Explainable artificial intelligence of DNA methylation-based brain tumor diagnostics

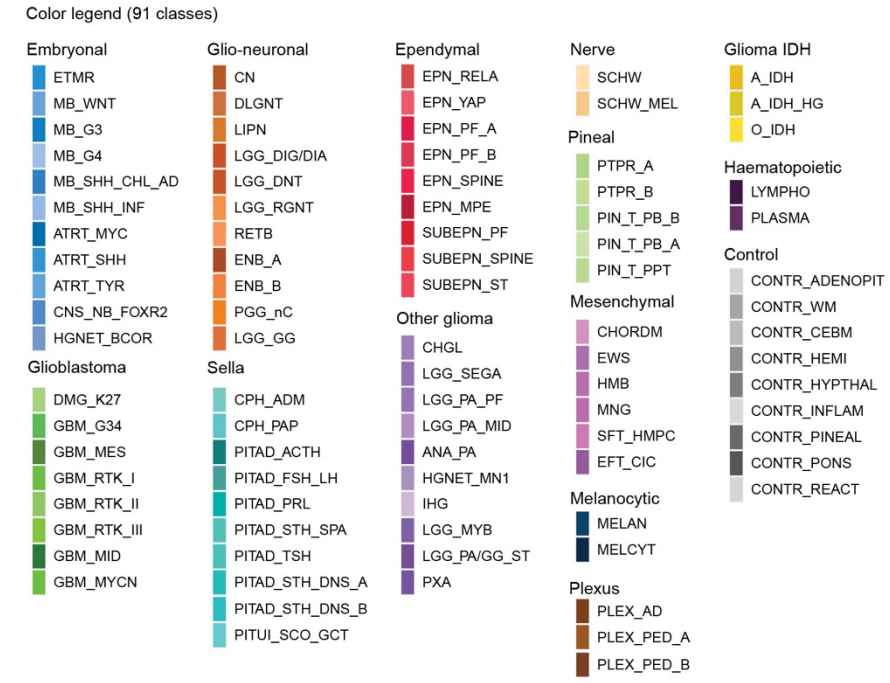
Supplementary Information

Salvatore Benfatto[1,2,3], Martin Sill[4,5,6], David T.W. Jones[5,6,7], Stefan M. Pfister[4,5,6,8], Felix Sahm[9,10], Andreas von Deimling[9,10], David Capper [11,12], Volker Hovestadt[1,2,3*]

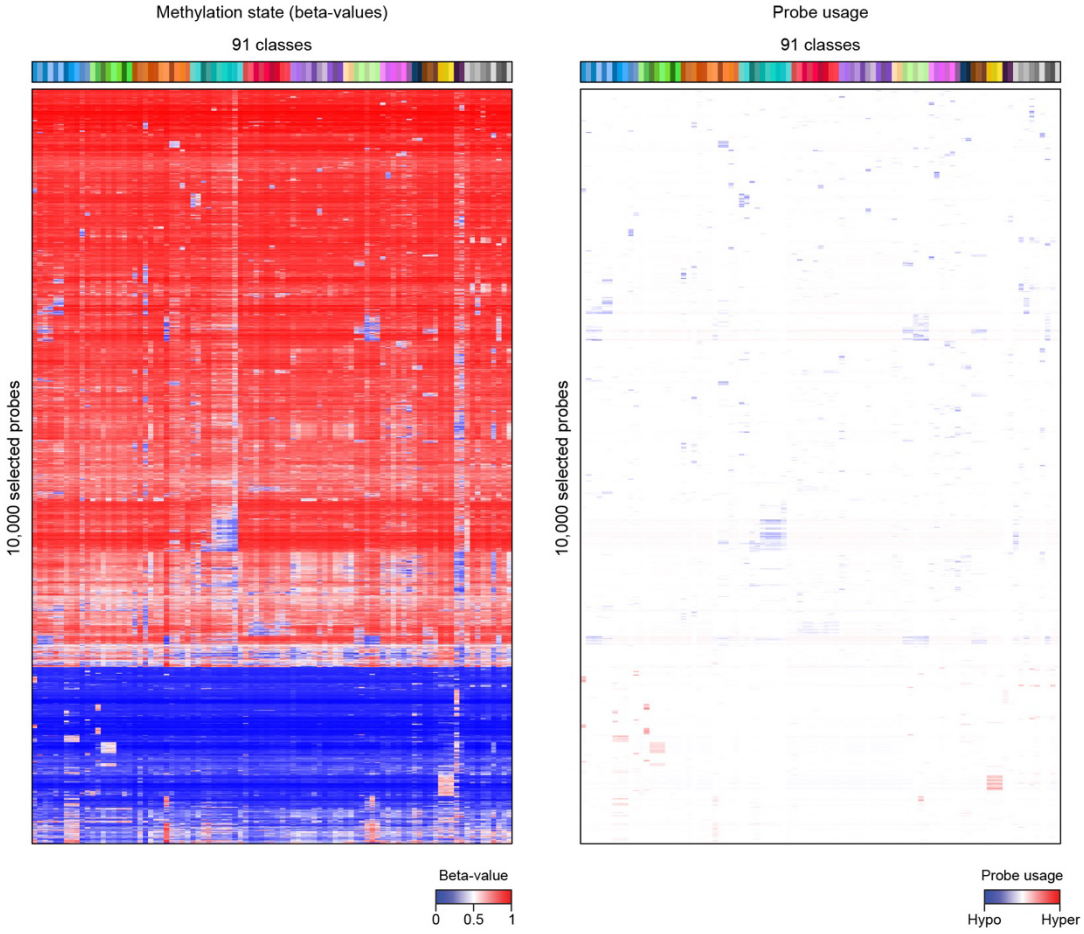
[1] Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA.
[2] Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA.
[3] Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[4] Division of Pediatric Neurooncology, Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany
[5] National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and Heidelberg University Hospital, Germany
[6] German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany.
[7] Division of Pediatric Glioma Research, Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany
[8] Department of Pediatric Oncology, Hematology & Immunology, Heidelberg University Hospital, Heidelberg, Germany.
[9] Department of Neuropathology, Heidelberg University Hospital, Heidelberg, Germany.
[10] Clinical Cooperation Unit Neuropathology, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany.
[11] Department of Neuropathology, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany.
[12] German Cancer Consortium (DKTK), Partner Site Berlin, German Cancer Research Center (DKFZ), Heidelberg, Germany.

*Correspondence: volker_hovestadt@dfci.harvard.edu

a


Color legend (91 classes)
Embryonal
ETMR
MB_WNT
MB_G3
MB_G4
MB_SHH_CHL_AD
MB_SHH_INF
ATRT_MYC
ATRT_SHH
ATRT_TYR
CNS_NB_FOXR2
HGNET_BCOR
Glioblastoma
DMG_K27
GBM_G34
GBM_MES
GBM_RTK_I
GBM_RTK_II
GBM_RTK_III
GBM_MID
GBM_MYCN
Glio-neuronal
CN
DLGNT
LIPN
LGG_DIG/DIA
LGG_DNT
LGG_RGNT
RETB
ENB_A
ENB_B
PGG_nC
LGG_GG
Sella
CPH_ADM
CPH_PAP
PITAD_ACTH
PITAD_FSH_LH
PITAD_PRL
PITAD_STH_SPA
PITAD_TSH
PITAD_STH_DNS_A
PITAD_STH_DNS_B
PITUI_SCO_GCT
Ependymal
EPN_RELA
EPN_YAP
EPN_PF_A
EPN_PF_B
EPN_SPINE
EPN_MPE
SUBEPN_PF
SUBEPN_SPINE
SUBEPN_ST
Other glioma
CHGL
LGG_SEGA
LGG_PA_PF
LGG_PA_MID
ANA_PA
HGNET_MN1
IHG
LGG_MYB
LGG_PA/GG_ST
PXA
Nerve
SCHW
SCHW_MEL
Pineal
PTPR_A
PTPR_B
PIN_T_PB_B
PIN_T_PB_A
PIN_T_PPT
Mesenchymal
CHORDM
EWS
HMB
MNG
SFT_HMPC
EFT_CIC
Melanocytic
MELAN
MELCYT
Plexus
PLEX_AD
PLEX_PED_A
PLEX_PED_B
Glioma IDH
A_IDH
A_IDH_HG
O_IDH
Haematopoietic
LYMPHO
PLASMA
Control
CONTR_ADENOPIT
CONTR_WM
CONTR_CEBM
CONTR_HEMI
CONTR_HYPTHAL
CONTR_INFLAM
CONTR_PINEAL
CONTR_PONS
CONTR_REACT

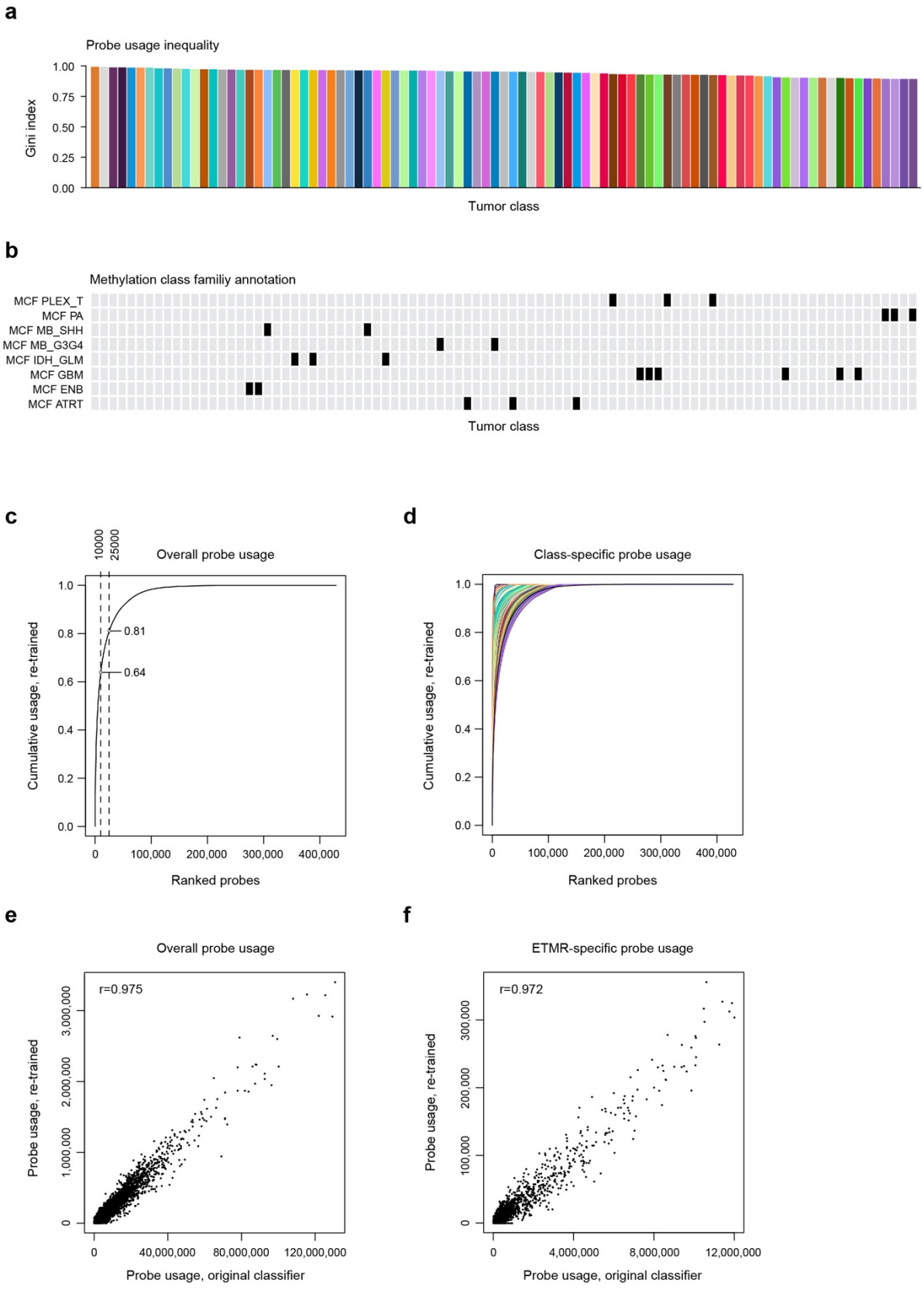
b


Methylation state (beta-values)
91 classes
10,000 selected probes
Beta-value
0
0.5
1
Probe usage
91 classes
10,000 selected probes
Probe usage
Hypo
Hyper

Supplementary Figure 1 | Overview of the reference cohort.

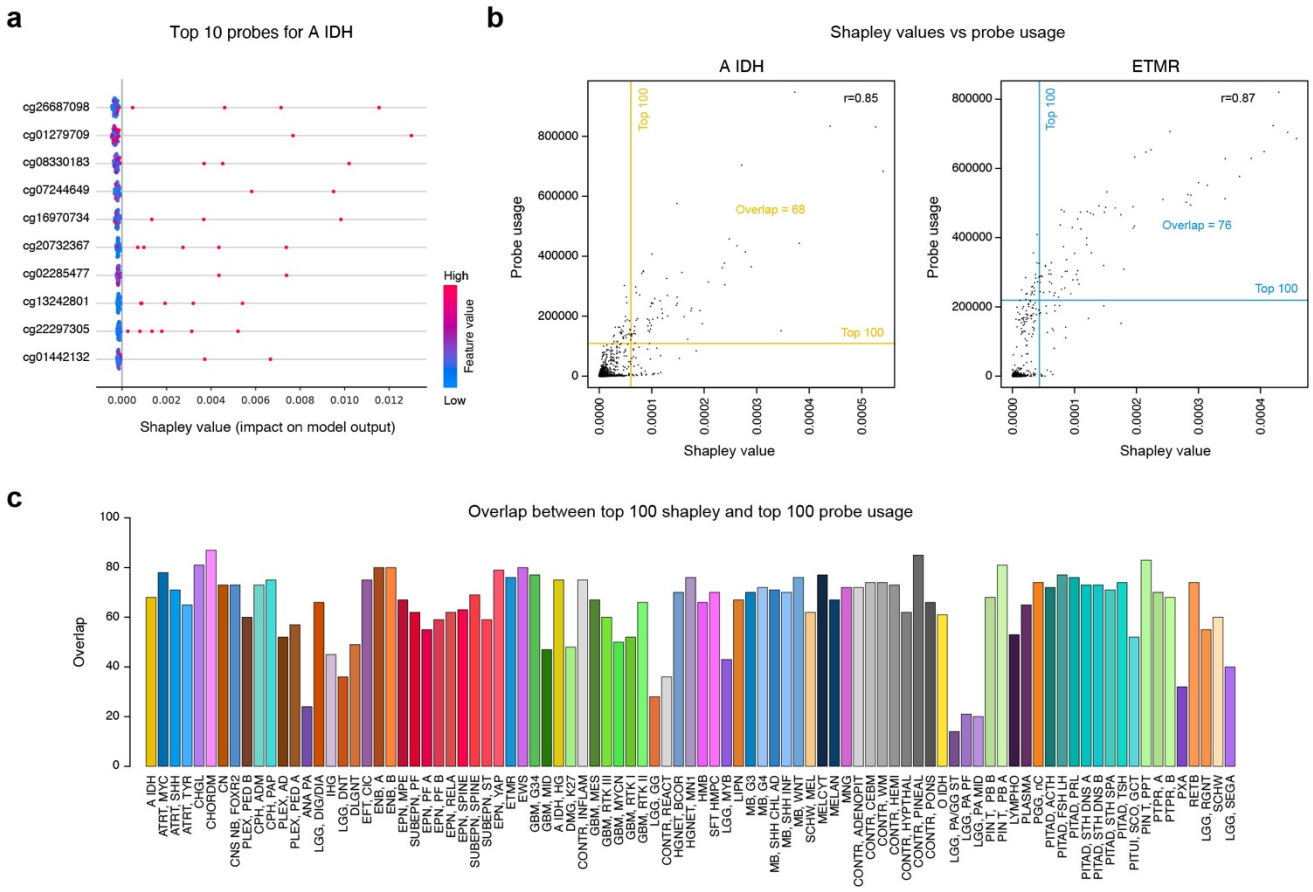
a Color legend of 82 tumor classes and 9 control classes in the Heidelberg brain tumor classifier. Colors are used throughout the manuscript.

b Average DNA methylation beta-values (left) and probe usage (right) of the 10,000 selected probes of the inner classifier (rows) for each of the 91 classes (columns).


a
Probe usage inequality
Gini index
1.00
0.75
0.50
0.25
0.00
Tumor class
b
Methylation class familiy annotation
MCF PLEX_T
MCF PA
MCF MB_SHH
MCF MB_G3G4
MCF IDH_GLM
MCF GBM
MCF ENB
MCF ATRT
Tumor class
c
Overall probe usage
10000
25000
0.81
0.64
Cumulative usage, re-trained
Ranked probes
d
Class-specific probe usage
Cumulative usage, re-trained
Ranked probes
e
Overall probe usage
r=0.975
Probe usage, re-trained
Probe usage, original classifier
f
ETMR-specific probe usage
r=0.972
Probe usage, re-trained
Probe usage, original classifier

Supplementary Figure 2 | Global probe usage patterns and probe usage stability in an independent model.

a Barplot shows the Gini index of the probe usage for each tumor class. A higher Gini index indicates that fewer probes are used more frequently.
b Table shows tumor classes (columns) that are part of methylation class families (MCFs, rows). Tumor classes are sorted as in panel (a).
c Plot shows the overall probe usage of a re-trained independent RF model with probes ranked from the most to least used. Vertical lines indicate the fraction of the total usage of the first 10,000 and 25,000 probes.
d Plot shows the cumulative probe usage of an independent RF model for each tumor class. The probes are ranked by their usage.
e Scatterplot shows the overall probe usages of the re-trained RF model against the original RF model.
f Same as **e** for the ETMR tumor class.

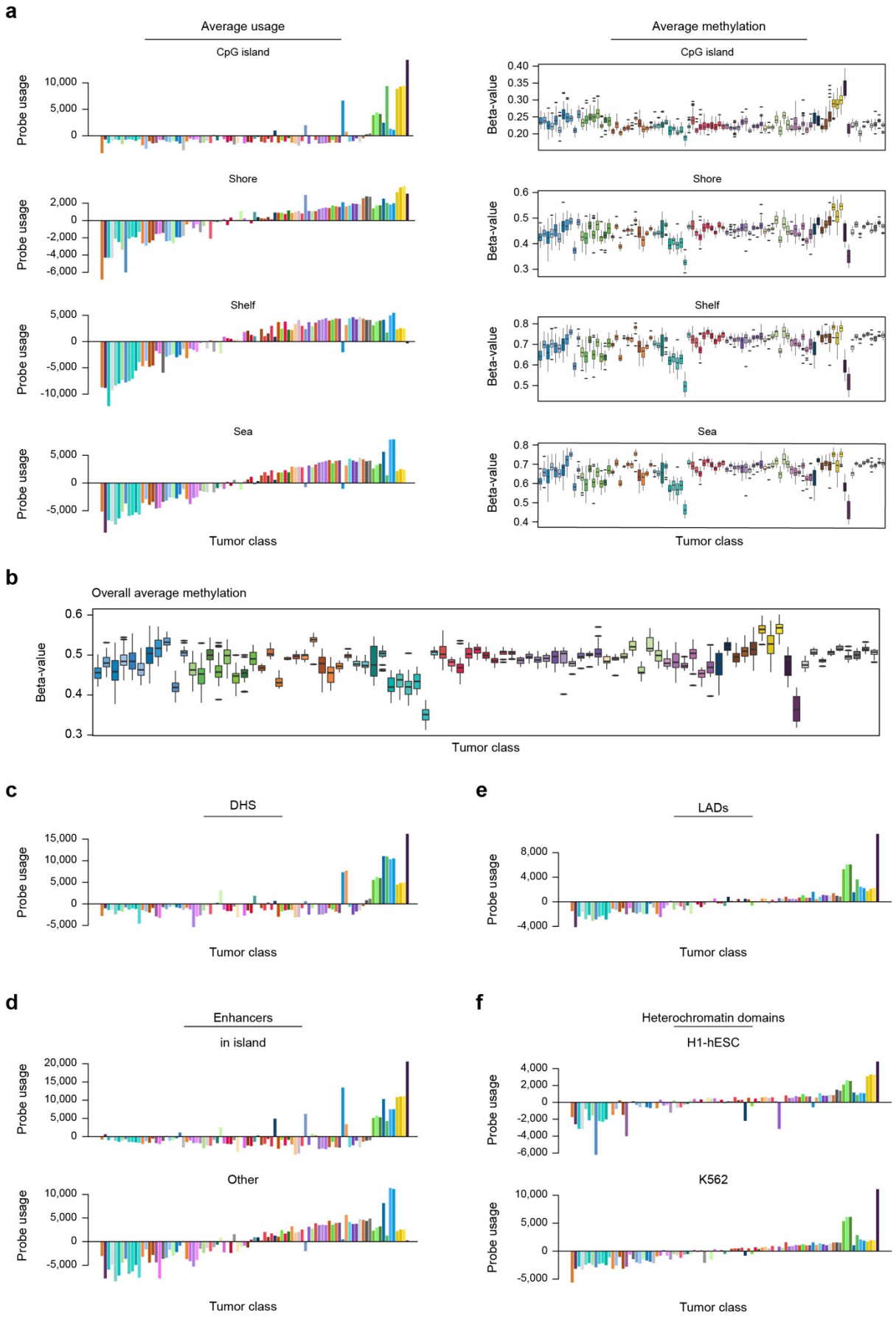


Supplementary Figure 3 | Results of the SHAP approach applied to the brain tumor classifier dataset.

a Summary plot of the top 10 probes for the A_IDH class (y-axis). Shapley values are indicated on the x-axis. Results indicate that probes hypermethylated in IDH-mutant classes are frequently selected during training of the random forest model to separate these classes.

b Left scatter plot compares A_IDH-specific Shapley values to our previously generated probe usage values. Results indicate a high agreement between the two approaches (Pearson's r=0.85). Overlap between the top 100 probes from each approach are indicated. Right scatter plot similarly compares ETMR-specific values.

c Bar plot shows overlap between the top 100 probes using the Shapley approach and our previous probe usage approach for all 91 classes. High concordance can be observed.


a
Average usage
CpG island
Probe usage
10,000
5,000
0
Shore
2,000
0
-2,000
-4,000
-6,000
Shelf
5,000
0
-5,000
-10,000
Sea
5,000
0
-5,000
Tumor class
Average methylation
CpG island
Beta-value
0.40
0.35
0.30
0.25
0.20
Shore
0.6
0.5
0.4
0.3
Shelf
0.8
0.7
0.6
0.5
Sea
0.8
0.7
0.6
0.5
0.4
Tumor class
b
Overall average methylation
Beta-value
0.6
0.5
0.4
0.3
Tumor class
c
DHS
Probe usage
15,000
10,000
5,000
0
-5,000
Tumor class
e
LADs
Probe usage
8,000
4,000
0
-4,000
Tumor class
d
Enhancers
in island
Probe usage
20,000
15,000
10,000
5,000
0
Other
10,000
5,000
0
-5,000
Tumor class
f
Heterochromatin domains
H1-hESC
Probe usage
4,000
2,000
0
-2,000
-4,000
-6,000
K562
10,000
5,000
0
-5,000
Tumor class

Supplementary Figure 4 | Global probe usage patterns by functional genomic regions of different sizes.

a Barplots on the left show the average probe usage by tumor class. Positive and negative values indicate that probes are predominantly hyper- and hypomethylated, respectively. Probes are grouped by different genomic regions: CpG islands, Shore (0 to 2 kb from island), Shelf (2 to 4 kb from island and Sea (greater than 4 kb from island) regions. Boxplots on the right indicate average DNA methylation values in the same regions for each tumor class.

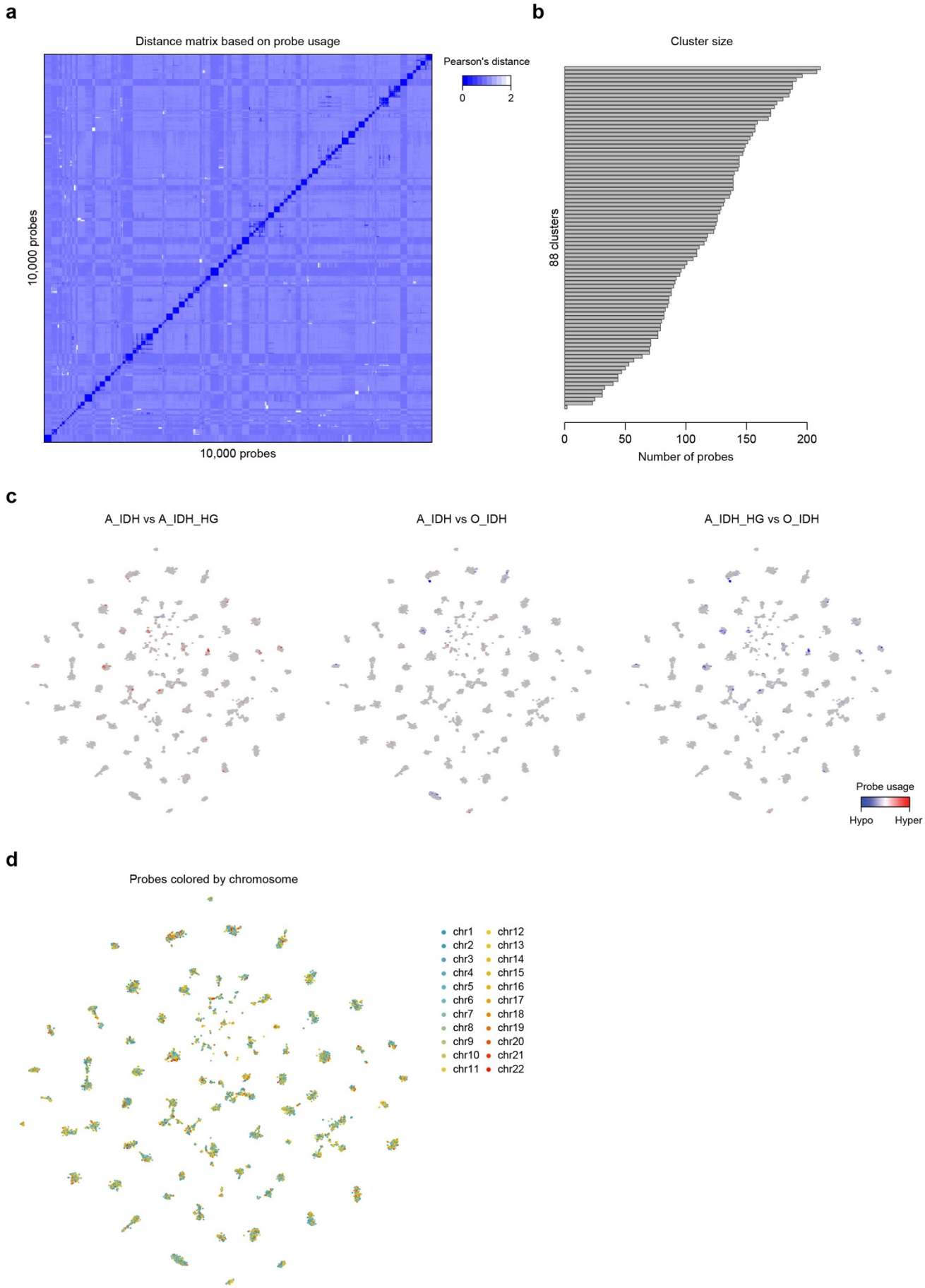
b Boxplot indicates the genome-wide average DNA methylation value for each sample in the reference cohort, grouped by tumor class.

c Barplot shows the average probe usage by tumor class within DHS regions. Positive and negative values indicate that probes are predominantly hyper- and hypomethylated, respectively.

d Similar barplots show the average probe usage within annotated enhancer regions that overlap CpG islands (top) or not (bottom).

e Similar barplot shows the average probe usage within broad lamina-associated domains (LADs).

f Similar barplots show the average probe usage within heterochromatic domains in two different cell lines (top: H1 embryonic stem cells, bottom: K562 cells).

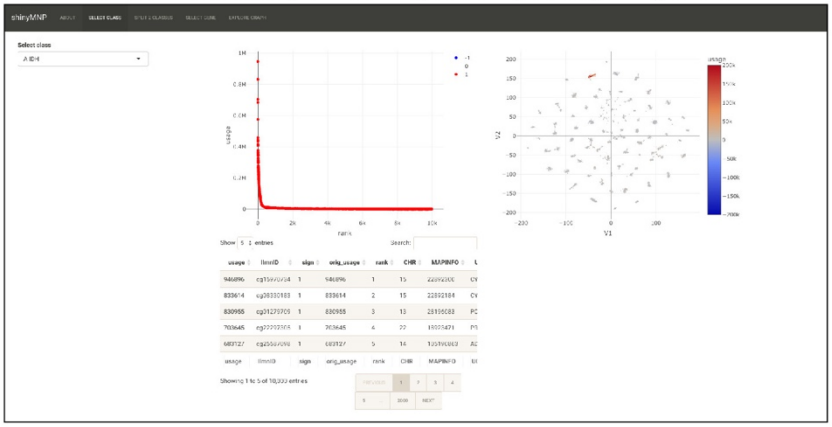

a
Distance matrix based on probe usage
Pearson's distance
0
2
10,000 probes
10,000 probes
b
Cluster size
88 clusters
0
50
100
150
200
Number of probes
c
A_IDH vs A_IDH_HG
A_IDH vs O_IDH
A_IDH_HG vs O_IDH
Probe usage
Hypo
Hyper
d
Probes colored by chromosome
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22

Supplementary Figure 5 | Clustering of probes with similar usage.

a Heatmap shows the Pearson's distance between the 10,000 selected probes. Distance is calculated on the probe usage. If usage between probes is very similar (they are redundant), they have a low distance (indicated by dark blue color).
b Barplot indicates the number of probes in each of the 88 identified clusters.
c t-SNE visualizations showing the probe usage between classes of the IDH-mutation glioma methylation class family. Red color indicates that the probe is hypermethylated and blue color indicates that the probe is hypomethylated in the tumor class that is named first.
d t-SNE visualization as in Fig. 3a colored by the chromosome the probe is mapping to.

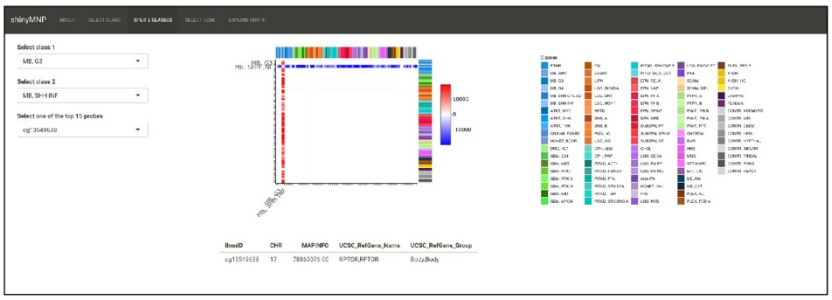
a

shinyMNP interface - Probes ranked by cumulative usage and tSNE probe usage for a given class



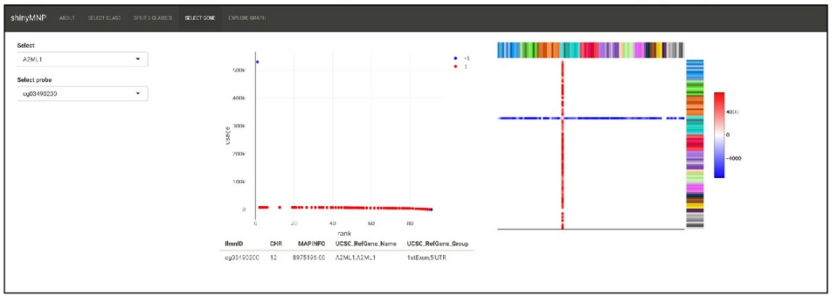
b

shinyMNP interface - Top N probes to split any pairwise combination of tumor classes

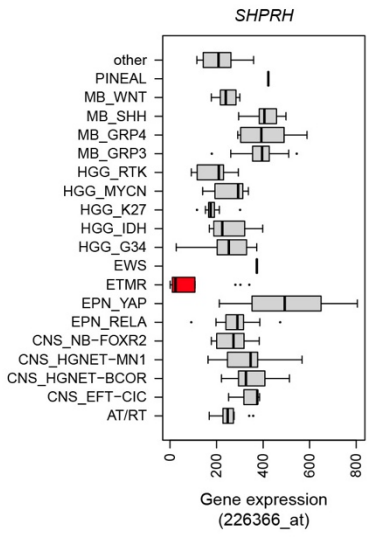


c

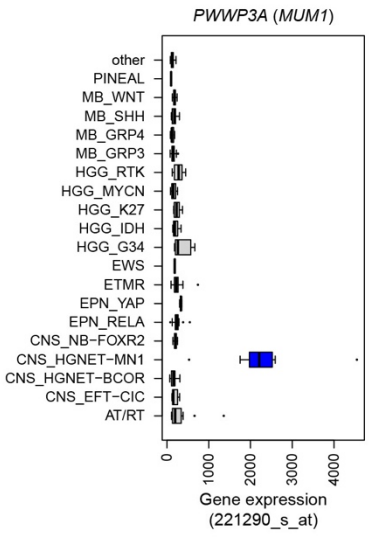
shinyMNP interface - Probes associated to a gene of interest, ranked cumulative usage plot by class and heatmap



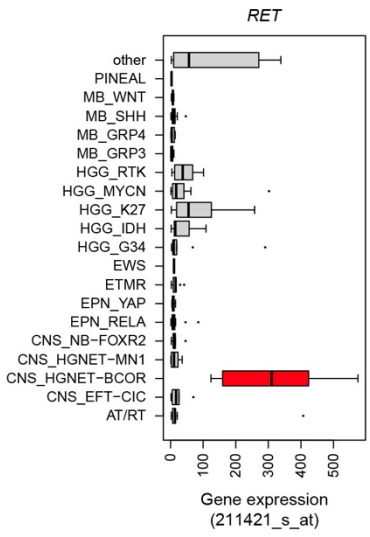
d



e



f



Supplementary Figure 6 | Overview of the shinyMNP web application and gene expression of class-specifically methylated genes.

a Screenshot of the first panel of the shinyMNP web application. For the selected class, the total usage of the 10,000 probes is shown as an interactive ranked plot. The same information is depicted in the t-SNE by the color gradient. Probe usage is proportional to the color intensity, red and blue indicate hyper- and hypomethylation, respectively.

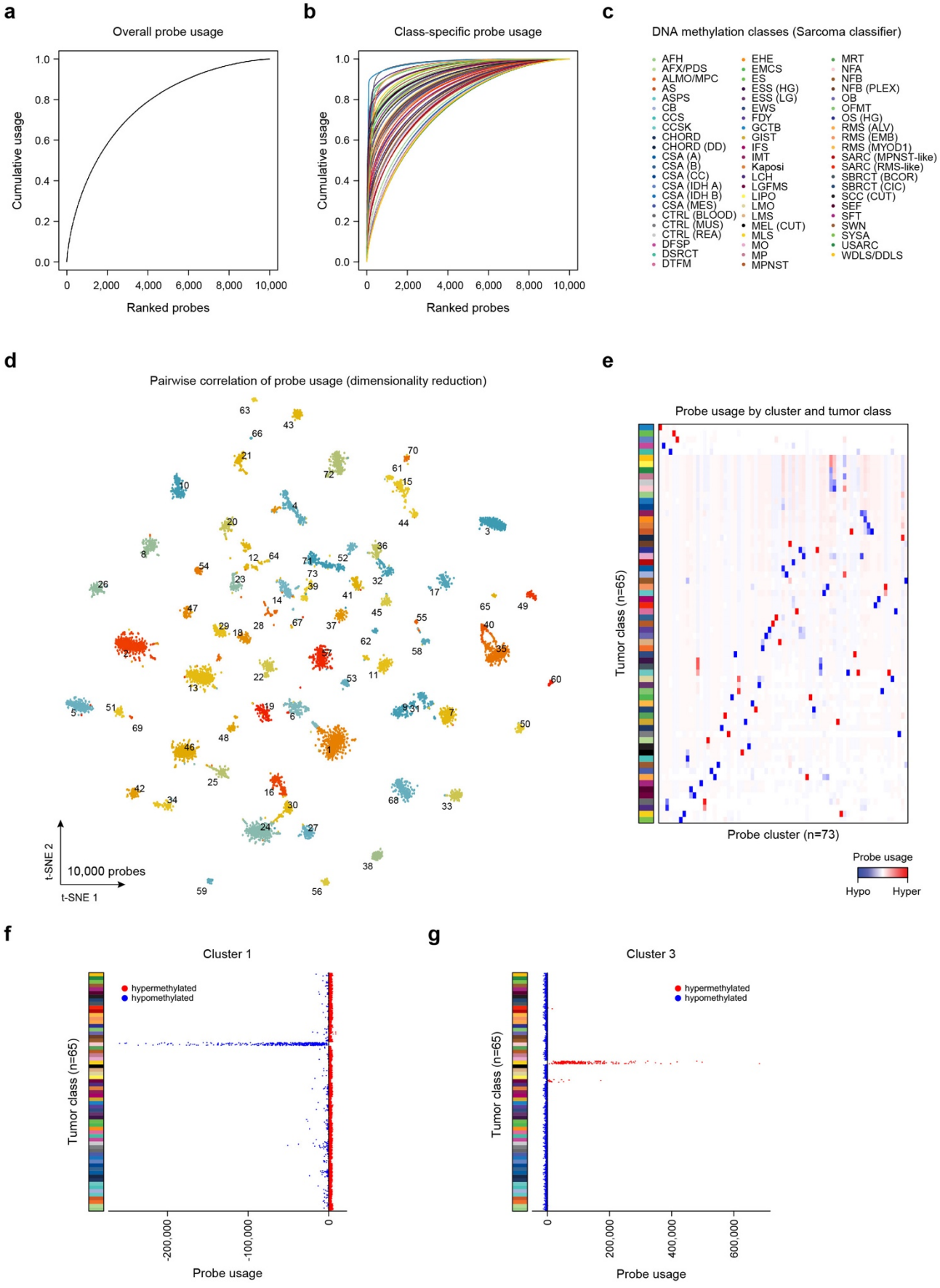
b Screenshot of the second panel in which users can identify the probes with the highest usage that split any two classes. The interactive heatmap shows the probe usage in the 8,281 possible class combinations.

c Screenshot of the third panel of the shinyMNP app. After selecting a gene of interest, users can select one of the multiple associated probes. An interactive ranked plot displays the total probe usage by class and the interactive heatmap shows the probe usage in the 8,281 possible class combinations.

d Boxplot shows expression of the *SHPRH* gene, as measured by Affymetrix gene expression array across different tumor classes. *SHPRH* is specifically silenced in ETMR.

e Boxplot shows high expression of *PWWP3A* (also known as *MUM1*) specifically in the HGNET_MN1 tumor class.

f Boxplot shows high expression of the *RET* oncogene specific to the HGNET_BCOR tumor class.


a
Overall probe usage
Cumulative usage
Ranked probes
b
Class-specific probe usage
Cumulative usage
Ranked probes
c
DNA methylation classes (Sarcoma classifier)
AFH
AFX/PDS
ALMO/MPC
AS
ASPS
CB
CCS
CCSK
CHORD
CHORD (DD)
CSA (A)
CSA (B)
CSA (CC)
CSA (IDH A)
CSA (IDH B)
CSA (MES)
CTRL (BLOOD)
CTRL (MUS)
CTRL (REA)
DFSP
DSRCT
DTFM
EHE
EMCS
ES
ESS (HG)
ESS (LG)
EWS
FDY
GCTB
GIST
IFS
IMT
Kaposi
LCH
LGFMS
LIPO
LMO
LMS
MEL (CUT)
MLS
MO
MP
MPNST
MRT
NFA
NFB
NFB (PLEX)
OB
OFMT
OS (HG)
RMS (ALV)
RMS (EMB)
RMS (MYOD1)
SARC (MPNST-like)
SARC (RMS-like)
SBRCT (BCOR)
SBRCT (CIC)
SCC (CUT)
SEF
SFT
SWN
SYSA
USARC
WDLS/DDLS
d
Pairwise correlation of probe usage (dimensionality reduction)
t-SNE 2
t-SNE 1
10,000 probes
e
Probe usage by cluster and tumor class
Tumor class (n=65)
Probe cluster (n=73)
Probe usage
Hypo
Hyper
f
Cluster 1
hypermethylated
hypomethylated
Tumor class (n=65)
Probe usage
g
Cluster 3
hypermethylated
hypomethylated
Tumor class (n=65)
Probe usage

Supplementary Figure 7 | Explainable AI of the sarcoma classifier.

a Plot shows the overall probe usage of the inner RF sarcoma model with probes ranked from the most to least used. The model was trained using the 10,000 most important features from the outer RF classifier.

b Plot shows the cumulative probe usage for each tumor class. The probes are ranked from the most to least used within each class. Different classes show a degree of probe usage inequality, i.e. few probes that are selected most of the time to distinguish one class from all others.

c Tumor class legend of the sarcoma cohort.

d t-SNE dimensionality reduction and unsupervised clustering of 10,000 probes according to their usage (Pearson's distance). A total of 73 clusters are identified. Individual clusters of probes are numbered. This analysis reveals a high redundancy of probes, similar to the CNS tumor classifier.

e Heatmap shows the total probe usage by class and by cluster. Blue color indicates clusters in which probes are predominantly hypomethylated in a given tumor class, red color indicates clusters which are predominantly hypermethylated. Clusters (columns) and tumor classes (rows) are ordered by hierarchical clustering. Values are scaled by column. Results indicate that the total probe usage in each cluster is predominantly high for one or few tumor classes, demonstrating that most clusters are tumor class specific.

f Scatterplot showing the usage of probes belonging to cluster 1 (n=469) for each tumor class. Probes of cluster 1 are often used to separate the tumor class dermatofibrosarcoma protuberans (DFSP) from all others.

e Scatterplot showing the usage of probes belonging to cluster 3 (n=231). Probes belonging to cluster 3 separate the tumor class epithelioid sarcoma (ES) from all others.

Methylation Class Name (v.11b2 classifier)	Methylation Class Name Abbreviated	Methylation Class Family (MCF)	MCF Name Abbreviated	Group Size (n)
methylation class IDH glioma, subclass astrocytoma	A IDH	methylation class family glioma, IDH mutant	MCF IDH GLM	78
methylation class IDH glioma, subclass high grade astrocytoma	A IDH, HG	methylation class family glioma, IDH mutant	MCF IDH GLM	46
methylation class anaplastic pilocytic astrocytoma	ANA PA			21
methylation class atypical teratoid/rhabdoid tumor, subclass MYC	ATRT, MYC	methylation class family atypical teratoid/rhabdoid tumor	MCF ATRT	29
methylation class atypical teratoid/rhabdoid tumor, subclass SHH	ATRT, SHH	methylation class family atypical teratoid/rhabdoid tumor	MCF ATRT	46
methylation class atypical teratoid/rhabdoid tumor, subclass TYR	ATRT, TYR	methylation class family atypical teratoid/rhabdoid tumor	MCF ATRT	37
methylation class chordoid glioma of the third ventricle	CHGL			12
methylation class chordoma	CHORDM			9
methylation class central neurocytoma	CN			21
methylation class CNS neuroblastoma with FOXR2 activation	CNS NB, FOXR2			39
methylation class control tissue, pituitary gland anterior lobe	CONTR, ADENOPIT			9
methylation class control tissue, cerebellar hemisphere	CONTR, CEBM			8
methylation class control tissue, hemispheric cortex	CONTR, HEMI			13
methylation class control tissue, hypothalamus	CONTR, HYPTHAL			9
methylation class control tissue, inflammatory tumor microenvironment	CONTR, INFLAM			24
methylation class control tissue, pineal gland	CONTR, PINEAL			12
methylation class control tissue, pons	CONTR, PONS			12
methylation class control tissue, reactive tumor microenvironment	CONTR, REACT			23
methylation class control tissue, white matter	CONTR, WM			9
methylation class craniopharyngioma, adamantinomatous	CPH, ADM			25
methylation class craniopharyngioma, papillary	CPH, PAP			20
methylation class diffuse leptomeningeal glioneuronal tumor	DLGNT			8
methylation class diffuse midline glioma H3 K27M mutant	DMG, K27			78
methylation class CNS Ewing sarcoma family tumor with CIC alteration	EFT, CIC			13
methylation class esthesioneuroblastoma, subclass A	ENB, A	methylation class family esthesioneuroblastoma	MCF ENB	23
methylation class esthesioneuroblastoma, subclass B	ENB, B	methylation class family esthesioneuroblastoma	MCF ENB	16
methylation class ependymoma, myxopapillary	EPN, MPE			28
methylation class ependymoma, posterior fossa group A	EPN, PF A			91
methylation class ependymoma, posterior fossa group B	EPN, PF B			51
methylation class ependymoma, RELA fusion	EPN, RELA			70
methylation class ependymoma, spinal	EPN, SPINE			27
methylation class ependymoma, YAP fusion	EPN, YAP			11
methylation class embryonal tumor with multilayered rosettes	ETMR			43
methylation class Ewing sarcoma	EWS			14
methylation class glioblastoma, IDH wildtype, H3.3 G34 mutant	GBM, G34			41
methylation class glioblastoma, IDH wildtype, subclass mesenchymal	GBM, MES	methylation class family glioblastoma, IDH wildtype	MCF GBM	56
methylation class glioblastoma, IDH wildtype, subclass midline	GBM, MID	methylation class family glioblastoma, IDH wildtype	MCF GBM	14
methylation class glioblastoma, IDH wildtype, subclass MYCN	GBM, MYCN	methylation class family glioblastoma, IDH wildtype	MCF GBM	16
methylation class glioblastoma, IDH wildtype, subclass RTK I	GBM, RTK I	methylation class family glioblastoma, IDH wildtype	MCF GBM	64
methylation class glioblastoma, IDH wildtype, subclass RTK II	GBM, RTK II	methylation class family glioblastoma, IDH wildtype	MCF GBM	143
methylation class glioblastoma, IDH wildtype, subclass RTK III	GBM, RTK III	methylation class family glioblastoma, IDH wildtype	MCF GBM	13
methylation class CNS high grade neuroepithelial tumor with BCOR alteration	HGNET, BCOR			23
methylation class CNS high grade neuroepithelial tumor with MN1 alteration	HGNET, MN1			21
methylation class hemangioblastoma	HMB			25
methylation class infantile hemispheric glioma	IHG			10
methylation class low grade glioma, desmoplastic infantile astrocytoma / ganglioglioma	LGG, DIG/DIA			8
methylation class low grade glioma, dysembryoplastic neuroepithelial tumor	LGG, DNT			44
methylation class low grade glioma, ganglioglioma	LGG, GG			21
methylation class low grade glioma, MYB/MYBL1	LGG, MYB			22
methylation class low grade glioma, subclass midline pilocytic astrocytoma	LGG, PA MID	methylation class family pilocytic astrocytoma	MCF PA	38
methylation class low grade glioma, subclass posterior fossa pilocytic astrocytoma	LGG, PA PF	methylation class family pilocytic astrocytoma	MCF PA	114
methylation class low grade glioma, subclass hemispheric pilocytic astrocytoma and ganglioglioma	LGG, PA/GG ST	methylation class family pilocytic astrocytoma	MCF PA	24
methylation class low grade glioma, rosette forming glioneuronal tumor	LGG, RGNT			9
methylation class low grade glioma, subependymal giant cell astrocytoma	LGG, SEGA			21
methylation class cerebellar liponeurocytoma	LIPN			10
methylation class lymphoma	LYMPHO			13
methylation class medulloblastoma, subclass group 3	MB, G3	methylation class family medulloblastoma group 3 and 4	MCF MB G3G4	77
methylation class medulloblastoma, subclass group 4	MB, G4	methylation class family medulloblastoma group 3 and 4	MCF MB G3G4	138
methylation class medulloblastoma, subclass SHH A (children and adult)	MB, SHH CHL AD	methylation class family medulloblastoma, SHH	MCF MB SHH	84
methylation class medulloblastoma, subclass SHH B (infant)	MB, SHH INF	methylation class family medulloblastoma, SHH	MCF MB SHH	52
methylation class medulloblastoma, WNT	MB, WNT			39
methylation class melanoma	MELAN			12
methylation class melanocytoma	MELCYT			15
methylation class meningioma	MNG			90
methylation class IDH glioma, subclass 1p/19q codeleted oligodendroglioma	O IDH	methylation class family glioma, IDH mutant	MCF IDH GLM	80
methylation class paraganglioma, spinal non-CIMP	PGG, nC			19
methylation class pineoblastoma group A / intracranial retinoblastoma	PIN T, PB A			9
methylation class pineoblastoma group B	PIN T, PB B			22
methylation class pineal parenchymal tumor	PIN T, PPT			19
methylation class pituitary adenoma, ACTH	PITAD, ACTH			18
methylation class pituitary adenoma, FSH/LH	PITAD, FSH LH			21
methylation class pituitary adenoma, prolactin	PITAD, PRL			8
methylation class pituitary adenoma, STH densely granulated, group A	PITAD, STH DNS A			9
methylation class pituitary adenoma, STH densely granulated, group B	PITAD, STH DNS B			12
methylation class pituitary adenoma, STH sparsely granulated	PITAD, STH SPA			17
methylation class pituitary adenoma, TSH	PITAD, TSH			10
methylation class pituicytoma / granular cell tumor / spindle cell oncocytoma	PITUI, SCO, GCT			29
methylation class plasmacytoma	PLASMA			8
methylation class plexus tumor, subclass adult	PLEX, AD	methylation class family plexus tumor	MCF PLEX T	22
methylation class plexus tumor, subclass paediatric A	PLEX, PED A	methylation class family plexus tumor	MCF PLEX T	15
methylation class plexus tumor, subclass paediatric B	PLEX, PED B	methylation class family plexus tumor	MCF PLEX T	46
methylation class papillary tumor of the pineal region group A	PTPR, A			8
methylation class papillary tumor of the pineal region group B	PTPR, B			22
methylation class (anaplastic) pleomorphic xanthoastrocytoma	PXA			44
methylation class retinoblastoma	RETB			19
methylation class schwannoma	SCHW			23
methylation class melanotic schwannoma	SCHW, MEL			8
methylation class solitary fibrous tumor / hemangiopericytoma	SFT HMPC			16
methylation class subependymoma, posterior fossa	SUBEPN, PF			37
methylation class subependymoma, spinal	SUBEPN, SPINE			9
methylation class subependymoma, supratentorial	SUBEPN, ST			19

Supplementary Table 1 | Overview of tumor classes and methylation class families