Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/27/eaba1862/DC1

Supplementary Materials for

Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns

Marina Salvadores, Francisco Fuster-Tormo, Fran Supek*

*Corresponding author. Email: fran.supek@irbbarcelona.org

Published 1 July 2020, *Sci. Adv.* **6**, eaba1862 (2020)
DOI: 10.1126/sciadv.aba1862

The PDF file includes:

Figs. S1 to S9

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/27/eaba1862/DC1)

Table S1

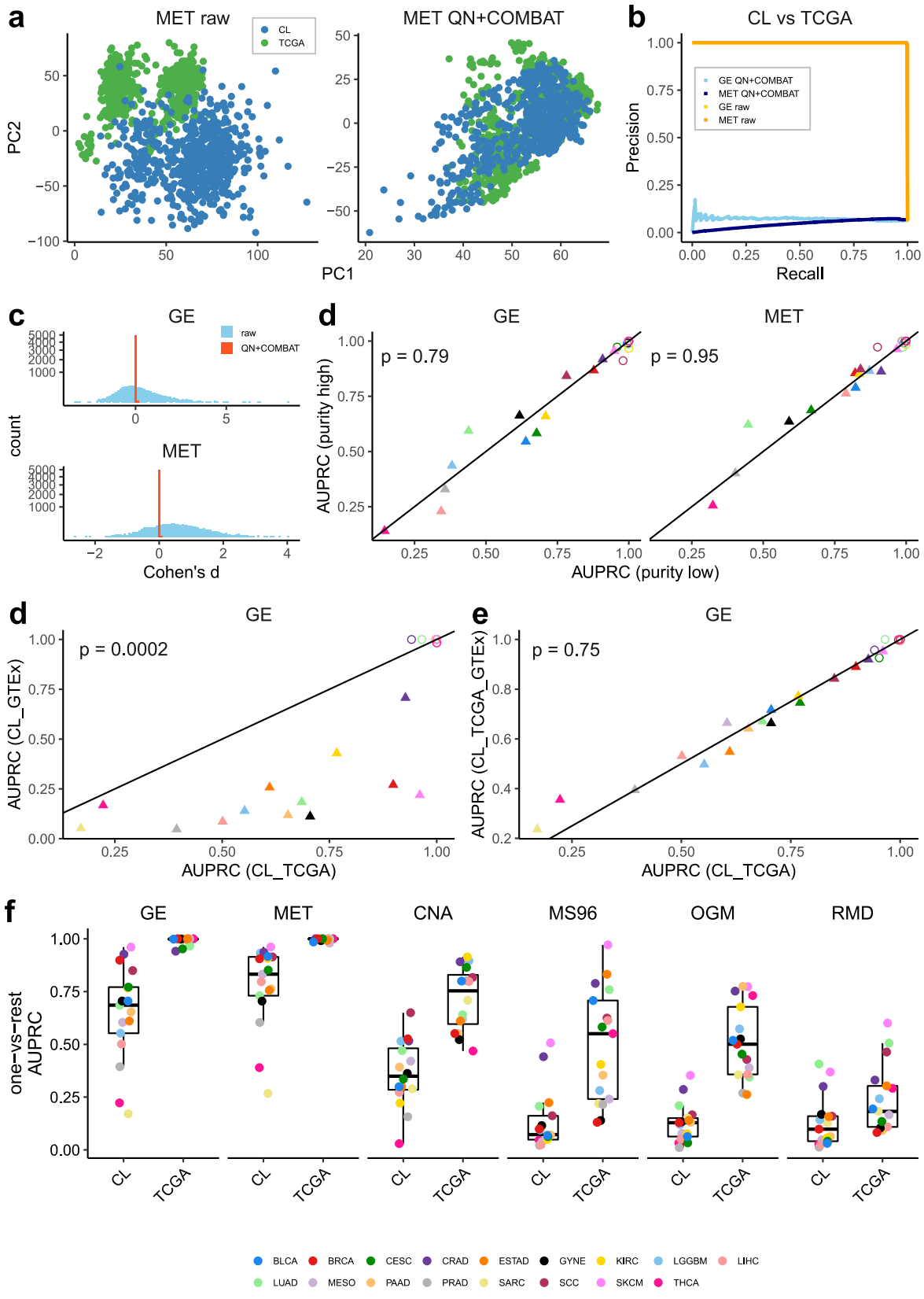

Fig S1. Quality control measures for adjustments of gene expression / DNA methylation data to remove tumor-cell line differences. ***(a)*** *Principal component (PC) 1 and PC2 of a PC analysis in the MET pre-adjustment data (raw) and in the post-adjustment data (QN+ComBat). Colors represent the batch effect labels (CL are the cell lines and TCGA are the human tumors).* ***(b)*** *PR curve for classifying tumors versus cell lines in the data pre-adjustment and post-adjustment for GE and MET.* ***(c)*** *Distribution of the cohen's d between CL and TCGA for gene expression/DNA methylation values before and after adjustment.* ***(d)*** *Comparison between the Area Under the Precision Recall Curve (AUPRC) obtained when testing in TCGA test set (circles) and cell lines (triangles) and training in high versus low purity samples from TCGA.* ***(e)*** *Comparison between the AUPRC yielded when testing in TCGA test set (circles) and cell lines (triangles) and training in healthy samples (GTEx) versus tumor samples (TCGA).* ***(f)*** *Comparison between the AUPRC yielded when testing in TCGA test set (circles) and cell lines (triangles) and training in data adjusted with the previous batch effects (GDSC/CLLE/TCGA) versus adjusted with batch effects including healthy data only (GDSC/CLLE /TCGA/GTEx).* ***(g)*** *Comparison between the AUPRC obtained when training on TCGA train set and testing on TCGA test set and on the cell lines dataset (CL) using 6 different types of features. P-values calculated for cell lines using a Wilcoxon-test (paired, two-tailed) comparing AUPRCs in x-axis versus AUPRCs in y-axis.*

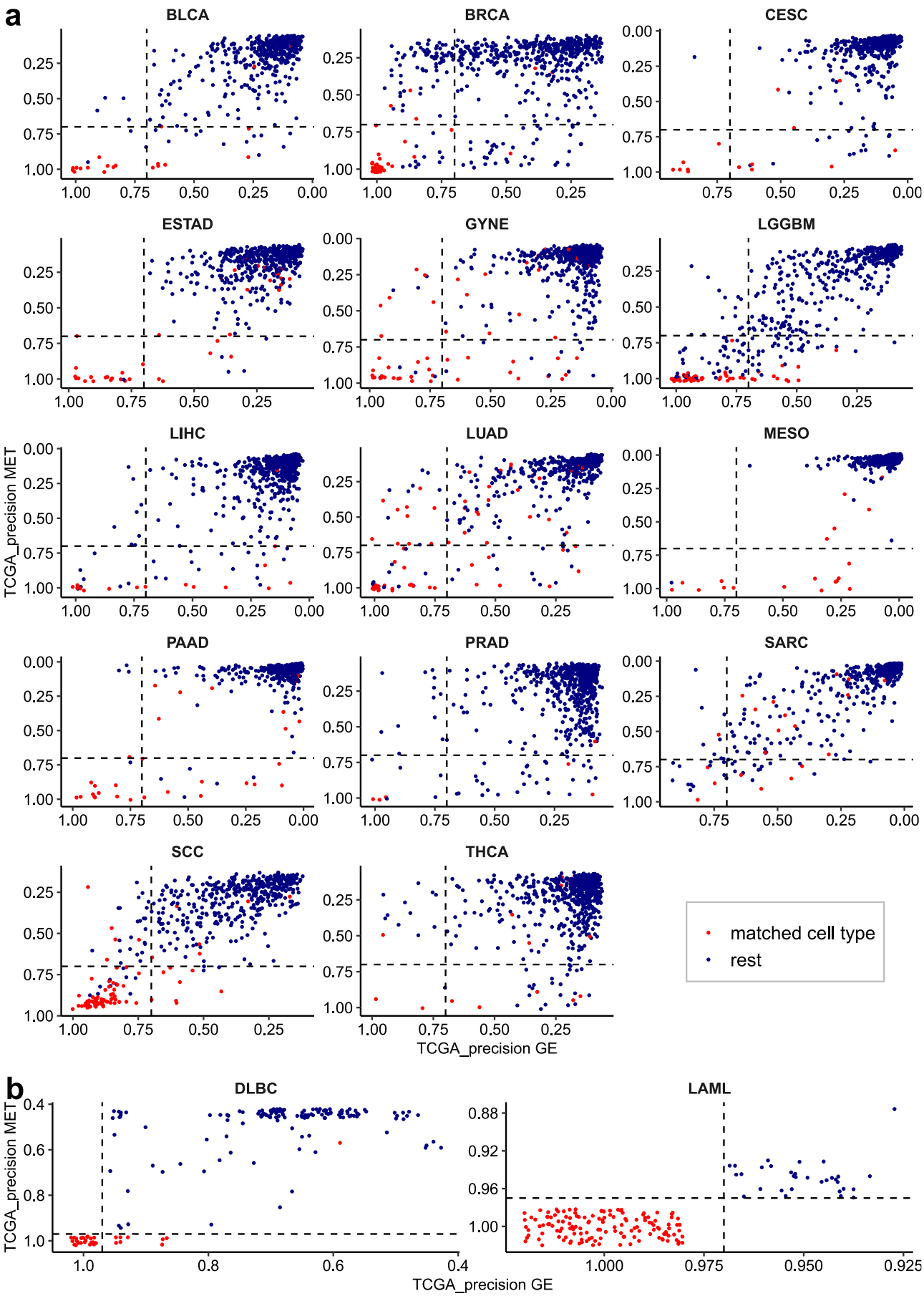

Fig S2. Reclassification of cancer cell lines in various cancer types according to classifiers based on TCGA tumor data*. TCGA-based precision scores (see Methods) for cell lines calculated in gene expression (GE, x-axis) and DNA methylation-based (MET) cancer type classifiers, using a one-vs-rest classification scheme. The higher the TCGA precision, the more likely it is that a cell line belongs to that particular cancer type (shown on top of each plot; the acronyms are from TCGA or described in Methods). The cell lines that were originally annotated as the cancer type that is being tested in each plot are shown in red in that plot, the rest of the cell lines are shown in blue. Solid tumor classifier shown in* ***(a)*** *and blood tumor classifiers in* ***(b).***

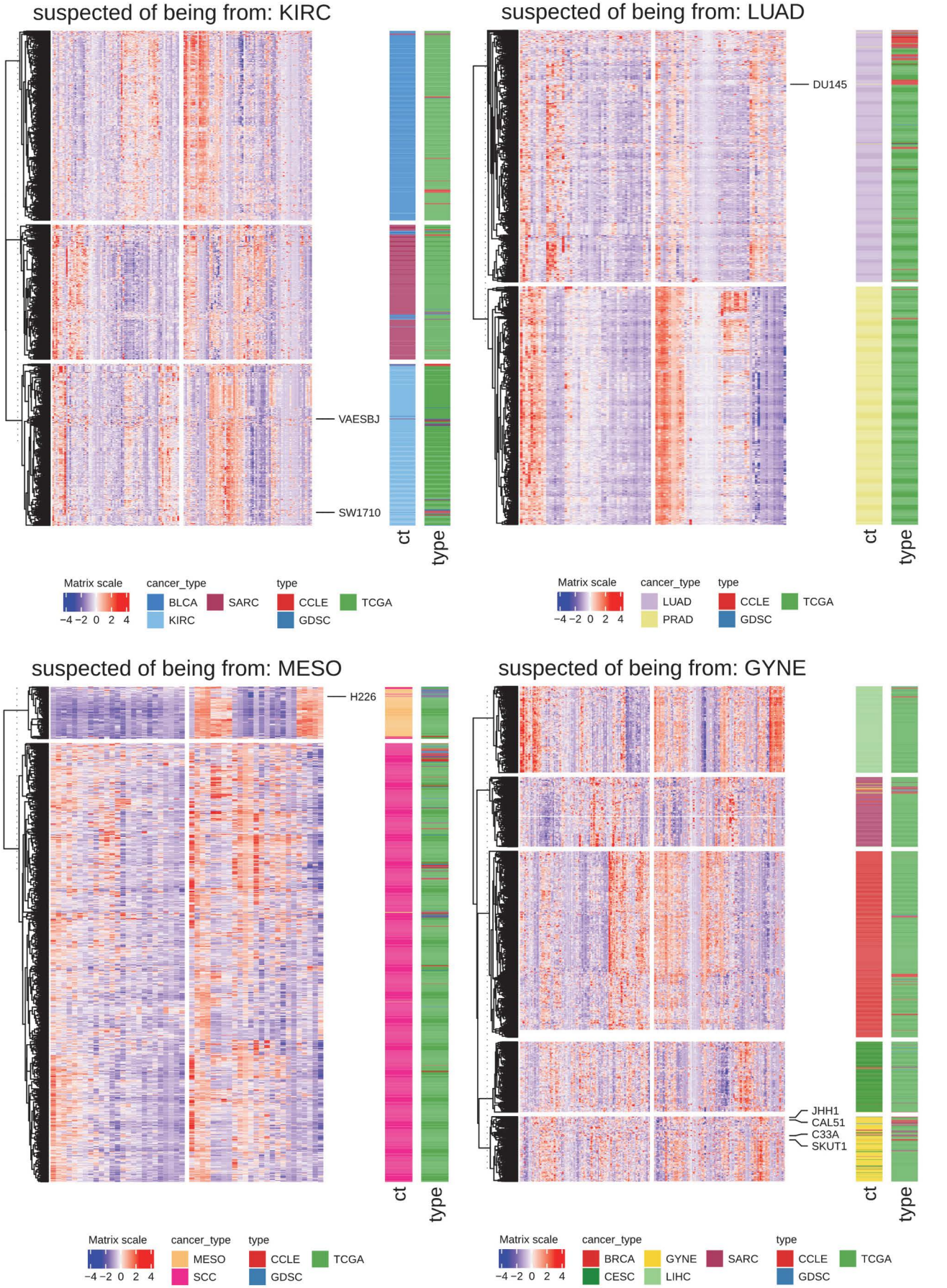

Fig S3. Visual inspection of cell lines mislabelled to a different cancer type. Heatmaps for the 25 genes (in GE) and probes (in MET) with highest absolute value of ridge regression coefficients for each of the cancer types in the plot (suspect cancer type for each case) versus rest classifiers. There is one heatmap per cancer type of those that that presented at least 1 cell line suspected of mislabeling (the suspected cancer type is written in the tittle). In each heatmap, the cell lines suspected to belong to that particular cancer type are labeled on the right-hand side of the plot. In each case, the cancer types included are the suspected cancer type, and the original cancer types of the suspect cell lines.

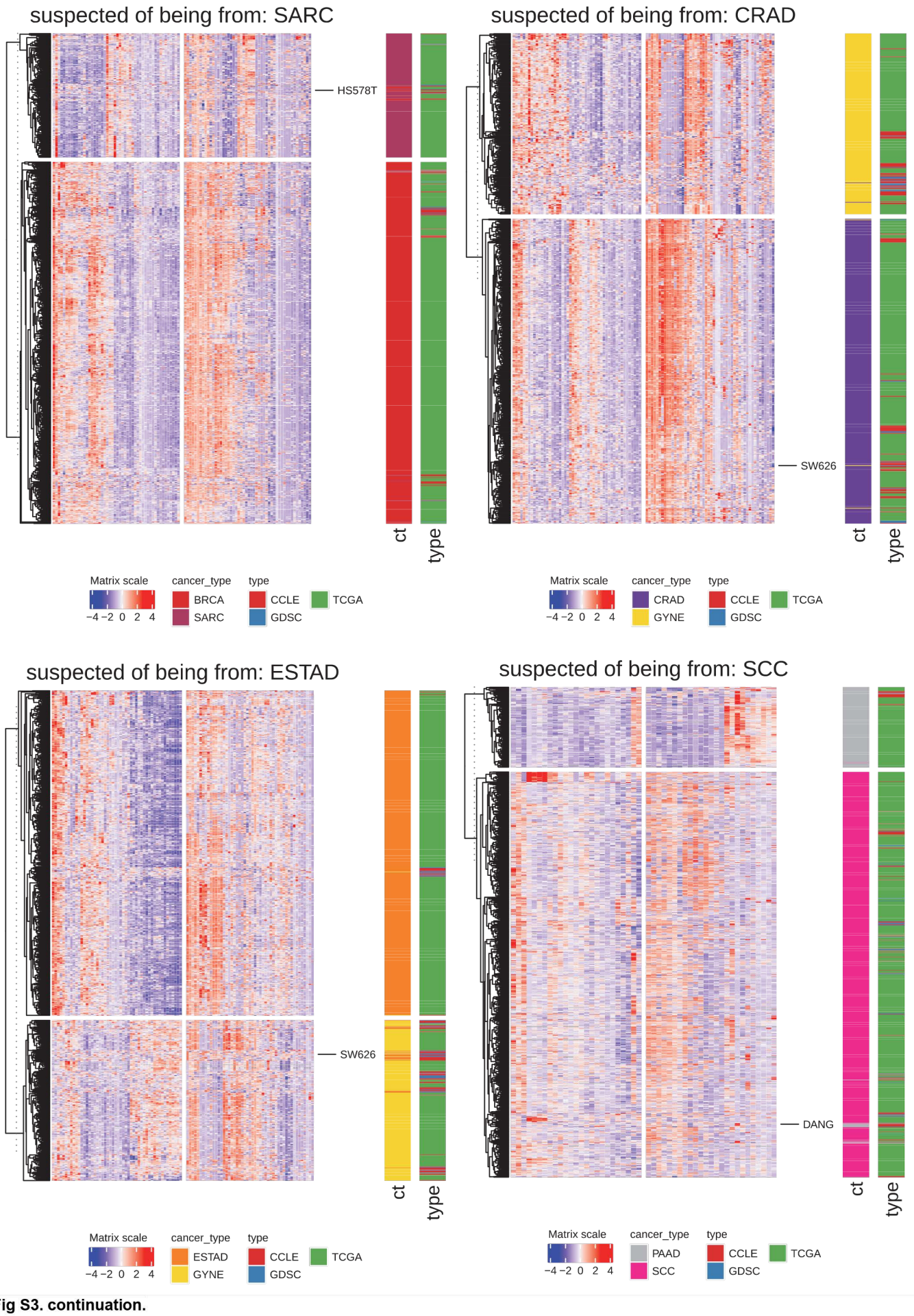

Fig S3. continuation.

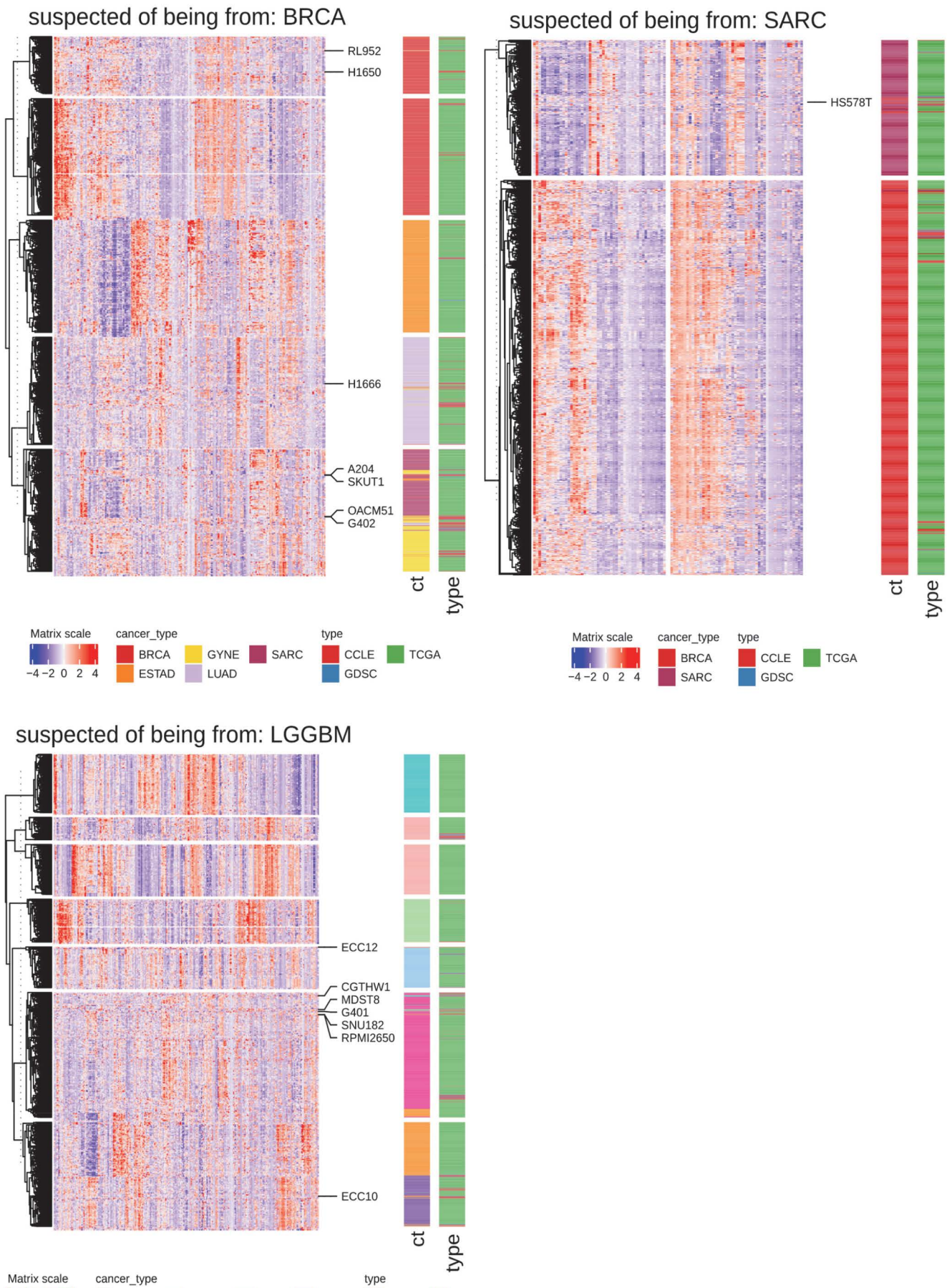

Fig S3. Continuation.

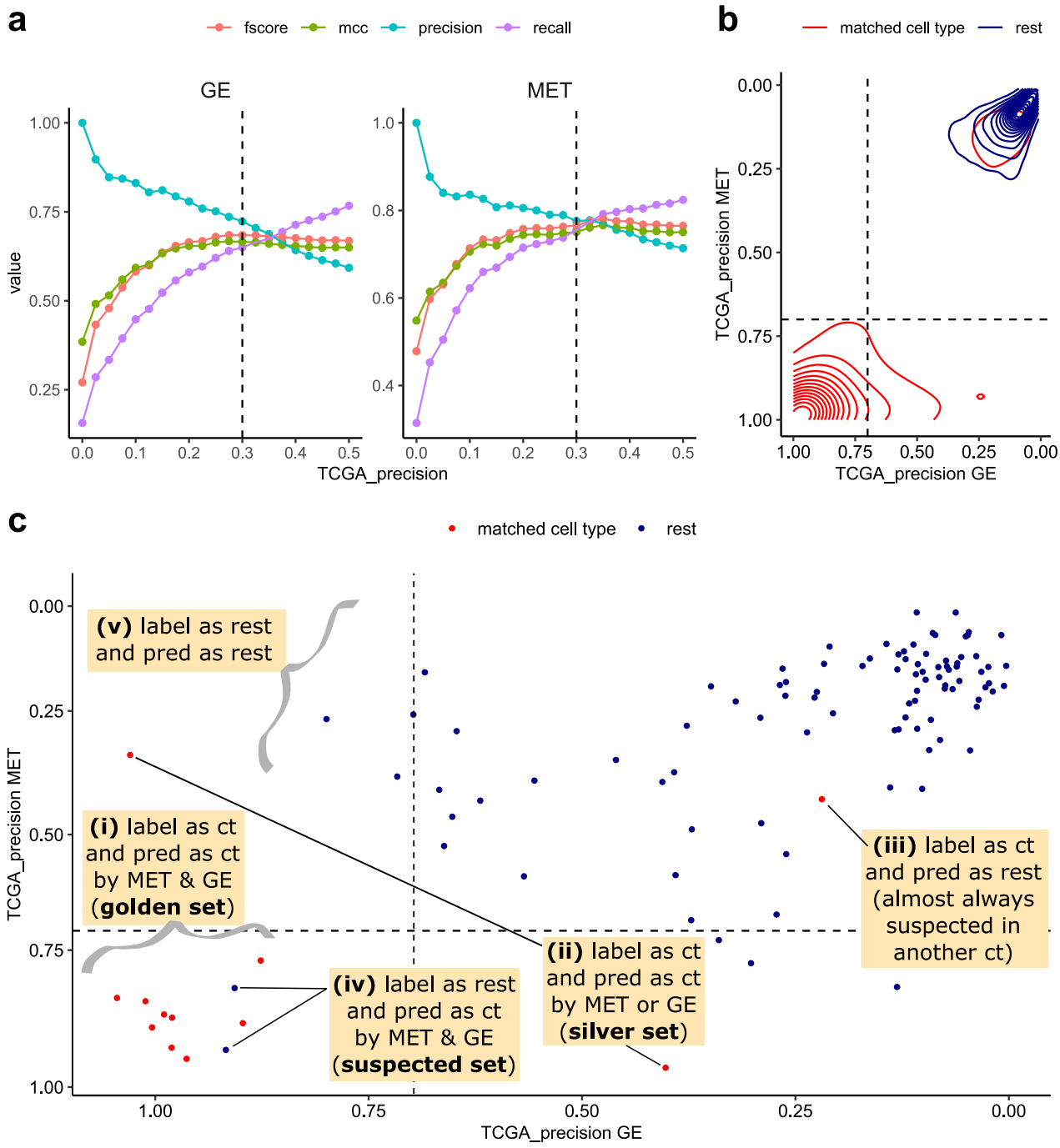

Fig S4. Decision thresholds for reclassifying a cell line to another tissue/cell type of origin. (a) *Accuracy measures (y axis) for cutting at different TCGA-based precision score thresholds (x axis) in gene expression (GE, top) and DNA methylation (MET, bottom) classifiers. "fscore" is the F1 score (harmonic mean of precision and recall), "mcc" is the Matthews correlation coefficient. Shown accuracy measures are derived only from tissue-of-origin labels on cell line data.* ***(b)*** *Distribution (two-dimensional histogram) of data points, collected across classifiers for all tissues, for matched cancer types (red) and the remainder (in blue).* ***(c)*** *A schematic diagram outlining different cases of how, for the selected threshold (TCGA-based precision score ≥0.7), a cell line is classified to the original cancer type or to an alternate cancer type. The shows shows an example of a one-versus-rest classifier, where "ct" denotes that one cancer type that the classifier is trained to recognize from the rest of the cancer types. "pred" is shorthand for "predicted". "label" implies the original label.*

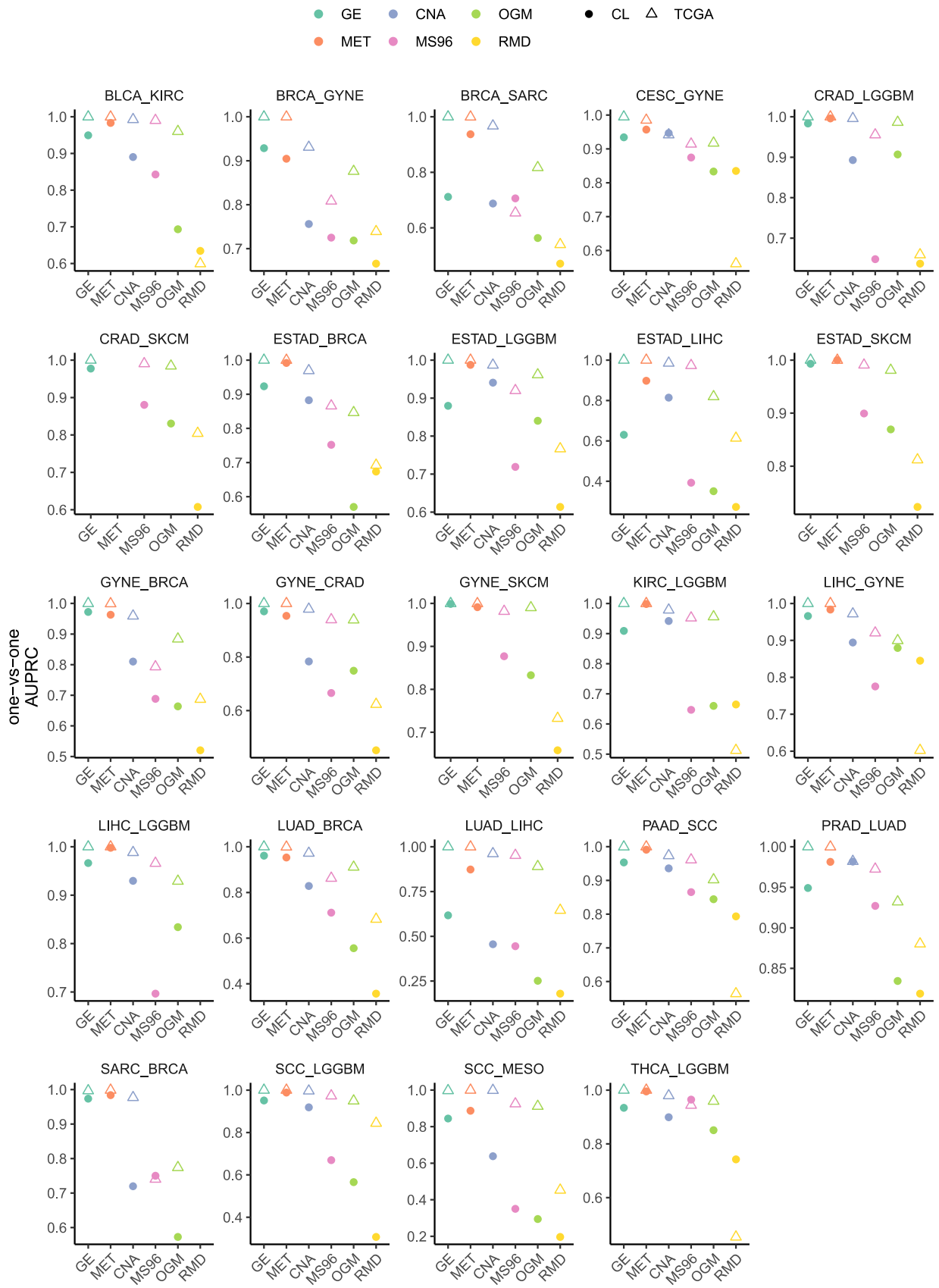

Fig S5. Accuracy of classifiers that distinguish between pairs of cancer types of interest. *Plots show the Area Under the Precision Recall curve (AUPRC) statistic for predicting cancer type in one-versus-one classifiers (the two tested cancer types are named in the top of the plot) using different types of features. GE, gene expression. MET, DNA methylation. CNA, copy number alterations. MS96, trinucleotide mutation spectrum. OGM, oncogenic mutations. RMD, regional mutation density. The classifier is trained on a training set of tumors from TCGA and the AUPRC score is calculated either on a held-out testing set of tumors (TCGA, shown with a triangle), and on all cell lines (CL, shown with a circle).*

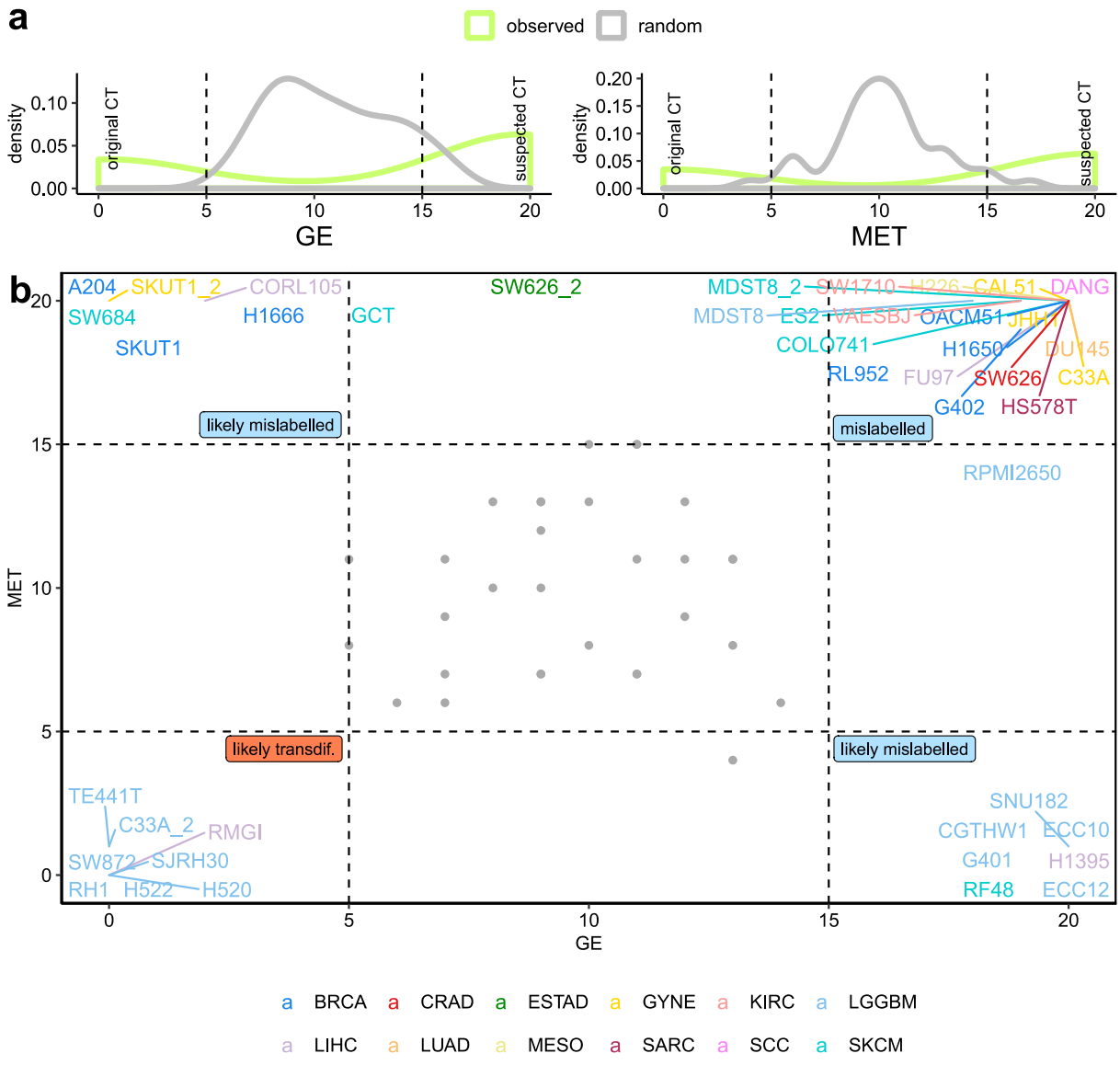

Fig S6. Additional evidence supporting tissue identity of the suspected mislabelled cell lines. (a) *Distribution of the consistency scores (outcome of classifier that predicts suspected versus the original cancer type, run 20 times), for GE and MET features on real data and on randomized data). "CT", cancer type.* ***(b)*** *Prediction consistency scores for the gene expression (GE, x axis) and DNA methylation (MET, y axis) classifiers, for the suspected mislabelled cell lines. Colors represent the suspected cancer type. Grey dots represent randomized data.*

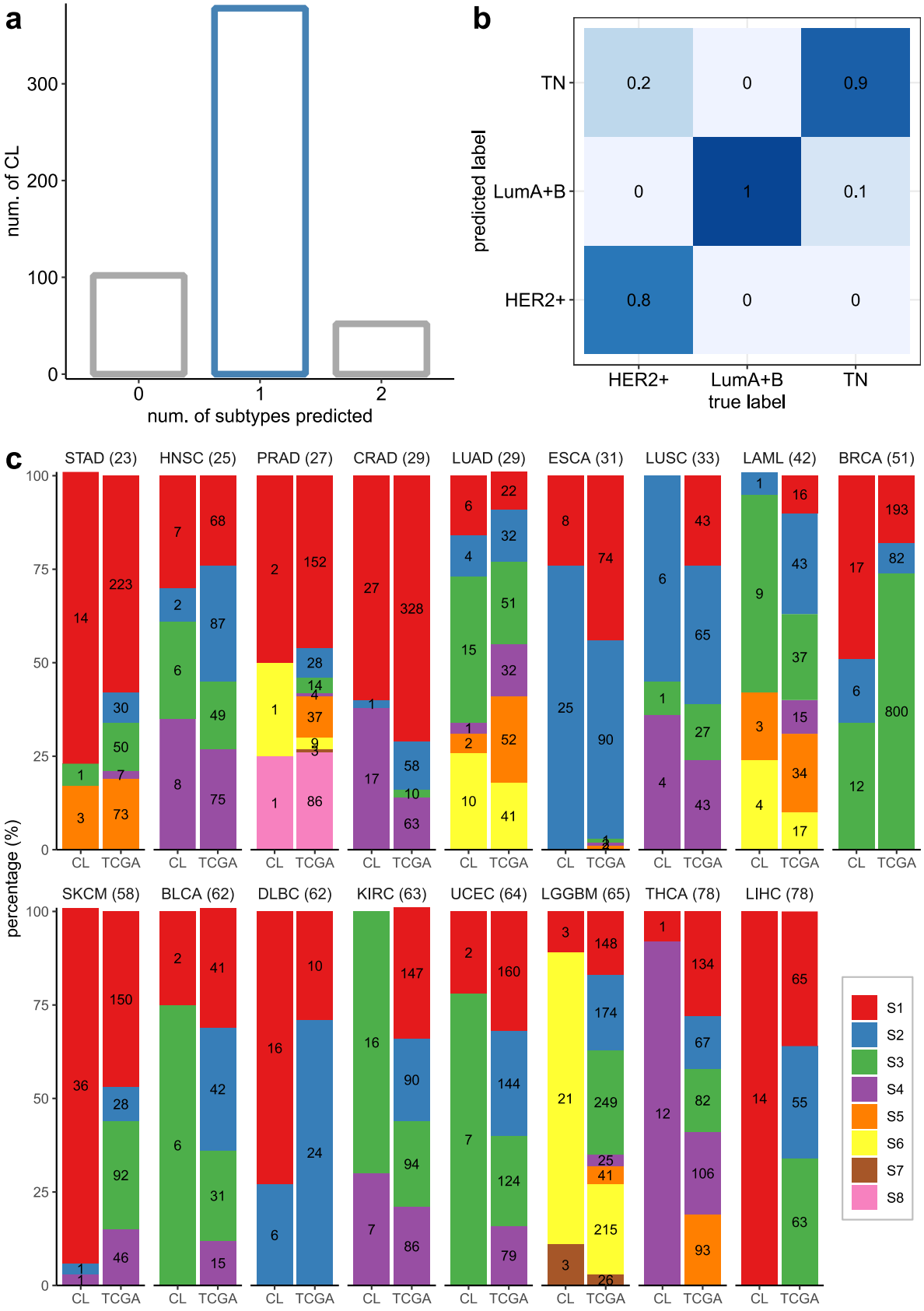

Fig S7. Assigning tumor subtypes to cancer cell lines. (a) *Number of cell lines that are assigned to 0, 1, or 2 subtypes (using one-vs-rest subtype classifiers, within each cancer type).* ***(b)*** *Normalized confusion matrix for the prediction of the breast cancer subtypes.* ***(c)*** *Subtype proportions (raw counts shown overlaid on the bars) observed in the cell lines (CL) and the tumors (TCGA) for each cancer type. The cancer types are ordered from more similar to less similar (left to right) according to concordance between subtype proportions in CL and TCGA (by Euclidian distance, shown above bars; lower scores are better, meaning that examined cell line panels are more representative of tumors in TCGA in terms of subtype representation). Chi-square statistic, p-value and Euclidean distance for the differences between CL and TCGA are provided in Supplementary Table 4.*

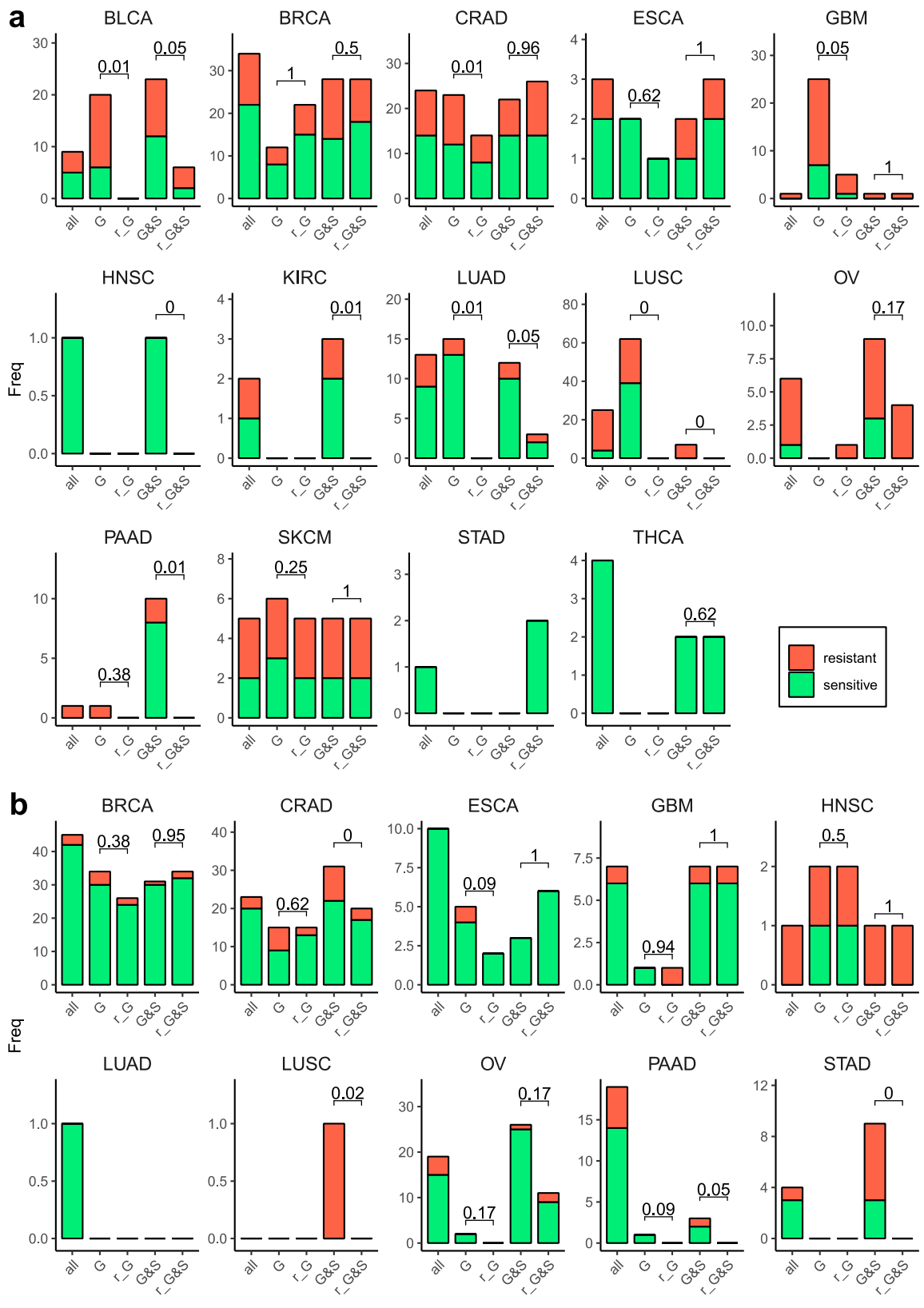

Fig S8. Tally of significant associations between cancer functional events and drug activity or gene dependency on focused panels of cell lines.* (a)** *Associations with drug activity were detected in an ANOVA test (at FDR 25%) for all cell lines (all), cell lines in the golden set (G), cell lines in the golden and silver sets (G&S) and a random subset of cell lines that match the number of cell lines in the golden set (r_G) and in the golden and silver set (r_G&S). For the random subsets, the number of significant associations is calculated from 10 random samplings and the median shown. P-values shown over the bars are from a sign test (one-tailed) between the associations in the G/G&S and the associations in the 10 runs of random_G/random_G&S.* ***(b) *Associations with gene dependency (from CRISPR/Cas9 k.o. screens) were detected in an ANOVA test (at FDR 25%). Labels and statistical tests as in (a).*

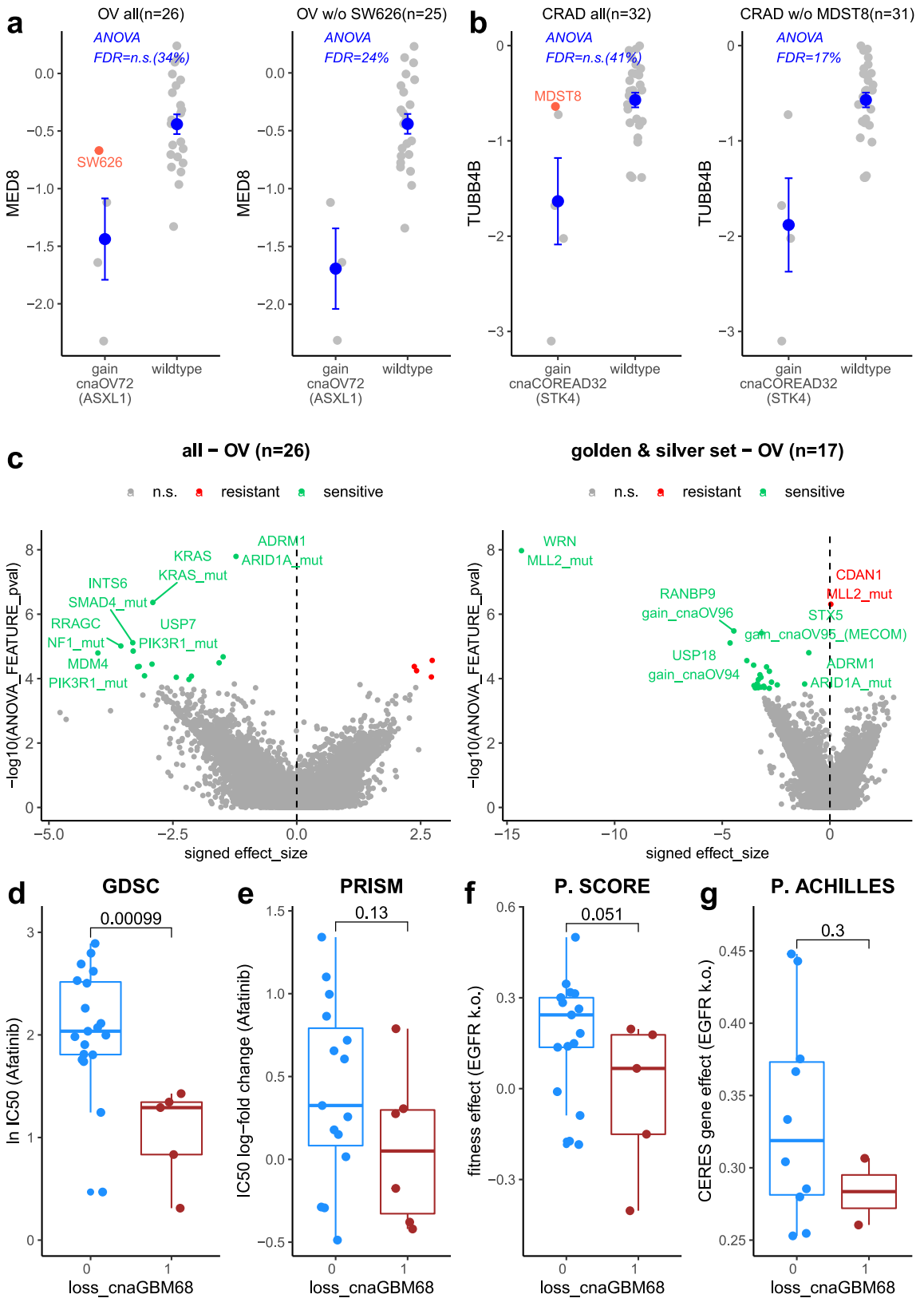

Fig S9. Analyses of genetic screening data using high-confidence cell lines and validation of an illustrative example with independent datasets. (a) *Fitness effect (fold change) for MED8 gene k.o. in all ovarian cancer (OV) cell lines (left) and all OV cell lines except SW626, which is suspected to originate from colorectal cancer (right). Cell lines with copy number gain in a region containing ASXL1 (gain_cnaOV72), and without (wild-type) are compared. ANOVA FDR for this association (MED8 k.o. and gain_cnaOV72) is shown in blue for both datasets.* ***(b)*** *Fitness effect (fold change) for TUBB4B k.o. in all CRAD cell lines (left) and all CRAD cell lines except MDST8, which is suspected to originate from skin (right). Cell lines with copy number gain in region containing STK4 ("COREAD32") and without (wild-type) are compared. ANOVA FDR for this association (TUBB4B k.o. and "COREAD32") shown in blue for both datasets.* ***(c)*** *Differential dependency biomarkers were analysed by ANOVA for all OV cell lines (left) and those in the golden and silver sets only (right). Each point is an association between the fitness effect of a gene and a genetic feature (CFE).* ***(d-g)*** *Drug sensitivity to afatinib in glioblastoma (GBM) cell lines in the 'golden set'. Two groups are compared: cell lines with a particular copy number loss (ID="cnaGBM68", coordinates=chr1:51169045-51472178 (1p32.3), genes affected are CDKN2C and FAF1) (labelled as 1) and wild-type (0). Afatinib IC50 obtained from GDSC in* ***(d)*** *and from PRISM* ***(e).*** *Gene dependency to EGFR k.o. in GBM cell lines (golden set only). Two groups are compared: cell lines with loss ofcnaGBM68 (1) and wild-type cells (0). Gene dependency data obtained from Project Score* ***(f)*** *and from Project Achilles* ***(g).*** *P-values calculated with Wilcoxon test, one-tailed.*