# DNA Methylation Signatures Reveal the Diversity of Processes Remodeling Hepatocellular Carcinoma Methylomes

Léa Meunier, Theo Z Hirsch, Stefano Caruso, Sandrine Imbeaud, Quentin Bayard, Amélie Roehrig, Gabrielle Couchy, Jean-charles Nault, Josep Llovet, Jean-frédéric Blanc, et al.

MISS LÉA MEUNIER (Orcid ID : 0000-0003-1855-7467)
PROF. JESSICA ZUCMAN-ROSSI (Orcid ID : 0000-0002-5687-0334)
DR. ERIC LETOUZÉ (Orcid ID : 0000-0002-6369-2839)

Article type : Original Article

# DNA Methylation Signatures Reveal the Diversity of Processes Remodeling Hepatocellular Carcinoma Methylomes

Léa Meunier,[1] Théo Z. Hirsch,[1] Stefano Caruso,[1] Sandrine Imbeaud,[1] Quentin Bayard,[1] Amélie Roehrig,[1] Gabrielle Couchy,[1] Jean-Charles Nault[1-3] Josep M. Llovet,[4-6] Jean-Frédéric Blanc,[7-9] Julien Calderaro,[10] Jessica Zucman-Rossi,[1,11*] and Eric Letouzé[1*]

From the [1]Centre de Recherche des Cordeliers, Sorbonne Université, INSERM, Université de Paris, Université Paris Nord, Functional Genomics of Solid Tumors Laboratory, Equipe Labellisée Ligue Contre le Cancer, Paris, France; [2]Service d'Hépatologie, Hôpital Jean Verdier, Hôpitaux Universitaires Paris-Seine-Saint-Denis, Assistance-Publique Hôpitaux de Paris, Bondy, France; [3]Unité de Formation et de Recherche Santé Médecine et Biologie Humaine, Université Paris 13, Communauté d'Universités et Etablissements Sorbonne Paris Cité, Paris, France; [4]Mount Sinai Liver Cancer Program, Division of Liver Diseases, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY; [5]Translational Research in Hepatic Oncology, Liver Unit, IDIBAPS, Hospital Clinic, University of Barcelona, Barcelona, Catalonia, Spain; [6]Institució Catalana d'Estudis Avançats (ICREA), Barcelona, Catalonia, Spain; [7]Department of Hepato-Gastroenterology and Digestive Oncology, CHU de Bordeaux, Haut-Lévêque Hospital, Bordeaux, Aquitaine, France; [8]Department of Pathology, CHU de Bordeaux, Pellegrin Hospital, Bordeaux, Aquitaine, France; [9]Bordeaux Research in Translational Oncology, Université Bordeaux, Bordeaux, Aquitaine, France; [10]Service d'Anatomopathologie, Hôpital Henri Mondor; Université Paris Est, INSERM U955, Team 18, Institut Mondor de Recherche Biomédicale, Créteil, France; [11]Hôpital Européen Georges Pompidou, Assistance Publique-Hôpitaux de Paris, Paris, France.

*These authors contributed equally to this work.

**Keywords:** hepatocellular carcinoma; DNA methylation; epigenetic reprogramming; integrated genomic analysis; independent component analysis.

**ADDRESS CORRESPONDENCE AND REPRINT REQUESTS TO:**
Eric Letouzé, PhD
Centre de Recherche des Cordeliers
15 Rue de l'École de Médecine
75006 Paris, France
E-mail: eric.letouze@inserm.fr
Tel.: +33 (0)1 44 27 80 84
OR
Jessica Zucman-Rossi, MD, PhD
Centre de Recherche des Cordeliers
15 Rue de l'École de Médecine
75006 Paris, France
E-mail: jessica.zucman-rossi@inserm.fr
Tel.: +33 (0)1 44 27 80 84

**Abbreviations:** ARID1A, AT-rich interactive domain-containing protein 1A; BAP1, BRCA1-associated protein 1; bp, base pair; bval, beta value; CCN, cyclin; CCNA2/E1, cyclin A2/E1; CGI, CpG island; CTNNB1, catenin beta 1; HBV, hepatitis B virus; HCC, hepatocellular carcinoma; HCV, hepatitis C virus; HMD, highly methylated domain; H3K27ac, acetylation of histone H3 on lysine 27; H3K4Me1, monomethylation of histone H3 at lysine 4; H3K4Me3, trimethylation of histone H3 at lysine 4; ICA, independent component analysis; ICF, immune-mediated cancer field; LICA-FR, Liver Cancer (France); LMR, lowly methylated region; MC, methylation component; MethICA, methylation signature analysis with independent component analysis; MRCpG, most representative CpG; NT, non-tumor; PMD, partially methylated domain; RNA-seq, RNA sequencing; RPS6KA3, ribosomal protein S6 kinase A3; TCGA, The Cancer Genome Atlas; TCGA-LIHC, TCGA Liver Hepatocellular Carcinoma; t-SNE, t-distributed stochastic neighbor

embedding; TERT, telomerase reverse transcriptase; TSS, transcription start site; UMR, unmethylated region.

**Financial Support:** This study was supported by grants from the Institut du Cancer (INCa) within the framework of the ICGC project and MUTHEC project (INCa translationnel PRTK2014), France Génomique, Cancéropole Ile de France (ExhauTrans project), ITMO Cancer AVIESAN (Alliance Nationale pour les Sciences de la Vie et de la Santé, National Alliance for Life Sciences and Health) within the framework of the Cancer Plan ("HTE program-HetColi network" and "Cancer et environnement program"), BPI France (ICE project), Agence Nationale de Recherche sur le Sida et les hépatites virales (ANRS), and the French Liver Biobanks network (INCa, BB-0033-00085, Hepatobio bank). The group is supported by the Ligue Nationale Contre le Cancer (Equipe Labellisée), Labex OncoImmunology (investissement d'avenir), Coup d'Elan de la Fondation Bettencourt-Shueller, the SIRIC CARPEM, and Fondation Mérieux. The study was also supported by a fellowship from the HOB doctoral school and the ministry of Education and Research to L.M. and Q.B.; a fellowship from Cancéropole Ile de France and Fondation d'Entreprise Bristol-Myers Squibb pour la Recherche en Immuno-Oncologie to T.Z.H.; CARPEM and the Labex OncoImmunology to S.C.; Fondation pour la Recherche Médicale to A.R.; and the Accelerator Award (CRUCK, AEEC, AIRC) (HUNTER, Ref. C9380/A26813), National Cancer Institute (P30-CA196521), US Department of Defense (CA150272P3), Samuel Waxman Cancer Research Foundation, Spanish National Health Institute (PID2019-105378RB-100), and the Generalitat de Catalunya/AGAUR (SGR-1358) to J.M.L.

**ABSTRACT**

DNA methylation patterns are highly rearranged in hepatocellular carcinomas (HCCs). However, diverse sources of variation are intermingled in cancer methylomes, precluding the precise characterization of underlying molecular mechanisms. We developed a computational framework (methylation signature analysis with independent component analysis [MethICA]), leveraging independent component analysis (ICA) to disentangle the diverse processes contributing to DNA methylation changes in tumors. Applied to a collection of 738 HCCs, MethICA unraveled 13 stable methylation components (MCs) preferentially active in specific chromatin states, sequence contexts, and replication timings. These included signatures of general processes associated with gender and age but also new signatures related to specific driver events and molecular subgroups. Catenin beta 1 (*CTNNB1*) mutations were major modulators of methylation patterns in HCC, characterized by a targeted hypomethylation of transcription factor 7 (TCF7)-bound enhancers in the vicinity of Wnt target genes as well as a widespread hypomethylation of late-replicated partially methylated domains (PMDs). By contrast, demethylation of early-replicated highly methylated domains (HMDs) was a signature of replication stress, leading to an extensive hypomethylator phenotype in cyclin (CCN)-activated HCC. Inactivating mutations of the chromatin remodeler AT-rich interactive domain-containing protein 1A (ARID1A) were associated with epigenetic silencing of differentiation-promoting transcriptional networks, also detectable in cirrhotic liver. Finally, a hypermethylation signature targeting Polycomb-repressed chromatin domains was identified in the G1 molecular subgroup with progenitor features. *Conclusion:* This study elucidates the diversity of processes remodeling HCC methylomes and reveals the epigenetic and transcriptional impact of driver alterations.

Hepatocellular carcinoma (HCC), the third most deadly cancer worldwide, is a heterogeneous disease that usually develops in a context of cirrhosis, related to diverse risk factors, such as hepatitis B virus (HBV) or hepatitis C virus (HCV) infection, alcohol intake, or metabolic syndrome.[1] HCCs are heterogenous at the molecular level, with up to six distinct transcriptional subgroups[2-4] and >30 driver genes belonging to 11 major pathways.[4-6] After telomerase reverse transcriptase (*TERT*) promoter (60% of HCC cases), *TP53,* and catenin beta 1 (*CTNNB1*) (25%-30%), chromatin remodeling is the most frequently altered pathway with recurrent mutations in AT-rich interactive domain-containing protein 1A (*ARID1A*) (13%) and *ARID2* (7%) genes. Epigenetic regulation is also strongly altered in HCC. In particular, DNA methylation changes are widespread, including hypermethylation of CpG islands (CGIs) and an extensive hypomethylation in open sea regions.[7-10] HCCs display heterogeneous methylation landscapes. DNA methylation-based classifications revealed between three and seven HCC subgroups showing more or less widespread hypo- and hypermethylation changes.[4,11,12] Methylation markers are also valuable for early HCC detection[13] and prognosis.[14] However, the molecular mechanisms causing these changes and the relationship between DNA methylation signatures and driver genes, including epigenetic regulators, remain largely unknown.

The DNA methylation landscape of human cancers is modulated by various factors, including the cell of origin,[15] age-related processes,[16,17] environmental exposures,[18] driver alterations,[19] deregulated oncogenic pathways,[20] and stromal cell composition.[21] Thus, the DNA methylation profile of each tumor reflects the addition of many processes, operative with different strengths and during different time windows in tumor history. The DNA methylation signatures of these processes are intermingled in the final tumor methylome, precluding the precise characterization of underlying molecular mechanisms. Blind source separation methods are dedicated to the deconvolution of independent signals intermingled in a data set, and these methods have shown promising applications to cancer biology.[22] Non-negative matrix factorization (NMF) is widely used to uncover signatures of mutational processes in cancer genomes.[23] Independent component analysis (ICA) has been shown to outperform principal component analysis and clustering-based methods to identify biologically meaningful transcriptomic components in cancers.[24] However, these methods have not yet been applied to analyze DNA methylation changes in cancer.

Here, we present the methylation signature analysis with independent component analysis (MethICA) statistical framework, leveraging ICA to disentangle independent sources of variation in methylation data. Applying MethICA to a collection of 738 HCCs with extensive clinical and molecular data, we show that the methylome of each tumor reflects a unique combination of 13 ubiquitous and tumor-specific processes. We unravel DNA methylation signatures induced by several driver genes and their transcriptional consequences, providing insights into the causes and roles of DNA methylation changes in the pathogenesis of HCC.

## MATERIAL AND METHODS

### *Liver Cancer (France) cohort*

A series of 274 samples—239 HCCs, including 4 fibrolamellar carcinomas (FLCs), and 35 adjacent non-tumor (NT) liver tissues—were collected from patients surgically treated in four French hospitals located in the Bordeaux and Paris regions. The study was approved by the institutional review board committees (CCPRB Paris Saint-Louis, 1997, 2004, and 2010, approval number 01–037; Bordeaux, 2010, A00498–31). Written informed consent was obtained in accordance with French legislation. Of the 239 HCC cases, 105 (44%) developed in non-fibrotic (METAVIR F0-F1), 55 (23%) in chronic hepatitis (F2-F3), and 78 (33%) in cirrhotic liver (F4). Clinicopathological data were available for all cases. The Liver Cancer (France) (LICA-FR) cohort mostly comprises males (81%), with a median age at sampling of 65 years, related to diverse risk factors, including alcohol (45%), HBV (18%), and HCV (16%). The 274 samples were analyzed using Illumina Infinium HumanMethylation450 BeadChip arrays for this study (see below). Somatic mutations were available for 209 samples previously analyzed by whole-genome sequencing (WGS) or whole-exome sequencing (WES).[5,25-28] For samples that were not analyzed using these techniques, gene mutation data were completed using MiSeq or Sanger sequencing, as described.[29] RNA sequencing (RNA-seq) data were also available for 145 tumor and 5 NT samples.[25,28] Gene expression of a panel of 190 genes was also analyzed in 229 HCCs by quantitative reverse-transcription polymerase chain reaction (qRT-PCR) on Fluidigm 96 dynamic arrays to classify HCC in the G1-G6 molecular groups, as described.[2,29]

Detailed clinical characteristics and sequencing details for each sample are provided in Supporting Table S1.

The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC)[4] and genomic predictors and oncogenic drivers in HCC (HEPTROMIC)[14] cohorts are described in the Supporting Methods.

### DNA methylation arrays

We analyzed the 274 samples from the LICA-FR cohort using Illumina Infinium HumanMethylation450 BeadChip arrays. Microarray experiments were performed by Integragen SA (Evry, France). In brief, genomic DNA was bisulfite-converted using the EZ-96 DNA

Methylation Kit (Zymo Research, Irvine, CA, USA), whole-genome amplified, enzymatically fragmented, and hybridized to the BeadChip arrays in accordance with the manufacturer's instructions. The beta value (bval) DNA methylation scores for each locus were extracted together with detection *P* values from Illumina GenomeStudio software. The bval gives an estimate of the methylation level of each CpG locus using the ratio of intensities between methylated and unmethylated probes. We removed CpGs with "NA (not available)" values or a detection *P* value >0.05 in more than 20% of the samples, leaving 351,509 probes for analysis.

The two other cohorts (TCGA-LIHC and HEPTROMIC) were analyzed with the same methylation array. We retrieved the bval and detection *P* value matrices for these two data sets and selected reliable CpGs as we did for the LICA-FR cohort.

**RNA-seq data processing**

RNA-seq read counts per gene were obtained for the LICA-FR cohort, as described,[25] and directly from TCGA website for TCGA-LIHC cohort. We then applied the same pipeline to the raw counts of the two series to obtain normalized fragments per kilobase of exon per million reads mapped (FPKM) and variance stabilizing transformation (VST) matrices. We used DESeq2[30] to import raw read counts into R statistical software and apply VST to the raw count matrix. FPKM scores were calculated by normalizing the count matrix for the library size and the coding length of each gene.

**ICA**

We restricted each data set to the 200,000 most variant CpGs based on their standard deviation. We computed 20 independent methylation components (MCs) in each cohort using the FastICA algorithm,[31] as implemented in the sklearn.decomposition Python library, with a first step of whitening of the matrix, the function of approximation to neg-entropy *logcosh*, and *parallel* algorithm. Because the FastICA algorithm involves random initialization, we performed 100 iterations and kept the results from the most stable iteration. A component was considered "stable" when a similar component (Pearson correlation of CpG contribution >0.9) was identified in 50% or more of the iterations. We selected the iteration giving the highest number of stable components and the highest average Pearson correlation score among stable components.

We next compared the results obtained for the three data sets. The similarity of two components from two different data sets was determined by calculating the absolute value of the Pearson correlation coefficient from the contribution of their common CpGs. For further analysis, we selected the 13 most reliable components found in at least two of the three HCC data sets with a Pearson correlation score >0.45.

**Association between MCs and (epi)genomic features**

To better understand the preferential activity of each component toward specific regions, we analyzed the enrichment of their most contributing CpG sites across diverse types of (epi)genomic features. We selected the most representative CpG (MRCpG) sites of each MC by thresholding their absolute projections onto the MC: abs(projection) >0.005. We next estimated the enrichment of these MRCpGs across diverse (epi)genomic features (see the description of features and their sources in the Supporting Methods). To do so, we calculated an enrichment score (ES) for each feature corresponding to the ratio between the proportion of the most contributing CpGs being located within the feature and the proportion of the 200,000 analyzed CpGs being located within the feature:

$$ES = \frac{\frac{N_{feature}^{contrib}}{N^{contrib}}}{\frac{N_{feature}^{all}}{200{,}000}}$$

$N_{feature}^{contrib}$ indicates the number of most contributing CpGs located within the feature; $N^{contrib}$, the number of most contributing CpGs; and $N_{feature}^{all}$, the number of CpGs located within the feature among the 200,000 analyzed CpGs.

**Association between MCs and clinico-molecular annotations**

We analyzed the association of each MC with more than 50 clinical and molecular features. For this part, we chose to focus on the LICA-FR and TCGA-LIHC cohorts, for which extensive clinical and molecular data were available. The full list of clinical and molecular features included in the analysis is provided in the Supporting Methods. We first used linear models to identify features significantly correlated with sample contributions, using the *lm* function in R statistical software: *lm(sample contribution ~ annotation)*. Only positive associations were considered (i.e., features associated with an increased activity of a component). For example, mutation of a given

driver gene was considered to be associated with a component only if mutated cases displayed a higher activity of the MC, to favor the identification of causal factors rather than indirect associations. This step was done separately in the LICA-FR and TCGA-LIHC cohorts. Clinico-molecular features that were significant (*P* value <0.005) in both cohorts were then included in multivariate analyses also using the *lm* function: *lm(sample contribution ~ all selected annotations)*. We defined the most contributing features of each MC as those that remained significant (*P* value <0.05) in multivariate analysis in both cohorts.

**Data availability**

The DNA methylation data generated for this study (274 tumor and NT liver tissues analyzed with Illumina Infinium HumanMethylation450 BeadChip arrays) have been deposited to the Gene Expression Omnibus database (accession number: GSE157341).

MethICA is an open-source collaborative initiative available in the GitHub repository FunGeST/MethICA.

## RESULTS

### Independent component analysis of liver cancer methylomes

To unravel the diverse epigenetic processes remodeling liver cancer methylomes, we analyzed three independent data sets (LICA-FR, n = 274; TCGA-LIHC,[4] n = 325; HEPTROMIC,[14] n = 243) totaling 738 HCC and 104 non-tumor (NT) liver samples, all profiled with Illumina Infinium HumanMethylation450 BeadChip arrays (Supporting Table S1). We first performed ICA within each cohort to decompose the DNA methylation matrix as a mixture of 20 independent methylation components (MCs), each characterized by a specific pattern of activation across samples and across CpG sites (Fig. 1A). To evaluate the reproducibility of the results, we quantified the correlation of MCs across the three data sets based on the contributions of CpG sites. A total of 13 components (MC1-MC13) were highly reproducible and shared by at least two data sets (Pearson correlation >0.45), 11 of which were identified in the three data sets (Supporting Fig. S1).

Then, we identified for each component a set of most representative CpG sites (MRCpGs, i.e., CpGs with the strongest contribution to the component), and we examined the DNA methylation changes across these MRCpGs in the 5% of tumors with the strongest deviation from NT liver tissues (Fig. 1B). MC1-MC3 were dominated by hypermethylation, MC10-MC13 were dominated by hypomethylation, and MC4-MC9 showed a combination of hyper- and hypomethylation. The range of methylation changes also varied strongly across components. MC10 and MC11 involved hypomethylation of CpG sites that are highly methylated in NT liver (median bval >0.87), whereas MC12 and MC13 involved hypomethylation of CpG sites with intermediate methylation levels (median bval ~0.7). MC1 and MC2 both involved hypermethylation of CpG sites with low methylation in NT liver (median bval = 0.14), but the median methylation increase was only 0.36 in MC1 versus 0.52 in MC2. Thus, each component displays its own dynamics of methylation changes.

### Methylation components are preferentially active in specific chromatin states and sequence contexts

We next examined whether the MRCpGs of each component were preferentially located within specific CGI-based features (island, shore, shelf, or outside CGI), gene-based features (transcription start site [TSS] ± 500 bases, gene body, or intergenic), or chromatin states (Fig. 2A

and Supporting Fig. S2). Chromatin states were defined by the Roadmap consortium based on the chromatin immunoprecipitation (ChIP)-seq analysis of six different histone modifications in normal liver tissue.[32] Although histone marks are altered in cancer cells, we observed a good agreement between chromatin states defined in normal liver tissue and in the liver cancer cell line HepG2 (Supporting Fig. S3). Thus, normal liver chromatin states likely reflect reasonably well the actual chromatin state at the time DNA methylation changes occur. We also investigated the methylation domains and sequence contexts of the MRCpGs of each component (Fig. 2B). We first used normal liver whole-genome bisulfite sequencing (WGBS) data[33] to identify CpGs located in large (megabase-scale) partially methylated domains (PMDs)[34] and highly methylated domains (HMDs) or in short (hundreds to thousands of base pairs [bps]) lowly methylated regions (LMRs) and unmethylated regions (UMRs). LMRs and UMRs correspond respectively to distal and proximal regulatory elements.[35] We then classified the sequence context around each CpG dyad into 12 categories as described by Zhou et al.,[17] taking into account the local CpG density (number of CpG sites within 35 bps on each side of the dyad) and the nucleotides directly flanking the CpG (S = C or G; W = A or T).

The MRCpGs of hypermethylation components (MC1-MC3) were preferentially located in CGIs, TSSs, and UMRs but displayed different chromatin state enrichment patterns: mostly bivalent and Polycomb-repressed chromatin for MC2, active TSS for MC3, and a mixture for MC1 (Fig. 2A). Hypomethylation components (MC10-MC13) were associated with inactive chromatin domains (Fig. 2A) but with different methylation contexts: MC10 and MC11 were mostly active in HMDs and MC12 and MC13 in PMDs (Fig. 2B). MC4-MC8, characterized by a more balanced combination of hyper- and hypomethylation events, were enriched in enhancer regions and LMRs. These components had the greatest transcriptional impact with, on average, 20% of their MRCpGs linked with the expression of a gene versus 8.6% among hypermethylation components MC1 and MC2 and 5.6% among hypomethylation components MC10-MC13 (Supporting Fig. S4). The enrichment patterns of MCs within chromatin states and CpG sequence contexts were reproducible across the three cohorts (Supporting Figs. S5 and S6), suggesting that MCs correspond to genuine biological processes preferentially active in specific epigenomic contexts.

**Gender- and age-related components**

To unravel the origin of each process, we analyzed the activity of components across tumor samples in two independent series (LICA-FR and TCGA-LIHC cohorts) for which extensive clinical and molecular data were available. We performed univariate (Fig. 3A; Supporting Table S2) and multivariate (Fig. 3B; Supporting Table S3) linear regression analyses to identify the main contributing features.

Several components were associated with general patient characteristics, such as gender and age. MC3 was perfectly associated with gender ($P = 5.0 \times 10^{-64}$; Fig. 4A). Of its MRCpGs, 96% were located within active TSS regions (Fig. 2A) of X chromosome genes (Fig. 4B). These CpGs were unmethylated in males and hemi-methylated in females (Fig. 4C). Thus, MC3 corresponds to the signature of X chromosome inactivation in females, illustrating the ability of MethICA to extract signatures of well-defined epigenetic processes, even when they involve a limited number of CpG sites.

Hypermethylation component MC1 involved CpG-dense islands enriched at bivalent promoters and enhancers (Fig. 2). These regions display a coexistence of active (monomethylation of histone H3 at lysine 4 [H3K4Me1] and/or trimethylation (H3K4Me3) and inactive (trimethylation of histone H3 on lysine 27 [H3K27Me3]) histone marks and have been shown to be prone to hypermethylation in cancer[36,37] and aging.[16,38] Consistently, the most contributing CpG sites of MC1 were progressively hypermethylated with age, both in HCC and NT liver (Fig. 4D). However, the gain of methylation at these CpG sites was considerably faster in tumors (+0.32% per year on average) than in NT liver (+0.024% per year), and this observation was validated in cancers from several other tissues (Supporting Fig. S7). Thus, MC1 reflects the progressive hypermethylation of bivalent chromatin domains that occurs naturally with age but is sharply increased in tumors.

Hypomethylation components MC12 and MC13 also increased linearly with age in both LICA-FR and TCGA-LIHC series (Fig. 4D). These components were particularly active in late-replicated PMDs, known to be prone to hypomethylation in cancer and aging,[17,39] but displayed different sequence context preferences (Fig. 5A). MC13 was more active in CpG-dense sequences, whereas MC12 was more active in sequences of low CpG density, particularly in the "solo-WCGW" context (CpG dyads surrounded by A/T and with no other CpG within 35 bps) prone to

methylation loss along cell divisions.[17] Thus, MC12 and MC13 suggest the existence of two distinct processes associated with the loss of methylation in late-replicated PMDs in liver cancer, operative in different sequence contexts.

**_CTNNB1_ mutation is a major modulator of DNA methylation in HCC**

Among HCC driver genes, *CTNNB1* showed the greatest impact on methylation, being significantly associated with four distinct components (Fig. 3).

First, age-related hypermethylation (MC1) and hypomethylation (MC12 and MC13) components were markedly increased in *CTNNB1*-mutated tumors. This observation is partly explained by the fact that *CTNNB1*-mutated cases tend to be older (mean age, 66 years versus 61 for non-mutated cases; $P = 0.017$). However, these associations remained significant independently from age, *CTNNB1* mutation being the most significant feature in multivariate analysis for MC12 and MC13 (Fig. 5B). As a result, *CTNNB1*-mutated tumors display a massive hypomethylation of PMDs as compared with other HCCs (Fig. 5A), with an average methylation in these regions of 45% versus 53% in other HCCs and 71% in NT liver.

In addition to age-related processes, MC8 was the most strongly associated with *CTNNB1* activating mutations ($P = 1.5 \times 10^{-21}$). In addition, different types of *CTNNB1* mutations activate ß-catenin with different strengths,[40] and the activity of MC8 followed this gradient of activation (Fig. 6A). The most contributing CpG sites, preferentially located in active enhancers, were strongly correlated to the expression of adjacent genes (Fig. 6B) enriched in Wnt/ß-catenin target genes (Fig. 6C). Motif analysis revealed an enrichment of transcription factor 7 (TCF7)-binding sites in the vicinity of MC8 MRCpGs (Fig. 6D). TCF7 is a member of the TCF/lymphoid enhancer–binding factor (LEF) family of transcription factors, the main downstream effectors of Wnt signaling pathway. Thus, MC8 reveals a coordinated hypomethylation of enhancers bound by TCF7 in *CTNNB1*-mutated HCC, associated with the up-regulation of Wnt/ß-catenin pathway genes. A representative example is shown in Fig. 6E,F where the hypomethylation of a cluster of CpG sites, overlapping intragenic H3K27Ac and TCF7 ChIP-seq peaks, accompanies the overexpression of *AXIN2* in *CTNNB1*-mutated tumors. These methylation changes likely play an active role in tumorigenesis by stabilizing the transcriptional changes induced by *CTNNB1* mutations.

**Hypomethylation of HMD is a signature of cyclin-activated HCC with intense replication stress**

Unexpectedly, hypomethylation components MC10 and MC11 were enriched in early replicated, CpG-dense regions within HMDs (Fig. 5A), which have been shown to be hypomethylation-resistant in previous studies. These components were strongly associated with the cyclin (CCN)-HCC subgroup of highly proliferative tumors, driven by cyclin A2/E1 (*CCNA2*/*E1*) activation (Fig. 5B). In these tumors, CCNA2 or CCNE1 activation by viral insertion, gene fusion, or enhancer hijacking leads to premature S phase entry and intense replication stress.[25] We hypothesize that, in CCN-HCC, cancer cells are pushed to replicate so fast that even early replicated HMDs become hypomethylated. As a result, this subgroup displays a striking hypomethylator phenotype involving all chromatin domains and sequence contexts (Fig. 5A). MC4, characterized by hypermethylation of partially methylated CpGs in early replicated regions, was also associated with CCN-HCC and may be another consequence of replication stress.

Altogether, our data indicate that several epigenetic processes are involved in the loss of DNA methylation in liver cancer cells. These processes are modulated by oncogenic alterations and lead to more or less extended hypomethylation patterns between molecular subgroups (Fig. 5A). CpG sites within PMDs are hypomethylated in all HCCs, but the methylation decrease is particularly strong in *CTNNB1*-mutated tumors. By contrast, CpG sites within HMD seem resistant to demethylation, except in CCN-HCC that are highly proliferative and subject to intense replication stress.

**Methylation signatures related to cellular differentiation**

MC2 and MC7 were encountered in tumors with a progenitor phenotype (Fig. 3), associated with diverse molecular features.

MC7 was significantly associated with *ARID1A* mutations in both the LICA-FR ($P = 0.0012$) and TCGA-LIHC ($P = 1.2 \times 10^{-5}$) cohorts (Fig. 7A). *ARID1A*, a member of the SWltch/Sucrose Non-Fermentable (SWI/SNF) chromatin remodeling complex, is recurrently mutated in HCC (13%, the fourth most frequently altered gene[5]). In mice, Arid1a interacts with several transcription factors that repress proliferation and maintain liver differentiation (CCAAT enhancer–binding protein

alpha [CEBPA], hepatocyte nuclear factor 4 alpha [Hnf4a], and forkhead box A2 [Foxa2]), and these pathways are down-regulated in *Arid1a*-deficient cells.[41] Consistently, MC7 was characterized by a hypermethylation of enhancers enriched in several transcription factor binding motifs (Fig. 7B), including CEBPA, FOXA2, and HNF4A, but also nuclear factor I A (NFIA) implicated in the differentiation of several cell types.[42-45] In addition, genes paired with hypermethylated CpGs related to MC7 were enriched in liver-specific genes (Gene Set Enrichment Analysis [GSEA]; $P < 2.2 \times 10^{-16}$; normalized enrichment score [NES] = 4.0). This methylation signature suggests that *ARID1A* deficiency impairs the DNA binding of several transcription factors and promotes the dedifferentiation of liver cancer cells.

MC2 was characterized by the hypermethylation of CGIs and CpG shores in chromatin regions repressed by Polycomb proteins (marked by the repressive H3K27Me3 histone mark only), in addition to bivalent TSS and enhancers (Fig. 2). Contrary to MC1, MC2 was not active in all HCCs but essentially in the G1 transcriptional subgroup ($P = 8.1 \times 10^{-8}$; Fig. 7C,D). This subgroup, enriched in young patients of African origin, is characterized by a progenitor phenotype with an overexpression of fetal liver genes.[2,46] G1 tumors display frequent alterations in *AXIN1*, ribosomal protein S6 kinase A3 (*RPS6KA3*), and BRCA1-associated protein 1 (*BAP1*) genes, all of which were significantly associated with MC2, but not independently from G1 subgroup (Fig. 7C). The specific hypermethylation signature of G1 tumors may thus reflect the epigenetic state of a progenitor cell of origin or the consequence of driver alterations enriched in this molecular subtype.

**DNA methylation-based classification of HCC reflects the combination of several components**

We next explored the relationships between MCs and methylation-based HCC classifications. Consensus clustering revealed relatively stable partitions of both LICA-FR and TCGA-LIHC cohorts into eight tumor clusters. These clusters, highly consistent in the two series, defined seven common subgroups ($M_1^{HCC}$ to $M_7^{HCC}$), with $M_6^{HCC}$subdivided in two ($M_{6a}^{HCC}$and $M_{6b}^{HCC}$) in the LICA-FR cohort (Fig. 8A,B**)** and $M_1^{HCC}$subdivided in two ($M_{1a}^{HCC}$and $M_{1b}^{HCC}$) in TCGA-LIHC cohort (Fig. 8C,D**)**.

DNA methylation-based HCC subgroups were significantly associated with age, geographical origin, transcriptional subgroups, and driver alterations. Cluster $M_1^{HCC}$displayed the least methylation changes with respect to NT liver tissues. Cluster $M_2^{HCC}$, characterized by a high activity of MC2, comprised tumors of the G1 transcriptional subgroup, of Asian or African origin, with high frequencies of *BAP1*, *AXIN1,* and *RPS6KA3* mutations. Clusters $M_3^{HCC}$to $M_7^{HCC}$ comprised older patients, with high activity of age-related hypermethylation (MC1) and hypomethylation (MC12 and MC13) components. Cluster $M_4^{HCC}$was enriched in *TP53*-mutated tumors of the G3 transcriptional subgroup. Clusters $M_5^{HCC}$and $M_6^{HCC}$, with a high activity of hypomethylation components MC12 and MC13, were enriched in well-differentiated *CTNNB1*-mutated tumors of the G5 and G6 transcriptional subgroups. Cluster $M_7^{HCC}$, with a high activity of MC4, MC10, and MC11, was strongly associated with *CCNA2*/*E1* activation and displayed the most striking hypomethylator phenotype.

Thus, MCs capture variations that are either widespread in the data set (e.g., MC1), restricted to a precise cluster (e.g., MC2), or dispersed across tumors belonging to distinct clusters (e.g., MC3), as clearly illustrated in t-distributed stochastic neighbor embedding (t-SNE) plots (Supporting Fig. S8). For example, *ARID1A*-mutated tumors are not enriched in a particular cluster, but MethICA was able to extract their common signature within MC7. Thus, ICA reveals individual sources of variation that are intermingled in cancer methylomes and highlights subtle methylation signatures beyond the main methylation clusters that reflect the activity of a few dominant processes.

**Methylation components reveal pre-neoplastic changes in cirrhotic liver**

We next examined the methylation profiles of 35 NT liver tissues of the LICA-FR cohort, comprising 5 non-fibrotic (METAVIR F0-F1), 14 chronic hepatitis (F2-F3), and 16 cirrhotic (F4) livers. Hierarchical clustering revealed four homogeneous subgroups strongly associated with fibrosis stage ($P = 1.6 \times 10^{-6}$). The two main groups corresponded to cirrhotic and non-cirrhotic livers. Non-cirrhotic livers were further divided in three subgroups distinguishing F0-F1 from F2-F3 samples (Supporting Fig. S9A). To identify methylation changes accompanying cirrhosis, we compared the intensity of our 13 MCs between different fibrosis stages. MC6 and MC7 were significantly more active in cirrhotic liver (Supporting Fig. S9B), which was validated in TCGA-LIHC cohort (Supporting Fig. S9C).

MC6 increased progressively in F2-F3 and F4 livers. This component was correlated with the level of immune infiltration estimated from gene expression data (Fig. 3) and with the immune-mediated cancer field (ICF) signature, a signature of deregulated immune response associated with risk of HCC development in patients with cirrhosis.[47] (Supporting Fig. S9D,E) DNA methylation changes related to MC6 involve two anti-correlated sets of CpGs. On one side, CpG sites located within hepato-specific enhancers, enriched in hepatocyte nuclear factor–binding motifs, are hypermethylated in samples with a stronger immune infiltrate (Supporting Fig. S9F). On the other side, CpG sites located within immune cell–specific enhancers, enriched in JUN/FOS-binding motifs, are demethylated in more infiltrated samples (Supporting Fig. S9G). Thus, MC6 is an epigenetic signature of the immune response that occurs in fibrotic / cirrhotic liver and promotes carcinogenesis.[47]

In addition, *ARID1A*-associated MC7 was activated in cirrhotic liver although to a lesser extent than in HCC (Supporting Fig. S9B,C). Interestingly, ultra-deep sequencing revealed *ARID1A* mutations in cirrhotic nodules, and *Arid1a* depletion was shown to promote clonal expansion and regeneration in chronic liver disease.[48] In agreement with these findings, our results suggest that the coordinated hypermethylation of enhancers implicated in liver differentiation may drive hepatocytes to a more proliferative state, favoring the clonal expansion of cirrhotic nodules.

## DISCUSSION

Independent component analysis of the largest HCC series analyzed so far revealed 13 different methylation components operative with different strengths across HCC and NT liver tissues. This represents a much greater diversity of signatures than identified in previous methylation studies. Early reports described global changes in HCC as compared with NT liver tissue, including hypermethylation of CpG islands enriched in Polycomb-repressive complex 2 (PRC2) target genes, and a widespread hypomethylation in open sea regions.[7-10] Previous unsupervised classifications revealed between three and seven HCC subgroups.[4,11,12] In particular, TCGA described four tumor subgroups based on hypermethylated probes and three subgroups (largely overlapping) based on hypomethylated probes. These subgroups, strongly associated with our consensus clusters (Fig. 8D), display varying levels of hyper- and hypomethylation with respect to NT samples. However, ICA allowed us to define more DNA methylation signatures, related to precise biological processes, and to disentangle age- and gender-related processes from changes associated with specific tumor subgroups and driver alterations.

MC1 captured the hypermethylation of CGIs located in bivalent chromatin domains, known to occur naturally with aging. This component increases with age in both NT liver and HCC, but the slope of this increase is much sharper in tumors. By contrast, MC2 is associated with the G1 transcriptomic subgroup and defines a strongly hypermethylated HCC entity. Further studies are required to determine whether this methylation signature reflects a different cell of origin for this subgroup or is acquired during tumorigenesis.

Global loss of DNA methylation has been described in most cancer types,[49] including HCC,[50] but the mechanisms by which this hypomethylation occurs remain incompletely understood. We show here that four independent processes are involved in this process. MC12 and MC13 are preferentially active in late-replicated PMD, known to be prone to hypomethylation along cell divisions.[17] Hypomethylation of MC12 and MC13 MRCpGs was correlated with age in HCC but, surprisingly, not in NT liver (Fig. 4D). We hypothesize that the loss of methylation in PMD occurs stochastically at different CpG sites in each normal cell and is thus barely detectable in NT tissue. By contrast, clonal expansion amplifies the hypomethylation pattern of the cell of origin that becomes visible in the tumor, just like somatic mutations. In addition, these components were significantly more active in *CTNNB1*-mutated HCC. This might reflect differences in terms of cell

of origin or tumor growth dynamics. *CTNNB1*-mutated HCCs are usually well differentiated and less proliferative than other HCC subgroups. Thus, these tumors may have a longer development, leaving more time for methylation changes to occur.

The two other hypomethylation components (MC10 and MC11) affect HMDs and are particularly active in the CCN-HCC subgroup, driven by CCNA2/E1 activation. We previously found that, in CCN-HCC, replication stress induces a massive accumulation of structural rearrangements, preferentially located in early replicated regions.[25] Here we show that replication stress also impacts the methylome of these tumors, presumably because cells enter S phase prematurely, before the newly synthesized strand has been properly methylated.

Finally, we identified five MCs (MC4-MC8) of coordinated enhancer methylation reprogramming. These components have been missed in previous studies, possibly because they involve fewer CpG sites and can be dispersed across the main DNA methylation subgroups. However, they have the strongest transcriptional impact and constitute valuable markers of transcriptional network activity. MC8 reflects the precise level of activation of the Wnt/ß-catenin pathway induced by diverse *CTNNB1* mutations.[40] To our knowledge, MC7 is the first methylation signature associated with *ARID1A* mutations. Motif analysis shed light on the transcription factors impacted by *ARID1A* deficiency, including several key regulators of liver differentiation.

Although our approach was not primarily designed for biomarker identification, several components may have therapeutic implications. First, a high activity of MC10 and MC11 indicates highly proliferative tumors with replication stress, who might benefit from ATR pathway inhibitors.[25] Second, MC6 provides a precise estimation of the level of immune infiltration in the tumor. A high activity of MC6 may thus highlight good candidates for immunotherapy. Future studies are required to determine if MCs constitute valuable biomarkers for treatment response.

Overall, ICA appears as a powerful tool for the analysis of DNA methylation signatures. All the utilities we developed for extracting and interpreting MCs are included in the MethICA package, applicable to both microarray and bisulfite sequencing data.

**Acknowledgment:** We thank the principal investigators of the liver cancer TCGA (Lewis Roberts, David Wheeler) and HEPTROMIC (Augusto Villanueva, Josep Llovet) projects for providing the high-quality data sets used in this study; all the clinicians, surgeons, pathologists, hepatologists, and oncologists who contributed to the tissue collection and clinical annotations; and the Réseau national CRB Foie (BB-0033-0085) and the tumor banks of CHU Bordeaux (BB-0033-00036), Jean Verdier Hospital (APHP), and CHU Henri Mondor (APHP) for contributing to the tissue collection.

**REFERENCES**

1) Llovet JM, Zucman-Rossi J, Pikarsky E, Sangro B, Schwartz M, Sherman M, et al. Hepatocellular carcinoma. Nat Rev Dis Primers 2016;2:16018. doi: 10.1038/nrdp.2016.18.

2) **Boyault S**, **Rickman DS**, **de Reyniès A**, Balabaud C, Rebouissou S, Jeannot E, et al. Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. HEPATOLOGY 2007;45:42-52.

3) Hoshida Y, Nijman SMB, Kobayashi M, Chan JA, Brunet J-P, Chiang DY, et al. Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. Cancer Res 2009;69:7385-7392.

4) Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell 2017;169:1327-1341.e23.

5) **Schulze K**, **Imbeaud S**, **Letouzé E**, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet 2015;47:505-511.

6) **Fujimoto A**, **Furuta M**, **Totoki Y**, **Tsunoda T**, **Kato M**, Shiraishi Y, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet 2016;48:500-509.

7) Zhang C, Li Z, Cheng Y, Jia F, Li R, Wu M, et al. CpG island methylator phenotype association with elevated serum alpha-fetoprotein level in hepatocellular carcinoma. Clin Cancer Res 2007;13:944-952.

8) Stefanska B, Huang J, Bhattacharyya B, Suderman M, Hallett M, Han Z-G, et al. Definition of the landscape of promoter DNA hypomethylation in liver cancer. Cancer Res 2011;71:5891-5903.

9) Neumann O, Kesselmeier M, Geffers R, Pellegrino R, Radlwimmer B, Hoffmann K, et al. Methylome analysis and integrative profiling of human HCCs identify novel protumorigenic factors. HEPATOLOGY 2012;56:1817-1827.

10) Shen J, Wang S, Zhang Y-J, Kappil M, Wu H-C, Kibriya MG, et al. Genome-wide DNA methylation profiles in hepatocellular carcinoma. HEPATOLOGY 2012;55:1799-1808.

11) Mah W-C, Thurnherr T, Chow PKH, Chung AYF, Ooi LLPJ, Toh HC, et al. Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. PLoS One 2014;9:e104158. doi: 10.1371/journal.pone.0104158.

12) Cheng J, Wei D, Ji Y, Chen L, Yang L, Li G, et al. Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. Genome

Med 2018;10:42. doi: 10.1186/s13073-018-0548-z.
13) Kisiel JB, Dukek BA, Kanipakam RVSR, Ghoz HM, Yab TC, Berger CK, et al. Hepatocellular carcinoma detection by plasma methylated DNA: discovery, phase I pilot, and phase II clinical validation. HEPATOLOGY 2019;69:1180-1192.
14) Villanueva A, Portela A, Sayols S, Battiston C, Hoshida Y, Méndez-González J, et al. DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma. HEPATOLOGY 2015;61:1945-1956.
15) Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell 2018;173:291-304.e6.
16) Rakyan VK, Down TA, Maslau S, Andrew T, Yang T-P, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res 2010;20:434-439.
17) Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet 2018;50:591-602.
18) Vandiver AR, Irizarry RA, Hansen KD, Garza LA, Runarsson A, Li X, et al. Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. Genome Biol 2015;16:80. doi: 10.1186/s13059-015-0644-y.
19) **Letouzé E**, **Martinelli C**, Loriot C, Burnichon N, Abermil N, Ottolenghi C, et al. SDH mutations establish a hypermethylator phenotype in paraganglioma. Cancer Cell 2013;23:739-752.
20) Yao L, Shen H, Laird PW, Farnham PJ, Berman BP. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. Genome Biol 2015;16:105. doi: 10.1186/s13059-015-0668-3.
21) Chakravarthy A, Furness A, Joshi K, Ghorani E, Ford K, Ward MJ, et al. Pan-cancer deconvolution of tumour composition using DNA methylation. Nat Commun 2018;9:3220. doi: 10.1038/s41467-018-05570-1.
22) Zinovyev A, Kairov U, Karpenyuk T, Ramanculov E. Blind source separation methods for deconvolution of complex signals in cancer biology. Biochem Biophys Res Commun 2013;430:1182-1187.
23) Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415-421.
24) Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C. Elucidating the altered

transcriptional programs in breast cancer using independent component analysis. PLoS Comput Biol 2007;3:e161. doi: 10.1371/journal.pcbi.0030161.
25) Bayard Q, Meunier L, Peneau C, Renault V, Shinde J, Nault J-C, et al. Cyclin A2/E1 activation defines a hepatocellular carcinoma subclass with a rearrangement signature of replication stress. Nat Commun 2018;9:5235. doi: 10.1038/s41467-018-07552-9.
26) Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet 2012;44:694-698.
27) **Letouzé E**, **Shinde J**, Renault V, Couchy G, Blanc J-F, Tubacher E, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nat Commun 2017;8:1315. doi: 10.1038/s41467-017-01358-x.
28) **Hirsch TZ**, **Negulescu A**, Gupta B, Caruso S, Noblet B, Couchy G, et al. BAP1 mutations define a homogeneous subgroup of hepatocellular carcinoma with fibrolamellar-like features and activated PKA. J Hepatol 2020;72:924-936.
29) Nault J-C, Martin Y, Caruso S, Hirsch TZ, Bayard Q, Calderaro J, et al. Clinical impact of genomic diversity from early to advanced hepatocellular carcinoma. HEPATOLOGY 2020;71:164-182.
30) Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550. doi: 10.1186/s13059-014-0550-8.
31) Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 1999;10:626-634.
32) Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al.; Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317-330.
33) Salhab A, Nordström K, Gasparoni G, Kattler K, Ebert P, Ramirez F, et al. A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. Genome Biol 2018;19:150. doi: 10.1186/s13059-018-1510-5.
34) Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009;462:315-322.
35) Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 2011;480:490-495.
36) Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, et al. Polycomb-

mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet 2007;39:232-236.
37) Easwaran H, Johnstone SE, Van Neste L, Ohm J, Mosbruger T, Wang Q, et al. A DNA hypermethylation module for the stem/progenitor cell signature of cancer. Genome Res 2012;22:837-849.
38) Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res 2010;20:440-446.
39) Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet 2011;44:40-46.
40) **Rebouissou S**, **Franconi A**, Calderaro J, Letouzé E, Imbeaud S, Pilati C, et al. Genotype-phenotype correlation of CTNNB1 mutations reveals different ß-catenin activity associated with liver tumor progression. HEPATOLOGY 2016;64:2047-2061.
41) Sun X, Chuang J-C, Kanchwala M, Wu L, Celen C, Li L, et al. Suppression of the SWI/SNF component Arid1a promotes mammalian regeneration. Cell Stem Cell 2016;18:456-466.
42) Piper M, Barry G, Hawkins J, Mason S, Lindwall C, Little E, et al. NFIA controls telencephalic progenitor cell differentiation through repression of the Notch effector Hes1. J Neurosci 2010;30:9127-9139.
43) Hiraike Y, Waki H, Yu J, Nakamura M, Miyake K, Nagano G, et al. NFIA co-localizes with PPARγ and transcriptionally controls the brown fat gene program. Nat Cell Biol 2017;19:1081-1092.
44) Singh PNP, Yadav US, Azad K, Goswami P, Kinare V, Bandyopadhyay A. NFIA and GATA3 are crucial regulators of embryonic articular cartilage differentiation. Development 2018;145. doi: 10.1242/dev.156554.
45) Chen K-S, Bridges CR, Lynton Z, Lim JWC, Stringer BW, Rajagopal R, et al. Transcription factors NFIA and NFIB induce cellular differentiation in high-grade astrocytoma. J Neurooncol 2020;146:41-53.
46) Calderaro J, Couchy G, Imbeaud S, Amaddeo G, Letouzé E, Blanc J-F, et al. Histological subtypes of hepatocellular carcinoma are related to gene mutations and molecular tumour classification. J Hepatol 2017;67:727-738.
47) Moeini A, Torrecilla S, Tovar V, Montironi C, Andreu-Oller C, Peix J, et al. An immune gene

expression signature associated with development of human hepatocellular carcinoma identifies mice that respond to chemopreventive agents. Gastroenterology 2019;157:1383-1397.e11.
48) Zhu M, Lu T, Jia Y, Luo X, Gopal P, Li L, et al. Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease. Cell 2019;177:608-621.e12.
49) Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics 2009;1:239-259.
50) Lin CH, Hsieh SY, Sheen IS, Lee WC, Chen TC, Shyu WC, et al. Genome-wide hypomethylation in hepatocellular carcinogenesis. Cancer Res 2001;61:4238-4243.
Author names in bold designate shared co-first authorship.

**Figure Legends**

**FIG. 1.** Identification of 13 stable methylation components in liver cancers. (A) MethICA workflow. Three independent HCC data sets analyzed with the same methylation array are included in this study. ICA is used to decompose the methylation bval matrix X of dimension n × 200,000 (n samples and 200,000 most variant CpGs) as the product of a matrix A (size n × 20 MCs) giving the contributions of the samples to each MC (or activities of the MC in the n samples) and a matrix S (size 20 × 200,000) giving the projections of the CpGs onto each MC. CpGs having the largest projection onto a component (providing the greatest contribution) are the most strongly influenced by the epigenetic process underlying the MC. To unravel the biological meaning of each component, we analyzed the clinical and molecular annotations of the most contributing samples and the (epi)genomic features of the most contributing CpGs. (B) Major DNA methylation changes associated with each component in the LICA-FR cohort. For each MC, the most contributing CpG sites were selected and their methylation compared between NT samples and the 5% of tumors with the strongest methylation changes. Abbreviations: bval, beta value; CGI, CpG island; HCC, hepatocellular carcinoma; ICA, independent component analysis; LICA-FR, Liver Cancer (France); MC, methylation component; MethICA, methylation signature analysis with independent component analysis; T, tumor; NT, non-tumor; RNA-seq, RNA sequencing; TCGA-LIHC, The Cancer Genome Atlas Liver Hepatocellular Carcinoma; WES, whole-exome sequencing; WGS, whole-genome sequencing.

**FIG. 2.** Methylation components preferentially affect specific methylation domains and chromatin states. (A) Epigenomic features associated with each component. The most contributing CpG sites of each component were extracted. The first two lines indicate the proportion of these CpG sites falling within each CGI- and gene-based feature. Enrichment scores (ESs) in active/inactive chromatin and across the 18 chromatin states defined by the Roadmap consortium in normal liver are represented below, with a color code for each chromatin state as displayed in Supporting Fig. 2B. (B) Methylation contexts associated with each component. Each CpG of the array was classified into one of 48 categories based on the methylation domain in normal liver (HMD, PMD, LMR, UMR), local CpG density (number of flanking CpGs within 35 bps on each side of the dyad), and sequence context (SCGS, SCGW, or WCGW, with S denoting C or G and W denoting A or T). The distribution of CpG methylation in each category in 35 NT liver tissues is represented as a violin plot. ESs of the most contributing CpG sites of each component across the

48 categories are represented as bar plots. Displayed ESs were computed in the LICA-FR cohort except for MC10 (TCGA). ESs were highly reproducible in the three cohorts (Supporting Figs. S5 and S6). Abbreviations: bps, base pairs; ES, enrichment score; HMD, highly methylated domain; LMR, lowly methylated region; PMD, partially methylated domain; TSS, transcription start site; UMR, unmethylated region; ZNF, zinc finger.

**FIG. 3.** Clinical and molecular features associated with each component. (A) Results of the univariate analysis in the LICA-FR (top) and TCGA-LIHC (bottom) series are shown. All clinical and molecular features significantly associated (*P* value <0.05) with at least one component are shown. The size of each circle indicates the *P* value of the association as represented in the legend below, and its color indicates the type of feature (clinical, molecular, or phenotypic). Associations that are significant in both series are squared. MC10 was not identified in LICA-FR. See also Supporting Table S2. * Molecular subgroups correspond to the G1-G6 transcriptomic groups defined by Boyault *et al.* ** Gene expression signatures were previously described by Nault *et al.* (differentiation, proliferation), Caruso *et al.* (liver progenitor, stem cell, EMT/metastasis) and Becht *et al.* (immune infiltrate). See Supplementary Methods for more details. (B) Results of the multivariate analysis in LICA-FR (top) and TCGA-LIHC (bottom) series. Only features significant in univariate analyses in both series were included; others are colored in light gray. See also Supporting Table S3.
Abbreviations: ARID1A/2, AT-rich interactive domain-containing protein 1A/2; BAP1, BRCA1-associated protein 1; CCNA2/E1, cyclin A2/E1; CTNNB1, catenin beta 1; EMT, epithelial-to-mesenchymal transition; KEAP1, kelch-like ECH-associated protein 1; RB1, RB transcriptional corepressor 1; RPS6KA3, ribosomal protein S6 kinase A3; TERT, telomerase reverse transcriptase.

**FIG. 4.** Gender- and age-related MCs. This figure describes MCs associated with gender (MC3) and age (MC1, MC12 and MC13). (A) Sample contribution to MC3 allows to perfectly split males and females. (B) Enrichment of MC3 most contributing CpG sites in sexual chromosomes. (C) Average methylation of MC3 most contributing CpG sites in males (*x* axis) and females (*y* axis). (D) Correlation of hypermethylation component MC1 and hypomethylation components MC12 and MC13 with age in cancerous (HCC) and NT liver tissue. Abbreviations: ChrX, chromosome

X; ChrY, chromosome Y; F, female; M, male; MRCpGs, most representative CpGs; NS, not significant.

**FIG. 5.** Four components shape the hypomethylation landscapes of HCC subgroups. (A) Top: methylation domains and sequence contexts enriched in the four hypomethylation components (MC10-MC13). Bottom: distribution of methylation levels per CpG sequence context in non-tumor liver and different HCC molecular subgroups. (B) Heatmaps showing the methylation of the MRCpGs of MC10-MC13 across tumor and non-tumor samples, ordered by component intensity, with associated clinico-molecular features. Cyclin status refers to the presence or of genomic alterations (structural rearrangements, gene fusions or viral insertions) activating *CCNA2* or *CCNE1*, as described in Bayard *et al*. Proliferation and differentiation scores refer to the mean expression of markers of liver differentiation and cell proliferation previously established by Nault *et al.* (see Supplementary Methods for more details). Non-tumor liver tissues are represented in the heatmaps but were not used in association tests. Abbreviations: NT, non-tumor; FLC, fibrolamellar carcinoma; M, mutated; NM, non-mutated; multiv., multivariate; univ., univariate; WT, wild-type.

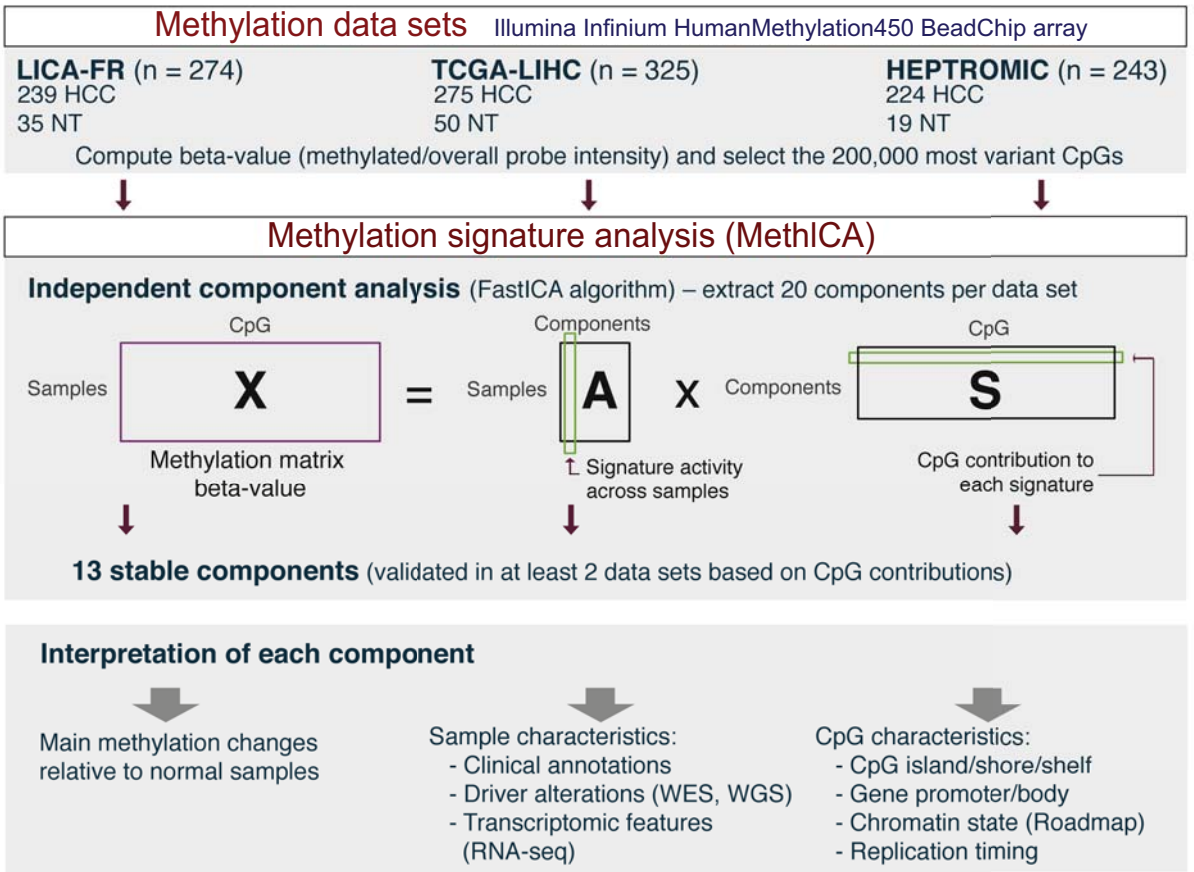
**FIG. 6.** Coordinated hypomethylation of TCF7-bound enhancers in *CTNNB1*-mutated HCC. This figure describes MC8, the most strongly associated with *CTNNB1* mutations. (A) MC8 is strongly correlated to somatic alterations that activate ß-catenin with different strengths. *CTNNB1*-mutated tumors were stratified according to the mutated amino acid hotspot (K335, N387, S45, T41). Mutations affecting the ß-Trcp binding site (D32-S37) were grouped. Large deletions correspond to activating in-frame deletions of exon 3. *CTNNB1* mutations were ordered by the level of ß-catenin activation established by Rebouissou *et al.* (B) Hypomethylation of its most contributing CpG sites is associated with the up-regulation of target genes, enriched in (C) Wnt/ß-catenin targets. (D) Motif analysis reveals an enrichment of TCF7 targets, exemplified by (E) cg11122009 associated with *AXIN2* regulation. (F) Epigenomic features and transcriptomic regulation at *AXIN2* locus. Tracks display, from top to bottom, the chromatin states inferred by Roadmap consortium in normal liver, methylation beta-values measured by Illumina Infinium HumanMethylation450 array and RNA-seq coverage in non-tumor liver and *CTNNB1*-mutated HCC samples, H3K27 acetylation and TCF7 binding from ENCODE ChIP-seq data. In *CTNNB1*-mutated HCC, *AXIN2* overexpression is accompanied by the demethylation of a group of CpGs

comprising cg11122009 and overlapping H3K27Ac and TCF7 binding peaks. Abbreviations: NES, Normalized enrichment score; FDR, False Discovery Rate; H3K27ac, acetylation of histone H3 on lysine 27; TCF7, transcription factor 7; TxWk, Weak transcription; Enha2, Active enhancer 2; Tx, Strong transcription; EnhWk, Weak enhancer; TssFlnkD, Flanking TSS downstream; bval, beta-value; NT, non-tumor; HCC, hepatocellular carcinoma.

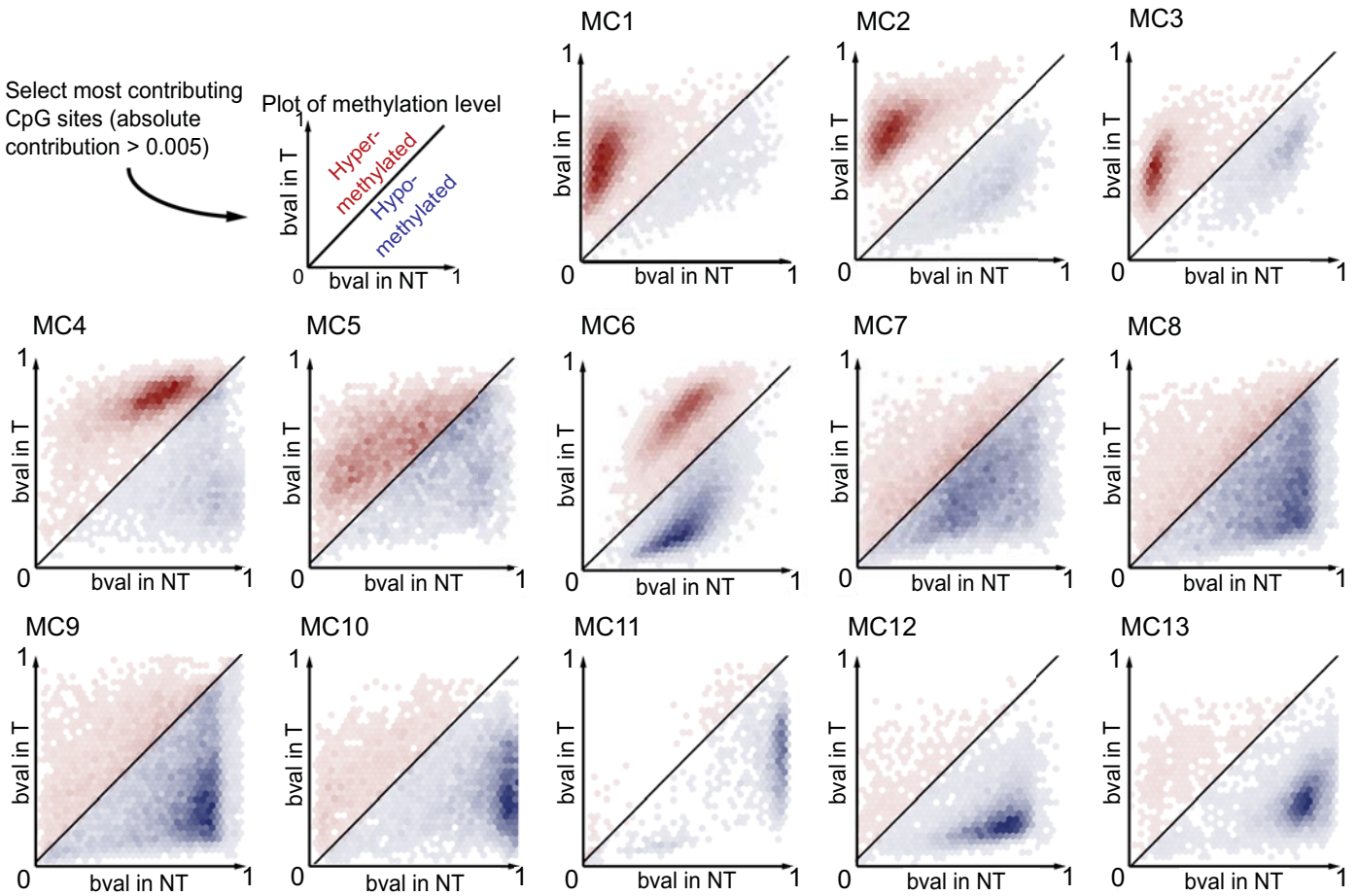
**FIG. 7.** MCs related to cellular dedifferentiation. This figure describes components related to *ARID1A* inactivation (MC7) and the G1 molecular subgroup (MC2). (A,B) MC7, correlated with *ARID1A* inactivation, involves the hypermethylation of binding sites for several transcription factors related to liver differentiation. (C) Hypermethylation component MC2 is particularly active in the G1 progenitor subgroup, associated with *AXIN1*, *RPS6KA3,* and *BAP1* mutations. (D) Box plot showing the distribution of MC2 activity in the G1-G6 molecular subgroups defined by Boyault *et al.* Abbreviations: CEBPA, CCAAT enhancer–binding protein alpha; FOXA2, forkhead box A2; geo., geographical; HNF4A, hepatocyte nuclear factor 4 alpha; NF1A, nuclear factor I A; TFBS, transcription factor–binding site; transcr., transcriptomic.

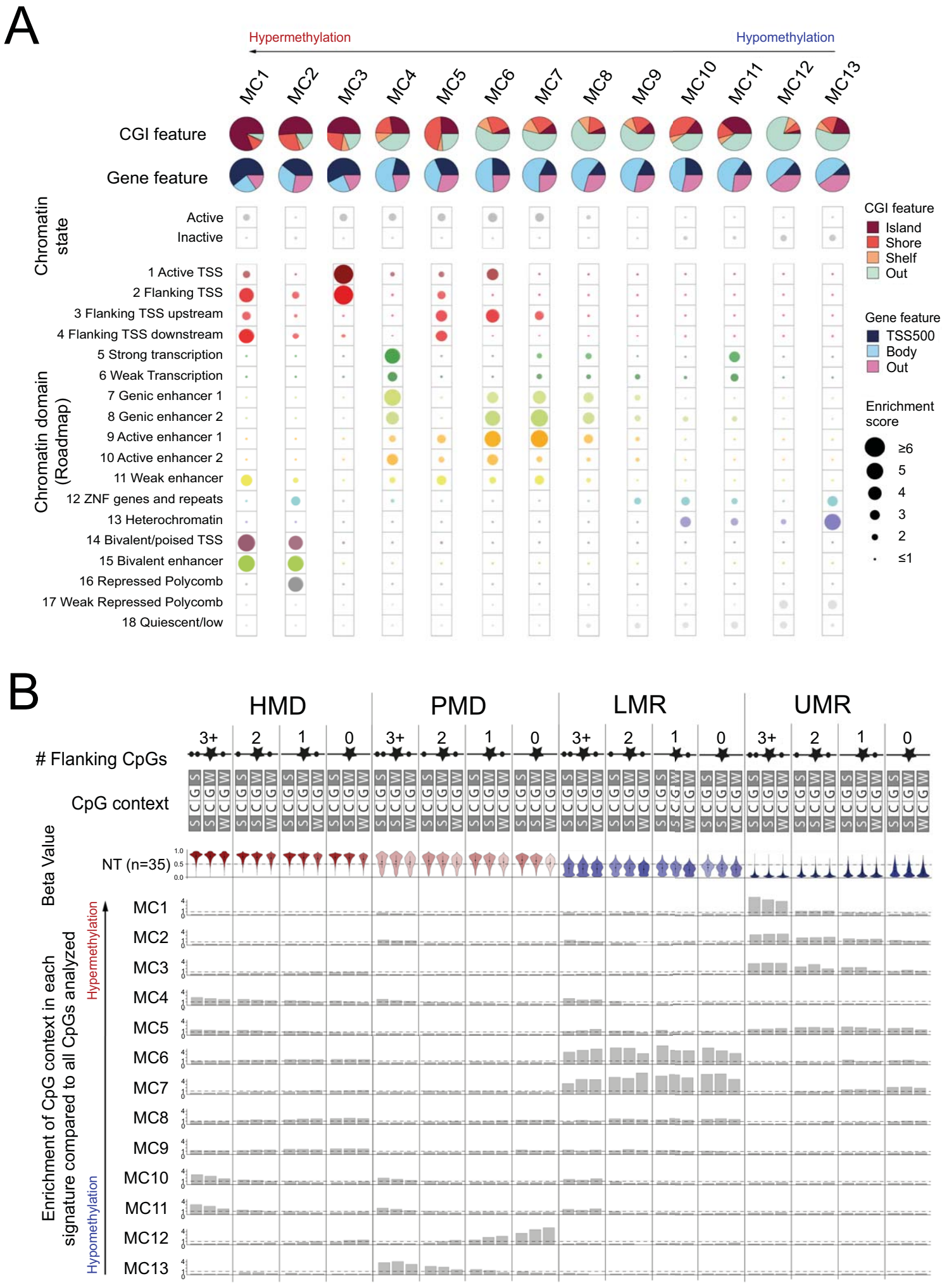
**FIG. 8.** DNA methylation-based classification of HCC. Tumors from the (A,B) LICA-FR and (C,D) TCGA-LIHC cohorts were classified according to the methylation levels of their most variant CpGs. (A) and (C) display the consensus matrices representing the similarity between tumors. Consensus index values range from 0 (highly dissimilar profiles, white) to 1 (highly similar profiles, dark blue). Samples are ordered on the *x* and *y* axes by the consensus clustering, which is depicted above the heatmap. (B) and (D) display heatmap representations of DNA methylation profiles. The degree of DNA methylation (bval) for each probe (row) in each sample (column) is represented with a color scale (dark blue, nonmethylated; yellow, methylated). Tumors are ordered by methylation cluster. Probes are arranged by similarity and in the same order in (B) and (D). Clinical and molecular annotations are indicated above the heatmap with *P* values showing their association with the clusters. The activity of each MC is represented below the heatmap with a color scale (blue, low activity; red, high activity). Abbreviations: hyper., hypermethylated; hypo., hypomethylated.

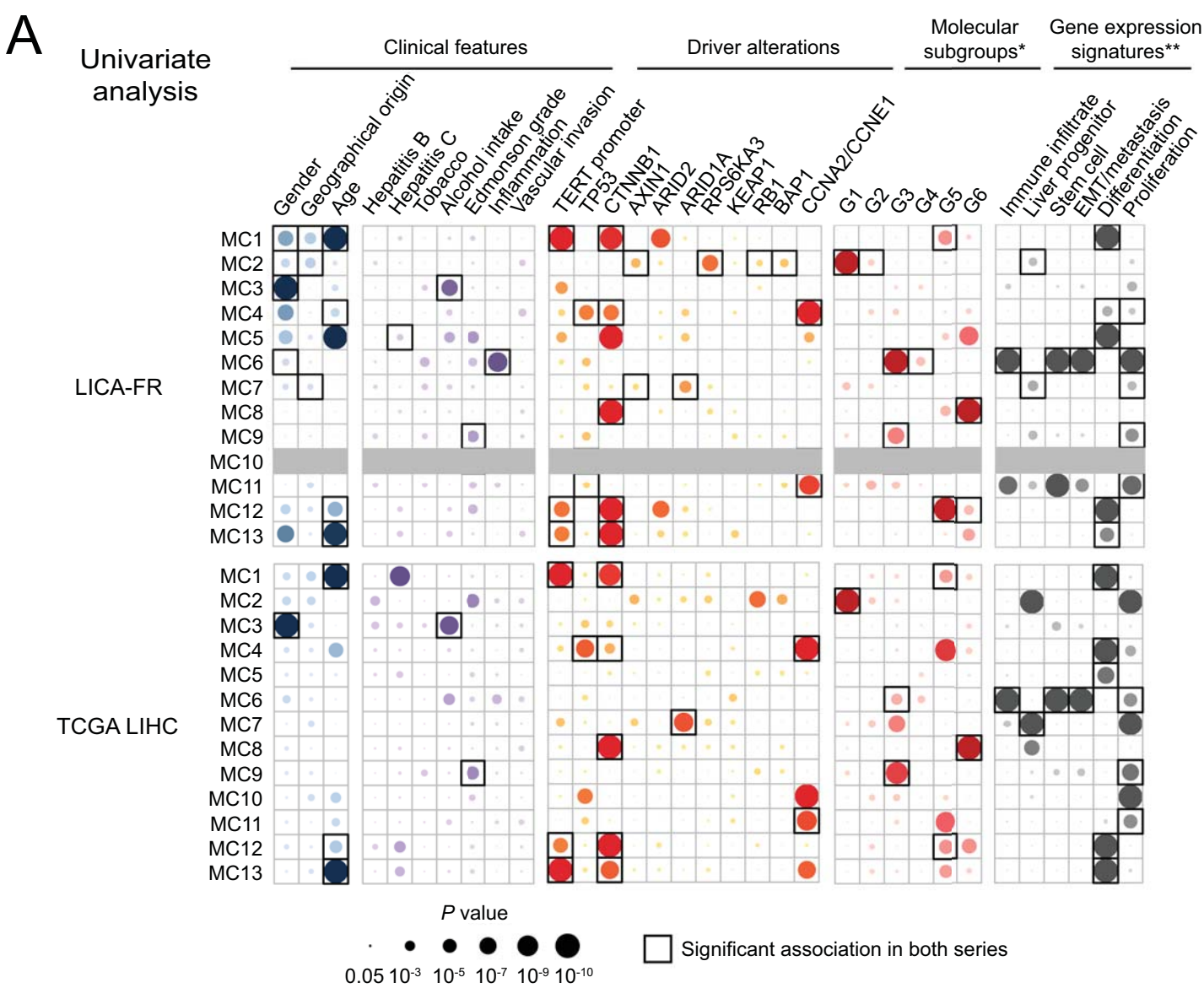
A

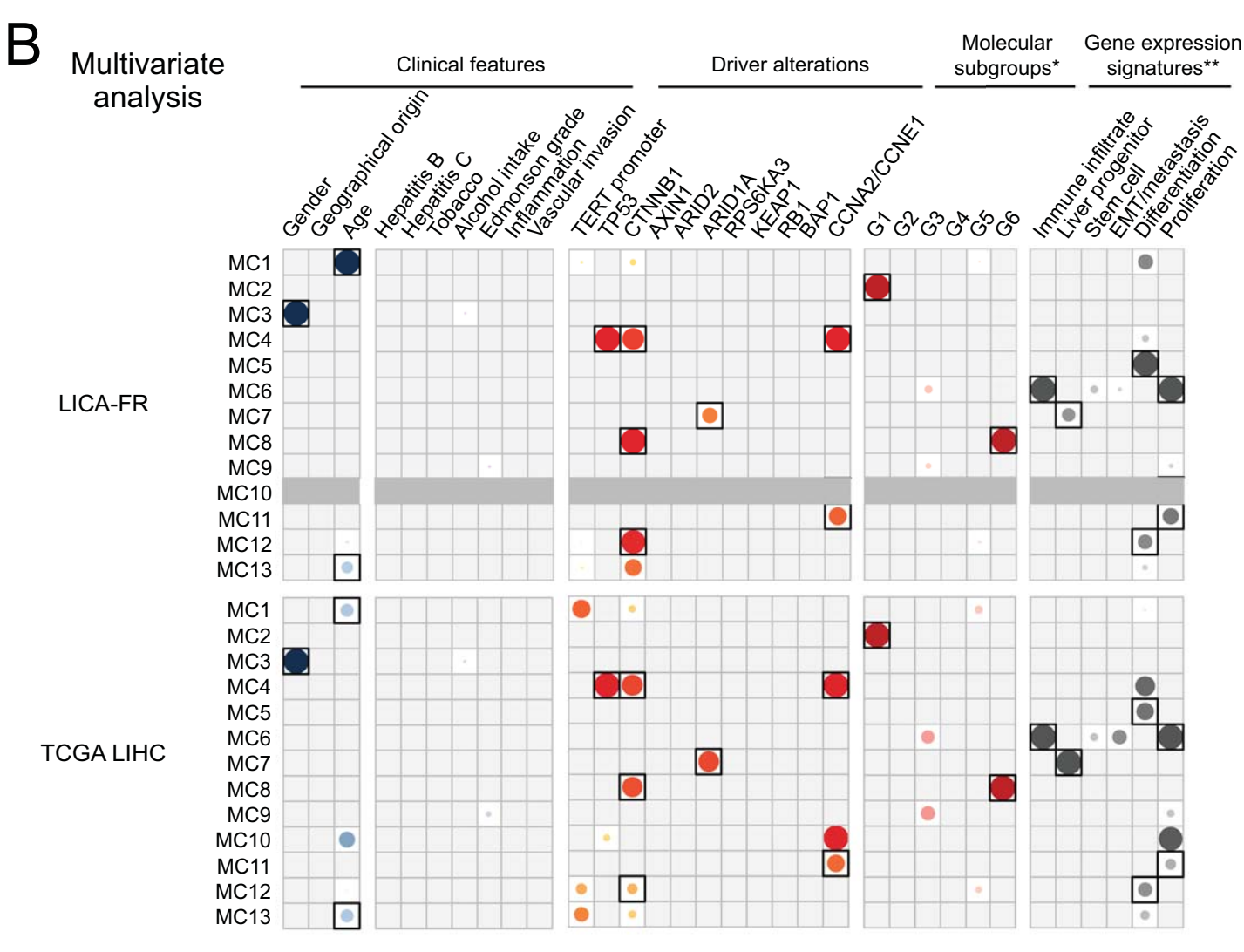

Methylation data sets
Illumina Infinium HumanMethylation450 BeadChip array
LICA-FR (n = 274)
239 HCC
35 NT
TCGA-LIHC (n = 325)
275 HCC
50 NT
HEPTROMIC (n = 243)
224 HCC
19 NT
Compute beta-value (methylated/overall probe intensity) and select the 200,000 most variant CpGs
Methylation signature analysis (MethICA)
Independent component analysis (FastICA algorithm) – extract 20 components per data set
CpG
Samples
X
Methylation matrix
beta-value
=
Components
Samples
A
Signature activity
across samples
x
Components
CpG
S
CpG contribution to
each signature
13 stable components (validated in at least 2 data sets based on CpG contributions)
Interpretation of each component
Main methylation changes
relative to normal samples
Sample characteristics:
- Clinical annotations
- Driver alterations (WES, WGS)
- Transcriptomic features
(RNA-seq)
CpG characteristics:
- CpG island/shore/shelf
- Gene promoter/body
- Chromatin state (Roadmap)
- Replication timing

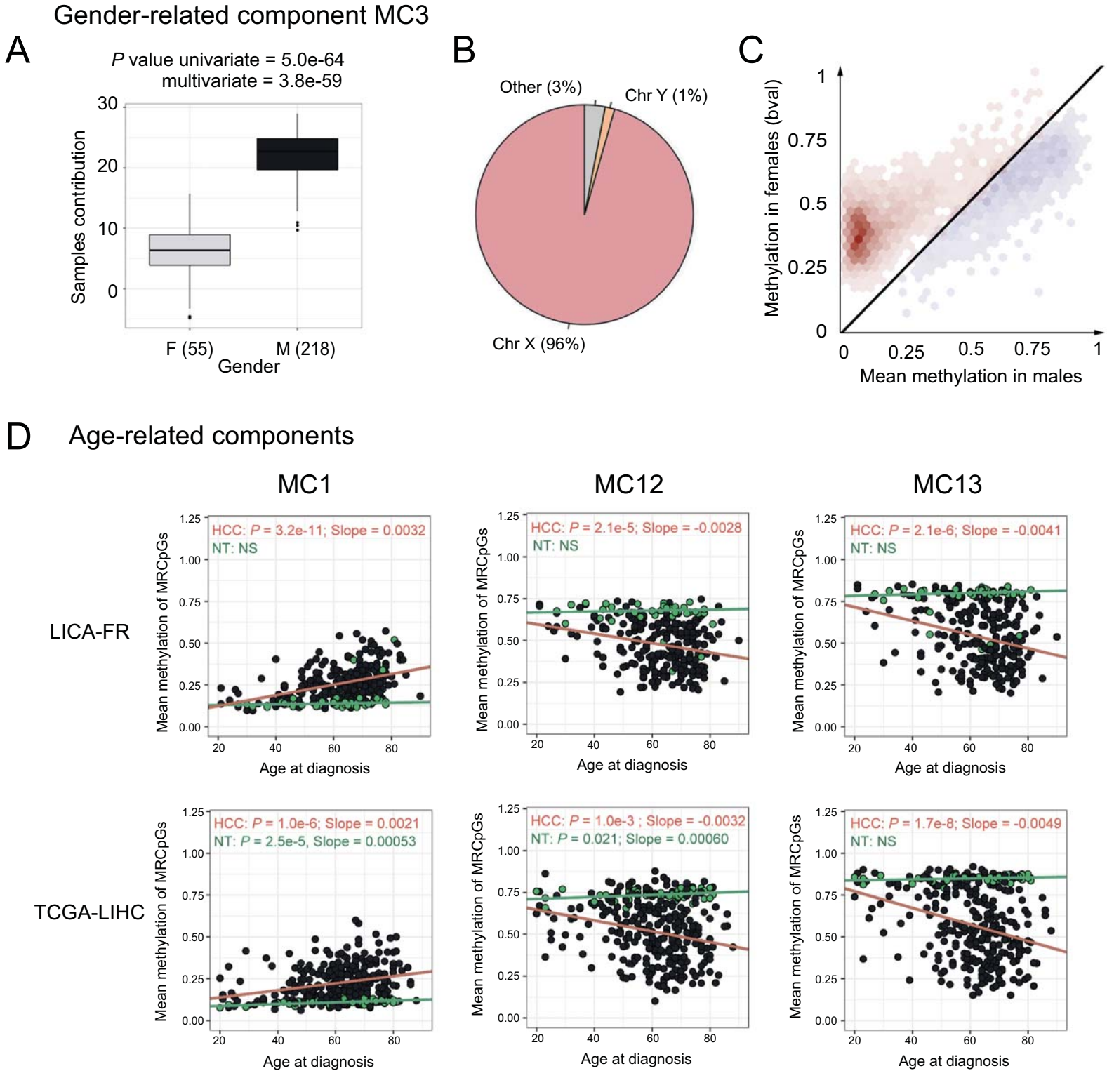
B

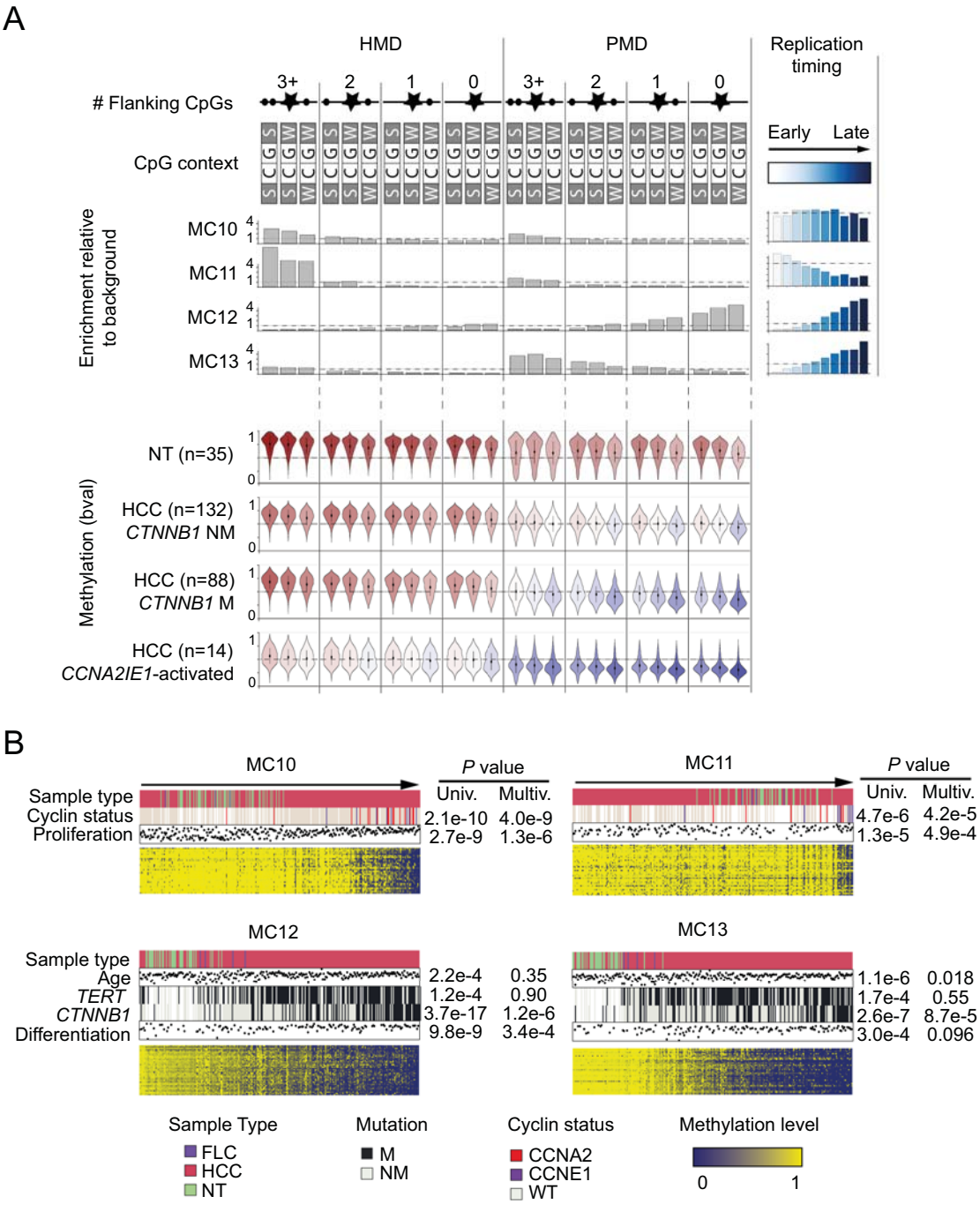

Select most contributing
CpG sites (absolute
contribution > 0.005)
Plot of methylation level
Hyper-
methylated
Hypo-
methylated
bval in T
bval in NT
0
1
MC1
MC2
MC3
MC4
MC5
MC6
MC7
MC8
MC9
MC10
MC11
MC12
MC13

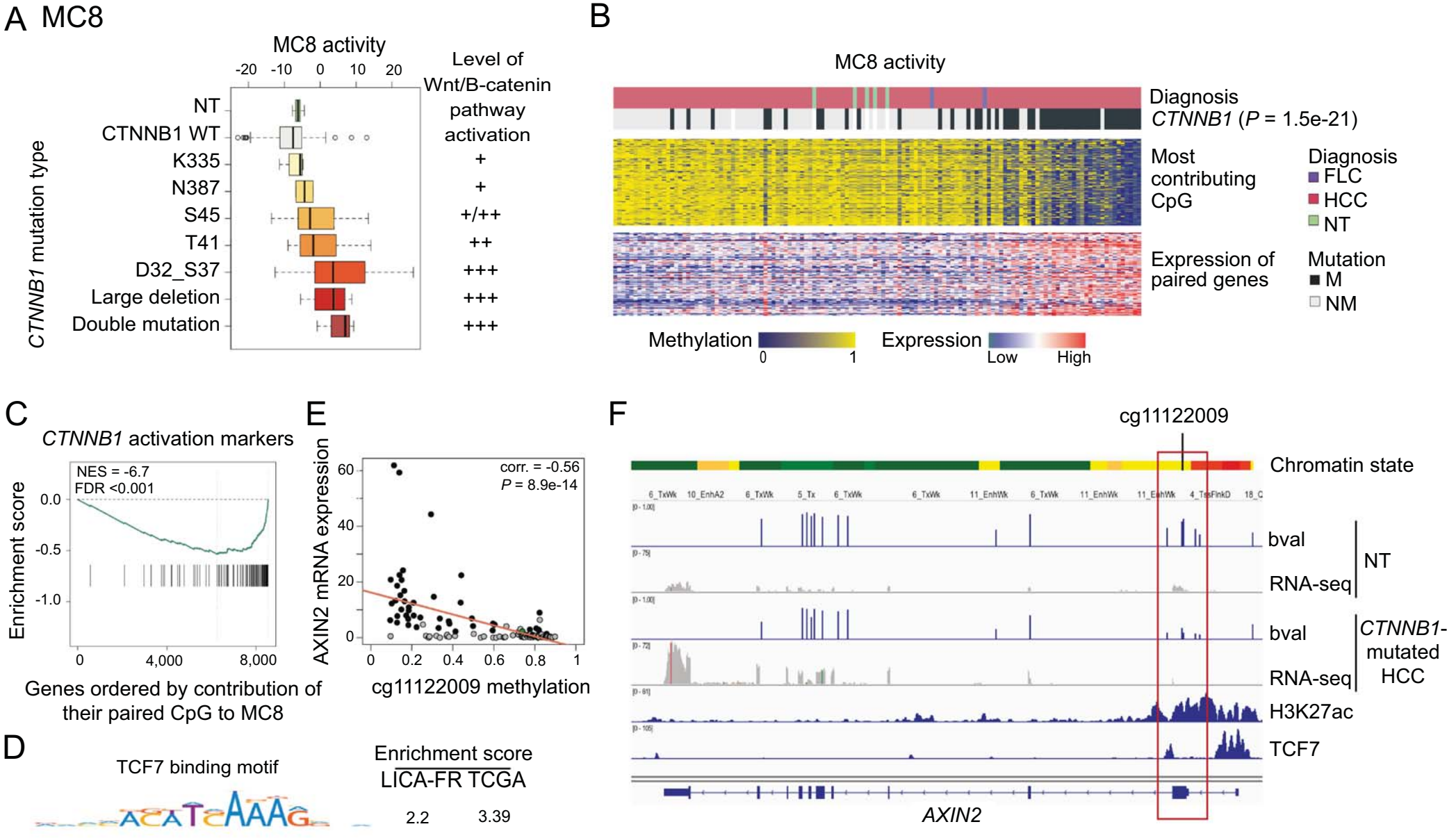

A
Hypermethylation
Hypomethylation
MC1
MC2
MC3
MC4
MC5
MC6
MC7
MC8
MC9
MC10
MC11
MC12
MC13
CGI feature
Gene feature
Chromatin state
Active
Inactive
Chromatin domain (Roadmap)
1 Active TSS
2 Flanking TSS
3 Flanking TSS upstream
4 Flanking TSS downstream
5 Strong transcription
6 Weak Transcription
7 Genic enhancer 1
8 Genic enhancer 2
9 Active enhancer 1
10 Active enhancer 2
11 Weak enhancer
12 ZNF genes and repeats
13 Heterochromatin
14 Bivalent/poised TSS
15 Bivalent enhancer
16 Repressed Polycomb
17 Weak Repressed Polycomb
18 Quiescent/low
CGI feature
Island
Shore
Shelf
Out
Gene feature
TSS500
Body
Out
Enrichment score
≥6
5
4
3
2
≤1
B
HMD
PMD
LMR
UMR
# Flanking CpGs
3+
2
1
0
CpG context
S C G S
S C G W
W C G W
W C G
Beta Value
NT (n=35)
1.0
0.5
0.0
Enrichment of CpG context in each signature compared to all CpGs analyzed
Hypermethylation
Hypomethylation
MC1
MC2
MC3
MC4
MC5
MC6
MC7
MC8
MC9
MC10
MC11
MC12
MC13

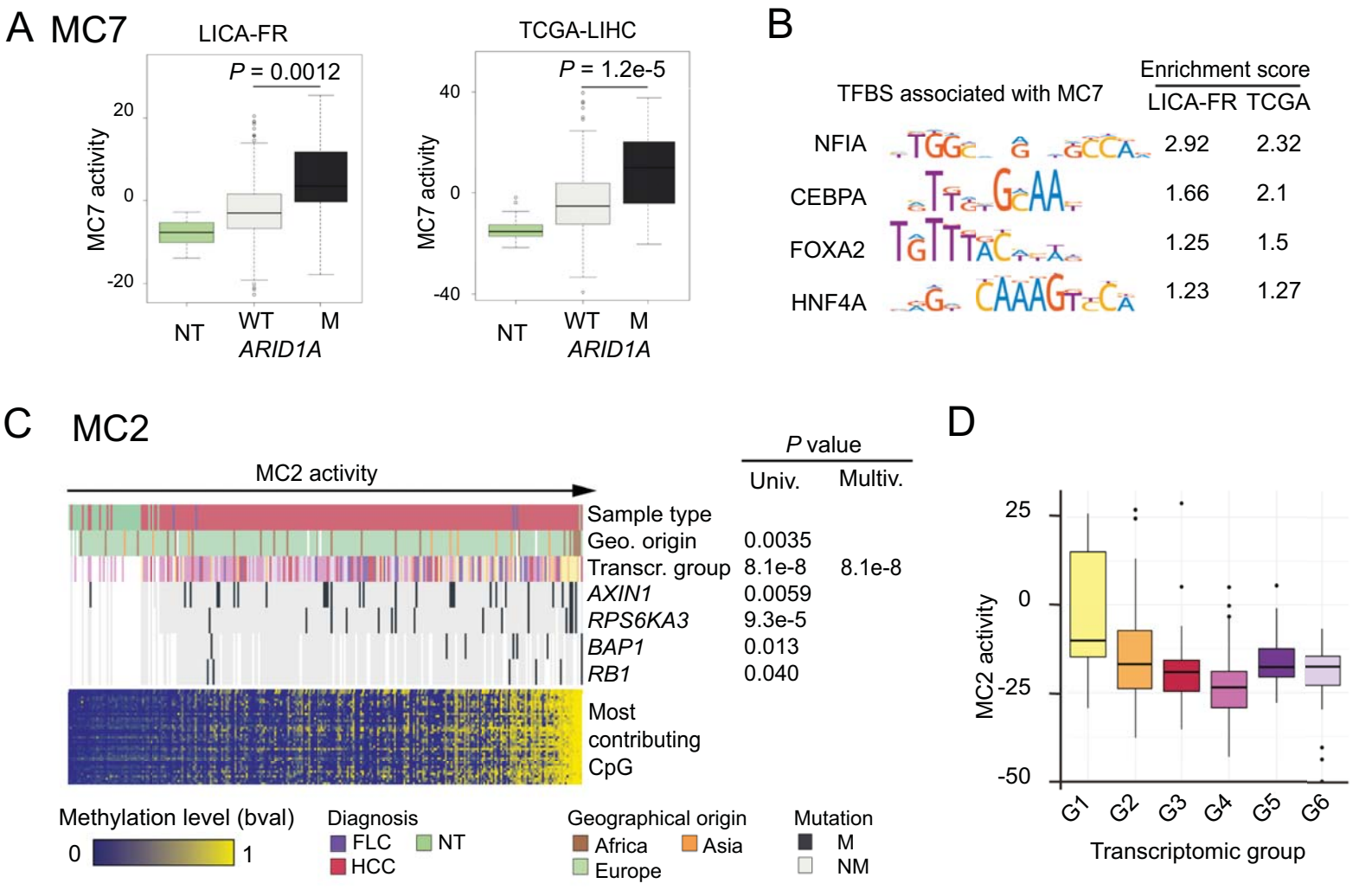

A
Univariate analysis
Clinical features
Driver alterations
Molecular subgroups*
Gene expression signatures**
Gender
Geographical origin
Age
Hepatitis B
Hepatitis C
Tobacco
Alcohol intake
Edmonson grade
Inflammation
Vascular invasion
TERT promoter
TP53
CTNNB1
AXIN1
ARID2
ARID1A
RPS6KA3
KEAP1
RB1
BAP1
CCNA2/CCNE1
G1
G2
G3
G4
G5
G6
Immune infiltrate
Liver progenitor
Stem cell
EMT/metastasis
Differentiation
Proliferation
LICA-FR
TCGA LIHC
MC1
MC2
MC3
MC4
MC5
MC6
MC7
MC8
MC9
MC10
MC11
MC12
MC13
P value
0.05 10-3 10-5 10-7 10-9 10-10
Significant association in both series


B
Multivariate analysis
Clinical features
Driver alterations
Molecular subgroups*
Gene expression signatures**
Gender
Geographical origin
Age
Hepatitis B
Hepatitis C
Tobacco
Alcohol intake
Edmonson grade
Inflammation
Vascular invasion
TERT promoter
TP53
CTNNB1
AXIN1
ARID2
ARID1A
RPS6KA3
KEAP1
RB1
BAP1
CCNA2/CCNE1
G1
G2
G3
G4
G5
G6
Immune infiltrate
Liver progenitor
Stem cell
EMT/metastasis
Differentiation
Proliferation
LICA-FR
TCGA LIHC
MC1
MC2
MC3
MC4
MC5
MC6
MC7
MC8
MC9
MC10
MC11
MC12
MC13


Gender-related component MC3
A
P value univariate = 5.0e-64
multivariate = 3.8e-59
Samples contribution
30
20
10
0
F (55)
M (218)
Gender
B
Other (3%)
Chr Y (1%)
Chr X (96%)
C
Methylation in females (bval)
1
0.75
0.5
0.25
0
0
0.25
0.5
0.75
1
Mean methylation in males
D
Age-related components
MC1
MC12
MC13
LICA-FR
TCGA-LIHC
Mean methylation of MRCpGs
Age at diagnosis
HCC: P = 3.2e-11; Slope = 0.0032
NT: NS
HCC: P = 2.1e-5; Slope = -0.0028
NT: NS
HCC: P = 2.1e-6; Slope = -0.0041
NT: NS
HCC: P = 1.0e-6; Slope = 0.0021
NT: P = 2.5e-5, Slope = 0.00053
HCC: P = 1.0e-3 ; Slope = -0.0032
NT: P = 0.021; Slope = 0.00060
HCC: P = 1.7e-8; Slope = -0.0049
NT: NS


A
HMD
PMD
Replication timing
# Flanking CpGs
3+
2
1
0
CpG context
S C G S
S C G W
W C G W
Early
Late
Enrichment relative to background
MC10
MC11
MC12
MC13
Methylation (bval)
NT (n=35)
HCC (n=132) CTNNB1 NM
HCC (n=88) CTNNB1 M
HCC (n=14) CCNA2IE1-activated
B
MC10
P value
Univ.
Multiv.
Sample type
Cyclin status
Proliferation
2.1e-10
4.0e-9
2.7e-9
1.3e-6
MC11
4.7e-6
4.2e-5
1.3e-5
4.9e-4
MC12
Sample type
Age
TERT
CTNNB1
Differentiation
2.2e-4
0.35
1.2e-4
0.90
3.7e-17
1.2e-6
9.8e-9
3.4e-4
MC13
1.1e-6
0.018
1.7e-4
0.55
2.6e-7
8.7e-5
3.0e-4
0.096
Sample Type
FLC
HCC
NT
Mutation
M
NM
Cyclin status
CCNA2
CCNE1
WT
Methylation level
0
1

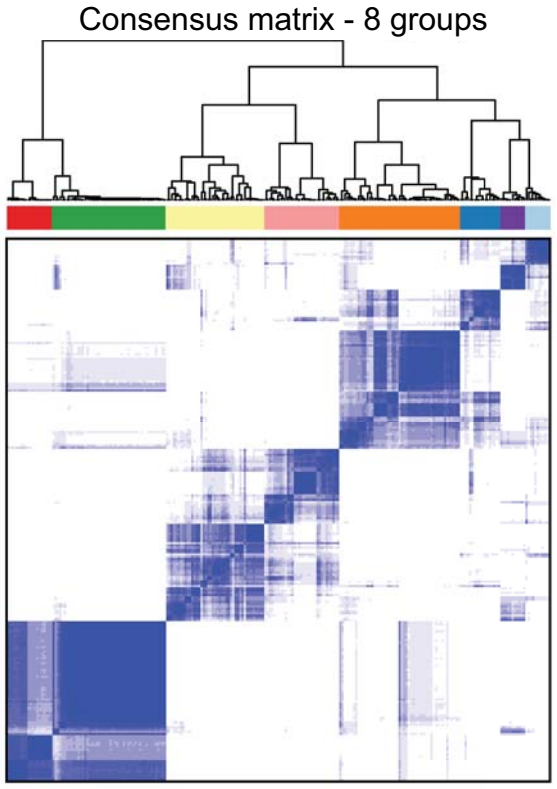

A MC8
MC8 activity
-20 -10 0 10 20
Level of Wnt/B-catenin pathway activation
CTNNB1 mutation type
NT
CTNNB1 WT
K335
N387
S45
T41
D32_S37
Large deletion
Double mutation
+
+
+/++
++
+++
+++
+++
B
MC8 activity
Diagnosis
CTNNB1 (P = 1.5e-21)
Most contributing CpG
Expression of paired genes
Diagnosis
FLC
HCC
NT
Mutation
M
NM
Methylation
0
1
Expression
Low
High
C
CTNNB1 activation markers
NES = -6.7
FDR <0.001
Enrichment score
0.0
-0.5
-1.0
0
4,000
8,000
Genes ordered by contribution of their paired CpG to MC8
E
AXIN2 mRNA expression
corr. = -0.56
P = 8.9e-14
60
40
20
0
0 0.2 0.4 0.6 0.8 1
cg11122009 methylation
D
TCF7 binding motif
Enrichment score
LICA-FR TCGA
2.2
3.39
F
cg11122009
Chromatin state
6_TxWk
10_EnhA2
6_TxWk
5_Tx
6_TxWk
6_TxWk
11_EnhWk
6_TxWk
11_EnhWk
11_EnhWk
4_TssFlnkD
18_C
bval
RNA-seq
NT
bval
RNA-seq
CTNNB1-mutated HCC
H3K27ac
TCF7
AXIN2


A MC7
LICA-FR
P = 0.0012
MC7 activity
20
0
-20
NT
WT
M
ARID1A
TCGA-LIHC
P = 1.2e-5
40
0
-40
B
TFBS associated with MC7
Enrichment score
LICA-FR TCGA
NFIA 2.92 2.32
CEBPA 1.66 2.1
FOXA2 1.25 1.5
HNF4A 1.23 1.27
C MC2
MC2 activity
P value
Univ. Multiv.
Sample type
Geo. origin 0.0035
Transcr. group 8.1e-8 8.1e-8
AXIN1 0.0059
RPS6KA3 9.3e-5
BAP1 0.013
RB1 0.040
Most contributing CpG
Methylation level (bval)
0 1
Diagnosis
FLC NT
HCC
Geographical origin
Africa Asia
Europe
Mutation
M
NM
D
MC2 activity
25
0
-25
-50
G1 G2 G3 G4 G5 G6
Transcriptomic group

A

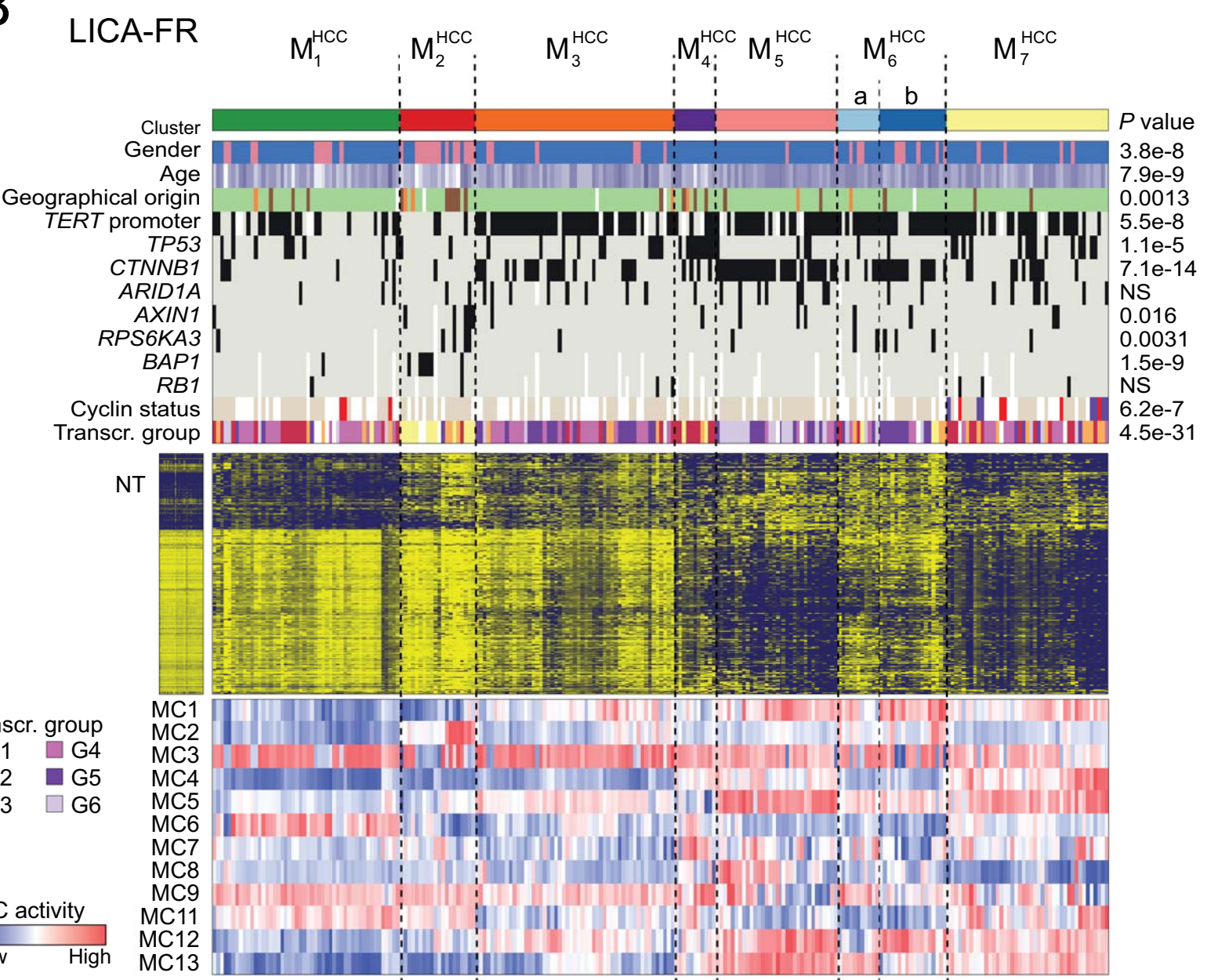
LICA-FR

Consensus matrix - 8 groups



B

LICA-FR




Gender: F, M

Geo. origin: Africa, Asia, Europe

Cyclin status: CCNA2, CCNE1, WT

Transcr. group: G1, G2, G3, G4, G5, G6

Mutation: M, NM

Age: 16 – 88

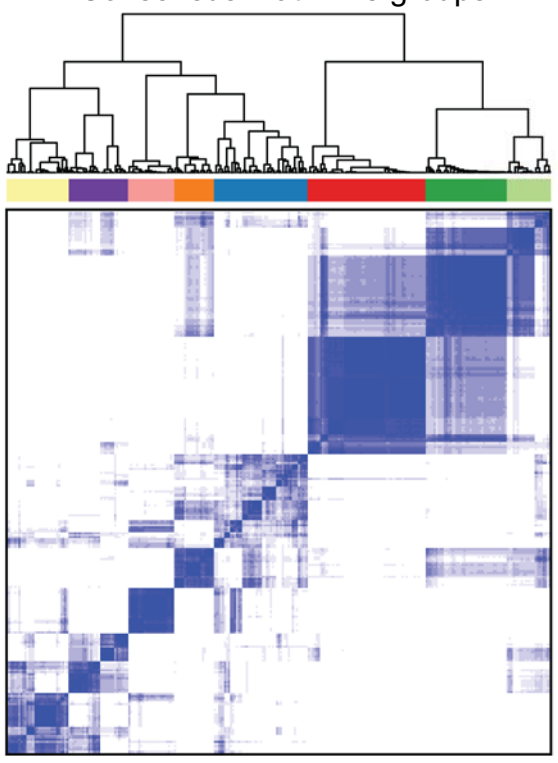
Methylation level: 0 – 1

MC activity: Low – High

C

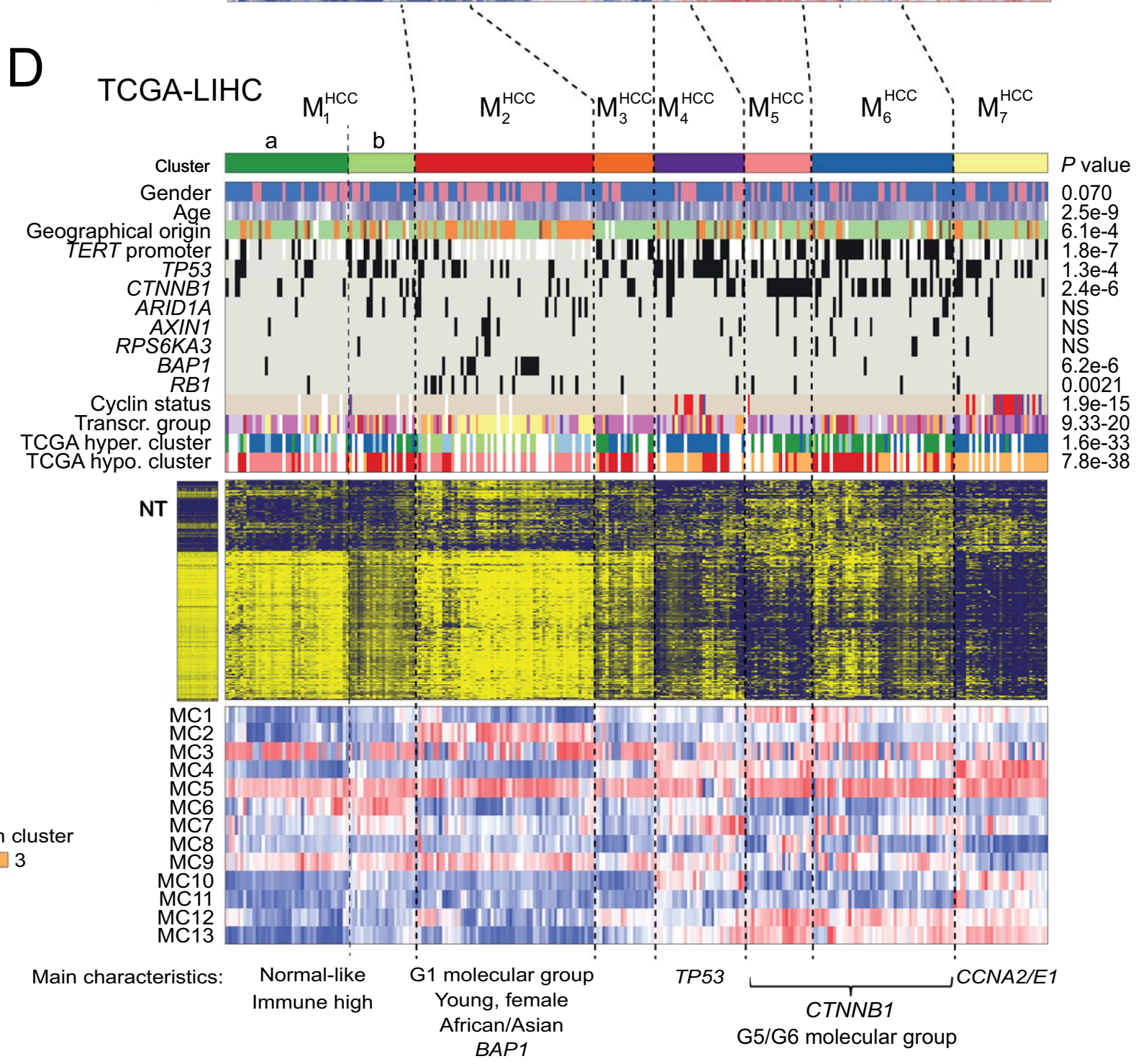
TCGA-LIHC

Consensus matrix - 8 groups



D

TCGA-LIHC




TCGA classification

Hypermethylation cluster: 1, 2, 3, 4

Hypomethylation cluster: 1, 2, 3