**Transcriptomic analysis of cutaneous squamous cell carcinoma reveals a multi-gene prognostic signature associated with metastasis**

Wang et al., JAAD manuscript, 2023 Accepted

Jun Wang, PhD,a Catherine A. Harwood, MD PhD,a,b Emma Bailey, PhD,a Findlay Bewicke-Copley, PhD,a Chinedu Anthony Anene, PhD,a,c Jason Thomson, MD,a,b Mah Jabeen Qamar,a,b Rhiannon Laban,a,b Craig Nourse, PhD,d Christina Schoenherr, PhD,d Mairi Treanor-Taylor, MD,d,e, Eugene Healy, MB PhD,f,g Chester Lai, BM PhD,f,g Paul Craig,h Colin Moyes,i William Rickaby,j Joanne Martin, PhD,a Charlotte Proby, MD,k Gareth J. Inman, PhD,d,e Irene M. Leigh, CBE, DSca

**Supplemental Materials**

**Pathological review and sample selection**

A standard H and E section was marked to identify tumour regions (ideally 10mm2 with a minimum 2mm2) and a similar sized region of normal ‘perilesional’ skin at least 2mm away from the tumour region. A second slide (10mm section) slide was matched and marked on the reverse of the unstained section and was submitted for the transcriptomics analysis.

The H&E sections were digitally scanned with a Leica scanner and Aperio software and images were reviewed independently by two expertdermatopathologists (PC, WR) who possess the Royal College of Pathologists UK (RCPath) Diploma in Dermatopathology where the standard is of an expert referral level dermatopathologist. The diagnosis of cSCC was confirmed and a consensus formed in any case where there was discrepancy for any of these features.

**TempO-seq data analysis**

Samples with a total raw read count less than 1M reads (note that the average raw read count was ~5M across samples) were excluded from our analysis. Raw read counts of captured probes / genes were processed for the following analysis using limma R package. First, probes that consistently had zero or very low counts across all samples were excluded using the default setting in limma gene filtering, leading to 21,701 retained probes. Raw read counts were further normalised using the limma voom function. Next, a gene expression matrix using the gene symbols as the unique identifiers was created using the following strategy (1). For genes with ≤2 probes, we calculated the median value of the probes as the expression value of the gene. For genes with >2 probes, we calculated the correlation among probes, selected the most correlated probes (correlation coefficient, r>0.5) and used the median value of these correlated probes as the expression value of the gene. This led to 19,072 genes sufficiently profiled in our data set. Principle component analysis (PCA) was used to inspect the full data set in terms of sample clustering and grouping based on clinical annotation. Differential expression (DE) analysis was performed using limma (2). Significant DE genes were defined as those with the absolute value of log2 fold change >1 and adjusted p-value < 0.05 in the limma analysis.

**Machine learning process**

A range of ML techniques were used and compared, including Support Vector Machines with Radial Basis Function Kernel (SVM Radial), random forest, PAM with the Nearest Shrunken Centroids, Linear Discriminant Analysis and K-nearest neighbours (KNN). The best performing set of genes for each ML algorithm (i.e., feature selection) was determined using the Recursive Feature Elimination procedure, with 10-fold repeated cross validation of five repeats. Briefly, the method evaluates multiple models using procedures that add and/or remove predictors to find the optimal combination that maximises the model performance, to determine a top performing model based only on a subset of predictive genes. The final number of predictors (genes) for each ML algorithm was determined using the elbow plot of accuracy against the number of variables (genes) in the training set.

**Generation of the linear predictor**

A weighted linear predictor was generated for each sample based on the expression of the final set of genes in the model and their fold changes in the DE analysis, as

$$linear predictor= \sum\_{i=1}^{n}(log\_{2}FC\_{i}×Exp\_{i})$$

where $log\_{2}FC\_{i}$ is the log2 fold change of the *i*th gene comparing metastasising cSCC and non-metastasising cSCC; $Exp\_{i}$ is the log2 normalised expression value of the *i*th gene in this sample, and *n* is the number of genes in the model. Linear predictors were compared between clinical groups and correlated with classes. The area under the ROC curve (AUC) was calculated using the pROC package (3).

**Code and data availability**

All the code used to analyse the data are available at our Github page, https://github.com/BioInforCore-BCI/20gene\_signature. The data used in this study are still part of an ongoing effort of multi-omic integration to develop a joint genomic-transcriptomic prognostic signature. However, requests to access the transcriptomic data can be made to the corresponding author for non-commercial research purposes. Access will be granted upon review of a project proposal, which will be evaluated by a PERMEDID data access committee, and entering into an appropriate data access agreement, subject to any applicable ethical approvals.

1. Sha C, Barrans S, Cucco F, Bentley MA, Care MA, Cummin T, et al. Molecular High-Grade B-Cell Lymphoma: Defining a Poor-Risk Group That Requires Different Approaches to Therapy. J Clin Oncol. 2019;37(3):202-12.

2. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47.

3. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Mar 17;12:77.



**Supplemental Fig 1.** Principal component analysis of 433 samples from four clinical groups sufficiently profiled by TempO-seq. Tum\_met: metastasising primary cSCC; Tum\_no\_met: non-metastasising primary cSCC; Norm\_with\_met: normal adjacent samples to metastasising primary cSCC; Norm\_no\_met: normal adjacent samples to non-metastasising primary cSCC.

 

**Supplemental Fig 2.** Gene set over-representation test on canonical pathways and cellular signatures for 1,274 DE genes between metastasising and non-metastasising primary cSCC. (A) Selected highly significantly over-represented pathways, (B) highly significantly over-represented cellular signatures for DE genes. All analysis was performed using the clusterProfiler R package.

****

**Supplemental Fig 3.** Heatmap of the 20 genes derived from the K-nearest neighbours model across all primary tumour samples. Tum\_met: metastasising primary cSCC; Tum\_no\_met: non-metastasising primary cSCC.

****

**Supplemental Fig 4.** Boxplots of linear predictors between metastasising and non-metastasising primary cSCC in the discovery (training) and testing data sets. The statistical p-value based on the Wilcoxon Rank Sum test is shown. Y: metastasising cSCC; N: non-metastasising cSCC.

 

**Supplemental Fig 5.** Boxplots of linear predictors across the four clinical groups in our cohort. The statistical p-value based on the Wilcoxon Rank Sum test is shown. Tum\_met: metastasising primary cSCC; Tum\_no\_met: non-metastasising primary cSCC; Norm\_with\_met: normal adjacent samples to metastasising primary cSCC; Norm\_no\_met: normal adjacent samples to non-metastasising primary cSCC.

**Supplemental Table I.** The number of samples successfully profiled in TempO-seq used for downstream analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Annotation | # samples | # patients |
| Norm\_no\_met | Normal adjacent to non-metastasising cSCC | 132 | 127 |
| Norm\_with\_met | Normal adjacent to metastasizing cSCC | 71 | 69 |
| Tum\_no\_met | Non-metastasising primary cSCC | 146 | 143 |
| Tum\_met | Metastasising primary cSCC | 84 | 82 |

**Supplemental Table II.** Top 25 dysregulated pathways based on the GSEA analysis (Q value) comparing normal samples adjacent to metastasising and normal adjacent to non-metastasising primary cSCC.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Description | Enrichment Score | Normalised Enrichment Score\* | P value | Adjusted p-value | Q value |
| CD22 Mediated BCR Regulation | 0.815 | 2.734 | 1.00E-10 | 4.98E-09 | 3.44E-09 |
| FCGR Activation | 0.783 | 2.696 | 1.00E-10 | 4.98E-09 | 3.44E-09 |
| Initial Triggering of Complement | 0.752 | 2.654 | 1.00E-10 | 4.98E-09 | 3.44E-09 |
| Role of Phospholipids in Phagocytosis | 0.728 | 2.601 | 1.00E-10 | 4.98E-09 | 3.44E-09 |
| Signaling by the B Cell Receptor BCR | 0.577 | 2.210 | 1.00E-10 | 4.98E-09 | 3.44E-09 |
| Chemokine Receptors bind Chemokines | 0.565 | 1.975 | 1.14E-06 | 3.17E-05 | 2.18E-05 |
| Interleukin 10 Signaling | 0.571 | 1.921 | 2.62E-05 | 5.09E-04 | 3.51E-04 |
| RhoH GTPase Cycle | 0.581 | 1.885 | 8.88E-05 | 1.26E-03 | 8.67E-04 |
| TCR Signaling | 0.494 | 1.868 | 3.29E-08 | 1.24E-06 | 8.57E-07 |
| Elastic Fibre Formation | 0.549 | 1.850 | 1.04E-04 | 1.42E-03 | 9.80E-04 |
| Cellular Response to Hypoxia | 0.498 | 1.798 | 3.82E-05 | 6.81E-04 | 4.69E-04 |
| mRNA Splicing | 0.454 | 1.761 | 4.95E-08 | 1.81E-06 | 1.25E-06 |
| DNA Double Strand Break Response | 0.525 | 1.760 | 4.27E-04 | 4.22E-03 | 2.91E-03 |
| Interleukin 1 Family Signaling | 0.454 | 1.729 | 2.13E-06 | 5.52E-05 | 3.80E-05 |
| Dissolution of Fibrin Clot | 0.682 | 1.728 | 2.16E-03 | 1.50E-02 | 1.04E-02 |
| Synthesis of DNA | 0.456 | 1.720 | 1.73E-05 | 3.79E-04 | 2.61E-04 |
| DNA Replication | 0.444 | 1.686 | 2.91E-05 | 5.50E-04 | 3.79E-04 |
| Activation of Matrix Metalloproteinases | 0.524 | 1.657 | 2.84E-03 | 1.83E-02 | 1.26E-02 |
| Extracellular Matrix Organization | 0.412 | 1.633 | 2.74E-08 | 1.07E-06 | 7.36E-07 |
| DNA Repair | 0.411 | 1.625 | 5.30E-08 | 1.89E-06 | 1.30E-06 |
| Signaling by NOTCH4 | 0.440 | 1.607 | 1.33E-03 | 1.07E-02 | 7.36E-03 |
| Cell Cycle Mitotic | 0.398 | 1.602 | 1.00E-10 | 4.98E-09 | 3.44E-09 |
| Integrin Cell Surface Interactions | 0.428 | 1.569 | 2.62E-03 | 1.72E-02 | 1.18E-02 |
| Platelet Activation Signaling and Aggregation | 0.383 | 1.507 | 2.08E-05 | 4.25E-04 | 2.93E-04 |
| Golgi to ER Retrograde Transport | 0.379 | 1.443 | 4.68E-03 | 2.71E-02 | 1.87E-02 |

\* Positive normalised enrichment scores indicate upregulated activities in normal samples adjacent to metastasising compared to normal adjacent to non-metastasising primary cSCC. The higher the score, the more upregulated.

**Supplemental Table III.** A 20-gene GEP signature derived from the KNN model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene Symbol | Description | Log2 fold change\* | P value | Adjusted p-value |
| STMN1 | stathmin 1 | 1.447 | 1.41E-10 | 1.62E-08 |
| CALML3 | calmodulin-like 3 | -2.077 | 3.38E-19 | 1.61E-15 |
| CDSN | corneodesmosin | -3.027 | 5.02E-18 | 1.06E-14 |
| FAM25C | family with sequence similarity 25 member C | -2.897 | 2.11E-19 | 1.34E-15 |
| SPNS2 | SPNS lysolipid transporter 2, sphingosine-1-phosphate | -1.694 | 2.35E-20 | 2.24E-16 |
| LCE3C | late cornified envelope 3C | -3.453 | 2.80E-21 | 5.34E-17 |
| TREX2 | three prime repair exonuclease 2 | -1.902 | 9.61E-17 | 1.08E-13 |
| LCE2B | late cornified envelope 2B | -3.742 | 1.28E-18 | 4.89E-15 |
| HAL | histidine ammonia-lyase | -2.787 | 2.55E-18 | 6.65E-15 |
| CYP4F22 | cytochrome P450 family 4 subfamily F member 22 | -2.596 | 1.07E-16 | 1.13E-13 |
| FAM25E | family with sequence similarity 25 member E | -2.934 | 1.70E-14 | 7.71E-12 |
| LCE1C | late cornified envelope 1C | -3.003 | 1.74E-14 | 7.73E-12 |
| FAM25G | family with sequence similarity 25 member G | -3.142 | 1.52E-17 | 2.51E-14 |
| MDK | midkine | 1.316 | 1.39E-14 | 6.45E-12 |
| ASPRV1 | aspartic peptidase retroviral like 1 | -2.654 | 5.05E-17 | 6.97E-14 |
| PLA2G4F | phospholipase A2 group IVF | -2.064 | 4.29E-16 | 3.15E-13 |
| FAM25A | family with sequence similarity 25 member A | -3.106 | 1.26E-17 | 2.40E-14 |
| C10orf99 | G protein-coupled receptor 15 ligand | -2.914 | 1.17E-14 | 5.74E-12 |
| LCE2C | late cornified envelope 2C | -2.954 | 6.05E-16 | 3.98E-13 |
| PLA2G4E | phospholipase A2 group IVE | -2.126 | 4.69E-15 | 2.48E-12 |

\*Log2 fold change in the comparison of metastasising versus non-metastasising primary cSCC. This was used to calculate the linear predictor.

**Supplemental Table IV.** Matrices of BWH T-staging based on original pathology reports without central consensus review (BWH v1) were also shown in “poorly differentiated” and “BWH T stage” fields.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | All (n=237) | No metastasis (n=151)\* | Metastasis (n=86)\*\* | P value |
| Poorly differentiated, n (%) | 79 (33.2) | 20 (13.2) | 59 (68.6) | <.0001 |
| BWH T stage, n (%)¶¶ |  |  |  | <.0001 |
| T1 | 104 (45.6) | 101 (68.2) | 3 (3.8) |  |
| T2a | 58 (25.4) | 33 (21.6) | 25 (31.25) |  |
| T2b | 60 (28) | 14 (9.46) | 46 (57.5) |  |
| T3 | 6 (2.6) | - | 6 (7.5) |  |
|  |  |  |  |  |