Supplementary Methods

Surgical sample processing

Formalin-fixed, paraffin-embedded (FFPE) blocks or unstained cut sections from gastrectomy specimens were collected from Japanese sites participating in the SAMIT study. The samples were collected by the Kanagawa Cancer Center Data Center, Yokohama, Japan and subsequently shipped to Yokohama City University, Yokohama, Japan, for RNA extraction. Extracted RNA was transferred to Duke-NUS, Singapore for NanoString analysis. The translational study analysis was approved by the Domain Specific Review Board (DSRB), Singapore (Ethics approval Ref: 2019/00429).

RNA Extraction

Haematoxylin/Eosin stained slides were reviewed and the area with the highest tumor content was outlined manually. After manual microdissection, total RNA was isolated from the FFPE GC tissues using the NucleoSpin FFPE RNA XS kit according to the instructions of the manufacturer (MACHEREY-NAGEL GmbH & Co. KG, Düren, Germany). RNA quality control was performed by measuring the OD260/OD280 ratio using NanoDrop 2000 (Thermo Fisher Scientific Inc., MA, USA), and determining the total RNA Integrity Number (RIN) using the Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Waldbronn, Germany).

NanoString analysis

Raw counts were normalized using the geometric mean of the internal positive control probes included in the CodeSet and housekeeping genes using the vendor-provided nCounter software (nSolver, Nanostring Technologies, USA). The normalized gene expression data were then used for analysis.

Gene Signature Development using Machine Learning Models

Commonly used machine learning methods for developing predictive models such as support vector machines (SVM) and random forests were evaluated. The random forest method utilizes an ensemble of classification trees of several variables.[1] Various SVM kernels (polynomial, radial and linear) as well as different parameters of random forests such as number of trees (between 1000 and 5000 trees) and number of variables tested in each split (mtry = 2, 3 and 4) were tested, and the best performing model based on 10-fold cross validation was selected for further development. Metrics

for measuring the predictive performance included accuracy, precision, recall, F-measure and area under curve (AUC). AUC was calculated using the R package pROC. The F-measure is the harmonic mean of a model's precision and recall, with higher F-measure implying higher positive predictive value (PPV) and sensitivity.[2] The F-measure has been shown to handle class imbalances in the dataset better than PPV and accuracy.[3] For SVM, F-measures ranged from 0.39 to 0.51, accuracy ranged from 0.55 to 0.66, and AUC ranged from 0.46 to 0.47, compared to random forest, where F-measures ranged from 0.40 to 0.64, accuracy ranged from 0.66 to 0.68 and AUC ranged from 0.60 to 0.74.

Gene expression from RNA-Seq (Pac-Ram cohort) and NanoString were normalized for comparison using DeSeq2 and COMBAT. To calculate F-measure, accuracy and AUC for the Pac-Ram cohort, a surrogate equivalent of 2-year DFS of 4 months was used.[4]

We initially attempted application of the machine-learning approach on the entire SAMIT data set of paclitaxel treated samples by creating randomly selected samples for the training and validation cohorts. The dataset was randomly divided into a training cohort, n = 188 (75%) and validation cohort, n = 63 (25%). A signature using the top fourteen genes was the best performing model. The trained model was applied on the validation cohort and correctly predicted a survival benefit for *Pac-Sensitive* patients (Hazards Ratio (HR): 0.28, 95% CI: 0.13 to 0.62, p = 0.0016). However, when the classifier was applied on the independent external validation Pac-Ram cohort, it was unable to accurately identify patients who benefited from paclitaxel (*Pac-Sensitive* vs. *Pac-Resistant* HR 0.58, 95% CI: 0.27 to 1.22, logrank p =0.15).

References

1 Breiman L. Random Forests. Machine Learning 2001;**45**:5-32.

2 Rijsbergen CJV. Information Retrieval: Butterworth-Heinemann, 1979.

3 Forman G, Scholz M. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. SIGKDD Explor Newsl 2010;**12**:49-57.

4 Wilke H, Muro K, Van Cutsem E, Oh SC, Bodoky G, Shimada Y*, et al.* Ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (RAINBOW): a double-blind, randomised phase 3 trial. The lancet oncology 2014;**15**:1224-35.