ACCSCIENCE PUBLISHING

Tumor Discovery

ORIGINAL RESEARCH ARTICLE

Highly specific and sensitive gene panels for cancer screening: First application of only-normal and only-tumor genes

Supplementary File

1. Perfect gene panel inventory

Hereafter, we list the perfect gene panels obtained for the cancer types under study. Genes were identified using Entrez symbols, whenever available, or alternatively, by their Ensembl codes. For each panel, genes are itemized in descending order of the number of type II errors incurred when classification is performed using their individual expression levels.

1.1. Cancer type: Breast invasive carcinoma

- Only-T-above panel (six genes): *MMP11*, *FLAD1*, *SLC7A8*, *CCN4*, *PTTG1IP*, and *RRP1B*
- Only-N-above panel (10 genes): *ARHGAP20*, *VEGFD*, *PAMR1*, *AC084759.3*, *LEPR*, *SLC2A4*, *DST*, *CBX7*, *ELOVL7*, and *FAM236D*
- Only-T-below panel (six genes): *SPRY2*, *CA4*, *TMEM220-AS1*, *PGM5P3-AS1*, *FP325317.1*, and *GABARAPL1*

1.2. Cancer type: Colon adenocarcinoma

- Only-T-above panel (two genes): *KRT80* and *ESM1*
- Only-N-above panel (one gene): *SCARA5*
- Only-T-below panel (one gene): *SCARA5*
- Only-N-below panel (two genes): *AJUBA* and *KRT80*

1.3. Cancer type: Head-and-neck squamous cell carcinoma

- Only-T-above panel (four genes): *CDCA5*, *GPRIN1*, *LINC01633*, and *OFCC1*
- Only-N-above panel (eight genes): *EMP1*, *PIP*, *KRTAP13-1*, *CIDEC*, *LINC00443*, *ANXA1*, *CLEC3B*, and *CYP3A4*
- Only-T-below panel (five genes): *ADIPOQ*, *CYP4B1*, *TPT1*, *GPT2*, and *B4GALT1-AS1*

1.4. Cancer type: Kidney renal clear cell carcinoma

- Only-T-above panel (four genes): *DDB2*, *PARVB*, *SLC15A4*, and *GACAT2*
- Only-N-above panel (three genes): *AQP2*, *SOST*, and *TSPAN6*
- Only-T-below panel (three genes): *AC104237.2*, *PAQR7*, and *LY86-AS1*

1.5. Cancer type: Kidney renal papillary cell carcinoma

- Only-T-above panel (three genes): *HK2*, *AC124798.1*, and *NME1*
- Only-N-above panel (one gene): *UMOD*
- Only-T-below panel (one gene): *UMOD*

1.6. Cancer type: Liver hepatocellular carcinoma

- Only-T-above panel (three genes): *GABRD*, *SEPTIN7P2*, and *TOMM40L*
- Only-T-below panel (five genes): *ANGPTL6*, *LCAT*, *UGT2B4*, *LINC02027*, and *MIP*
- Only-N-below panel (five genes): *LRRC14*, *MSTO1*, *AURKA*, *APLN*, and *FLAD1*

1.7. Cancer type: Lung adenocarcinoma

- Only-T-above panel (three genes): *ALDH18A1*, *TRIM27*, and *PYCR1*
- Only-N-above panel (five genes): *STX11*, *CALCRL*, *SFTPC*, *FHL1*, and *EDNRB*
- Only-T-below panel (four genes): *AGER*, *ST7*, *FABP4*, and *GYPE*

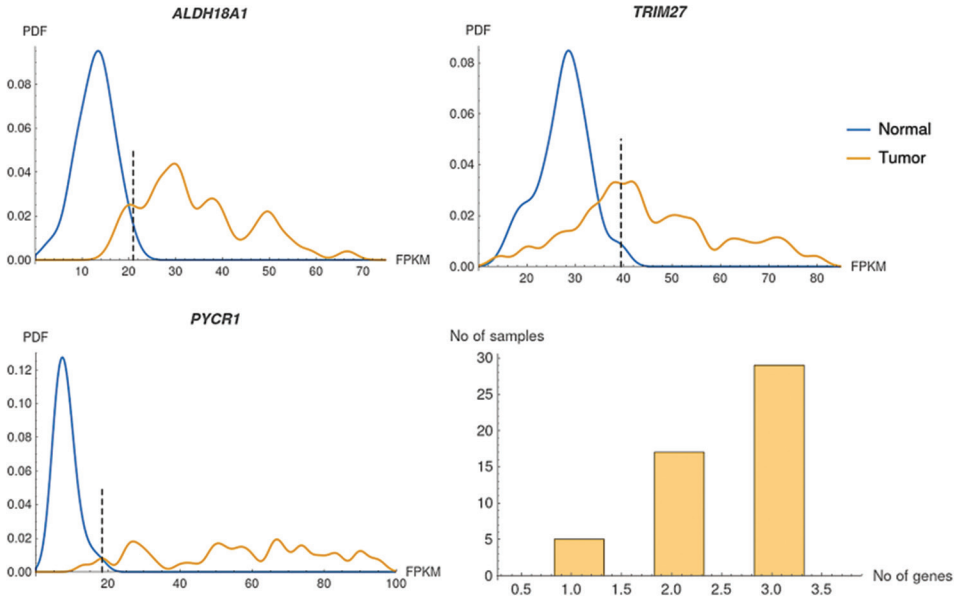

Figure S1. Validation of the three-gene panel for lung adenocarcinoma using the data of Xu *et al.*[1] for a Chinese cohort. The expression data show that the genes remain in the only-T-above class, whereas the last histogram proves that there is at least one deregulated gene for any tumor sample.
Abbreviations: FPKM: Fragments per kilobase of transcript per million mapped reads; PDF: Probability density functions.

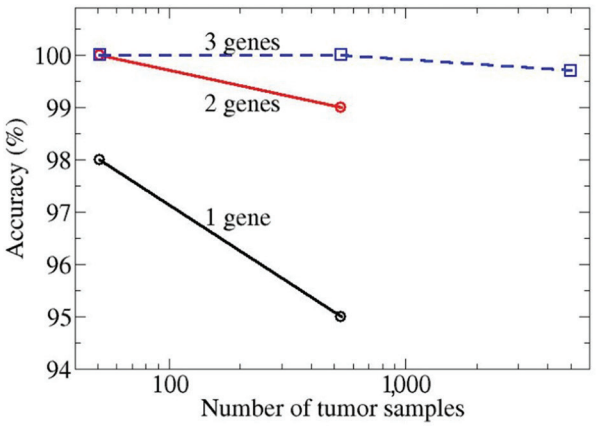

Figure S2. Performance of the one-, two-, and three-gene panels for lung adenocarcinoma as a function of the number of tumor samples. The data from Xu *et al.*[1] (51 samples) and the Cancer Genome Atlas-Lung adenocarcinoma data (535 samples) are used. A hypothetical point corresponding to a dataset with 5,000 samples is also included.

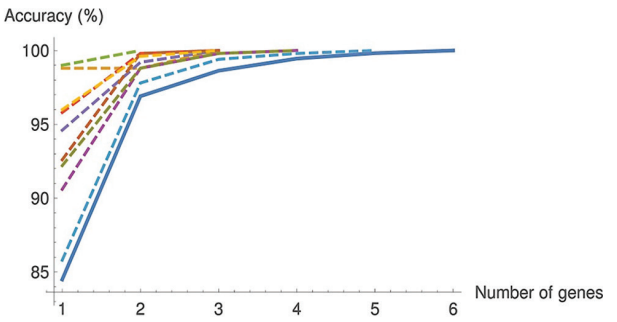

Figure S3. Performance of the six-gene only-T-above panel found for breast invasive carcinoma in imbalanced situations. The thick line indicates the original set of 112 normal and 1,096 tumor samples. Dashed lines show the random subsets of 20 normal and 500 tumor samples. Only nine of these sets are included in the plot for clarity.

1.8. Cancer type: Lung squamous cell carcinoma

- Only-T-above panel (two genes): *MRGBP* and *F12*
- Only-N-above panel (three genes): *EMP2*, *ESAM*, and *GKN2*
- Only-T-below panel (two genes): *ADGRF5* and *ADGRD1*
- Only-N-below panel (three genes): *SLC2A1*, *RFC4*, and *PPP1R14BP3*

1.9. Cancer type: Prostate adenocarcinoma

- Only-T-above panel (eight genes): *EPHA10*, *UCN*, *TMEM86A*, *RPL7AP31*, *FRMPD3*, *RP11-658F2.8*, *HOXA10-AS*, and *UBXN6*
- Only-N-above panel (14 genes): *SEPTIN10*, *KLHL4*, *CD82*, *LINC01546*, *HCAR2*, *FGD5P1*, *MYH11*, *SLC18A2*, *RNA5SP342*, *AL161668.4*, *TES*, *ASS1P1*, *TRMT1L*, and *AC093536.1*
- Only-T-below panel (15 genes): *CTF1*, *SH3GLB1*, *C19orf47*, *DLEU1*, *NRSN2-AS1*, *RHOQP3*, *YWHAZP10*, *CRYAB*, *AP2B1P1*, *TGM5*, *DNAJB4*, *CLCA4-AS1*, *CASQ1*, *GBP6*, and *DCTN6*

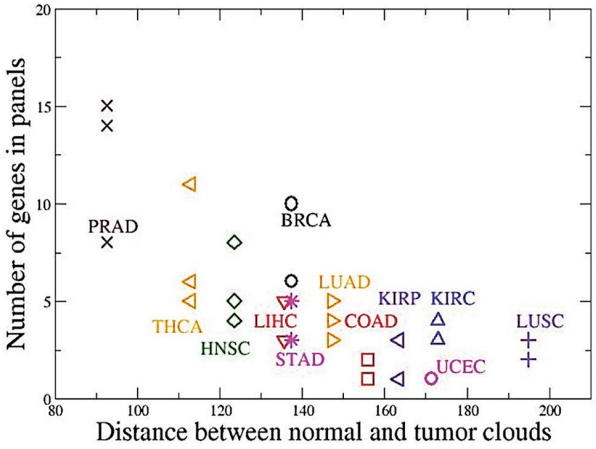

Figure S4. A sketch of the perfect panels found for 12 tissues as a function of the distance between the normal and tumor clouds in gene expression space. As expected, tissues with distant, less-overlapping normal and tumor clouds are more easily classified.
Abbreviations: BRCA: Breast invasive carcinoma; COAD: Colon adenocarcinoma; HNSC: Head-and-neck squamous cell carcinoma; KIRC: Kidney renal clear cell carcinoma; KIRP: Kidney renal papillary cell carcinoma; LIHC: Liver hepatocellular carcinoma; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; PRAD: Prostate adenocarcinoma; STAD: Stomach adenocarcinoma; THCA: Thyroid carcinoma; UCEC: Uterine corpus endometrial carcinoma.

1.10. Cancer type: Stomach adenocarcinoma

- Only-T-above panel (three genes): *ESM1*, *AMH*, and *ZNF761*
- Only-N-above panel (five genes): *IGHV3OR16-13*, *DPT*, *MALL*, *MT-TY*, and *MYZAP*
- Only-T-below panel (five genes): *PLP1*, *ALDOC*, *NSG1*, *FZD9*, and *KRT222*
- Only-N-below panel (five genes): *CENPL*, *COL10A1*, *XPO5*, *ACAN*, and *HSPD1P1*

1.11. Cancer type: Thyroid carcinoma

- Only-T-above panel (five genes): *METTL7B*, *GJC1*, *TYW1*, *UNC5B-AS1*, and *FHOD1*
- Only-N-above panel (11 genes): *AC109326.1*, *AL034374.1*, *PIK3C2G*, *BCL2L11*, *LINC01589*, *RIF1*, *AC105105.1*, *GPM6A*, *UGT2B11*, *AC239804.2*, and *PDXDC2P*
- Only-T-below panel (six genes): *TFF3*, *GBA3*, *SLC6A16*, *TRMO*, *AC010834.2* and, *AC131206.1*

1.12. Cancer type: Uterine corpus endometrial carcinoma

- Only-T-above panel (one gene): *TBC1D7*
- Only-N-above panel (one gene): *PLSCR4*
- Only-T-below panel (one gene): *PLSCR4*
- Only-N-below panel (one gene): *TBC1D7*

Reference

1. Xu JY, Zhang C, Wang X, *et al.* Integrative proteomic characterization of human lung adenocarcinoma. *Cell.* 2020;182(1):245-261.e17.

doi: 10.1016/j.cell.2020.05.043