellular composition of these tumours and to uncover any common genomic and transcriptomic alterations that are associated with immune responses or immune evasion.

Although flow cytometry of disaggregated tumour biopsies is commonly used for investigating cellular composition, this is often unfeasible for several reasons; difficulty in obtaining fresh tumour tissue; lack of defined markers for poorly characterised cell types (e.g. CAFs); and high cost of labour, reagents and equipment required for such analyses. Cellular disaggregation of collagen-rich tumours is also problematic, where cells are embedded in a dense extracellular matrix. To overcome these difficulties, multiple reference-free or reference-based methods have recently been developed to

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

permit the *in-silico* deconvolution of complex cellular mixtures or to estimate tumour purity [15-22]. For example, accurate deconvolution of complex cellular mixtures, including tumours, has recently been achieved by application of support vector regression modelling (CIBERSORT) to gene expression microarray data[21,23]. Notably, DNA methylation data are also suitable for deconvolution of tissue mixtures, although studies so far have focussed primarily on simple tissues such as blood, where cell type differences are a major confounder in Epigenome Wide Association Studies[17].

Here we apply CIBERSORT-based deconvolution to genome-wide DNA methylation data from whole tumour tissue (hereafter referred to as 'MethylCIBERSORT') which permits accurate tumour cell deconvolution of both fresh and archival samples. Notably, unlike other available methods, MethylCIBERSORT estimates the tumour cell content (tumour purity) of a sample, in addition to performing deconvolution, thus providing absolute rather than relative estimates of infiltrating cell fractions. Initially focussing on head and neck cancer (HNSCC), a tumour type where we have previously demonstrated the prognostic significance of tumour-infiltrating lymphocytes (TILs), particularly in those cancers driven by human papillomavirus (HPV)[1,24], we extended our technique to 21 further solid malignancies. As expected, the proportion of immune hot tumours varies widely by cancer type, but immune hot tumours are found even among those malignancies that typically display very little immune infiltration, such as pancreatic ductal adenocarcinoma or prostate adenocarcinoma. We leverage both DNA methylation and gene expression based deconvolution to reveal differences in the phenotype, from Th1/M1 pro-inflammatory in hot tumours, to Th2/M2 pro-fibrotic in cold tumours. Our genomic analysis reveals multiple copy number alterations enriched in cold tumours, including deletions in *PTEN* and amplifications in *MYC* and *EGFR*. We show that responses to PD1-blockade are associated with a transcriptional signature for hot tumours post-treatment, while the cold signature, and specifically a gene expression module we previously linked to increased aerobic glycolysis downstream of EGFR in HNSCC[25], is enriched in non-responders.

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Importantly however, defining whether a tumour is hot or cold is not sufficient to accurately predict response to immune checkpoint blockade. By interrogating matched genomic data, we find that (presumably due to selective pressures imposed by the adaptive immune system during their evolution) treatment-naive hot tumours frequently display genomic alterations known to confer immunotherapy resistance. Our findings provide an explanation for the failure of immune checkpoint blockade in a subset of well-infiltrated tumours and identify genomic biomarkers for patient stratification. Building upon recent analyses of immune infiltrates and tumour gene expression profiles[20,23] or molecular correlates of cytolytic activity[26], we reveal fundamental, cross-cancer patterns of cellular infiltration and their relationships with genomic make-up with important implications for immunotherapy.

Results

Development and validation of methylation-based deconvolution using CIBERSORT.

To develop a DNA methylation based deconvolution pipeline for application in tumours, we developed a custom R interface to develop basis matrices for use with CIBERSORT and generated a reference using fibroblasts and seven different immune cell types (see methods for details). We then evaluated the ability of our feature selection heuristic to accurately deconvolute mixtures of leukocytes using publicly available methylation data from mixtures of PBMCs with composition verified by flow-cytometry (gold standard). This showed extremely high correlation between the estimated and gold-standard fractions (Pearson's $R = 0.986$, $p < 2.2e\text{-}16$, Figure 1A). We also carried out benchmarking against the performance of RNA-based CIBERSORT using the LM22 basis matrix against leukocyte mixtures of similar resolution originally profiled in Newman et al[21]. This revealed that MethylCIBERSORT estimates demonstrate higher correlations, both at the cell-type and the sample level (Figure 1B, C) and significantly lower absolute error (Figure 1D). Thus, methylation data

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

coupled to CIBERSORT is highly accurate and may offer distinct advantages relative to expression-based CIBERSORT.

To validate the method on real tumour samples, especially the tumour content in order to permit absolute quantification of tumour composition, we applied our pipeline to generate an HNSCC specific basis matrix and applied it to the set of 464 HNSCCs that have both RNA-sequencing and DNA methylation profiles available from The Cancer Genome Atlas (TCGA) project [27]. Upon comparing cancer cell proportion (purity) estimates derived using MethylCIBERSORT with estimates derived from ABSOLUTE [28] (which jointly estimates purity and ploidy using mutation and copy number data) relative to other previously published methods of estimating purity (LUMP[29] and ESTIMATE[30]) with data aggregated in[29], MethylCIBERSORT displayed the highest correlation (R=0.82) and better concordance with ABSOLUTE than other methods **(Figure 1E).** Analysis of residuals (method estimate – ABSOLUTE estimate) suggested close concordance with ABSOLUTE estimates for MethylCIBERSORT, with larger deviations only seen when samples were of very high purity (>80%), while other methods tended to overestimate tumour cell content in samples of low purity (**Figure S1A**), resulting in statistically significant differences in distributions (FDR < 2.2e-16).

We also compared the mRNA expression of a panel of cellular lineage markers with MethylCIBERSORT estimates and found significant associations for multiple cell types **(Figure S1B)** even though they are derived from different samplings of the same tumour. Many of these marker genes demonstrated more variable expression in tumours with lower estimates of infiltrating cell fraction, suggesting that low coverage on either or both platforms (RNA-seq and methylation array) at the lower end of cellular abundance may result in poorer concordance. Taken together, these observations confirm that MethylCIBERSORT can accurately deconvolute the mixed cell populations in tumour samples using DNA methylation data.

Detection of increased B- and T-lymphocyte infiltration in HPV-associated HNSCC

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Having established the potential of MethylCIBERSORT to identify patterns of cellular infiltration in solid tumours, we tested its ability to detect the elevated TIL levels previously documented in HPV-driven (HPV+) HNSCC[1]. MethylCIBERSORT detected not only the increased TIL levels in HPV+ HNSCC compared with HPV- HNSCC (p= 2.167e-05 , Wilcoxon's Rank Sum Test) but more specifically attributed this to increased numbers of B (CD19+) and cytotoxic T (CD8+) lymphocytes (CTLs, **Figure 1D**), in agreement with observations made using other methods including immunohistochemistry and gene expression analysis[31], potentially also helping to explain favourable prognosis displayed by this subgroup, independent of treatment modality[32-34].

Deep deconvolution highlights novel associations between infiltrating cell types and identifies two distinct patterns of infiltration in HNSCC

Next, we extended our analysis to HPV-negative (HPV-) HNSCC, a heterogeneous, anatomically-diverse group of tumours in which prognosis is typically much poorer than in HPV+ disease. Again, using TCGA data (available for 398 HPV- HNSCCs) we observed interesting relationships between multiple cell types, with 24/36 pairs of cell types showing significant correlations (Spearman's rank correlations, FDR<0.01**; Figure 1D)**. CTLs are associated with both CD14+ (monocytes / macrophages / myeloid-derived suppressor cells) and B-lymphocytes (Rho =0.14 and 0.55). CD4+/FoxP3- T-lymphocytes (CD4 + effector T-lymphocytes), meanwhile display inverse correlations with CTLs (R= -0.27) and Tregs (R=-0.49). CD56+ Natural Killer (NK) cell abundance is also inversely correlated with CTLs (R=-0.48). Of note, CTLs are inversely correlated with fibroblast abundance (R=-0.16) and to validate this latter finding, we analysed data from two large studies in which these parameters had been quantified in HNSCC[1,7]. In a pooled analysis of these data, TIL content and SMA expression (a CAF marker) are inversely correlated (r=-0.322 and -0.344 for CD8 and CD3 IHC in the Ward (oropharyngeal SCC) cohort (**Figure 1E**); -0.4 and -0.424 for TIL scoring of H&E sections in the Ward (oropharyngeal SCC) and Marsh (oral SCC) cohorts respectively). They are also strongly prognostic (**Figure 1F**; p<0.001, Log Rank Test).

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Given the complex nature of associations between different cell types, we performed consensus PAM clustering on the estimated cellular fractions to define subgroups by infiltration patterns. We derived two clusters ('immune cold' (C1) and 'immune hot (C2)' that show markedly different distributions of multiple cell types, most notably CTLs, CD4+ effector T-lymphocytes, CD19+ B-lymphocytes and NK cells, all of which are implicated in antitumour immunity **(Figure 2A)**. Consistent with our previous observations, estimates of fibroblast content are higher in the immune cold group (mean fold change 1.31, FDR < 1.8e-6, Wilcoxon's Rank Sum Test) and in multivariate Cox regression analysis, controlling for stage and age at diagnosis, membership of the Immune-cold group is associated with significantly shorter overall survival (**Figure 2B;** HR=1.42, CI=1.04–1.96, p=0.03). To explore the functional significance of our observations, we tested for associations between individual cellular fractions or immune cluster and a recently defined measure of local cytolytic activity based on the expression of Granzyme A and Perforin 1 (*GZMA* and *PRF1*; markers of activated T-cells)[26]. All infiltrating cell fractions display significant correlations with cytolytic activity, with CD8+ cells showing the maximum positive correlation (**Figure 2C**, FDR < 0.05, Spearman's Rank Correlation). Accordingly, the immune hot cluster displays significantly higher cytolytic activity (**Figure 2D**, p = 2e-16, Wilcoxon's Rank Sum test), and increased ratios of CTLs to Tregs (**Figure 2E**, p < 2e-16, Wilcoxon's Rank Sum test); a metric that is prognostic in multiple settings[35-37].

Together, our analyses suggest a tilting of the balance towards CTL activity in the microenvironment of tumours from the immune hot cluster that may explain the increased overall survival in this subgroup. Integrated analysis of the impact of different cell populations on cytolytic activity using linear modelling identified CD8+ (coef = 0.12, p < 2e-16), CD14+ (coef = 0.11, p <3.3e-8), Tregs (coef = 0.06, p < 0.003) and CD56+ (coef = 0.18, p < 8.8e-10) cell abundance as positive predictors and fibroblast abundance as a negative predictor (coef = -0.07, p <1.9e-10).

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Infiltrating cell patterns are associated with distinct transcriptional and proteomic profiles in HNSCC.

Having established that these infiltrating cell patterns are of prognostic relevance in HNSCC we investigated if they are associated with different molecular profiles. Using limma-trend analysis, we identified 457 genes differentially expressed (DEGs) between the immune hot and immune cold clusters at a fold-change of greater than 2 (FDR<0.01**, Table S1**, genes highlighted in bold). Multiple DEGs are consistent with the MethylCIBERSORT-derived estimates of lymphocyte infiltration; *CD8A, ZAP70* and *CD3D* (CD8 lymphocyte markers), *CD79A* and *CD19* (B-lymphocyte markers), are all upregulated in the lymphocyte-enriched cluster, as are multiple chemokines and their receptors (*CCL5, CCR5, CXCR2, CXCR6, CCL19, CXCL11)*), immune checkpoint gene transcripts (*LAG3, PD1, IDO1, CTLA4*), immunosuppressive enzymes *(ADORA2A, IDO1)* and as expected, the cytolytic markers *PRF1* and *GZMA*. In extended analyses of all genes at FDR < 0.01 **(Table S1)**, multiple other genes, including the Class 1 MHC gene *B2M* (FC = 1.3), PD-1 ligand *CD274* (FC = 1.5) and *ACTA2*, which encodes SMA *(*FC = 0.71), are also differentially expressed between the two clusters, the latter validating fibroblast estimates from MethylCIBERSORT (**Figure 2A**). Of note, the increased expression of PD-1 ligand (PD-L1) the immune hot group suggests tumours of this subtype are more likely to respond to anti-PD1 checkpoint inhibition[38].

Ingenuity Pathway Analysis further confirmed observations made using MethylCIBERSORT estimates, identifying differential regulation of multiple canonical pathways associated with immune function and inflammatory conditions (**Table S2)**, consistent with differential lymphocyte infiltration and activity. Diseases and functions ontology (**Table S3)** indicated that the top few pathways activated in the immune hot cluster were associated with leukocyte and lymphocyte migration. Upstream regulatory analysis implicated increased activation of the chromatin-modifying factors EHF and EZH2, and inhibition of Interferon-stimulated transcription mediated by IRF4, in the lymphocyte-rich tumours **(Table S4)**. In addition, OX40/OX40L signalling is predicted to be upregulated in immune cold tumours (**Table S2**). OX40 is a co-stimulatory molecule expressed on T-lymphocytes and is one

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of a number of targets currently under early clinical investigation for immune checkpoint modulation therapy[39]. OX40 signalling opposes differentiation of CD4+ cells into Tregs and antagonizes Treg function, potentially explaining its reduction in Treg-rich immune hot tumours[40]. Finally, analysis of RPPA data identified 7 differentially abundant (FDR<0.1) proteins or phospho-proteins **(Table S5)**. Higher levels of cleaved Caspase 7 (FC=1.46) in the immune hot subgroup indicates increased apoptosis**,** whereas Fibronectin and PAI1 upregulation in immune cold tumours indicate a distinct pattern of TGFβ-driven extracellular matrix remodelling in what may be a CAF-linked phenomenon.

Distinct mutations are associated with HNSCC immune cluster.

Having established that immune cluster is associated with distinct transcriptional patterns, we then sought to identify individual mutations in driver genes (MutSig CV [41] q.value < 0.01) associated with immune cluster using Negative Binomial regression. This identified *CASP8, NSD1, NOTCH1, EP300, HLA-A, RAC1, and RB1* as significantly more mutated in immune hot HNSCC and *TP53* to be less mutated (Figure 2F).

CASP8 mutations are implicated in subverting apoptosis induced by lymphocytes; they are enriched in tumours with high immune cytolytic activity and likely reflect an increased selective pressure exerted by the presence of adaptive immune cells[26,42]. Fas-ligand *(FASLG)*, an upstream activator of pro-apoptotic signalling through Caspase 8[43] is also upregulated in the immune hot tumours, further highlighting the importance of this pathway (**Table S1**). Identification of this lymphocyte-rich, good prognosis group displaying *CASP8* mutations and a relative lack of *TP53* mutations is striking, since TCGA previously identified a subset of good-prognosis oral cavity tumours bearing the same genomic hallmarks, which were reported to co-occur with *HRAS* mutations[27].

Neoantigen burden has previously been identified as a predictor of anti-tumour immune responses[26,44,45] and consistent with this, we identified significantly higher neoantigen burdens in the

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

immune hot cluster (OR = 1.54, p < 1.8e-6, Negative Binomial GLM) and a smaller increase in overall mutational burden (OR = 1.23, p = 0.008). Strikingly, the differentially mutated driver genes by themselves tended to encode predicted neoantigens significantly more often in the Immune-hot cancers (OR = 1.44, p = 0.01, Fisher's Exact Test). Moreover, in 16 tumours from the immune hot cluster versus 4 in the immune cold cluster, *CASP8* mutations themselves encoded at least one neoantigenic peptide (**Figure 2F**), demonstrating the existence of mutations that both contribute to the development of a potential selective constraint, and serve as an adaptive mechanism to evade it.

Our findings provide new insight into the *CASP8*-mutant / *TP53*-WT / *HRAS*-mutant prognostic subgroup identified by TCGA and suggest increased CTL infiltration as a potential mechanism to explain the improved outcomes seen in these tumours. The ability of our approach to rediscover a prognostic subgroup previously defined by genomic profiles using an independent approach highlights the potential of DNA methylation based immune cell fraction deconvolution. We then sought to use our inferences in HNSCC as a tool for investigating the immune microenvironment pan-cancer, permitting analyses where variation between tumour types could be accounted for to produce a comprehensive molecular portrait of cancer immunosurveillance.

HNSCC-derived immune clusters are reproducible across tumour types.

To examine whether the relationships between tumour composition, genomic alterations and clinical behaviour we observed in HNSCC are generally applicable, we derived cancer-type specific basis matrices) and conducted deconvolution on 18 further tumour types for which cancer cell line methylation data have recently been published[46]. For 9 of these we were able to compare our predictions of tumour purity with ABSOLUTE estimates and observed strong correlations and significantly lower error margins compared to LUMP and ESTIMATE(**Figure S2A, S2B**). Further, we observed a robust preservation of positive correlations between MethylCIBERSORT and marker expression pan-cancer (**Figure S2C**), again with the caveat that the samples were taken from

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

different aliquots of the tumour. Taken together, these findings attest to the general pan-cancer applicability of MethylCIBERSORT. An important potential advantage of DNA-methylation over gene expression-based deconvolution methods is the ease with which accurate DNA methylation profiles can be obtained from formalin-fixed, paraffin-embedded (FFPE) samples [47-49]. We therefore compared estimates pertaining to fresh frozen and matched FFPE samples (n = 21 from 3 tumours)[50] and recorded very high correlations, indicating our method is applicable also to archival material (**Figure S2D**).

We then trained an elastic-net classifier using 5-fold cross-validation for tuning on the HNSCC cellular abundance data, returning highly accurate recapitulation of clustering (Kappa = 0.9), and predicted immune cluster membership for the validation set of 7269 samples representing 21 further tumour types from TCGA (**Figure 3A**). As expected, we observed strong enrichment for CTLs, Tregs, and B-lymphocytes in immune hot tumours pan-cancer, while CD4-effectors, NK cells, eosinophils and CAFs were enriched in immune cold tumours (**Figure 3B**). Different tumour types also display markedly varying degrees of lymphocyte infiltration, with the majority of pancreatic ductal adenocarcinomas, colorectal, thyroid, uterine corpus endometrial, kidney, prostate, hepatocellular cancers and sarcomas belonging to the immune cold cluster (**Figure 3A**). We again observed increased CTL:Treg ratios in immune hot tumours (**Figure 3C**) and similar relationships between tumour composition and CYT to those seen in HNSCC.

Immune hot tumours display greater evidence of immunoediting and are marked by Th1/M1 responses.

To further determine if the immune infiltrate was active in these tumours, we assayed immunoediting by testing for reductions from the expected ratio (as previously defined by Rooney et al[26]) of observed neoantigens to total nonsilent mutations per tumour and adapted this approach to derive the estimated number of neoepitopes lost through immune editing while controlling for tumour type. Accordingly, we found significant enrichment for editing in immune hot tumours

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

compared to immune cold tumours (OR = 1.23, p = 0.008, Negative Binomial GLM). Additionally, upon integration with T-cell receptor (TCR) repertoire data from Li et al[51], we found more diversity (Number of TCR clones / Total number of TCR reads) in the immune hot tumours (**Figure 3D, p < 2.2e-16**), suggesting that broader immune responses may underlie the greater depletion of neoantigens in this group.

Given the evidence for divergent infiltration patterns and activity between the immune clusters across cancer types, we then investigated the determinants of this response by identifying differentially expressed genes after adjusting for tumour type. We identified 275 genes at FDR < 0.01, 2FC and in pathway analysis, the top pathways were significantly associated with Th1 vs Th2 responses (**Figure 3E, Table S6**). Multiple Th1 cytokines and downstream targets were overexpressed in hot tumours (*IFNG, CCL4, CCL5, CXCL9, CXCL10*), along with costimulatory and coinhibitory receptors, suggesting these tumours were marked by a state of lymphocyte activation and counter-responses thereto. We next scored proinflammatory (Th1, Th17) and suppressive (Th2)) CD4+ cell populations using RNA-seq reference profiles from purified cells to derive relative estimates using CIBERSORT[21]. Consistent with our inferences from pathway analysis, we found enrichment for Th1 and Th17 cells in immune hot and Th2 cells in immune cold cancers (Figure 3F) , translating to markedly elevated Th1/Th2 ratios in hot tumours (p < 2.2e-16, Wilcoxon's Rank Sum Test), with smaller but significant increases in Th17/Th2 ratios (p = 1.2e-5, Wilcoxon's Rank Sum Test). Importantly, T-helper 2 (Th2) cells have been linked to poor prognosis in multiple studies, while Th1 cells are associated with good prognosis and aiding CTL responses[52].

We also used expression-based CIBERSORT to derive estimates for different myeloid cell populations (n = 2346 tumours at deconvolution P < 0.05), and identified substantially higher fractions of M1 relative to M2 macrophages in hot tumours (p = 2.2e-16, Wilcoxon's Rank Sum Test, **Figure 3G**). Notably, M2-like polarisation is associated not only with Th2 immune responses but also with immune-suppressive Myeloid Derived Suppressor Cells (MDSCs)[53]. Taken together, our analyses

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

implicate Th1/Th17cytokine signalling programmes as responsible for establishing an immune-hot state and define MDSCs and Th2 programmes as targets for efforts to switch immune cold tumours to an immune hot state.

Transcriptional correlates of immune clusters predict responses to checkpoint blockade and suggest TCGA hot tumour transcriptomes are linked to active antitumour immunity.

We reasoned that if the signature for immune hot tumours represented active immunity, it would be applicable to the prediction of immunotherapy responses and evaluated this hypothesis using tumour gene expression data from three melanoma cohorts: post-sequential aCTLA4 and aPD1 treatment[54,55]; pre-aCTLA4 treatment[14] and post-aPD1 (Nivolumab) treatment[56]. Analysis of these transcriptional patterns indicated differential expression between responders and non-responders (**Figure 4A**), and accordingly, ssGSEA scores for the hot transcriptional signature showed significant enrichment in responders for the latter two datasets (**Figure 4B, C**). Moreover, a similar association emerged from comparing the probability of response to hot/cold class prediction, inferred using a logistic regression fit on TCGA hot/cold transcriptional signature ssGSEA scores (**Figure 4D**). Finally, we evaluated the ability of the hot-signature to stratify patients by response relative to mutational load and Class I neoepitope burden using elastic nets coupled to cross-validation for each dataset (**Figure 4E**). This showed that in the post treatment data, the immune-hot signature outperforms neoantigen and mutational burdens respectively, and for pretreatment aCTLA4 data, performs similarly to mutational burden and neoantigen burden. Notably, however, even in the post-treatment setting, the presence of the hot signature does not translate to a guaranteed response, indicating potential prior selection for genetic and epigenetic alterations that confer resistance to T-cell mediated killing in otherwise immunogenic tumours

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Taken together, these findings establish the fundamental similarity between cancers responsive to immunotherapy and immune-hot tumours across a wide range of cancer types, and suggest that genomic features that determine immunotherapy response should also be enriched in immune hot TCGA tumours. We reasoned that the heterogeneity in responses to checkpoint blockade among hot tumours might be driven by intrinsic resistance to T-cell mediated destruction due to pre-existing genomic alterations within the tumour cells. We set out to test this hypothesis by constructing a pan-cancer catalogue of genomic alterations enriched in hot tumours, with the additional aim of finding those alterations enriched in cold tumours which might drive lymphocyte exclusion or reduce tumour immunogenicity.

Genomic features of hot and cold tumours.

Consistent with our observations in HNSCC, and with estimates from gene expression-based deconvolution[20], immune hot tumours harboured higher overall mutation loads (OR = 1.4, p = 1.64e-31, negative binomial GLM controlling for cancer type) and more predicted neoantigens than immune cold tumours (**Figure 4A**).

Given the recent finding that in addition to the presence of neoantigens, their clonality (i.e. presence in all tumour cells as opposed to minor subclones) is associated with prognosis and response to Pembrolizumab in lung adenocarcinoma[12], we analysed immune microenvironment composition as a function of neoantigen clonality (as denoted by The Cancer Immunome Atlas[57]). We found that the abundance of both CTLs and Tregs is correlated with clonal neoantigen load pan-cancer (**Figure 4B**), while the relationship is much weaker when subclonal neoantigens are considered. CD4+/FOXP3- effector lymphocytes display a striking inverse correlation with clonal neoantigens (**Figure 4C**). Consistent with our earlier observation they are enriched in CTL / Treg low immune cold tumours CAFs are inversely correlated with both clonal and subclonal neoantigen loads. Immune hot tumours

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

display a significantly higher clonal neoantigen burden (OR = 1.236, p < 2.2e-16, Negative Binomial GLM) as well as a skew in the neoantigen burden towards clonal neoantigens after adjusting for tumour type (OR = 1.03, p = 1.6e-5). These findings provide the first evidence for a direct link between Class I MHC clonal neoantigen burden and patterns of TIL abundance and help to explain the observations of McGranahan and colleagues, that high clonal neoantigen burden predicts favourable response to immune checkpoint modulation using Pembrolizumab[12]. While it is unclear why clonal neoantigens should elicit a proinflammatory Th1 response, some previous work has suggested that the antigen dose may determine the nature of the subsequent immune response[58], or alternatively, clonal neoantigens may simply have been subjected to immunosurveillance for far longer, and this is borne out by improved immunosurveillance in mouse models when mismatch repair deficiency is induced clonally instead of subclonally[59].

Finally, we examined if the genomic features associated with immune cluster were also reproducible across cancer types, performing adjusted binomial regressions to estimate cluster association after controlling for tumour type for genes previously implicated as pan-cancer drivers based on signatures of positive selection[60] and recorded 114 hits at FDR < 0.1 (Table S7). Interestingly, these putative drivers of hot tumours were significantly enriched (OR = 8.43, p < 0.002, Fisher's Exact Test) in a list of genes demonstrated to confer resistance to CD8 T-cell mediated killing in a recent CRISPR-Cas9 screen [61]. These functionally-verified immune-resistance genes included components of the MHC Class I complex such as *B2M* and *HLA-A,* apoptosis pathway genes such as *CASP8*, and *JAK1*, which is required for interferon mediated signalling that in turn is associated with resistance to checkpoint blockade[62-64]. We also show the recently-identified immunotherapy sensitizers (*ARID2, BRD7*[65]*)* to be disproportionately mutated in hot tumours, together with tumour suppressors such as *RB1* [66] , *TP53* [67],*LATS2* [68], *ATRX* [69] and *SETD2* [70], all of which have been associated directly or indirectly with interferon responses, additional epigenetic regulators *(KMT2A, TET2, IDH1, NSD1, KDM6A, KMT2B, BCOR, and NCOR)*, and finally, multiple DNA-damage associated proteins, including *BRCA1*, *BAP, TOP2A* and *CDK12*. Taken together, these point to a model where mutations in certain genes

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

render tumours hot as a consequence (and therefore susceptible to checkpoint blockade), or may enable tumours to survive in a hot tumour microenvironment, potentially also bestowing resistance to checkpoint blockade. We sought to test this model by linking immune-resistance mutations to lack of immune checkpoint blockade response in hot tumours and although we observed a trend, (odds ratio= 0.26), the number of treated tumours with sequence data available is currently too low (41 hot tumours across four studies) to gain a definitive answer.

To complement these analyses, we also carried out copy number analyses, calling copy numbers across 11,000 tumours and testing for differential association of peaks with immune cluster after adjusting for tumour type for the subset with Immune Cluster Assignment available. This led to the identification of multiple events that occurred at different frequencies between cold and hot tumours (FDR < 0.1, **Figure 4B**). Of these, prominent examples included amplifications targeting the epidermal growth factor receptor (*EGFR*) (7p11.2) and *MYC* (8q24.3) and deletions at 10q23.31, encompassing the *PTEN* tumour suppressor gene in cold tumours and *JAK2* (9p24.1) amplifications in hot tumours. Some of these candidates already have known associations with immune evasion; *MYC* has been linked to an immune evasion phenotype that is amenable to targeting through gene-body demethylation [71] and *PTEN* deletion has recently been linked to failure of immunotherapy and decreased cytotoxic T-lymphocyte infiltration in patients and in a mouse model of melanoma[55,72,73]. Among the genomic alterations we identify (for full list of predicted driver events see **Table S8**), it is likely that some establish, while others are selected for, in different immune microenvironments. In either case, alongside *PTEN* deletion, these alterations warrant further investigation as candidate genomic markers for response to immune checkpoint blockade. The enrichment of *EGFR* and *MYC* amplification, together with *PTEN* deletion in cold tumours pan-cancer was striking given the co-expression module linked to increased tumour cell glycolysis and immune evasion in HNSCC, which includes *EGFR* and in which pathway analysis also predicts increased c-MYC and mTORC1 activity[25]. A similar relationship has been observed in triple negative breast cancer[74] and we therefore investigated this relationship further, initially interrogating the link between *EGFR* protein levels and

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TILs in two HNSCC cohorts (n=518)[1,7] we found that samples classified as EGFR high and moderate were significantly more likely to be TIL low than EGFR low cancers after accounting for anatomic site and HPV status (Figure 5E, $p < 0.05$ and 0.01 for EGFR moderate and high cancers, Logistic Regression). The positive correlation between EGFR levels (which are themselves correlated with EGFR phosphorylation (activation), Figure S3) and the glycolytic signature is maintained across TCGA when matched RPPA profiles and RNA-seq data are compared (Fig 5E). Notably, high levels of the glycolytic signature are present in progressing melanomas after PD-1 blockade (Fig 5F, $p = 0.06$, t-test, $p = 0.02$ when excluding stable disease) and are inversely associated with expression of the immune-hot transcriptional signature (Rho = -0.44) we associated with response (Fig 5G).

Discussion

Of the methods developed to deconvolve cell mixtures into multiple cell types from methylation data, none have been comprehensively employed across cancer types. Methods such as LUMP and the Leukocyte Methylation Score estimate only the overall leukocyte fraction, while methods based on expression data either produce relative estimates of abundance within the immune fraction or enrichment scores(CIBERSORT, TIMER[20]) or perform low resolution deconvolution (ESTIMATE)[21,29,75,76]. Combining methylation-based feature selection from both stromal and cancer cells with the robust performance of CIBERSORT previously displayed on gene expression microarray data[21] allowed us to derive estimates for different infiltrating cell populations as a fraction of the overall sample.

While approaches using RNA-sequencing or other transcriptional profiling, such as the construction of an index of cytolytic activity, have been useful in predicting immunotherapy response[77] and in identifying the role of mutations in genes like *CASP8* in immune evasion[26], the deeper level of deconvolution made feasible using DNA methylation data allows the roles of distinct cellular subsets and their interdependencies to be dissected. Here, by applying the method to HNSCC, which is

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

447 marked by a great degree of clinical heterogeneity, we identified lymphocyte-rich and stromal-rich
448 prognostic subgroups consistent with those discovered previously using a variety of independent
449 methods [1,7,31,78-81] and derived novel insights into the microenvironmental alterations that modulate
450 prognosis. In the process, we showed that our scheme for classifying cancers correlates with well-
451 established immune metrics such as cytolytic activity, neoantigen/mutational load, and CTL:Treg
452 ratios, and then demonstrated that tumours similar to the HNSCC Immune-hot subgroup, which by
453 association exist in varying fractions across the vast majority of cancer types, and the congruence of
454 our classification with the aforementioned metrics is maintained throughout, and translates to
455 broader TCR responses which in turn correspond to greater depletion of neoantigens.

456 We then demonstrated that the transcriptional correlates of our classification scheme are indicative
457 of active antitumour immunity through Th1 responses and reinforced this theme through integration
458 with RNA-based CIBERSORT that indicated differences in CD4 and macrophage polarisation. We
459 went on to show that the hot-tumour transcriptional programme, if induced upon checkpoint
460 blockade, is strongly associated with response, indicating that it represents active antitumour
461 immunity and establishing the TCGA as a resource to study genomic alterations that may associate
462 with checkpoint blockade resistance/sensitivity. Upon extending our analysis pan-cancer, we made
463 several observations that give novel insight into the interplay between tumour genomics and the
464 immune microenvironment. Genomic analysis identified significant enrichment for events that
465 confer resistance to T-cell mediated destruction in hot tumours as well as potential sensitisers. Our
466 copy number analysis revealed that *PTEN* deletion, *MYC* amplification and *EGFR* amplification are
467 associated with immune depletion.

468 All of these mediate increased glycolysis, which we have previously linked to immune evasion[82]. Our
469 finding that *PTEN* deletion is associated with poor CTL infiltration in this pan-cancer cohort adds
470 substantial support and mechanistic rationale for its proposed role as a determinant of response to
471 immune checkpoint blockade. Taken together with our identification of *EGFR* and *MYC* amplification

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in cold tumours, our analysis suggests that pharmacological inhibition of EGFR/mTORC1/MYC-driven glycolysis could be an effective means by which to 'warm-up' these tumours and potentially enhance responses to immune checkpoint blockade. Additionally, our analysis of neoantigen clonality and immune infiltration patterns adds mechanistic insight to the value of clonal neoantigen burden in predicting response to immune checkpoint blockade[12]. In particular, we show that clonal neoantigens are associated with infiltration of CTLs and Tregs, while Th2 cells and CAFs are enriched in tumours with lower clonal neoantigen loads. These findings support recent evidence suggesting that differentiation of naïve CD4+ T-lymphocytes into Tregs occurs within tumours[83], since a microenvironment favouring differentiation into Tregs would likely be selected for in tumours with increased neoantigens and more infiltrating CTLs. Why these relationships between neoantigen loads and T-lymphocytes are apparent only when one considers clonal neoantigens is an intriguing question. It could be that since many subclonal neoantigens are expressed by a small minority of cells within the tumour, these evade effective presentation to the immune system. Indeed, in a previous study by several of the authors, it was possible to isolate T-lymphocytes reactive against clonal but not subclonal neoantigens from lung cancer patients[12]. Our data suggest that this is due to a relative paucity of CTLs in tumours with low clonal neoantigen loads and that this is true across a wide range of cancer types. In summary, the development of a stand-alone method to estimate both tumour purity and stromal composition from DNA methylation data has allowed us to make a number of novel insights that shed light on potential biomarkers for immunotherapy response and the way in which tumour genomes influence, and are shaped by, the immune microenvironment. Finally, the association between neoantigen clonality, the probability of being immunologically hot, and enrichment for known mediators of immune evasion/resistance in hot tumours before treatment with any checkpoint blockade also raises profound questions about the use of immunotherapy; the features that render tumours susceptible to immune destruction are likely to have existed far back into the evolutionary history of these tumours, increasing the likelihood that natural selection will have already produced genomic alterations that doom eventual checkpoint

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

blockade to failure. Thus, we can stratify immune-hot tumours into two groups: those without aberrations in immune-resistance genes that we may expect to respond to checkpoint blockade, and those with resistance mutations, in which these either these aberrant pathways will likely have to be co-targeted, or alternative therapies may be more suitable. Finally, the lack of immune-resistance mutations in cold tumours (presumably due to the absence of a selection pressure for them) suggests that if we can induce lymphocyte infiltration (e.g. by targeting glycolysis or CAFs[84]), we may improve the effectiveness of checkpoint blockade across a broader range of patients.

Methods

Code Availability.

All scripts and functions developed for our method will be made freely available in an R-package upon publication of this manuscript. R markdowns for analysis code will be available by request upon publication.

Development of a methylation signature for* in-silico *deconvolution.

Dataset Assembly and Preprocessing

Raw data were obtained in the form of IDAT files from the following sources (the number of samples from which each profile was derived is shown in parentheses): Granulocytes (12), CD8+ (cytotoxic T-lymphocytes) (6), CD19+ (B-lymphocytes) (6), CD56+ (Natural Killer cells) (6), CD14+ (monocyte lineage) (6), Eosinophils (6) were from the Blood.450k Bioconductor package[85]. CD4+ cells were removed from the Blood.450k dataset and CD4+ T-cells from the Zhang dataset[86] (data kindly provided by Dr Alicia Oshlack) were further divided into FOXP3+ (Tregs) (4) or FOXP3- (6) groups. Fibroblast profiles (4) were from the Gene Expression Omnibus (GSE74877). Neutrophils are the

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

most abundant subset of granulocytes and these samples were therefore aggregated into a single category for further analysis. To generate a DNA methylation signature for cancer cells, we used 450k methylation profiles we previously obtained from a series of 6 HNSCC cell lines: UM-SCC47; 93VU147T; UPCI:SCC090; PCI-30; UPCI:SCC036 and UPCI:SCC003 (GSE38270, described in[87]) and additionally those from Iorio et al (GSE68379) .

The files were parsed into R using the *minfi*[88] Bioconductor package and were normalised using single sample Noob as implemented in *minfi*.

Derivation of signature features

A custom limma based wrapper function was used to fit a series of linear models for all pairwise comparisons between candidate cell types. Features from this set of analyses were then restricted to MVPs that showed a median beta-value difference of 0.25 at an FDR of 0.01 for that fit or less, with a maximum of 100 MVPs per pairwise comparison. Finally, for use with CIBERSORT, data were transformed from beta values (bound between 0 and 1) to percentages (0 – 100). Type-wise means were estimated for each probe and cell type and the matrix exported for upload to CIBERSORT.

Benchmarking using PBMC mixtures.

We applied the feature selection pipeline to the matrix of stromal cells that we assembled and then tested performance against 450k profiles of PBMC mixtures with flow-cytometry gold standards. We also applied *LM22* (Expression-based CIBERSORT) to datasets consisting of PBMC samples and Follicular Lymphoma biopsies originally evaluated in CIBERSORT[21]. Wilcoxon's rank sum tests were used to test for differences in correlations with flow-cytometry for cell types and samples, and absolute errors between flow-cytometry and deconvolution estimate.

For the Expression CIBERSORT estimates, we performed comparisons against both calibrated (i.e, enforcing a sum-to-one constraint as reflected in the flow cytometry) and uncalibrated (straight estimates of cell fractions from CIBERSORT) estimates.

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Running Deconvolution Experiments on HNSCC using CIBERSORT.

Data for 464 methylation profiled TCGA HNSCC samples were downloaded in the form of raw IDAT files for the 450k array from the TCGA data. Data were normalised using functional normalisation[89] in the minfi[88] package and BMIQ[90], with 10,000 reference probes for Expectation Maximisation fitting. HPV status was determined using VirusSeq[91] based on detection of viral gene transcripts.

Beta values for deconvolution associated features, and the signature matrix derived in the previous step, were uploaded to CIBERSORT at https://cibersort.stanford.edu. The data were not quantile normalised due to the potential for global methylation shifts in cancers, and CIBERSORT was run using 1000 permutations. Output files were downloaded as tab-delimited text files and custom parsers were used to import results into R for downstream analysis. FFPE methylation profiles for 42 HNSCC were obtained from Gene Expression Omnibus (Accession GSE38266) using the GEOquery R package, and beta values were BMIQ normalised and analysed using CIBERSORT as described for the TCGA cohort. Wilcoxon's Rank Sum Tests were used to test for differences in total TIL abundance and TIL subsets.

Estimating accuracy of MethylCIBERSORT in tumour deconvolution

In the absence of flow-cytometry based estimates for the different cell types in the analyzed tumours, the estimated fraction of cancer cells from MethylCIBERSORT was compared to sequencing-data based estimates from ABSOLUTE available for 466 HNSCCs from previously published work[29] using Spearman's Rank Correlation. Correlations were between ABSOLUTE and other methods of estimating purity/immune cell fraction in this subset of tumours; (LUMP, ESTIMATE[76] and H&E staining assessment of tumour purity (data available in[29]). Residuals were computed by subtracting the method estimate from the ABSOLUTE value. Distributions were compared using Wilcoxon's Rank Sum Test. Spearman's Rank Correlation was used to estimate

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

correlations between expression of marker transcripts and MethylCIBERSORT estimates for multiple cellular populations. Where applicable, multiple testing correction was performed using the Benjamini Hochberg approach.

Clustering and correlation analyses.

Estimates of immune cell fractions in HPV- HNSCC (HPV-transcript negative) were examined for correlations with other infiltrating cell types using Spearman's Rank Correlation with BH correction for multiple testing. Clustering was carried out using the clusterCons package with 100 iterations using a manhattan distance metric. The most robust number of clusters was then selected.

Differences in the distribution of infiltrating cell types by immune cluster were summarised using mean fold changes and tested using Wilcoxon's Rank Sum Test with BH-correction for multiple testing.

Differentially expressed genes were identified using Limma-trend and were defined at a threshold of a 2-fold change and BHFDR < 0.01. Pathway analysis was carried out using Ingenuity Pathway Analysis, with findings restricted to experimentally confirmed direct interactions in human cells/tissues. Cytolytic activity (CYT) was calculated as the geometric mean of *GZMA* and *PRF1* expression as defined previously[26]. To estimate the contributions of cell population abundances to this, a linear model was fit against log2(CYT) with the different populations as predictors. Wilcoxon's Rank Sum tests were used to test differences in CYT and CD8:Treg ratios between the immune clusters.

Survival Analyses

Multivariate Cox Regression was used to estimate the prognostic utility of clusters derived using infiltration patterns with age and stage as covariates. The survival effect of estimated purity was regressed with the same covariates using a Cox regression with coefficients defined per percent increase in purity.

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genomic Correlates

We obtained a list of driver genes inferred by MutSigCV[41] in TCGA HNSCC cohort from the Broad Institute's GDAC. GISTIC Copy number estimates thresholded by genes were also obtained from this source. MAF files were obtained from the TCGA data portal. MutSigCV drivers were filtered at a q.value threshold of 0.01 and mutations in this set were tested for differences in frequencies of occurrence using a chi-squared test for differences in proportion. Multiple testing correction was carried out using the Benjamini-Hochberg method. Tables of predicted neoantigens were downloaded from The Cancer Immunome Atlas (http://tcia.at).

Benchmarking performance across other tumour types

Signature features were derived from 450k profiles using the aforementioned heuristic (delta-Beta and FDR cutoffs) with a maximum of 100 features per cell type for a wide range of tumour types, using cell lines allocated to the corresponding tissue in GSE6837916[46] (**Table S10**) and the aforementioned infiltrating cell types (**Table 1**). These signatures were applied to deconvolve methylation profiles and estimates of purity were derived using TCGA samples for which ABSOLUTE, ESTIMATE and LUMP purity estimates were available[29].

The cell line data were functionally normalized with the infiltrating cell types described earlier before signature extraction was carried out. 450k data for the aforementioned tumour types were loaded from a pan-cancer freeze derived from SAGE synapse for TCGA pan-cancer (syn2812961) and a custom function was used to extract signature probes and generate methylation percentage matrices for deconvolution with CIBERSORT. CIBERSORT was run as described previously. Correlation and residuals analysis were carried out as described above with MethylCIBERSORT purity estimates vs ABSOLUTE, and between previously published methods and ABSOLUTE. Wilcoxon's

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Rank Sum Test with Benjamini-Hochberg correction for multiple testing were used to compare distributions, with these estimates sourced from[29].

Pan-cancer analyses of immune cluster assignment.

An elastic net model was fit using cellular abundance estimates for HPV- HNSCC using 3 iterations of 5-fold cross-validation to identify the optimal values of lambda and alpha with Kappa values being the selection criterion. The classifier was then applied to MethylCIBERSORT estimates from 18 further tumour types for which corresponding cancer cell line methylation profiles were available[46] **(Table S10)** to allocate immune cluster. Deconvolution was performed as described above and class allocations were made using the elastic net classifier derived from HNSCC.

For immunoediting analyses, we estimated the number of nonsynonymous mutations encoding at least one immunogenic peptide empirically by summing coefficients across each of six base change contexts as well as the number of non-neoepitope nonsynonymous mutations. Together , these were applied to silent mutation counts in each cancer to derive an expected fraction of neoantigens to nonneoantigens. Comparing the observed fraction to the expected fraction yielded the percentage of neoantigens depleted, and using this in combination with the number of observed neoantigens yielded the count of neoantigen-encoding mutations lost specifically to immunoediting. This was then modelled using a negative binomial framework to estimate the influence of immune cluster on immunoediting.

MAF files for mutations were again downloaded from SAGE synapse for all tumours from the MC3 calling effort (syn7214402). Driver mutations were defined based on pan-cancer MutSig analyses previously published[41] and binomial GLMs were used to estimate coefficients for mutation frequencies for immune cluster with tumour type as a covariate. Significant genes were defined at BHFDR < 0.1. Survival analyses were performed using data downloaded from Synapse (syn7343873) using Cox proportional hazards regression with stage and cancer type as covariates. Substages were

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

aggregated into stages and only Stages I-IV were considered. Neoantigen abundance and clonality data were downloaded from The Cancer Immunome Atlas[92].

Negative binomial modelling was used to model all count data, cytolytic activity was modelled using linear models, and binomial GLMs were used to model proportions. Details of covariates, hypotheses and response variables are presented inline. For copy number analyses, SNP6 data were downloaded from the GDC data portal and processed using GISTIC 2.0[93] on the GenePattern Public Server (arm-level peel off, noise threshold 0.3, FDR < 0.01, driver-gene confidence > 95%) and modelled similarly to mutation data.

Further resolution of cell-types using expression-based CIBERSORT.

RNA-seq data were downloaded from the European Nucleotide Archive for the following datasets: PRJEB11844[94]; GSE60424[95]; and E-MTAB-2319[96]. Kallisto[97] was used to quantify gene expression with a reference transcriptome consisting of Gencode Grch37 assembly of protein coding and lincRNA transcripts. Data were then modelled using limma trend and the top 50 markers by t-statistics were selected for each cell subset from one versus all comparisons after thresholding with a 2-fold change and FDR < 0.05. These cell types were used to generate a reference profile and CIBERSORT was run to deconvolute samples. For M1/M2 macrophage analyses we used LM22 from the CIBERSORT server as the reference. In both cases, Wilcoxon's Rank Sum Test was used to estimate differences in distributions.

Analysis of Immunotherapy response

Nanostring data for a panel of immune genes and and exome sequencing data were obtained from Chen et al[54] and Roh et al[55] respectively for patients treated using sequential anti-CTLA4 and anti-PD1 checkpoint blockade.

Clustering and machine learning were carried out using the subset of genes intersecting with the Hot-vs-Cold pancancer signature. .632 bootstrapping was used for hyperparameter tuning and ROC

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

estimation. Negative binomial regression was used to model neoantigen, mutation and subclone numbers, and logistic regression to estimate predictive performance of count data on response.

The number of subclones present in each tumour from the Roh cohort, derived from the EXPANDS algorithm, were obtained from the associated publication[98]. RNAseq data were obtained for aCTLA4 pretreatment biopsies by personal communication with Eliezer Van Allen and genomic data from the associated publication. Data for post-treatment Nivolumab treated melanomas were obtained from [56].

Valiation of EGFR association with cold tumours.

RPPA data were downloaded for TCGA cancers from the TCPA portal. IHC data were derived from[1,7] for comparison of EGFR protein levels vs TIL levels, previously defined in[1,7]. ssGSEA scores were used to summarise the activity of the glycolytic gene signature (described in) and standard statistical procedures were used to assess interrelationships.

Acknowledgments

This work was funded by grants from Rosetrees Trust and Debbie Fund (TRF) and a UCL graduate scholarship (AC). We are grateful to Dr Alicia Oshlack for raw DNA methylation data from CD4+ and CD4+/FoxP3+ lymphocytes, Elizier Van Allen for aCTLA4 expression data, and Bo Li and Xiaole Shirley Liu for TCR repertoire data. The results published here are, in part, based upon data generated by The Cancer Genome Atlas (TCGA) project established by the National Cancer Institute and National Human Genome Research Institute. Information about TCGA and the investigators and institutions who constitute TCGA research network can be found at http://cancergenome.nih.gov.

References

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1. Ward, M.J., *et al.* Tumour-infiltrating lymphocytes predict for outcome in HPV-positive oropharyngeal cancer. *Br J Cancer* **110**, 489-500 (2014).
2. Galon, J., *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313**, 1960-1964 (2006).
3. Piras, F., *et al.* The predictive value of CD8, CD4, CD68, and human leukocyte antigen-D-related cells in the prognosis of cutaneous malignant melanoma with vertical growth phase. *Cancer* **104**, 1246-1254 (2005).
4. Brambilla, E., *et al.* Prognostic Effect of Tumor Lymphocytic Infiltration in Resectable Non-Small-Cell Lung Cancer. *J Clin Oncol* **34**, 1223-1230 (2016).
5. Quigley, D.A. & Kristensen, V. Predicting prognosis and therapeutic response from interactions between lymphocytes and tumor cells. *Mol Oncol* **9**, 2054-2062 (2015).
6. Hanley, C.J., *et al.* A subset of myofibroblastic cancer-associated fibroblasts regulate collagen fiber elongation, which is prognostic in multiple cancers. *Oncotarget* **7**, 6159-6174 (2016).
7. Marsh, D., *et al.* Stromal features are predictive of disease mortality in oral cancer patients. *The Journal of Pathology* **223**, 470-481 (2011).
8. Orimo, A., *et al.* Stromal Fibroblasts Present in Invasive Human Breast Carcinomas Promote Tumor Growth and Angiogenesis through Elevated SDF-1/CXCL12 Secretion. *Cell* **121**, 335-348 (2005).
9. Furness, A.J., Vargas, F.A., Peggs, K.S. & Quezada, S.A. Impact of tumour microenvironment and Fc receptors on the activity of immunomodulatory antibodies. *Trends Immunol* **35**, 290-298 (2014).
10. Patel, S.P. & Kurzrock, R. PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. *Mol Cancer Ther* **14**, 847-856 (2015).
11. Le, D.T., *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* **372**, 2509-2520 (2015).
12. McGranahan, N., *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463-1469 (2016).
13. Rizvi, N.A., *et al.* Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* **348**, 124 (2015).
14. Van Allen, E.M., *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207 (2015).
15. Houseman, E.A., *et al.* Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* **17**, 1-15 (2016).
16. Houseman, E.A., Molitor, J. & Marsit, C.J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30**, 1431-1439 (2014).
17. Jaffe, A.E. & Irizarry, R.A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* **15**, R31 (2014).
18. Koestler, D.C., *et al.* Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics* **8**, 816-826 (2013).
19. Li, B. & Li, J.Z. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol* **15**, 473 (2014).
20. Li, B., *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol* **17**, 174 (2016).
21. Newman, A.M., *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015).
22. Teschendorff, A.E., Breeze, C.E., Zheng, S.C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18**, 105 (2017).
23. Gentles, A.J., *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat Med* **21**, 938-945 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

24. Chakravarthy, A., *et al.* Human Papillomavirus Drives Tumor Development Throughout the Head and Neck: Improved Prognosis Is Associated With an Immune Response Largely Restricted to the Oropharynx. *J Clin Oncol* **34**, 4132-4141 (2016).
25. Ottensmeier, C.H., *et al.* Upregulated Glucose Metabolism Correlates Inversely with CD8+ T-cell Infiltration and Survival in Squamous Cell Carcinoma. *Cancer Res* **76**, 4136-4148 (2016).
26. Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48-61 (2015).
27. Network, T.C.G.A.R. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576-582 (2015).
28. Carter, S.L., *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**, 413-421 (2012).
29. Aran, D., Sirota, M. & Butte, A.J. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**, 8971 (2015).
30. Yoshihara, K., *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications* **4**, 2612 (2013).
31. Wood, O., *et al.* Gene expression analysis of TIL rich HPV-driven head and neck tumors reveals a distinct B-cell signature when compared to HPV independent tumors. *Oncotarget* (2016).
32. Fischer, C.A., *et al.* Is the improved prognosis of p16 positive oropharyngeal squamous cell carcinoma dependent of the treatment modality? *International Journal of Cancer* **126**, 1256-1262 (2010).
33. Hong, A.M., *et al.* Human papillomavirus predicts outcome in oropharyngeal cancer in patients treated primarily with surgery or radiation therapy. *Br J Cancer* **103**, 1510-1517 (2010).
34. Ward, M.J., *et al.* Staging and treatment of oropharyngeal cancer in the human papillomavirus era. *Head & Neck* **37**, 1002-1013 (2015).
35. deLeeuw, R.J., Kost, S.E., Kakal, J.A. & Nelson, B.H. The Prognostic Value of FoxP3+ Tumor-Infiltrating Lymphocytes in Cancer: A Critical Review of the Literature. *Clinical Cancer Research* **18**, 3022 (2012).
36. Sato, E., *et al.* Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 18538-18543 (2005).
37. Suzuki, H., *et al.* Intratumoral CD8+ T/FOXP3+ cell ratio is a predictive marker for survival in patients with colorectal cancer. *Cancer Immunology, Immunotherapy* **59**, 653-661 (2010).
38. Patel, S.P. & Kurzrock, R. PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. *Molecular Cancer Therapeutics* **14**, 847 (2015).
39. Aspeslagh, S., *et al.* Rationale for anti-OX40 cancer immunotherapy. *European Journal of Cancer* **52**, 50-66 (2016).
40. Croft, M., So, T., Duan, W. & Soroosh, P. The significance of OX40 and OX40L to T-cell biology and immune disease. *Immunol Rev* **229**, 173-191 (2009).
41. Lawrence, M.S., *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
42. Li, C., Egloff, A.M., Sen, M., Grandis, J.R. & Johnson, D.E. Caspase-8 mutations in head and neck cancer confer resistance to death receptor-mediated apoptosis and enhance migration, invasion, and tumor growth. *Mol Oncol* **8**, 1220-1230 (2014).
43. Aguirre, A., Shoji, K.F., Saez, J.C., Henriquez, M. & Quest, A.F. FasL-triggered death of Jurkat cells requires caspase 8-induced, ATP-dependent cross-talk between Fas and the purinergic receptor P2X(7). *Journal of cellular physiology* **228**, 485-493 (2013).
44. Brown, S.D., *et al.* Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Research* **24**, 743-750 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

45. Matsushita, H., *et al.* Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* **482**, 400-404 (2012).
46. Iorio, F., *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754 (2016).
47. Moran, S., *et al.* Validation of DNA methylation profiling in formalin-fixed paraffin-embedded samples using the Infinium HumanMethylation450 Microarray. *Epigenetics : official journal of the DNA Methylation Society* **9**, 829-833 (2014).
48. Dumenil, T.D., *et al.* Genome-wide DNA methylation analysis of formalin-fixed paraffin embedded colorectal cancer tissue. *Genes, chromosomes & cancer* **53**, 537-548 (2014).
49. Thirlwell, C., *et al.* Genome-wide DNA methylation analysis of archival formalin-fixed paraffin-embedded tissue using the Illumina Infinium HumanMethylation27 BeadChip. *Methods (San Diego, Calif.)* **52**, 248-254 (2010).
50. Siegel, E.M., *et al.* Expanding epigenomics to archived FFPE tissues: an evaluation of DNA repair methodologies. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **23**, 2622-2631 (2014).
51. Li, B., *et al.* Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat Genet* **48**, 725-732 (2016).
52. Fridman, W.H., Pages, F., Sautes-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nat Rev Cancer* **12**, 298-306 (2012).
53. Elliott, L.A., Doherty, G.A., Sheahan, K. & Ryan, E.J. Human Tumor-Infiltrating Myeloid Cells: Phenotypic and Functional Diversity. *Frontiers in Immunology* **8**, 86 (2017).
54. Chen, P.-L., *et al.* Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer Discovery* (2016).
55. Roh, W., *et al.* Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Science Translational Medicine* **9**(2017).
56. Riaz, N., *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934-949.e915 (2017).
57. Charoentong, P., *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep* **18**, 248-262 (2017).
58. Ruedl, C., Bachmann, M.F. & Kopf, M. The antigen dose determines T helper subset development by regulation of CD40 ligand. *European Journal of Immunology* **30**, 2056-2064 (2000).
59. Germano, G., *et al.* Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. *Nature* **552**, 116-120 (2017).
60. Martincorena, I., *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e1021 (2017).
61. Patel, S.J., *et al.* Identification of essential genes for cancer immunotherapy. *Nature* **548**, 537-542 (2017).
62. Shin, D.S., *et al.* Primary Resistance to PD-1 Blockade Mediated by JAK1/2 Mutations. *Cancer discovery* **7**, 188-201 (2017).
63. Gao, J., *et al.* Loss of IFN-gamma Pathway Genes in Tumor Cells as a Mechanism of Resistance to Anti-CTLA-4 Therapy. *Cell* **167**, 397-404.e399 (2016).
64. Zaretsky, J.M., *et al.* Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. *The New England journal of medicine* **375**, 819-829 (2016).
65. Pan, D. & Kobayashi, A. A major chromatin regulator determines resistance of tumor cells to T cell-mediated killing. (2018).

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

66. Ishak, C.A., *et al.* An RB-EZH2 Complex Mediates Silencing of Repetitive DNA Sequences. *Molecular cell* **64**, 1074-1087 (2016).
67. Leonova, K.I., *et al.* p53 cooperates with DNA methylation and a suicidal interferon response to maintain epigenetic silencing of repeats and noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E89-98 (2013).
68. Moroishi, T., *et al.* The Hippo Pathway Kinases LATS1/2 Suppress Cancer Immunity. *Cell* **167**, 1525-1539.e1517 (2016).
69. Sadic, D., *et al.* Atrx promotes heterochromatin formation at retrotransposons. *EMBO reports* **16**, 836-850 (2015).
70. Chen, K., *et al.* Methyltransferase SETD2-Mediated Methylation of STAT1 Is Critical for Interferon Antiviral Activity. *Cell* **170**, 492-506.e414 (2017).
71. Topper, M.J., *et al.* Epigenetic Therapy Ties MYC Depletion to Reversing Immune Evasion and Treating Lung Cancer. *Cell* **171**, 1284-1300.e1221 (2017).
72. George, S., *et al.* Loss of PTEN Is Associated with Resistance to Anti-PD-1 Checkpoint Blockade Therapy in Metastatic Uterine Leiomyosarcoma. *Immunity* **46**, 197-204 (2017).
73. Peng, W., *et al.* Loss of PTEN Promotes Resistance to T Cell–Mediated Immunotherapy. *Cancer Discovery* **6**, 202 (2016).
74. Lim, S.O., *et al.* EGFR Signaling Enhances Aerobic Glycolysis in Triple-Negative Breast Cancer Cells to Promote Tumor Growth and Immune Escape. *Cancer Res* **76**, 1284-1296 (2016).
75. Carter, S.L., *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotech* **30**, 413-421 (2012).
76. Yoshihara, K., *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**, 2612 (2013).
77. Van Allen, E.M., *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207-211 (2015).
78. Balermpas, P., *et al.* Tumour-infiltrating lymphocytes predict response to definitive chemoradiotherapy in head and neck cancer. *Br J Cancer* **110**, 501-509 (2014).
79. Keck, M.K., *et al.* Integrative Analysis of Head and Neck Cancer Identifies Two Biologically Distinct HPV and Three Non-HPV Subtypes. *Clinical Cancer Research* **21**, 870-881 (2015).
80. Nguyen, N., *et al.* Tumor infiltrating lymphocytes and survival in patients with head and neck squamous cell carcinoma. *Head & Neck* **38**, 1074-1084 (2016).
81. The Cancer Genome Atlas Research, N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576-582 (2015).
82. Ottensmeier, C.H., *et al.* Upregulated Glucose Metabolism Correlates Inversely with CD8+ T-cell Infiltration and Survival in Squamous Cell Carcinoma. *Cancer research* (2016).
83. Su, S., *et al.* Blocking the recruitment of naive CD4+ T cells reverses immunosuppression in breast cancer. *Cell Res* (2017).
84. Hanley, C.J., *et al.* Targeting the Myofibroblastic Cancer-Associated Fibroblast Phenotype Through Inhibition of NOX4. *Journal of the National Cancer Institute* **110**(2018).
85. Reinius, L.E., *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One* **7**, e41361 (2012).
86. Zhang, Y., *et al.* Genome-wide DNA methylation analysis identifies hypomethylated genes regulated by FOXP3 in human regulatory T cells. *Blood* **122**, 2823-2836 (2013).
87. Lechner, M., *et al.* Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. *Genome medicine* **5**, 15 (2013).
88. Aryee, M.J., *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369 (2014).
89. Fortin, J.P., *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome biology* **15**, 503 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

888 90. Teschendorff, A.E., *et al.* A beta-mixture quantile normalization method for correcting probe
889 design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189-196
890 (2013).
891 91. Chen, Y., *et al.* VirusSeq: software to identify viruses and their integration sites using next-
892 generation sequencing of human cancer tissue. *Bioinformatics* **29**, 266-267 (2013).
893 92. Charoentong, P., *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-
894 Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell*
895 *Reports* **18**, 248-262 (2017).
896 93. Mermel, C.H., *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of
897 focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
898 94. De Simone, M., *et al.* Transcriptional Landscape of Human Tissue Lymphocytes Unveils
899 Uniqueness of Tumor-Infiltrating T Regulatory Cells. *Immunity* **45**, 1135-1147.
900 95. Linsley, P.S., Speake, C., Whalen, E. & Chaussabel, D. Copy number loss of the interferon
901 gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis.
902 *PLoS One* **9**, e109760 (2014).
903 96. Bonnal, R.J.P., *et al.* De novo transcriptome profiling of highly purified human lymphocytes
904 primary cells. *Scientific Data* **2**, 150051 (2015).
905 97. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
906 quantification. *Nature biotechnology* **34**, 525-527 (2016).
907 98. Andor, N., Harness, J.V., Müller, S., Mewes, H.W. & Petritsch, C. EXPANDS: expanding ploidy
908 and allele frequency on nested subpopulations. *Bioinformatics* **30**, 50-60 (2014).

909

910 **Figure Legends**

911

912 **Figure 1. Validation of DNA methylation-based deconvolution for the analysis of tumour**
913 **composition (a)** Correlation between MethylCIBERSORT fractions and flow cytometry for PBMC
914 mixtures in independent data. **(b-d)** Boxplots showing comparisons between MethylCIBERSORT and
915 flow cytometry versus Expression-CIBERSORT and flow cytometry in mixtures of similar complexity
916 for correlations by cell type, correlations within samples, and finally absolute error. **(e)** Correlations
917 between ABSOLUTE and MethylCIBERSORT versus other previously published purity estimation
918 methods. **(f)** Validation of previously reported associations between CD8 T-cells and B-cells and HPV
919 status by HPV status. **(g)** Correlation plot showing Spearman's Rho between cell-types in HPV-
920 HNSCC, red boxes indicate nonsignificance at $q < 0.1$. **(h)** IHC showing representative image of CD8
921 and SMA and Kaplan-Meier curves confirming the prognostic impact of TILs and fibroblasts in HPV-
922 negative HNSCC.

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2: Classification of HNSCC into hot and cold tumour subgroups on the basis of immune cell infiltration patterns. (a) Boxplot of cell-types based on clustering HNSCC. **(b)** Kaplan-Meier curves of overall survival by HNSCC immune cluster. HR and p.values from multivariate Cox regression controlling for age and stage. **(c)** Bar-graph showing associations between cytolytic activity and cell types. **(d)** Cytolytic activity is elevated in immune-hot HNSCC. **(e)** CD8/Treg ratios by HNSCC Immune Cluster. **(f)** Mutations significantly associated with HNSCC immune cluster.

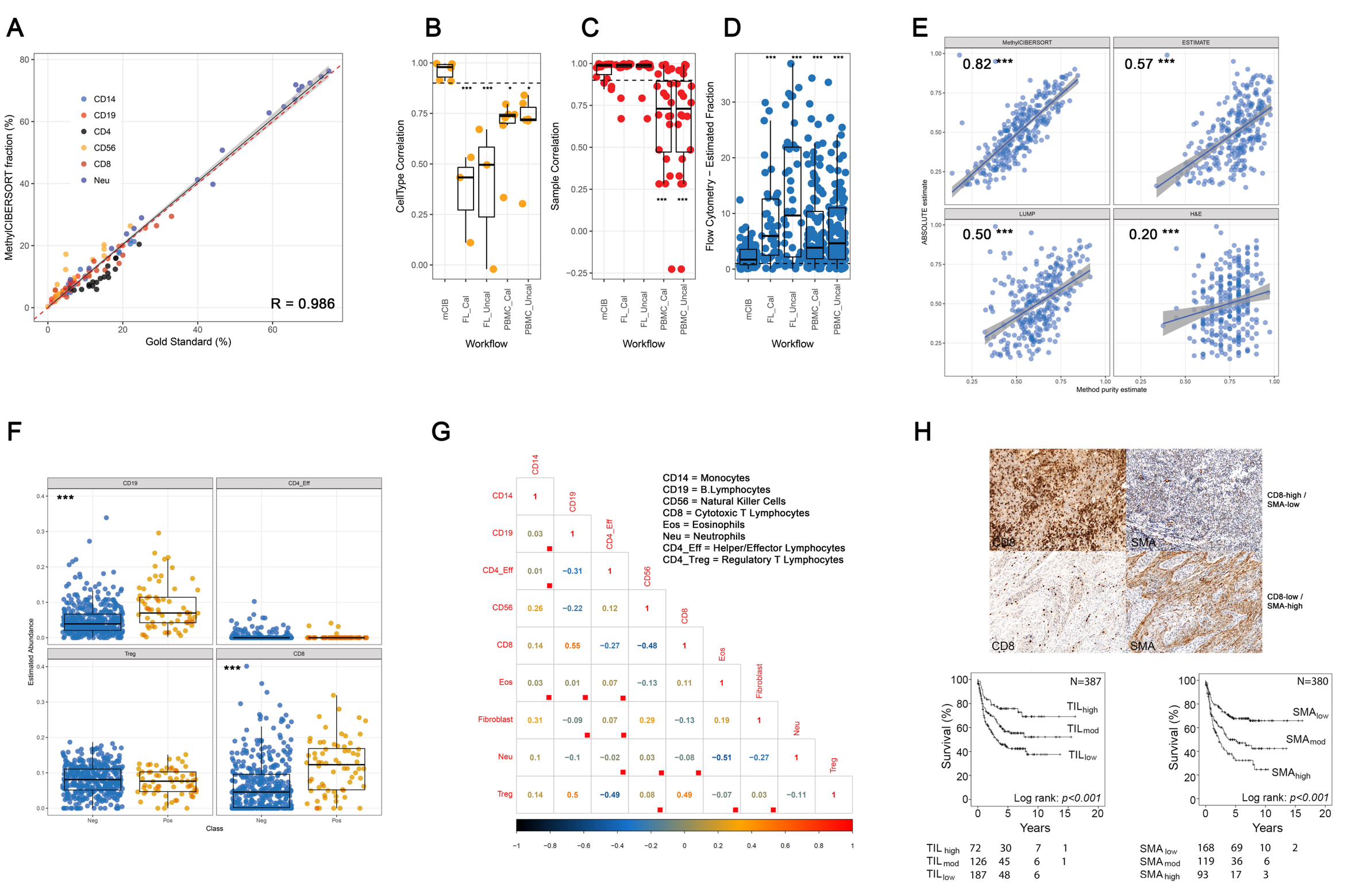
Figure 3. Identification and characterization of hot and cold tumours pan-cancer. (a) Barplot of distribution of Immune-hot and cold tumours across TCGA. Cancers known to respond favourably to checkpoint blockade, such as lung cancer and melanomas, show high fractions of hot tumours. **(b)** Boxplot of cell-type estimates by immune cluster. All at $q < 0.05$. Numbers represent mean fold changes. **(c)** CD8:Treg ratio is elevated in hot tumours pan-cancer. **(d)** Increased breadth of TCR sequences in Immune Hot tumours. **(e)** Results of IPA canonical pathway analysis comparing hot and cold tumours pan-cancer after adjusting for tumour type. **(f)** Transcriptional deconvolution by Expression-based CIBERSORT shows immune cluster is associated with distinct CD4 polarisation and **(g)** macrophage polarisation.

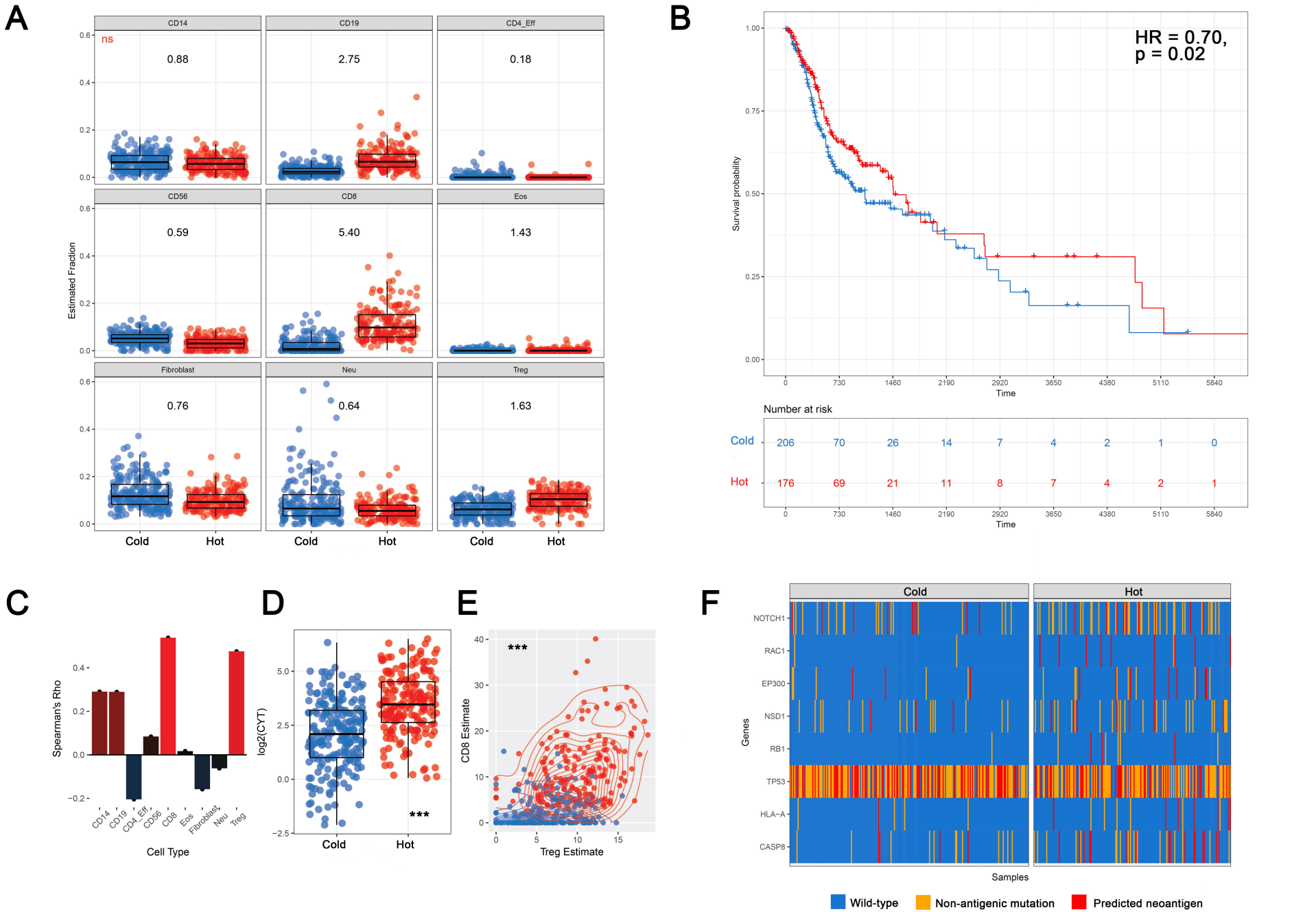
Figure 4. The immune-hot signature is associated with response to immune checkpoint blockade in melanoma. (a) heatmap showing expression of the hot-tumour transcriptional signature in Nanostring data from posttreatment biopsies of immunotherapy patients. **(b)** heatmaps showing the same signature in RNAseq data of aCTLA4 (pre-treatment) and aPD1 (post-treatment) respectively. **(c)** boxplots highlighting significant differences in ssGSEA scores for the hot-tumour transcriptional signature in the datasets featured in (b). **(d)** barplots display similarity to TCGA hot and cold tumours

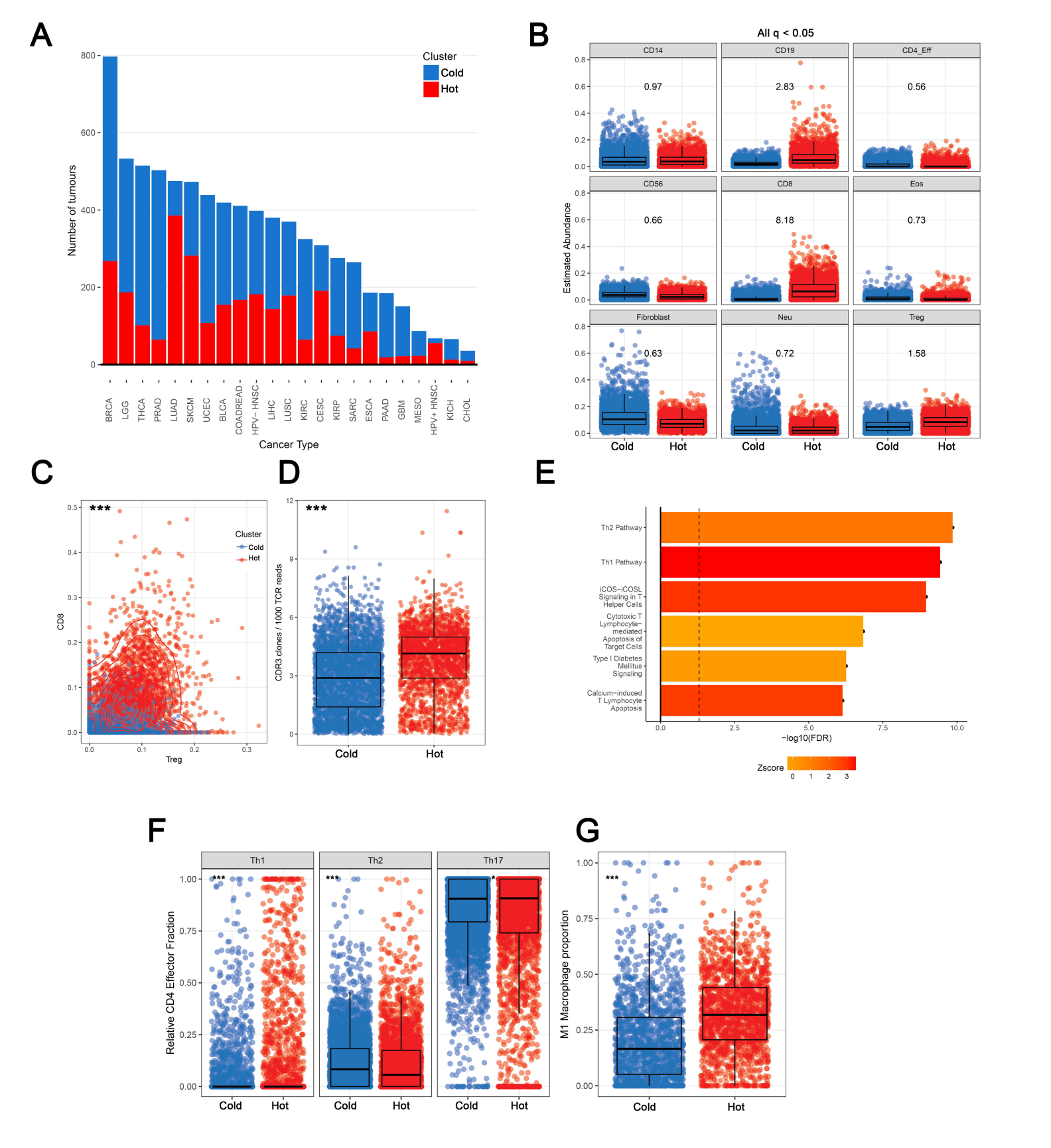
bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

based on logistic regression class probabilities from a model fit to TCGA data, which are associated with response. **(e)** boxplots showing Kappa values from cross-validation for models examining the performance of the Immune-hot signature, Class I neoepitope burden, and finally mutational load on immunotherapy response classification.

Figure 5. Genomic features of hot and cold tumours. (a) Density plots showing differences in neoantigen burden by immune cluster pan-cancer. P.value from negative binomial regression that accounts for tumour type. **(b)** Clonal neoantigens and subclonal neoantigens are correlated with different infiltration profiles. Volcanoplot shows Spearman's Rho on the X-axis and –log10(FDR) on the y-axis. **(c)** Volcanoplot showing results of binomial regression testing for associations between Immune-hot cancers and mutation frequencies in candidate cancer driver genes. Those genes implicated in resistance to T-cell mediated destruction are highlighted in orange. **(d)** Volcanoplot showing associations between GISTIC candidate driver copy number peaks and immune cluster. **(e)** Plot showing results of logistic regression in a cohort of HNSCCs where the probability of being classified TIL-high was regressed against anatomic subsite, EGFR IHC (low/moderate/high) and HPV status. **(f)** Correlation between glycolytic coexpression signature ssGSEA scores and EGFR levels by RPPA. **(g)** Association of glycolytic signature post-Nivolumab with response and **(h)** Inverse correlation between the glycolytic signature and the immune-hot expression signature, Spearman's correlation has been plotted.

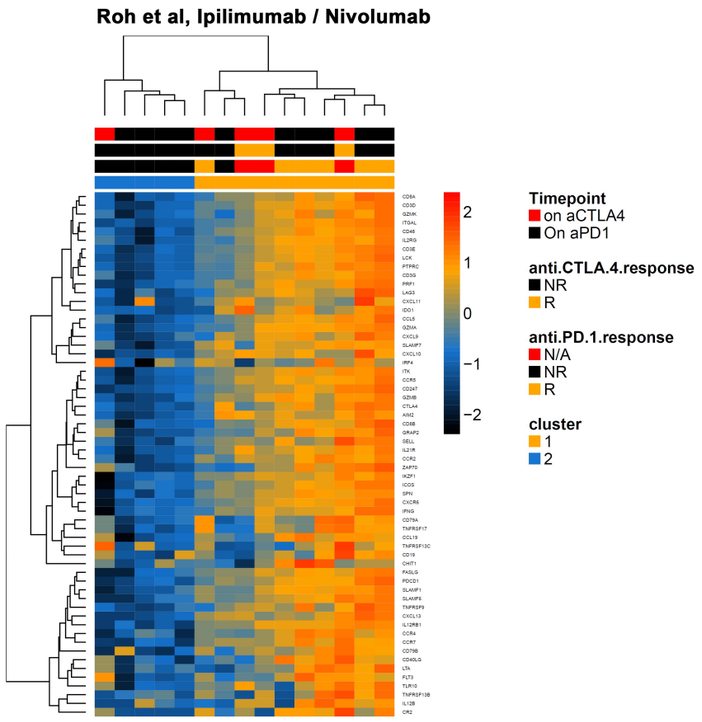

A
MethylCIBERSORT fraction (%)
Gold Standard (%)
CD14
CD19
CD4
CD56
CD8
Neu
R = 0.986
B
CellType Correlation
Workflow
mCIB
FL_Cal
FL_Uncal
PBMC_Cal
PBMC_Uncal
C
Sample Correlation
Workflow
D
Flow Cytometry − Estimated Fraction
Workflow
E
MethylCIBERSORT
ESTIMATE
LUMP
H&E
0.82 ***
0.57 ***
0.50 ***
0.20 ***
ABSOLUTE estimate
Method purity estimate
F
CD19
CD4_Eff
Treg
CD8
Estimated Abundance
Neg
Pos
Class
G
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
H
CD8-high / SMA-low
CD8-low / SMA-high
CD8
SMA
Survival (%)
Years
N=387
TIL high
TIL mod
TIL low
Log rank: p<0.001
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
CD19
CD4_Eff
CD56
CD8
Eos
Fibroblast
Neu
Treg
ns
0.88
2.75
0.18
0.59
5.40
1.43
0.76
0.64
1.63
Estimated Fraction
Cold
Hot
B
HR = 0.70,
p = 0.02
Survival probability
Time
Number at risk
Cold 206 70 26 14 7 4 2 1 0
Hot 176 69 21 11 8 7 4 2 1
C
Spearman's Rho
Cell Type
D
log2(CYT)
***
E
CD8 Estimate
Treg Estimate
F
Cold
Hot
NOTCH1
RAC1
EP300
NSD1
RB1
TP53
HLA-A
CASP8
Genes
Samples
Wild-type
Non-antigenic mutation
Predicted neoantigen

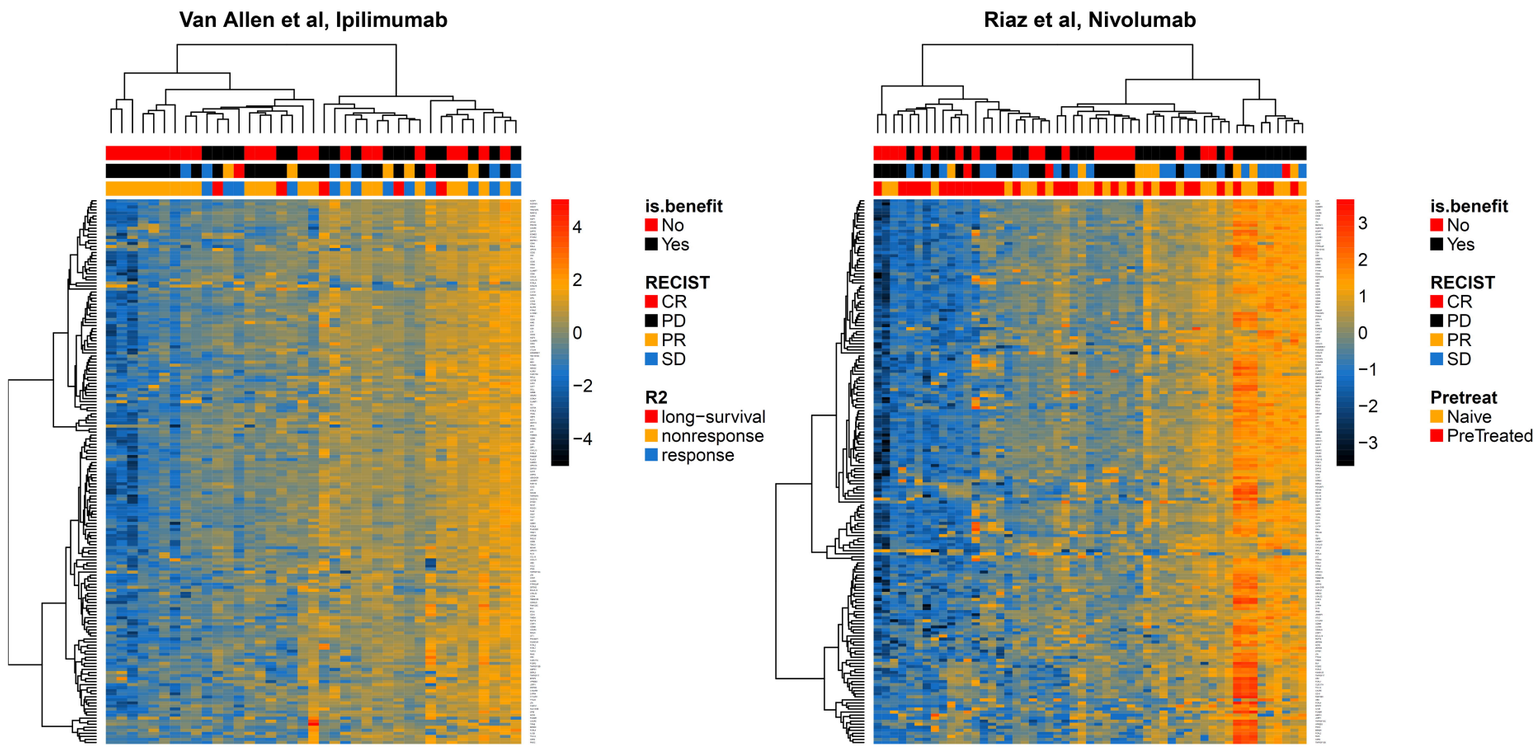

A
Cluster
Cold
Hot
Number of tumours
800
600
400
200
0
BRCA
LGG
THCA
PRAD
LUAD
SKCM
UCEC
BLCA
COADREAD
HPV− HNSC
LIHC
LUSC
KIRC
CESC
KIRP
SARC
ESCA
PAAD
GBM
MESO
HPV+ HNSC
KICH
CHOL
Cancer Type
B
All q < 0.05
CD14
CD19
CD4_Eff
CD56
CD8
Eos
Fibroblast
Neu
Treg
0.97
2.83
0.56
0.66
8.18
0.73
0.63
0.72
1.58
Estimated Abundance
Cold
Hot
C
***
Cluster
Cold
Hot
CD8
Treg
D
***
CDR3 clones / 1000 TCR reads
Cold
Hot
E
Th2 Pathway
Th1 Pathway
iCOS−iCOSL Signaling in T Helper Cells
Cytotoxic T Lymphocyte−mediated Apoptosis of Target Cells
Type I Diabetes Mellitus Signaling
Calcium−induced T Lymphocyte Apoptosis
−log10(FDR)
Zscore
F
Th1
Th2
Th17
***
*
Relative CD4 Effector Fraction
Cold
Hot
G
***
M1 Macrophage proportion
Cold
Hot

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

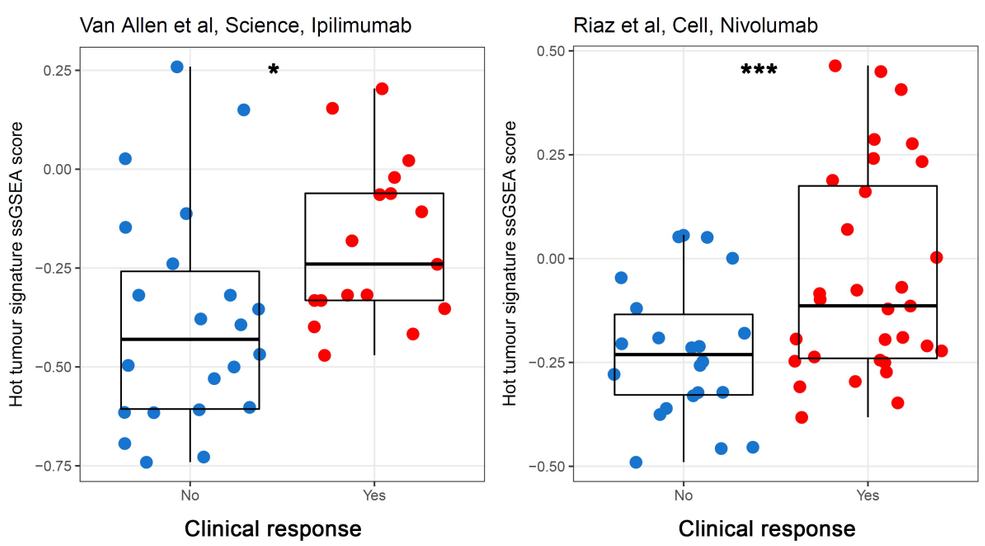
A



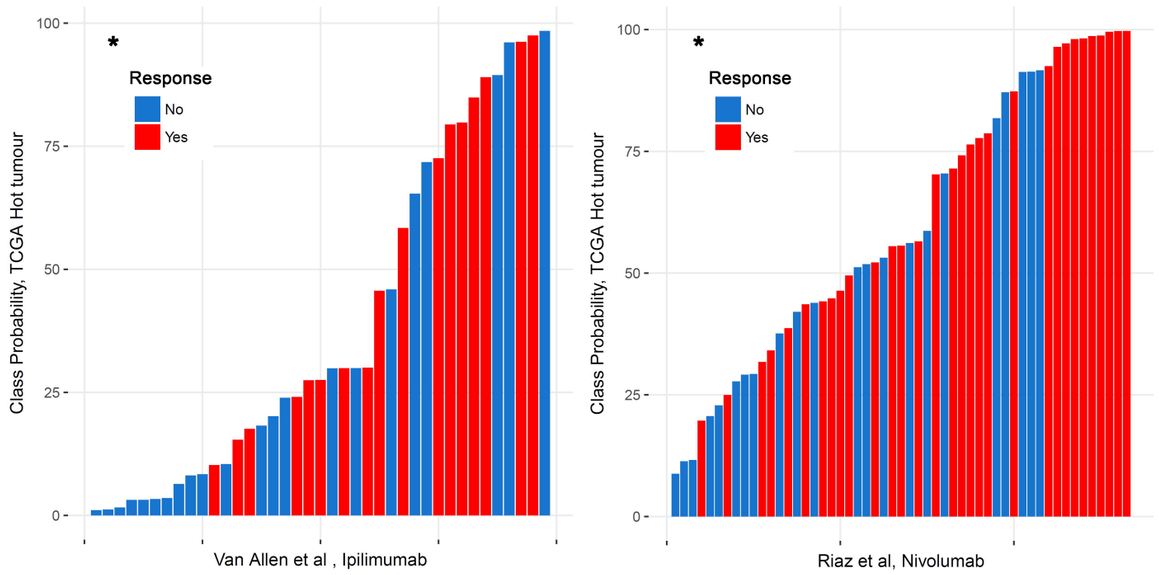
B



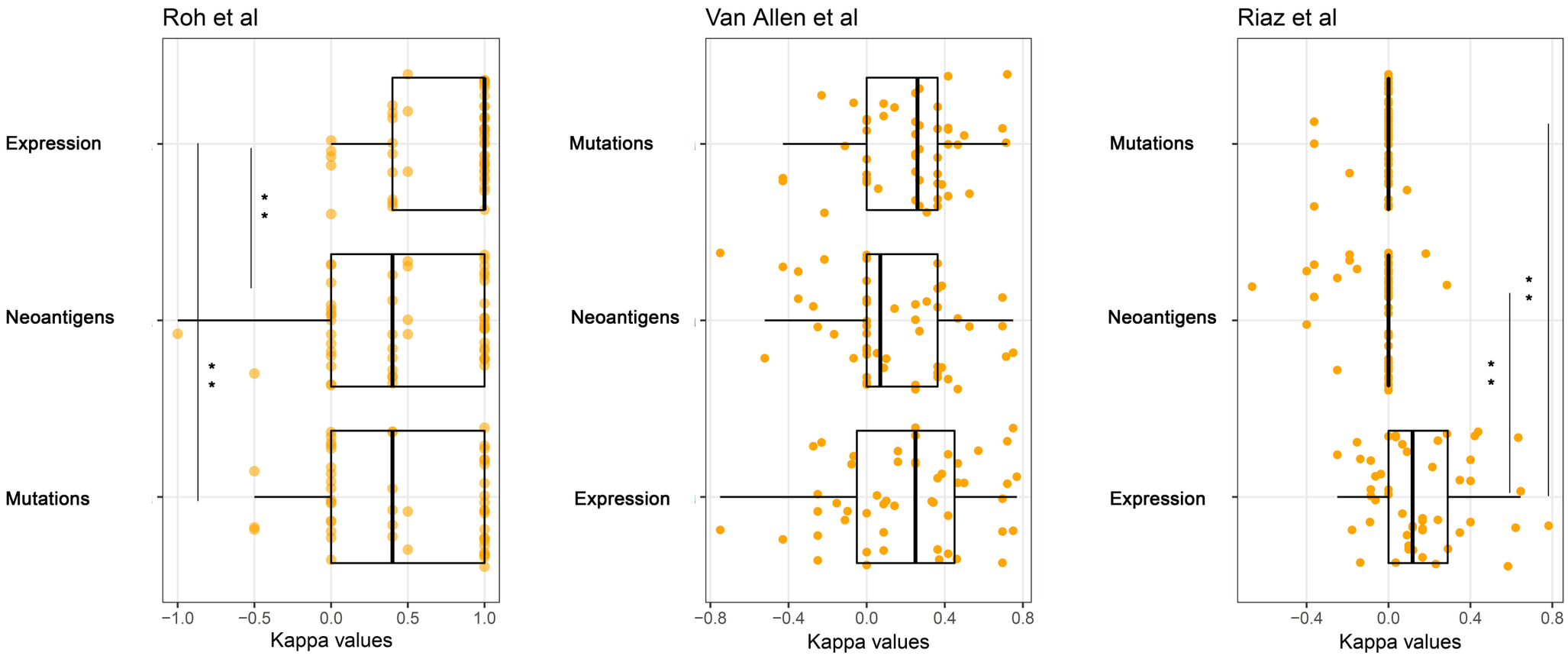
C



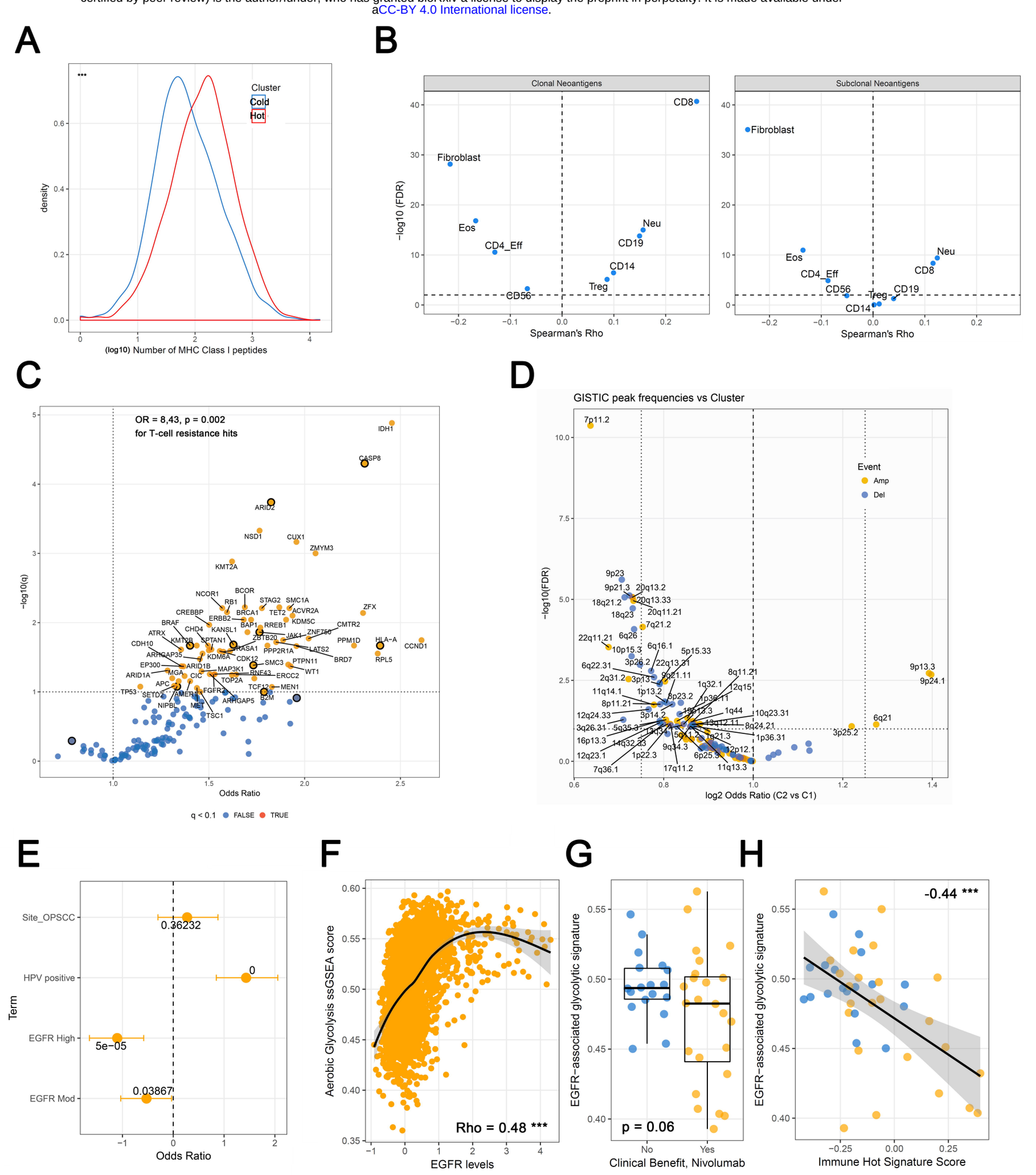
D



E



certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under
aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figures

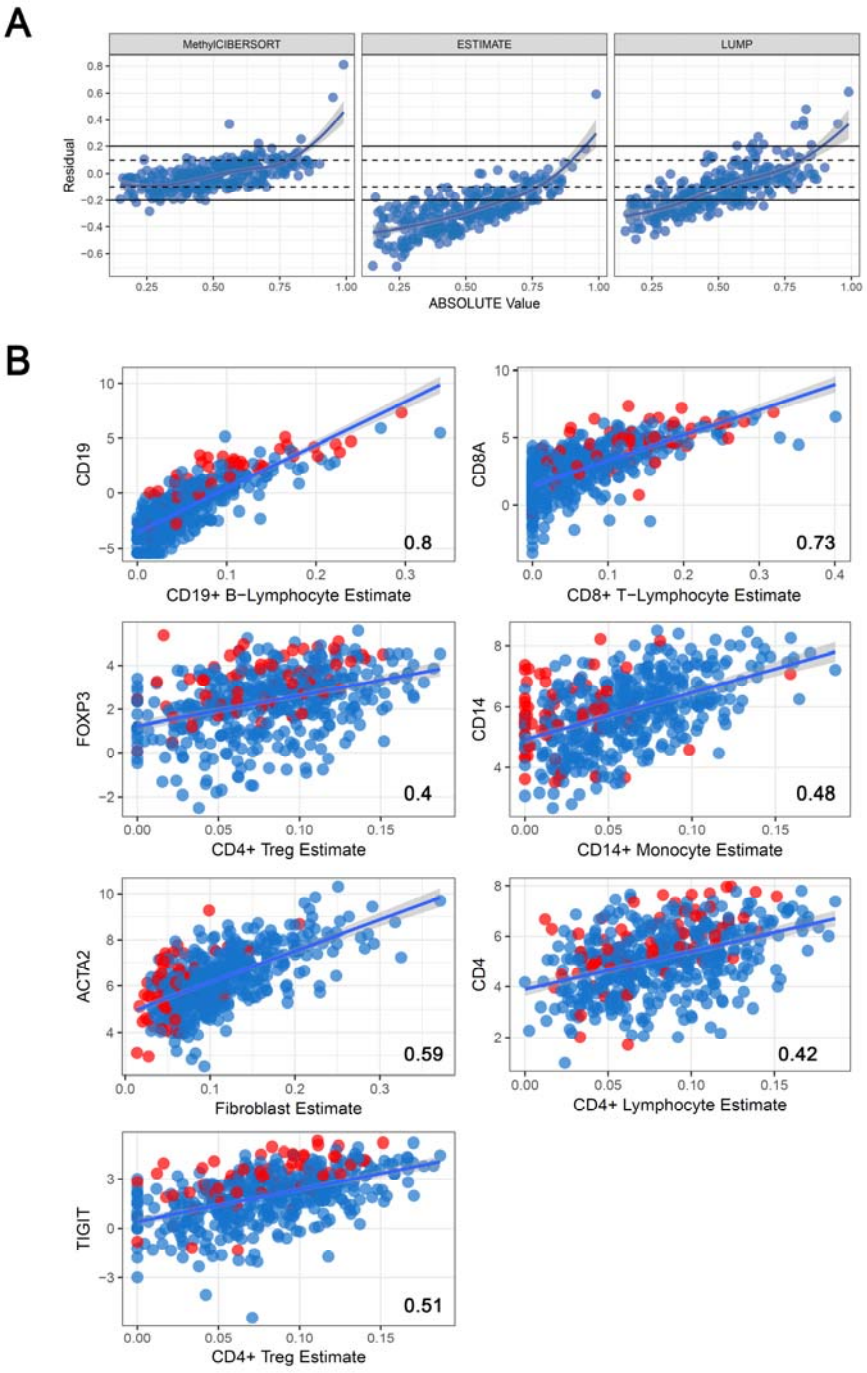


Figure S1: (A) Analysis of ABSOLUTE estimate (x-axis) and error from MethylCIBERSORT, ESTIMATE and LUMP in estimating purity in relation (y-axis). **(B)** Correlations (Spearman's Rho) between MethylCIBERSORT estimates and marker gene expression in TCGA HNSCC.

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

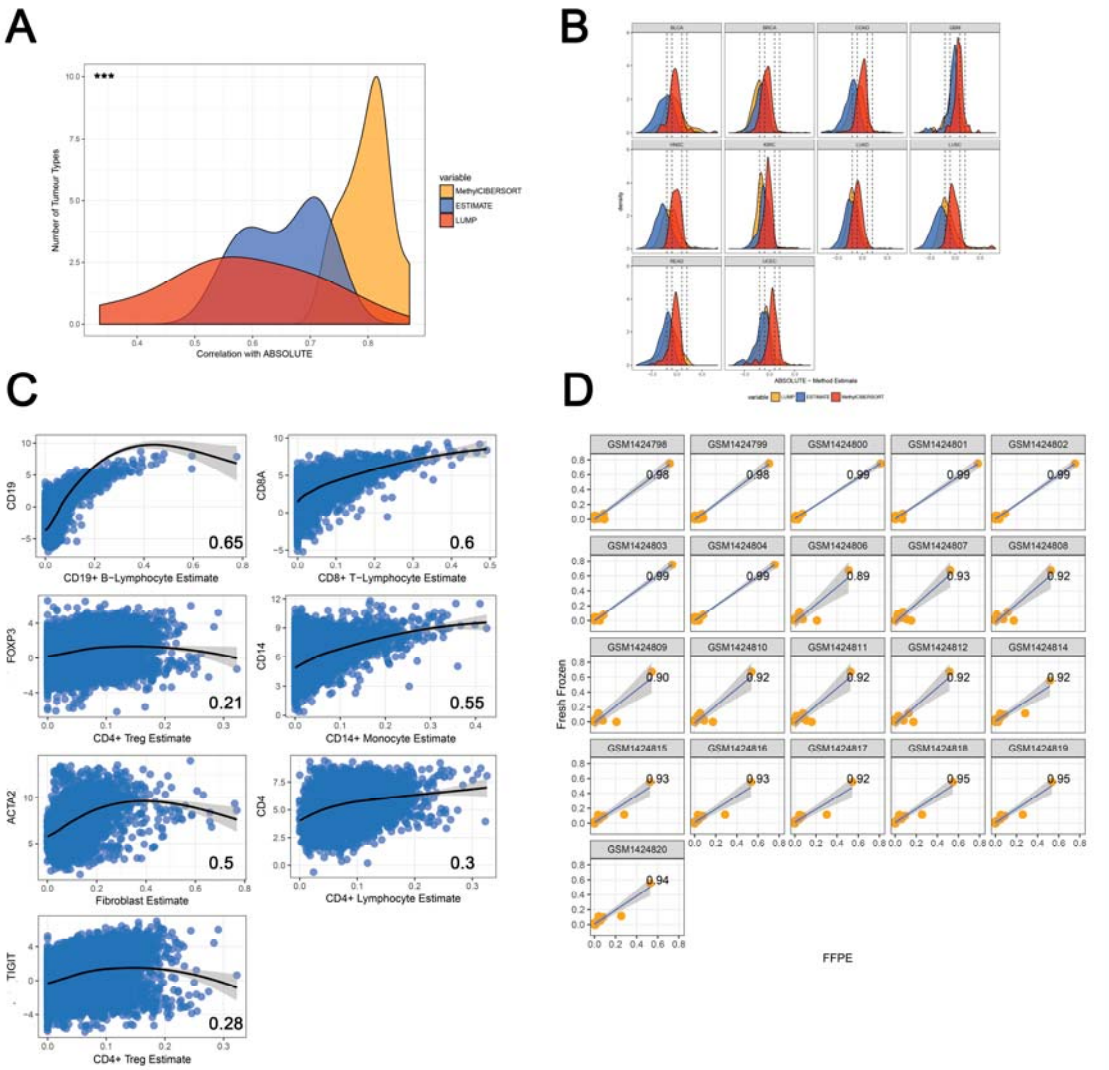
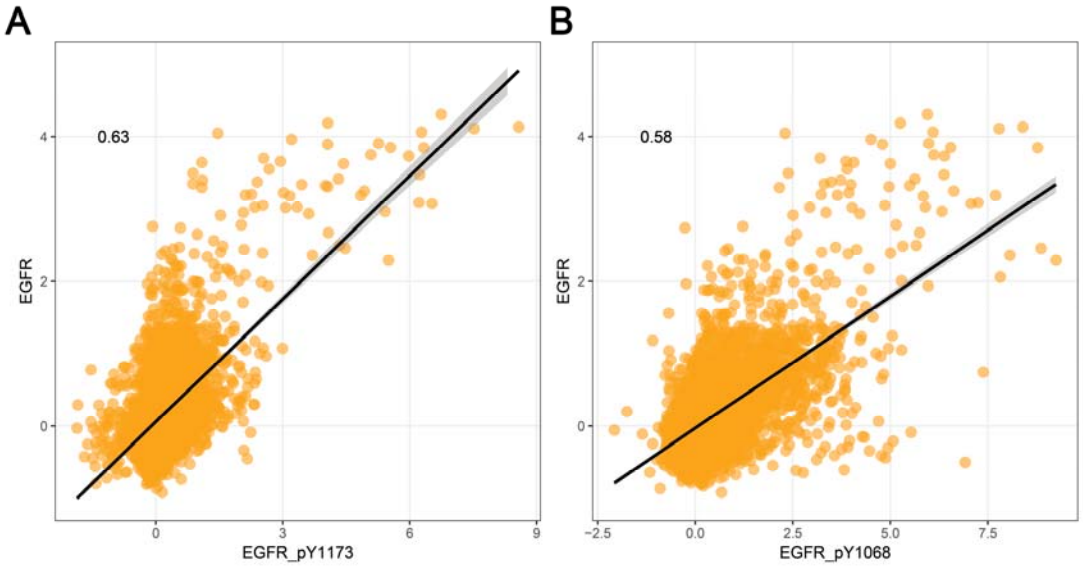


Figure S2: (A) Correlation densities for associations with ABSOLUTE purity for MethylCIBERSORT, ESTIMATE and LUMP across tumour types. **(B)** Density plots showing error relative to ABSOLUTE in individual tumour types for MethylCIBERSORT, ESTIMATE and LUMP. **(C)** Marker correlation plots between MethylCIBERSORT estimates and expression of marker genes. **(D)** Correlation plots for 21 450k methylomes relative from FFPE samples relative to their fresh frozen counterparts.

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

16



17

18 **Figure S3:** Scatterplots showing association between EGFR levels by RPPA and phosphorylation at

19 key activating residues.

20

bioRxiv preprint doi: https://doi.org/10.1101/281592; this version posted March 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

List of Supplementary Tables – Chakravarthy et al

Table S1 - Genes differentially expressed between HNSCC Hot and Cold clusters.

Tables S2, S3, S4 – IPA canonical pathway analysis, ontology analysis, and upstream regulator analysis for genes in Table S1, respectively.

Table S5 - Differentially bound antibodies from RPPA data for HNSCC Hot vs Cold cluster comparison

Table S6 – Canonical pathway analysis for pan-cancer hot vs cold transcriptional signature.

Table S7 – Results of association analysis between immune cluster and mutation frequency.

Table S8 – Results of association analysis between immune cluster and copy number alteration frequency.