



## Article

<https://doi.org/10.1038/s41588-023-01321-1>

# Multiomic analysis of malignant pleural mesothelioma identifies molecular axes and specialized tumor profiles driving intertumor heterogeneity

Received: 25 December 2021

A list of authors and their affiliations appears at the end of the paper

Accepted: 26 January 2023

Published online: 16 March 2023

Check for updates

Malignant pleural mesothelioma (MPM) is an aggressive cancer with rising incidence and challenging clinical management. Through a large series of whole-genome sequencing data, integrated with transcriptomic and epigenomic data using multiomics factor analysis, we demonstrate that the current World Health Organization classification only accounts for up to 10% of interpatient molecular differences. Instead, the MESOMICS project paves the way for a morphomolecular classification of MPM based on four dimensions: ploidy, tumor cell morphology, adaptive immune response and CpG island methylator profile. We show that these four dimensions are complementary, capture major interpatient molecular differences and are delimited by extreme phenotypes that—in the case of the interdependent tumor cell morphology and adapted immune response—reflect tumor specialization. These findings unearth the interplay between MPM functional biology and its genomic history, and provide insights into the variations observed in the clinical behavior of patients with MPM.

Malignant pleural mesothelioma (MPM) is a rare and aggressive disease associated with asbestos exposure<sup>1</sup>. The World Health Organization (WHO) histological classification distinguishes three major types with prognostic value: epithelioid (MME), biphasic (MMB) and sarcomatoid (MMS)<sup>2</sup>. In the past decade, genomic studies uncovered molecular profiles (clusters) related to MPM's histopathological classification, each enriched for somatic alterations in known cancer genes (for example, *BAP1* in MME and *TP53* in MMS)<sup>3–5</sup>. We and others undertook unsupervised analyses of these data, revealing a molecular continuum of types that explained the prognosis of the disease more accurately than any reported discrete cluster<sup>6,7</sup>. MPM interpatient heterogeneity at the biological and clinical level is therefore expected to be sufficiently explained by the histopathological classification, with phenotypes ranging from MME to MMS<sup>8,9</sup>.

Nevertheless, the full extent of MPM phenotypes and the mechanisms by which they evolved are poorly understood. Histopathological

features (such as architectural subtypes) and molecular features (such as aneuploidy and immune infiltration) were shown to be independent of histopathological type<sup>8,9</sup>, suggesting that there are additional sources of heterogeneity that remain unexplained. In addition, although malignant transformation and cancer development can depend on a wide range of genomic aberrations<sup>10–12</sup>, genomic events have not been fully described in MPM as previous efforts have been restricted to profiling only exomes or a reduced representation of genomes<sup>3–5,13</sup>. As a result, biological functions performed by tumor cells, and the role of genomic events in shaping these functions, remain largely unknown, hindering any meaningful progress in the diagnosis, classification and treatment of the disease<sup>8</sup>.

We designed the MESOMICS study to uncover the main sources of molecular variation explaining MPM intertumoral heterogeneity, and to identify the underlying biological functions. Using multiomic

analyses combining genomic, transcriptomic and epigenomic data on a novel cohort of 120 MPM tumors (Supplementary Tables 1–3), we show that the current histopathological classification only explains a fraction of the molecular heterogeneity of the disease, while ploidy, adaptive immune response and CpG island methylation are as important. Taking advantage of a large cohort of whole-genome sequencing (WGS) data, we map the molecular landscape of 120 MPMs and elucidate the link between genotype and phenotype.

## Results

### Multomic analyses uncover four axes of molecular variation

We first found that the current histopathological classification only accounts for up to 10% of the interpatient molecular differences (2–10%, depending on the molecular layer, with an average of 6%), leaving 90% unexplained (Fig. 1a). We then undertook an unsupervised decomposition of the interpatient molecular heterogeneity using Multi-Omics Factor Analysis (MOFA)<sup>14</sup>, integrating genomic, transcriptomic and epigenomic data. We identified four independent and reproducible latent factors individually explaining more than 10% of molecular variation in at least one molecular layer, and collectively up to 61% of interpatient differences (19–61%, depending on the molecular layer, with an average of 33%; Fig. 1a, Extended Data Figs. 1–3, Supplementary Fig. 1 and Supplementary Tables 4–7). Only latent factor 2 (LF2) was associated with the histopathological classification, the recent artificial intelligence score based on digital pathology<sup>15</sup> and the previously proposed molecular classifications<sup>3–7</sup> (median  $q$  value =  $6.94 \times 10^{-11}$ ; Fig. 1b). Therefore, LF1, LF3 and LF4 capture three prominent sources of biological variation overlooked by previous histopathological and genomic studies.

LF1 (the ploidy factor) is largely explained by tumor ploidy ( $r = 0.87$ ; Fig. 1c,d). LF2 (the morphology factor) separates the main histopathological types and thus summarizes the morphological and related molecular classifications (Fig. 1a–c). LF3 (the adaptive response factor) summarizes immune infiltration with adaptive response effectors (lymphocytes) (Fig. 1c). For LF2 and LF3, enhancer methylation was the major molecular layer captured (Fig. 1a), partly explained by its implication in the tumor–immune interaction phenotype captured by LF3, and its variability in MPM samples is probably driven by cell-type heterogeneity (Supplementary Fig. 2 and Supplementary Tables 5, 6 and 8). The major feature captured by LF4 (the CpG island methylator phenotype (CIMP) factor) was methylation at gene body and promoter regions, and most of its molecular variation was strongly associated with the CIMP index ( $r = 0.92$ ; Fig. 1c,e). We then identified proxies to facilitate the interpretation of the latent factors and their implementation in the clinical setting: aneuploidy for LF1; the percentage of sarcomatoid component as reported by pathologists for LF2; an adaptive versus innate immune response score (Methods) for LF3; and a five-gene CIMP index proxy (Methods) for LF4. LF1, LF3, LF4 and their proxies were statistically independent of histopathological type (that is, all histological types can be either high or low ploidy, have high or low adaptive immune responses and have a high or low CIMP index), further confirming that these latent factors represent independent sources of molecular variation (Extended Data Fig. 4a–c).

In line with our previous observations<sup>6</sup>, tumor samples did not form clusters in MOFA but rather gradients between extreme molecular profiles (Fig. 1d,e). The ploidy factor ranged between a genomic near-haploidization (GNH) and a whole-genome doubling (WGD) profile, with a gradient of intermediate ploidies due to various levels of chromosome arm and focal amplifications and deletions (Fig. 1d). In contrast with the features found associated with the GNH subtype identified in the The Cancer Genome Atlas (TCGA) cohort<sup>4</sup>, the single near-haploid sample, MESO\_108, had a ploidy of 1.10, almost no copy-neutral loss of heterozygosity (LOH) (<1%) and no *SETDB1/TP53* mutations and did not undergo WGD. Therefore, this sample does not correspond to the GNH subtype as described by Hmeljak and colleagues<sup>4</sup>, but to another possible genomic trajectory, where genomic instability is driven by

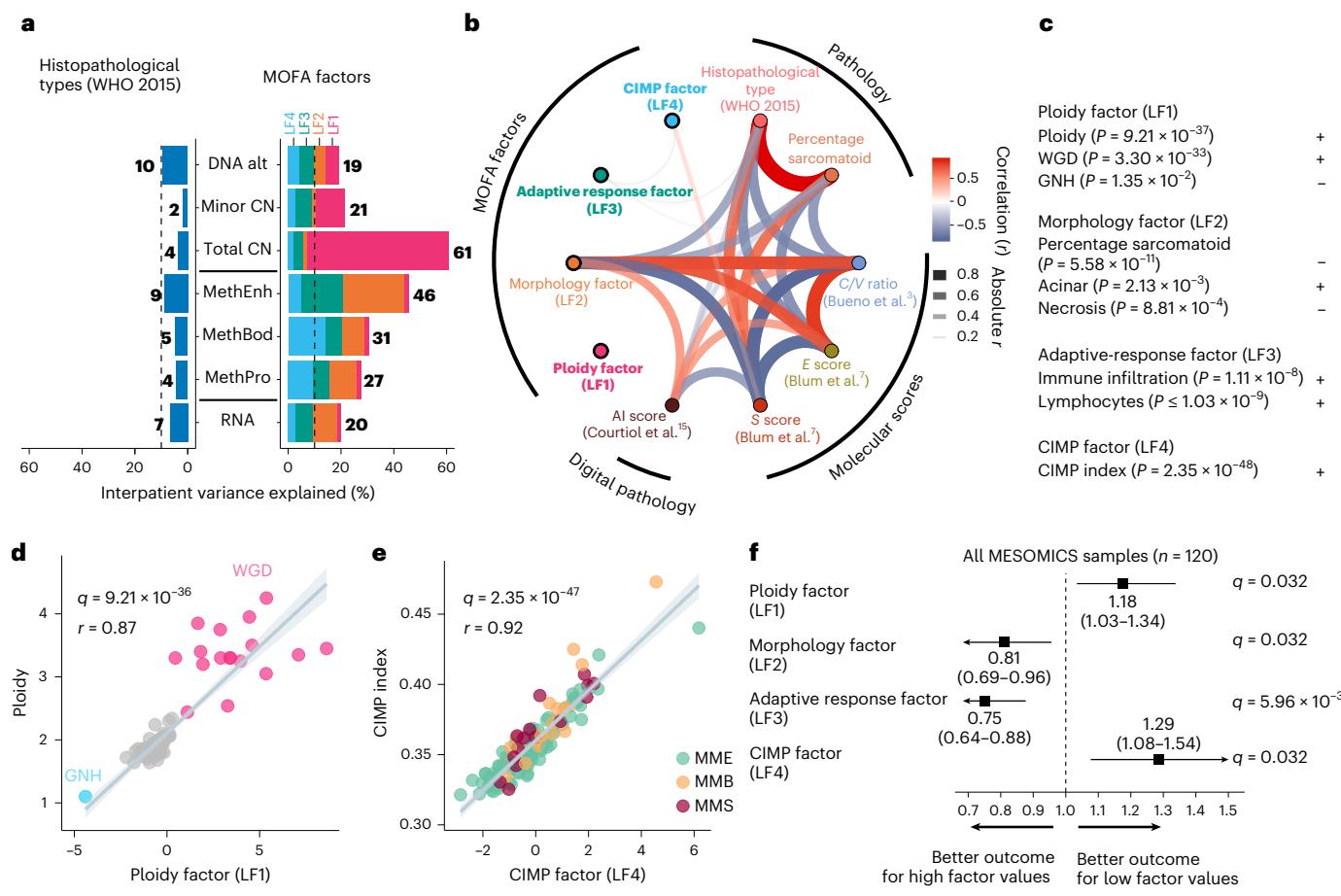
alternative pathways. Differential gene expression analyses showed that, as reported in other tumor types<sup>12</sup>, the most upregulated enriched pathway in WGD-positive (WGD<sup>+</sup>) versus WGD-negative (WGD<sup>-</sup>) cases was E2F targets ( $q$  value = 0.048; Supplementary Tables 9 and 10), although we could not replicate this result in the TCGA cohort<sup>4</sup>, possibly due to the difficulty of replicating such findings in low-sample-size series ( $n = 11$  WGD<sup>+</sup> samples). The CIMP factor also ranged between two extreme profiles: CIMP-low and CIMP-high (Fig. 1e). A well-known effect of the CIMP-high phenotype is epigenetic silencing of tumor suppressor genes<sup>16</sup>. In line with this, we identified five Catalogue of Somatic Mutations in Cancer (COSMIC) tumor suppressor genes<sup>17</sup>, whose expression was negatively correlated with both the CIMP index and the methylation level of their CpG island(s): *CBFA2T3*, *FBLN2*, *PRF1*, *SLC34A2* and *WT1* (median  $q$  value =  $2.6 \times 10^{-3}$ ; Supplementary Table 11).

We trained latent factor-based survival models and tested their performance over previously proposed prognostic factors to evaluate to what extent each latent factor captured variability predictive of prognosis (Methods). While individually they provided a prediction value similar to each other, when combining the four latent factors there was an increase in their area under the receiver operating characteristic curve value, suggesting that they capture molecular characteristics with independent prognostic value, being informative of MPM progression in a complementary manner (Extended Data Fig. 5, Supplementary Fig. 3 and Supplementary Tables 12–20). In line with evidence from multiple cancer types<sup>12</sup>, survival was lowest for the greatest ploidy (Fig. 1f). As expected, samples in the lower extreme of the morphology factor, enriched for sarcomatoid tumors, presented the worst prognosis. The adaptive response factor linked hot tumors (tumors with a high level of immune infiltration) with better survival, whereas CIMP-low tumors had better survival than CIMP-high tumors (Fig. 1f). The previously described proxies also demonstrated prognostic value in the MESOMICS cohort, and allowed for validation of the prognostic value of the latent factors in the validation cohorts (Extended Data Fig. 4d–g). Probably due to the limited power and a potential effect of histology, the prognostic value of the ploidy and CIMP factors was not statistically significant when analyzing MME samples only; however, their respective effect size remained similar to those identified in the entire cohort (Supplementary Fig. 3). We additionally validated the existence of the four dimensions as well as their prognostic values in previously published cohorts (Supplementary Tables 21 and 22).

Finally, combining molecular and drug response data for 59 MPM cell lines from Iorio et al.<sup>18</sup>, de Reyniès et al.<sup>5</sup> and Blum et al.<sup>7</sup>, we were able to evaluate the therapeutic value of the ploidy, morphology and CIMP factors (the lack of microenvironment in cell culture models did not allow for replication of the adaptive response factor), by assessing the impact that cell line position along each latent factor had on the response to candidate drugs (Extended Data Fig. 6, Supplementary Fig. 4 and Supplementary Tables 23–26). Significant drug responses associated with the different factors were entirely orthogonal (Extended Data Fig. 6a), highlighting the fact that MOFA latent factors capture independent axes of heterogeneity in both tumoral mechanisms and therapeutic responses. Therefore, both survival and cell line analyses showed that these axes of variation are clinically relevant and have the potential for translation into clinical practice.

### Task specialization analyses reveal diverse tumor strategies

Samples along the interdependent morphology and adaptive response factors formed a triangular shape delimited by three extremes (Fig. 2a and Supplementary Fig. 5). The well-established Pareto optimum theory<sup>19</sup> (ParetoTI method) predicted that this pattern results from natural selection for cancer tasks, with specialist tumors close to the vertices of the triangle and generalists in the center (triangle fit  $P$  value = 0.001; Fig. 2b). Integrative gene set enrichment analysis (IGSEA) pointed to the following cancer tasks and tumor phenotypes: cell division,



**Fig. 1 | MOFA of whole genomes, transcriptomes and methylomes of the MESOMICS cohort ( $n = 120$ ). a**, Proportion of interpatient variance within a molecular layer explained by WHO-defined histopathological type (left) and MOFA latent factors 1–4 (right). For example, 7% of variation present in RNA expression can be explained by mesothelioma types, in contrast with 20% explained by integrative MOFA. CN, segmental copy number; DNA alt, rearrangements and mutations; MethBod, DNA methylation level at body regions; MethEnh, DNA methylation level at enhancer regions; MethPro, DNA methylation level at promoter regions; RNA, gene expression level. **b**, Network of the correlations between latent factors, tumor histopathology and previously published molecular scores. The arc colors, widths and transparency correspond to Pearson correlation coefficients. Features uncorrelated with any

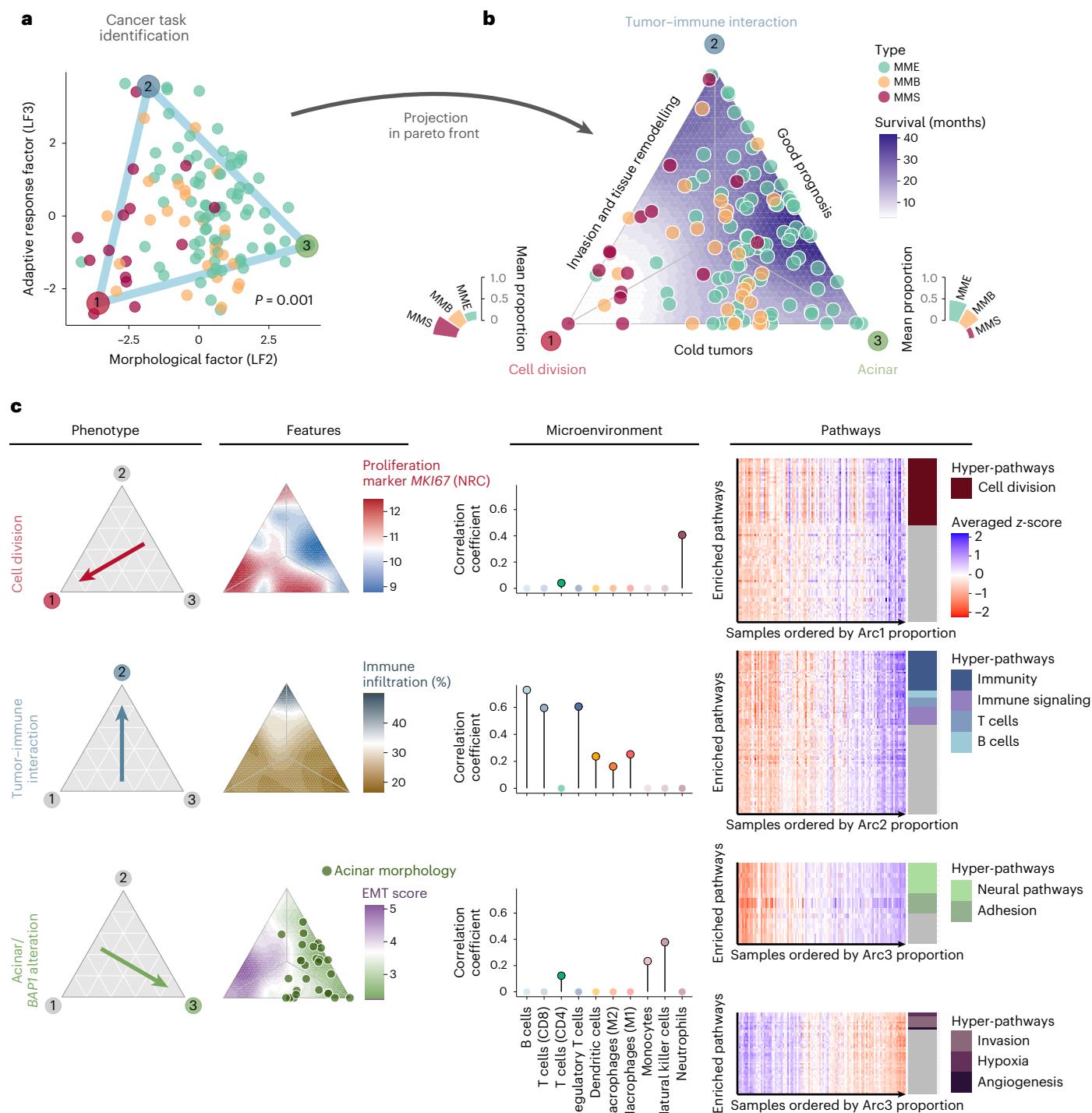
other features are highlighted in bold. AI, artificial intelligence; C/V ratio, log<sub>2</sub> ratio of *CLDN15* to *VIM* gene expression; S score, sarcomatoid gene expression score; E score, epithelioid gene expression score. **c**, Interpretation of MOFA latent factors. Plus signs indicate positive correlations and minus signs indicate negative correlations. **d**, Correlation between the ploidy factor (LF1) and ploidy. **e**, Correlation between the CIMP factor (LF4) and CIMP index. The samples are colored by histological type. **f**, Forest plot of the hazard ratios of MOFA latent factors for overall survival. The squares correspond to estimated hazard ratios and the segments correspond to their 95% confidence intervals. In **b–e**,  $P$  values,  $q$  values and  $r$  coefficients were determined by two-sided Pearson correlation tests. In **d** and **e**, the gray bands represent 95% confidence intervals.

tumor-immune interaction and acinar phenotype (Fig. 2c and Supplementary Tables 27–30 for archetypes, IGSEA significant pathways and  $q$  values).

Tumors specialized in the cell division task displayed upregulation of these pathways, as reported by Hausser et al. in multiple tumor types<sup>20</sup>. This phenotype was enriched for nonepithelial tumors and presented higher levels of necrosis, higher grade and a greater percentage of infiltrating innate immune response cells (neutrophils) (median  $q$  value = 0.005). Cell division specialization was supported by high expression levels of the proliferation marker *MKI67* and increased genomic instability (estimated from genomic, transcriptomic and epigenomic data; median  $q$  value =  $1.97 \times 10^{-4}$ ). Tumors specialized in the tumor-immune interaction task carried upregulated immune-related pathways, high expression of immune checkpoint genes and high immune infiltration with an enrichment for adaptive response cells: B lymphocytes, CD8<sup>+</sup> T cells and regulatory T cells (median  $q$  value =  $2.73 \times 10^{-3}$ ). The cell division and tumor-immune interaction specialists also showed high expression of hypoxia

response pathways and common enrichment for pathways in the invasion and tissue remodeling universal cancer task. Indeed, we found a higher epithelial-to-mesenchymal transition (EMT) score among tumors in this area of the Pareto triangle, driven by upregulation of mesenchymal genes and hypomethylation of their associated enhancers (median  $q$  value =  $1.61 \times 10^{-6}$ ). In line with in vitro studies showing that asbestos may induce EMT in MPM<sup>21</sup>, we found a positive correlation between the expression of mesenchymal genes and asbestos exposure score ( $r = 0.44$  and  $q$  value = 0.01) and a negative correlation between this score and enhancer methylation of mesenchymal genes ( $r = -0.33$  and  $q$  value = 0.02). We also observed overexpression of neoangiogenesis-related genes, corroborating the ability of these tumors to remodel their environment.

The last extreme phenotype was characterized by samples with acinar morphology, presenting a very structured tissue organization with epithelial cells tightly linked into tubular structures, and correlated with the presence of monocytes and natural killer cells (innate immune response cells) (median  $q$  value = 0.022). This phenotype presented



**Fig. 2 | Cancer task inference from the morphology and adaptive response factors ( $n = 120$ ).** **a**, Sample positions along the morphological (LF2) and adaptive response factors (LF3) are contained within a triangle formed by three phenotypic archetypes (colored vertices). The  $P$  value corresponds to a one-sided test from the Pareto fit. **b**, Ternary plot representing the sample's distance from the three specialized profiles. The bar plots represent the association between archetypes and histopathological types. **c**, Summary table of the main phenotypes, features and overexpressed pathways (columns) identified in each profile (rows). Left, arrows indicate the focal profile of each row. Middle left,

ternary plots with a color-filled background representing key features for each profile. NRC, normalized read count. Middle right, lollipop plots presenting the correlation between RNA-seq-estimated immune cell infiltration and the proportion of archetypes. Right, expression heatmaps of cancer tasks inferred from each phenotype. The rows represent enriched pathways and the columns represent the samples, ordered by increasing phenotype proportion. The heatmap color scale corresponds to the averaged z score of each gene set. The colored tiles on the right annotate the gene sets that belong to the hyper-pathways inferred from each phenotype.

the lowest EMT score, with overexpression of epithelial markers such as cell adhesion molecules (median  $q$  value =  $1.21 \times 10^{-3}$ ), corroborating the importance of tissue organization in this phenotype, and also low

levels of *MKI67* expression, indicating slow growth. This phenotype showed no particular tumoral specialization in any task based on the few IGSEA upregulated pathways. In line with the better prognosis

reported for this subtype<sup>8</sup>, the acinar phenotype is characterized by the highest levels of global methylation<sup>22</sup> ( $q$  value =  $5.58 \times 10^{-10}$ ). Altogether, these data provide a biological understanding of the molecular and phenotypic heterogeneity characteristic of MPM tumors.

### WGS uncovers a diverse genomic landscape

We found 97% (111/115) of MPM tumors harboring at least one large genomic event (copy number variant (CNV), amplicon, homologous recombination deficiency (HRD), chromothripsis or aneuploidy; Fig. 3a). As captured by the ploidy factor, MPM samples ranged from haploid to tetraploid (Fig. 1d). The average CNV profile was highly consistent between cohorts (Supplementary Fig. 6), with several recurrent chromosome arm-level CNVs, as well as focal alterations encompassing known cancer genes (Fig. 3b and Supplementary Tables 31–35). As previously reported<sup>23</sup>, all of the *MTA* alterations co-occurred with *CDKN2A/B* (Fig. 3a and Supplementary Tables 36 and 37). We also found recurrent deletions of a prominent immune recognition gene, *B2M* (chr15q14; Fig. 3b).

A comprehensive analysis of mutational signatures, encompassing single-base substitutions, CNVs and structural variants<sup>24,25</sup>, allowed us to identify the processes leading to particular somatic alteration patterns (Extended Data Fig. 7). A total of ten active single-base substitution signatures were detected in MPM genomes (Extended Data Fig. 7b); all corresponded to known COSMIC signatures and none was associated with asbestos exposure, as was previously reported<sup>3,4</sup>. Six tumors were found to have extrachromosomal DNA (ecDNA) (Supplementary Fig. 7 and Supplementary Table 38), and in the one sample with transcriptomic data we found increased expression of the genes predicted to be present on the ecDNA, including the known oncogene *BRIP1* (Fig. 3c). We observed that the aforementioned ecDNA sample co-occurred with, and may be fueled by, kataegis<sup>26</sup> (Supplementary Fig. 8). Overall, kataegis was rarely seen in our cohort, contributing to only 2% of the MPM clustered mutations (Supplementary Tables 39 and 40). The identified complex mutational processes included a pattern compatible with chromothripsis. This was observed in 20% of the samples (Fig. 3a, Supplementary Fig. 9 and Supplementary Table 39) and also at the transcriptomic level, as fusion transcripts, in half of the positive samples (Supplementary Fig. 10a and Supplementary Tables 41–43). A signature of clustered structural variants was detected and significantly associated with a high structural variant load and chromothripsis (Supplementary Fig. 10b,c and Supplementary Tables 41 and 42). For one sample (MESO\_019), the chromothripsis region overlapped with an ecDNA region, suggesting that chromothripsis may have been the source of the circular amplification (Fig. 3c). Finally, 23% of the samples showed a HRD phenotype, identified either by copy number signatures<sup>25</sup> or structural variant pattern-based methods<sup>27</sup> (Supplementary Fig. 11 and Supplementary Table 40). Among these samples, five harbored pathogenic germline mutations (from the ClinVar database) in one of 26 genes known to be involved in homologous recombination<sup>28</sup>—significantly more than the two mutations reported in the 77% of samples without HRD (Fisher's exact test,  $P$  value = 0.00587).

We detected an HRD signature in nine out of 21 MPM cell lines from Iorio et al.<sup>18</sup>, thus validating the high rate of this pattern in MPM.

In addition, the sensitivity of these cell lines to the clinically approved olaparib showed a tendency toward higher sensitivity in HRD samples compared with non-HRD samples (Supplementary Fig. 12). This may be linked with the results of a clinical trial suggesting a highly complex mechanism between the response to this drug and markers for DNA repair pathway activity<sup>29</sup>. Indeed, in contrast with their original hypothesis, patients with *BAP1* mutations had poorer survival when treated with olaparib than wild-type patients. In line with this observation, the olaparib response was positively associated with the prognostic CIMP index factor ( $r = 0.65$ ; Extended Data Fig. 6), meaning that CIMP-low samples were more sensitive to this poly-ADP ribose polymerase inhibitor than CIMP-high samples (which are enriched for *BAP1* alterations (Fig. 5a) and associated with poorer survival (Supplementary Fig. 3)).

Despite the low mutational rate (0.98 nonsynonymous small variants per megabase; Supplementary Fig. 13a and Supplementary Tables 44–46), MPM tumors carry a particularly high number of structural variants relative to tumors with similarly low mutational burden (Fig. 4 and Supplementary Fig. 13b). The top genes altered by structural variants ( $\geq 5\%$ ) were *RBFOX1*, *NF2*, *BAP1*, *MTAP* and *PCDH15* (Supplementary Fig. 14a). For *RBFOX1*, 13 out of 39 samples have two separate events, with most deleting part of the RNA-binding protein domain (Supplementary Fig. 14b). Many of these genomic rearrangements resulted in fusion transcripts detected at the transcriptomic level (Supplementary Figs. 10a and 15).

Combining the MESOMICS dataset with the two other large datasets from Bueno et al.<sup>3</sup> and the TCGA<sup>4</sup>, we reached the sample size ( $n \approx 300$ ) needed to detect rare driver alterations (1%). The IntOGen pipeline<sup>30</sup> discovered 30 MPM driver genes based on small variants (Supplementary Fig. 14c). *BAP1*, *NF2*, *SETD2*, *TP53* and *LATS2* are all known MPM driver genes. Among the other 25 genes, some were previously reported as recurrently mutated in MPM (*PBRM1*, *KMT2D*, *DDX3X*, *PIK3CA*, *FBXW7*, *MGA*, *NFI*, *SETDB1*, *MYH9*, *PTCH1*, *RHOA* and *TRAF7*)<sup>31–33</sup> or altered by structural variants (*PTPRD* and *LRP1B*)<sup>34</sup>, two were found overexpressed in MPM cell lines (*DNMT3B* and *EZH2*)<sup>35</sup> and, for another two, germline mutations have been discovered, suggesting genetic susceptibility (*NCOR1* (ref. <sup>36</sup>) and *MYO5A*<sup>37</sup>). The remaining seven driver genes have, to our knowledge, not been previously reported in MPM, but they are all known cancer genes, as reported in COSMIC: *FAT3*, *NN*, *ARHGAP5*, *HLA-A*, *NCOR2*, *SRGAP3* and *WNK2*. Of note, *NF2* and *MYH9* (IntOGen drivers) are located within the significantly deleted chr22q region, along with *TTC28*—a gene frequently altered by structural variants (Figs. 3a,b and 4). Beyond extending the list of putative MPM drivers, combining point mutations with structural variants allowed for refinement of the frequency of alterations in key MPM genes (Fig. 4 and Supplementary Tables 41–46).

### Genomic alterations tune the molecular profiles of MPM

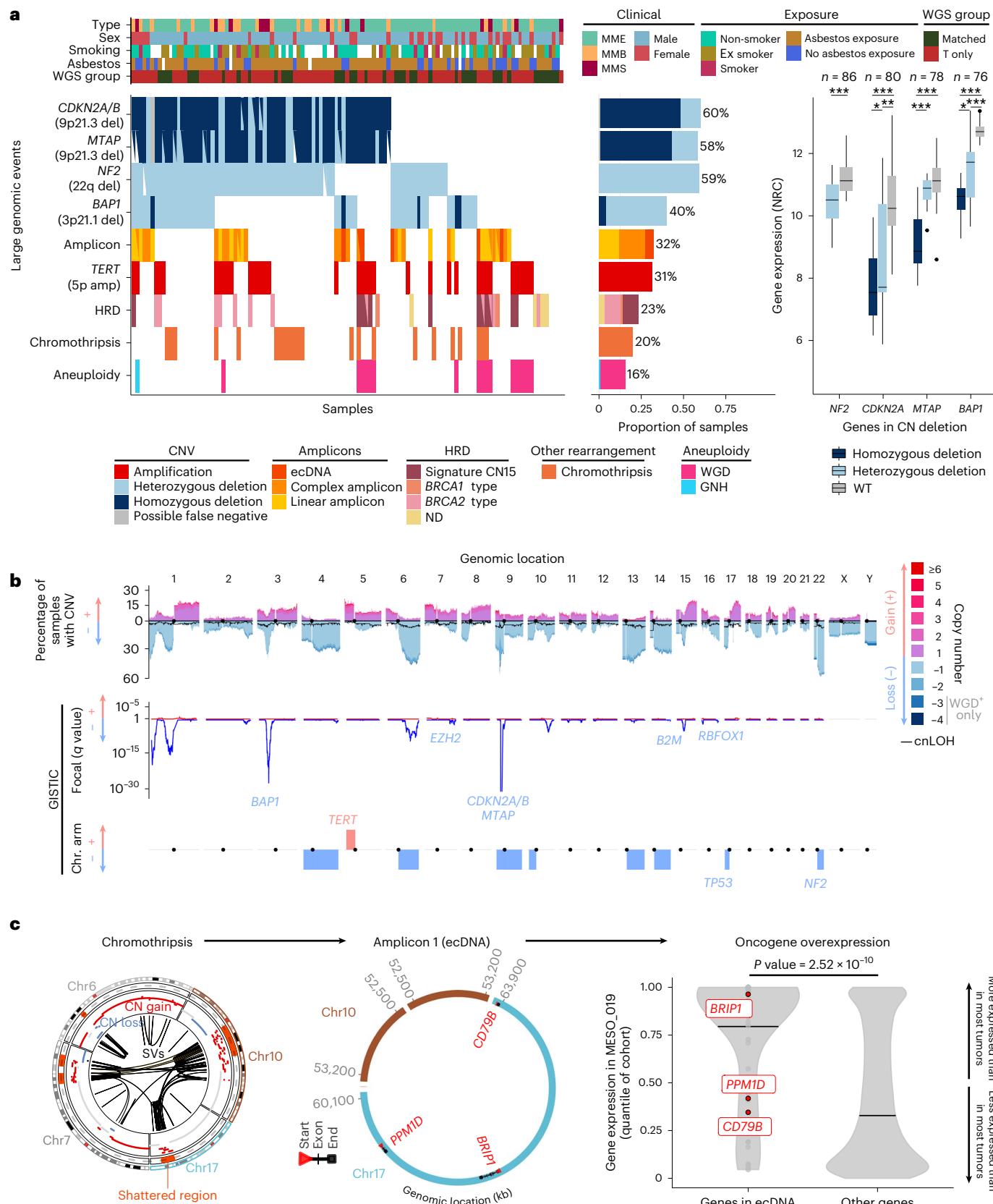
Genomic events were associated with all MOFA latent factors and the extreme profiles that they encapsulated, as well as with the phenotypic specialists captured by the morphology and adaptive response factors (Fig. 5a and Supplementary Tables 47 and 48). Associated alterations significantly tuned tumor specialization ( $P$  value = 0.003; Methods and Extended Data Fig. 8). In addition to ploidy, *NCOR2* alterations

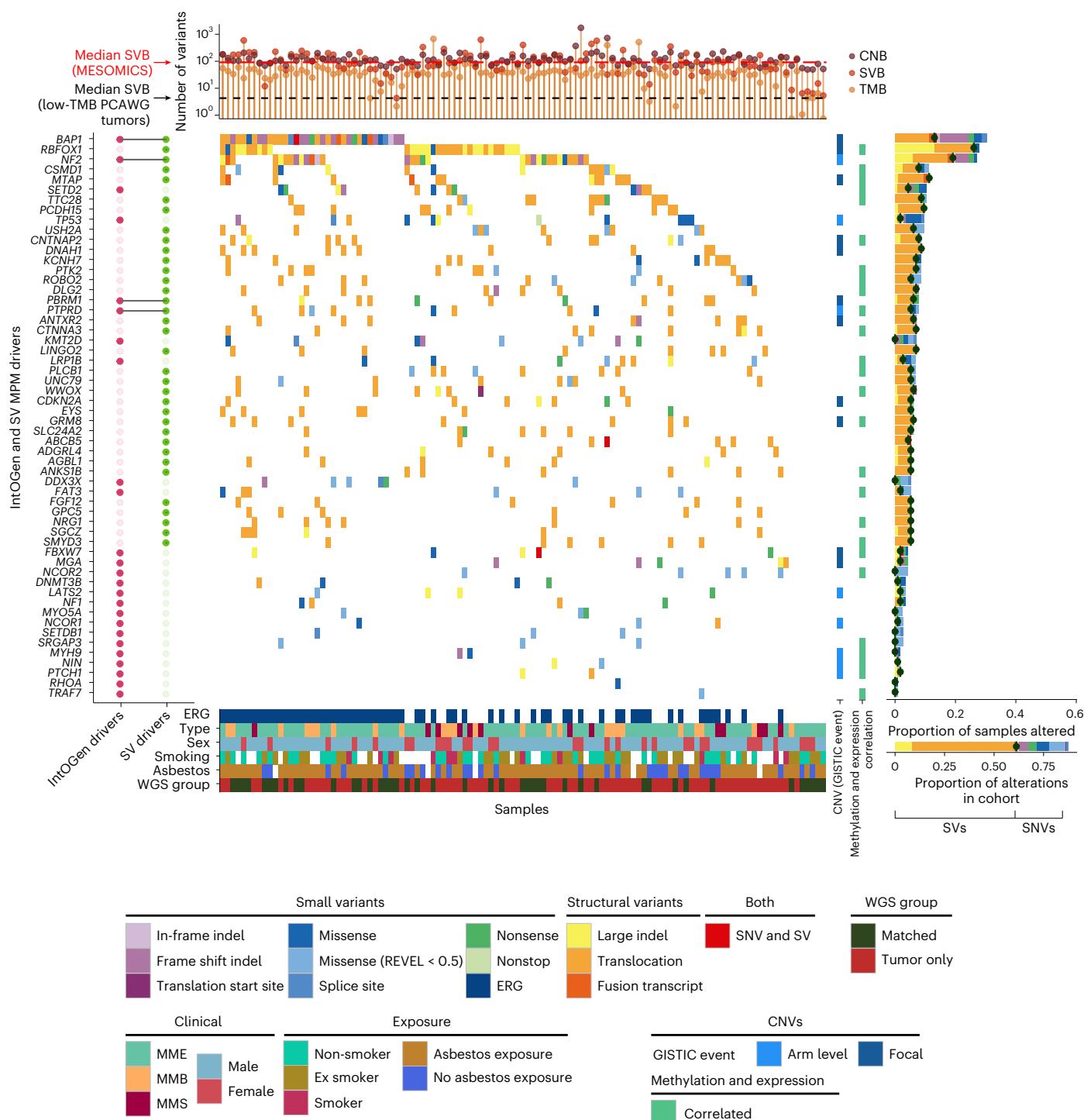
**Fig. 3 | Genomic characterization of MPM from the MESOMICS cohort.** **a**, Recurrent large genomic events. Top, clinical, epidemiological, morphological and technical features per sample. T only represents samples with WGS on the tumor sample only. Bottom left, oncoplot describing the genomic events per sample. amp, amplification; del, deletion; ND, HRD type not determined. Bottom middle, barplot of the frequency of each event within the cohort. Bottom right, comparison of the gene expression of cancer-relevant genes belonging to frequent deletions detected by GISTIC, with regards to their copy number (CN) status. Wild-type (WT) cases correspond to samples without copy number, structural variant or single-nucleotide variant events detected. The box plots represent the median and interquartile range and the whiskers the maximum and minimum values, excluding outliers. The  $n$  number above represents the

number of biologically independent samples for each test. \* $0.01 < q$  value  $\leq 0.05$ ; \*\* $0.001 < q$  value  $\leq 0.01$ ; \*\*\* $q$  value  $\leq 0.001$ . NRC, normalized read count. **b**, Cohort-level copy number profile (top), with significantly altered regions identified by GISTIC in focal peaks (middle) and at the chromosome (chr.) arm level (bottom). cnLOH, copy-neutral LOH. **c**, Data from a patient with oncogene amplification due to a chromothripsis event (MESO\_019). Left, chromosomes involved in the chromothripsis event (outer circle, shattered regions; intermediate circle, copy number gain and loss; inner circle, structural variants (SVs)). Middle, reconstructed ecDNA structure. Right, gene expression in MESO\_019 relative to the expression in other tumors of the cohort (quantile). Oncogenes found within the ecDNA region are represented in red. The  $P$  value was determined by two-sided Wilcoxon rank-sum test. kb, kilobases.

and *TERT* amplification were associated with the ploidy factor ( $q$  values =  $4.3 \times 10^{-18}$  and  $3.3 \times 10^{-4}$ , respectively; Fig. 5a). Thirty-six samples (31%) displayed *TERT* amplification, resulting in a significant increase in *TERT* expression ( $P$  value =  $1.8 \times 10^{-5}$ ; Supplementary Fig. 16a,b). *TERT* amplification was accompanied by an underlying amplification of

chr5p in 81% of the positive cases. While no association was previously detected between *TERT* promoter mutations and WGD<sup>38</sup>, here we found that both *TERT* amplification and its increased expression were associated with WGD events ( $P$  value =  $1.6 \times 10^{-10}$  and 0.009, respectively; Supplementary Fig. 16c).



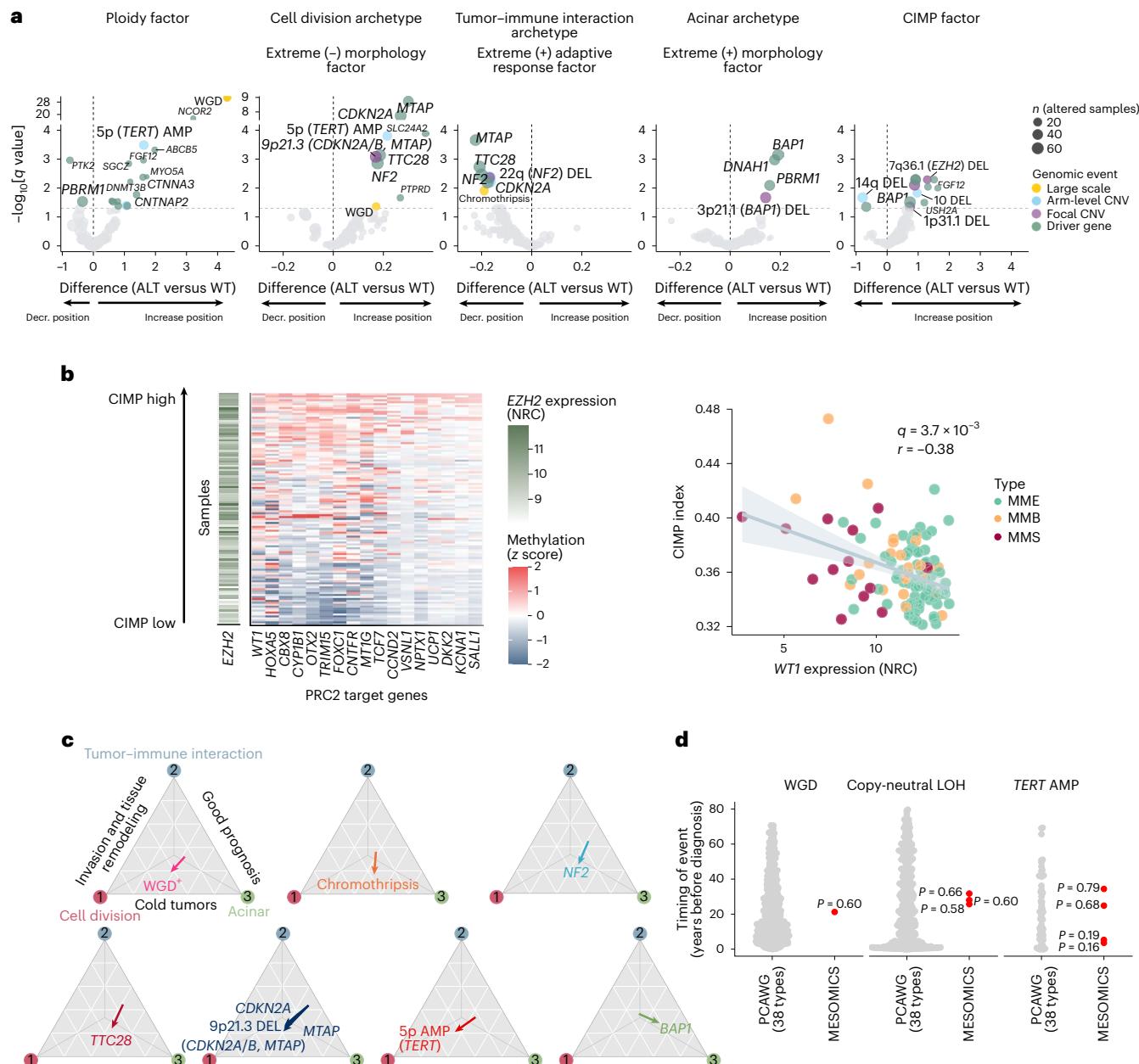


**Fig. 4 | MPM driver genes in the MESOMICS cohort.** Top, tumor mutational burden (TMB), number of segments or copy number burden (CNB) and structural variant burden (SVB) of each sample. Main, oncoplot describing genomic alterations in IntOGen and structural variant MPM driver genes per sample. These genomic events can co-occur with copy number changes. Large indels and translocations refer to structural variant events detected by structural variant callers while fusion transcripts are detected at the transcriptomic level. Each gene

is also annotated as belonging to one focal or arm-level GISTIC event, as well as for being regulated by DNA methylation (right bars). Right, frequency of alterations within the cohort. For each gene, the dark green dot represents the frequency of structural variants. In the legend, ERG indicates whether the sample has one or more alteration in an ERG. Key clinical, epidemiological, morphological and technical features are given for each sample. PCAWG, Pan-Cancer Analysis of Whole Genomes; SNV, single-nucleotide variant; SV, structural variant.

Genomic alterations in epigenetic regulatory genes (ERGs) have previously been shown to drive CIMP in cancer<sup>39</sup>. In line with this, we found enrichment for ERGs ( $P$  value =  $3.4 \times 10^{-3}$ ; Methods and Supplementary Fig. 17), including the mesothelioma drivers *NCOR2* and *EZH2*, among the genes highly expressed in CIMP-high tumors, and more

generally in the list of MPM drivers ( $q$  value =  $2.1 \times 10^{-5}$ ). Chr7q36.1del, encompassing *EZH2*, further tuned the position of the samples along the CIMP factor ( $q$  value =  $5.2 \times 10^{-3}$ ; Fig. 5a). *EZH2* (enhancer of zeste homolog 2) is a histone methyltransferase that functions as part of the Polycomb repressive complex 2 (PRC2) complex to promote gene



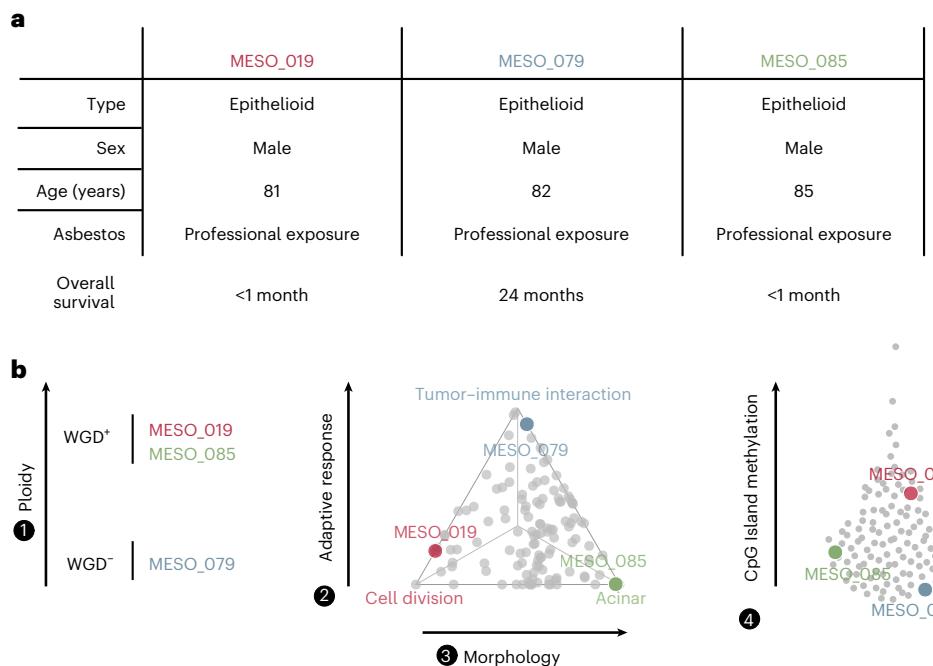
**Fig. 5 | Impact of genomic events on MPM molecular profiles.** **a**, Association between genomic events and MOFA factors. For each event, the ALT (altered) versus wild-type difference corresponds to the difference between the mean factor value of wild-type samples and that of altered samples. The  $q$  values correspond to an adjusted analysis of variance  $P$  value; the dashed horizontal line represents the  $q$  value threshold of 0.05. AMP, amplification; Decr., decrease; DEL, deletion. **b**, Association between CIMP index, EZH2 expression ( $n = 109$  samples) and PRC2 target gene methylation ( $n = 119$  samples). Left, heatmap of EZH2 gene expression (NRC) and CpG island methylation ( $z$  score) of PRC2 target genes whose methylation level was significantly positively correlated with CIMP index

( $q < 0.05$ ), for samples ordered by CIMP index. Right: correlation between  $WT1$  expression and CIMP index. The  $q$  value was determined by Pearson correlation test and the gray band corresponds to the 95% confidence interval. **c**, Effect vector of key alterations affecting specialization in the tumor tasks from Fig. 2. The effect vector corresponds to the difference in position on the Pareto front between the centroid of altered samples and the centroid of wild-type samples. **d**, Comparison of the timing of large-scale amplifications in the MESOMICS and PCAWG cohorts. The points represent estimates of the timing of genomic events. The empirical  $P$  values (red data points) were determined by one-sided outlier tests.

silencing of specific targets<sup>40</sup>. Indeed, genes whose CpG island methylation level was highest in CIMP-high tumors were enriched for PRC2 target genes ( $P$  value = 0.01; Fig. 5b).  $WT1$ , which is found downregulated in CIMP-high tumors, is particularly interesting and a vaccine against this PRC2 target is currently being assessed in clinical trials for mesothelioma<sup>41</sup>. Cancers frequently associated with a CIMP-high phenotype include colorectal cancer (CRC) and glioma<sup>42,43</sup>, with  $BRAF$  (CRC) and  $IDH1$  (glioma) mutations also associated with this phenotype, as well

as with microsatellite instability in CRC<sup>42</sup>. Microsatellite instability and  $BRAF/IDH1$  mutations were rare or absent events in our series and unrelated to the CIMP phenotype (Supplementary Tables 7, 44 and 49), suggesting that the mutational processes linked with CIMP phenotype in MPM may differ from those of other cancers.

WGD and chromothripsis seemed to push tumors away from the tumor–immune interaction phenotype ( $q$  values = 0.042 and 0.012, respectively; Fig. 5c); indeed, both cell division and acinar phenotypes



**Fig. 6 | Added value of the four-factor molecular classification in understanding intertumor heterogeneity in three example patients.** **a**, Patients MESO\_019, MESO\_079 and MESO\_085 had nearly identical clinical characteristics.

**b**, The three patients had vastly different profiles based on our four-factor morphomolecular classification: different WGD status (left), opposite positions on the Pareto front (middle) and variable levels of CpG island methylation (right).

were characterized by low immune cell infiltration (cold tumors), which may be explained by the downregulation of the interferon response pathway and *B2M* expression seen in WGD<sup>+</sup> MPM tumors ( $q$  value =  $7.4 \times 10^{-17}$ ; Supplementary Fig. 18a,b,e and Supplementary Tables 9 and 10). These may represent important mechanisms for WGD<sup>+</sup> tumors to avoid the immune response<sup>12,44</sup>. Chromothripsis has also been associated with low immune infiltration as part of the chromosomal chaos that silences immune surveillance<sup>45</sup>.

*CDKN2A*, *MTAP* and *NF2* alterations also converged on cold tumors (median  $q$  value = 0.003). Within this cold phenotype, *TERT* amplification and alterations in *TTC28*, involved in the mitotic cell cycle, moved tumors towards cell division specialization ( $q$  values =  $1.6 \times 10^{-4}$  and  $7.4 \times 10^{-4}$ , respectively; Fig. 5c), whereas chr3p21.1del (*BAP1*, *DNAH1* and *PBRM1*) and *BAP1* mutations moved tumors toward the better-prognosis acinar phenotype ( $q$  values = 0.021 and  $7.1 \times 10^{-4}$ , respectively; Fig. 5c), as expected given the previously reported association between *BAP1* alterations and better survival in MPM<sup>36</sup>. A loss of *BAP1* (BRCA1-associated protein-1) expression, measured by immunohistochemistry, was also associated with this phenotype ( $r = -0.38$  and  $q$  value =  $4.61 \times 10^{-5}$ ; Supplementary Fig. 19). Interestingly, an analysis of splicing variation found that the morphology factor and acinar phenotype were significantly associated with alternative splicing events (Supplementary Fig. 20a–f). Major contributions came from events in cell adhesion genes, and neuronal progenitor BAF, neuron-specific BAF and SWI/SNF complexes, potentially affecting the alternative splicing pattern of genes such as *BCL11A* and *SMARCE1* (Supplementary Fig. 20g,h). The fact that these genes (just like *BAP1*) have important roles in chromatin remodeling suggests that disruption of chromatin remodeling pathways may molecularly define the acinar phenotype.

The specialization of tumors can be influenced by early genomic events. Estimates of the timing of WGD, *TERT* amplification and copy-neutral LOH in the few samples ( $n = 6$ ) with such events where a subclonal deconvolution was possible showed that our samples fall well within the values observed across >2,500 tumors of the Pan-Cancer Analysis of Whole Genomes Consortium<sup>46</sup> (empirical  $P$  values = 0.16–0.79; Fig. 5d and Supplementary Fig. 21). Thus, these genomic events

may indeed have occurred more than 10 years before diagnosis. Three out of the six patients were exposed to asbestos (of the other three patients, two had no known exposure and one had unknown exposure), among whom two had well-documented periods of exposure, from 56 to 21 years before diagnosis for MESO\_048 (including the estimated timing of LOH) and from 54 to 50 years before diagnosis for MESO\_057, more than 50 years before the estimated timing of *TERT* amplification, suggesting that genomic events can occur both concomitantly with and subsequent to asbestos exposure, although conclusive evidence of the timing of these alterations will need to be investigated in hypothesis-driven studies. Using a multiregional subcohort from 13 patients, we found intratumor heterogeneity in all factors except the ploidy factor, further suggesting that genomic events are mostly early and thus do not vary much across regions (Extended Data Fig. 9, Supplementary Fig. 22 and Supplementary Tables 50–52). Finally, we detected neutral tumor evolution close to the acinar phenotype ( $P$  value = 0.0024; Supplementary Fig. 23) at extreme values of the morphology and adaptive response factor, suggesting that tumors with this profile were even less influenced by recent genomic events.

## Discussion

The MESOMICS project represents a substantial advancement toward the comprehensive molecular characterization of MPM, made possible by inclusion of a large WGS dataset<sup>3,4,34</sup> and by the depth of the multiomic integrative analyses undertaken. We demonstrated that ploidy, adaptive immune response and CpG island methylation constitute independent sources of molecular variation with quantitatively similar impacts on interpatient MPM heterogeneity as the histological classification. Despite some individual observations made in previous studies<sup>6,7,13</sup>, these three sources of molecular variation have been mostly unexplored or unknown because of the major focus that was put on refining the histological groups, and the lack of comprehensive analysis of a large multiomics dataset. In this sense, the unifying framework aspect of our research approach allowed us to capture the entire molecular landscape of MPM, summarized in four dimensions.

Aneuploidy is one of the morphology-independent features previously reported in MPM<sup>4</sup> but poorly characterized. The ploidy factor identified tumors that underwent WGD, previously described in multiple cancer types as an early transformative event that dramatically destabilizes cell genetics and fuels tumor development<sup>47</sup>. WGD tends to be favored along the evolutionary course of low-mutational-burden tumors like MPM<sup>12</sup> and is suspected to serve as a genetic spare tire in case of lethal alterations<sup>48</sup>. As a consequence, this event shapes the cellular phenotype associated with specific vulnerabilities<sup>12</sup>.

The CIMP has been reported in several cancer types, most notably CRC and glioblastoma, with inconsistent associations with survival<sup>49–51</sup>. Here we provide further evidence, to that of Blum et al.<sup>7</sup>, of distinct variation in CIMP index within mesothelioma tumors, and have shown that a high CIMP index is independent of morphology and predictive of poorer outcome. While a universal cause for a CIMP-high phenotype has not been established, it has been previously associated with alterations in ERGs<sup>39,52</sup>. Indeed, our data suggest that some mesothelioma tumors may acquire a CIMP-high phenotype through the activity of the ERG EZH2, to hypermethylate and silence specific target genes. Such a strategy may be warranted to promote malignant transformation in a lowly mutated tumor such as mesothelioma<sup>35</sup>.

Pareto task inference uncovered three specialized tumor profiles in the space delimited by the interdependent morphology and adaptive response factors, presumably resulting from pressures of the microenvironment, each selecting for adaptive alterations and phenotypic traits. Cell division specialists adopted a fast reproduction strategy that was expected to result from unfavorable and unpredictable environments<sup>53</sup>, with their genomic instability suggesting adaptation through evolutionary leaps<sup>54,55</sup>. Immune interaction specialists adopted an immune evasion or camouflage strategy. Both phenotypes also presented characteristics of invasion and tissue remodeling specialists<sup>20</sup>. These tumors tended to occur in intensely asbestos-exposed individuals, suggesting that chronic inflammation (promoted by asbestos exposure<sup>56</sup>) may have created the unfavorable environment responsible for selective pressure. Finally, acinar phenotype specialists adopted a structured tissue organization and slow growth strategy. This suggests an equilibrium strategy that is expected to be favorable in stable, resource-rich environments with limited predation<sup>57</sup>, in line with the lower level of asbestos exposure and limited inflammation and immune infiltration observed in these tumors. Consistent with limited environmental pressures, acinar tumors were enriched for neutral evolution and *BAP1* alterations—an event that, when combined with weak asbestos exposure in mice, greatly increased mesothelioma occurrence over weak asbestos exposure alone<sup>58</sup>.

Overall, the four molecular factors are highly informative and capture specific profiles that are complementary in predicting tumor phenotype and aggressiveness. The fact that they are all independent and mostly unrelated to the morphology factor (histology) means that disregarding them might not only jeopardize the success of any treatment but also miss opportunities to stratify patients based on their molecular profile (Fig. 6). The tightly correlated proxies that we have identified could serve as biomarkers for response to specific therapies (such as immunotherapy for LF3) and could be easily tested in a hypothesis-driven study design. Subsequently, integrating these complementary factors would help to stratify patients for preselected-cohort clinical trials<sup>59</sup>, a process that has proven to be beneficial in small-cell lung cancer, another aggressive recalcitrant cancer<sup>60–62</sup>. The results of the MESOMICS project pave the way for the establishment of a more clinically relevant morpho-molecular classification of MPM tumors.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-023-01321-1>.

## References

- Carbone, M. et al. Mesothelioma: scientific clues for prevention, diagnosis, and therapy. *CA Cancer J. Clin.* **69**, 402–429 (2019).
- WHO Classification of Tumours, Thoracic Tumours (5th edn) (International Agency for Research on Cancer, 2020).
- Bueno, R. et al. Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. *Nat. Genet.* **48**, 407–416 (2016).
- Hmeljak, J. et al. Integrative molecular characterization of malignant pleural mesothelioma. *Cancer Discov.* **8**, 1548–1565 (2018).
- De Reyniès, 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.* **20**, 1323–1334 (2014).
- Alcalá, N. et al. Redefining malignant pleural mesothelioma types as a continuum uncovers immune–vascular interactions. *EBioMedicine* **48**, 191–202 (2019).
- Blum, Y. et al. Dissecting heterogeneity in malignant pleural mesothelioma through histo-molecular gradients for clinical applications. *Nat. Commun.* **10**, 1333 (2019).
- Nicholson, A. G. et al. EURACAN/IASLC proposals for updating the histologic classification of pleural mesothelioma: towards a more multidisciplinary approach. *J. Thorac. Oncol.* **15**, 29–49 (2020).
- Fernandez-Cuesta, L., Mangiante, L., Alcalá, N. & Foll, M. Challenges in lung and thoracic pathology: molecular advances in the classification of pleural mesotheliomas. *Virchows Arch.* **478**, 73–80 (2021).
- Cortés-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* **52**, 331–341 (2020).
- ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
- Quinton, R. J. et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* **590**, 492–497 (2021).
- Creaney, J. et al. Comprehensive genomic and tumour immune profiling reveals potential therapeutic targets in malignant pleural mesothelioma. *Genome Med.* **14**, 58 (2022).
- Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
- Courtial, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
- Baylin, S. B. & Jones, P. A. Epigenetic determinants of cancer. *Cold Spring Harb. Perspect. Biol.* **8**, a019505 (2016).
- Sondka, Z. et al. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
- Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
- Haussler, J. & Alon, U. Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nat. Rev. Cancer* **20**, 247–257 (2020).
- Haussler, J. et al. Tumor diversity and the trade-off between universal cancer tasks. *Nat. Commun.* **10**, 5423 (2019).
- Turini, S., Bergandi, L., Gazzano, E., Prato, M. & Aldieri, E. Epithelial to mesenchymal transition in human mesothelial cells exposed to asbestos fibers: role of TGF-β as mediator of malignant mesothelioma development or metastasis via EMT event. *Int. J. Mol. Sci.* **20**, 150 (2019).
- Shipony, Z. et al. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–119 (2014).

23. Chapel, D. B. et al. MTAP immunohistochemistry is an accurate and reproducible surrogate for CDKN2A fluorescence in situ hybridization in diagnosis of malignant pleural mesothelioma. *Mod. Pathol.* **33**, 245–254 (2020).
24. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
25. Steele, C. D. et al. Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022).
26. Bergstrom, E. N. et al. Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. *Nature* **602**, 510–517 (2022).
27. Ladan, M. M., van Gent, D. C. & Jager, A. Homologous recombination deficiency testing for BRCA-like tumors: the road to clinical validation. *Cancers* **13**, 1004 (2021).
28. Toh, M. & Ngeow, J. Homologous recombination deficiency: cancer predispositions and treatment implications. *Oncologist* **26**, e1526–e1537 (2021).
29. Ghafoor, A. et al. Phase 2 study of olaparib in malignant mesothelioma and correlation of efficacy with germline or somatic mutations in BAP1 gene. *JTO Clin. Res Rep.* **2**, 100231 (2021).
30. Martínez-Jiménez, F. et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
31. De Rienzo, A. et al. Gender-specific molecular and clinical features underlie malignant pleural mesothelioma. *Cancer Res.* **76**, 319–328 (2016).
32. Kato, S. et al. Genomic landscape of malignant mesotheliomas. *Mol. Cancer Ther.* **15**, 2498–2507 (2016).
33. Shukuya, T. et al. Identification of actionable mutations in malignant pleural mesothelioma. *Lung Cancer* **86**, 35–40 (2014).
34. Mansfield, A. S. et al. Neoantigenic potential of complex chromosomal rearrangements in mesothelioma. *J. Thorac. Oncol.* **14**, 276–287 (2019).
35. McLoughlin, K. C., Kaufman, A. S. & Schrump, D. S. Targeting the epigenome in malignant pleural mesothelioma. *Transl. Lung Cancer Res.* **6**, 350–365 (2017).
36. Pastorino, S. et al. A subset of mesotheliomas with improved survival occurring in carriers of BAP1 and other germline mutations. *J. Clin. Oncol.* **36**, 3485–3494 (2018).
37. Hylebos, M. et al. Molecular analysis of an asbestos-exposed Belgian family with a high prevalence of mesothelioma. *Fam. Cancer* **17**, 569–576 (2018).
38. Bielski, C. M. et al. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**, 1189–1195 (2018).
39. Turcan, S. et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479–483 (2012).
40. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343–349 (2011).
41. Zauderer, M. G. et al. A randomized phase II trial of adjuvant galinpepimut-S, WT-1 analogue peptide vaccine, after multimodality therapy for patients with malignant pleural mesothelioma. *Clin. Cancer Res.* **23**, 7483–7489 (2017).
42. Phipps, A. I. et al. Association between molecular subtypes of colorectal cancer and patient survival. *Gastroenterology* **148**, 77–87.e2 (2015).
43. Malta, T. M. et al. Glioma CpG island methylator phenotype (G-CIMP): biological and clinical implications. *Neuro. Oncol.* **20**, 608–620 (2018).
44. Sreejit, G. et al. The ESAT-6 protein of *Mycobacterium tuberculosis* interacts with beta-2-microglobulin (β2M) affecting antigen presentation function of macrophage. *PLoS Pathog.* **10**, e1004446 (2014).
45. Zanetti, M. Chromosomal chaos silences immune surveillance. *Science* **355**, 249–250 (2017).
46. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
47. Fujiwara, T. et al. Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. *Nature* **437**, 1043–1047 (2005).
48. López, S. et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat. Genet.* **52**, 283–293 (2020).
49. Advani, S. M. et al. Clinical, pathological, and molecular characteristics of CpG island methylator phenotype in colorectal cancer: a systematic review and meta-analysis. *Transl. Oncol.* **11**, 1188–1201 (2018).
50. Noushmehr, H. et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510–522 (2010).
51. Hughes, L. A. E. et al. The CpG island methylator phenotype: what's in a name? *Cancer Res.* **73**, 5858–5868 (2013).
52. Moarri, M., Reyal, F. & Vert, J.-P. Integrative DNA methylation and gene expression analysis to assess the universality of the CpG island methylator phenotype. *Hum. Genomics* **9**, 26 (2015).
53. Maley, C. C. et al. Classifying the evolutionary and ecological features of neoplasms. *Nat. Rev. Cancer* **17**, 605–619 (2017).
54. Vendramin, R., Litchfield, K. & Swanton, C. Cancer evolution: Darwin and beyond. *EMBO J.* **40**, e108389 (2021).
55. Gould, S. J. & Eldredge, N. Punctuated equilibria: an alternative to phyletic gradualism. In Schopf, T.J.M. *Models in Paleobiology* 82–115 (Freeman Cooper, 1972).
56. Zolondick, A. A. et al. Asbestos-induced chronic inflammation in malignant pleural mesothelioma and related therapeutic approaches—a narrative review. *Precis. Cancer Med.* **4**, 27–27 (2021).
57. Southwood, T. R. E., May, R. M., Hassell, M. P. & Conway, G. R. Ecological strategies and population parameters. *Am. Nat.* **108**, 791–804 (1974).
58. Napolitano, A. et al. Minimal asbestos exposure in germline BAP1 heterozygous mice is associated with deregulated inflammatory response and increased risk of mesothelioma. *Oncogene* **35**, 1996–2002 (2016).
59. Adashek, J. J., Goloubek, A., Kato, S. & Kurzrock, R. Missing the target in cancer therapy. *Nat. Cancer* **2**, 369–371 (2021).
60. Gay, C. M. et al. Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities. *Cancer Cell* **39**, 346–360. e7 (2021).
61. Dora, D. et al. Neuroendocrine subtypes of small cell lung cancer differ in terms of immune microenvironment and checkpoint molecule distribution. *Mol. Oncol.* **14**, 1947–1965 (2020).
62. Owonikoko, T. K. et al. YAP1 expression in SCLC defines a distinct subtype with T-cell-inflamed phenotype. *J. Thorac. Oncol.* **16**, 464–476 (2021).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Lise Mangiante<sup>1,2,47</sup>, Nicolas Alcala<sup>1,47,48</sup>, Alexandra Sexton-Oates<sup>1,47</sup>, Alex Di Genova<sup>1,3,4,47</sup>, Abel Gonzalez-Perez<sup>1,5,6</sup>, Azhar Khandekar<sup>7</sup>, Erik N. Bergstrom<sup>7</sup>, Jaehee Kim<sup>8,9</sup>, Xiran Liu<sup>8</sup>, Ricardo Blazquez-Encinas<sup>10,11,12,13</sup>, Colin Giacobi<sup>1</sup>, Nolwenn Le Stang<sup>14</sup>, Sandrine Boyault<sup>15</sup>, Cyrille Cuenin<sup>16</sup>, Severine Tabone-Eglinger<sup>14</sup>, Francesca Damiola<sup>14</sup>, Catherine Voegele<sup>1</sup>, Maude Ardin<sup>17</sup>, Marie-Cecile Michallet<sup>17</sup>, Lorraine Soudade<sup>1</sup>, Tiffany M. Delhomme<sup>1,5</sup>, Arnaud Poret<sup>1</sup>, Marie Brevet<sup>18</sup>, Marie-Christine Copin<sup>19</sup>, Sophie Giusiano-Courcambeck<sup>20</sup>, Diane Damotte<sup>21,22</sup>, Cecile Girard<sup>23</sup>, Veronique Hofman<sup>24</sup>, Paul Hofman<sup>24</sup>, Jérôme Mouroux<sup>25</sup>, Charlotte Cohen<sup>26</sup>, Stephanie Lacomme<sup>27</sup>, Julien Mazieres<sup>28</sup>, Vincent Thomas de Montpreville<sup>29</sup>, Corinne Perrin<sup>30</sup>, Gaetane Planchard<sup>31</sup>, Nathalie Rousseau<sup>31</sup>, Isabelle Rouquette<sup>32</sup>, Christine Sagan<sup>23</sup>, Arnaud Scherpereel<sup>33</sup>, Francoise Thivolet<sup>30</sup>, Jean-Michel Vignaud<sup>34,35</sup>, Didier Jean<sup>36</sup>, Anabelle Gilg Soit Ilg<sup>37</sup>, Robert Olaso<sup>10</sup>, Vincent Meyer<sup>38</sup>, Anne Boland-Auge<sup>10</sup>, Jean-Francois Deleuze<sup>10</sup>, Janine Altmuller<sup>39</sup>, Peter Nuernberg<sup>39</sup>, Alejandro Ibáñez-Costa<sup>10,11,12,13</sup>, Justo P. Castaño<sup>10,11,12,13</sup>, Sylvie Lantuejoul<sup>14,40</sup>, Akram Ghantous<sup>16</sup>, Charles Maussion<sup>10</sup>, Pierre Courtiol<sup>41</sup>, Hector Hernandez-Vargas<sup>10</sup>, Christophe Caux<sup>17</sup>, Nicolas Girard<sup>44,45</sup>, Nuria Lopez-Bigas<sup>10,5,6,46</sup>, Ludmil B. Alexandrov<sup>7</sup>, Françoise Galateau-Salle<sup>14</sup>, Matthieu Foll<sup>1,48</sup>✉ & Lynnette Fernandez-Cuesta<sup>1,48</sup>✉

<sup>1</sup>Rare Cancers Genomics Team, Genomic Epidemiology Branch, International Agency for Research on Cancer/World Health Organization, Lyon, France.

<sup>2</sup>Department of Medicine, Stanford University, Stanford, CA, USA. <sup>3</sup>Instituto de Ciencias de la Ingeniería, Universidad de O'Higgins, Rancagua, Chile.

<sup>4</sup>Centro de Modelamiento Matemático UMI-CNRS 2807, Universidad de Chile, Santiago, Chile. <sup>5</sup>Institute for Research in Biomedicine (IRB Barcelona),

The Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>6</sup>Centro de Investigación Biomédica en Red en Cáncer, Instituto de Salud Carlos III, Madrid, Spain. <sup>7</sup>Department of Cellular and Molecular Medicine, Department of Bioengineering and Moores Cancer Center, University of California,

San Diego, La Jolla, CA, USA. <sup>8</sup>Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. <sup>9</sup>Department of

Computational Biology, Cornell University, Ithaca, NY, USA. <sup>10</sup>Maimonides Biomedical Research Institute of Cordoba, Córdoba, Spain. <sup>11</sup>Department of Cell Biology, Physiology and Immunology, University of Cordoba, Córdoba, Spain. <sup>12</sup>Reina Sofia University Hospital, Córdoba, Spain. <sup>13</sup>CIBER Fisiopatología de la Obesidad y Nutrición, Córdoba, Spain. <sup>14</sup>UMR INSERM 1052, CNRS 5286, Cancer Research Center of Lyon, MESOPATH-MESOBANK, Department of Biopathology, Cancer Centre Léon Bérard, Lyon, France. <sup>15</sup>Cancer Genomic Platform, Translational Research and Innovation Department, Centre Léon Bérard, Lyon, France. <sup>16</sup>EpiGenomics and Mechanisms Branch, International Agency for Research on Cancer/World Health Organization, Lyon, France.

<sup>17</sup>Tumor Escape, Resistance and Immunity Department, Centre de Recherche en Cancérologie de Lyon, Centre Léon Bérard, Université de Lyon, Université Claude Bernard Lyon 1, INSERM 1052, CNRS 5286, Lyon, France. <sup>18</sup>Cypath and Cypath-rb, Villeurbanne, France. <sup>19</sup>University of Lille, Centre Hospitalier Universitaire Lille, Institut de Pathologie, Tumorothèque du Centre de Référence Régional en Cancérologie, Lille, France. <sup>20</sup>Department of Pathology,

Centre Hospitalier Universitaire Nord, Marseille, France. <sup>21</sup>Centre de Recherche des Cordeliers, Inflammation, Complement and Cancer Team, Sorbonne Université, INSERM, Université de Paris, Paris, France. <sup>22</sup>Department of Pathology, Hôpitaux Universitaire Paris Centre, Tumorothèque/CRB Cancer, Cochin Hospital, Assistance Publique-Hôpitaux de Paris, Paris, France. <sup>23</sup>Tumorothèque Centre Hospitalier Universitaire de Nantes, Nantes, France. <sup>24</sup>Université Côte d'Azur, Laboratory of Clinical and Experimental Pathology, Nice Center Hospital, FHU OncoAge, Biobank BB-0033-00025 and IRCAN Inserm U1081/CNRS 7284, Nice, France. <sup>25</sup>Université Côte d'Azur, Department of Thoracic Surgery, Nice Center Hospital, FHU OncoAge and IRCAN Inserm U1081/CNRS 7284, Nice, France. <sup>26</sup>Department of Thoracic Surgery, FHU OncoAge, Nice Pasteur Hospital, Université Côte d'Azur, Nice, France. <sup>27</sup>Nancy Regional University Hospital, Centre Hospitalier Régional Universitaire, CRB BB-0033-00035, Inserm U1256, Nancy, France. <sup>28</sup>Toulouse University Hospital, Université Paul Sabatier, Toulouse, France. <sup>29</sup>Department of Pathology, Marie Lannelongue Hospital, Le Plessis Robinson, France. <sup>30</sup>Hospices Civils de Lyon, Institut de Pathologie, Centre de Ressources Biologiques des HCL, Tissu-Tumorothèque Est, Lyon, France. <sup>31</sup>Centre Hospitalier Universitaire de Caen, MESOPATH Regional Center, Caen, France. <sup>32</sup>Centre de Pathologie des Côtes, Centre de Ressources Biologiques (CRB Cancer), IUCT Oncopole, Toulouse, France. <sup>33</sup>University of Lille, Centre Hospitalier Universitaire Lille, INSERM, OncoThAI, NETMESO Network, Lille, France. <sup>34</sup>Department of Biopathology, Centre Hospitalier Régional Universitaire de Nancy, Vandoeuvre-les-Nancy, France. <sup>35</sup>BRC, BB-0033-00035, Centre Hospitalier Régional Universitaire de Nancy, Vandoeuvre-les-Nancy, France. <sup>36</sup>Centre de Recherche des Cordeliers, INSERM, Sorbonne Université, Université de Paris, Functional Genomics of Solid Tumors, Paris, France. <sup>37</sup>Direction Santé Environnement Travail, Santé Publique France, Paris, France. <sup>38</sup>Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, Evry, France. <sup>39</sup>Cologne Centre for Genomics, Cologne, Germany. <sup>40</sup>Grenoble Alpes University, Saint-Martin-d'Hères, France. <sup>41</sup>Owkin, New York, NY, USA. <sup>42</sup>UMR INSERM 1052, CNRS 5286, UCBL1, Centre Léon Bérard, Lyon, France.

<sup>43</sup>Centre de Recherche en Cancérologie de Lyon, Lyon, France. <sup>44</sup>Institut Curie, Institut du Thorax Curie Montsouris, Paris, France. <sup>45</sup>Université de Versailles Saint-Quentin-en-Yvelines, Université Paris-Saclay, Versailles, France. <sup>46</sup>Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain.

<sup>47</sup>These authors contributed equally: Lise Mangiante, Nicolas Alcala, Alexandra Sexton-Oates, Alex Di Genova. <sup>48</sup>These authors jointly supervised this work: Nicolas Alcala, Matthieu Foll, Lynnette Fernandez-Cuesta. ✉ e-mail: [follm@iarc.who.int](mailto:follm@iarc.who.int); [fernandezcuestal@iarc.who.int](mailto:fernandezcuestal@iarc.who.int)

## Methods

This section briefly describes the main methods (see Supplementary Information for details on the data, processing and analyses).

### Ethics

All of the methods were carried out in accordance with relevant guidelines and regulations. This study is part of a larger study, the MESOMICS project, aiming to perform comprehensive molecular characterization of MPM, and was approved by the International Agency for Research on Cancer (IARC) Ethics Committee (project number 15-17). The samples used in this study belong to the virtual biorepository French MESOBANK. Written, informed consent was obtained from all participants and no participant compensation was provided.

### Clinical data

Age at diagnosis (in years), sex (male or female), smoking status (non-smoker, ex smoker or smoker), asbestos exposure (exposed or non-exposed), previous treatment with chemotherapy drugs (yes or no), treatment information (surgery, chemotherapy, radiotherapy, immunotherapy or cancer history) and survival data (calculated in months from surgery to the last day of follow-up or death) were collected for all 123 patients. The median age at diagnosis was 67.5 years and 73.3% of patients were male.

### MESOMICS cohort

The MESOMICS cohort includes biological material from 123 patients with MPM (including three nonchemonaive patients who were excluded from all analyses unless explicitly mentioned) kindly provided by the French MESOBANK and annotated with detailed clinical, epidemiological and morphological data. Samples were collected from chemonaive surgically resected tumors, applying local regulations and rules at the collecting site, and included patient consent for molecular analyses, as well as the collection of de-identified data. Samples underwent an independent pathological review by the French MESOPATH reference panel, who determined that of the 120 MPM tumor samples, 79 belonged to the MME type, 26 were MMB and 15 were MMS. Of the 105 samples with an epithelioid component (79 MME and 26 MMB), solid, acinar, trabecular and tubulopapillary architectural patterns were the most frequent in the series ( $n = 37, 31, 16$  and 14, respectively).

### Discovery and intratumoral heterogeneity cohorts

Among the 123 patients with MPM, 13 had two tumor specimens collected for the study of intratumoral heterogeneity (ITH). The one with the highest tumor content, estimated by pathological review, was selected for this descriptive study and is reported in Supplementary Tables 1–3, and the other region is described in Supplementary Tables 50–52. Additionally, three patients have been reported as nonchemonaive and they were excluded from the analyses except if explicitly mentioned otherwise in the Methods.

### Pathological review

For all 136 samples (123 tumors plus 13 additional regions), a hematoxylin and eosin stain from a representative formalin-fixed, paraffin embedded block was collected for pathological review. Our pathologist (F.G.-S.) performed a detailed pathological review and classified all tumors according to the 2015 WHO classification<sup>63,64</sup>. The hematoxylin and eosin stain was also used to assess the quality of the frozen material selected for molecular analyses and to confirm that all frozen samples were at least 70% tumor cells.

### Artificial Intelligence analysis

Whole-slide image-based artificial intelligence prognostic scores were computed using the artificial intelligence MesoNet model based on morphological features, developed by Owkin—an artificial intelligence for medical research company<sup>15</sup>.

### Statistical analyses

All analyses were performed in R version 4.1.2. All tests involving multiple comparisons were adjusted using the Benjamini–Hochberg procedure, controlling the false discovery rate using the p.adjust R function (stats package version 3.4.4). To limit false discoveries, we took a conservative  $q$  value threshold of 0.05. In addition, in line with the American Statistical Association statement on the misuse of  $P$  values<sup>65</sup>, which intends to ‘steer research into a “post  $P < 0.05$  era”’, we report all  $P$  and  $q$  values, even those that may be closer to arbitrary thresholds such as the 5% threshold. To improve the reproducibility of our results, we summarize in Supplementary Tables 21 and 22 all  $P$  and  $q$  values reported in the text and main figures, along with details about the tests performed (hypothesis, model and sample size) and replication performed with additional cohorts.

### Survival analysis

Survival analysis has been performed using Cox’s proportional hazard model from which the significance of the hazard ratio between the reference and the other levels has been evaluated using Wald tests. We assessed the global significance of the model using the logrank test statistic (R package survival version 2.41-3) and drew Kaplan–Meier and forest plots using the R package survminer (version 0.4.2).

### DNA extraction

Included samples were extracted using the Gentra Puregene Tissue Kit (4 g) (158667; Qiagen), following the manufacturer’s instructions. All DNA samples were quantified using the fluorometric method (Quant-iT PicoGreen dsDNA Assay; Life Technologies) and assessed for purity by NanoDrop (Thermo Scientific) 260/280 and 260/230 ratio measurements. The DNA integrity of the fresh frozen samples was checked with a TapeStation system (Agilent Biotechnologies) using Genomic DNA ScreenTape (Agilent Biotechnologies).

### RNA extraction

Included samples were extracted using the AllPrep DNA/RNA extraction kit (Qiagen) following the manufacturer’s instructions. All RNA samples were treated with DNase I for 15 min at 30 °C. The RNA integrity of the frozen samples was checked with a TapeStation system (Agilent Biotechnologies) using RNA ScreenTape (Agilent Biotechnologies).

Because of unsuccessful extraction (impacting either the quality or the quantity), we obtained different numbers of MPM samples for which WGS, DNA methylation or RNA sequencing (RNA-seq) data are available (Supplementary Tables 1–3).

### DNA sequencing

**Sequencing.** WGS was performed by the Centre National de Recherche en Génomique Humaine (Institut de Biologie François Jacob, CEA) on 130 fresh frozen MPMs, 54 of which with matched normal tissue or blood samples. We used an Illumina TruSeq DNA PCR-Free Library Preparation Kit (20015963; Illumina) according to the manufacturer’s instructions and sequenced them on a HiSeq X Five platform (Illumina) as paired-end 150-base pair reads. Samples paired with matched normal tissue or blood had a target sequencing depth of 60× and other samples had a target depth of 30×.

**Data processing.** WGS reads were mapped to the reference genome GRCh38 (with ALT and decoy contigs) using our in-house workflow (<https://github.com/IARCbioinfo/alignment-nf>; release version 1.0)<sup>66</sup>. In summary, this workflow relies on the Nextflow domain-specific language<sup>67</sup> version 20.10.0.5430 and consists of four steps: read mapping (software BWA<sup>68</sup>; version 0.7.15), duplicate marking (software samblaster<sup>69</sup>; version 0.1.24), read sorting (software sambamba<sup>70</sup>; version 0.6.6) and base quality score recalibration using GATK<sup>71</sup> (version 4.0.12).

**Variant calling and filtering on DNA.** We performed somatic variant calling using the software Mutect2 (ref.<sup>72</sup>) from GATK version 4.1.5.0, as implemented in our Nextflow workflow (<https://github.com/IARCBioinfo/mutect-nf>; release version 2.2b). Multiregion samples were processed jointly using the multisample calling mode of Mutect2. We called germline variants using Strelka2 (ref.<sup>73</sup>) version 2.9.10-0 using our Nextflow workflow (<https://github.com/IARCBioinfo/mutect2-nf>; release version 1.2a). Annotation was performed with ANNOVAR<sup>74</sup> (16 April 2018) using the GENCODE version 33 annotation, COSMIC version 90 and REVEL databases. To call somatic variants on tumor-only samples (72/115), a similar procedure was performed (Mutect2 tumor-only mode) but including further germline-filtering steps using a random forest classifier.

**CNV calling.** Somatic CNVs were called using the PURPLE software<sup>75</sup> version 2.52, as implemented in our Nextflow workflow (<https://github.com/IARCBioinfo/purple-nf>; version 1.0). We used a total of 57 matched WGS samples of MPM (including multiregion samples) for benchmarking the tumor-only mode of PURPLE. We ran PURPLE twice for each matched sample: first using the matched WGS normal/tumor pair as input and second using only the tumor WGS sample as input.

**Structural variant calling.** To identify somatic structural variants, including insertions, deletions, duplications, inversions and translocations, we built a consensus structural variants call set by integrating SvABA<sup>76</sup> version 1.1.0, Manta<sup>77</sup> version 1.6.0 and DELLY<sup>78</sup> version 0.8.3 calls with SURVIVOR<sup>79</sup> version 1.0.7. Somatic structural variants (minimum structural variant size = 50 base pairs) identified by at least two callers and single-caller predictions with a minimum read support of 15 pairs (including paired-end and split-read evidence) were included in the consensus set of each matched sample.

## RNA-seq

**Sequencing.** RNA-seq was performed on 126 fresh frozen MPM samples in the Cologne Center for Genomics, of which 109 MPM samples belonged to the discovery cohort (Supplementary Tables 1–3). Libraries were prepared using the Illumina TruSeq Stranded mRNA Sample Preparation Kit (20020595; Illumina) and the pool was sequenced using an Illumina NovaSeq 6000 sequencing device and a paired-end 100-nucleotide protocol.

**Data processing.** The 126 raw read files from the MESOMICS cohort and the 21 files from the Iorio and colleagues<sup>18</sup> mesothelioma cohort (downloaded from the European Genome-phenome Archive (EGA) and Sequence Read Archive websites; datasets EGAS00001000828 and PRJNA523380, respectively) were processed in three steps using the RNA-seq processing workflow based on the Nextflow language and accessible at <https://github.com/IARCBioinfo/RNAseq-nf> (release version 2.3)<sup>66</sup>. Then, reads were realigned locally using ABRA2 (ref.<sup>80</sup>); (workflow <https://github.com/IARCBioinfo/abra-nf>; release version 3.0) and base quality scores were recalibrated using GATK (workflow <https://github.com/IARCBioinfo/BQSR-nf>; release version 1.1). Once processed, expression was quantified using StringTie software (version 2.1.2; Nextflow pipeline accessible at <https://github.com/IARCBioinfo/RNAseq-transcript-nf>; release version 2.2).

The raw read counts of the 59,607 genes in the expression data matrix, from the MESOMICS, TCGA and Bueno cohorts<sup>3,4</sup>, from which we removed non-chimera samples, were normalized using the variance-stabilizing transform (vst function from R package DESeq2 version 1.14.1); this transformation enables comparisons between samples with different library sizes and different variances in expression across genes.

## DNA methylation

**EPIC 850K methylation array.** Epigenome analysis was performed on 119 MPMs (Extended Data Fig. 1 and Supplementary Tables 1–3), two

technical replicates and three adjacent normal tissues. Epigenomic studies were performed at the IARC with the Infinium EPIC DNA methylation beadchip platform (Illumina) used for the interrogation of over 850,000 CpG sites (dinucleotides that are the main target for methylation).

**Data processing.** The resulting IDAT raw data files were preprocessed using the R packages minfi (version 1.34.0) and ENmix (version 1.25.1). Raw data were then normalized using functional normalization (function preprocessFunnorm; minfi), to reduce technical variation within the data, and probe removal steps were performed to ensure reliability and accuracy of the final dataset. This resulted in a normalized, filtered dataset of 781,245 probes for 139 samples. Finally, beta and *M* values were extracted (functions getBeta and getM; minfi). Nine probes recorded *M* values of  $\sim$  for at least one sample, and these values were replaced with the next lowest *M* value in the dataset. The three normal tissues and one remaining technical replicate were then removed from the beta and *M* matrices for the subsequent analyses. This resulted in 135 samples: 122 for discovery and an additional 13 for ITH analyses.

**CIMP index.** A CIMP index value was calculated for all samples as follows. The mean beta value across all probes located within CpG islands was calculated per sample, resulting in beta values for 24,891 and 24,924 CpG islands, MESOMICS (EPIC array), TCGA<sup>4</sup> and Iorio and colleagues<sup>18</sup> cell lines (HM450K array), respectively. The CIMP index was then calculated as the proportion of these 24,891 or 24,924 islands with  $\geq 30\%$  methylation (beta value  $\geq 0.3$ ) per sample.

## Integrative unsupervised analyses

We performed four series of analyses with different subsets of samples: (1) discovery analyses with all of our discovery cohort (MESOMICS cohort; 120 samples), for which WGS, RNA-seq and/or 850K methylation array data were available; (2) and (3) replication analyses with the already published data from Bueno<sup>3</sup> (181 samples after exclusion of nonchimera samples) and Hmeljak and colleagues<sup>4</sup> (TCGA cohort; 73 samples in the curated list), respectively; (4) combined analyses integrating the MESOMICS, Bueno and TCGA cohorts<sup>3,4</sup> with a total of 374 samples; and (5) replication combining cell lines from the Iorio study<sup>18</sup> (for which whole-exome sequencing, expression arrays and RNA-seq, 450K methylation arrays and drug responses in the form of half-maximum inhibitory concentration scores are available (21 samples; 265 drugs)) and the de Reyniès<sup>5</sup> and Blum et al.<sup>7</sup> datasets (for which expression arrays and drug responses are available (38 samples; three drugs)). In addition, some single-omic analyses are also described in this section.

**Preprocessing of expression data.** We used normalized read count matrices (see the section ‘RNA-seq’) for subsets (1)–(4), encompassing 59,607 genes. Among these genes, those having less than one fragment per kilobase of exon per million mapped fragments (FPKM) difference across the samples were excluded from the unsupervised analyses. Also, to mitigate sex influence on the expression profiles, we removed genes from the sex chromosomes. For each analysis, the top 5,000 most variable genes were selected. Similarly, the 5,000 most variable genes from the normalized array expression of cell lines (see the section ‘Processing of publicly available expression array processing’ in Supplementary Methods) were selected. Whenever several probes were available for the same gene, the one with the highest intensity was selected.

**Preprocessing of methylation data.** DNA methylation was available for both the MESOMICS and TCGA cohorts. First, we extracted the *M* values of the CpGs from the MESOMICS, TCGA<sup>4</sup>, combined MESOMICS/TCGA and Iorio<sup>18</sup> cell line cohorts, respectively<sup>81</sup>. We excluded sex chromosome CpGs, CpGs that did not pass quality control (see the

section 'DNA methylation' in Supplementary Methods) and those having less than 0.1 beta value difference across the (1) 119, (3) 73, (4) 192 and (5) 59 samples. Based on this annotation, the CpG list representing the methylation data was divided according to their association with promoters, enhancers or the gene body using the EPIC 850K array manifest B5 (see the section 'Regional methylation analysis' in Supplementary Methods), resulting in three datasets, respectively named MethPro, MethEnh and MethBod. For each analysis and dataset, the top 5,000 most variable CpGs (calculated from  $M$  values) were selected.

**Preprocessing of copy number changes.** Copy number change data were available for the MESOMICS, TCGA and MPM cell line cohorts. We assessed the global (total) and minor (minor) allele copy number states at the gene level using, respectively, the total (total) and minor (minor) copy number estimate given by PURPLE (see the section 'CNV calling') on the hg38 genome for the MESOMICS cohort and SNP array estimates downloaded from the Genomic Data Commons portal for the TCGA–MESO cohort<sup>4</sup> and from the Cell Model Passports portal for the MPM cell lines.

For the three analyses, the resulting value assigned to each gene is an average of the copy number estimate of the tumor by taking into account the tumor purity (purity) estimated by PURPLE. To avoid redundancy, genes with exactly the same resulting copy number value in all samples (because of their genome location proximity) were grouped as one single feature in the dataset. Only the genes or groups of genes altered in at least three samples were selected. To ensure continuity of the data, which is technically necessary for the algorithm, the copy number estimates were centered and scaled before being integrated into the MOFA algorithm. For consistency, somatic CNVs occurring on sex chromosomes were removed and the top 5,000 most variable genes or groups of genes were selected to be integrated.

**Preprocessing of genomic alterations data.** Somatic structural variants data were used only for integrative analyses (1) and (4), while somatic mutations were used in all analyses. Each gene, altered by somatic splicing, structural variants or exonic, damaging mutations (see the section 'Damaging variants and driver detection' in Supplementary Methods) was integrated in a common dataset. Of note, for missense mutations, we used the REVEL annotation included in ANNOVAR for predicting the pathogenicity of these variants and we used a 0.5 cut-off to restrict to the most likely damaging missense events. We also removed genes altered in fewer than three samples. For consistency, we selected genes in non-sex chromosomes, protein-coding or long noncoding RNA genes, and with expression greater than or equal to 0.01 fragment per kilobase of exon per million mapped fragments (FPKM) in at least one sample of the cohort, to be sure to include genes expressed in mesothelioma. We integrated the resulting datasets as a Boolean variable in the following analyses.

**Multomic integrative analyses.** To provide an integrative low-dimensional summary of the molecular variation across the samples, we performed continuous latent factors identification using the software MOFA (R package MOFA2, version 1.7.0). Indeed, MOFA is able to integrate different molecular datasets (layers) by generating independent continuous variables, named latent factors, that explain most variation from the joint datasets. In total, we performed five analyses: (1) MOFA–MESOMICS ( $n = 120$ ; Fig. 1 and Extended Data Fig. 1a); (2) MOFA–Bueno ( $n = 181$ ; Extended Data Fig. 1c); (3) MOFA–TCGA ( $n = 73$ ; Extended Data Fig. 1b); (4) MOFA–3 cohorts ( $n = 374$ ; Extended Data Fig. 1d) and (5) MOFA–cell lines, as described above ( $n = 59$ ; Supplementary Fig. 4). Additionally, we ran MOFA on our discovery cohort, including the ITH samples (MOFA–ITH;  $n = 134$ ) to evaluate the ITH within MPM samples.

MOFA was performed independently for each analysis, setting the number of latent factors to ten (function runMOFA from the R

package MOFA2). A summary of all of these runs is given in Extended Data Figs. 1 and 2, Fig. 1 and Supplementary Figs. 1 and 4 and coordinates and proportions of variance explained for models (1)–(4) are given in Supplementary Tables 4–8, while those for MOFA–ITH are given in Supplementary Tables 50–52 and those for the cell lines (model (5)) are given in Supplementary Tables 23–26. A comparison with other multiomic methods is provided in Extended Data Fig. 10 (see section 'Multomic integrative analyses details' in Supplementary Methods).

## Evolutionary tumor trade-off analyses

**Pareto task identification.** The Pareto front model was fitted to different sets of samples using the ParetoTlR package (<https://github.com/vitkl/ParetoTl>; release version 0.1.13), following the above-mentioned analyses (1)–(4), and additionally on two different kinds of molecular maps: using MOFA (restricting to LF1, LF2, LF3 and LF4) and using expression principal component analysis as technical validation (see the section 'RNA-seq'). In brief, the algorithm tries to find polyhedra by testing successively 1 to  $n$  axes, adding them one after another in decreasing order of transcriptomic variance explained. For this technical reason, the MOFA latent factors were ordered as follows by decreasing transcriptomic variance explained: morphology factor (LF2), adaptive response factor (LF3), CIMP factor (LF4) and ploidy factor (LF1). For each number  $n$  of axes used, ParetoTl identifies the position of the  $n + 1 = k$  vertices (archetypes) in the molecular map defined, and we used 200 bootstraps, each taking 75% of the data to measure the variability in archetype position and infer archetype positions robust to outliers (function fit\_pch\_bootstrap with the parameters bootstrap = T and bootstrap\_N = 200; see our code at [https://github.com/IARCbioinfo/MESOMICS\\_data/blob/main/phenotypic\\_map/MESOMICS/PhenotypicMap\\_MESOMICS.md](https://github.com/IARCbioinfo/MESOMICS_data/blob/main/phenotypic_map/MESOMICS/PhenotypicMap_MESOMICS.md)).

**Interpretation of tumor archetypes.** To further characterize the phenotype of each archetype, we used the proportion of each archetype for each sample estimated by ParetoTl. These proportions were used as continuous variables to further test the association between each archetype and clinical, epidemiological and morphological variables, as well as molecular data (Supplementary Tables 27–30).

More specifically, we inferred each archetype phenotype by performing IGSEA on the expression data. To do so, we used the ActivePathways R package (<https://github.com/reimandlab/ActivePathways>; release version 1.1.0), which is a tool able to integrate different sources of molecular variation to assess the enrichment of Gene Ontology terms by combining  $P$  values from different association tests between sources and gene-level data. Here we integrated these proportions as different axes of molecular variation. We restricted the Gene Ontology terms to a minimum size of 20 genes and a maximum size of 1,000 genes as the default parameters of ActivePathways. To infer the pathways specifically altered in each archetype, we integrated the Pearson's  $P$  value correlation of each gene from the expression matrix of 59,607 genes with the proportion from each archetype and we selected the pathways for which the enrichment source only corresponded to the tested archetype. We performed two kinds of analyses: one restricted to the genes positively correlated with the proportion (to obtain the upregulated pathways) and the other restricted to the negatively correlated genes (to identify the downregulated pathways).

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The genome sequencing, RNA-seq and methylation data have been deposited in the EGA database, which is hosted at the European Bioinformatics Institute and Centre for Genomic Regulation under accession number [EGAS00001004812](https://ega.ebi.ac.uk/study/EGAS00001004812). Because raw omics datasets

derived from humans are at risk of re-identification when combined with information from other public sources, access must be requested from the MESOMICS data access committee, as detailed at <https://ega-archive.org/studies/EGAS00001004812>. Minimum datasets of processed somatic alterations for genomic, transcriptomic and epigenomic data, sufficient to reproduce, interpret and extend our main results, are publicly available at [https://github.com/IARCBioinfo/MESOMICS\\_data/tree/main/phenotypic\\_map/MESOMICS](https://github.com/IARCBioinfo/MESOMICS_data/tree/main/phenotypic_map/MESOMICS). A data note manuscript detailing all of the quality controls of the dataset is available at <https://www.biorxiv.org/content/10.1101/2022.07.06.499003v1> (ref. <sup>82</sup>). TCGA whole-exome sequencing, RNA-seq and methylation array data are available from the Genomic Data Commons portal (TCGA-MESO cohort<sup>4</sup>). Whole-exome sequencing and RNA-seq data from the Bueno and colleagues cohort<sup>3</sup> are available from the EGA under accession number [EGAS00001001563](https://ega-archive.org/studies/EGAS00001001563). Small variant lists, RNA-seq, expression array and methylation data for the Iorio and colleagues cohort<sup>18</sup> are available from the Gene Expression Omnibus (accession number [GSE29354](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29354)), EGA (accession number [EGAS00001000828](https://ega-archive.org/studies/EGAS00001000828)) and Sequence Read Archive (accession number [PRJNA523380](https://www.ncbi.nlm.nih.gov/sra/PRJNA523380)). Corresponding drug responses are available from the cancerrxgene.org website ([https://www.cancerrxgene.org/downloads/drug\\_data?issue=MESO](https://www.cancerrxgene.org/downloads/drug_data?issue=MESO); accessed July 2021). Expression array data for the de Reyniès and colleagues cohort<sup>5</sup> are available from the ArrayExpress platform ([E-MTAB-1719](https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1719)) and corresponding drug response data are available from the supplementary material of Blum et al.<sup>7</sup> All of the other data supporting the findings of this study are available within the article and its Supplementary Information files. Further information and requests for resources should be directed to and will be fulfilled by M.F. (follm@iarc.who.int). Source data are provided with this paper.

## Code availability

All bioinformatics pipelines are available at <https://github.com/IARCBioinfo> (see Methods for details about which pipelines and versions were used for each analysis). A detailed R notebook allowing reproduction of the MOFA and Pareto tumor task inference results for the MESOMICS cohort is available at [https://github.com/IARCBioinfo/MESOMICS\\_data](https://github.com/IARCBioinfo/MESOMICS_data).

## References

63. Galateau-Salle, F., Churg, A., Roggli, V., Travis, W. D. & World Health Organization Committee for Tumors of the Pleura. The 2015 World Health Organization Classification of Tumors of the Pleura: advances since the 2004 classification. *J. Thorac. Oncol.* **11**, 142–154 (2016).
64. WHO Classification of Tumours of the Lung, Pleura, Thymus and Heart (4th edn) (International Agency for Research on Cancer, 2015).
65. Wasserstein, R. L. & Lazar, N. A. The ASA statement on P-values: context, process, and purpose. *Am Stat.* **70**, 129–133 (2016).
66. Alcalá, N. et al. Integrative and comparative genomic analyses identify clinically relevant pulmonary carcinoid groups and unveil the supra-carcinoids. *Nat. Commun.* **10**, 3407 (2019).
67. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
68. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
69. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
70. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
71. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O'Reilly Media, 2020).
72. Benjamin, D. et al. Calling somatic SNVs and indels with Mutect2. Preprint at *bioRxiv* <https://doi.org/10.1101/861054> (2019).
73. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
74. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
75. Cameron, D. L. et al. GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. Preprint at *bioRxiv* <https://doi.org/10.1101/781013> (2019).
76. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
77. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
78. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
79. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
80. Mose, L. E., Perou, C. M. & Parker, J. S. Improved indel detection in DNA and RNA via realignment with ABRA2. *Bioinformatics* **35**, 2966–2973 (2019).
81. Du, P. et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
82. Genova, A. D. et al. A molecular phenotypic map of malignant pleural mesothelioma. *Gigascience* **12**, giac128 (2022).

## Acknowledgements

The MESOMICS project is part of the Rare Cancers Genomics initiative ([www.rarecancersgenomics.com](https://www.rarecancersgenomics.com)) led by the Rare Cancers Genomics team at the IARC (<https://www.iarc.who.int/teams-gem-rcg/>). We thank the patients for donating tumor specimens. The human biological samples and associated data were obtained from the French MESOBANK. We also thank R. Argelaguet for advice on using MOFA, H. Begueret, N. Rousseau, D. Bozonnet, E. Wasielewski, G. Clapison, C. Bonnetaud, K. Washetine, A. Lupo Mansuet, C. Cuenin and E. Clermont for their contribution to the biorepository. We acknowledge the American Association for Cancer Research and its financial and material support in the development of the AACR Project GENIE registry, as well as members of the consortium for their commitment to data sharing. Interpretations are the responsibility of the study authors. The results published here are in part based on data generated by the TCGA Research Network (<https://www.cancer.gov/tcg>). We also thank the French National Mesothelioma Surveillance Program and Santé Publique France. This work has been funded by the French National Cancer Institute (PRT-K 2016-039 to L.F.-C. and M.F.) and the Ligue Nationale Contre le Cancer (LNCC 2017 and 2020 to L.F.-C. and M.F.). L.M. has a fellowship from the Ligue Nationale Contre le Cancer. This work also benefited from support from the France Génomique national infrastructure, funded as part of the Investissements d'Avenir program managed by the Agence Nationale de la Recherche (contract ANR-10-INBS-09). Other funding was provided by the Spanish Ministry of Science and Innovation (PID2019-105201RB-I00 to J.P.C.), the Instituto de Salud Carlos III, co-funded by the European Union (ERDF/ESF; Investing in Your Future), a Sara Borrell postdoctoral grant (CD19/00255 to A.I.-C.), the Spanish Ministry of Universities (predoctoral contract FPU18/02275 to R.B.-E.), the Junta de Andalucía (BIO-0139) and the Universidad de Córdoba-FEDER (UCO-202099901918904) (to J.P.C. and A.I.-C.), a GETNE2019 Research grant to J.P.C. and the CIBER Fisiopatología de la

Obesidad y Nutrición (CIBER is an initiative of the Instituto de Salud Carlos III). We finally thank the reviewers and the editor for taking the time to provide very useful and constructive feedback.

## Author contributions

L.F.-C. and M.F. conceived the study idea. L.F.-C., M.F., L.M., N.A., A.D.G. and A.S.-O. developed the study methodology. L.M., N.A., A.D.G., A.S.-O., A.G.-P., A.K., E.N.B. and C.V. developed software. L.M., N.A., A.D.G. and A.S.-O. validated the results. L.M., N.A., A.D.G., A.S.-O., A.G.-P., A.K., E.N.B., C.V., M.A., C.M., P.C., A.G.-P. and F.G.-S. performed the formal analyses. L.M., N.A., A.D.G., A.S.-O., A.G.-P., A.K., E.N.B., J.K., X.L., R.B.-E., A.I.-C., J.P.C., C. Giacobi, M.A., L.S., T.M.D., A.P., C.M. and P.C. performed the investigation. N.L.S., S.B., S.T.-E., F.D., M.B., M.-C.C., S.G.-C., D.D., C. Girard, V.H., P.H., J. Mouroux., C. Cohen, S. Lacomme, J. Mazieres, V.T.d.M., C.P., G.P., N.R., I.R., C.S., A.S., F.T., J.-M.V., A.G.S.I., R.O., V.M., S. Lantuejoul and F.G.-S. provided resources. C. Cuenin performed the methylation experiments. L.M., N.A., A.D.G., A.S.-O. and C.V. curated the data. L.F.-C., M.F., L.M., N.A., A.D.G. and A.S.-O. wrote the original draft of the manuscript. L.F.-C., M.F., L.M., N.A., A.D.G., A.S.-O., A.G.-P., A.K., D.J., H.H.-V., C. Caux, N.G., N.L.-B., L.B.A. and F.G.-S. reviewed and edited the manuscript. L.M., N.A., A.D.G. and A.S.-O. created the visualizations of the results. L.F.-C., M.F. and N.A. supervised the project. L.F.-C., M.F., L.M., N.A., M.-C.M., A.B.-A., J.-F.D., J.A., P.N. and A.G. administered the project. L.F.-C., M.F. and N.A. acquired funding. L.F.-C., M.F., L.M., N.A. and A.S.-O. revised the manuscript.

## Competing interests

Where authors are identified as personnel of the IARC/WHO, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the IARC/WHO. Where authors are identified as personnel of the Centre de Recherche en Cancérologie de Lyon, the authors declare no competing interests. A.S. participated in expert boards and clinical trials with AstraZeneca, Bristol-Myers Squibb, MSD and Roche. N.G. declares consultancy for and research support from Bristol-Myers Squibb, AstraZeneca, Roche and MSD. S Lantuejoul

declares research support from AstraZeneca, Sanofi, Bristol-Myers Squibb, Janssen and Eli Lilly and has participated in expert boards for MSD and Bristol-Myers Squibb. D.D. declares research support from AstraZeneca. J. Mazieres declares consultancy for and research support from Roche, AstraZeneca, Bristol-Myers Squibb, MSD and Pierre Fabre. M.B. declares consultancy for and research support from AstraZeneca, Bristol-Myers Squibb and Amgen. I.R. participated in expert boards for AstraZeneca, MSD and Bristol-Myers Squibb. L.B.A. is a compensated consultant and has equity interest in io9. His spouse is an employee of Biotheranostics. L.B.A. is also an inventor on a US patent (10,776,718) relating to source identification by non-negative matrix factorization. E.N.B. and L.B.A. declare US provisional patent applications with the serial numbers 63/289,601 and 63/269,033. L.B.A. declares US provisional patent applications with the serial numbers 63/366,392, 63/367,846 and 63/412,835. C.M. is employed by and has equity interest in Owkin. C.M., P.C. and F.G.-S. are inventors on the US patent 17185924 'Systems and methods for mesothelioma feature detection and enhanced prognosis or response to treatment'. All other authors declare no competing interests.

## Additional information

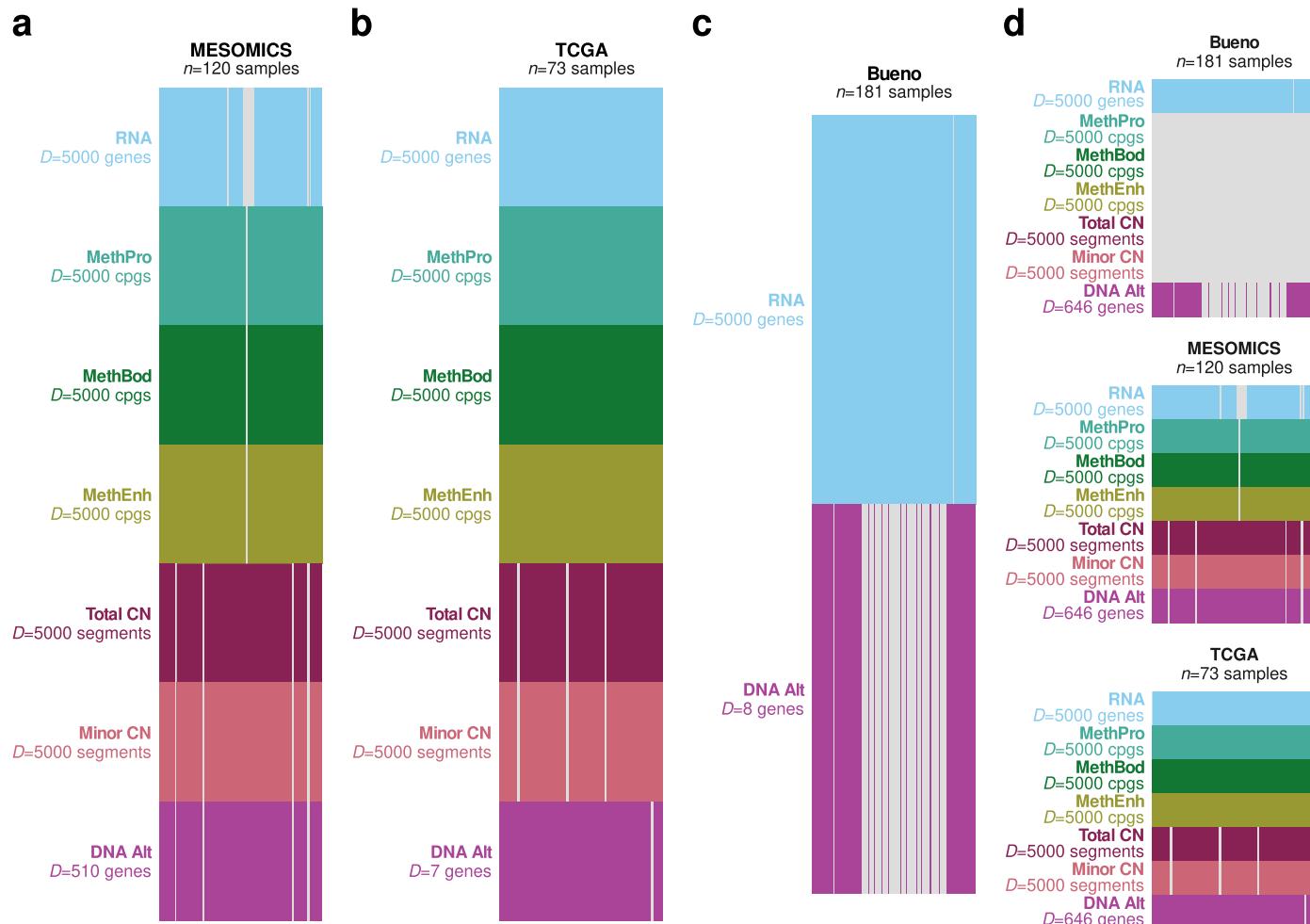
**Extended data** is available for this paper at <https://doi.org/10.1038/s41588-023-01321-1>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41588-023-01321-1>.

**Correspondence and requests for materials** should be addressed to Matthieu Foll or Lynnette Fernandez-Cuesta.

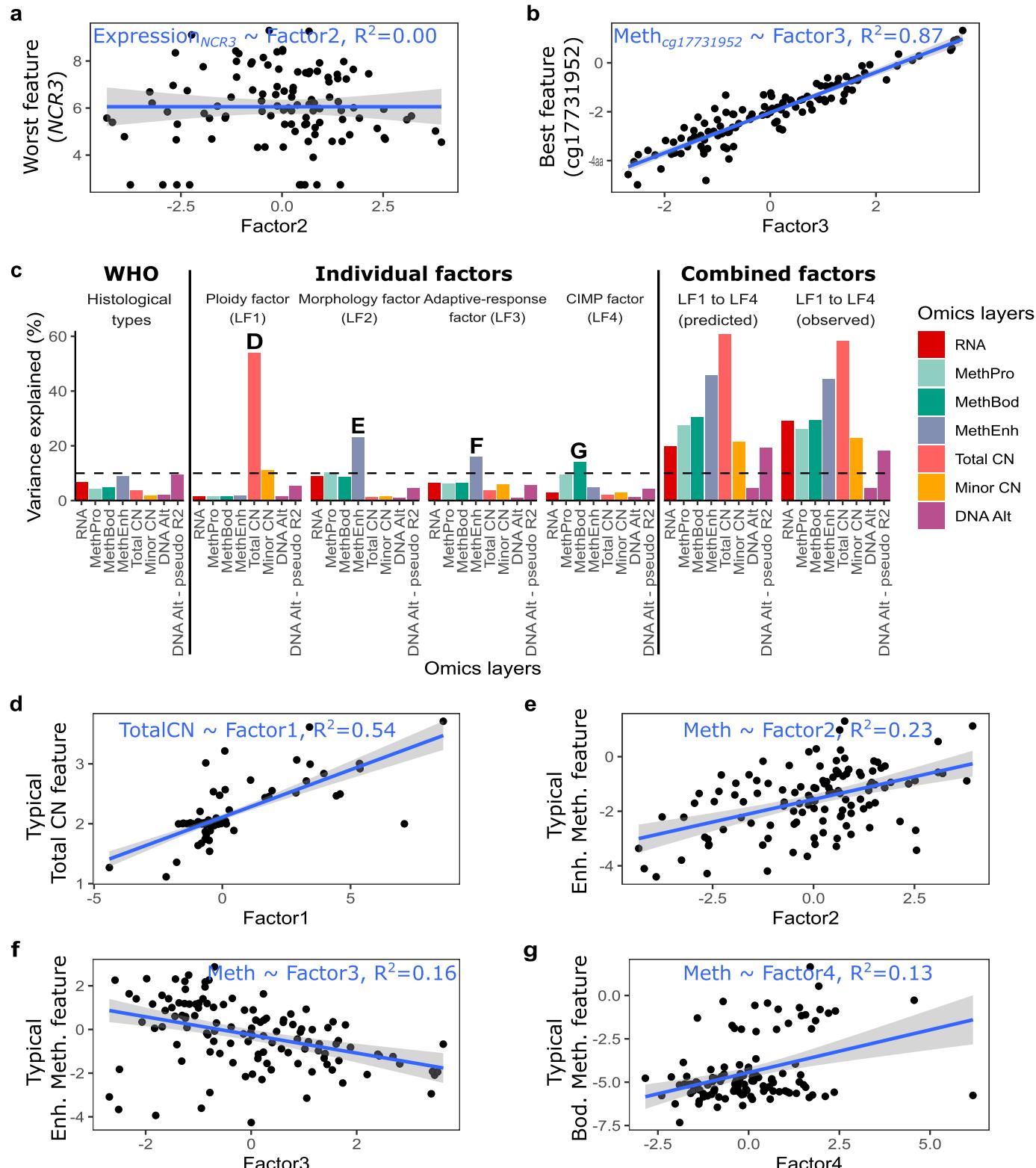
**Peer review information** *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



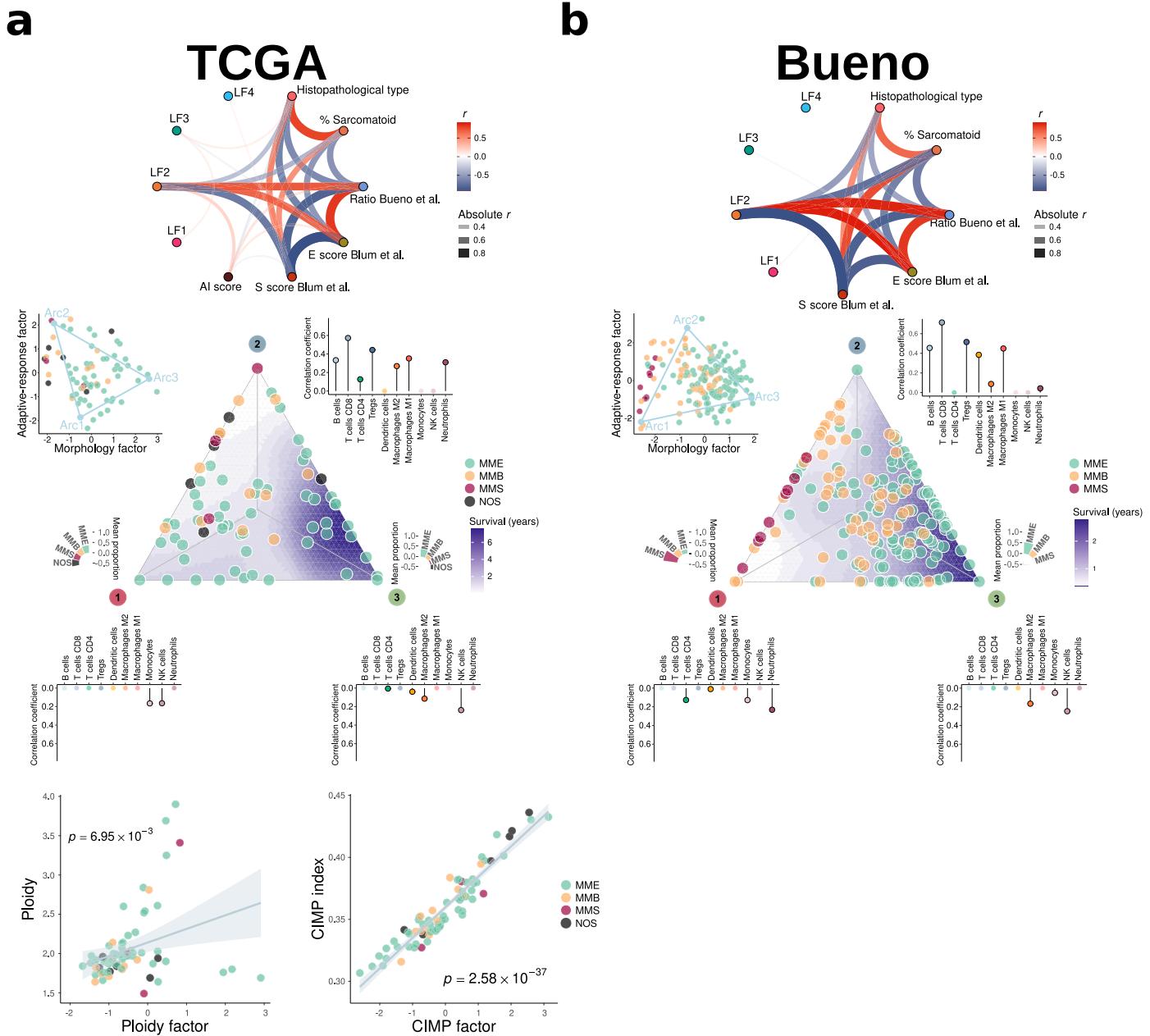
**Extended Data Fig. 1 | Overview of the MPM datasets for multiomic integration with MOFA.** Overview of the omic data sets integrated into multiomics factor analyses (MOFAs), for (a) a MOFA of the MESOMICS cohort (*n* = 120), (b) MOFA of the TCGA cohort (*n* = 73), (c) MOFA of the Bueno cohort (*n* = 181), and (d) MOFA of the 3 cohorts (*n* = 120 + 73 + 181). *D* is the number

of integrated omic features from genomic (rearrangements and mutations within *DNA Alt*; allele-specific copy number (CN) in *Total CN* and *Minor CN*), transcriptomic (RNA), and epigenomic data at promoter (*MethPro*), gene body (*MethBod*), and enhancer regions (*MethEnh*).



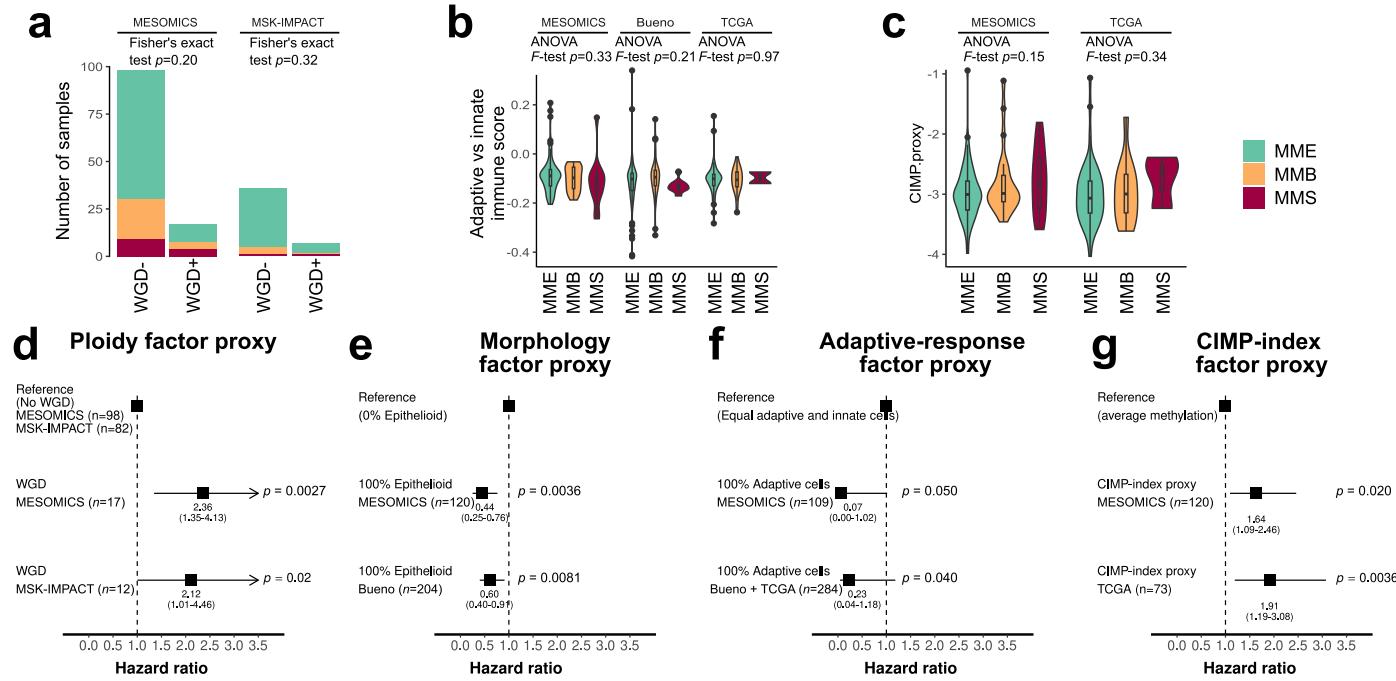
**Extended Data Fig. 2 | Proportion of interpatient variance explained by MOFA latent factors.** **a**) Example feature for which a latent factor explains 0% of interpatient variance (here factor 2 explains no variance at all in the expression of gene *NCR3*— $R^2 = 0$ ). **b**) Example feature for which a latent factor explains most of the variance (here factor 3 explains 87% of the variance of methylation site *cgl17731952*— $R^2 = 0.87$ ). **c**) Variance explained by the three histopathological types, each latent factor (LF) independently, predicted total variance explained by all latent factors together if they were completely independent (LF1 to LF4

predicted), and actual variance explained by a model including the four latent factors as covariates (LF1 to LF4 observed). CIMP: CpG island methylator phenotype. **d)** Typical Total copy number (CN) feature associated with Factor 1. **e)** Typical Enhancer Methylation feature associated with Factor 2. **f)** Typical Enhancer Methylation feature associated with Factor 3. **g)** Typical Gene Body Methylation feature associated with Factor 4. In **(a)-(b)** and **(d)-(g)**, the gray band corresponds to 95% confidence intervals.



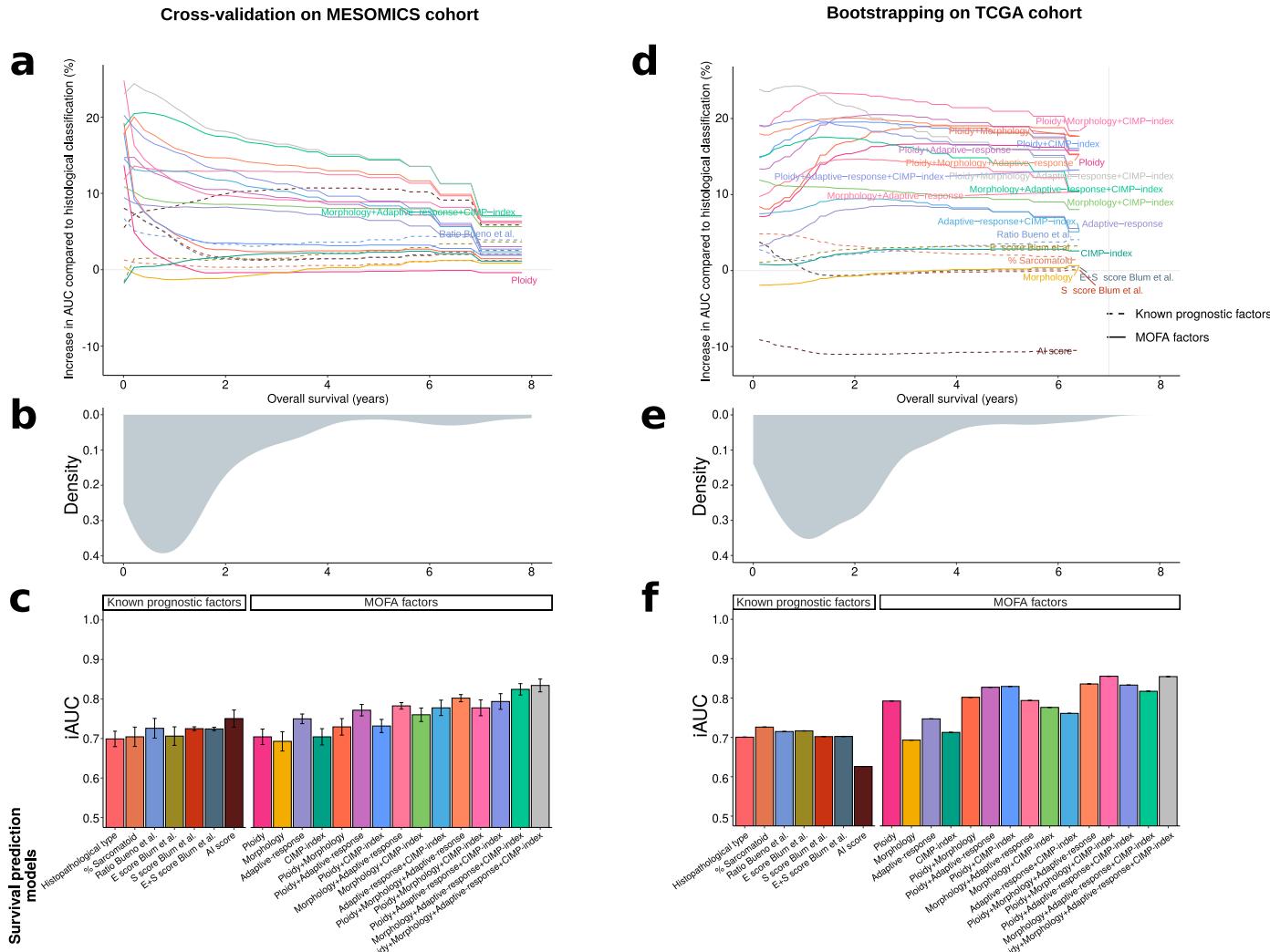
**Extended Data Fig. 3 | Replication of MOFAs latent factors and tumor tasks in major MPM cohorts.** MESOMICS MOFA latent factors and tumor task replicated in the TCGA (a) and Bueno (b) cohorts. The gray band corresponds to 95%

confidence intervals. In (a),  $P$  values correspond to Pearson correlation tests ( $n = 73$ ). MME: epithelioid; MMB: biphasic; MMS: sarcomatoid; NOS: malignant pleural mesothelioma not otherwise specified.



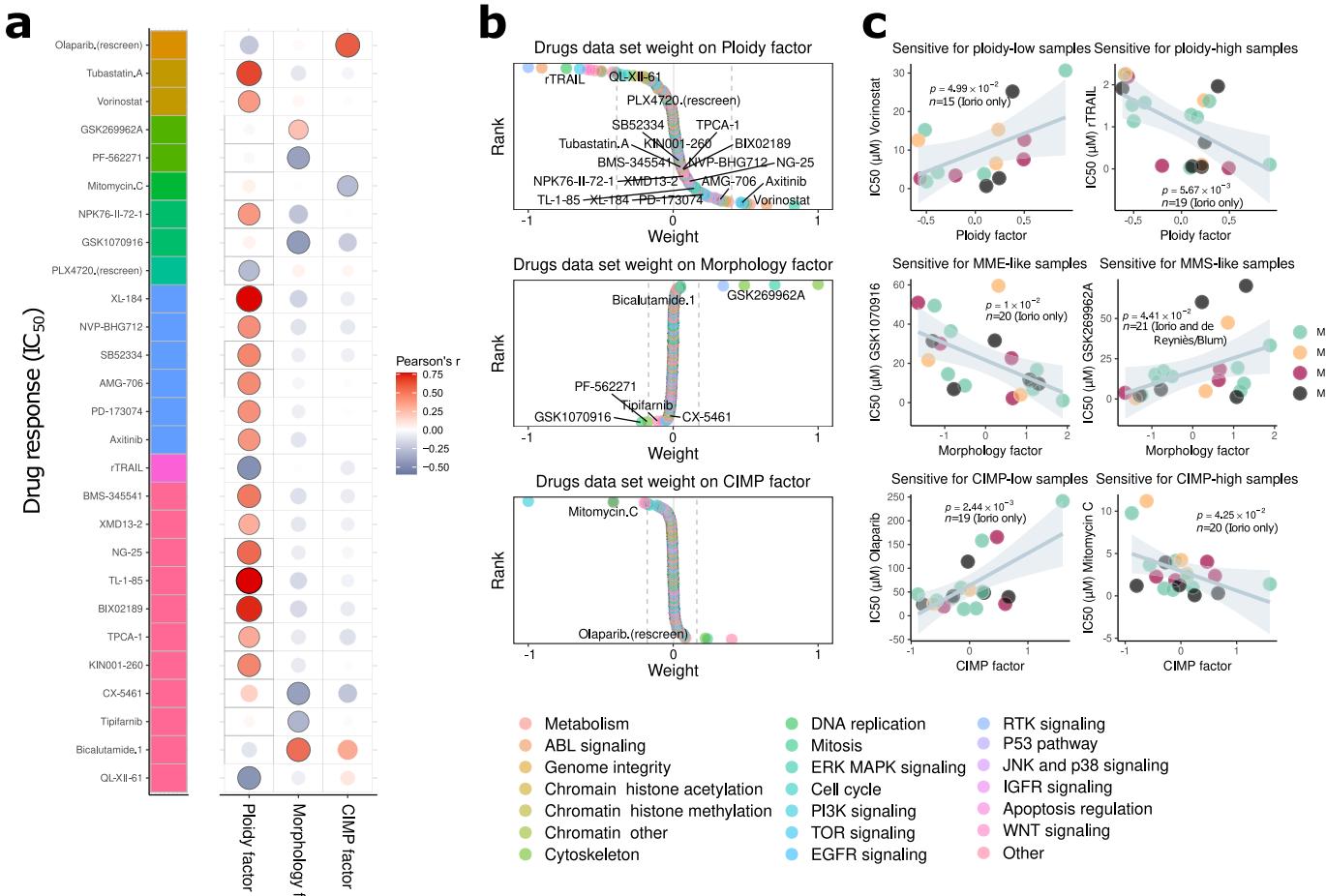
**Extended Data Fig. 4 | Replication of the prognostic value of MOFA's latent factors in other MPM cohorts.** **a)-(c)** Association between histological types and proxies for the MOFA latent factors. **a)** Association between whole genome-doubling (WGD) status and histological types in the MESOMICS and MSK-IMPACT cohorts, as determined by Fisher's exact tests. **b)** Association between the Adaptive versus innate response score and histological types, in the MESOMICS ( $n = 120$ ) and Bueno ( $n = 211$ ) and TCGA ( $n = 73$ ) cohorts, as determined by ANOVA tests. **c)** Association between the CIMP-index proxy computed on a five-gene panel and histological types in the MESOMICS and TCGA cohorts, as determined by ANOVA tests. **d)-(g)** Forest plots of hazard ratios for overall survival showing the replication of latent factors' prognostic value, using a Cox proportional hazards model. In **(b)-(c)**, boxplots represent the median and interquartile range and whiskers the maximum and minimum values, excluding outliers. **d)** WGD status (proxy for the ploidy factor) in the MESOMICS and MSK-IMPACT

cohorts. **e)** Percentage of epithelioid estimated by pathologists from H&E slides (proxy for the morphology factor) in the MESOMICS and Bueno cohorts. **f)** Adaptive versus innate response score (proxy for the adaptive-response factor), in the MESOMICS and Bueno and TCGA cohorts, computed as the difference between the proportion of lymphocyte B and T-cells minus the proportion of macrophages, monocytes, and neutrophils, estimated from gene expression data (quantITseq software). **g)** CIMP-index proxy computed on a five-gene panel (proxy for the CIMP-index factor), in the MESOMICS and TCGA cohorts. In all panels,  $P$  values indicate the significance of tests. In **(d)-(g)**, squares correspond to estimated hazard ratios and segments to their 95% confidence intervals; tests in the MESOMICS cohort (discovery) are two-sided while tests in validation cohorts (MSK-IMPACT, TCGA, or Bueno cohorts) are one-sided, in the direction found in the discovery cohort.



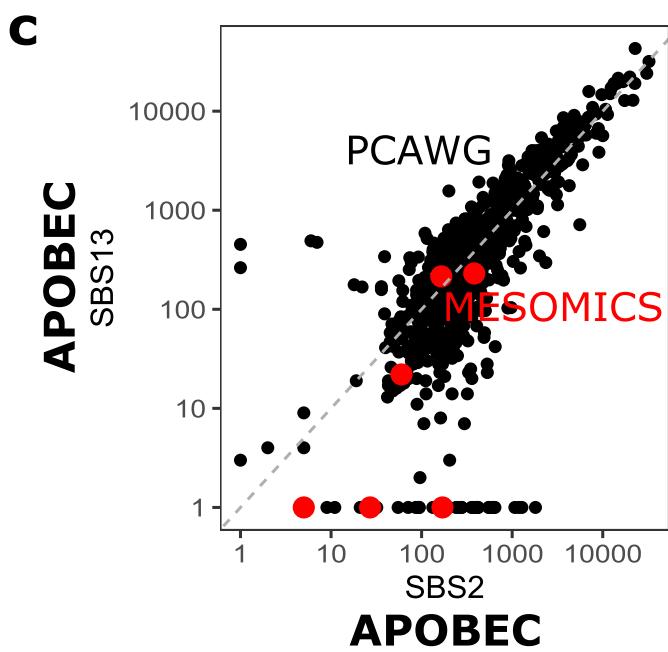
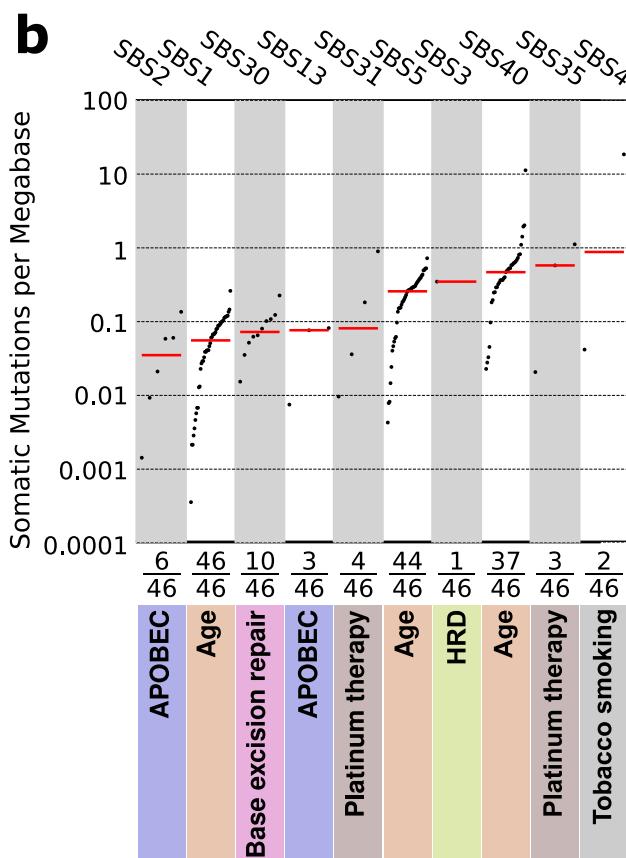
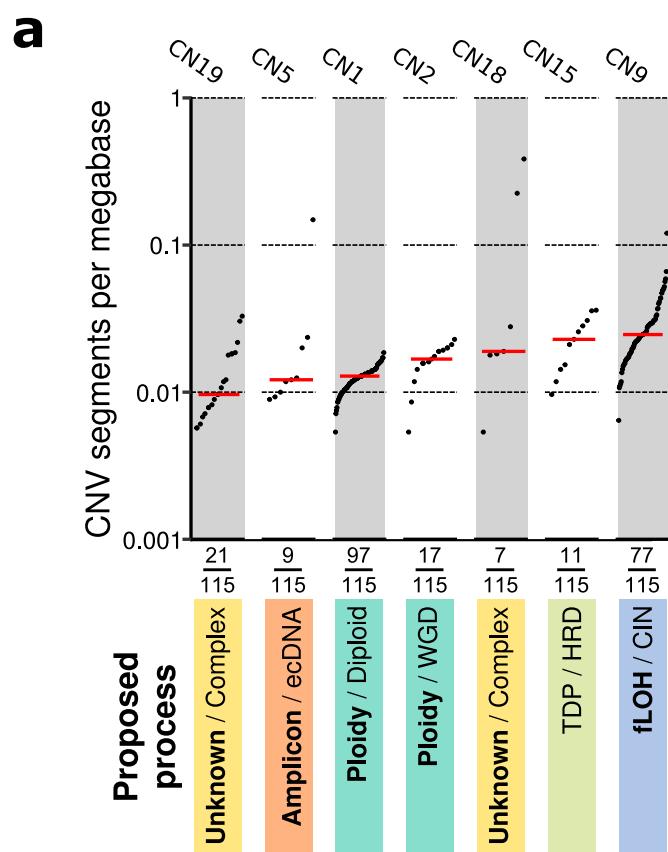
**Extended Data Fig. 5 | Performance of MOFA factors to predict survival. a)** Increase in area under the curve (AUC) as a function of percentage of change compared to histological classification. **b)** Density of survival time within the MESOMICS cohort. **c)** Integral AUC (iAUC) of twenty-two Cox proportional hazards survival models based on: (i) the three histopathological types (MME, MMB, and MMS); (ii) the proportion of sarcomatoid content; (iii) the log<sub>2</sub> ratio of CLDN15/VIM (C/V) expression proposed by Bueno and colleagues; (iv), (v) and (vi) the E score, S score, and combining both scores from Blum and colleagues, respectively; (vii) an Artificial Intelligence (AI) prognostic score; (viii-xi) the one-dimensional summary of molecular data using LFs as a continuous variable;

(xii-xvii), the two-dimensional summary of molecular data using either each combination of 2 LFs as continuous variables, respectively; (xviii-xxi), the three-dimensional summary of molecular data using each combination of 3 LFs as continuous variables; and (xxii), the four-dimensional summary of molecular data using all 4 LFs. Bars represent the mean values and error bars their standard error. Panels (a-c) present the out-of-sample accuracy within the MESOMICS cohort (4-fold cross-validation on  $n = 120$  individuals), while (d-f) present the out-of-sample accuracy within the TCGA cohort (2000 bootstraps on  $n = 73$  individuals). The model fit accuracy (no split between training and test sets) on MESOMICS and TCGA cohort are presented in Supplementary Table 17.

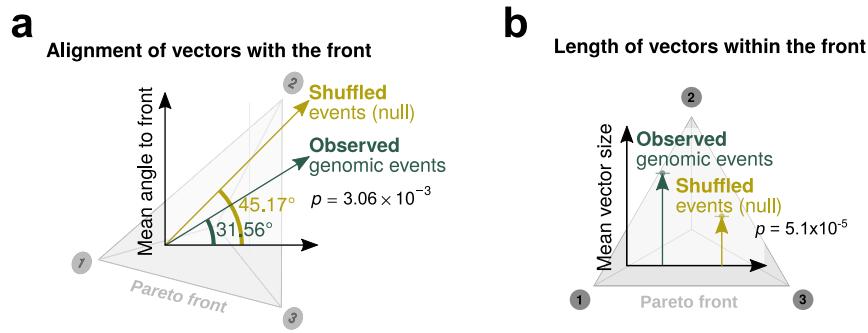


**Extended Data Fig. 6 | MOFA LFs of MPM cell lines and drug response.** **a)** Correlations between drug responses (measured by half maximal inhibitory concentration,  $\text{IC}_{50}$  in  $\mu\text{M}$ ) and MOFA LFs of cell lines. Significant associations are annotated by black point border. **b)** Distribution of drug response weights from the Drugs data set, with drugs for which the response is significantly correlated with the given LF annotated in black. Targeted pathways are

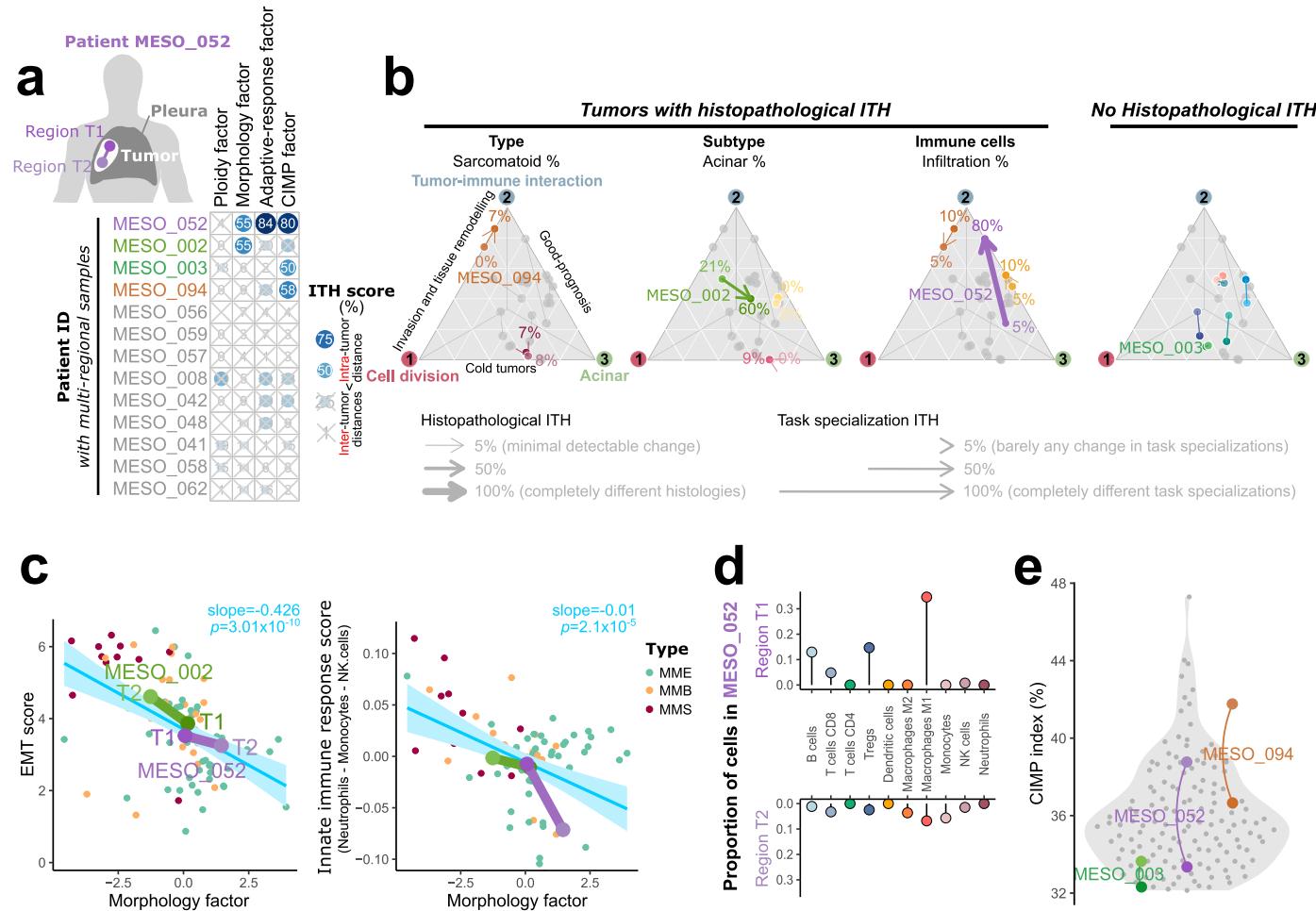
represented in **(a)** by a color bar (left), and in **(b)** by point colors. **c)** Correlations between representative drug responses significantly correlated with MOFA LFs from cell lines (left: negative correlations, right: positive associations). MPM: malignant pleural mesothelioma not otherwise specified. Gray bands correspond to 95% confidence intervals. Pearson correlation coefficients and the associated two-sided  $P$  values are displayed in **(a)** and **(c)**.



**Extended Data Fig. 7 | Tumor burden of mutational signatures.** Tumor Mutational Burden of **a**) 7 copy number signatures ( $n = 115$  biologically independent samples) and **b**) 10 Single Nucleotide Variant Signatures detected in the MESOMICS cohort ( $n = 46$  biologically independent samples). Note that although SBS40 is associated with age in many cancers, its etiology is still unknown. TDP: tandem duplicator phenotype; HRD: homologous recombination deficiency; fLOH: focal loss of heterozygosity; CIN: chromosomal instability.

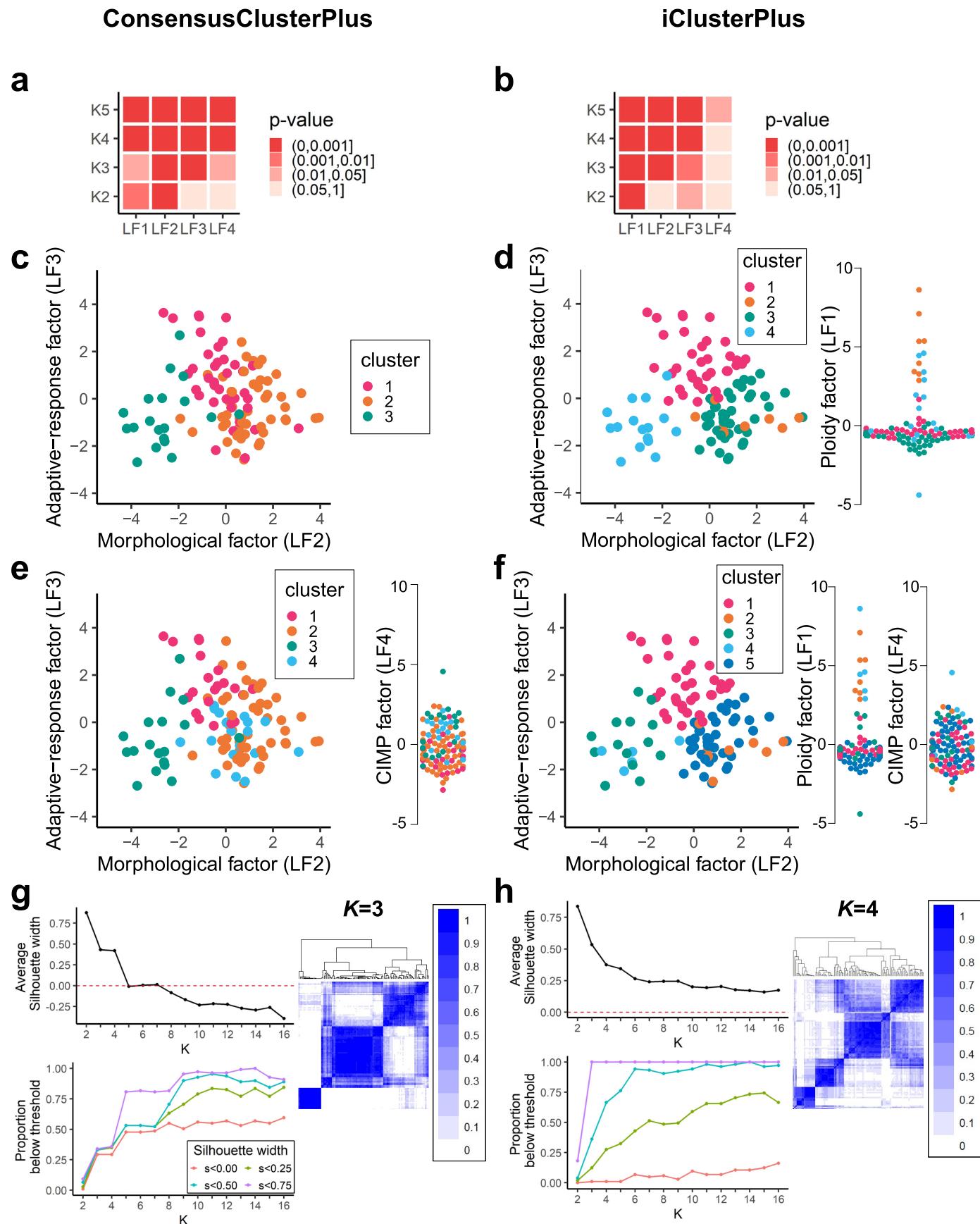


**Extended Data Fig. 8 | Test of the impact of genomic events on cancer task specialization.** **a)** Alignment of vectors with the Pareto front in degree (0°: perfectly aligned, 90°: completely orthogonal) and **(b)** length of the vector. *P* values correspond to two-sided Wilcoxon tests between observed and shuffled vector distributions.



**Extended Data Fig. 9 | Multomics intra-tumor heterogeneity (ITH) of 13 multiregion samples.** **a**) Intratumor heterogeneity (ITH) score, ranging from 0% (no ITH) to 100% (ITH greater than the maximum observed intertumor heterogeneity in the cohort), for each sample (row) and each MOFA latent factor (column). The score is computed as the percentage of inter-tumor distances in a MOFA factor that are lower than the observed intratumor distance between regions. The four samples with ITH score greater than 50% are highlighted in color. **b**) Relationship between histopathological heterogeneity and cancer task specialization. Ternary plots depicting task specialization in three cancer tasks (see Fig. 2). For each histopathological feature, a colored arrow connects regions from tumors with differences in this feature. Numbers correspond to the percentage of this feature in the tumor as estimated by our pathologist. The right

ternary plot represents all samples with no histopathological ITH. **c**) Epithelial to mesenchymal transition (EMT) score and innate immune composition score as a function of MOFA's Morphology factor. Small points correspond to all samples from the MESOMICS cohort, and large points connected by segments to regions from the 3 patients with CIMP factor ITH highlighted in (a). Blue bands correspond to 95% confidence intervals, and  $P$  values to two-sided  $t$ -tests. **d**) Lollipop plot of the estimated proportion of immune cells in two regions of a sample with ITH in the adaptive-response factor highlighted in (a). **e**) CIMP index in regions of two tumors with substantial ITH in the CIMP factor highlighted in (a) (colored points connected by an arc), compared to that of the rest of the cohort (grey points).



Extended Data Fig. 10 | See next page for caption.

**Extended Data Fig. 10 | Association between MOFA latent factors and the clusters identified by consensus clustering (a-d) and integrative clustering (e-h).** **a**) Kruskal-Wallis rank sum test significance ( $P$  value) between each  $K$  (row) and the LFs (column), for  $K$  from 2 to 5 from consensus clustering results and the first four LFs. **b**) Kruskal-Wallis rank sum test significance ( $P$  value) between each  $K$  (row) and the LFs (column), for  $K$  from 2 to 5 from integrative clustering results and the first four LFs. **c**) Consensus clustering results for  $K = 3$ . Samples are visualized in MOFA latent factor space of LF2 vs. LF3 and colored by the consensus clustering results. **d**) Integrative clustering results for  $K = 4$ . Samples are visualized in MOFA latent factor space of LF2 vs. LF3 and colored by the integrative clustering results. On the right, we show the samples in one-dimensional space of LF1 using beeswarm plot. **e**) Consensus clustering results for  $K = 4$ . **f**) Integrative clustering results for  $K = 5$ . Samples are visualized in MOFA

latent factor space of LF2 vs. LF3 and colored by the integrative clustering results. On the right, we show the samples in one-dimensional space of LF1 and LF4 using beeswarm plot. **g**) Top-left: average silhouette width for consensus clustering with different  $K$ . Bottom-left: proportion of samples below the selected silhouette width threshold for consensus clustering with different  $K$ . Right: consensus matrix heatmap for  $K = 3$ . Color gradient represents consensus values from 0–1. **h**) Top-left: average silhouette width for integrative clustering with different  $K$ . Bottom-left: proportion of samples below the selected silhouette width threshold for integrative clustering with different  $K$ . Right: heatmap of the frequencies of samples being clustered together among all clustering results using the set of iClusterPlus lambda values for  $K = 4$ . Color gradient represents consensus values from 0–1.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give P values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used for data collection.
Data analysis	<p>WGS reads mapping: in-house workflow (<a href="https://github.com/IARCbioinfo/alignment-nf">github.com/IARCbioinfo/alignment-nf</a> v1.0) that used bwa version 0.7.15, GATK version 4.0.12, samblaster version 0.1.24, sambamba version 0.6.6.</p> <p>Variant calling: Mutect2 from GATK version 4.1.5, annotation with ANNOVAR version of April 16th 2018.</p> <p>Copy number variant calling: PURPLE version 2.52 and Gistic2 version 2.20.23.</p> <p>Structural variant calling: SVaba version 1.1.0, Delly version 0.8.3, Manta version 1.6.0, and SURVIVOR version 1.0.7.</p> <p>Amplicon pattern calling: AmpliconArchitect version 1.2 and CNVkit version 0.9.7.</p> <p>RNA-seq data processing: in-house workflow (<a href="https://github.com/IARCbioinfo/RNAseq-nf">github.com/IARCbioinfo/RNAseq-nf</a> v2.3) that used Trim Galore (version 0.6.5 for expression quantification, and version 0.4.2 for alternative splicing analyses), STAR version 2.7.3a and StringTie version 2.1.2.</p> <p>Quality control with FastQC version 0.11.9 and RSeQC version 3.0.1.</p> <p>Fusion transcript discovery: realignment with STAR version 2.7.6a followed by Arriba version 2.1.0.</p> <p>Immune contexture quantification: quanTLseq version of July 2020.</p> <p>Statistical analyses with R version 4.0.3</p>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The genome sequencing data, RNA-seq data, and methylation data have been deposited in the European Genome-phenome Archive (EGA) database, which is hosted at the EBI and the CRG, under accession number EGAS00001004812. Because raw -omics datasets derived from humans are at risk of re-identification when combined with information from other public sources, access must be requested to the MESOMICS data access committee (DAC) as detailed at <https://ega-archive.org/studies/EGAS00001004812>. Minimum datasets of processed somatic alterations for genomic, transcriptomic, and epigenomic data, sufficient to reproduce, interpret and extend our main results, are publicly available at [https://github.com/IARCbioinfo/MESOMICS\\_data/tree/main/phenotypic\\_map](https://github.com/IARCbioinfo/MESOMICS_data/tree/main/phenotypic_map). A data note manuscript detailing all quality controls of the dataset is available at <https://www.biorxiv.org/content/10.1101/2022.07.06.499003v16>. TCGA whole-exome sequencing, RNA-seq, and methylation array data are available from the GDC portal (TCGA-MESO cohort), the whole-exome sequencing and RNA-seq data from the Bueno and colleagues cohort are available from the European Genome-phenome Archive, EGA:EGAS00001001563. Small variants lists, RNA-seq, expression array, and methylation data for the Iorio and colleagues cohort are available from the GEO (GSE29354), EGA (EGAS00001000828), and SRA (PRINAS23380) websites, and corresponding drug responses are available from the cancerrxgene.org website ([https://www.cancerrxgene.org/downloads/drug\\_data?tissue=MESO](https://www.cancerrxgene.org/downloads/drug_data?tissue=MESO); accessed July 2021). Expression array data for the de Reyniès and colleagues cohorts are available from the ArrayExpress platform (E-MTAB-1719), and corresponding drug response data from the supplementary material of Blum et al. All the other data supporting the findings of this study are available within the article and its supplementary information files.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

### Sample size

This study first includes an unsupervised analysis of genomes, RNA-seq, and methylation data using group factor analysis (software MOFA). For such multivariate analyses, no simple power calculations are available, but recommendations for stable latent factors and weights such that the results from the sample can accurately be generalized to the population suggest  $n>100$  (Saccetti and Timmerman J. Proteom. Res. 2016), which is in line with the sample size in our study (total of  $n=120$ ). This also ensured that the Pareto task inference relying on these factors had a sufficient sample size. All omic and clinical data were available for at least  $n=100$  samples, ensuring that correlation tests have a power of 80% to detect non-zero coefficients  $r>0.277$  given a type I error rate of 5% (Hulley et al. Lippincott Williams & Wilkins 2013), and thus ensuring that downstream analyses such as gene set enrichment analysis relying on correlation test p-values are also properly powered. Combining the MESOMICS, Bueno, and TCGA cohorts allows to reach  $n=300$  samples with exonic sequencing data, allowing to detect recurrent alterations with a prevalence of 1% in malignant pleural mesothelioma in at least 2 samples with a power of 80% (based on the binomial distribution).

### Data exclusions

Pre-established exclusion criteria were as follows: samples were excluded if they were derived from metastasis or a participant who had undergone chemotherapy for the treatment of malignant pleural mesothelioma. These samples were excluded as we aimed to identify whether particular molecular patterns or profiles in primary tumours were associated with disease progression and patient prognosis, therefore we examined only the molecular profile of primary untreated mesothelioma samples obtained at diagnosis.

### Replication

We replicated our main findings independently on three external datasets: two of similar malignant pleural mesothelioma cohorts of sizes  $n=181$  from Bueno et al. Nat. Genet. 2016 and  $n=73$  from Hmeljak et al. 2019, and one of  $n=59$  cancer cell lines from de Reynies et al. Clin. Cancer Res. 2014 and Iorio et al. Cell 2016.

### Randomization

Samples were assigned a histopathological class through central pathological review. We assessed the importance of covariables like age, sex, smoking status, and asbestos exposure in multivariate analyses using regression analyses. We controlled for these variables in statistical analyses by adding them as covariables in the regressions when appropriate.

### Blinding

The investigators were not blinded to the histopathological class during the unsupervised analyses, but this was unnecessary because these analyses do not take into account any sample information except the genomic alterations, gene expression, and methylation levels. The investigators could not be blinded to the subsequent differential analyses, which required knowledge of the group membership for each sample.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

- |     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| n/a | <input type="checkbox"/> Involved in the study<br><input checked="" type="checkbox"/> Antibodies<br><input checked="" type="checkbox"/> Eukaryotic cell lines<br><input checked="" type="checkbox"/> Palaeontology and archaeology<br><input checked="" type="checkbox"/> Animals and other organisms<br><input type="checkbox"/> Human research participants<br><input checked="" type="checkbox"/> Clinical data<br><input checked="" type="checkbox"/> Dual use research of concern |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## Methods

- |     |                                                                                                                                                                                                                    |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| n/a | <input type="checkbox"/> Involved in the study<br><input checked="" type="checkbox"/> ChIP-seq<br><input checked="" type="checkbox"/> Flow cytometry<br><input checked="" type="checkbox"/> MRI-based neuroimaging |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## Antibodies

### Antibodies used

Santa Cruz BAP1 antibody (cloneC-4) catalog number sc-28383

### Validation

BAP1 antibody was purchased from Santa Cruz (catalog number sc-28383). Immunohistochemistry was performed on FFPE tumor sections 3µm tissue sections after pretreatment using a Ventana's Immunostainer (BenchMark Ultra). After pre-treatment (CC1 solution), primary antibody was applied (dilution 1:50) during 30 min. The detection kit ultra-view (Universal DAB Detection kit, Ventana) was used following manufacturer's recommendations. BAP1 (clone C4) (Santa Cruz: dilution one to 50) nuclear staining was considered positive (when nuclear expression was retained) or negative (complete loss of staining of all tumor cells with a positive internal control on the slides [fibroblast, lymphocytes, etc.]).

## Human research participants

### Policy information about [studies involving human research participants](#)

#### Population characteristics

Age: median of 67.5  
 Sex: 32 females and 88 males  
 n = 120 chemonaive at the moment of sample collection and three no chemonaive removed from the analyses.  
 The samples used in this study belong to the virtual biorepository French MESOBANK, whose guidelines include obtaining the informed consent from all subjects and no participant compensation.

#### Recruitment

There has not been a prospective recruitment of patients specifically done for this study. The biological specimens were previously collected for clinical routine and stored in the biobanks of the collaborative hospitals, which made the de-identified samples available for research.

#### Ethics oversight

International Agency for Research on Cancer ethics committee

Note that full information on the approval of the study protocol must also be provided in the manuscript.