SUPPLEMENTARY METHODS

Datasets and samples

A systematic search was performed in January 2018 to identify publicly available datasets that included gene expression profiles of malignant pleural mesothelioma (MPM) tumor samples. To perform this search, several platforms were used: Gene expression omnibus (GEO)[1] using the query “(("pleura"[MeSH Terms] OR pleural[All Fields]) AND ("mesothelioma"[MeSH Terms] OR mesothelioma[All Fields])) AND "Homo sapiens"[porgn] AND ("gse"[Filter] AND "Expression profiling by array"[Filter]))”; and ArrayExpress[2] using the query “pleural mesothelioma” and filtered by organism “Homo sapiens”, experiment type “rna assay”, and enabling the filter AE only to show results from within ArrayExpress. Moreover, we also used PubMed to search for datasets related with known biomedical literature studies.

To increase the accuracy and reproducibility of the analyses performed along this study, we included datasets that had at least 30 samples, and covered most of the transcriptome. As a result, eight datasets remained, yet one of them[3] was discarded to avoid potential sample redundancies due to same authorship and nearby publication period, we kept the dataset with higher sample size. Therefore, seven gene expression datasets were included in this study,[4–10] constituting a total of 516 MPM tumor samples.

Data downloading and processing

Data downloading

RNA-seq and whole exome sequencing raw data from Bueno et al. dataset was downloaded from the European Genome-Phenome Archive[11] upon request to the data access committee. Data from Hmeljak et al. dataset was downloaded from GDC data portal[12] for somatic variants (MAF file with MuTect2 algorithm from data release 10.0). Copy number alterations was downloaded from cBioPortal/GDAC Firehose (level 3 data; GISTIC2 lesions). Finally, gene expression quantification was downloaded from TCGA2BED FTP repository[13] in transcripts per million (TPM).

Gene expression profiles from Lopez-Rios et al. dataset were downloaded from the link supplied in supplementary materials and those of the remaining four datasets (Suraokar et al., De Reyniès et al., Bott et al., and Gordon et al.) were downloaded from GEO and ArrayExpress repositories. When available, raw data was prioritized to process all samples homogenously.

RNA-seq data processing and quantification

To obtain gene expression profiles from Bueno et al. dataset, raw reads were processed using the pipeline described hereafter. First of all, multiple quality control checks on raw sequenced data were performed to check any potential issues using FastQC version 0.11.4 (Simon Andrews, 2010). Adapter removal and trimming of low-quality quality reads (Q<25) was then performed using Trimmomatic version 0.32.[14] Furthermore, a custom Python script (run with Python version 2.7.13) was used to remove reads with undetermined bases. Alignment of processed reads was performed using STAR version 2.5.3a[15] using GENCODE release 26 (GRCh38.p10)[16] as the reference genome. Quantification of aligned reads to TPM was done with RSEM version 1.3.0.[17]

Expression array processing

When raw expression data from Affymetrix platforms was available, datasets were processed using robust multiarray average algorithm[18] implemented in the affy package version 1.56[19] available through the Bioconductor software project.[20] Probe-set to gene mapping was done using BioMart web services via biomaRt R package version 2.34,[21] selecting the most expressed probe as representative of gene expression when multiple mapping probes occurred to avoid duplicated genes.

Somatic variant calling

In order to have a homogenous set of somatic variants between Bueno et al. and Hmeljak et al. study, Bueno et al. reads were reprocessed from raw sequencing data. Like for RNA-seq data, low-quality (Q>25) reads filtering, adapter trimming, and malformed reads removal was performed. Then, following Broad Institute’s best practices,[22,23] variant discovery analysis was done using Genome Analysis Toolkit (GATK) version 3.7-0.[24] Processed reads were aligned using the reference genome from NCBI (GRCh38) and BWA-MEM algorithm from Burrows-Wheeler Alignment tool version 0.7.15.[25] SAMtools version 1.3[26] was used to convert file from SAM to BAM format and the set of command line tools Picard release 2.9.2 (http://broadinstitute.github.io/picard) were used to sort, index, and mark duplicate reads. In order to detect and avoid systematic errors in base quality scores, a base quality score recalibration was performed using GATK. Then, using the MuTect2 tool from GATK, somatic variants were called via local re-assembly of haplotypes.[27] In more detail, StrandOddsRatio, DepthPerAlleleBySample, BaseQualitySumPerAlleleBySample, TandemRepeatAnnotator, and OxoGReadCounts annotations were added to the MuTect2 results and were used for variant filtering in a subsequent step.

To get a set of high confidence somatic variants and reduce false positive calls due to technical artefacts, a set of filters were applied as described hereunder. Regarding allele frequency, only somatic variants having 0% in normal samples and more than 5% in tumor samples were kept. Moreover, the variants must be supported by at least 10 reads in either normal and tumor samples and at least 5 reads with the alternative allele in tumor samples. SQSS quality score is the sum of base quality scores for each allele divided by its allele depth, and we required score of at least 25 in the alternative allele in tumor samples to consider the variant. A filter to detect strand bias was also set using the symmetric odds ratio test implemented in MuTect2, and variants with a score greater than 3 in the test were discarded. Finally, variants with oxidative DNA damage during sample preparation were discarded, following authors' recommendations.[28] The final set of variants were functionally annotated with the web server tool wANNOVAR.[29]

Genomic analyses

Mutational signatures

Mutational signatures are patterns in the occurrence of somatic single-nucleotide variants that can reflect underlying mutational processes. The R package *deconstructSigs* (v.1.8.0)[30] was used to infer the mutational signatures from exonic single-nucleotide variants data in Bueno et al. and Hmeljak et al. datasets.

Neoepitope prediction

Four-digit HLA types were determined for each sample with raw exome sequencing data available, using OptiType version 1.3.1.[31] Variants from genes with low expression levels (i.e., $\log_2(\text{TPM}) < 4$) were excluded from the input VCF file. Mutant peptide sequences of 15 amino acids were obtained using Ensembl protein sequence file from release 93 (GRCh38). NetMHCcons[32] was used with default parameters to assess binding affinity of mutant peptides according to HLA-types.

Methylation data

β values from Illumina's Infinium HumanMethylation450K BeadChip were downloaded from GDC Data Portal.[12] Probes were summarized to gene level using the median beta value and according to probe location surrounding transcription start sites up to 200 bp upstream.[33] Analysis of variance was performed for each gene adjusting for sex, age, stage, and histology covariates. FDR was applied to correct for multiple testing.

All downstream statistical analyses were done with the free software environment R version 3.5.0.[34]

REFERENCES

1 Barrett T, Wilhite SE, Ledoux P, *et al.* NCBI GEO: archive for functional

genomics data sets--update. *Nucleic Acids Res* 2013;**41**:D991-995. doi:10.1093/nar/gks1193

2 Kolesnikov N, Hastings E, Keays M, *et al.* ArrayExpress update--simplifying data submissions. *Nucleic Acids Res* 2015;**43**:D1113-1116. doi:10.1093/nar/gku1057

3 De Rienzo A, Richards WG, Yeap BY, *et al.* Sequential binary gene ratio tests define a novel molecular diagnostic strategy for malignant pleural mesothelioma. *Clin Cancer Res* 2013;**19**:2493–502. doi:10.1158/1078-0432.CCR-12-2117

4 Bueno R, Stawiski EW, Goldstein LD, *et al.* Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. *Nat Genet* 2016;**48**:407–16. doi:10.1038/ng.3520

5 Hmeljak J, Sanchez-Vega F, Hoadley KA, *et al.* Integrative Molecular Characterization of Malignant Pleural Mesothelioma. *Cancer Discov* 2018;**8**:1548–65. doi:10.1158/2159-8290.CD-18-0804

6 Suraokar MB, Nunez MI, Diao L, *et al.* Expression profiling stratifies mesothelioma tumors and signifies deregulation of spindle checkpoint pathway and microtubule network with therapeutic implications. *Ann Oncol* 2014;**25**:1184–92. doi:10.1093/annonc/mdu127

7 de Reyniès A, Jaurand M-C, Renier A, *et al.* Molecular classification of malignant pleural mesothelioma: identification of a poor prognosis subgroup linked to the epithelial-to-mesenchymal transition. *Clin Cancer Res* 2014;**20**:1323–34. doi:10.1158/1078-0432.CCR-13-2429

8 Bott M, Brevet M, Taylor BS, *et al.* The nuclear deubiquitinase BAP1 is commonly inactivated by somatic mutations and 3p21.1 losses in malignant pleural mesothelioma. *Nat Genet* 2011;**43**:668–72. doi:10.1038/ng.855

9 López-Ríos F, Chuai S, Flores R, *et al.* Global gene expression profiling of pleural mesotheliomas: overexpression of aurora kinases and P16/CDKN2A deletion as prognostic factors and critical evaluation of microarray-based prognostic prediction. *Cancer Res* 2006;**66**:2970–9. doi:10.1158/0008-5472.CAN-05-3907

10 Gordon GJ, Rockwell GN, Jensen RV, *et al.* Identification of novel candidate oncogenes and tumor suppressors in malignant pleural mesothelioma using large-scale transcriptional profiling. *Am J Pathol* 2005;**166**:1827–40. doi:10.1016/S0002-9440(10)62492-3

11 Lappalainen I, Almeida-King J, Kumanduri V, *et al.* The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet* 2015;**47**:692–5. doi:10.1038/ng.3312

12 Grossman RL, Heath AP, Ferretti V, *et al.* Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* 2016;**375**:1109–12. doi:10.1056/NEJMp1607591

13 Cumbo F, Fiscon G, Ceri S, *et al.* TCGA2BED: extracting, extending,

integrating, and querying The Cancer Genome Atlas. *BMC Bioinformatics* 2017;**18**:6. doi:10.1186/s12859-016-1419-5

14 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20. doi:10.1093/bioinformatics/btu170

15 Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. doi:10.1093/bioinformatics/bts635

16 Harrow J, Frankish A, Gonzalez JM, *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012;**22**:1760–74. doi:10.1101/gr.135350.111

17 Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;**12**:323. doi:10.1186/1471-2105-12-323

18 Irizarry RA, Hobbs B, Collin F, *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;**4**:249–64. doi:10.1093/biostatistics/4.2.249

19 Gautier L, Cope L, Bolstad BM, *et al.* affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004;**20**:307–15. doi:10.1093/bioinformatics/btg405

20 Huber W, Carey VJ, Gentleman R, *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* 2015;**12**:115–21. doi:10.1038/nmeth.3252

21 Durinck S, Spellman PT, Birney E, *et al.* Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 2009;**4**:1184–91. doi:10.1038/nprot.2009.97

22 Van der Auwera GA, Carneiro MO, Hartl C, *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;**43**:11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

23 DePristo MA, Banks E, Poplin R, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491–8. doi:10.1038/ng.806

24 McKenna A, Hanna M, Banks E, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303. doi:10.1101/gr.107524.110

25 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60. doi:10.1093/bioinformatics/btp324

26 Li H, Handsaker B, Wysoker A, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. doi:10.1093/bioinformatics/btp352

27 Cibulskis K, Lawrence MS, Carter SL, *et al.* Sensitive detection of somatic point

mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;**31**:213–9. doi:10.1038/nbt.2514

28 Costello M, Pugh TJ, Fennell TJ, *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* 2013;**41**:e67. doi:10.1093/nar/gks1443

29 Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet* 2012;**49**:433–6. doi:10.1136/jmedgenet-2012-100918

30 Rosenthal R, McGranahan N, Herrero J, *et al.* DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* 2016;**17**:31. doi:10.1186/s13059-016-0893-4

31 Szolek A, Schubert B, Mohr C, *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 2014;**30**:3310–6. doi:10.1093/bioinformatics/btu548

32 Karosiene E, Lundegaard C, Lund O, *et al.* NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 2012;**64**:177–86. doi:10.1007/s00251-011-0579-8

33 Sandoval J, Heyn H, Moran S, *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 2011;**6**:692–702. doi:10.4161/epi.6.6.16196

34 R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: : R Foundation for Statistical Computing 2017. https://www.R-project.org/