SI Appendix

Inter-sample similarity/distance measures

The gene sets used in this paper are given in Supplementary Tables and described in Supplementary Materials. The distance measures used in this paper were:

1) **Similarity by SNAs** - inner (dot) product of per-gene indicator vectors (1=impacted by non-synonymous SNA, 0=not impacted).

2) **Similarity by CNAs** - inner (dot) product of per-gene thresholded GISTIC scores (-2=homozygous loss, -1=single copy loss, +1=single copy gain, +2=multi-copy amplification).

3) **Joint similarity by SNAs and CNAs** - sum of the *normalized* genome-wide SNA and CNA similarity scores such that the sum of all SNA scores equals the sum of all CNA scores, equals 1.

4) **Similarity by expression and CNA combined** - inner product of per-gene expression Z-scores & thresholded GISITC scores (as defined in (2) above).

5) **G-CIMP DNA methylation probes** – In 2010, a TCGA publication by Noushmehr et al (Cancer Cell. 2010 May 18;17(5):510-22) identified 1503 Illumina Infinium 27K array classifier probes that reliably identified hyper-methylated (G-CIMP) samples. We used 1444 of these probes, which had an exact match to the probe names in our Infinium 450K DNA methylation array data.

6) **Correlated methylation and expression** – We used the Bioconductor package "FDb.InfiniumMethylation.hg19" (http://bioconductor.org/packages/release/data/annotation/html/FDb.InfiniumMethylation.hg19.html) to identify CpG island probes within 2Kbp upstream of the transcription start sites of genes. Among these probes, we selected the top 8000 probes with the greatest Median Absolute Deviation. Probe beta values were averaged on a per gene basis and then used to calculate Spearman's rank correlation as a measure of sample similarity.

7) **Expression of stemness marker genes** – We used Manhattan distance (p=1 Minkowki distance).

8) **Expression of metabolic genes** – Sample similarities for the selected genes were visualized using Principle Component Analysis (PCA).

9) **Immune gene expression** – We performed Single Sample Gene Set Enrichment Analysis (ssGSEA, Bioconductor package http://www.bioconductor.org/packages/release/bioc/html/GSVA.html) using the 1910 gene sets from C7 collection of the MSigDB database (http://www.broadinstitute.org/gsea/msigdb/collections.jsp). The Manhattan distances of the ssGSEA score matrices were used as input to MDS.

Calculation of approximate p-values for detected sample clusters

We performed permutation-based tests to assess if visually identified sample clusters are significantly clustered compared to the rest of the samples. Briefly, for each identified cluster, we compute all pairwise distances within the cluster. Additionally, we compute all pairwise distances for all genes outside of the cluster. Then, we evaluate if these two sets of distances are statistically different from each other by computing a Z-score. Since pairwise distances are correlated, such Z-scores do not have any known statistical properties which can be used to derive a p-value. Hence, 10,000 permutations of the plot data are created under the null hypothesis, and are used to create a distribution of Z-scores.

The observed distance is then compared to null distribution to produce a p-value. For brevity, in Supplementary Materials, we refer to this measure of cluster significance as the “Within Clusters” comparison. Note that the computed p-value is an approximation and quantifies the strength of the visually observed clusters, since creating “clustering patterns by visualization” under the null is nearly impossible.

One potential weakness associated with the above measure is that it may fail to flag a user-selected cluster as significant if the samples outside the selected cluster are also tightly bunched together, resulting in comparable within-cluster distances for the selected cluster and control samples. To address this issue, we introduce a complementary measure: the difference between within-cluster distances for the selected cluster and the distances between every member of the selected cluster and every sample outside of the cluster. For brevity, in Supplementary Materials, we refer to this measure of cluster significance as the “Between Clusters” comparison. Again, we use a permutation strategy to compute an observed Z-score and a Z-score distribution for the null hypothesis.

For all of the eight clusters delineated in Figure 3, the observed Z-scores are well outside of their null distributions, and hence their p-values are below the low bound of the permutation exercise, i.e., 1/10000.

Differential expression/ methylation analysis

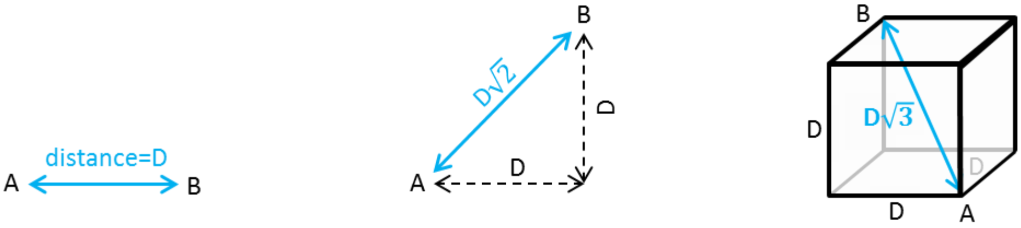
Differential expression analysis was performed with the Bioconductor package ‘limma’ (http://bioconductor.org/packages/release/bioc/html/limma.html) using the batch-effect corrected RNA-seq data. Differential methylation analysis was performed with the Bioconducto package ‘DMRcate’ (http://www.bioconductor.org/packages/release/bioc/html/DMRcate.html) using batch-effect corrected probe values. We found 1,808 differentially expressed genes and 620 differentially methylated regions containing 11,127 probes across 658 genes.

Supplementary Materials

Motivation for sample similarity plots

The millions of molecular data points (per patient) that are generated with high-throughput sequencing, array platforms, and proteomics, make brute-force, statistical, machine-learning, and other commonly-used 'unbiased' methods for discovering patient groups extremely inefficient. This limitation is fundamental. With only two dimensions (measurements), unsupervised clustering algorithms can identify distinct patient groups fairly reliably. But the efficiency of unsupervised clustering drops precipitously as the number of measurements grows, for the following reasons:

(1) Improvements in resolution when we measure multiple genes per patient only grow as the square root of the number of measurements. For example, measuring the expression level of 10,000 genes (instead of one) can increase the straight-line distance among samples in a scatter plot at most 100-fold. The figures below visually demonstrate this property for 1, 2 and 3 dimensions (measured values).



A remarkable consequence of this effect is that the distance difference between the nearest and farthest neighbors in a scatter plot shrinks as the number of dimensions (biomarkers) increases, thus making unambiguous assignment of points to clusters difficult and error prone.

Expert-guided Visual Exploration (EVE) addresses this challenge by relying on human visual pattern recognition capabilities, both when detecting sample clusters in 2 or 3D plots, and in selecting parameters for automated sample clustering.

(2) Clustering relies on calculating similarity distances among samples. The efficiency of computing distances falls with the number of measurements made. Measuring sample distances across millions of dimensions – as is the case for genomic medicine – becomes prohibitive for >O(1000) patients.

(3) As the number of measurements (N) increases, a larger fraction of the possible biomarker values for any given sample will be "tucked away" in the corners of an N-dimensional scatter plot. In other words, high-dimensional data is inherently fragmented. This fragmentation makes it difficult to identify subsets of biomarker groups that cluster patients in biologically meaningful ways.

The above effect is illustrated in the figures below. With only two measurements (left panel), the area of the largest circle inside the 2D box of all possible biomarker values covers more than 78% of the total area of the box. Thus, if all measured values were distributed uniformly across samples, more than 78% of the patients would fall inside this 'central' circle. With three biomarkers, the area of the largest sphere contained in the 3D cube of all possible biomarker combinations is only ~52% of the volume of the cube (right panel). For 10 biomarkers, this fraction becomes less than 2.5% of the measurement space. Thus when millions of molecular species are measured, there are virtually no 'average' patients; everybody is 'special' in some way (dimension), which makes clustering difficult.

EVE circumvents this challenge by relying on the domain-knowledge of expert users to nominate candidate gene/probes sets for delineation of specific sample subtypes.

(4) The common approach of combining measurements of different biomolecular entities (e.g. gene mutations and DNA-methylation probe β-values) into a single measure of similarity/distance among samples is problematic because in biology we do not currently have 'laws' governing the relationships among different measurable quantities (cf. volume and pressure, or energy, mass, and displacement in physics). To overcome this issue, clustering of biomarkers is often performed for one data type at a time, but then integration across multiple measurements is performed in arbitrary ways.

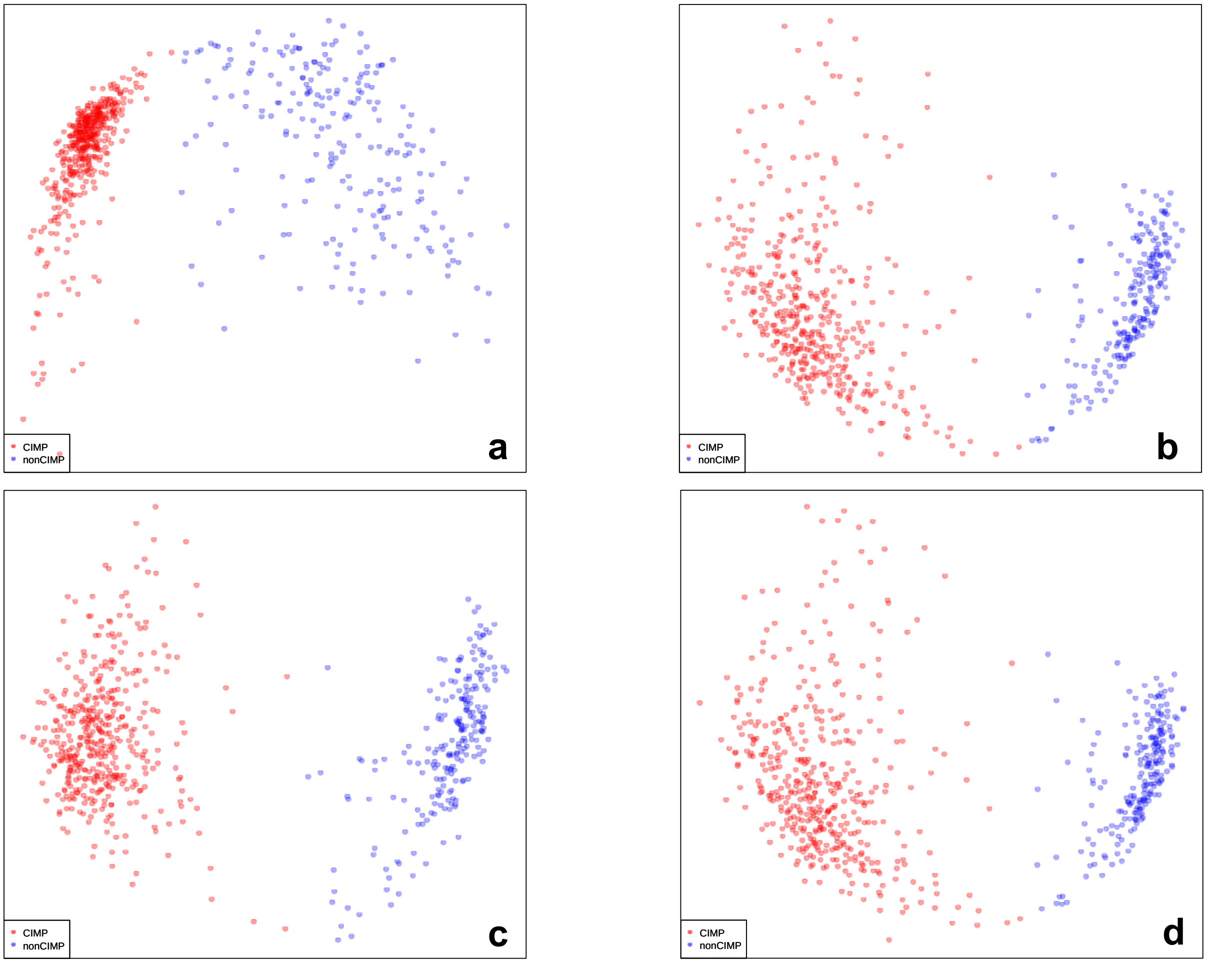
Because EVE allows users to select from among multiple methods, measures, and parameters, users can interactively explore the relative merits of different approaches to combining data types.

(5) Automated (statistical/algorithmic) clustering methods come in many flavors, each with its own limitations, such as sensitivity to outliers, inability to detect concave clusters, or using approximations that may not hold in some cases.

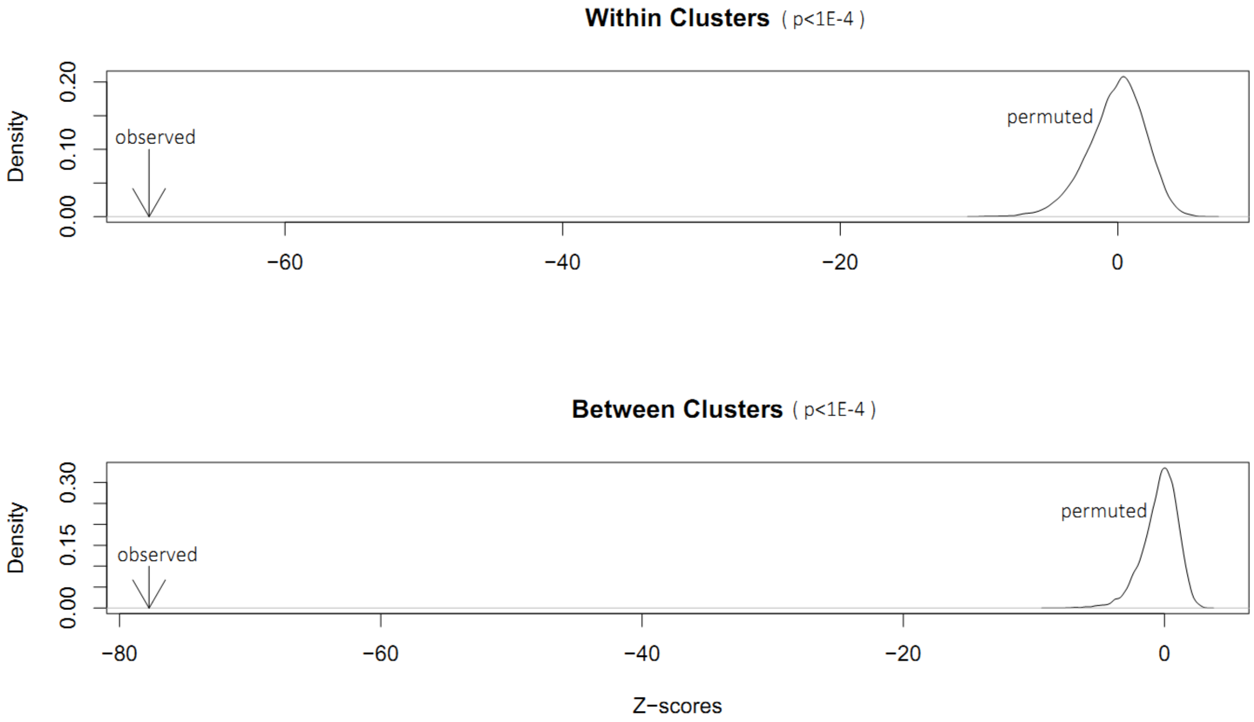
EVE provides a flexible framework in which users can explore the effect of changes in cluster methods, distance measures, etc.

Supplementary Figures

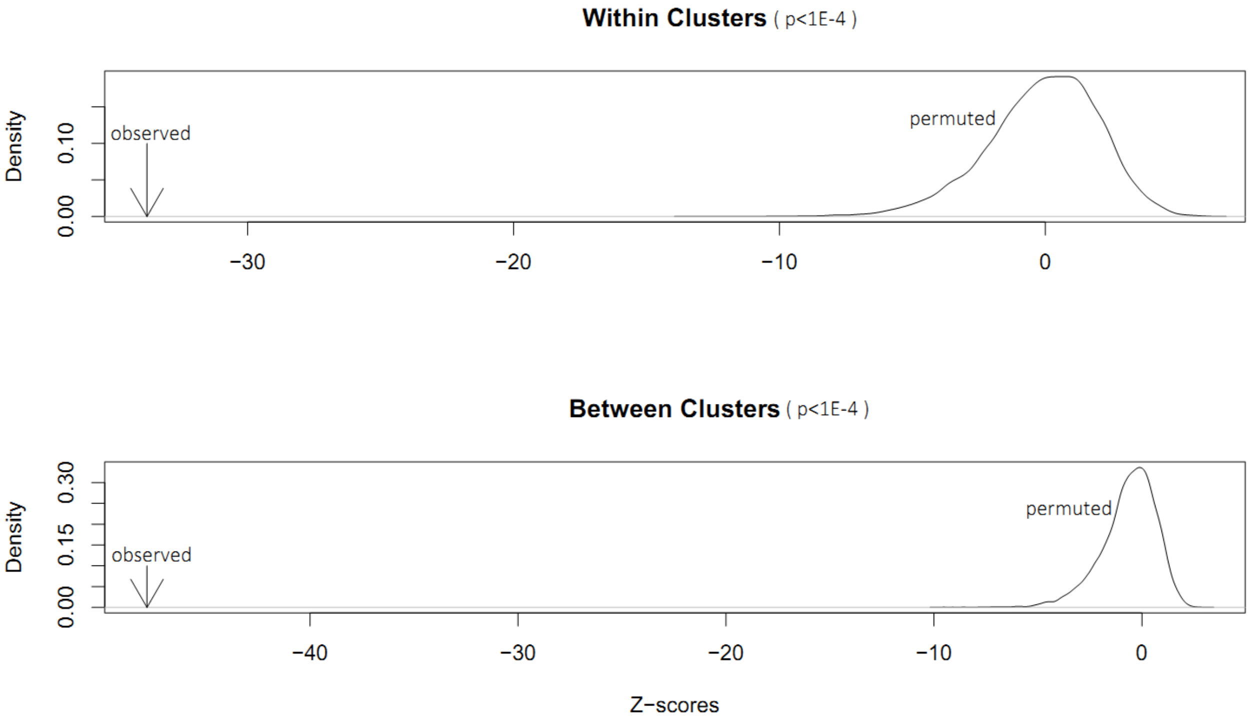
Supplementary Figure 1. Multiple subsets of genome-wide methylation probes unambiguously divide TCGA glioma samples into CIMP and non-CIMP groups. (a) Using the published 1503 CIMP classifier probes, the samples divide sharply into two groups: CIMP (red) and non-CIMP (blue). (b) Sample similarity using all probes located in gene bodies. Samples are colored according to panel (a). (c) Sample similarity using all probes located in promoters (defined as within 1000bp of transcription start sites). Samples are colored according to panel (a). (d) Sample similarity using only the 80000 most variable probes across all samples. Samples are colored according to panel (a).



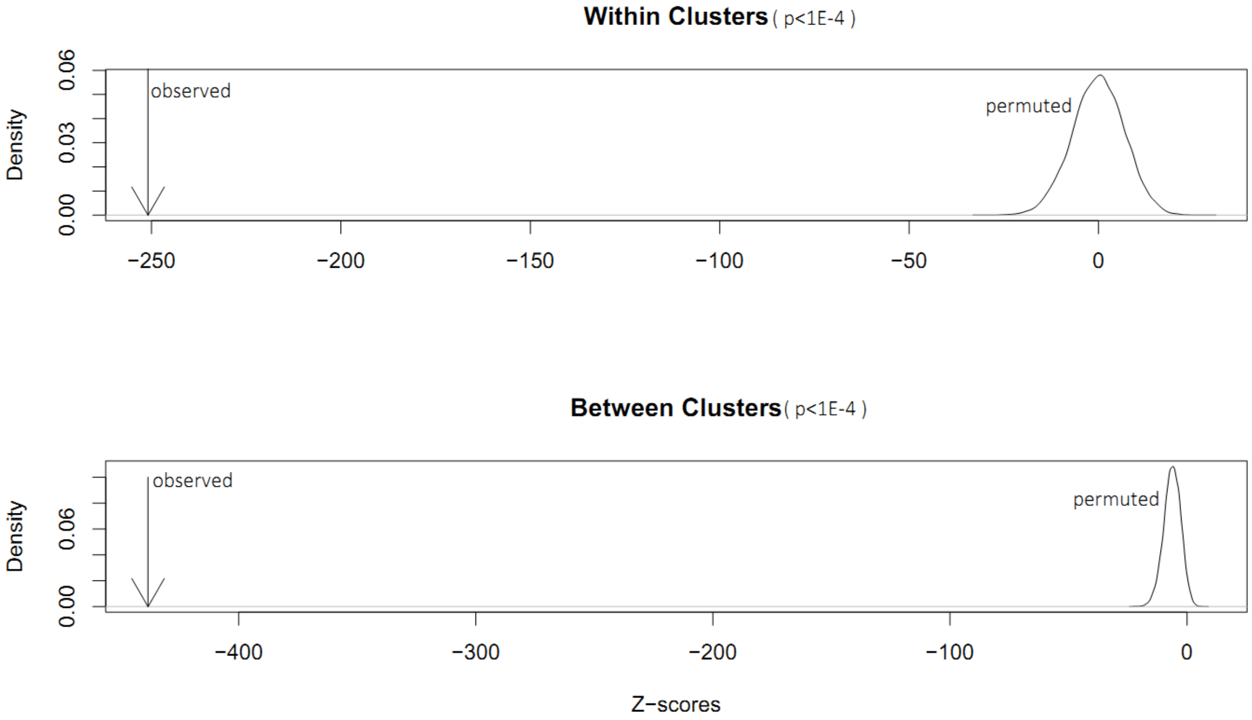
Supplementary Figure 2(a). Approximately computed P-values for cluster 1 of Figure 3.



Supplementary Figure 2(b). Approximately computed P-values for cluster 2 of Figure 3.



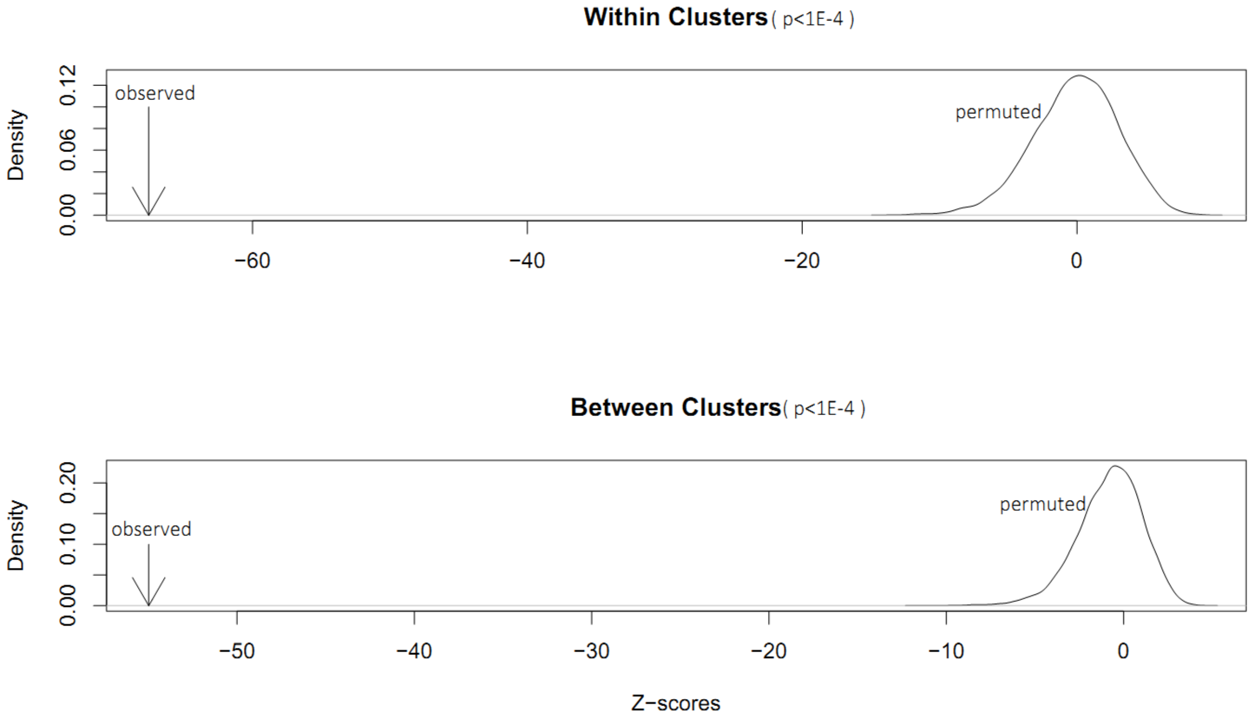
Supplementary Figure 2(c). Approximately computed P-values for cluster 3 of Figure 3.



Within Clusters (p<1E-4)

Between Clusters (p<1E-4)

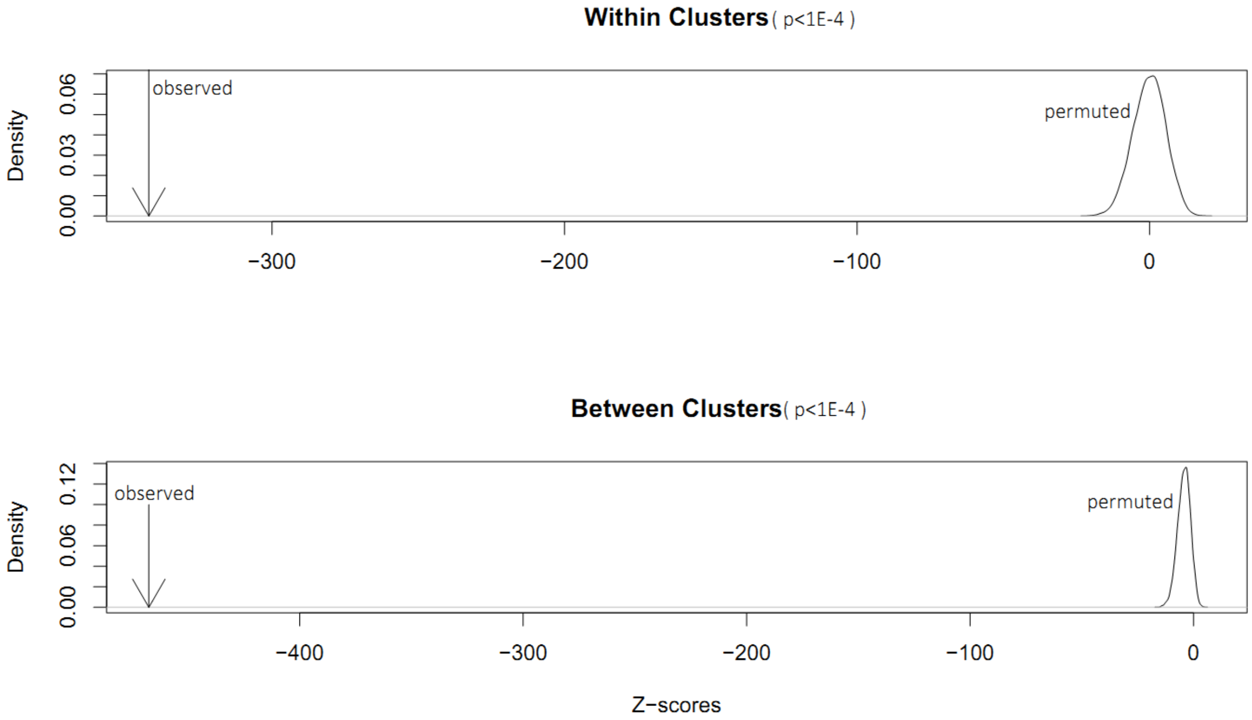
Supplementary Figure 2(d). Approximately computed P-values for cluster 4 of Figure 3.



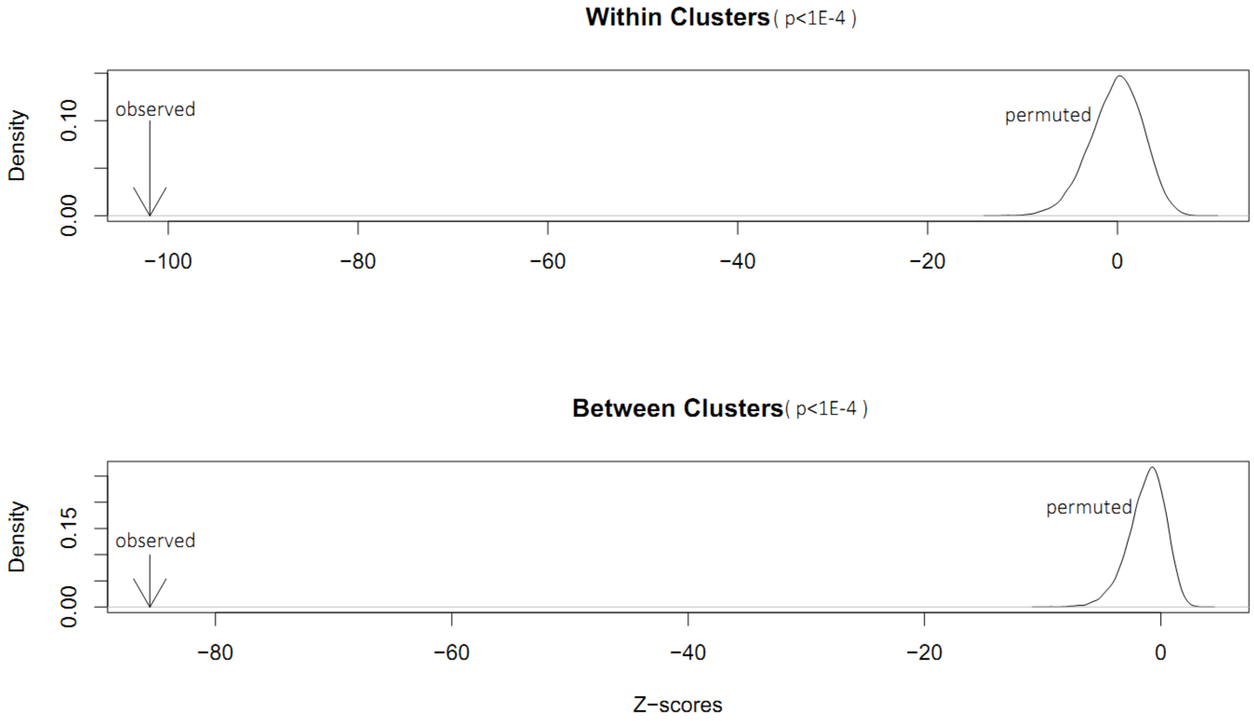
Within Clusters (p<1E-4)

Between Clusters (p<1E-4)

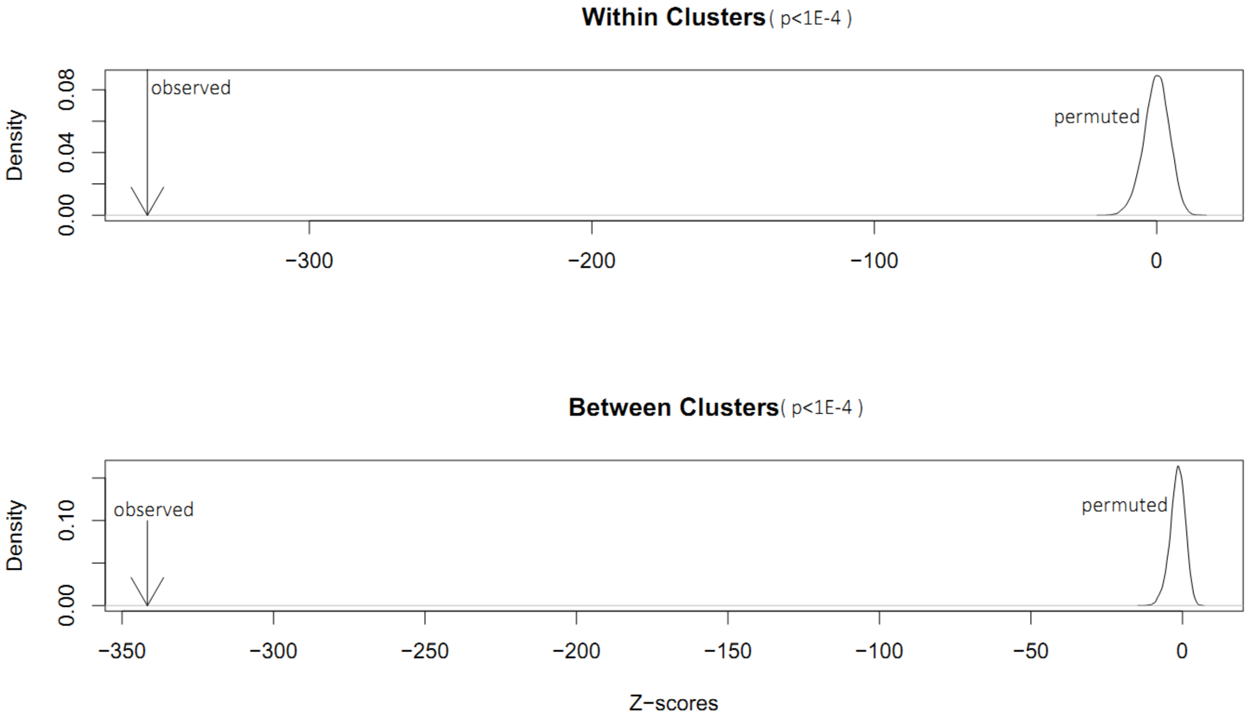
Supplementary Figure 2(e). Approximately computed P-values for cluster 5 of Figure 3.



Supplementary Figure 2(f). Approximately computed P-values for cluster 6 of Figure 3.



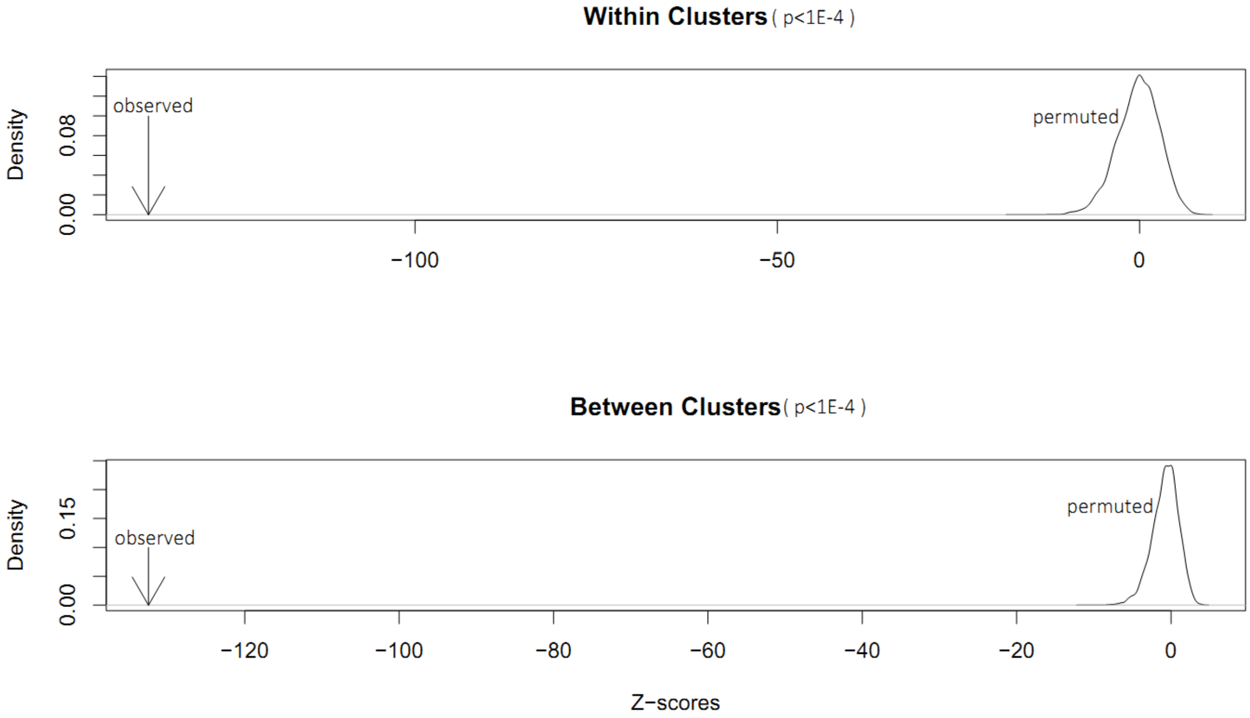
Supplementary Figure 2(g). Approximately computed P-values for cluster 7 of Figure 3.



Within Clusters (p<1E-4)

Between Clusters (p<1E-4)

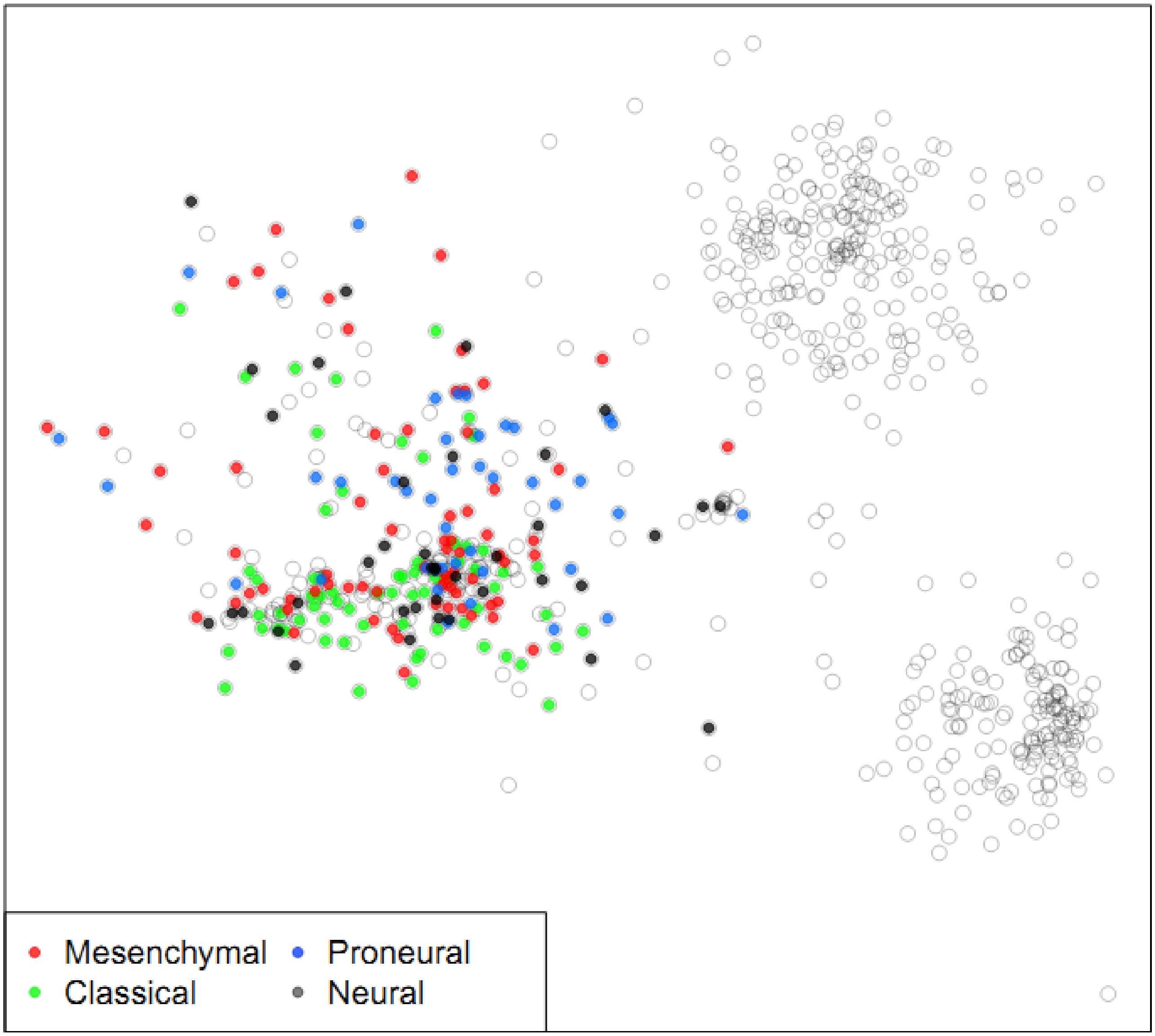
Supplementary Figure 2(h). Approximately computed P-values for cluster 8 of Figure 3.



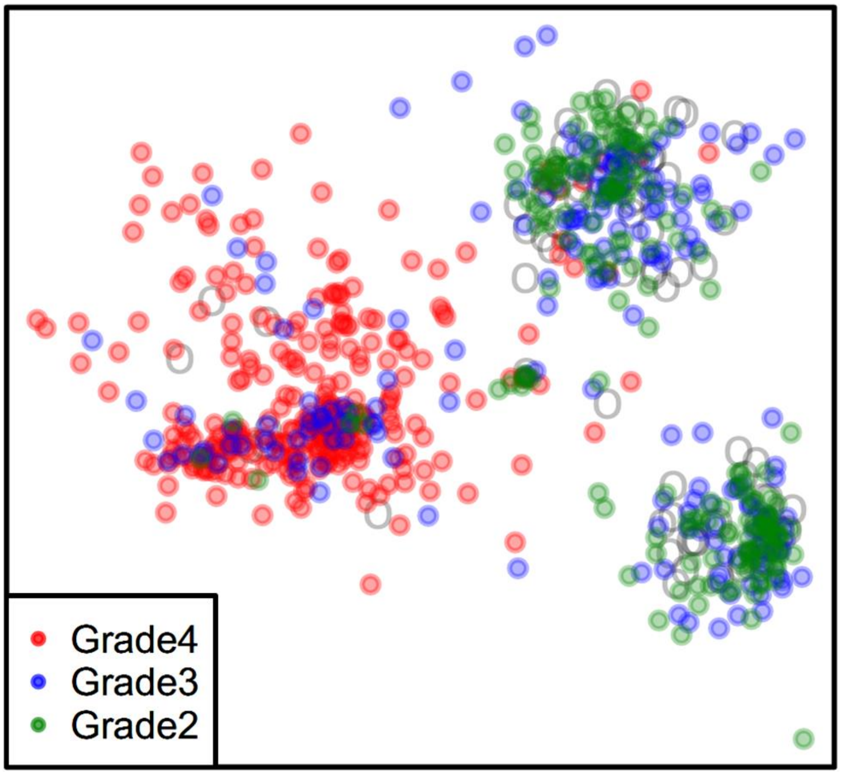
Within Clusters (p<1E-4)

Between Clusters (p<1E-4)

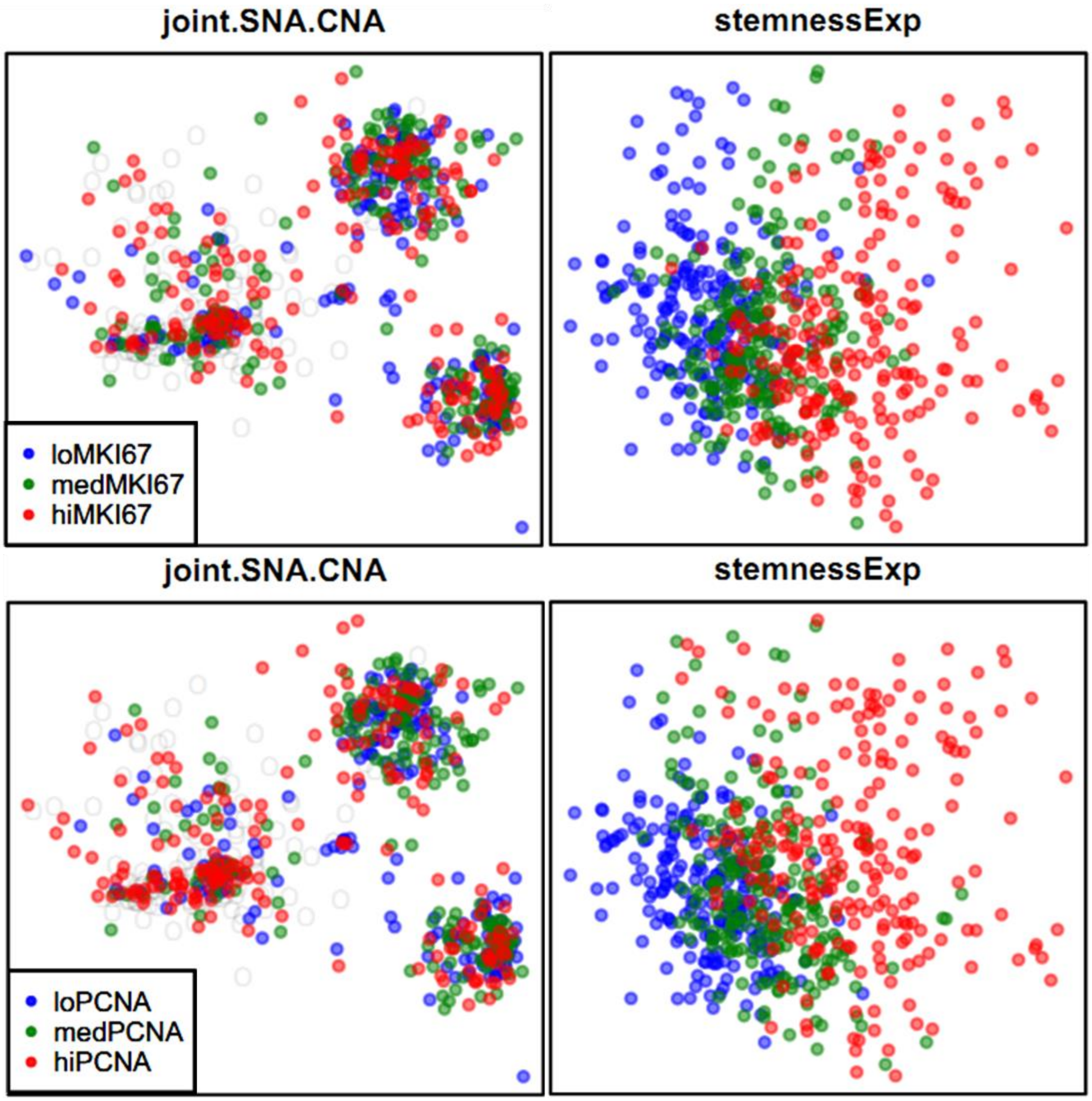
Supplementary Figure 3, The TCGA GBM expression subtypes are not regionally distributed in our genomic sample-similarity plot. The units of the horizontal and vertical axes are arbitrary.



Supplementary Figure 4. The distribution of tumor grades in the SNA/CNA sample similarity plot. The units of the horizontal and vertical axes are arbitrary.

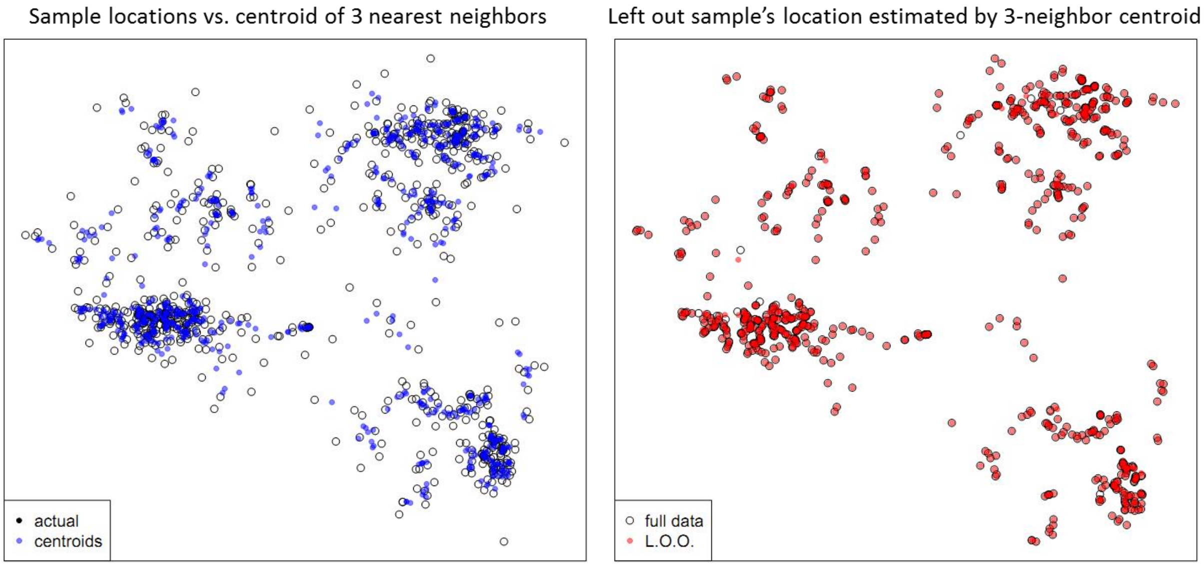


Supplementary Figure 5. The proliferation markers MKI67 and PCNA are not regionally distributed in the SNA.CNA sample similarity plot, in contrast to similarity by stemness marker gene expression. Expression levels were divided into 3 quatiles and colored differentially for ease of visualization.



Supplementary Figure 6. Sample placement on the sample similarity plot is robust and offers an intuitive approach to classifying new patients given a large number of 'training set' samples (e.g. TCGA data). To explore the ability of EVE to correctly classify new samples (generalization), we performed leave-one-out (LOO) validation. In each LOO iteration, one sample is left out of the 'training set' used to generate the SNA.CNA sample similarity plot. This sample is then super-imposed onto the plot by triangulation, and its position is compared to its position in the equivalent plots without any sample removal.

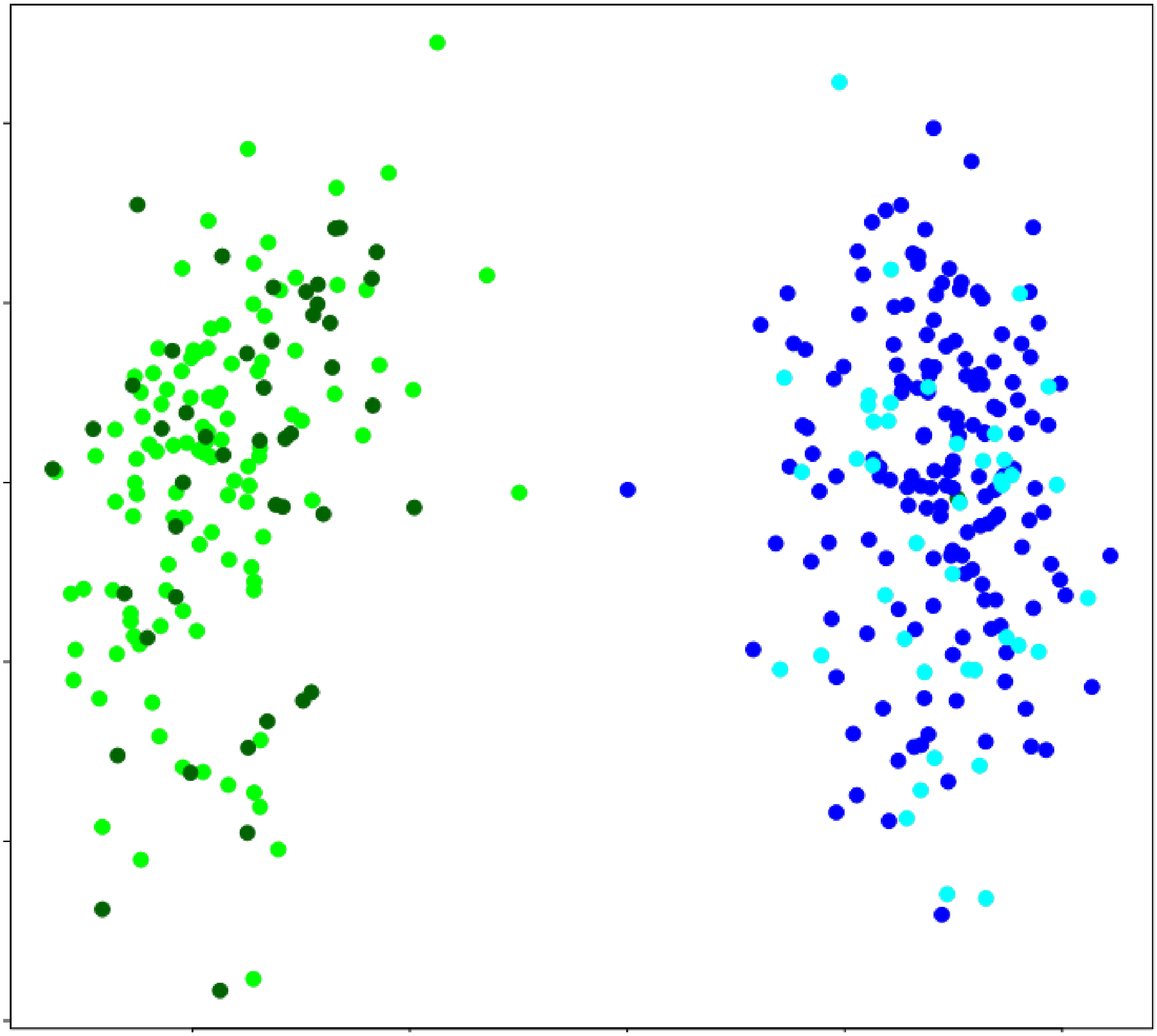
MDS projection of high dimensional data preserves inter-sample distances, not the locations of the samples in different plots. To accommodate this characteristic when comparing before and after sample coordinates, in the panel below left the coordinates of each plot point are calculated as the centroid of the three nearest neighbors in full dimensional space. We first estimated the location of each sample by the centroid of its three nearest neighbors using the full data. The result is shown in blue. We then removed one sample at a time from the plot, re-calculated the MDS projection of the resulting inter-sample distances, and then re-estimated the location of the missing sample from its previously marked three nearest neighbors. The resulting sample centroid locations are plotted in red (empty circles mark sample locations given full data). Overall, >96% of samples have identical actual and predicted nearest neighbors across all LOO runs. Even including outlier samples (for which neighbor-centroid calculations are inappropriate), the LOO and actual nearest neighbors are the same for >96% of all samples.



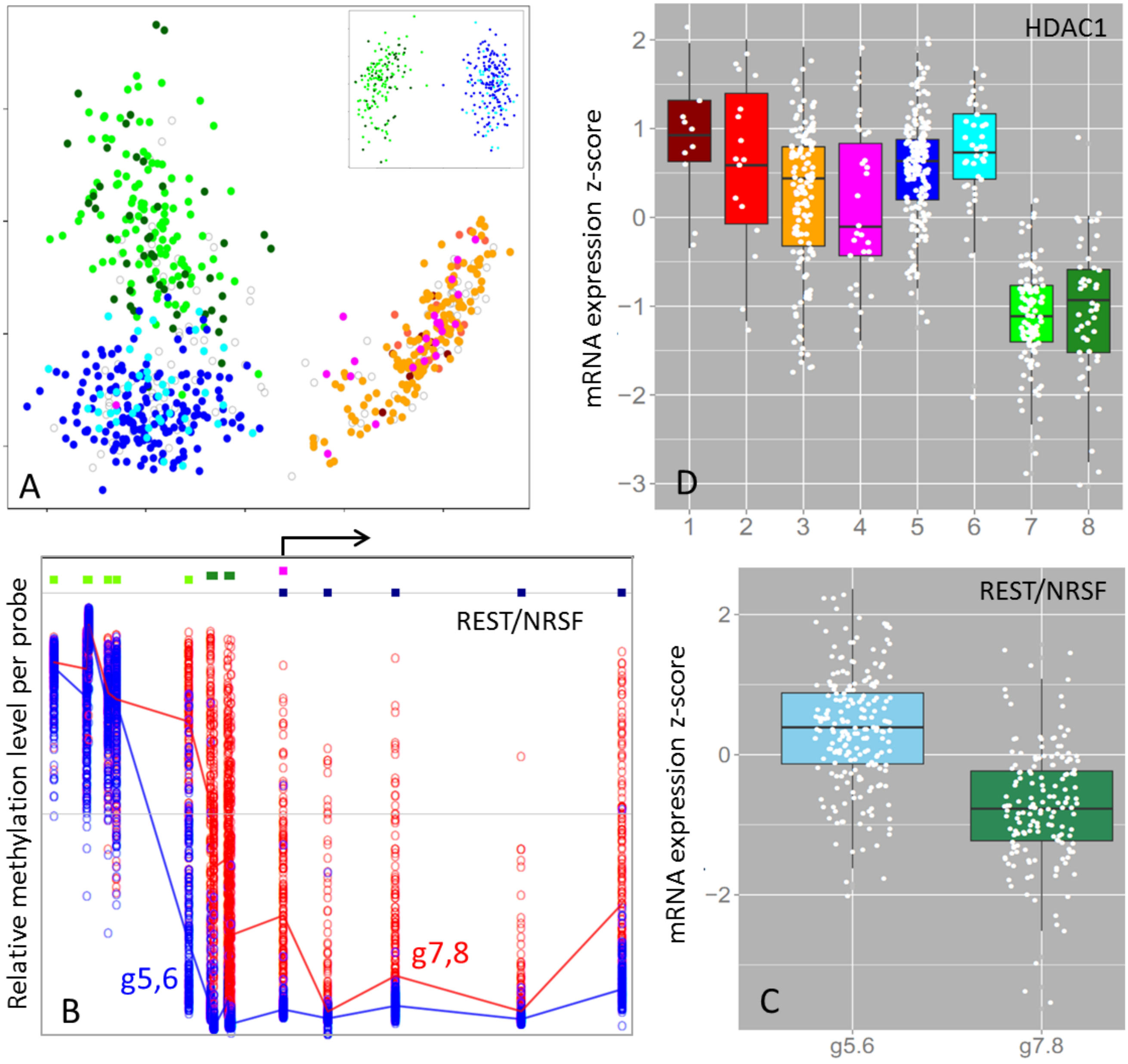
Sample locations vs. centroid of 3 nearest neighbors

Left out sample's location estimated by 3-neighbor centroid

Supplementary Figure 7. 1000 DNA methylation probes are sufficient to perfectly distinguish the two subtypes of CIMP LGGs. Colors are the same as in Figure 2f.



Supplementary Figure 8. The three major clusters detected by genomic data have distinct methylation expression profiles. (A) Methylation levels of 4,500 probes re-capitulate the three genomic clusters. Inset: 1000 probes are sufficient to segregate the Astro and Oligo CIMP LGG clusters. (B) REST/NRSF methylation levels are lower in the Astro cluster compared to the Oligo cluster. (C) Consistent with (B), REST/NRSF mRNA expression levels are higher in the Astro cluster compared to the Oligo cluster. (D) Consistent with (B) and (C), the NRSF co-repressor HDAC1 is expressed at significantly lower levels in the Oligo cluster (Benjamini and Hochberg FDR adjusted p-value = 0.03).



Supplementary Tables

Supplementary Table 1. A set of 396 stemness marker genes were constructed by combining gene sets from Wong et al. Cell Stem Cell. 2008 Apr 10; 2(4): 333–344, and

http://www.sabiosciences.com/rt_pcr_product/HTML/PAHS-404A.html.

ABCB7
ACAD8
ACHE
ADH5
ADORA1
ADORA2A
ADSL
AK3L1
ALDH7A1
ALDOC
ALK
AMOTL2
ANP32E
APBB1
APEX1
APOE
ARNT2
ASCL1
ATP5J
ATP5O
AURKA
AURKB
BAI1
BANF1
BAX
BCAT1
BDNF
BIRC5
BLM
BMP2
BMP4
BMP8B
BTF3
BUB1
BUB1B
BUB3

C11orf48
C2orf47
CBX3
CCNA2
CCNB2
CCNC
CCND1
CCND2
CCNF
CCT5
CDC20
CDC34
CDC6
CDC7
CDCA3
CDCA5
CDCA7
CDCA8
CDK4
CDK5R1
CDK5RAP1
CDK5RAP2
CDK5RAP3
CDKN1C
CDKN3
CHAF1A
CHEK1
CHEK2
CHRM2
CKAP2
CKS1B
CKS2
CLPP
COQ3
COX4NB
COX5B
CRABP2
CSE1L
CSRP2
CTNNA1
CTSC
CXCL1
CYCS

DAP3
DARS2
DBF4
DDX18
DEK
DHX9
DLAT
DLG4
DLL1
DNMT1
DPP3
DRD1
DRD2
DTL
DTYMK
DVL3
E2F3
EBNA1BP2
ECHS1
EEF1E1
EEF2
EFNB1
EGF
EIF2S2
EIF2S3
EIF4A1
EIF4B
EIF4EBP1
ELOVL6
ENO1
EP300
ERBB2
ERCC6L
ERP29
ETFA
EXO1
EXOSC7
FAM136A
FAM60A
FARSA
FBL
FDPS
FEZ1

FGF13
FGF2
FGFR1
FH
FLNA
G3BP1
GARS
GART
GDNF
GEMIN6
GJA1
GLDC
GLO1
GMNN
GNA14
GNAO1
GNL3
GNPDA1
GPI
GRIN1
GSPT2
GTSE1
HADH
HAT1
HDAC1
HDAC4
HDAC7
HELLS
HES1
HEY1
HEY2
HEYL
HMGB2
HN1
HNRNPA1
HSPA14
HSPA9
HSPE1
IARS
INHBA
IPO9
KIF11
KIF20A

KIF22
KIF23
KIF4A
KPNA2
KPNA6
KRAS
LMNB1
LSM10
LSM2
LSM4
LSM5
LYPLA1
MAD2L1
MAPK13
MCM2
MCM3
MCM4
MCM5
MCM7
MDK
MEF2C
MID1IP1
MKI67IP
MLL
MRPL11
MRPL12
MRPL13
MRPL15
MRPL16
MRPL37
MRPL39
MRPL4
MRPS17
MRPS18B
MRPS2
MRPS28
MRPS30
MRPS36
MRTO4
MSH2
MTF2
MTHFD2
MYBL2

MYC
NAP1L1
NASP
NCAPD2
NCAPH
NCBP2
NCL
NCOA6
NDC80
NDN
NDP
NDUFA11
NDUFA9
NDUFAB1
NDUFB10
NDUFB7
NDUFB8
NDUFS2
NEK2
NEUROD1
NIP7
NIPSNAP1
NLN
NME2
NME4
NOG
NONO
NOTCH2
NPTX1
NRCAM
NRG1
NRP1
NRP2
NT5DC2
NTHL1
NTN1
NUDCD2
NUP107
NUSAP1
ODZ1
ORC1L
OTX2
PA2G4

PABPC1
PAFAH1B1
PARD3
PARD6B
PARP1
PAX3
PAX5
PAX6
PCNA
PDCD2
PDHA1
PDIA4
PDPN
PHB
PHC1
PHF5A
PIPOX
PLK1
PLK4
POLD1
POLE2
POLR2F
POLR3K
POP7
POU3F3
POU4F1
PPM1G
PPP4C
PRDX1
PRIM1
PRIM2
PRMT1
PRMT3
PROM1
PRPS1
PSMA5
PSMA7
PSMB5
PSMB6
PSMD14
PSME3
PTN
PUS1

RAB34
RAC1
RACGAP1
RAD18
RAD23B
RCC1
RCC2
RCN2
RFC3
RNPS1
ROBO1
RPA2
RPA3
RPL10A
RPL13
RPL22
RPL27A
RPP40
RPS12
RPS16
RPS19
RPS23
RPS27
RPS3
RPS5
RPS8
RPSA
RRM1
RRM2
RTN4
RUVBL1
RUVBL2
S100A6
S100B
SARS
SDHC
SDHD
SEMA4D
SEPHS2
SERPINH1
SET
SHH
SIP1

SLC16A1
SLC25A5
SLC2A1
SLIT2
SMC2
SMC4
SNRPA
SNRPA1
SNRPD1
SNX5
SOX2
SOX8
SPAG5
SQLE
SS18
SSB
STAT3
STIP1
STOML2
SUMO1
TCF19
TCF7L1
TCOF1
TEAD2
TERF1
TGIF1
TGIF2
THOC3
TIMM13
TIMM44
TIMM8A
TIMM8B
TMEFF1
TNR
TOP2A
TP53
TRIP13
TRIP6
TTK
U2AF1
UBE2G1
UBE2V2
UGDH

UQCRH
UTP18
VBP1
VEGFA
VRK1
WBP11
WDHD1
WDR77
WEE1
XPO1
XRCC5
YAP1
YWHAH
YY1
ZIC3
ZNF22

Supplementary Table 2. Human metabolic genes were downloaded from the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/dbget-bin/www_bget?pathway+hsa01100. For our glioma analyses, to avoid confounding effects, we removed from this list genes associated with specific GBM expression subtypes (1157 genes remained).

A4GALT
AADAT
AANAT
AASS
ABAT
ABO
ACAA1
ACAA2
ACACA
ACACB
ACAD8
ACADL
ACADM
ACADS
ACADSB
ACADVL
ACAT1
ACAT2
ACER1
ACER2
ACLY
ACMSD
ACO1
ACO2
ACOT1
ACOT2
ACOT4
ACOT8
ACOX1
ACOX2
ACOX3
ACSBG1
ACSBG2
ACSL1
ACSL3
ACSL4
ACSL5
ACSL6

ACSM1
ACSM2A
ACSM2B
ACSM3
ACSM4
ACSM5
ACSS1
ACSS2
ACSS3
ACY1
ADA
ADAM29
adenylate
ADH1A
ADH1B
ADH1C
ADH4
ADH5
ADH6
ADH7
ADI1
ADK
ADO
ADPGK
ADSL
ADSS
ADSSL1
AFMID
AGK
AGL
AGMAT
AGPAT1
AGPAT2
AGPAT3
AGPAT4
AGPAT5
AGPAT6
AGPAT9
AGPS
AGXT
AGXT2
AHCY
AHCYL1

AHCYL2
AK1
AK2
AK4
AK5
AK6
AK7
AK8
AK9
AKR1A1
AKR1B1
AKR1B10
AKR1C3
AKR1C4
AKR1D1
ALAD
ALAS1
ALAS2
ALDH18A1
ALDH1A1
ALDH1A2
ALDH1A3
ALDH1B1
ALDH2
ALDH3A1
ALDH3A2
ALDH3B1
ALDH3B2
ALDH4A1
ALDH5A1
ALDH6A1
ALDH7A1
ALDH9A1
ALDOA
ALDOB
ALDOC
ALG1
ALG10
ALG10B
ALG11
ALG12
ALG13
ALG14

ALG2
ALG3
ALG5
ALG6
ALG8
ALG9
ALLC
ALOX12
ALOX12B
ALOX15
ALOX15B
ALOX5
ALPI
ALPL
ALPP
ALPPL2
AMACR
AMD1
AMDHD1
AMPD1
AMPD2
AMPD3
AMT
AMY1A
AMY1B
AMY1C
AMY2A
AMY2B
ANPEP
AOC2
AOC3
AOX1
APIP
APRT
ARG1
ARG2
ARSB
ASAH1
ASAH2
ASL
ASMT
ASNS
ASPA

ASS1
ATIC
ATP5A1
ATP5B
ATP5C1
ATP5D
ATP5E
ATP5F1
ATP5G1
ATP5G2
ATP5G3
ATP5H
ATP5I
ATP5J
ATP5J2
ATP5L
ATP5O
ATP6
ATP6AP1
ATP6V0A1
ATP6V0A2
ATP6V0A4
ATP6V0B
ATP6V0C
ATP6V0D1
ATP6V0D2
ATP6V0E1
ATP6V0E2
ATP6V1A
ATP6V1B1
ATP6V1B2
ATP6V1C1
ATP6V1C2
ATP6V1D
ATP6V1E1
ATP6V1E2
ATP6V1F
ATP6V1G1
ATP6V1G2
ATP6V1G3
ATP6V1H
ATP8
AUH

AZIN2
B3GALNT1
B3GALT1
B3GALT2
B3GALT4
B3GALT5
B3GALT6
B3GAT3
B3GNT2
B3GNT3
B3GNT4
B3GNT5
B3GNT6
B4GALNT1
B4GALT1
B4GALT2
B4GALT3
B4GALT4
B4GALT6
B4GALT7
B4GAT1
BAAT
BCAT1
BCAT2
BCKDHA
BCKDHB
BCO1
BDH1
BDH2
BHMT
BPGM
BPNT1
BST1
BTD
C1GALT1
C1GALT1C1
CAD
CBR1
CBR3
CBS
CCBL1
CCBL2
CD38

CDA
CDIPT
CDO1
CDS1
CDS2
CEL
CEPT1
CERS1
CERS2
CERS3
CERS4
CERS5
CERS6
CES1
CHDH
CHKA
CHKB
CHPF
CHPF2
CHPT1
CHSY1
CHSY3
CKB
CKM
CKMT1A
CKMT1B
CKMT2
CMAS
CMBL
CMPK1
CMPK2
CNDP1
CNDP2
COASY
COMT
COQ2
COQ3
COQ5
COQ6
COQ7
COX1
COX10
COX11

COX15
COX17
COX2
COX3
COX4I1
COX4I2
COX5A
COX5B
COX6A1
COX6A2
COX6B1
COX6B2
COX6C
COX7B
COX7B2
COX7C
COX8A
COX8C
CPOX
CPS1
CRLS1
CRYL1
CS
CSAD
CSGALNACT1
CSGALNACT2
CTH
CTPS1
CTPS2
CYC1
CYCS
CYP11A1
CYP11B1
CYP11B2
CYP17A1
CYP19A1
CYP1A1
CYP1A2
CYP21A2
CYP24A1
CYP26A1
CYP26B1
CYP26C1

CYP27A1
CYP27B1
CYP2A6
CYP2B6
CYP2C18
CYP2C19
CYP2C8
CYP2C9
CYP2E1
CYP2J2
CYP2R1
CYP2S1
CYP2U1
CYP3A4
CYP3A5
CYP3A7
CYP4A11
CYP4F2
CYP4F3
CYP4F8
CYP51A1
CYP7A1
CYP8B1
CYTB
DAD1
DAK
DAO
DBH
DBT
DCK
DCT
DCTD
DCTPP1
DCXR
DDC
DDOST
DEGS1
DEGS2
DGAT1
DGAT2
DGKA
DGKB
DGKD

DGKE
DGKG
DGKH
DGKI
DGKQ
DGKZ
DGUOK
DHCR24
DHCR7
DHFR
DHFRL1
DHODH
DHRS3
DHRS4
DHRS4L1
DHRS4L2
DHRS9
DLAT
DLD
DLST
DMGDH
DNMT1
DNMT3A
DNMT3B
DOLK
DPAGT1
DPM1
DPM2
DPM3
DPYD
DPYS
DSE
DTYMK
DUT
EARS2
EBP
ECHS1
EHHADH
ENO1
ENO2
ENO3
ENOPH1
ENPP1

ENPP3
ENPP7
EPHX2
EPRS
EPT1
ETNK1
ETNK2
ETNPPL
EXT1
EXT2
EXTL1
EXTL2
EXTL3
FAH
FAHD1
FAM213B
FASN
FAXDC2
FBP1
FBP2
FDFT1
FDPS
FECH
FH
FLAD1
FOLH1
FPGS
FPGT
FTCD
FUK
FUT1
FUT2
FUT3
FUT4
FUT5
FUT6
FUT7
FUT8
FUT9
G6PC
G6PC2
G6PC3
G6PD

GAA
GAD1
GAD2
GADL1
GAL3ST1
GALC
GALE
GALK1
GALM
GALNS
GALNT1
GALNT10
GALNT11
GALNT12
GALNT13
GALNT14
GALNT15
GALNT16
GALNT18
GALNT2
GALNT3
GALNT4
GALNT5
GALNT6
GALNT7
GALNT8
GALNT9
GALNTL5
GALNTL6
GALT
GAMT
GANAB
GANC
GAPDH
GAPDHS
GART
GATB
GATC
GATM
GBA
GBA2
GBA3
GBE1

GBGT1
GCDH
GCH1
GCK
GCLC
GCLM
GCNT1
GCNT2
GCNT3
GCNT4
GDA
GFPT1
GFPT2
GGPS1
GGT1
GGT5
GGT6
GGT7
GK
GK2
GLB1
GLCE
GLDC
GLS
GLS2
GLUD1
GLUD2
GLUL
GLYCTK
GMDS
GMPPA
GMPPB
GMPS
GNE
GNPDA1
GNPDA2
GNS
GOT1
GOT2
GPAA1
GPAM
GPAT2
GPI

GPT
GPT2
GRHPR
GSS
GSTZ1
GUK1
GUSB
H6PD
HAAO
HADH
HADHA
HADHB
HAL
HAO1
HAO2
HDC
HEXA
HEXB
HGD
HGSNAT
HIBADH
HIBCH
HK1
HK2
HK3
HKDC1
HLCS
HMBS
HMGCL
HMGCLL1
HMGCR
HMGCS1
HMGCS2
HOGA1
HPD
HPGDS
HPRT1
HPSE
HPSE2
HSD11B1
HSD17B1
HSD17B10
HSD17B12

HSD17B2
HSD17B3
HSD17B4
HSD17B6
HSD17B7
HSD17B8
HSD3B1
HSD3B2
HSD3B7
HYAL1
HYAL2
HYAL3
HYAL4
HYI
HYKK
IDH1
IDH2
IDH3A
IDH3B
IDH3G
IDI1
IDI2
IDNK
IDO1
IDO2
IDS
IDUA
IL4I1
IMPA1
IMPA2
IMPAD1
IMPDH1
IMPDH2
INPP1
INPP4A
INPP4B
INPP5A
INPP5B
INPP5E
INPP5J
INPP5K
IPPK
ISYNA1

ITPA
ITPK1
ITPKA
ITPKB
ITPKC
IVD
JMJD7-PLA2G4B
KDSR
KHK
KL
KLF15
KMO
KYNU
LALBA
LAP3
LCLAT1
LCT
LDHA
LDHAL6A
LDHAL6B
LDHB
LDHC
LIAS
LIPC
LIPF
LIPG
LIPT1
LIPT2
LPCAT1
LPCAT2
LPCAT4
LPIN1
LPIN2
LPIN3
LSS
LTA4H
LTC4S
MAN1A1
MAN1A2
MAN1B1
MAN1C1
MAN2A1
MAN2A2

MAOA
MAOB
MAT1A
MAT2A
MAT2B
MBOAT1
MBOAT2
MCAT
MCCC1
MCCC2
MCEE
MDH1
MDH2
ME1
ME3
MECR
MGAM
MGAT1
MGAT2
MGAT3
MGAT4A
MGAT4B
MGAT4C
MGAT4D
MGAT5
MGAT5B
MGLL
MINPP1
MLYCD
MMAB
MOCS1
MOCS2
MOGAT3
MOGS
MPI
MPST
MRI1
MSMO1
MTAP
MTHFD1
MTHFD1L
MTHFD2
MTHFD2L

MTHFR
MTHFS
MTM1
MTR
MUT
MVD
MVK
NADK
NADSYN1
NAGLU
NAGS
NAMPT
NANP
NANS
NAPRT
NAT1
NAT2
NAT8L
ND1
ND2
ND3
ND4
ND4L
ND5
ND6
NDST1
NDST2
NDST3
NDST4
NDUFA1
NDUFA10
NDUFA11
NDUFA12
NDUFA13
NDUFA2
NDUFA3
NDUFA4
NDUFA4L2
NDUFA5
NDUFA6
NDUFA7
NDUFA8
NDUFA9

NDUFAB1
NDUFB1
NDUFB10
NDUFB11
NDUFB2
NDUFB3
NDUFB4
NDUFB5
NDUFB6
NDUFB7
NDUFB8
NDUFB9
NDUFC1
NDUFC2
NDUFC2-KCTD14
NDUFS1
NDUFS2
NDUFS3
NDUFS4
NDUFS5
NDUFS6
NDUFS7
NDUFS8
NDUFV1
NDUFV2
NDUFV3
NFS1
NME1
NME1-NME2
NME2
NME3
NME4
NME5
NME6
NME7
NMNAT1
NMNAT2
NMNAT3
NMRK1
NMRK2
NNMT
NNT
NOS1

NOS2
NOS3
NSDHL
NT5C
NT5C1A
NT5C1B
NT5C1B-RDH14
NT5C2
NT5C3A
NT5C3B
NT5E
NT5M
NTPCR
OAT
OCRL
ODC1
OGDH
OGDHL
OLAH
OTC
OXSM
P4HA1
P4HA2
P4HA3
PAFAH1B1
PAFAH1B2
PAFAH1B3
PAFAH2
PAH
PAICS
PANK1
PANK2
PANK3
PANK4
PAPSS1
PAPSS2
PC
PCCA
PCCB
PCK1
PCK2
PCYT1A
PCYT1B

PCYT2
PDHA1
PDHA2
PDHB
PDHX
PDXK
PDXP
PEMT
PFAS
PFKL
PFKM
PFKP
PGAM1
PGAM2
PGAM4
PGAP1
PGD
PGK1
PGK2
PGLS
PGM1
PGM2
PGP
PGS1
PHGDH
PHOSPHO1
PHOSPHO2
PHYKPL
PI4K2A
PI4K2B
PI4KA
PI4KB
PIGA
PIGB
PIGC
PIGF
PIGH
PIGK
PIGL
PIGM
PIGN
PIGO
PIGP

PIGQ
PIGS
PIGT
PIGU
PIGV
PIGW
PIGX
PIGY
PIK3C2A
PIK3C2B
PIK3C2G
PIK3C3
PIP5K1A
PIP5K1B
PIP5K1C
PIP5KL1
PIPOX
PISD
PKLR
PKM
PLA2G10
PLA2G12A
PLA2G12B
PLA2G16
PLA2G1B
PLA2G2A
PLA2G2C
PLA2G2D
PLA2G2E
PLA2G2F
PLA2G3
PLA2G4A
PLA2G4B
PLA2G4C
PLA2G4D
PLA2G4E
PLA2G4F
PLA2G5
PLA2G6
PLA2G7
PLB1
PLCB1
PLCB2

PLCB3
PLCB4
PLCD1
PLCD3
PLCD4
PLCE1
PLCG1
PLCG2
PLCH1
PLCH2
PLCZ1
PLD1
PLD2
PLD3
PLD4
PMM1
PMM2
PMVK
PNLIP
PNLIPRP1
PNLIPRP2
PNLIPRP3
PNMT
PNP
PNPLA2
PNPLA3
PNPO
POC1B-GALNT4
POLA1
POLA2
POLD1
POLD2
POLD3
POLD4
POLE
POLE2
POLE3
POLE4
POLG
POLG2
POLR1A
POLR1B
POLR1C

POLR1D
POLR1E
POLR2A
POLR2B
POLR2C
POLR2D
POLR2E
POLR2F
POLR2G
POLR2H
POLR2I
POLR2J
POLR2J2
POLR2J3
POLR2K
POLR2L
POLR3A
POLR3B
POLR3C
POLR3D
POLR3F
POLR3G
POLR3GL
POLR3H
POLR3K
PON1
PON2
PON3
PPAP2A
PPAP2B
PPAP2C
PPAT
PPCDC
PPCS
PPOX
PPT1
PPT2
PRDX6
PRIM1
PRIM2
PRODH
PRODH2
proline

PRPS1
PRPS1L1
PRPS2
PSAT1
PSPH
PTDSS1
PTDSS2
PTGDS
PTGES
PTGES2
PTGES3
PTGIS
PTGS1
PTGS2
PTS
PYCR1
PYCR2
PYCRL
PYGB
PYGL
PYGM
QARS
QDPR
QPRT
QRSL1
RDH10
RDH11
RDH12
RDH16
RDH8
REV3L
RFK
RGN
RIMKLA
RIMKLB
RPE
RPEL1
RPIA
RPN1
RPN2
RRM1
RRM2
RRM2B

SARDH
SAT1
SAT2
SC5D
SCLY
SCP2
SDHA
SDHB
SDHC
SDHD
SDS
SDSL
SEPHS1
SEPHS2
SGMS1
SGMS2
SGPL1
SGSH
SHMT1
SHMT2
SI
SLC27A5
SLC33A1
SMPD1
SMPD2
SMPD3
SMPD4
SMS
SORD
SPAM1
SPHK1
SPHK2
SPR
SPTLC1
SPTLC2
SPTLC3
SQLE
SRM
ST20-MTHFS
ST3GAL1
ST3GAL2
ST3GAL3
ST3GAL4

ST3GAL5
ST3GAL6
ST6GAL1
ST6GAL2
ST6GALNAC1
ST6GALNAC3
ST6GALNAC4
ST6GALNAC5
ST6GALNAC6
ST8SIA1
ST8SIA5
STT3A
STT3B
SUCLA2
SUCLG1
SUCLG2
SYNJ1
SYNJ2
TALDO1
TAT
TBXAS1
TCIRG1
TDO2
TGDS
TH
THTPA
TK1
TK2
TKT
TKTL1
TKTL2
TM7SF2
TPH1
TPH2
TPI1
TPK1
TPO
TRAK2
TREH
TRIT1
TST
TSTA3
TUSC3

TWISTNB
TYMP
TYMS
TYR
TYRP1
UAP1
UAP1L1
UCK1
UCK2
UCKL1
UGCG
UGDH
UGP2
UGT1A1
UGT1A10
UGT1A3
UGT1A4
UGT1A5
UGT1A6
UGT1A7
UGT1A8
UGT1A9
UGT2A1
UGT2A2
UGT2A3
UGT2B10
UGT2B11
UGT2B15
UGT2B17
UGT2B28
UGT2B4
UGT2B7
UGT8
UMPS
UPB1
UPP1
UPP2
UPRT
UQCR10
UQCR11
UQCRB
UQCRC1
UQCRC2

UQCRFS1
UQCRH
UQCRHL
UQCRQ
URAD
UROC1
UROD
UROS
UXS1
WBSCR17
XDH
XYLB
XYLT1
XYLT2
ZNRD1

Supplementary Table 3. A list of the 45 genes with the highest impact on the layout of the SNA.CNA sample similarity plot.

ImpactScore was calculated as the change in the sum of all inter-sample distances in similarity plots obtained before and after removing the named gene. 45 genes with the highest impact scores are listed below.

Gene	ImpactScore (A.U.)
IDH1	94.62573562
TP53	55.72199052
ATRX	21.04492661
TTN	9.33324746
PTEN	6.245538775
EGFR	5.894697455
CIC	5.780078948
FRG1B	3.645510755
MUC16	3.128916005
PIK3CA	2.358006173
PIK3R1	1.465046184
RYR2	1.409495785
NBPF10	1.19814325
MUC17	1.198142084
HSD17B7P2	1.148019803
BAGE2	1.051021315
PCDHGC5	1.004144669
MTFR1	0.992275126
GPRIN1	0.992041129
SUSD2	0.991982686
PCSK7	0.991981885
BFSP1	0.9919601
SFRP2	0.991951325
APOL5	0.991946108
NTN3	0.991942754
AGXT2	0.991941791
ADIPOR1	0.991939614
FBXL20	0.991939231
KDM2A	0.991938837
SLC13A4	0.991938692
TMEM177	0.991938663
KHNYN	0.991938663
WTAP	0.991938663
SEMA6C	0.991938663

TNRC6B	0.991938663
RNF10	0.991938663
CDC16	0.991938663
PEX3	0.991938663
MSX1	0.991938663
HCN3	0.991938663
SEMA3F	0.991938663
FAM73A	0.991938663
TSPAN4	0.991938663
PHF13	0.991938663
DARS2	0.991938663

Supplementary Table 4. Genes that are both differentially expressed and differentially methylated between the two CIMP-LGG groups.

111 genes Diff. exp. & diff. Me. across 2 LGG clusters

ACCN1
AR
ASCL2
BARHL1
BARHL2
BATF3
BMP8B
BNC1
C14orf23
C2orf67
C5orf38
C6orf138
C7orf13
C8orf56
CACHD1
CBX2
CELSR1
CMTM7
CREB3L4
CRHR1
CRYBA2
CYGB
CYP24A1
D4S234E
DDC
DLX4
DUOXA1
FAM84A
FLJ45983
FOXA1
FOXE3
FXYD7
GAD2
GFPT2
GPR120
GPR6
GRPEL2
HAS2
HAS2AS

TF subset

AR
ASCL2
BARHL1
BARHL2
BATF3
BNC1
CREB3L4
DLX4
FOXA1
FOXE3
HOXA13
HOXC4
HOXD8
IKZF1
LHX1
LMX1A
MESP2
MKX
MLXIPL
NKX2-4
OSR2
PITX3
POU4F1
POU4F2
PYDC1
REST
SALL3
SIX2
TFAP2E
TLX1

HCN4
HCRTR1
HIST1H2AG
HLA-DMA
HOXA13
HOXC4
HOXD8
HTR6
IKZF1
IRX1
IRX2
IRX5
KCNB2
KCNS2
LEKR1
LHX1
LHX5
LHX9
LMX1A
LOC91149
LRRC10B
LRRC33
MAP3K9
MESP2
MKX
MLXIPL
NKX2-4
NOS1
OSR2
PCDH20
PDE8A
PHACTR1
PIK3R5
PITX3
PLD5
POU4F1
POU4F2
PRCD
PRDM13
PRLHR
PYDC1
REST
RIBC2

RNF182
RNF32
RNF39
SALL3
SCGBL
SELV
SIX2
SLC35D3
SLC35F1
SLC6A3
SLC7A14
SMC1B
SPRY4
ST8SIA2
SYCE2
SYT9
TAS1R1
TFAP2E
TLR5
TLX1
TMC8
TRIM67
TSPAN32
USP44
WNT9B
YBX2
ZIC5
ZNF662
ZNF876P

Supplementary Table 5. 22 genes associated with GPCR signaling. 22 GPCR genes.

Gene	**Organism**	**Collection**	**TargetMine Integrated Pathway**
AR	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
ASIC2	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
CREB3L4	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
CRHR1	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
DDC	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
FFAR4	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
GAD2	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
HCN4	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
HCRTR1	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
HIST1H2AG	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
HTR6	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
KCNB2	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
KCNS2	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
MLXIPL	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
NOS1	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
PDE8A	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
PIK3R5	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
PRLHR	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
SLC6A3	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
SMC1B	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
TAS1R1	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism
WNT9B	Homo sapiens	H002	GPCR ligand binding|Neuronal System|G alpha (i) signalling events|cAMP signaling pathway|Calcium signaling pathway|Alcoholism