Additional Methods

Leeds Melanoma Cohort (LMC|)

The LMC contains detailed information on clinico-histopathological variables and environmental exposures, and survival data is regularly updated from the national cancer registry and via primary and secondary care records. The H&E-stained slides of the blocks which were sampled for RNA extraction were also reviewed by a single observer (SJ O'S) to quantify the number of nuclear mitoses present within the sampled region of tumour. The core number of mitoses relates to the region of the tumour sampled for mRNA extraction. The number of mitoses were dichotomised into two categories as mitoses ≥1 and mitoses <1.

Nearest centroid classification (NCC)

To classify the tumour samples using the NCC method, gene expressions of each sample were correlated with the respective signature's class centroids and was then put in the class corresponding to the largest correlation. The samples which had a maximum correlation of less than 0.10 were considered unclassified.

Consensus PAM clustering

Partitioning around medoids (PAM), also known as k-medoid clustering, aims to minimise the within-cluster distance W(L).

$$W(L) = \sum_{i=1}^{L} \sum_{p \in c_i} D(p, m_i)$$

where L is the number of clusters, p are the elements of cluster c_i, m_i is the medoid (central element) of cluster c_i, and D(.) is the dissimilarity function [1, 2]. The clustering solution of the dataset is at the minimum value of W(L). At first run, PAM randomly assigns L samples as medoids from the dataset. The rest of the samples are classified into a cluster based on their nearest medoid and W(L) is calculated. In each iteration, a new medoid is determined for each cluster by finding a point that minimizes the dissimilarity with other samples of the cluster. The samples are then reclassified into the clusters based on the new medoids. This process is repeated until the W(L) value stabilizes. Since clusters in PAM are defined by medoids rather than centroids (such as in k-means), PAM clusters are less sensitive to outliers [3].

Consensus clustering

Clustering algorithms such as k-means or PAM do not indicate the optimal number of clusters in the data. This can be accomplished by integration of a consensus-based approach to KM and PAM algorithms [4]. Consensus-based clustering is an iterative algorithmic technique for identifying a stable number of clusters in high-dimensional datasets [4]. It sub-samples data, picking a random subset of samples and features, to generate multiple datasets over N iterations. The samples in these datasets are classified in a varying number of clusters k = [2, . . .,k_{max}], where k_{max} is input by the user (here we used k_{max}=12). For each k, and every dataset, every sample pair has consensus 1 if they are classified in the same cluster or 0 otherwise. Across the N iterations (re-samples), the sample pair consensus is the proportion of iterations in which the pair is put in the same cluster. This consensus information is stored in a n*n consensus matrix where n is the total number of samples.

Selecting k based on consensus matrices

The optimal number of clusters, k, was determined by visual examination of the consensus matrix heatmap, the cumulative distribution function (CDF) and delta CDF graphs of the consensus scores.[4] The elements M(i,j) in consensus matrix M are the proportion of times that a sample pair (i,j) are observed in the same cluster over number of times the sample pair were included in the same random sample subset. For any application of the clustering algorithm and resulting consensus matrix, the CDF can be defined, where CDF(c) is the proportion of pairs whose consensus index is less than or equal to c (0 ≤c ≤ 1). The area under the CDF corresponding to consensus matrices obtained at various k values, A(k), generally increases with k. The delta graph for relative change in area under the CDF was examined to select the maximum value of k that induces a substantial increase in the area under the CDF A(k) [4, 5].

The CDF and area under the CDF was calculated as:

$$CDF(c) = \frac{\sum_{i<j} 1\{M(i,j) \leq c\}}{n(n-1)/2}$$

$$A(k) = \sum_{i=2}^{m} [x_i - x_{i-1}] CDF(x_i)$$

where $1\{\ \ \}$ denotes the indicator function, n is total number of samples, k is the number of clusters, M is the consensus matrix obtained at given k and the set (x_1, x_2, . . .,x_m) is the sorted set of entries of the consensus matrix M (with m= n(n-1)/2). The relative change in the CDF, (Δ(k)), was computed as follows:

$$\Delta(k) = \begin{cases} A(k), & if\ k = 2 \\ \dfrac{[A(k) - A(k-1)]}{A(k-1)}, & if\ k > 2 \end{cases}$$

With this method, 6 tumour clusters were identified in the LMC dataset (6 LMC classes).

Selecting k based on cluster separation

A new objective measure, cluster separation index, calculated as the intra- to inter-cluster similarity ratio was used to further support the optimal k selected from visual examination of the CDF graphs. The intra-cluster similarity was calculated as the average consensus score across all pairs of samples within clusters; the inter-cluster similarity was calculated as the average consensus value across all pairs of samples in different clusters. The ratio of intra-cluster similarity to inter-cluster similarity was calculated for all values of k, and the optimum k was selected based on relative increase of this ratio. The intra-cluster and inter-cluster similarity calculations were done using R-package FPC [6]. The 6 clusters were confirmed by this method.

Refining the LMC signature

The LMC signature was refined following two steps: firstly, genes were ranked based on the P-values for their differential expression across the 6 LMC classes based on a Kruskal-Wallis test. After correction for multiple testing by Bonferroni method, the genes with adjusted $P< 0.0001$ were selected. Secondly, for each LMC class, these significant genes were ranked based on their mean expression, and the y most upregulated genes were selected, with y = 1, 5, 10, 25, 50, 100 giving 6 reduced signatures. The classification accuracy was compared between the initial tumour class labels and tumour class labels from each of the 6 reduced signatures. Classification accuracy was computed as follows: the mean expression of the top y genes was calculated for each class (using the initial class labels) to generate class centroids, which were then used in supervised NCC to reclassify the LMC tumours into new classes. The new classification labels were compared with the original tumour class labels using a contingency table. The proportion of overlap at diagonal values for the two classifiers represent classification accuracy. The classification accuracy was plotted for each of the reduced signatures.

Biological characterisation of the LMC signature

The differentially expressed genes (DEG) in each class of LMC signature versus all other classes were identified using the regularised t-test in R-package SAMR [7]. The DEGs which had a q-value equal to 0 were selected for network enrichment using Reactome Fiviz in Cytoscape [8, 9]. Pathways enrichment of DEG's was assessed using KEGG [10-12] and Reactome database [13] with hypergeometric test false discovery rate (FDR) < 0.01. The central nodes of the network were identified using a centrality measure (betweenness) in Gephi [14].

The biological associations of different tumour classes were also characterized using gene network modules reported by Cirenajwis et al. [15] in a Lund melanoma cohort, which we will refer to as "the Lund modules". Each of the Lund modules is a list of genes, identified using network analysis of highly correlated genes in metastatic tumours; the modules were named immune, MITF, stroma, cell cycle and interferon [15]. Of 231 genes in the immune module, 215 were present in the LMC dataset; similarly for 25 out of 26 genes from the MITF module, all genes of the cell cycle (11 genes), interferon (7 genes) and stroma (119 genes) modules. The mean of the module gene expression was calculated for each tumour. The correlation between the module scores was estimated using Spearman's correlation coefficient.

Applying a previously described approach of immune cell scoring [16], we calculated scores of 27 immune cell subsets using expressions of their specific genes as reported by Angelova et al [17]. We compared these immune cell scores across the identified tumour classes.

Copy number Alterations

The DNA extracted from tumours cores were quantified using Quant-iT™ broad range ds-DNA assay kit (Invitrogen™, Life Technologies, USA). Whole-genome DNA libraries were prepared using: 1) a previously described library preparation method [18, 19]; 2) the NEBNext® Ultra™ DNA library prep kit for Illumina® (indexed primers) (New England BioLabs, UK). The sequencing was performed on an Illumina GAII (initial 75 samples) or HiSeq sequencer (all subsequent samples) to produce >100 basepair paired-end reads. The sequence reads were trimmed to remove the low quality reads and adapters using cutadapt version 1.8.3. An average of 82.1 x 10^6 reads to 368.2 x 10^6 read pairs were obtained from the libraries. The reads were aligned using bwa mem 0.7.10 to GRCh38 human reference (no alternate contigs) [20]. Binning was performed (https://github.com/alastairdroop/bamwindow) and reads were assigned to bins using read midpoint, such that each read fell into exactly one window. The window read counts were normalized to reduce technical variation i.e. GC content and mappability bias. GC content was calculated from the reference genome while the mappability was estimated using gem-mappability software [21]. LOESS model was used to adjust for GC content and mappability bias in read counting.

1. Bandyopadhyay, S. and S. Bandyopadhyay, *Analysis of Biological Data: A Soft Computing Approach - Vol. 3*. 2007: World Scientific Publishing Co., Inc. 352.
2. Kaufman, L. and P. Rousseeuw, *Clustering by means of medoids. in 'Y. Dodge (editor) Statistical Data Analysis based on L1 Norm', 405-416*. 1987, Elsevier/North-Holland.
3. Crowley, J. and A. Hoering, *Handbook of statistics in clinical oncology.* 2012: p. 572-574.
4. Monti, S., et al., *Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data.* Machine learning, 2003. **52**(1-2): p. 91-118.
5. Wilkerson, M.D. and D.N. Hayes, *ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.* Bioinformatics, 2010. **26**(12): p. 1572-3.
6. Hennig, C., *fpc: Flexible procedures for clustering. R package version 2.1-5*. 2013.
7. Tibshirani, R., et al., *samr: SAM: Significance Analysis of Microarrays. R package version 2.0*. 2011.
8. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

9. Wu, G., et al., *ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis.* F1000Res, 2014. **3**: p. 146.

10. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res, 2017. **45**(D1): p. D353-D361.

11. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

12. Kanehisa, M., et al., *KEGG as a reference resource for gene and protein annotation.* Nucleic acids research, 2015. **44**(D1): p. D457-D462.

13. Fabregat, A., et al., *The Reactome Pathway Knowledgebase.* Nucleic Acids Res, 2018. **46**(D1): p. D649-D655.

14. Bastian, M., S. Heymann, and M. Jacomy, *Gephi: an open source software for exploring and manipulating networks.* Icwsm, 2009. **8**: p. 361-362.

15. Cirenajwis, H., et al., *Molecular stratification of metastatic melanoma using gene expression profiling: Prediction of survival outcome and benefit from molecular targeted therapy.* Oncotarget, 2015. **6**(14): p. 12297.

16. Nsengimana, J., et al., *beta-Catenin-mediated immune evasion pathway frequently operates in primary cutaneous melanomas.* J Clin Invest, 2018. **128**(5): p. 2048-2063.

17. Angelova, M., et al., *Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy.* Genome Biol, 2015. **16**(1): p. 64.

18. Wood, H.M., et al., *Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-*

fixed paraffin-embedded specimens. Nucleic acids research, 2010. **38**(14): p. e151-e151.

19. Craig, D.W., et al., *Identification of genetic variants using bar-coded multiplexed sequencing.* Nature methods, 2008. **5**(10): p. 887.

20. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* bioinformatics, 2009. **25**(14): p. 1754-1760.

21. Derrien, T., et al., *Fast computation and applications of genome mappability.* PloS one, 2012. **7**(1): p. e30377.