



DNA methylation markers for diagnosis and prognosis of common cancers

Xiaoke Hao^{a,1,2}, Huiyan Luo^{b,c,1}, Michal Krawczyk^{c,1}, Wei Wei^{b,c,1}, Wenqiu Wang^{c,d,1}, Juan Wang^{a,1}, Ken Flagg^c, Jiayi Hou^c, Heng Zhang^e, Shaohua Yi^c, Maryam Jafari^c, Danni Lin^c, Christopher Chung^c, Bennett A. Caughey^c, Gen Li^f, Debanjan Dhar^g, William Shi^c, Lianghong Zheng^f, Rui Hou^f, Jie Zhu^c, Liang Zhao^f, Xin Fu^c, Edward Zhang^c, Charlotte Zhang^c, Jian-Kang Zhu^e, Michael Karin^{g,2}, Rui-Hua Xu^{b,2}, and Kang Zhang^{c,h,2}

^aDepartment of Clinical Laboratory Medicine, Xijing Hospital, Fourth Military Medical University, Xi'an 710032, China; ^bState Key Laboratory of Oncology, Sun Yat-sen University Cancer Center, Guangzhou 510060, China; ^cInstitute for Genomic Medicine, University of California, San Diego, La Jolla, CA 92093; ^dShanghai Key Laboratory of Ocular Fundus Diseases, Shanghai General Hospital, Shanghai 200080, China; ^eShanghai Center for Plant Stress Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, Shanghai 210602, China; ^fGuangzhou Youze Biological Pharmaceutical Technology Company Ltd., Guangzhou 510005, China; ^gDepartment of Pharmacology, University of California, San Diego, La Jolla, CA 92328; and ^hVeterans Administration Healthcare System, San Diego, CA 92093

Contributed by Michael Karin, May 24, 2017 (sent for review March 3, 2017; reviewed by Hakon Hakonarson and Wei Zhang)

The ability to identify a specific cancer using minimally invasive biopsy holds great promise for improving the diagnosis, treatment selection, and prediction of prognosis in cancer. Using whole-genome methylation data from The Cancer Genome Atlas (TCGA) and machine learning methods, we evaluated the utility of DNA methylation for differentiating tumor tissue and normal tissue for four common cancers (breast, colon, liver, and lung). We identified cancer markers in a training cohort of 1,619 tumor samples and 173 matched adjacent normal tissue samples. We replicated our findings in a separate TCGA cohort of 791 tumor samples and 93 matched adjacent normal tissue samples, as well as an independent Chinese cohort of 394 tumor samples and 324 matched adjacent normal tissue samples. The DNA methylation analysis could predict cancer versus normal tissue with more than 95% accuracy in these three cohorts, demonstrating accuracy comparable to typical diagnostic methods. This analysis also correctly identified 29 of 30 colorectal cancer metastases to the liver and 32 of 34 colorectal cancer metastases to the lung. We also found that methylation patterns can predict prognosis and survival. We correlated differential methylation of CpG sites predictive of cancer with expression of associated genes known to be important in cancer biology, showing decreased expression with increased methylation, as expected. We verified gene expression profiles in a mouse model of hepatocellular carcinoma. Taken together, these findings demonstrate the utility of methylation biomarkers for the molecular characterization of cancer, with implications for diagnosis and prognosis.

DNA methylation | cancer diagnosis | cancer prognosis | gene expression | survival analysis

Accurate diagnosis of cancer based on histological subtype, as well as other markers identified via histology and immunohistochemistry, is crucial for choosing the proper treatment and for predicting survival (1). For some primary tumors, complex anatomy may prevent accurate identification of the tissue of origin or tumor type. Tissue must be obtained from these tumors either from surgical resection or from a tissue biopsy. Diagnosis in these cases may be limited by the patient's tolerance of surgery or by inaccessibility of the tumor, preventing acquisition of a tissue sample of adequate size and quality that preserves tissue architecture. Even when high-quality biopsy specimens are obtained, diagnostic uncertainty may persist, hindering treatment decisions and prognostication. Thus, there is a need for strategies to improve diagnostic certainty. Molecular characterization is increasingly used to predict tumor prognosis and response to therapy and offers great potential for improving understanding of an individual patient's tumor (2–4). Importantly, these methods may have specific utility in scenarios of limited tissue availability or quality.

Methylation of CpG sites is an epigenetic regulator of gene expression that usually results in gene silencing (5, 6). Extensive perturbations of DNA methylation have been noted in cancer, causing changes in gene regulation that promote oncogenesis (7–9). Understanding both epigenetic changes and somatic DNA mutations show promise for improving the characterization of malignancy to predict treatment response and prognosis (3, 10–12). Some changes in methylation are reproducibly found in nearly all cases of a specific type of cancer. In contrast, somatic mutations are often neither specific nor sensitive for a particular type of cancer. Even within commonly mutated genes, individual mutations may be found across tens or hundreds of kilobases, limiting the utility of targeted sequencing of molecular markers (10, 13, 14).

Consequently, to explore the utility of DNA methylation analysis for cancer diagnosis, we analyzed whole-genome methylation profiles of tumors and matched normal tissue from patients with four of the most common cancers to identify potential cancer-specific DNA methylation markers. We then verified these methylation markers in two other independent patient

Significance

The ability to identify a specific cancer using minimally invasive biopsy holds great promise for improving diagnosis and prognosis. We evaluated the utility of DNA methylation profiles for differentiating tumors and normal tissues for four common cancers (lung, breast, colon, and liver) and found that they could differentiate cancerous tissue from normal tissue with >95% accuracy. This signature also correctly identified 19 of 20 breast cancer metastases and 29 of 30 colorectal cancer metastases to the liver. We report that methylation patterns can predict the prognosis and survival, with good correlation between differential methylation of CpG sites and expression of cancer-associated genes. Their findings demonstrate the utility of methylation biomarkers for the molecular characterization, diagnosis, and prognosis of cancer.

Author contributions: X.H., M. Karin, R.-H.X., and K.Z. designed research; X.H., H.L., M. Krawczyk, W. Wei, W. Wang, J.W., K.F., H.Z., S.Y., M.J., D.L., C.C., G.L., W.S., L. Zheng, R.H., Jie Zhu, X.F., E.Z., and C.Z. performed research; J.H., B.A.C., D.D., L. Zhao, and Jian-Kang Zhu analyzed data; and X.H., M. Karin, R.-H.X., and K.Z. wrote the paper.

Reviewers: H.H., Children's Hospital of Philadelphia; and W.Z., Wake Forest Baptist Comprehensive Cancer Center.

The authors declare no conflict of interest.

¹X.H., H.L., M. Krawczyk, W. Wei, W. Wang, and J.W. contributed equally to this work.

²To whom correspondence may be addressed. Email: haoxkg@fmmu.edu.cn, mkarin@ucsd.edu, xurh@sysucc.org.cn, or kang.zhang@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1703577114/-DCSupplemental.

Table 1. Confusion table of the TCGA training cohort

Training cohort	Breast cancer	Colon cancer	Liver cancer	Lung cancer	Normal breast	Normal colon	Normal liver	Normal lung	Total
Breast cancer	520				6				
Colon cancer		275				5			
Liver cancer			238				9		
Lung cancer				584				6	
Normal breast			1	1	59	1			
Normal colon						21			
Normal liver							23		
Normal lung								43	
Total	520	275	239	585	65	27	32	49	1,792
Correct	520	275	238	584	59	21	23	43	1,763
False-positive					6	5	9	6	26
False-negative			1						3
Wrong tissue				1		1			3
Correct (%)	100	100	99.6	99.8	90.8	77.8	71.9	87.8	98.4

Orange indicates cancer sample, purple indicates normal sample, and gray indicates correctly diagnosed sample number of each training cohort.

cohorts. We also used methylation patterns to predict survival and analyzed the utility of combining methylation with mutational status in several tumor types. Finally, we correlated specific methylation patterns with gene expression in genes known to be important in cancer biology.

Results

Characteristics of Patients and Tissues. Clinical characteristics and molecular profiling, including methylation data for a training cohort of 1,619 tumor samples and 173 matched adjacent normal tissue samples, as well as a validation cohort of 791 tumor and 93 matched normal samples, were obtained from The Cancer Genome Atlas (TCGA). A separate validation cohort of 394 tumor samples and 324 matched normal samples was obtained from Chinese patients with cancer treated at the Sun Yat-sen University Cancer Center, West China Hospital, and Xijing Hospital. Matched adjacent normal tissue samples were collected simultaneously with tumor tissue from the same patient and were verified by histology to have no evidence of cancer. Clinical characteristics of all patients are summarized in *SI Appendix, Tables S1–S3*.

Methylation Profiling Identifies Cancer-Specific Methylation Signatures.

To identify a cancer type-specific signature, we randomly split the full TCGA dataset into training and test cohorts with a 2:1 ratio in each of the eight types of sample groups. We first performed the prescreening procedure to remove excessive

noise on the training data using the moderated *t* statistic (15). For multinomial classification, we used lasso (least absolute shrinkage and selection operator) under a multinomial distribution. A multiclass prediction system (16) was constructed to predict the group membership of samples using a panel of markers. Hierarchical clustering of these samples according to differential methylation of CpG sites in this fashion could distinguish the cancer tissue of origin, as well as differentiate cancer tissue from normal tissue in our TCGA training cohort (Table 1). The overall correct diagnosis rate was 98.4%. We then applied these markers to a TCGA validation cohort (Table 2), and found a slightly decreased but statistically similar correct rate of 97.1%. We also confirmed our results in an independent cohort of Chinese cancer patients (Table 3), which also showed a decreased but similar correct rate of 95.0%. Of note, the methylation analysis of the Chinese cohort was performed using an alternative bisulfite sequencing technique in a different ethnic and geographic background than the TCGA cohorts. Overall, these results demonstrate the robust nature of these methylation patterns in identifying the presence of malignancy as well as its site of origin (Fig. 1 and *SI Appendix, Table S4 and Fig. S1*).

Methylation Block Structure for Improved Allele Calling Accuracy. We used the well-established concept of genetic linkage disequilibrium to study the degree of comethylation among different DNA stands. We used paired-end Illumina sequencing reads to

Table 2. Confusion table of validation cohort 1

Validation cohort 1	Breast cancer	Colon cancer	Liver cancer	Lung cancer	Normal breast	Normal colon	Normal liver	Normal lung	Total
Breast Cancer	268			1	4				
Colon Cancer		129				7			
Liver Cancer			136				4		
Lung Cancer	1			252				5	
Normal Breast	1		1		28				
Normal Colon						11			
Normal Liver			1				14		
Normal Lung				1				20	
Totals	270	129	138	254	32	18	18	25	884
Correct	268	129	136	252	28	11	14	20	858
False-positive					4	7	4	5	20
False-negative	1		1	1					3
Wrong tissue	1		1	1					2
Correct (%)	99.3	100	98.6	99.2	87.5	61.1	77.8	80.0	97.1

Orange indicates cancer sample, purple indicates normal sample, and gray indicates correctly diagnosed sample number of each validation cohort.

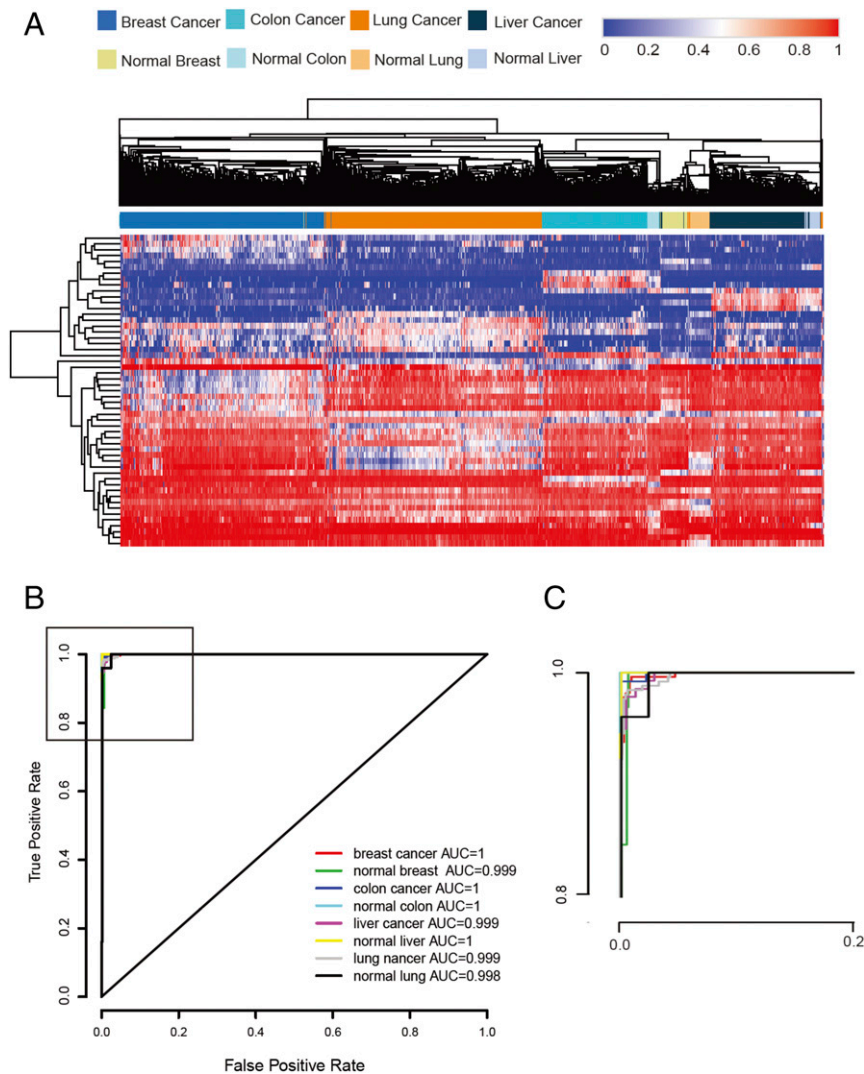


Fig. 1. Methylation signatures can differentiate different cancer types from corresponding normal tissues. (A) Unsupervised hierarchical clustering and heat map presentation associated with the methylation profile (according to the color scale shown) in different cancer types. (B) ROC curve showing the high sensitivity and specificity in predicting different cancer types. (C) Zoom-in view of the block diagram in B.

identify each individual methylation block (mBlock). We applied a Pearson correlation method to quantify the comethylation of mBlock. We compiled all common mBlocks of a region by calculating different mBlock fractions (*Methods*). We then partitioned the genome into blocks of tightly comethylated CpG sites that we termed methylation-correlated blocks (MCBs), using an R^2 cutoff of 0.5. We surveyed MCBs in cancer and normal tissues and found that MCBs were highly consistent among different cancer and normal tissues. Overall, we found ~3,600 MCBs, approximately one-half of which were incomplete/disrupted (*SI Appendix, Fig. S2*) owing to short a span of sequenced reads (~100 base pairs).

We next determined methylation values within MCBs. *SI Appendix, Fig. S3* shows an example of MCBs found on chromosome 1 in both normal tissues (breast, colon, liver, and lung) and corresponding tumor tissues: breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), liver hepatocellular carcinoma (LIHC), and lung adenocarcinoma (LUAD). We found similar β values across multiple CpG sites within a MCB, and thus calculated a compound methylation value for one entire MCB. We used them instead of single CpG sites in downstream

bioinformatics pipelines, which significantly enhanced the allele-calling accuracy.

Methylation Profiles Can Identify Cancer Metastases to Liver. Because identifying the tissue of origin is crucial in selecting the optimum treatment strategy for patients presenting with metastases, we investigated the utility of DNA methylation analysis for diagnosis of cancer metastases to liver and lung in our Chinese cohort. In addition to the aforementioned primary tumors, we analyzed 30 colorectal cancer metastases to liver and 34 colorectal cancer metastases to lung. We found that unsupervised hierarchical clustering could differentiate these metastases from colon cancer or normal tissue (Fig. 2). The methylation signature could correctly diagnose 29 of 30 colorectal cancer metastases to liver and 32 of 34 colorectal cancer metastases to lung (Table 3); one of the three misdiagnoses were identified as normal liver and two of the three misdiagnoses were identified as normal colorectal tissue, suggesting that the error was due to tissue contamination. These findings support the potential for using the DNA methylation signature to improve the diagnosis of metastatic disease in addition to primary cancers.

Table 3. Confusion table of validation cohort 2

Validation cohort 2	Breast cancer	Colorectal cancer	Colorectal cancer metastases to liver	Colorectal cancer metastases to lung	Liver cancer	Lung cancer	Normal breast	Normal colon	Normal liver	Normal lung	Total
Breast cancer	65						4			1	
Colorectal cancer	1	154	29	32							
Liver cancer		1			44	1			1	3	
Lung cancer						43				1	
Normal breast	7						41				
Normal colon		6		2				161			
Normal liver		2	1		2				72		
Normal lung		1				2				41	
Total	73	164	30	34	46	47	45	161	73	45	718
Correct	65	154	29	32	44	43	41	161	72	41	682
False-positive							4		1	1	6
False-negative	7	9	1	2	2	2					23
Wrong tissue	1	1				2				4	8
Correct (%)	89.4	93.9	96.7	94.1	95.6	91.5	91.1	100	98.6	91.1	95.0

Orange indicates cancer sample, purple indicates normal sample, and gray indicates correctly diagnosed sample number of each validation cohort.

Methylation Profiles Predict Prognosis and Survival. We next assessed the prognostic utility of a methylation signature for each type of cancer. Clinical and demographic characteristics, including age, sex, race, and American Joint Committee on Cancer stage, were included in the analysis as well, because the prognostic power can be greatly improved by combining this information with informative molecular data (17). For each cancer category, we used two different statistical learning algorithms, lasso and boosting, to reduce the dimensionality of markers and construct a predictive model. We evaluated the prognostic utility on TCGA training and validation cohorts at a 2:1 ratio. Our method performed well in differentiating low-risk and high-risk groups in Kaplan–Meier analyses and in associated log-rank tests with significant *P* values, demonstrating significant prognostic utility of the methylation signatures in BRCA and LUAD (Fig. 3 and *SI Appendix, Tables S5 and S6*).

A Cancer Methylation Profile Correlated with Its Gene Expression Pattern and Function. Given that DNA methylation is an essential epigenetic regulator of gene expression, we sought to investigate

how differential methylation of sites in genes in cancer versus normal tissue correlated with gene expression. Specifically, we were interested in those methylation sites that predicted the presence of malignancy in our aforementioned signatures. As described in *Methods*, we used both methylation and RNA sequencing data to select top CpG markers in LIHC for which methylation was significantly correlated with gene expression. As expected, we typically observed an inverse correlation between promoter methylation and gene expression and identified several genes known to be important in carcinogenesis, as well as genes with relatively unknown functional relevance in LIHC. Among the genes hypermethylated with decreased expression, we selected one gene for LIHC [fuzzy planar cell polarity protein (*Fuz*); Fig. 4]. Overexpression of FUZ suppressed LIHC cell line growth (Fig. 4).

We further attempted to validate a list of top genes whose methylation patterns were closely correlated with gene expression in a mouse model of LIHC. We found a good correlation between the gene expression profiles in human and mouse LIHC (*SI Appendix, Fig. S4 and Table S7*). These results support a

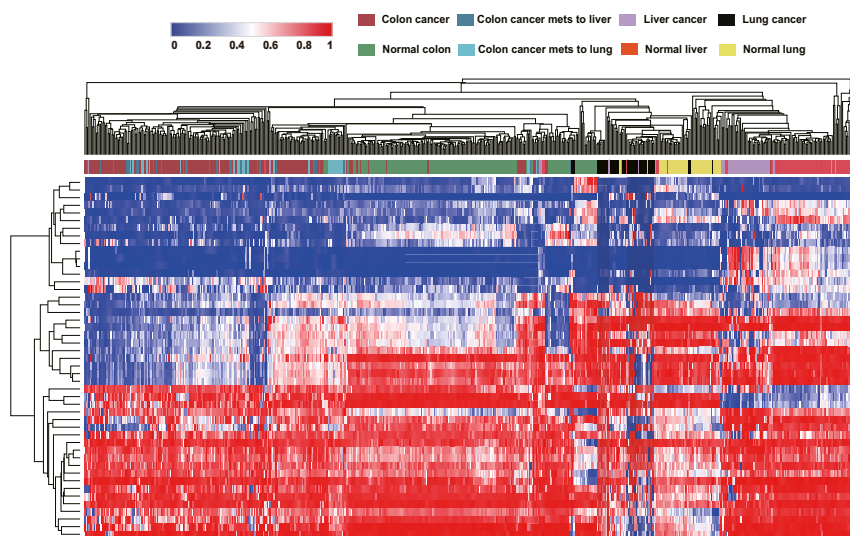


Fig. 2. DNA methylation signatures can identify the cancer of origin in metastasis of colon cancer. Shown are unsupervised hierarchical clustering and heat map associated with the methylation profile of 394 tumor samples and 324 normal samples of primary and metastatic colorectal cancer, liver cancer, and lung cancer in a Chinese cohort with a panel of 46 CpG markers. Each column represents an individual patient, and each row represents an individual CpG marker. The color scale shows relative methylation.

functional role of these methylation markers in promoting carcinogenesis and provide biological validation for their use in methylation studies to characterize cancers.

Discussion

The present study demonstrates the potential for using methylation signatures to identify cancer tissue of origin and predict prognosis. Although we focused on four common cancers here, we expect that DNA methylation analysis can be readily expanded to aid diagnosis of a much larger number of cancers. Our results may be particularly helpful for identifying cancers in cases with an inadequate tissue yield or quality for histological diagnosis, which requires preservation of the tissue architecture. In contrast, DNA methylation analysis requires only a small amount of

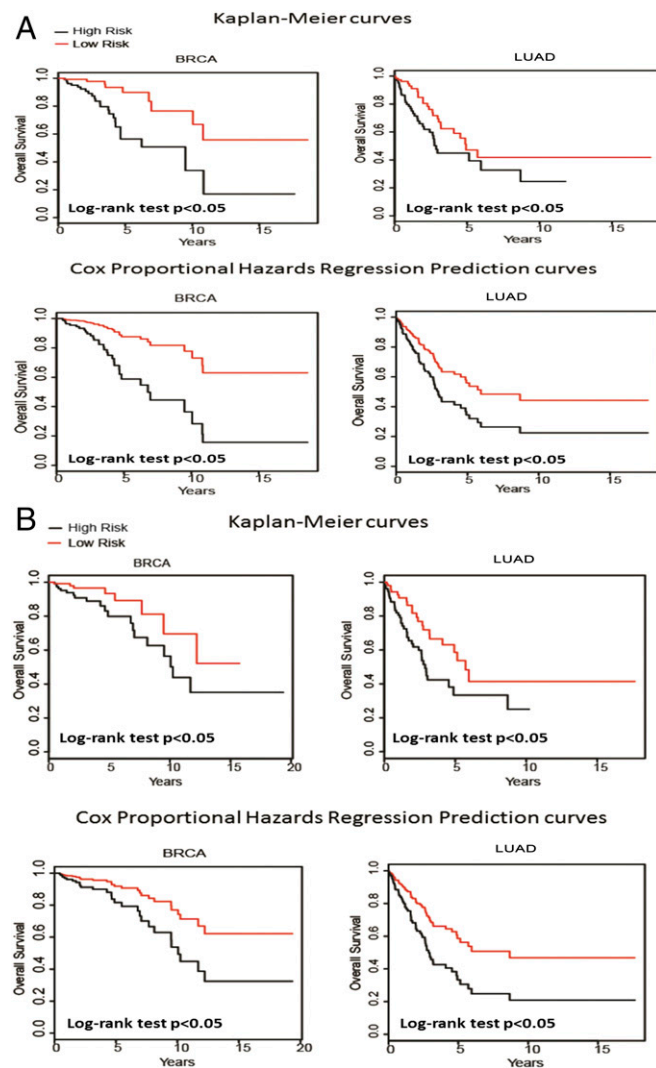


Fig. 3. Methylation markers can predict overall survival of patients in different types of cancers. (A) Overall survival curves of BRCA and LUAD patients with a low or high risk of death, according to a combined prognosis score from a lasso analysis. Shown are Kaplan–Meier curves (Upper) and Cox proportional hazards regression prediction curves (Lower) of overall survival in BRCA (Left) and LUAD (Right) patients with low or high risk of death. (B) Overall survival curves of BRCA and LUAD patients with a low or high risk of death, according to a combined prognosis score from a boosting analysis. Shown are Kaplan–Meier curves (Upper) and Cox proportional hazards regression prediction curves (Lower) of overall survival of BRCA (Left) and LUAD (Right) patients with low or high risk of death, according to a combined prognosis score from a boosting analysis.

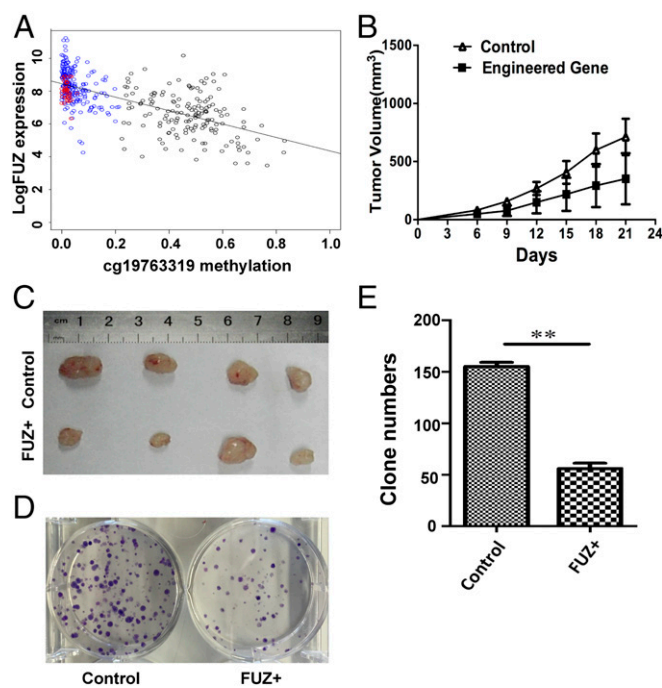


Fig. 4. Linking differentially methylated markers to gene expression in LIHC. (A) Relationship between methylation of CpG marker cg19763319 and expression of FUZ in liver cancer. Red dots indicate normal tissue samples; black dots, cancer samples. (B) Effect of FUZ expression on growth of liver cancer cell line HEP1. (C) Effect of FUZ expression on growth of HEP1 cells in a mouse xenograft model. (D) Effect of FUZ expression on colony formation of HEP1 cells. (E) Quantified colony formation by FUZ-transduced HEP1 cells compared with control. $**P < 0.001$.

tissue to obtain adequate DNA, thus potentially allowing the use of lower-quality biopsies. These studies also may have significant utility in assigning diagnoses from analysis of metastatic lesions, especially when the tumor is of an unknown primary cancer type.

Through sequencing of bisulfate-converted DNA (bis-DNA), we identified many previously unknown CpG markers differentially methylated in cancer tissues versus normal tissues. Lehmann-Werman et al. (18) described multiple adjacent CpG sites that share the same tissue-specific methylation pattern. We further explored this concept of the mBlock and found that many nearby methylation markers are highly correlated. This information allowed us to identify additional markers and improve the accuracy of sequencing for determining significant methylation differences. This method has substantial potential for improving the accuracy and utility of DNA methylation analysis for the four study cancer types and other cancers, as well as for expanding the number of diagnostic markers available for interrogation. However, the length of an MCB, which is related to how long a DNA methyl-transferase binds to and exerts its enzymatic effect on modifying adjacent and surrounding CpG sites on a DNA strand, is not clear, because its underlying biochemical basis is not fully defined.

DNA methylation analysis has the potential to improve outcomes, given that accurate diagnosis is often crucial to treatment selection. Our application of methylation signatures to prognosis revealed subsets of patients with positive and negative prognoses. This finding raises the possibility that methylation may help identify relatively indolent or aggressive tumors and may aid decision making regarding the selection of more aggressive or less aggressive treatment and monitoring. Further studies are warranted to fully explore the clinical applications of methylation sequencing to guide personalized care for patients with cancer.

Methods

Training and first validation cohorts were performed on patient data obtained from TCGA. Patient characteristics are summarized in *SI Appendix, Tables S1 and S2*. Complete clinical, molecular, and histopathological datasets are available at the TCGA website (<https://tcga-data.nci.nih.gov/docs/publications/tcga/>). Individual institutions that contributed samples coordinated the consent process and obtained informed written consent from each patient in accordance with their respective institutional review boards. A second independent (Chinese) cohort consisted of patients of the Sun Yat-sen University Cancer Center, the West China Hospital in Chengdu, China, and Xijing Hospital. Those who presented with lung adenocarcinoma, liver hepatocellular carcinoma, breast adenocarcinoma, and colorectal adenocarcinoma, including metastatic disease, were selected and enrolled in this

study. Patient characteristics are also summarized in *SI Appendix, Tables S1 and S3*. This project was approved by the IRB of the Sun Yat-sen University Cancer Center, West China Hospital, and Xijing Hospital. Informed consent was obtained from all patients. Tumor and normal tissues were obtained as clinically indicated for patient care and were retained for this study with patients' informed consent.

Information on data sources, statistical analyses, probe design, bis-DNA capture, sequencing and data analysis, DNA extraction, cell culture, colony formation assays, and tumor xenografts is provided in *SI Appendix*.

ACKNOWLEDGMENTS. This study was supported in part by the Carol and Dick Hertzberg Fund, the Richard Annesser Fund, and the Michael Martin Fund.

1. DeVita VT, et al. (2011) *DeVita, Hellman, and Rosenberg's Cancer: Principles and Practice of Oncology* (Lippincott Williams & Wilkins, Philadelphia), Ed 9.
2. Wang T, et al. (2015) Identification and characterization of essential genes in the human genome. *Science* 350:1096–1101.
3. Han L, et al. (2014) The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat Commun* 5:3963.
4. Akbani R, et al. (2014) A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun* 5:3887.
5. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat Genet* 33:245–254.
6. Vaissière T, Sawan C, Herceg Z (2008) Epigenetic interplay between histone modifications and DNA methylation in gene silencing. *Mutat Res* 659:40–48.
7. Egger G, Liang G, Aparicio A, Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 429:457–463.
8. Herman JG, Baylin SB (2003) Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* 349:2042–2054.
9. Feinberg AP, Tycko B (2004) The history of cancer epigenetics. *Nat Rev Cancer* 4: 143–153.
10. Kandath C, et al. (2013) Mutational landscape and significance across 12 major cancer types. *Nature* 502:333–339.
11. Paez JG, et al. (2004) EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. *Science* 304:1497–1500.
12. Ogino S, et al.; Alliance for Clinical Trials in Oncology (2013) Predictive and prognostic analysis of PIK3CA Mutation in Stage III Colon Cancer Intergroup Trial. *J Natl Cancer Inst* 105:1789–1798.
13. Koboldt DC, et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22:568–576.
14. Dees ND, et al. (2012) MuSiC: Identifying mutational significance in cancer genomes. *Genome Res* 22:1589–1598.
15. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:3.
16. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33:1–22.
17. Yuan Y, et al. (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* 32:644–652.
18. Lehmann-Werman R, et al. (2016) Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci USA* 113:E1826–E1834.

Supplementary Figure and Table

Supplementary Fig 1. Unsupervised hierarchical clustering of DNA methylation of COAD versus normal colon, BRCA versus normal breast, LIHC versus normal liver, and lung cancer versus normal lung using binary methylation markers in the TCGA validation cohort.

Supplementary Fig 2. DNA methylation signatures can identify cancer of origin in metastasis of colon cancer. Supervised hierarchical clustering and heat map associated with the methylation profile of 394 tumor and 324 normal samples of primary and metastatic colorectal cancer, liver cancer, and lung cancer in a Chinese cohort with a panel of 46 CpG markers. Each column represents an individual patient and each row represents an individual CpG marker. Color scale represents relative methylation.

Supplementary Fig 3. List of Blocks of Correlated Methylation (BCM). Data was compiled from both healthy tissue samples (lung, liver, colon, breast) and corresponding cancers (LUAD, LIHC, COAD, BRCA). Each row represents a CG dinucleotide analyzed in this study. Subset of those correspond to CG markers included on Illumina's Infinium HumanMethylation450 BeadChip (leftmost column), whereas the vast majority represent previously unknown potential novel markers that could be used in cancer diagnosis. Second column lists genomic distances between markers and third column shows genomic location of each analyzed CG. Graphs (eight columns, each corresponding to one normal or cancer tissue) illustrate Pearson correlation coefficients r^2 between β values of two closely positioned CGs calculated for samples from each tissue type separately. Correlation between any two markers is represented by a red square at the intersection of (virtual) perpendicular lines originating from these two markers. White color indicates no significant correlation, red color intensity marks r^2 values between 0.5 and 1. Black boxes indicate the ends of analyzed regions. A total of ~9800 CG positions have been analyzed.

Supplementary Fig 4. An example of a region within PRDM16 gene locus encompassing three Blocks of Correlated Methylation (BCM) in healthy tissue samples (lung, liver, colon, breast) and corresponding cancers (LUAD, LIHC, COAD, BRCA).

a: (left panel) a not-to-scale representation of a set of analyzed cg markers belonging to three BCMs in this region. Boundaries between identified blocks are defined as the extent of a probe and indicated by a black rectangle, whereas red squares indicate correlated methylation ($r > 0.5$) between two nearby markers. Correlation between any two markers is represented by a square at the intersection of (virtual) perpendicular lines originating from these two markers. White color indicates no significant correlation. 3 newly identified methylation markers in the left MCB anchored by marker cg23086843, 4 newly identified methylation markers in the middle MCB anchored by cg01940181 and 9 newly identified methylation markers in the right MCB anchored by cg06911744/ cg26000536 were highly consistent and correlated among all four normal and four cancer tissue DNAs. The location of all markers is indicated below the correlation graph. Using markers within the same MCB significantly improved measurement accuracy of methylation levels.

b: (three panels on the right) genomic neighborhood of the three BCM displayed within UCSC genome browser (genome.ucsc.edu). Upper portions of the subpanels show Pearson correlation data tracks for analyzed tissues by summing r values for a marker within a BCM. Cg marker names below the Pearson correlation graph are methylation markers used by TCGA. Gene name and common SNPs are also shown.

Supplementary Fig 5. Heatmap displaying positive fold changes (FC) in expression of top ten genes of interest in human LIHC versus normal liver tissue and its correlation in mouse LIHC (right).

Supplementary Table 1: Summary of study cohorts

Supplementary Table 2: Clinicopathological Characteristics of the TCGA cohort

Supplementary Table 3: Clinicopathological Characteristics of the Chinese cohort

Supplementary Table 4. List of markers presented in at least 7 out of 10 random split analyses in a multinomial analysis of four cancers and corresponding normal tissues

Supplementary Table 5. List of markers selected in 3 out of 10 training / test split analyses in survival analysis of breast cancer

Supplementary Table 6. List of markers selected in 3 out of 10 training / test split analyses in survival analysis of lung cancer (LUAD/LUSC)

Supplementary Table 7: Positive fold changes (FC) in expression of top ten genes of interest in human LIHC versus normal liver tissue and its correlation in mouse LIHC

Supplementary Table 8: Breast cancer: validation and prediction performance in a survival analysis

Supplementary Table 9: Lung cancer: validation and prediction performance in a survival analysis and its correlation in mouse LIHC

Methods

Data sources

DNA methylation data were obtained from both the TCGA analysis of 485,000 sites generated using the Infinium 450K Methylation Array and the following GSE datasets: GSE46306, GSE50192, GSE58298 and GSE41826. Methylation profiles for four unique cancer types and their corresponding normal tissues were analyzed. IDAT format files of the methylation data were generated containing the ratio values of each scanned bead. Using the minfi package from Bioconductor, these data files were converted into a score, referred to as a Beta value. Beta values for any markers that did not exist across all 20 datasets were excluded. Methylation data of the Chinese cohort was obtained by padlock based bisulfite sequencing and analyzed as described below.

Statistical Analysis

Diagnostic: building multiclass classifier

For each of the 8 types: BRCA, COAD, LIHC, LUNG (LUAD and LUSC combined) cancer and corresponding normal tissue samples, we randomly split the full dataset into training and validation sets with 2:1 ratio. We first performed the pre-screening procedure to remove excessive noise on the training data using the “moderated t-statistics” (1). For each set of comparison, one type of

sample was compared against all other 7 types of samples. A list of markers with significantly difference in mean among all 8 sets of comparisons were retained for future analysis. The Benjamini-Hochberg procedure (2) was used to control the FDR at significance level 0.05. For multinomial classification, we used LASSO under multinomial distribution. The tuning parameter was determined by the expected generalization error estimated from 10-fold cross-validation. We repeated the random split scheme for 10 times to stabilize the variable selection procedure. We constructed a composite panel by keeping markers with a high selection probability and disregard those with low selection probability from the 10 sets of markers selected from the aforementioned procedure. A multi-class prediction system (3) was constructed to predict the group membership of samples in the validation data using the panel of markers selected. A confusion matrix and ROC curves were also provided to evaluate sensitivity and specificity, in addition to prediction accuracy.

All the hypothesis testing are two-sided with p -value < 0.05 considered to be statistically significant. All the analyses were conducted in R version 3.3.2 with the following packages used: 'glmnet', 'lpc', 'CoxBoost', 'limma', 'ROCR'.

Prognostic: predicting survival outcomes

For each type of cancer: BRCA, COAD, LIHC and LUNG (LUAD and LUSC combined), we randomly split the full dataset into training and test sets with 2:1 ratio. We first performed the univariate pre-screening procedure on the training data to remove excessive noise and accelerate the computational procedure, which was generally recommended prior to applying any variable selection method (4) For each methylation marker, we fit a univariate Cox proportional hazards model by using each marker as the covariate. A marker with p -value < 0.05 from the Wald statistic was retained in the dataset.

We then applied four variable selection methods suitable for high-dimensionality on the prescreened training dataset: Least Absolute Shrinkage and Selection Operator (LASSO)(5), Elastic Net (6), Lassoed Principal Components (LPC) (7) and Boosting (8). For LASSO and Elastic Net, the tuning parameters (for LASSO and (for Elastic Net) were determined according to the expected generalization error estimated from 10-fold cross-validation and information-based criteria AIC/BIC. For LPC, the markers with p -value < 0.05 after False Discovery Rate (FDR)

correction were considered to be statistically significant. For boosting, the optimal step was determined by the expected generalization error estimated from 10-fold cross-validation. For all methods, the number of markers was also governed by the effective sample size in the training dataset, which equals to the number of events. We then fit a Cox proportional hazards model on the training data using markers selected at the optimal step as the covariates. The predictability of the model was evaluated by two criteria: on the training data and concordance probability (also known as C-index) on the test data. - the proportion of explained randomness (9) is a function of Kullback- Leibler information gain bounded between 0 and 1, with a larger value indicating larger proportion of randomness explained. C-index (10) calculates the proportions of concordant pairs among all pairs of observations with 1 indicating perfect prediction accuracy. To validate, we obtained a risk score for each patient in the test data by multiplying the unbiased coefficient estimates and the design matrix. By dividing the risk score according to its median, we formed a high and low risk groups with roughly equal number of observations. A figure of Kaplan-Meier estimator and log-rank test were included to determine if the median survival time was significantly different. A model-based prediction was also provided. The high concordance between the non-parametric and semi-parametric prediction curves indicated the possibility of accurately predicting a new patient's survival status for any future time point using the panel of markers selected.

As the results can depend strongly on the arbitrary choice of a random sample split for sparse high-dimensional data, we inherited the spirit of “multi-split” method (11) a remedy to improve variable selection consistency while controlling finite sample error. We repeated the “randomly split – screen – selection” procedure 10 times and ended up with 10 different sets of markers. The markers were then aggregated with the most common ones, determined from the frequency table to be used in the subsequent experiments. We validated the predictability of the panel of markers by randomly split another 10 sets of training and test groups with 2:1 ratio, and following the aforementioned validation procedure.

For BRCA patients, by using two different statistical learning algorithms, we identified a panel of 19 and 11 methylation markers present in at least 3 out of 10 “randomly split-screen-selection” procedure, respectively (Supplemental table 8). Among these markers identified, there were 10

overlapping markers. The average, the proportion of explained randomness calculated from the training data is 0.86 (min: 0.59; max: 0.98) for LASSO and 0.90 (min: 0.75; max: 0.99) for boosting, respectively. The average C-index calculated from the test data is 0.63 (min: 0.56; max: 0.71) for LASSO and 0.61(min:0.54; max:0.68) for boosting. The promising values from both criteria indicated good predictability from the methylation signatures. We also evaluated the prognostic utility on TCGA data on newly generated TCGA training and test datasets with 2:1 ratio. In a validation study, 5 out of 10 times the panel of aggregated methylation markers from LASSO can separate groups of high and low risk patients completely; 3 out of 10 times the panel of aggregated methylation markers from boosting can separate different risk groups (Fig. 3).

For LUNG (combined LUSC and LUAD), we detected 75 methylation markers by LASSO and 52 by boosting (45 markers overlapped) in at least 3 out of 10 “randomly split – screen – selection” procedure (Supplemental table 9). The average, the proportion of explained randomness calculated from the training data is 0.91 (min: 0.80; max: 0.96) for LASSO and 0.93 (min: 0.46; max: 0.95) for boosting, respectively. The average C-index calculated from the test data is 0.57 (min: 0.53; max:0.63) for LASSO and 0.55(min:0.50; max:0.66) for boosting. In a validation study, 2 out of 10 times the panel of aggregated methylation markers from LASSO can separate groups of high and low risk patients completely; 5 out of 10 times the panel of aggregated methylation markers from boosting can separate different risk groups (Fig 3).

Tumor DNA extraction

Genomic DNA extraction from pieces of freshly frozen healthy or cancer tissues was performed with eLiteDNA Mini Kit (EliteHealth, Guangzhou) according to manufacturer’s recommendations. DNA was extracted from roughly 0.5 mg of tissue. DNA was stored at -20°C and analyzed within one week of preparation.

DNA extraction from FFPE samples

Genomic DNA from frozen FFPE samples was extracted using eLiteDNA DNA FFPE Tissue Kit (DNA were stored at -20°C for further analysis (EliteHealth, Guangzhou)).

Bisulfite conversion of genomic DNA

1µg of genomic DNA was converted to bis-DNA using EZ DNA Methylation-Lightning™ Kit (Zymo Research) according to the manufacturer's protocol. Resulting bis-DNA had a size distribution of ~200-3000 bp, with a peak around ~500-1000 bp. The efficiency of bisulfite conversion was >99.8% as verified by deep-sequencing of bis-DNA and analyzing the ratio of C to T conversion of CH (non-CG) dinucleotides.

Determination of DNA methylation levels of the second validation cohort by deep sequencing of bis-DNA captured with molecular-inversion (padlock) probes

Padlock probes were designed to capture regions containing the CpG markers whose methylation levels significantly differed in any of the comparison between any cancer tissue and any normal tissue. Padlock-capture of bis-DNA was based on published techniques and protocols with modifications (12-14).

Probe design and synthesis

Padlock probes were designed using the ppDesigner software (13). The average length of the captured region was 100 bp, with the CpG marker located in the central portion of the captured region. Linker sequence between arms contained binding sequences for amplification primers separated by a variable stretch of Cs to produce probes of equal length. We incorporated a 6-bp unique molecular identifier (UMI) sequence in probe design to allow for the identification of unique individual molecular capture events and accurate scoring of DNA methylation levels.

Probes were synthesized as separate oligonucleotides using standard commercial synthesis methods (IDT, San Diego).

Bis-DNA capture

200 ng of bisulfite-converted DNA was mixed with padlock probes in 20 µl reactions containing 1X Ampligase buffer (Epicentre). To anneal probes to DNA, 10-minute denaturation at 95°C was followed by a slow cooling to 55°C. Hybridization was left to complete for 15 hrs at 55°C. To fill gaps between annealed arms, 5µl of the following mixture was added to each reaction: 2µl of HemoKlenTaq polymerase (NEB), 0.5U of Ampligase (Epicentre) and 250 pmol of each dNTP in 1X Ampligase buffer. 5 µl of exonuclease mix (20U of Exo I and 100U of ExoIII, both from

Epicentre) was added and single-stranded DNA degradation was carried out at 37°C for 2 hours, followed by enzyme inactivation for 2 minutes at 94°C.

Circular products of site-specific capture were amplified by PCR with concomitant barcoding of separate samples. Amplification was carried out using primers specific to linker DNA within padlock probes, one of which contained specific 6bp barcodes. Both primers contained Illumina next-generation sequencing adaptor sequences. PCR was done as follows: 1X Phusion Flash Master Mix, 3 µl of captured DNA and 200nM primers, using the following cycle: 10s @ 98°C, 8X of (1s @ 98°C, 5s @ 58°C, 10s @ 72°C), 25X of (1s @ 98°C, 15s @ 72°C), 60s @ 72°C. PCR reactions were mixed and the resulting library was size selected to include effective captures (~230bp) and exclude “empty” captures (~150bp) using Agencourt AMPure XP beads (Beckman Coulter). Libraries were sequenced using MiSeq and HiSeq2500 systems (Illumina) using a paired-end 150bp cycles.

Sequencing data analysis

Mapping of sequencing reads was done using the software tool bisReadMapper with some modifications(13). First, UMI were extracted from each sequencing read and appended to read headers within FASTQ files using a custom script. Reads were on-the-fly converted as if all C were non-methylated and mapped to in-silico converted DNA strands of the human genome, also as if all C were non-methylated, using Bowtie2(15). Methylation frequencies were calculated for all CpG dinucleotides contained within the regions captured by padlock probes by dividing the numbers of unique reads carrying a C at the interrogated position by the total number of reads covering the interrogated position.

Identification of Methylation Correlated Block (MCB)

In order to maximize the ability to measure small differences in DNA methylation we took advantage of the notion that closely positioned CpG tend to have similar methylation levels, what is believed to be a result of the processivity and lack of sequence-specificity of DNA methyltransferases and demethylases, as well as the concept of haplotype blocks in genetic linkage analysis (16). To investigate whether this indeed is evident in our data, we calculated Pearson correlation coefficients r^2 between β values of any two CpGs positioned within one kilobase of one another. We used a cutoff of $r^2 > 0.5$ to identify Methylation Correlated Block (MCB) within regions interrogated by our padlock probes. A value of Pearson's $r < 0.5$ was used to identify

transition spots (boundaries) between any two adjacent markers indicating uncorrelated methylation. Markers not separated by a boundary were combined into Methylation Correlated Block (MCB). This procedure identified a total of 3600 BCMs in each diagnostic category within our padlock data, combining between 2 and 22 CpG positions in each block. Methylation frequencies for entire MCBs were calculated by summing up the numbers of Cs at all interrogated CpG positions within a MCB and dividing by the total number of C+Ts at those positions.

Pearson correlation coefficients between methylation frequencies of each pair of CpG markers separated by no more than 200bp were calculated separately from 30 cancer and 30 corresponding normal tissue samples from each of the two diagnostic categories, ie normal liver and HCC. A value of Pearson's $r < 0.5$ was used to identify transition spots (boundaries) between any two adjacent markers indicating uncorrelated methylation. Markers not separated by a boundary were combined into Methylation Correlated Block (MCB). This procedure identified a total of ~1550 MCB in each diagnostic category within our padlock data, combining between 2 and 22 CpG positions in each block. Methylation frequencies for entire BCMs were calculated by summing up the numbers of Cs at all interrogated CpG positions within a BCM and dividing by the total number of C+Ts at those positions.

Linking differentially methylated markers to gene expression

TCGA DNA methylation and RNAseq expression data for LIHC samples was obtained from the TCGA website as above. Analysis to link differentially methylated markers to gene expression was performed as in Yao et al (17). Briefly, the degree of DNA methylation at each CpG was denoted as a beta value and was calculated as $(M/(M+U))$, where M and U are normalized values representing the methylated and unmethylated allele intensities respectively. Beta values range from 0 to 1 and reflect the fraction of methylated alleles at each CpG in each sample.

The methylation beta value was calculated for all 485,000 markers for each of the LIHC tumor and matched normal liver tissues in the TCGA data. CpG markers with a mean value less than .05 or greater than .95 were selected for further evaluation. We also selected markers with a difference between the mean methylation value for the tumor tissue and the mean methylation value of the corresponding normal tissue of greater than 0.5. At the intersection of these two groups, markers for which the mean methylation was $< .05$ for normal liver tissue samples and the difference between normal and tumor was greater than 0.5 were further selected and the genes associated with these markers were identified. For each marker, the tumor samples were then separated into

those with methylation values greater than the mean value of the tumor samples and those with methylation values less than the mean value of the tumor samples.

Next, we examined the RNAseq data in the TCGA data and calculated the relative expression of each gene. Because of the wide variation of the expression values, we adjusted the values as follows: $\log_2(\text{expressionValue} + 1)$

We then identified genes in which the difference in the methylation values correlated with variation in the associated gene expression levels. Genes for which there was a correlation were selected for further functional evaluation and validation.

Validation of genes differentially expressed in human LIHC in mouse LIHC.

Data on human genes with a good correlation between methylation and differential expression in LIHC versus normal liver tissue were compared to search for the mouse counterpart in a dataset previously reported(18). Fold change of expression comparing LIHC versus normal liver tissue were investigated and presented (supplemental Fig 3 and supplemental Table 7)

DNA/RNA isolation and Quantitative PCR

Tumor and corresponding far site samples of the same tissue were obtained from patients who underwent surgical tumor resection; samples were frozen and preserved at -80°C until use. Isolation of DNA and RNA from samples was performed using AllPrep DNA/RNA Mini kit (Qiagen, Valencia, CA) according to the manufacturer's recommendations. During RNA isolation, the sample was subjected to on-column DNase digestion. RNA was quantified using a Nanodrop 2000 (Thermo Scientific). 200ng RNA of each sample was used for cDNA synthesis using iScript cDNA synthesis kit (Bio-rad, Inc) according to the manufacturer's instructions. qPCR was performed by a standard 40-cycle amplification protocol using gene-specific primers (Supplemental Table 4) and a Power SYBR Green PCR Master Mix on a 7500 Real Time PCR system (Applied Biosystems). Experiments were carried out in triplicate and normalized to endogenous *ACTB* levels. Relative fold change in expression was calculated using the $\Delta\Delta\text{CT}$ method (cycle threshold values <30). Data are shown as Mean \pm SD based on three replicates.

Cell culture and gene transfections

Human liver cancer cell line HEP1 and the human embryonic kidney cell line HEK293a were obtained from American type culture collection (Manassas, VA, USA) and cultured according to their instructions. The expression construct for *FUZ* were purchased from Origene in a form of TrueORF® cDNA clones in pCMV6-Entry vector. cDNAs were shuttled into pLenti-C-mGFP (Origene) to create a lentivector encoding a fusion protein between the desired gene and mGFP. Lentiviral particles were made by co-transfection of HEK-293T cells with a *FUZ*-mGFP lentivector together with a third-generation packaging vector using calcium phosphate precipitation. Viral supernatants were collected 36 hours post-transfection. Human liver cancer cell line HEP1 were plated in a 6-well plate the day before transfection of *FUZ*-mGFP lentivector. Stable cell lines were generated by infecting cells with viral particles at the MOI of ~5 for 24hrs and collected and sorted to 100% GFP – positivity using FACS. The GFP positive cells were then used for a colony formation assay in cell culture and tumor xenograft in nude mice.

Colony formation assays

Cells were plated in 6-well plates at a density of 500 cells per well and cultured at 37°C with 5% CO₂ humidified air for 14 days. The colonies were fixed with 10% formaldehyde for 5 min and then stained with 0.1% crystal violet for 30 seconds. Colony consisting of 50 or more cells were counted. The experiment was performed in triplicate and repeated 3 times. Plate efficiency = (colony numbers /inoculated cell numbers) × 100%.

Tumor xenograft

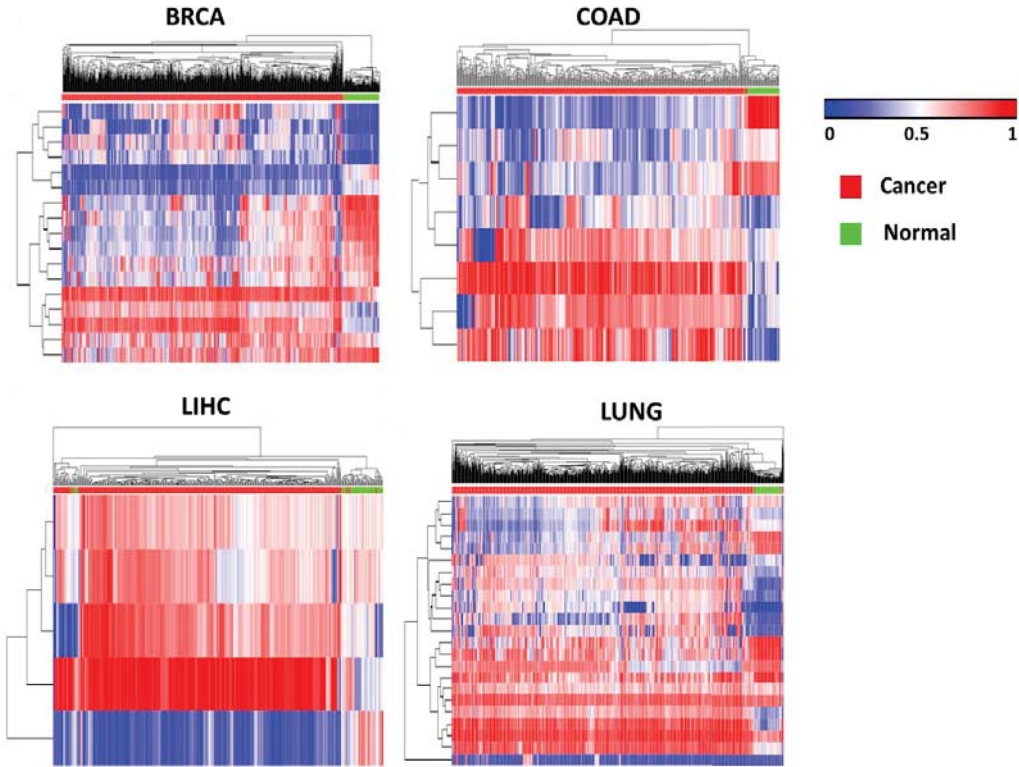
All animal studies were performed in accordance with institutional and international animal regulations. Animal protocols were approved by the Institutional Animal Care and Use Committee of Sun Yat- Sen University. Female athymic BALB/c nude mice (4–5 weeks of age, 18–20 g) were purchased from a vendor (Guangdong Province Laboratory Animal Center, Guangzhou, China). Mice were injected subcutaneously with 100 µl of tumor cells suspended in serum free medium. Tumor growth was monitored every 3 days by visual examination. Tumor sizes were measured using a caliper, and tumor volume was calculated according to the following equation: tumor volume (mm³) = (length (mm) × width (mm)²) × 0.5. All animals were sacrificed 3-4 weeks post-injection and the xenografts were harvested. Representative data were obtained from five mice per

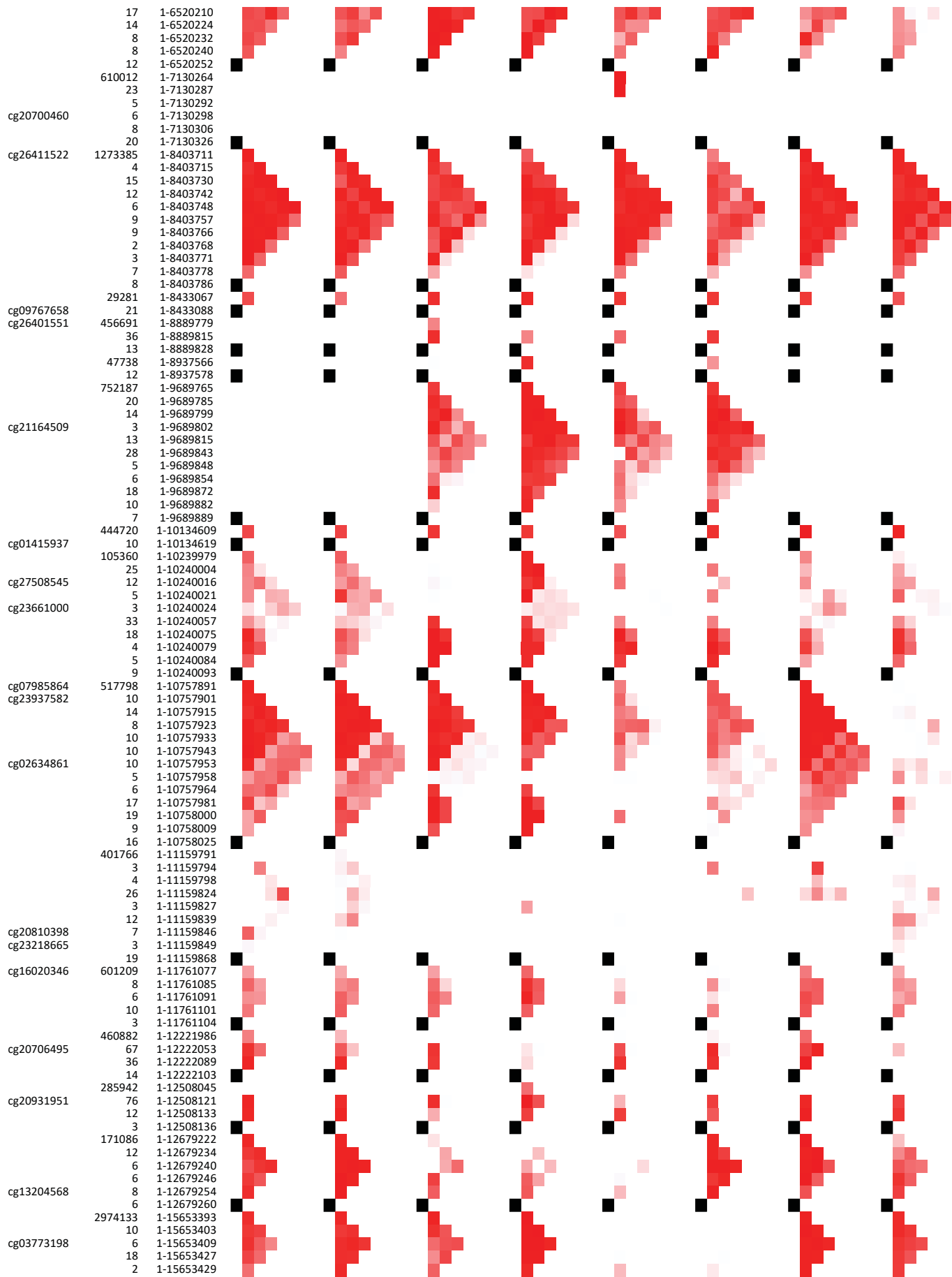
experimental group. Statistical analyses were performed with one-way repeated- measures ANOVA.

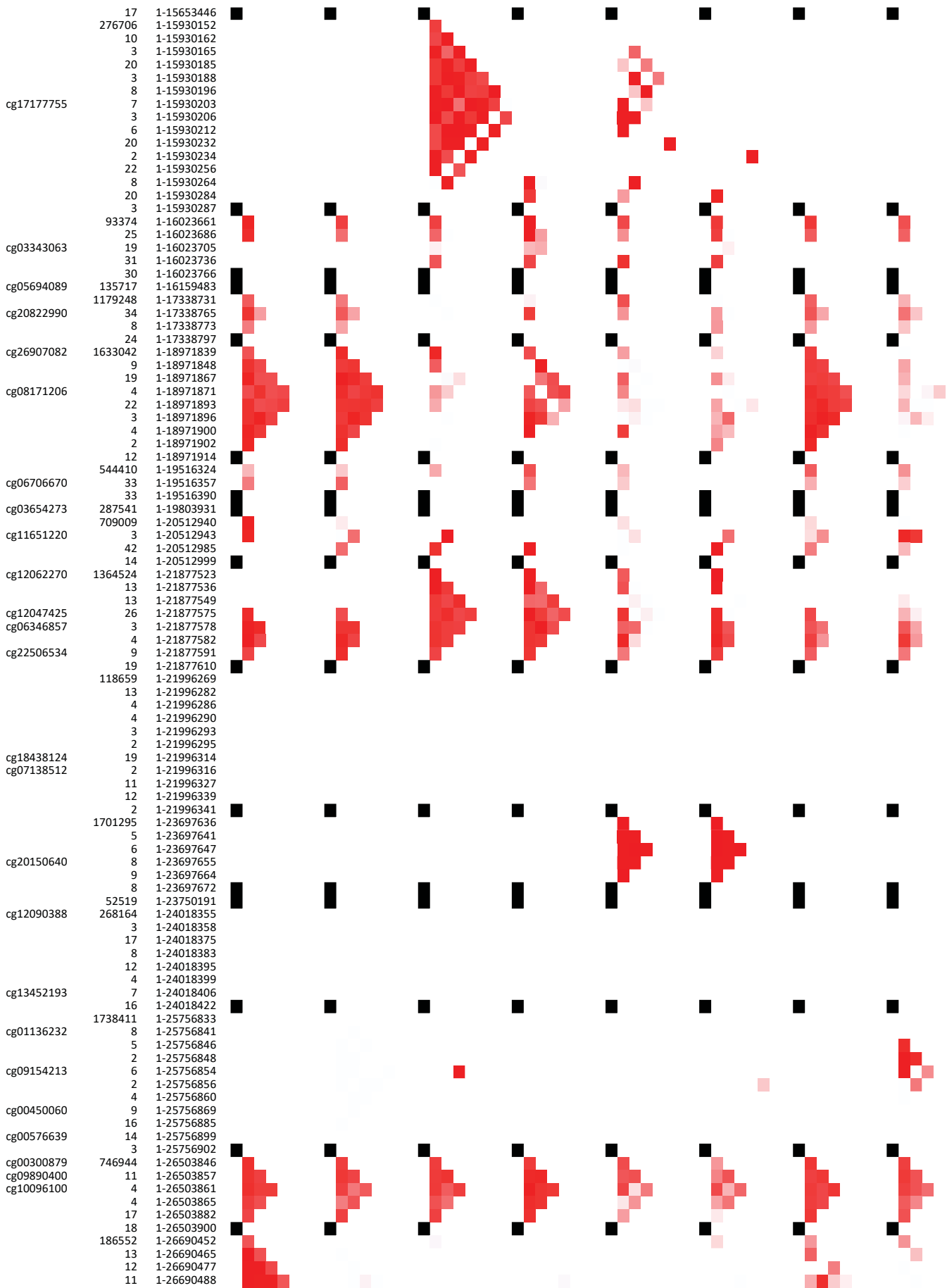
Reference

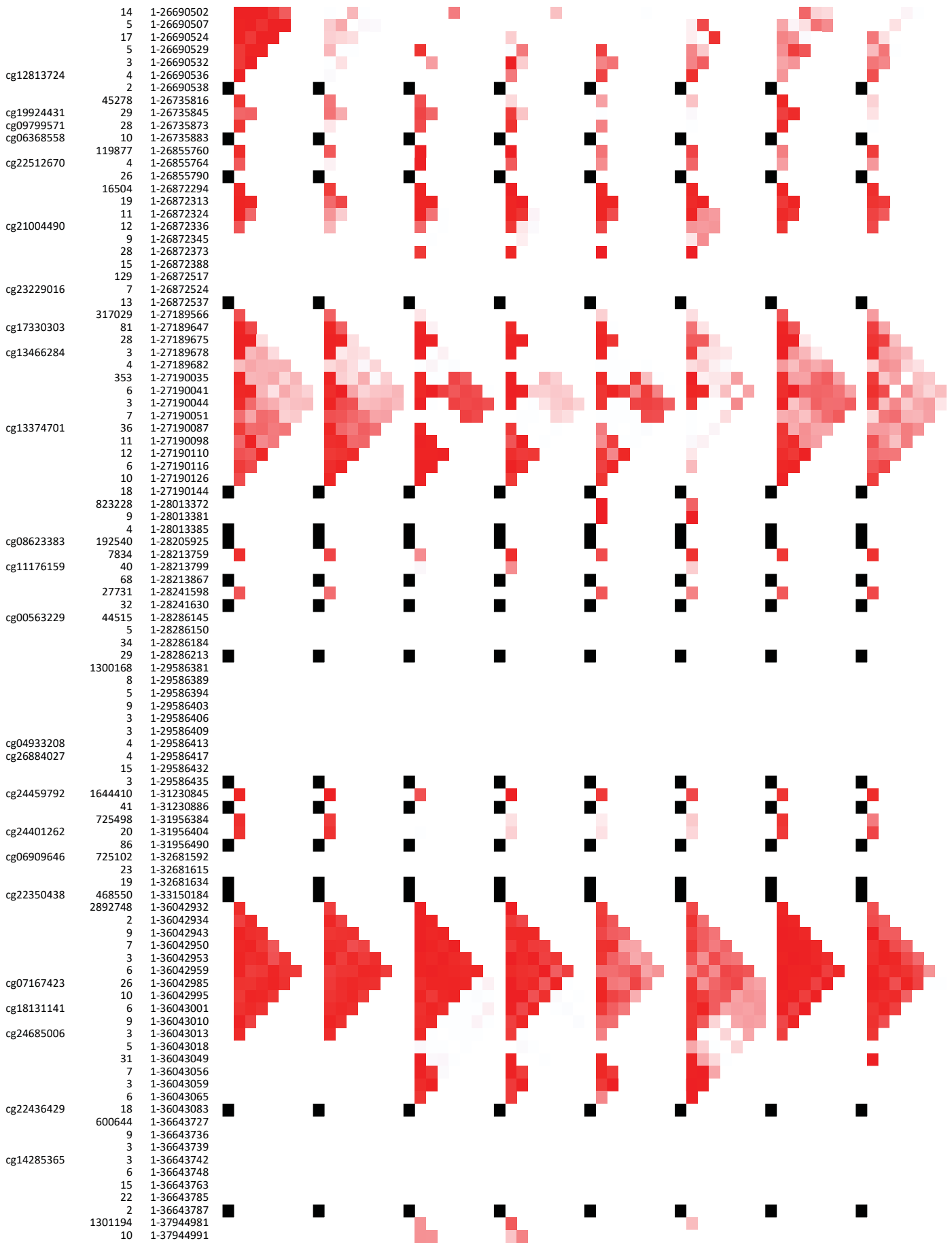
1. Berkeley C (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *E-book available at <http://www.bepress.com/sagmb/vol3/iss1/art3>. [PubMed].*
2. Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*:289-300.
3. Friedman J, Hastie T, & Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33(1):1.
4. Wasserman L & Roeder K (2009) High dimensional variable selection. *Annals of statistics* 37(5A):2178.
5. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*:267-288.
6. Zou H & Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301-320.
7. Witten DM & Tibshirani R (2008) Testing significance of features by lassoed principal components. *The annals of applied statistics* 2(3):986.
8. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of statistics*:1189-1232.
9. O'Quigley J, Xu R, & Stare J (2005) Explained randomness in proportional hazards models. *Statistics in medicine* 24(3):479-489.
10. Harrell FE, Lee KL, & Mark DB (1996) Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* 15:361-387.
11. Meinshausen N, Meier L, & Bühlmann P (2012) P-values for high-dimensional regression. *Journal of the American Statistical Association*.
12. Porreca GJ, *et al.* (2007) Multiplex amplification of large sets of human exons. *Nature methods* 4(11):931-936.
13. Diep D, *et al.* (2012) Library-free methylation sequencing with bisulfite padlock probes. *Nat Methods* 9(3):270-272.
14. Deng J, *et al.* (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature biotechnology* 27(4):353-360.
15. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357-359.
16. Wall JD & Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics* 4(8):587-597.
17. Yao L, Shen H, Laird PW, Farnham PJ, & Berman BP (2015) Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol* 16:105.
18. He G, *et al.* (2013) Identification of liver cancer progenitors whose malignant progression depends on autocrine IL-6 signaling. *Cell* 155(2):384-396.

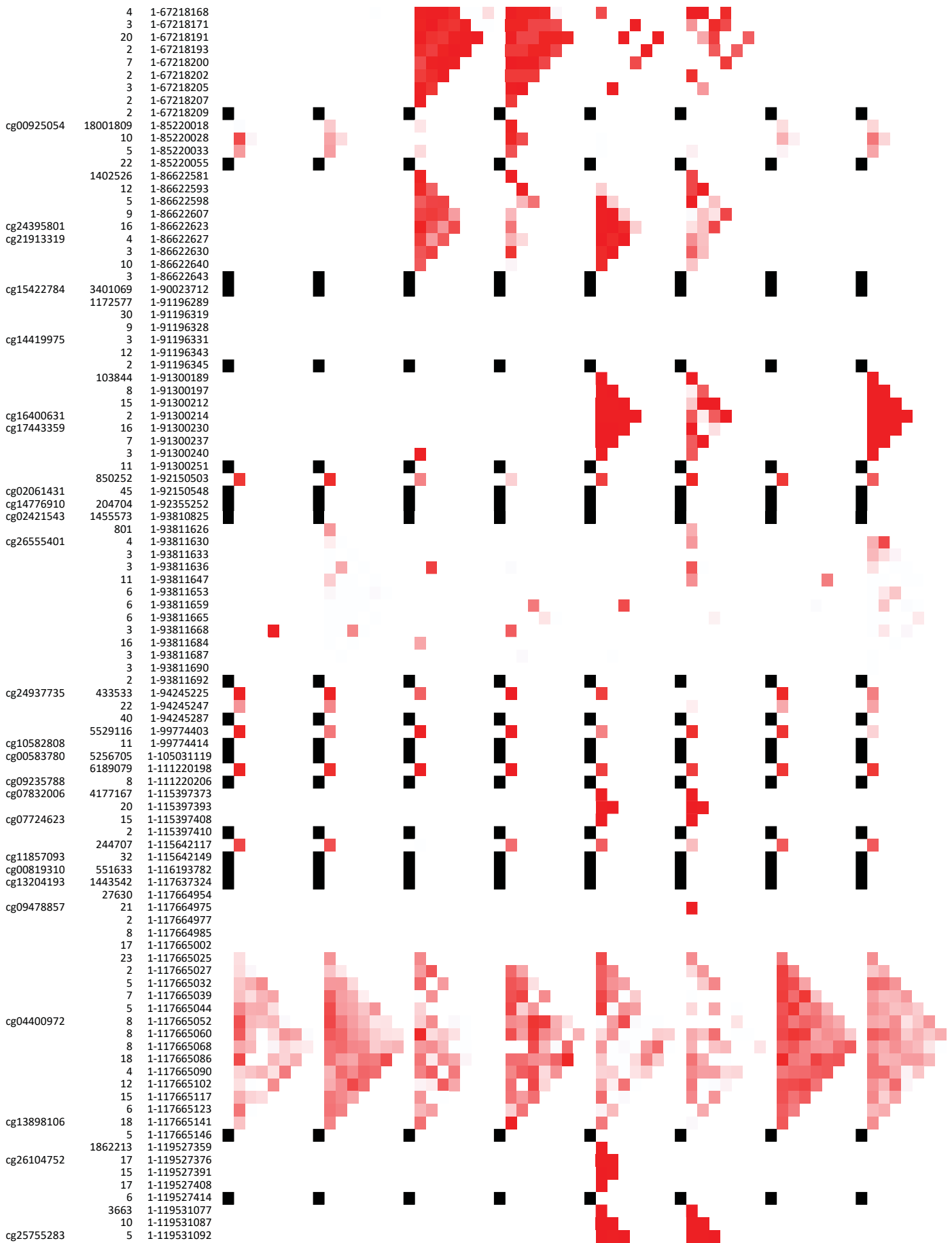
Suppl Fig 1

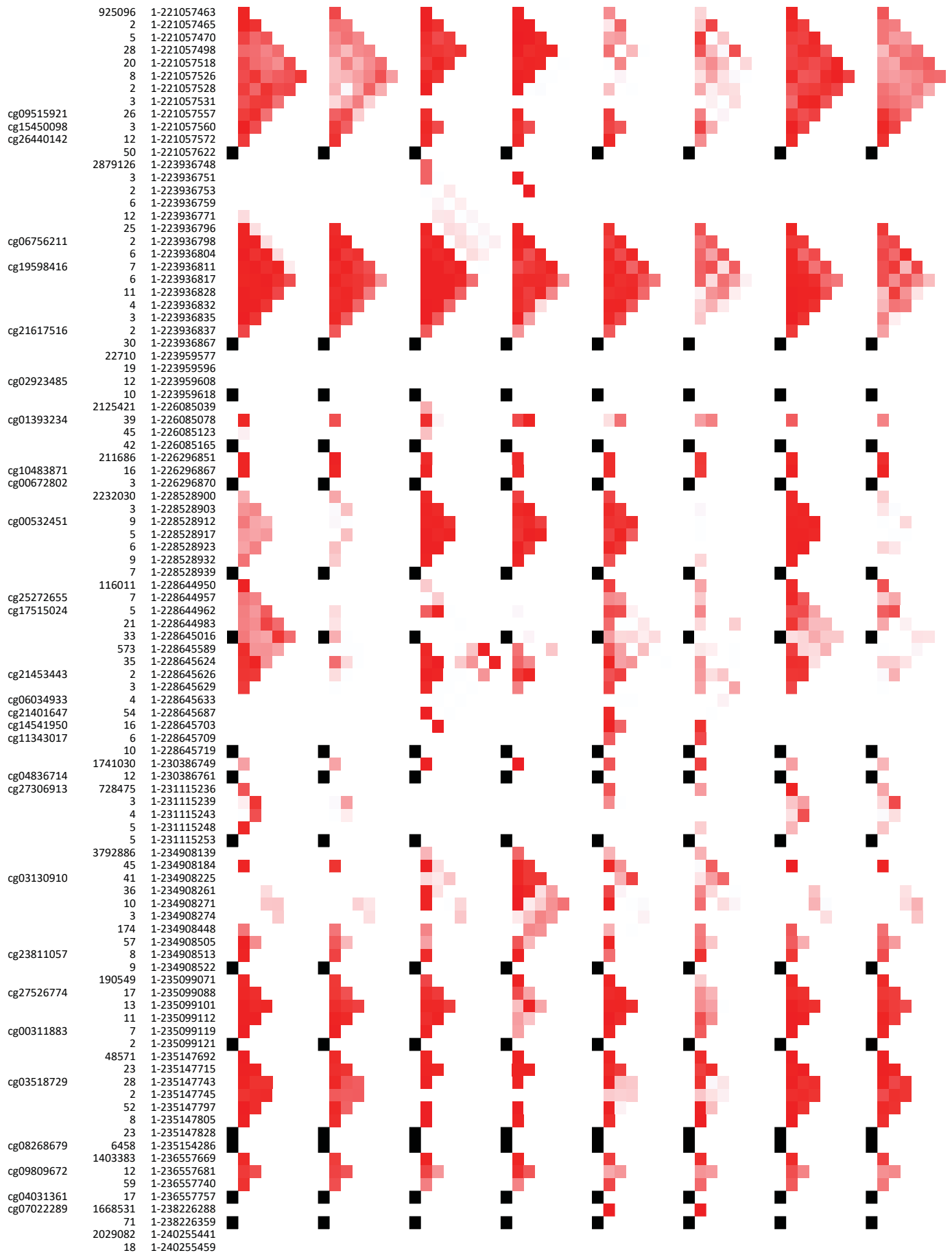


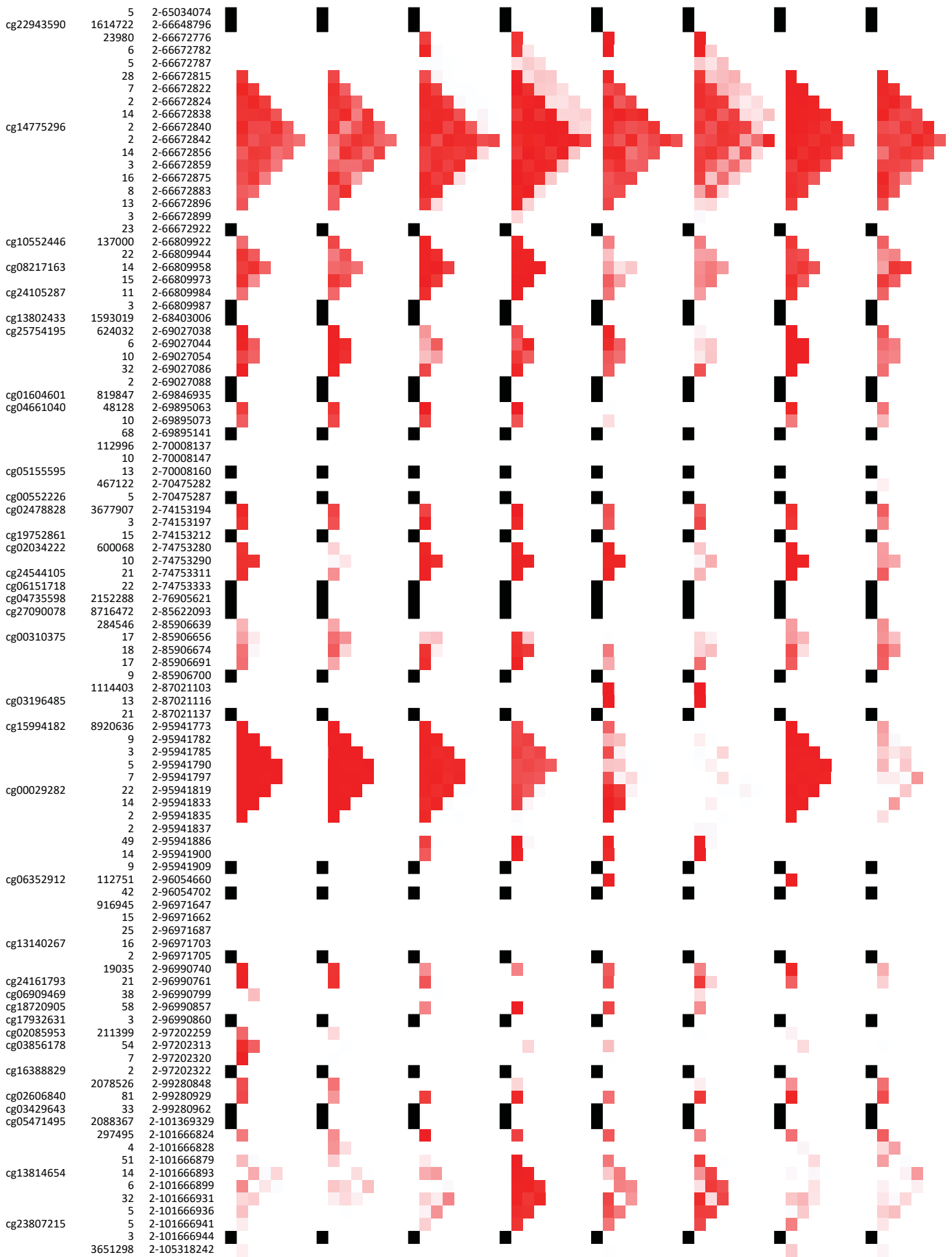


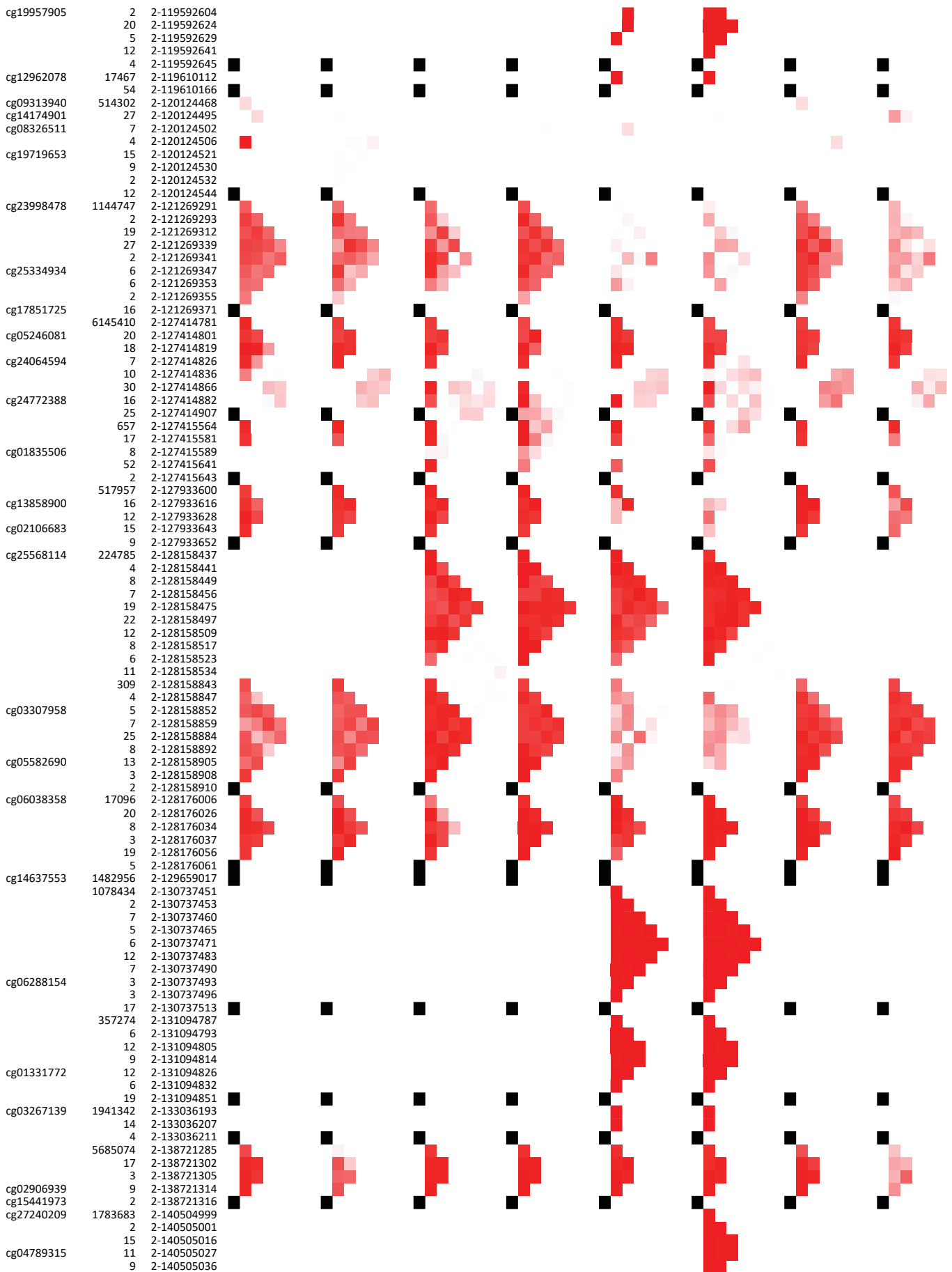


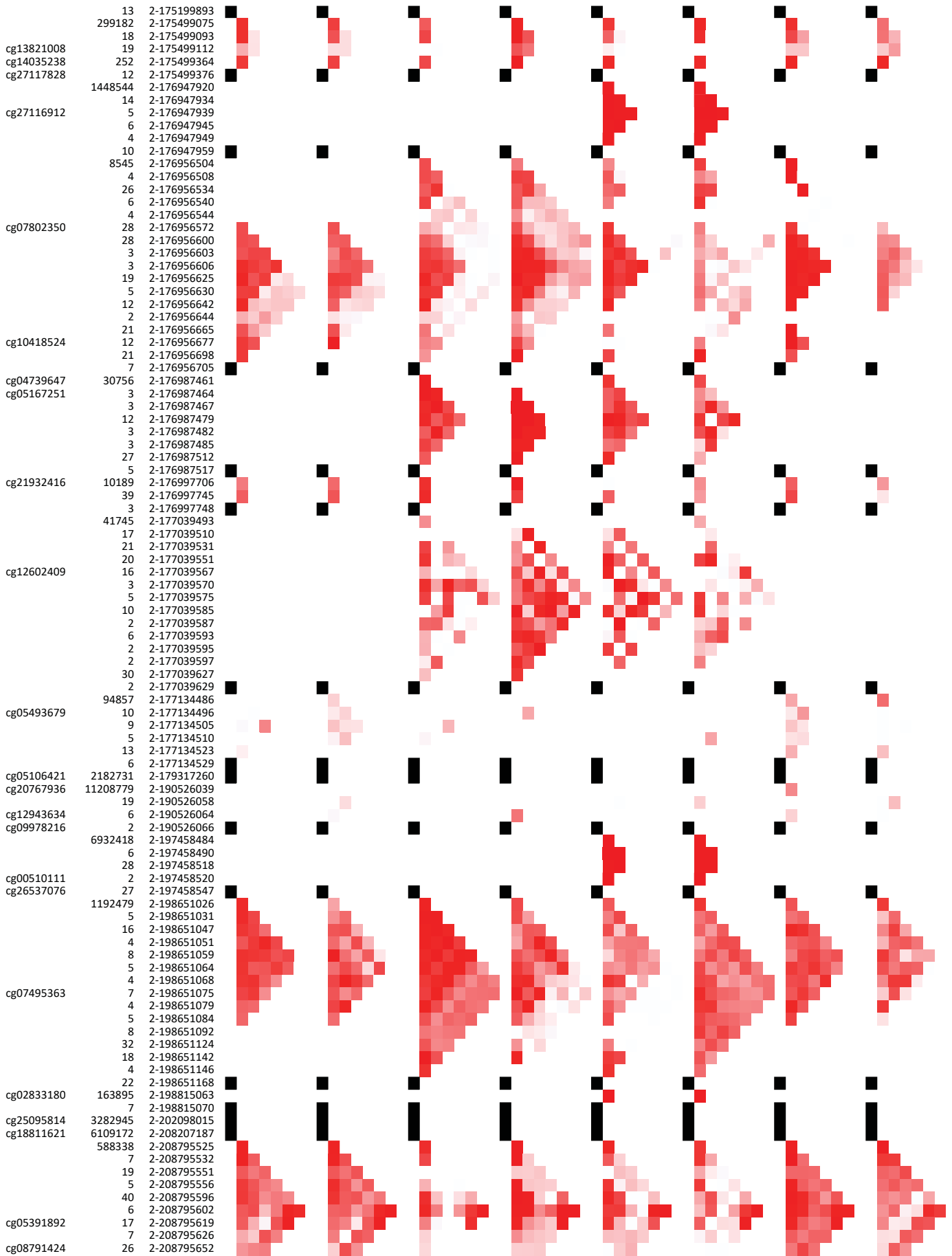


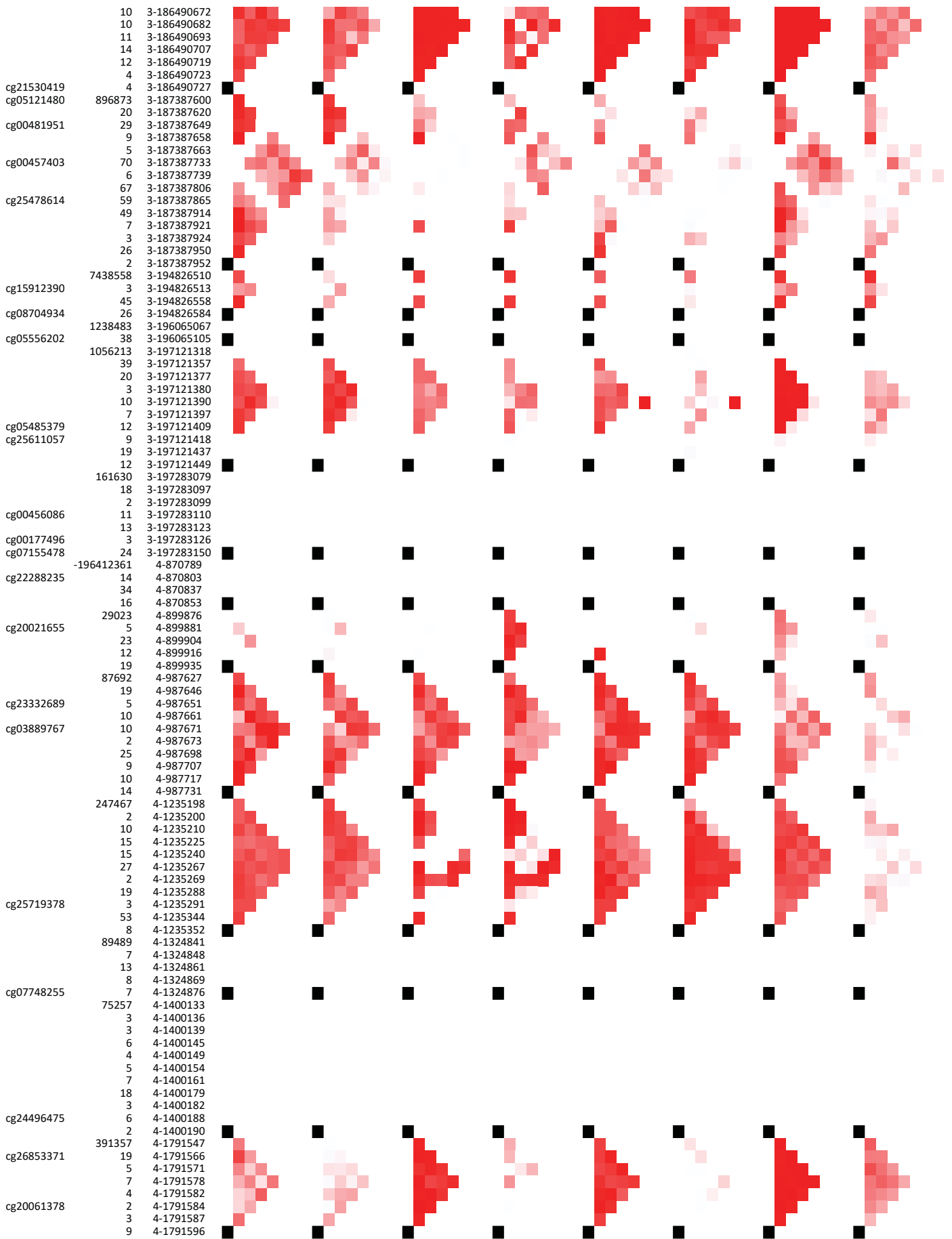


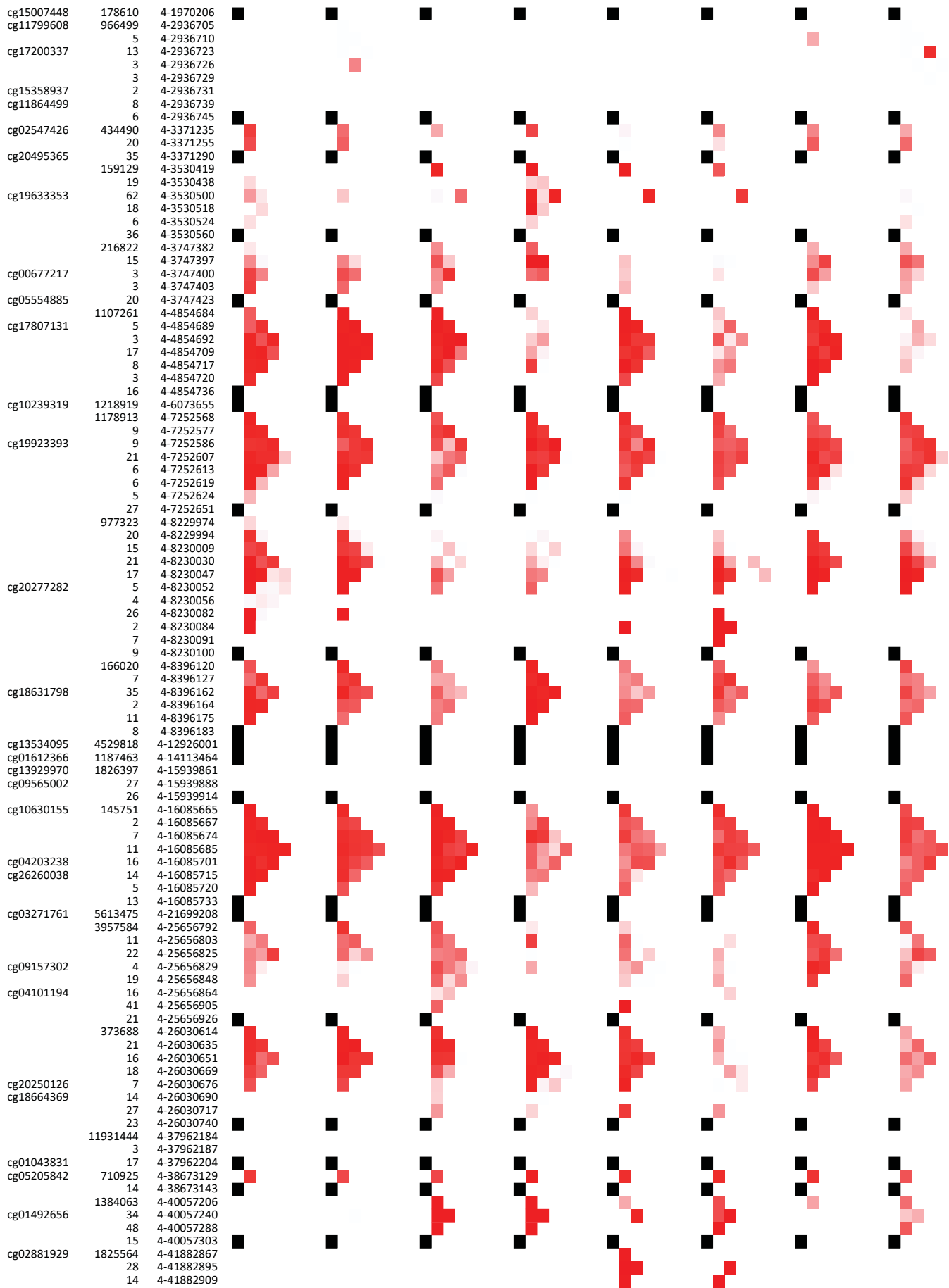


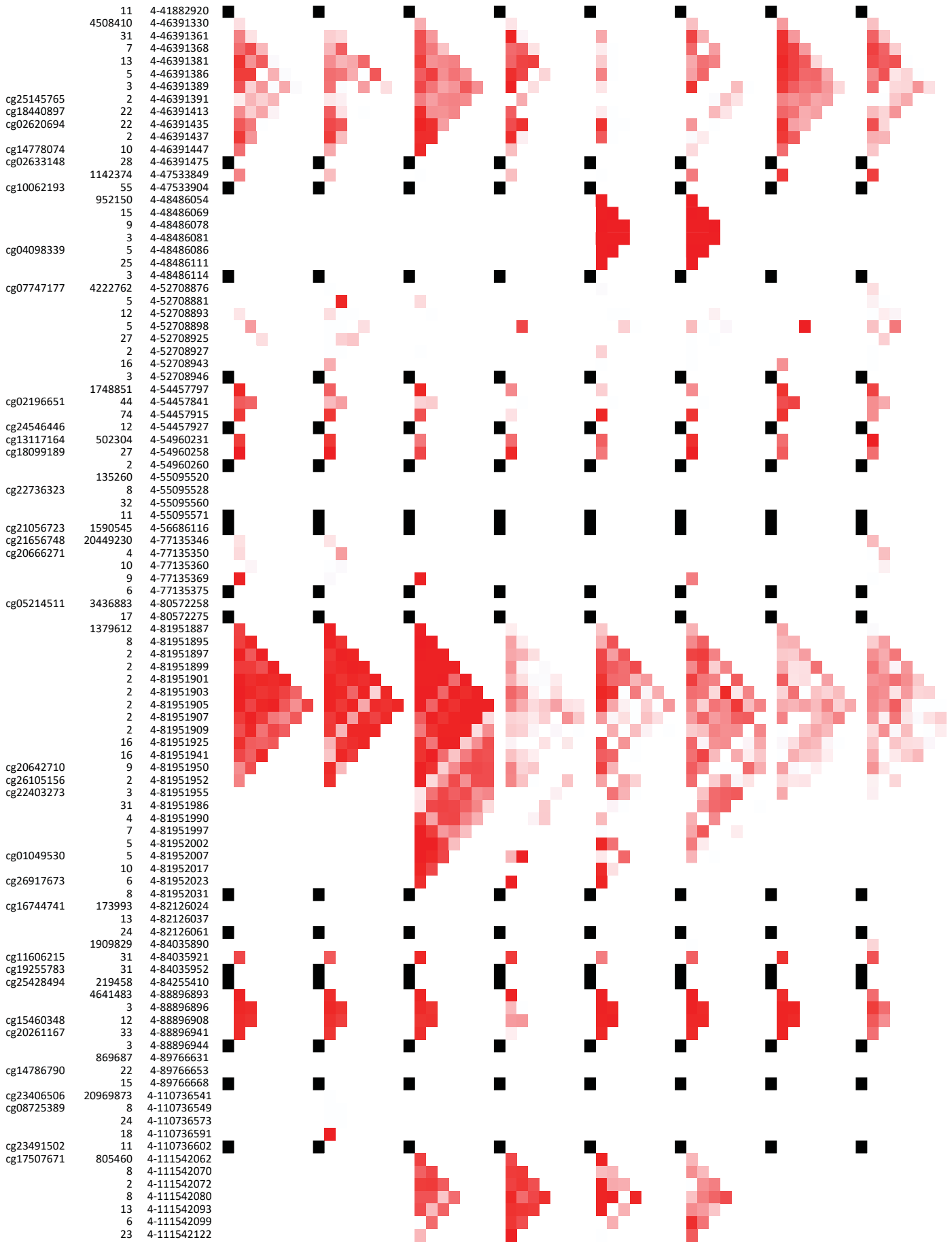


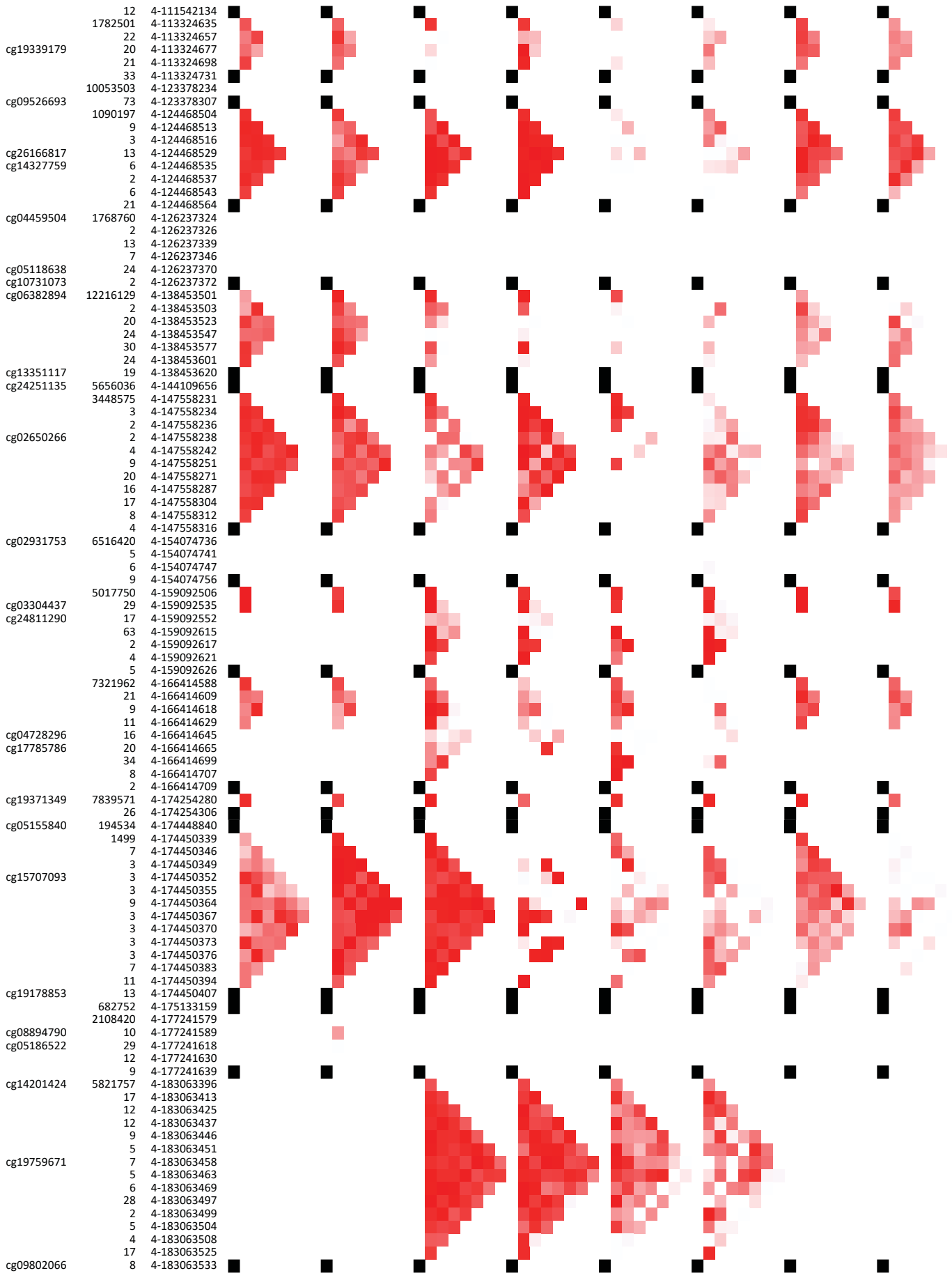


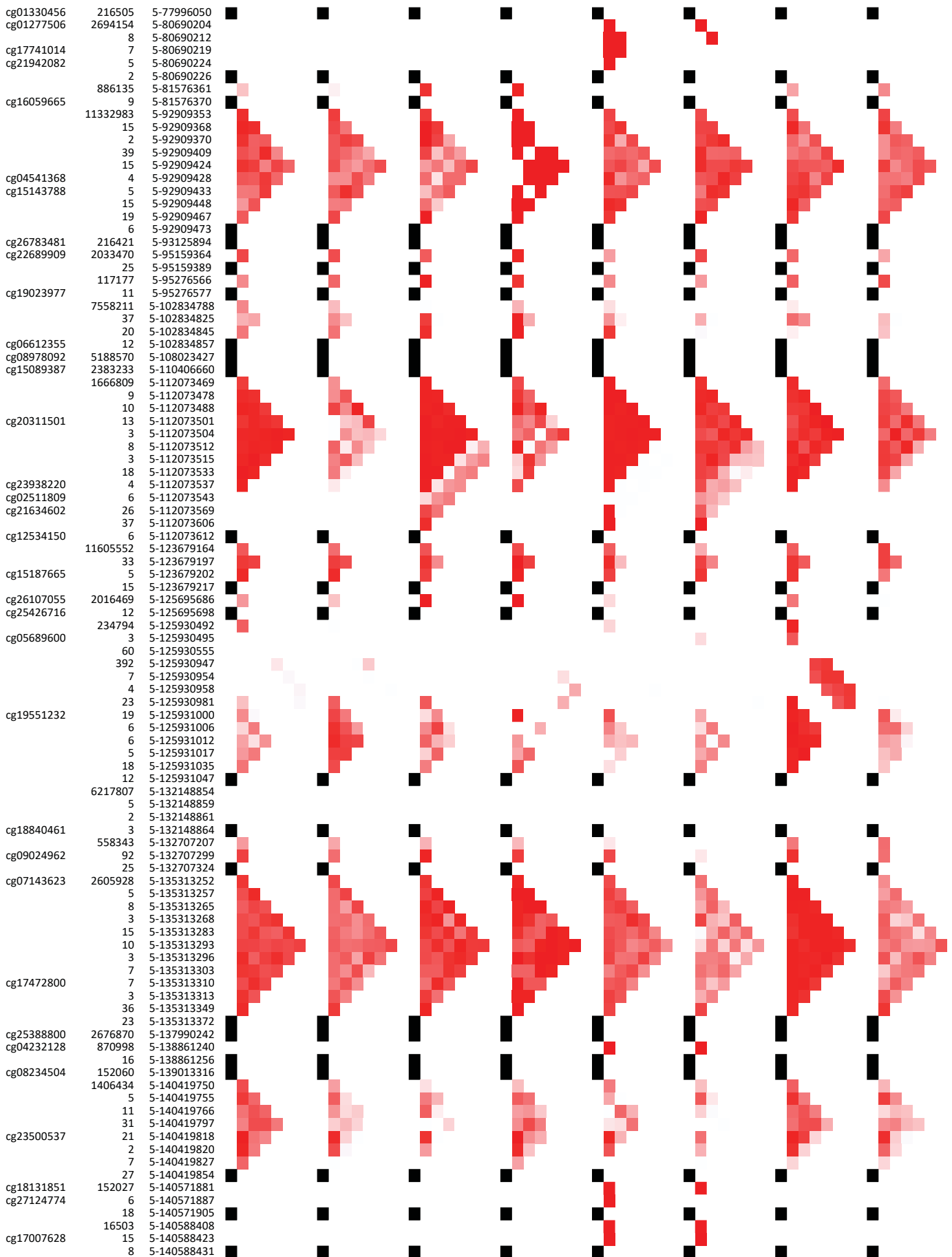


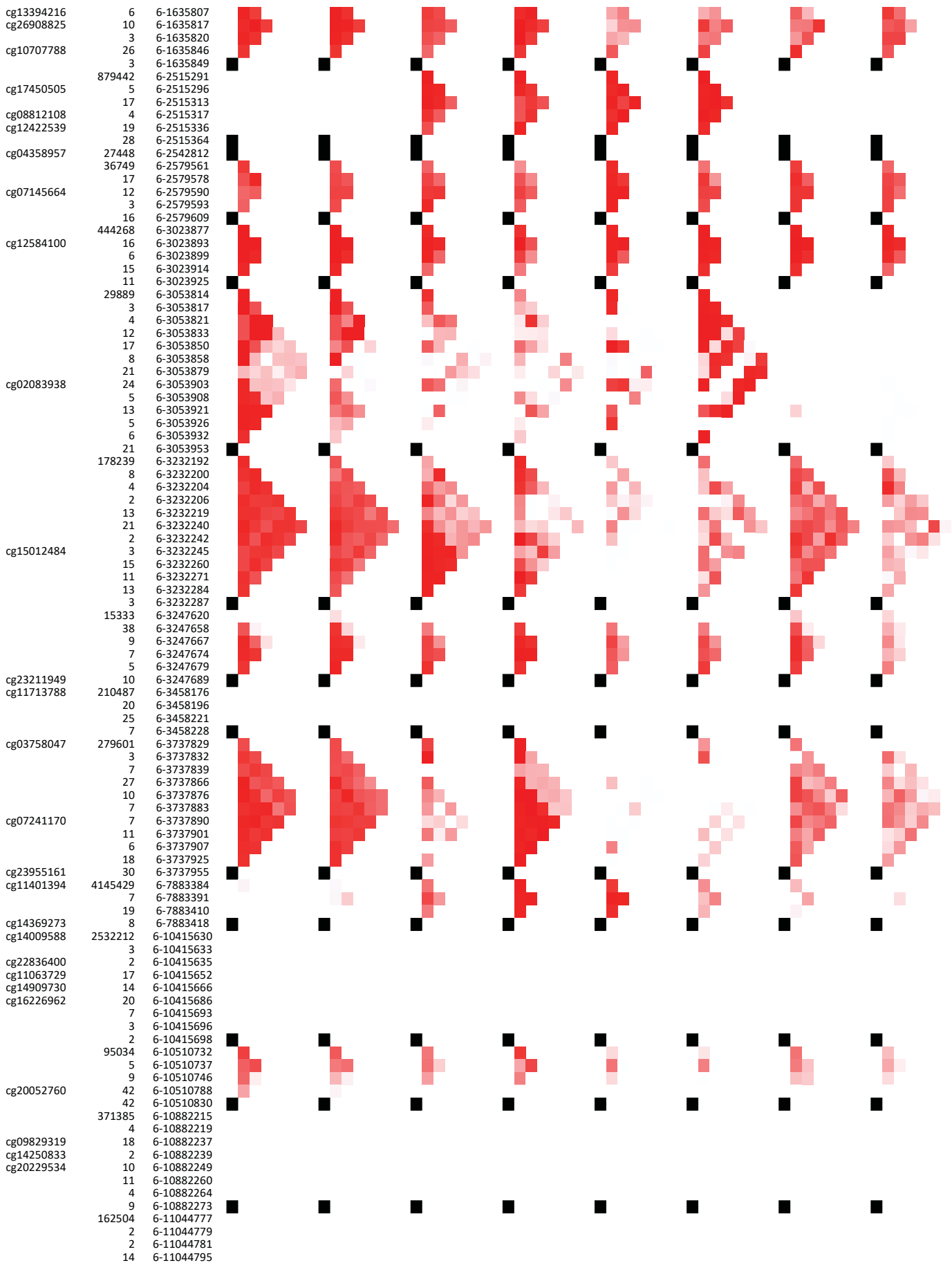


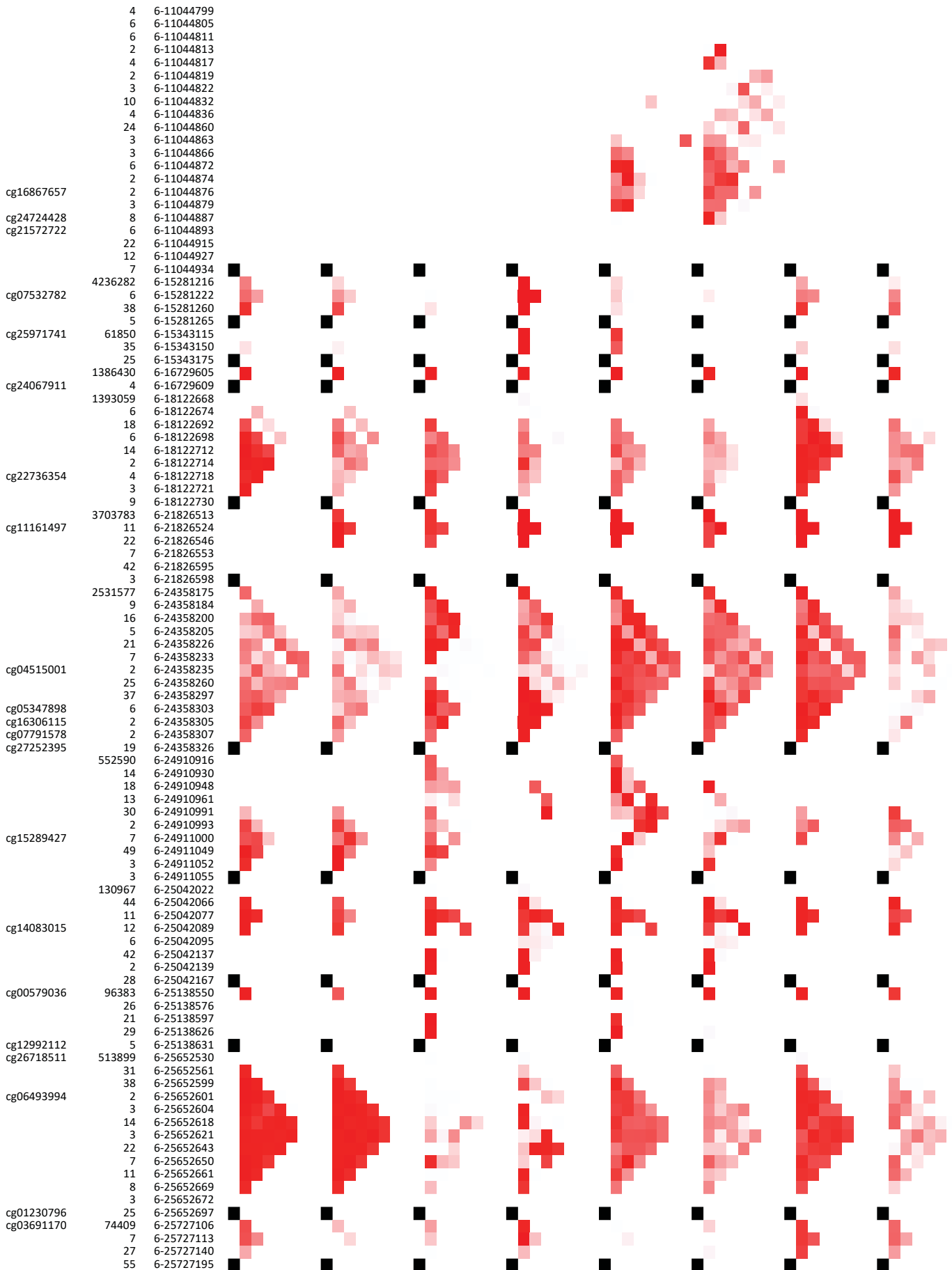


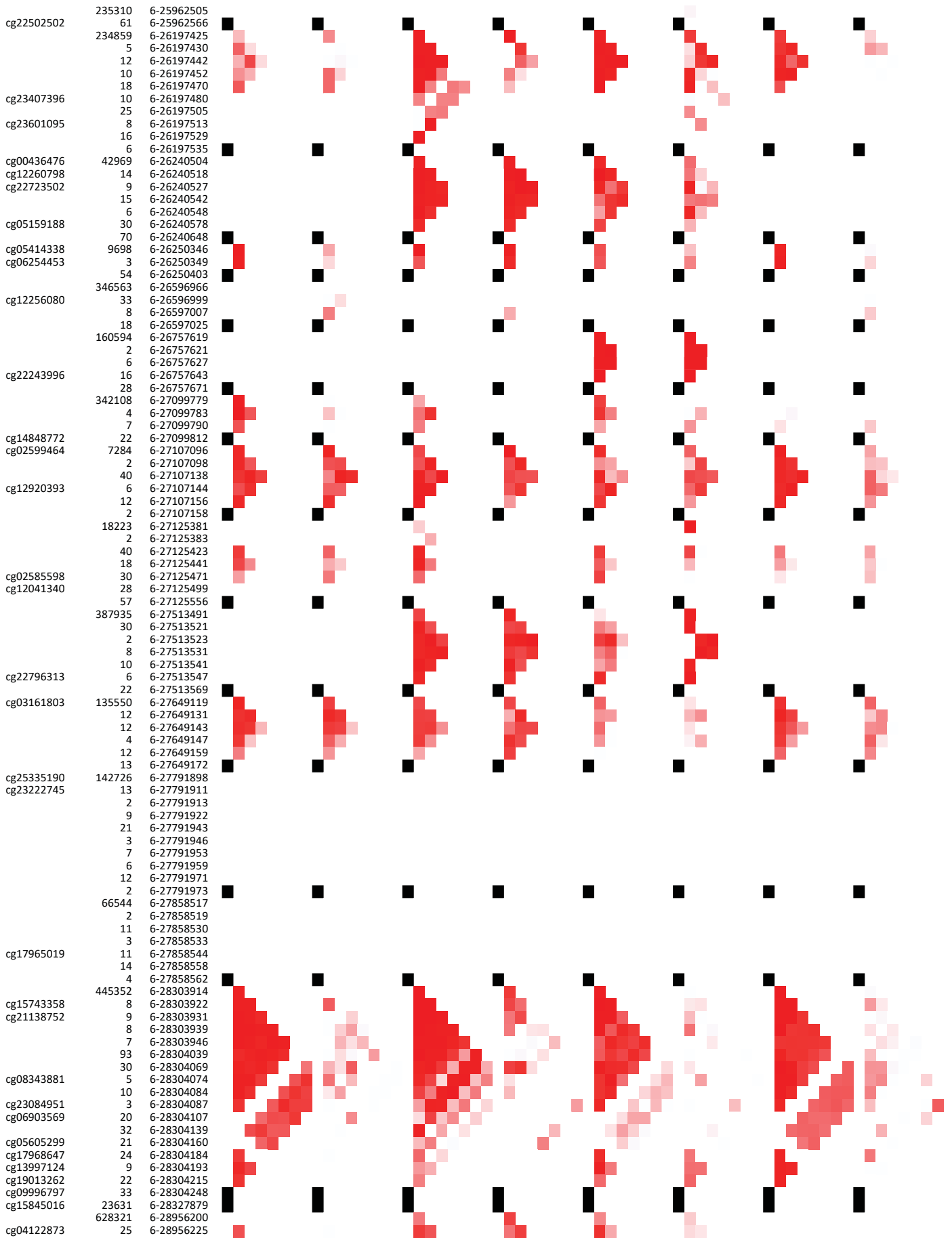


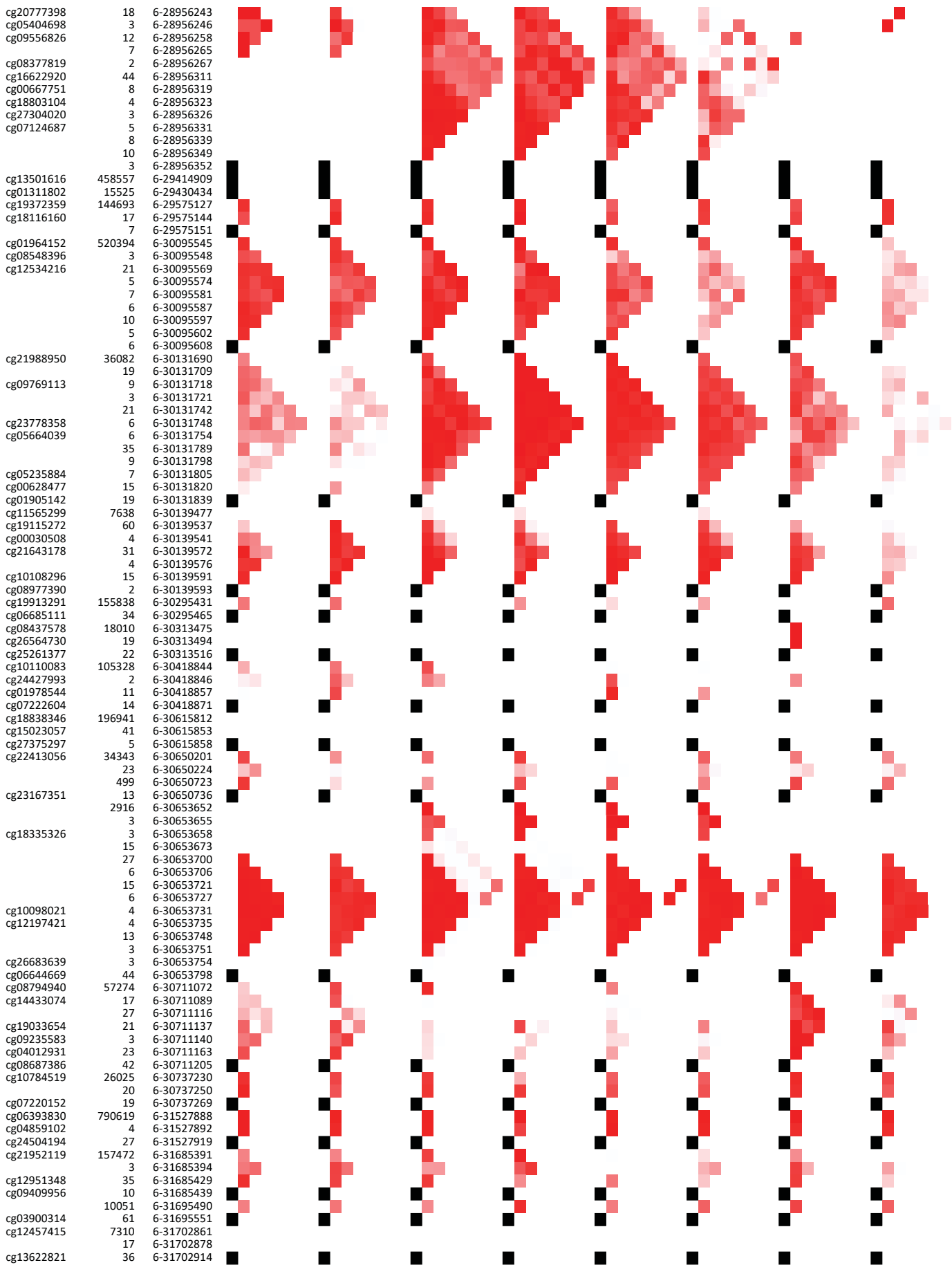


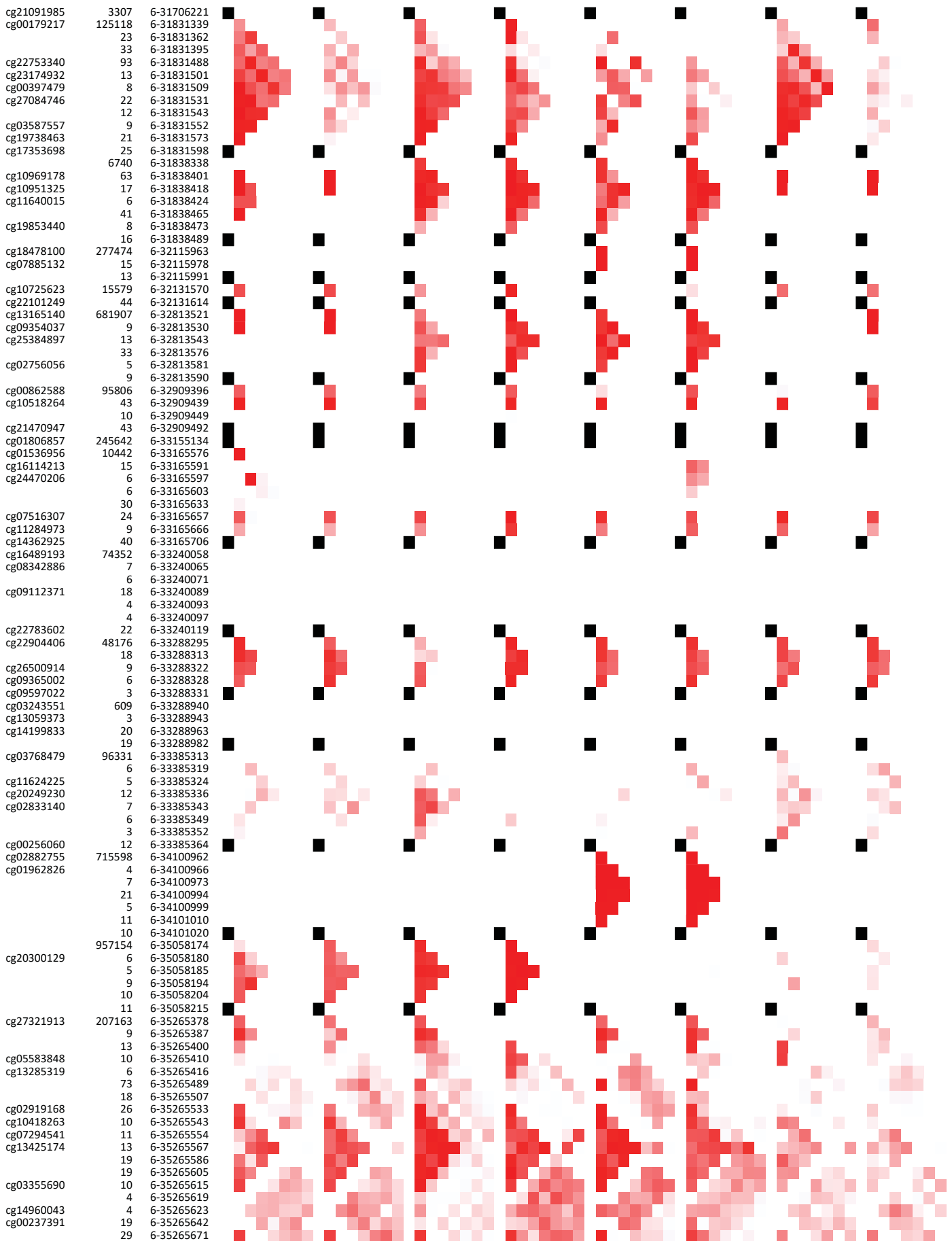


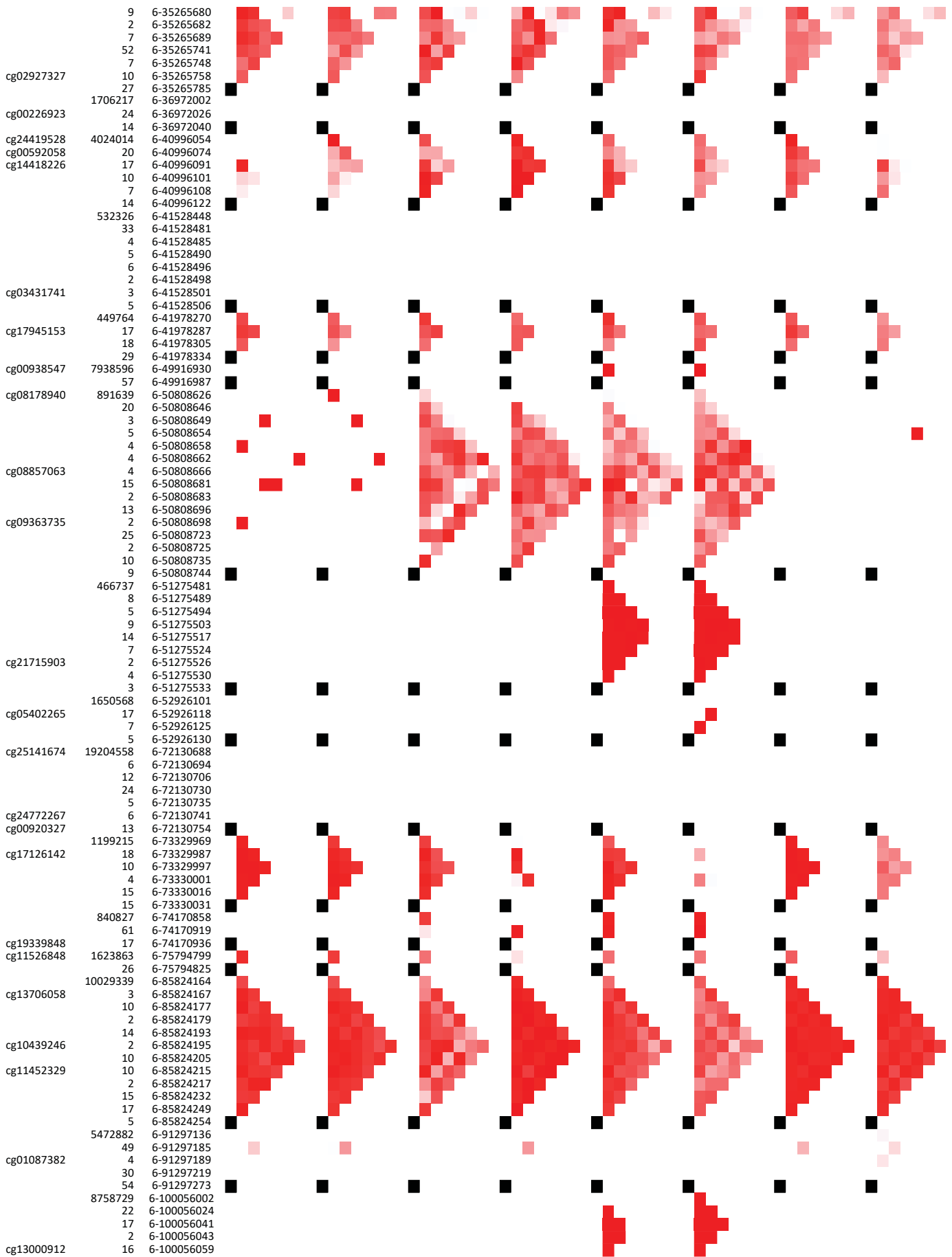


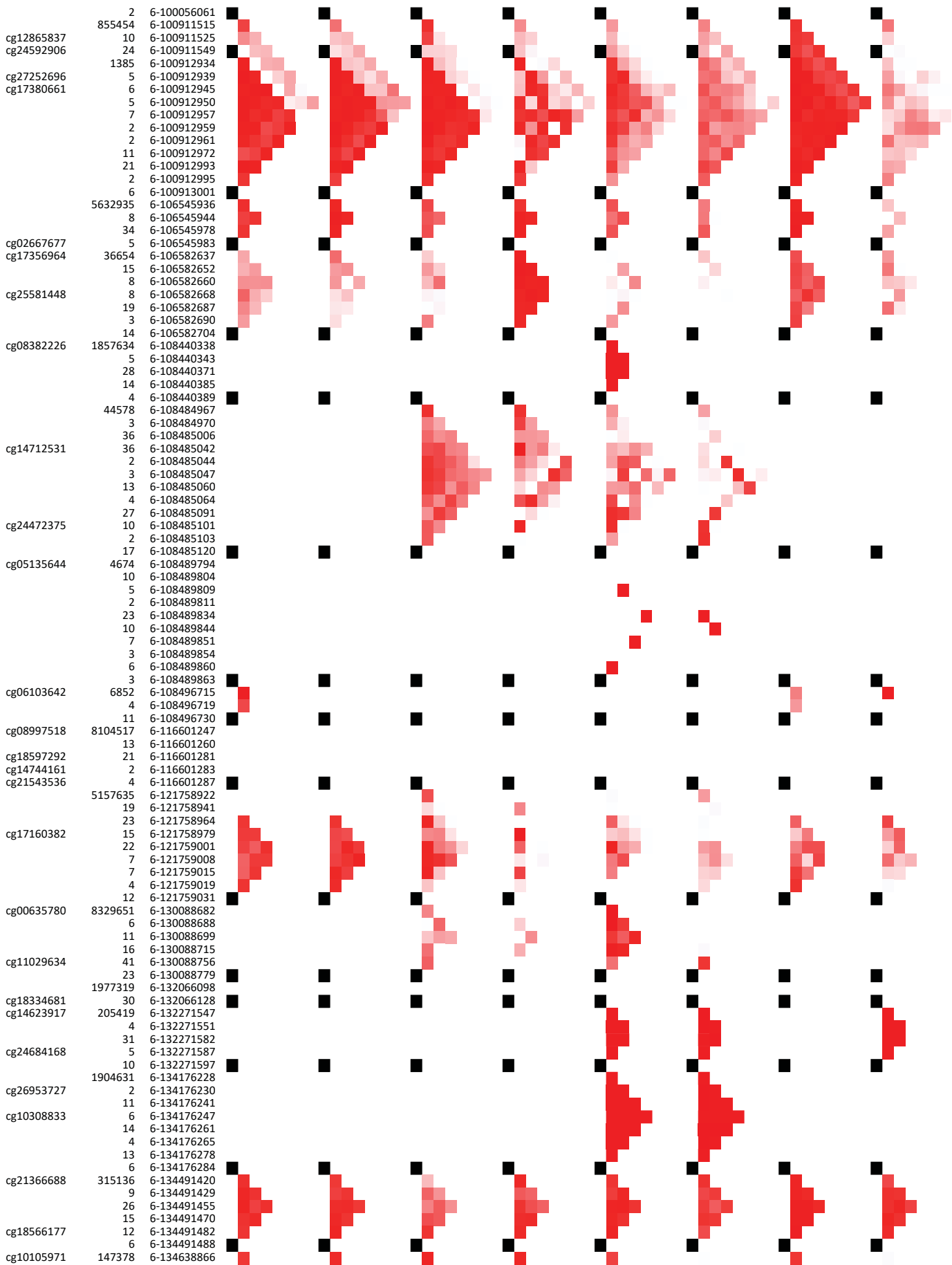


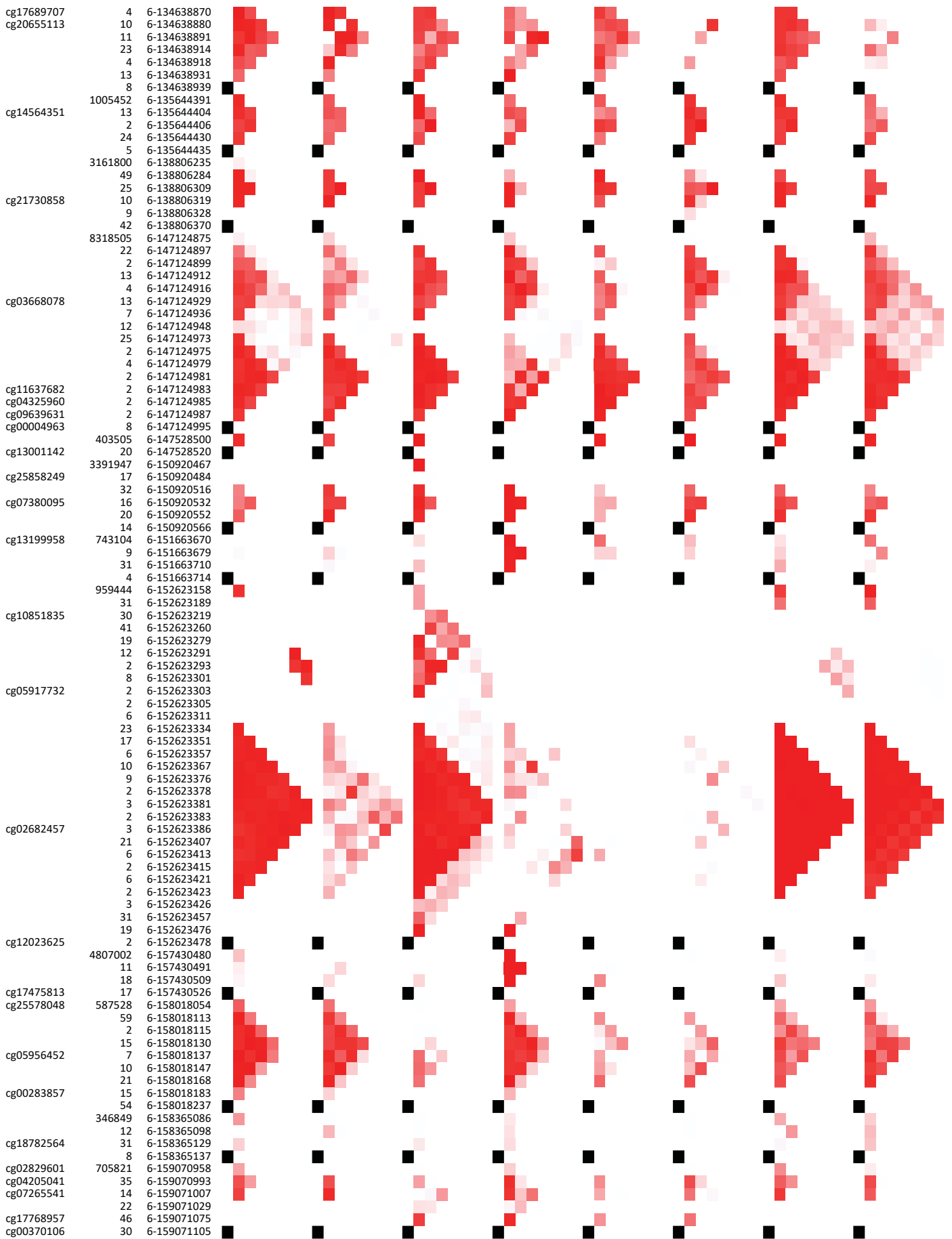


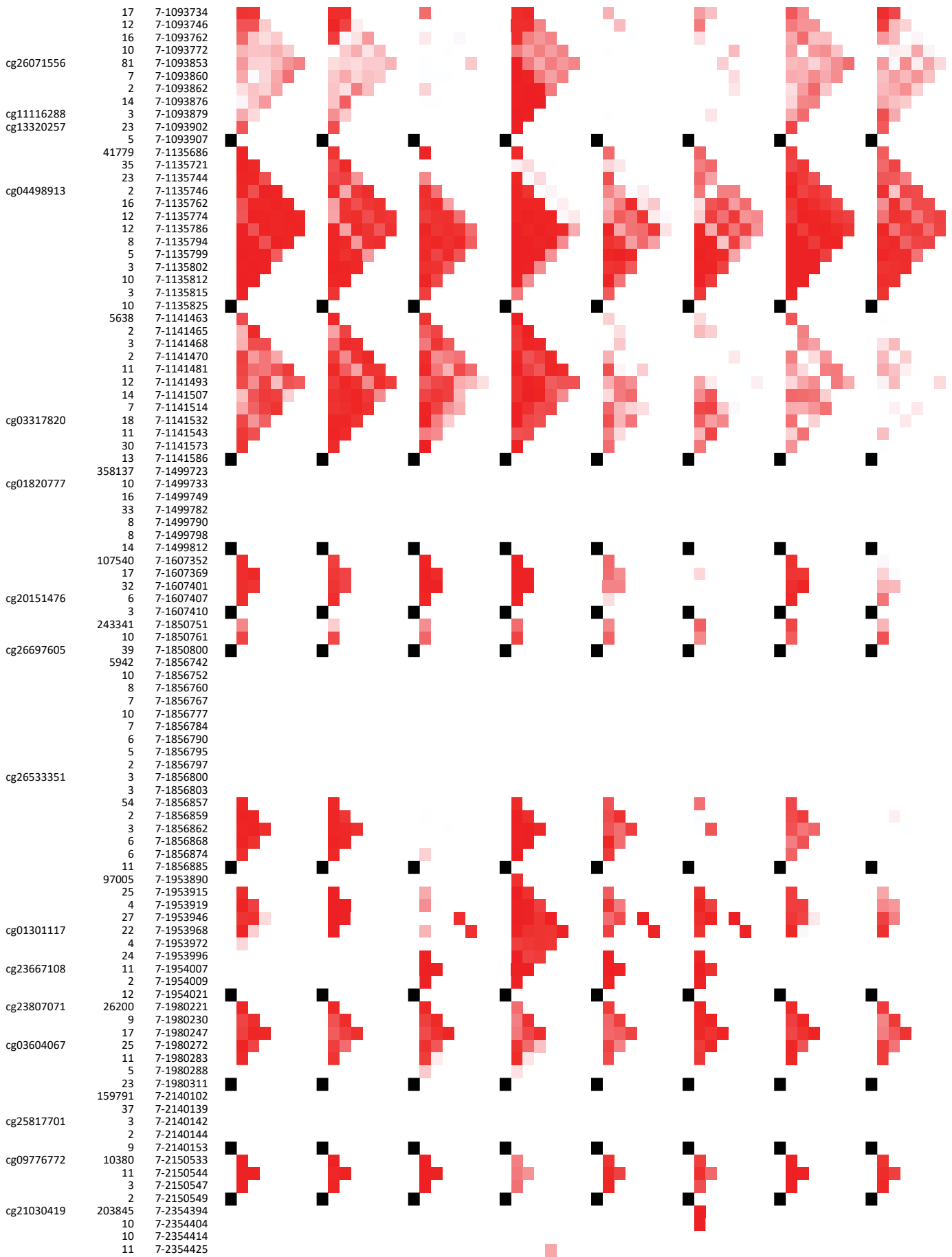


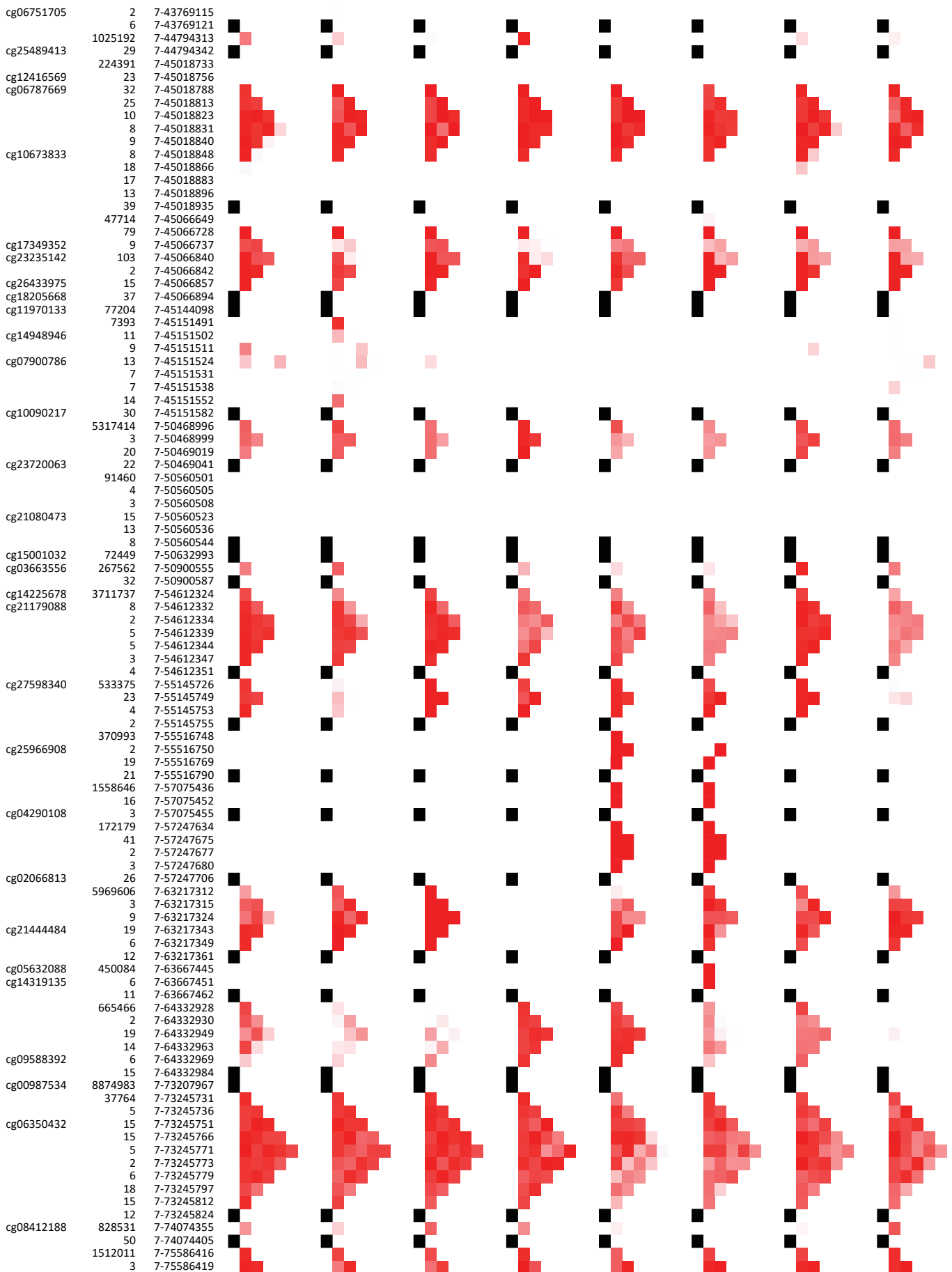


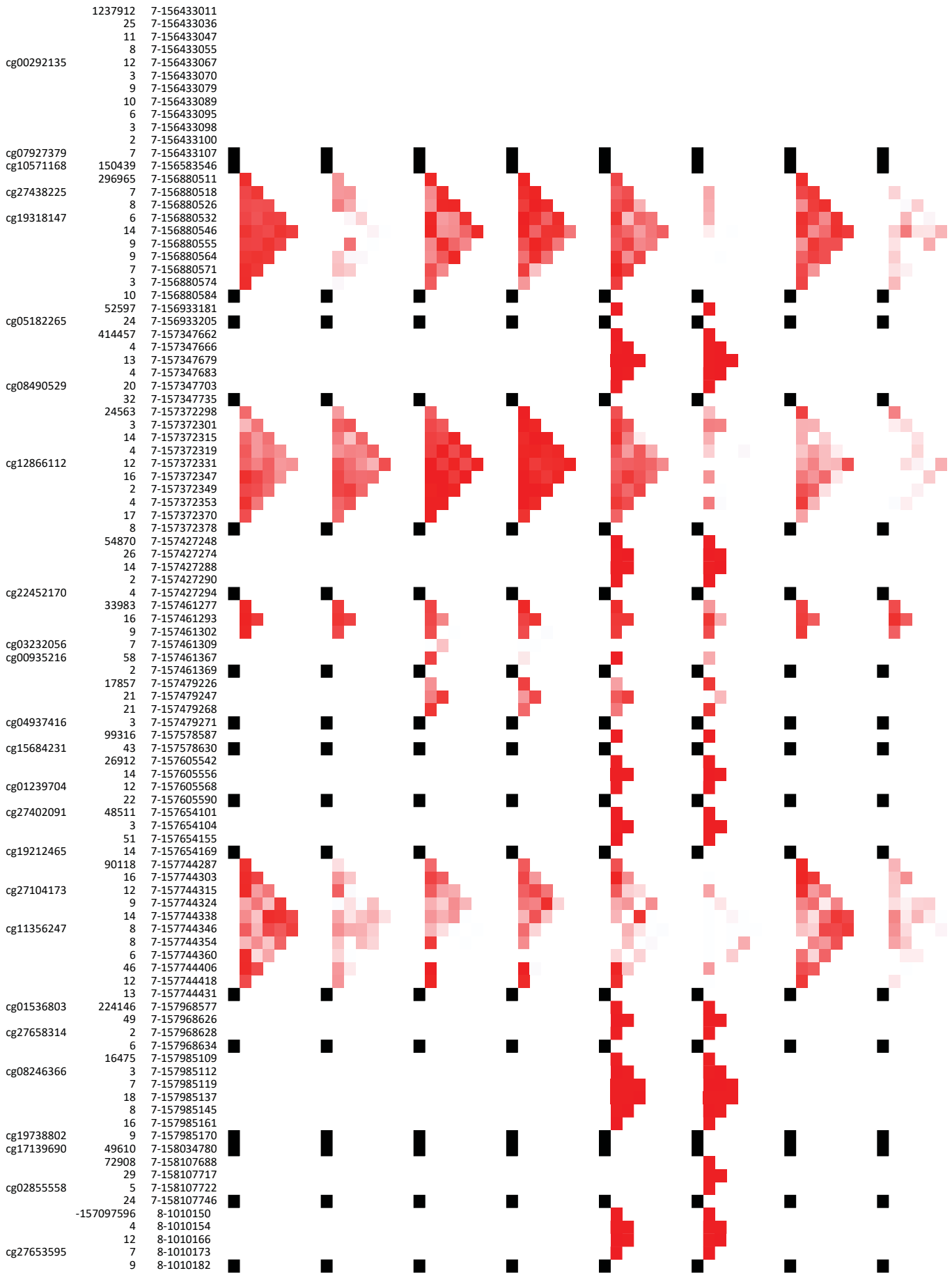


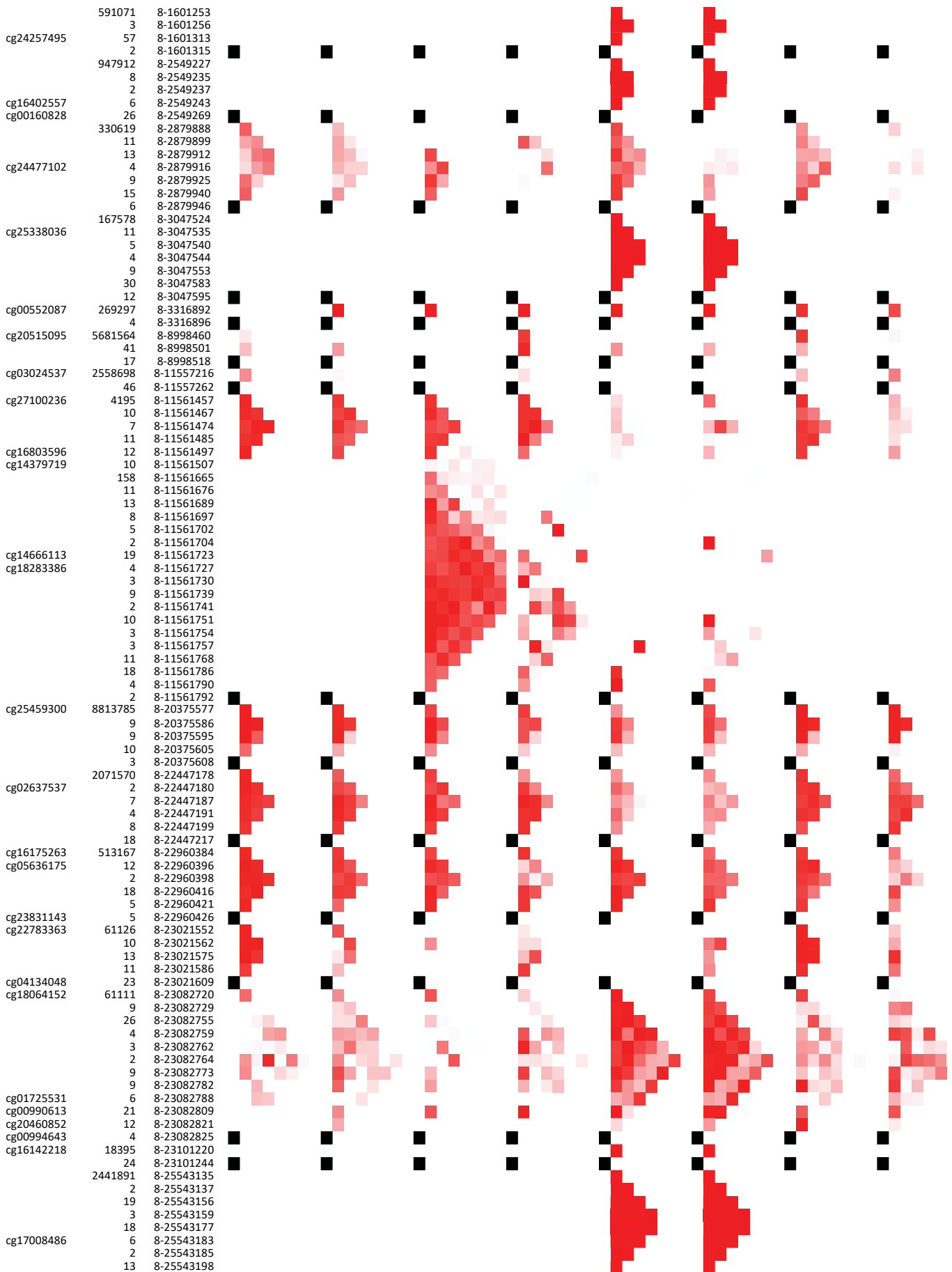


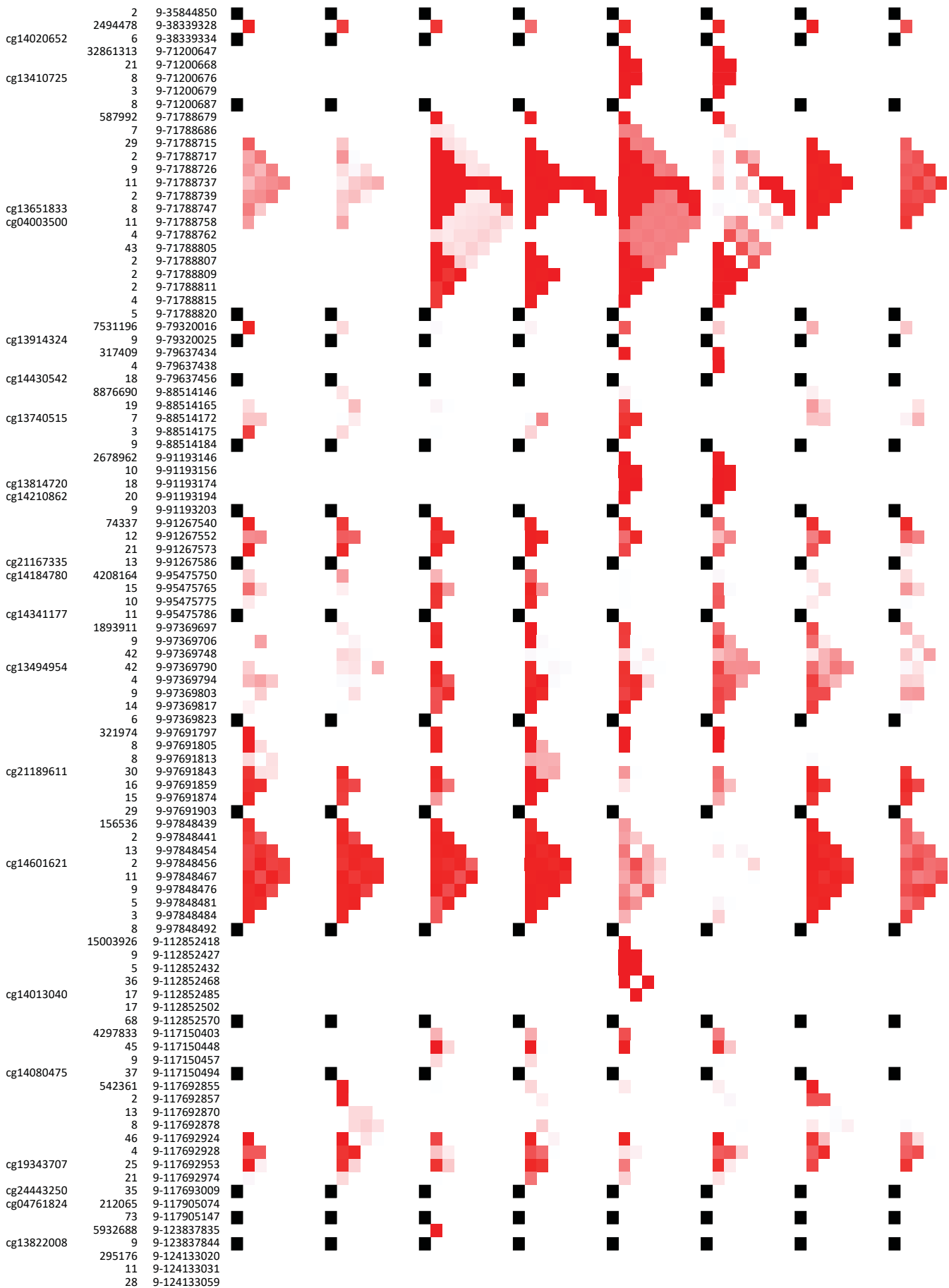


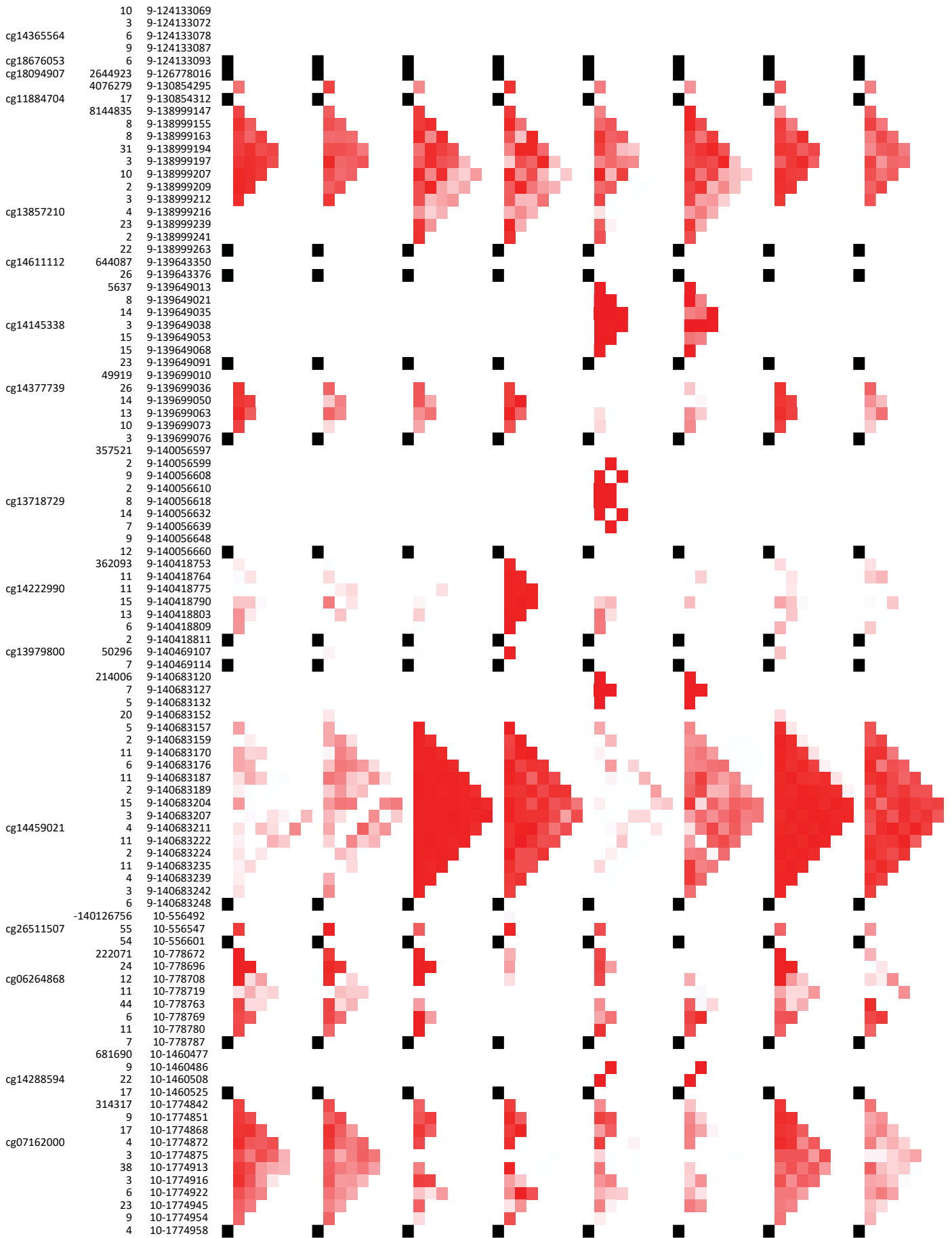


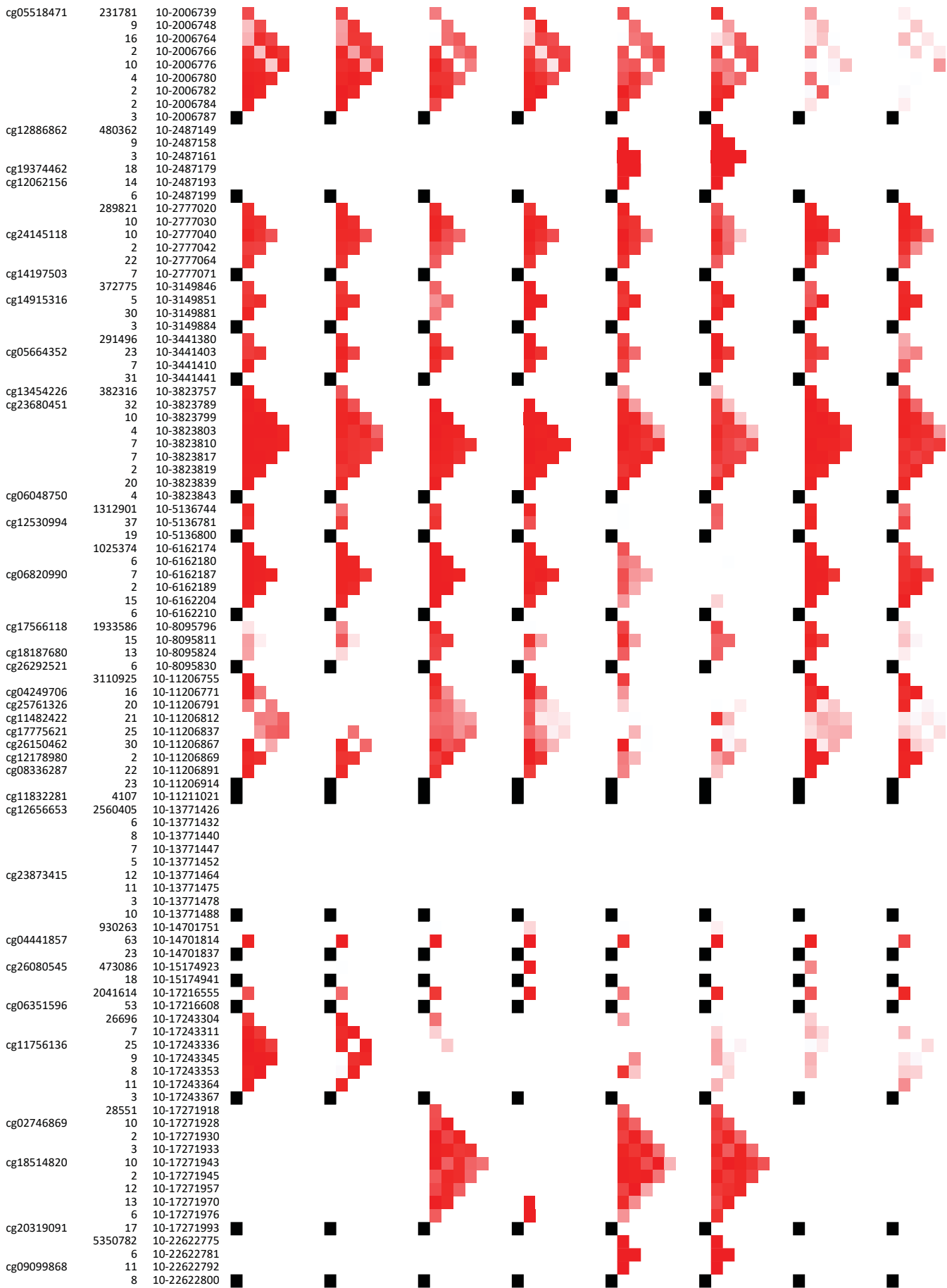


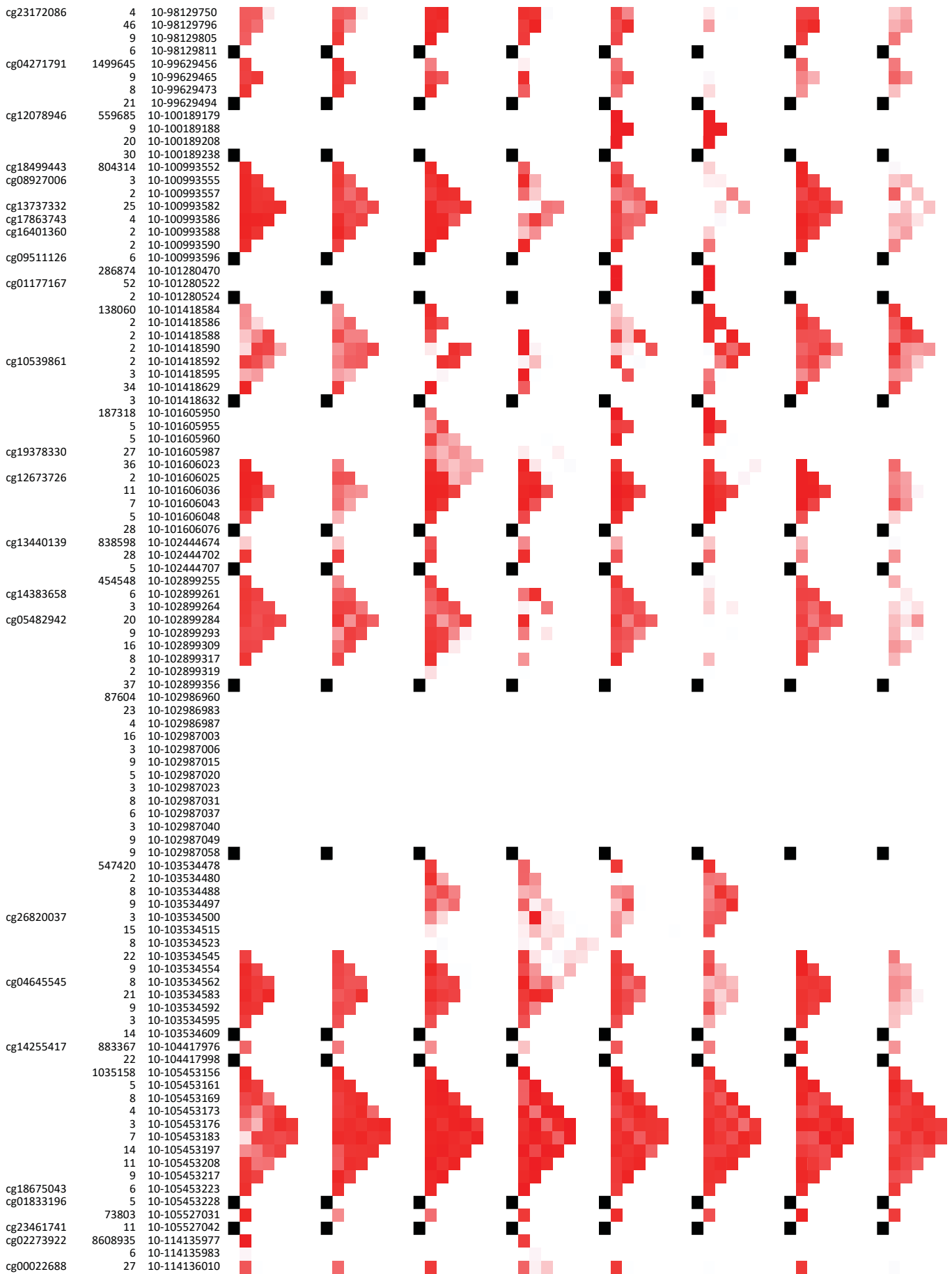


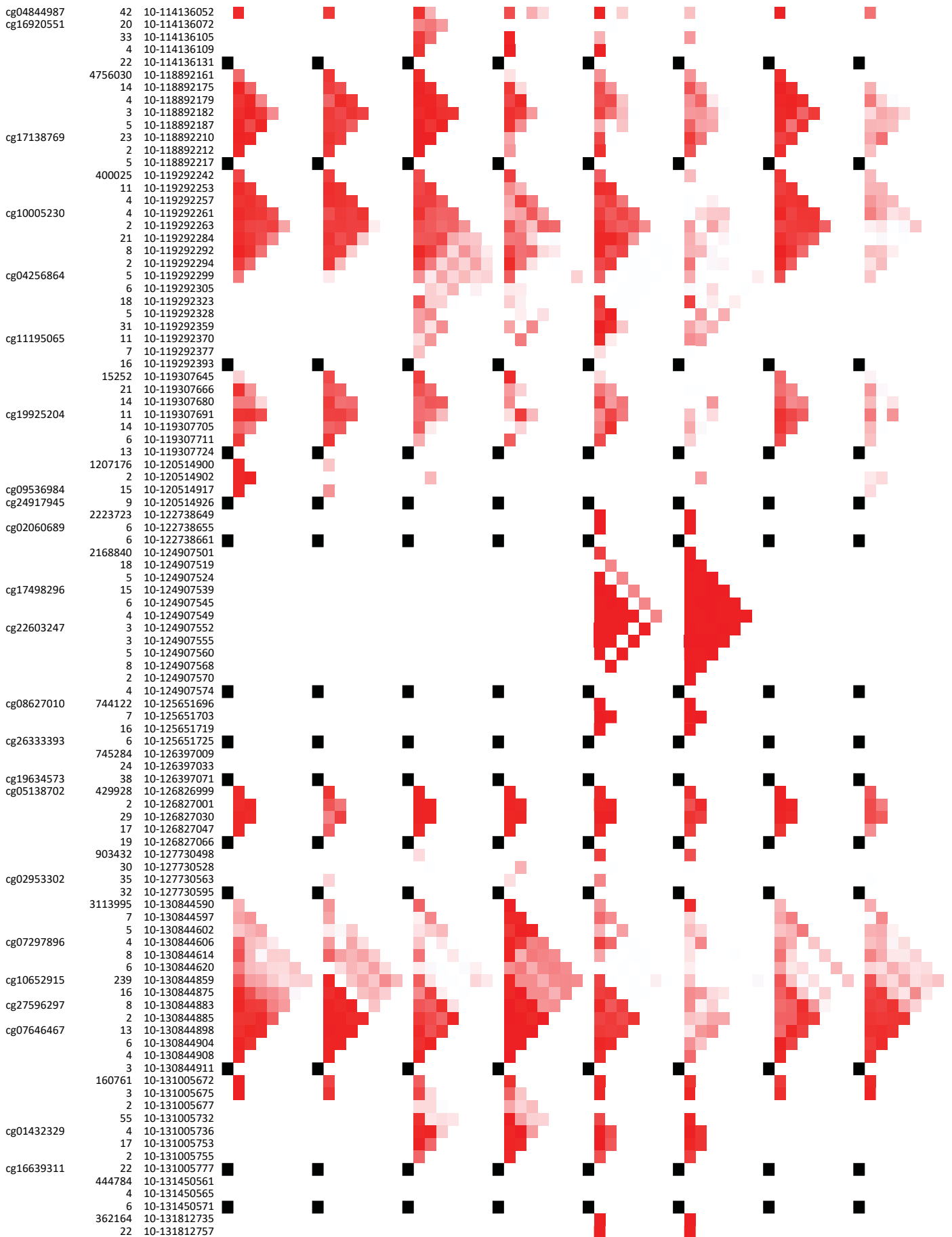


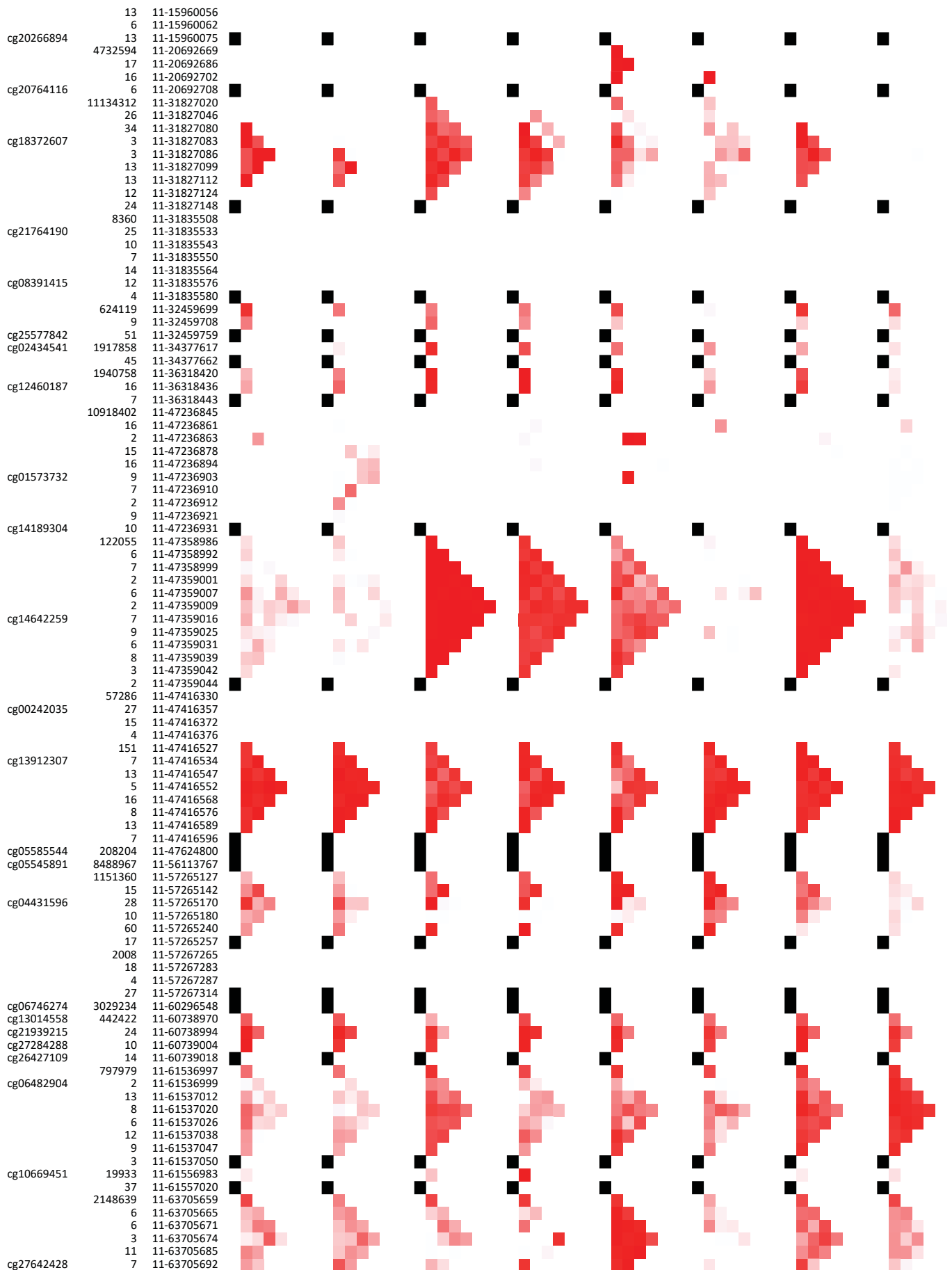


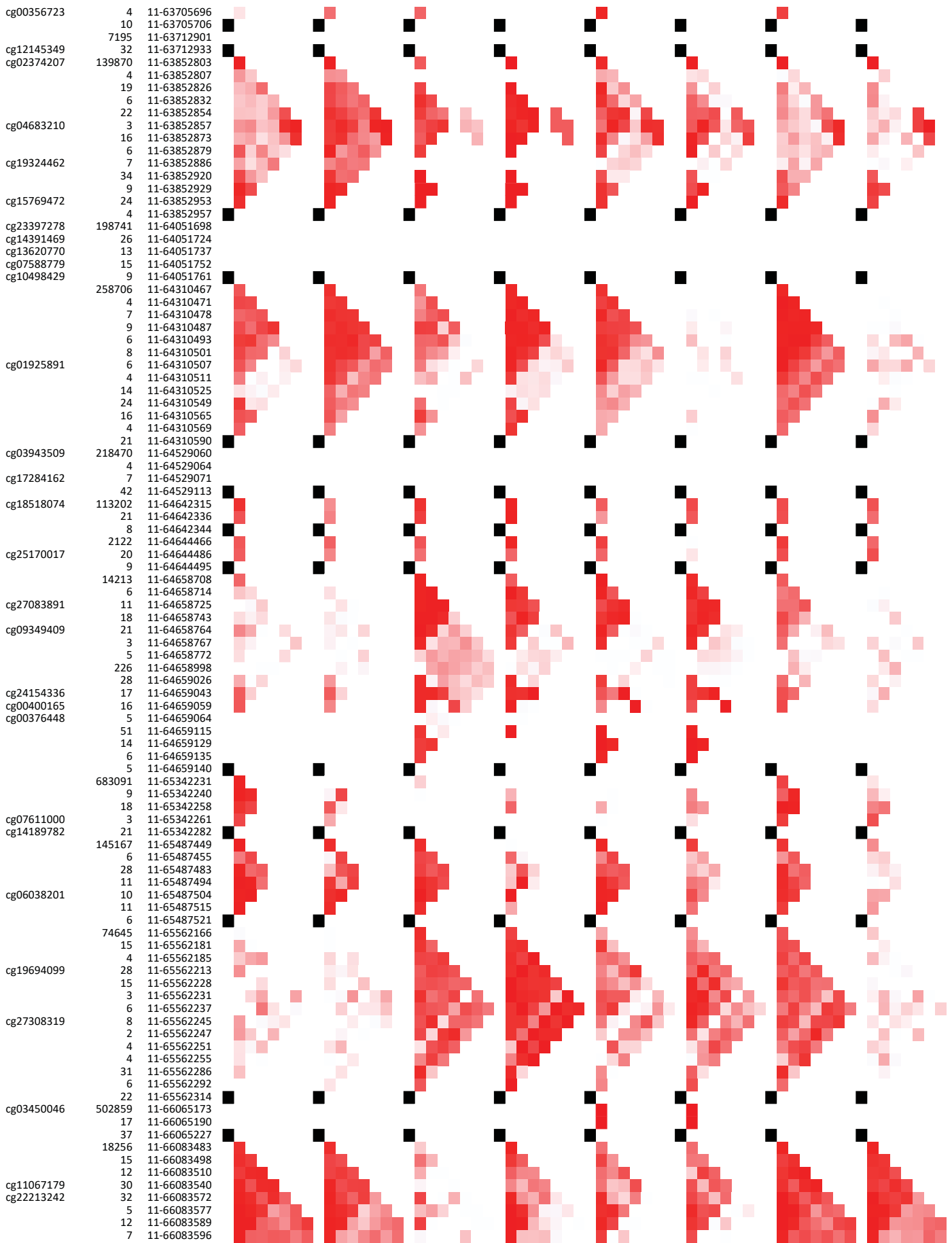


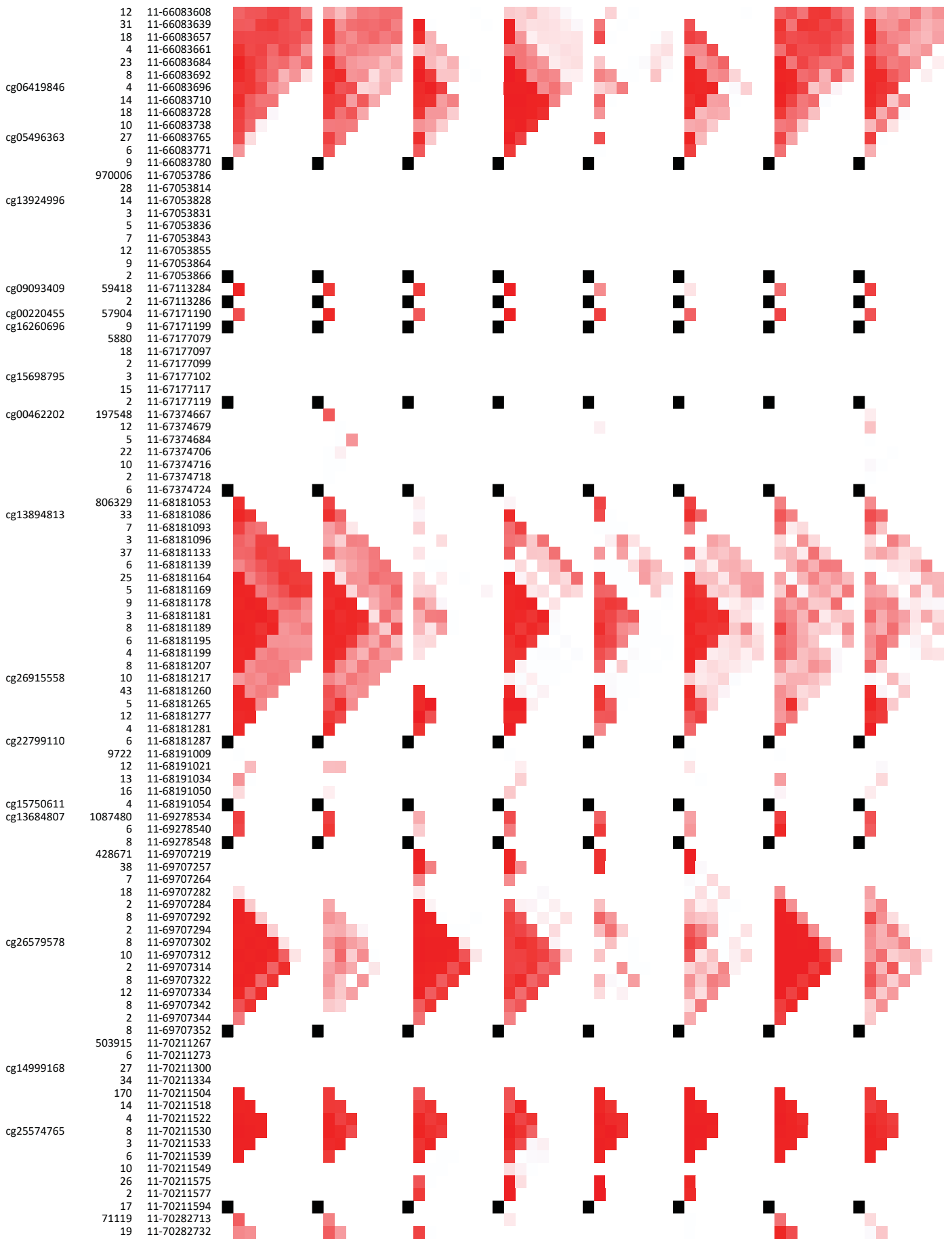


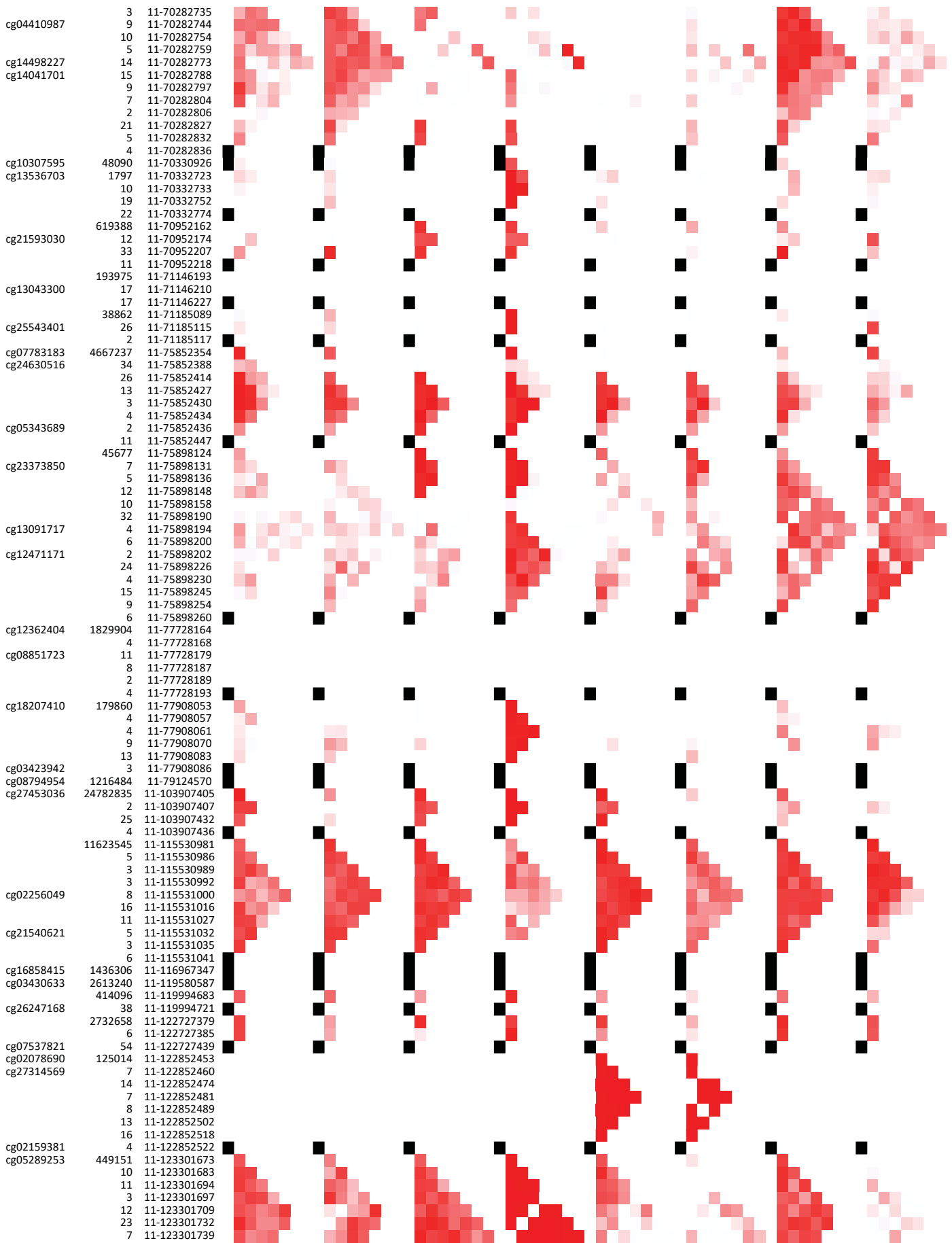


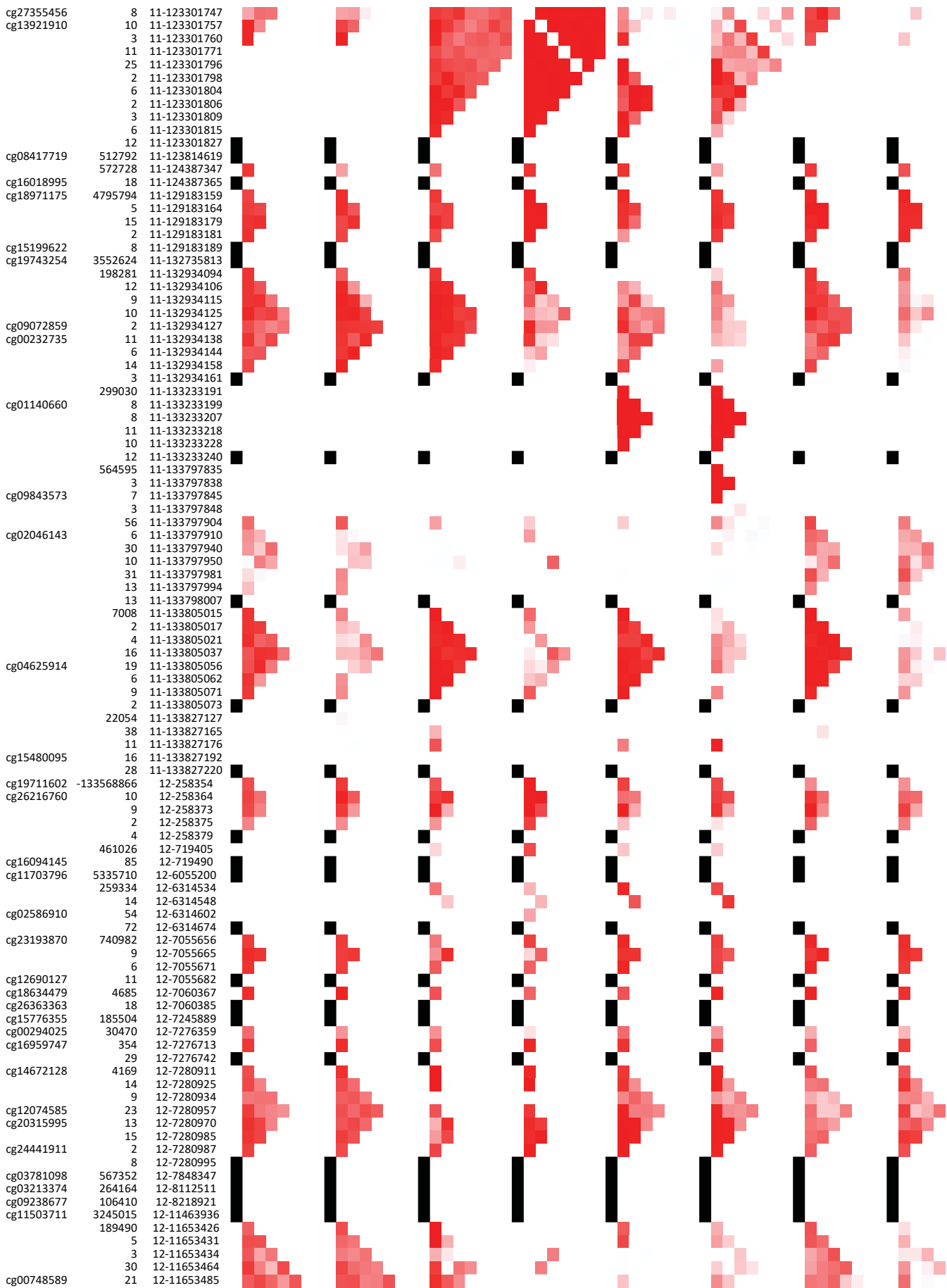


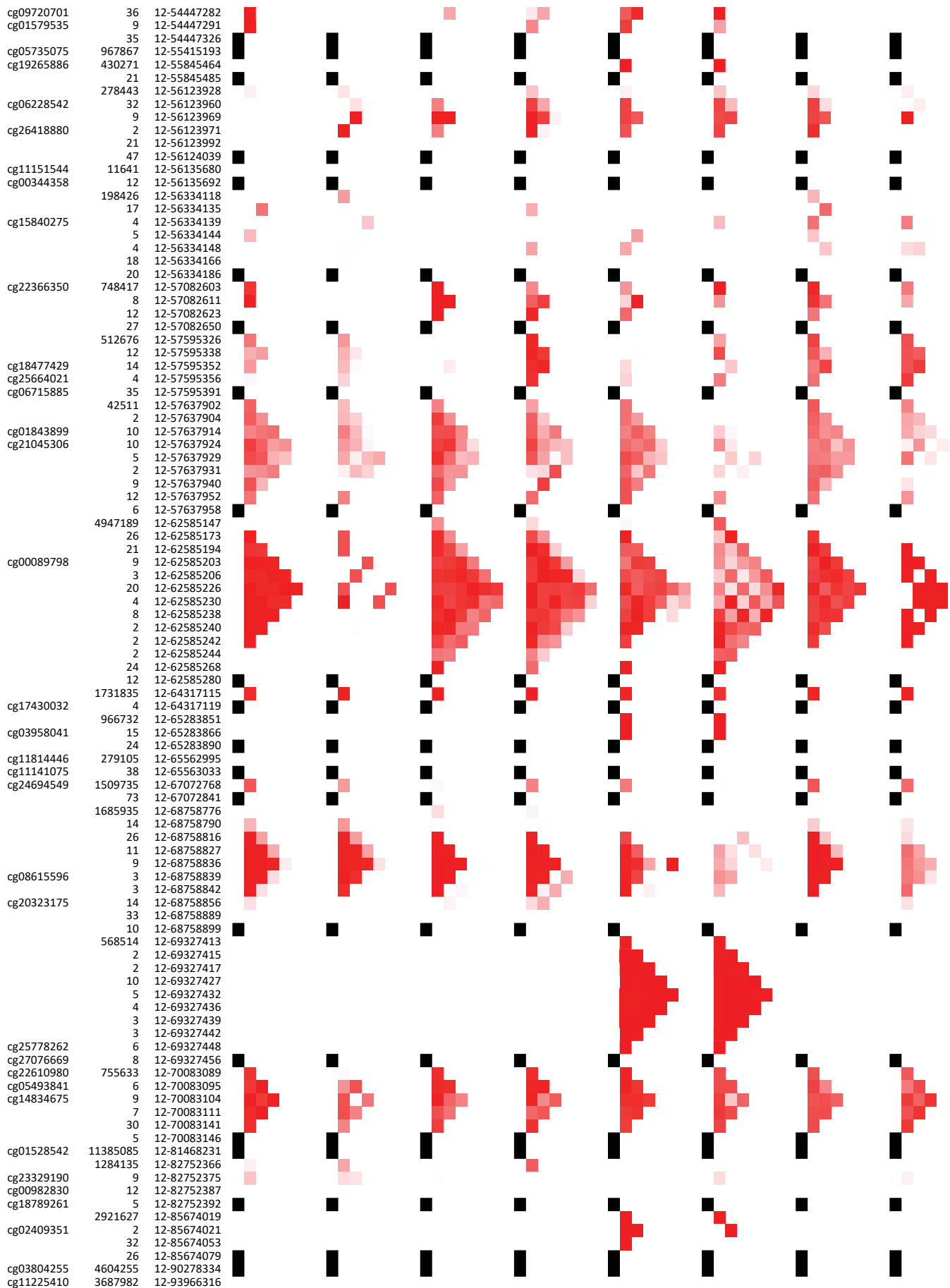


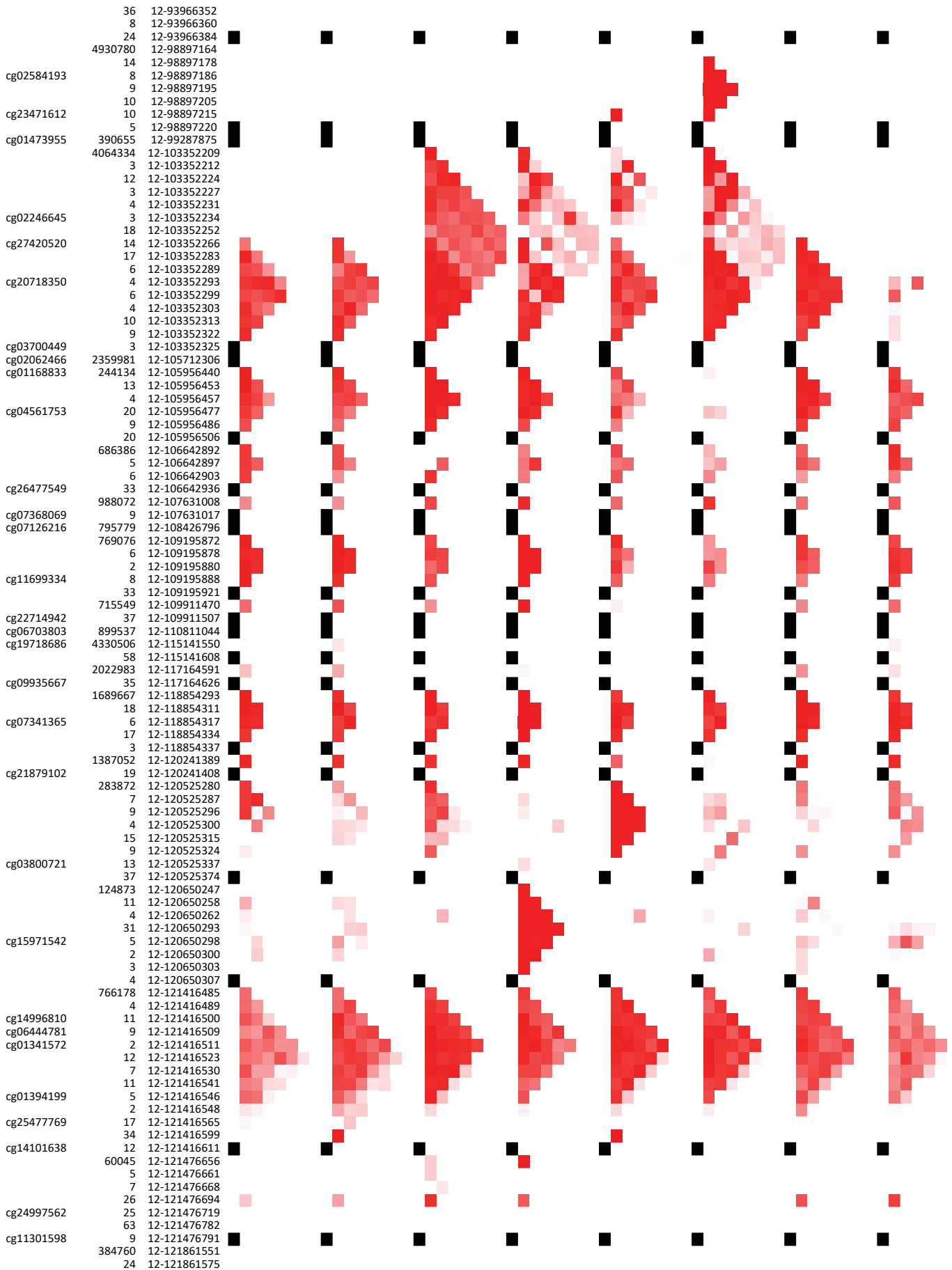


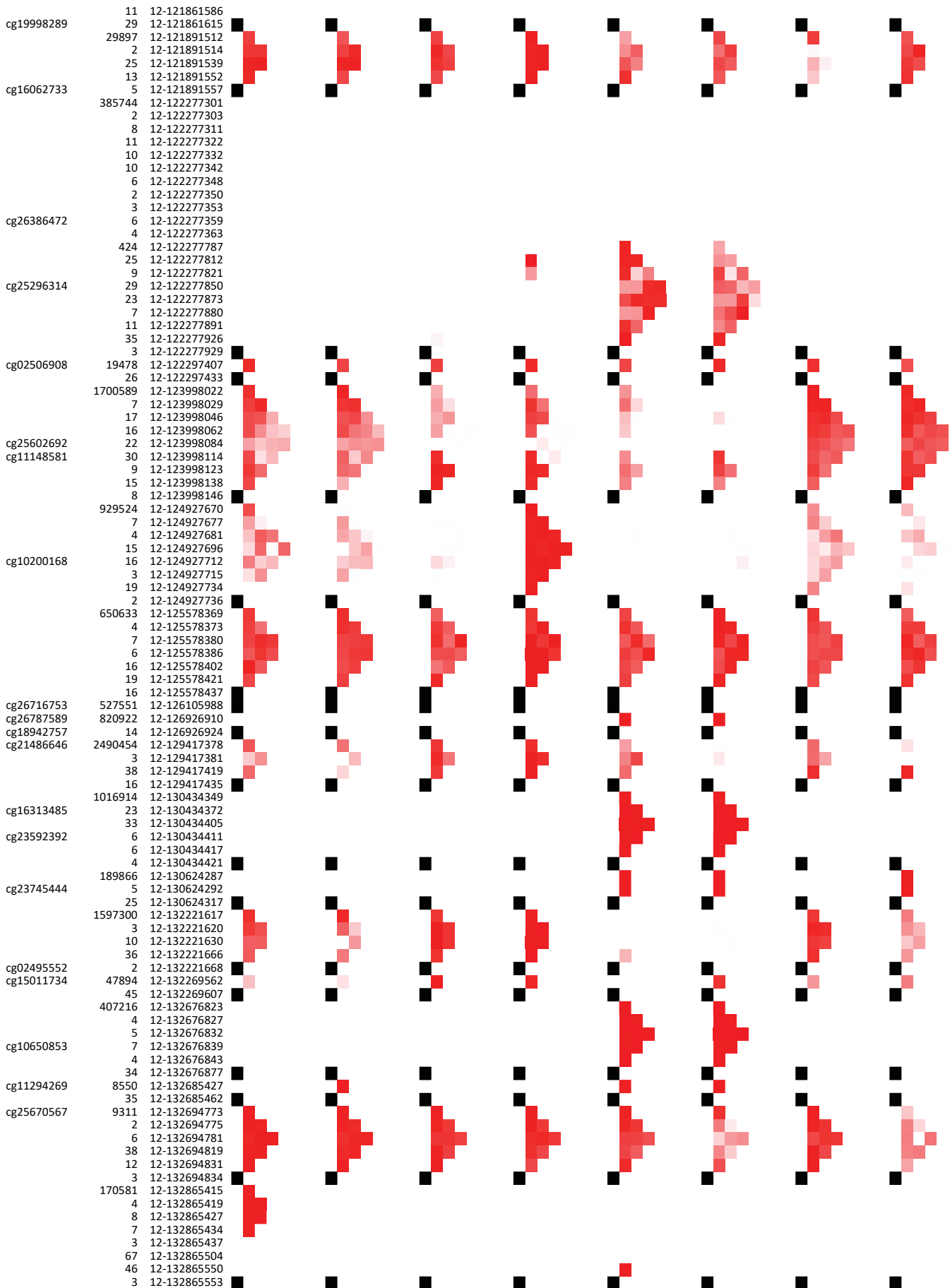


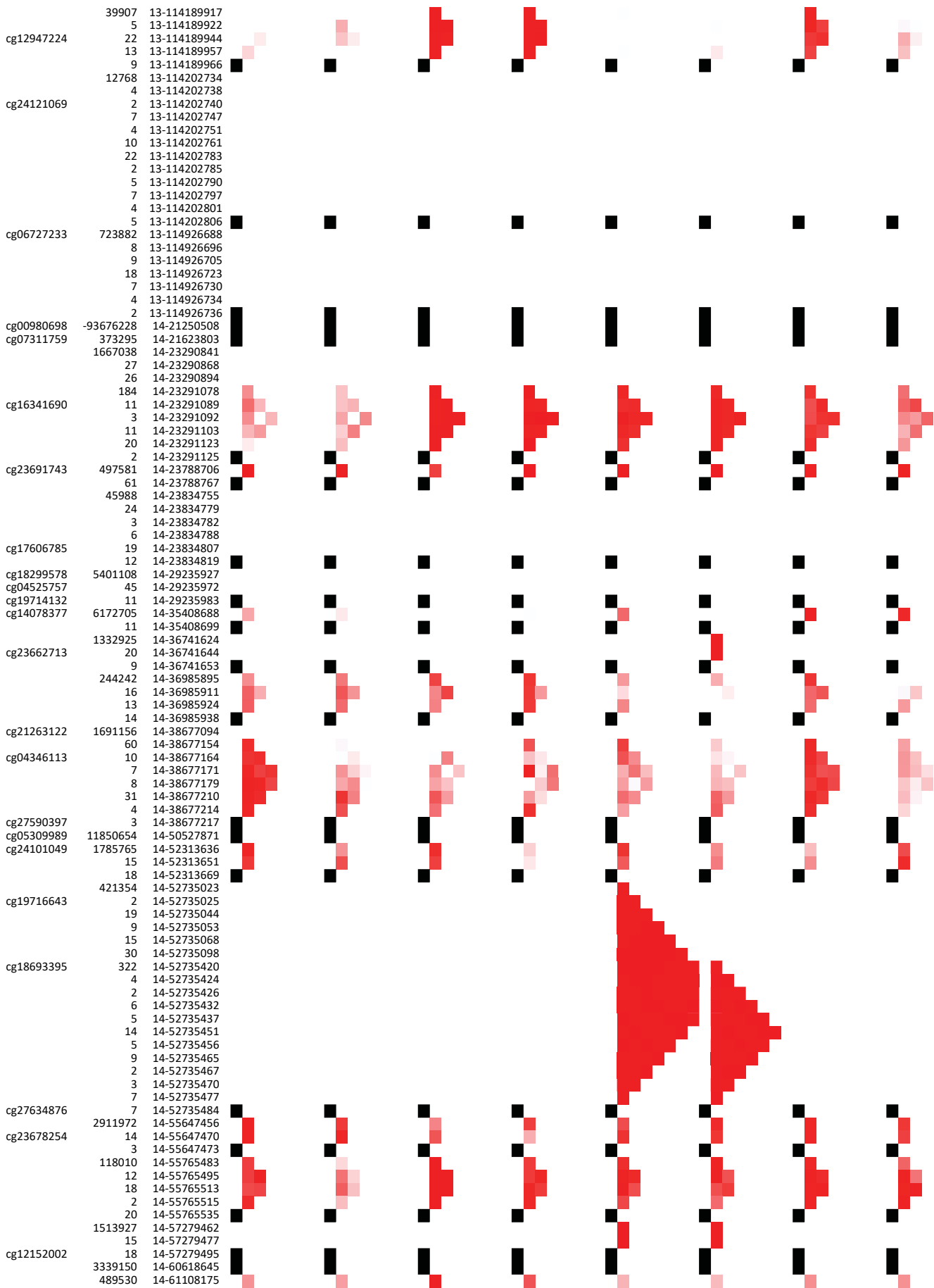


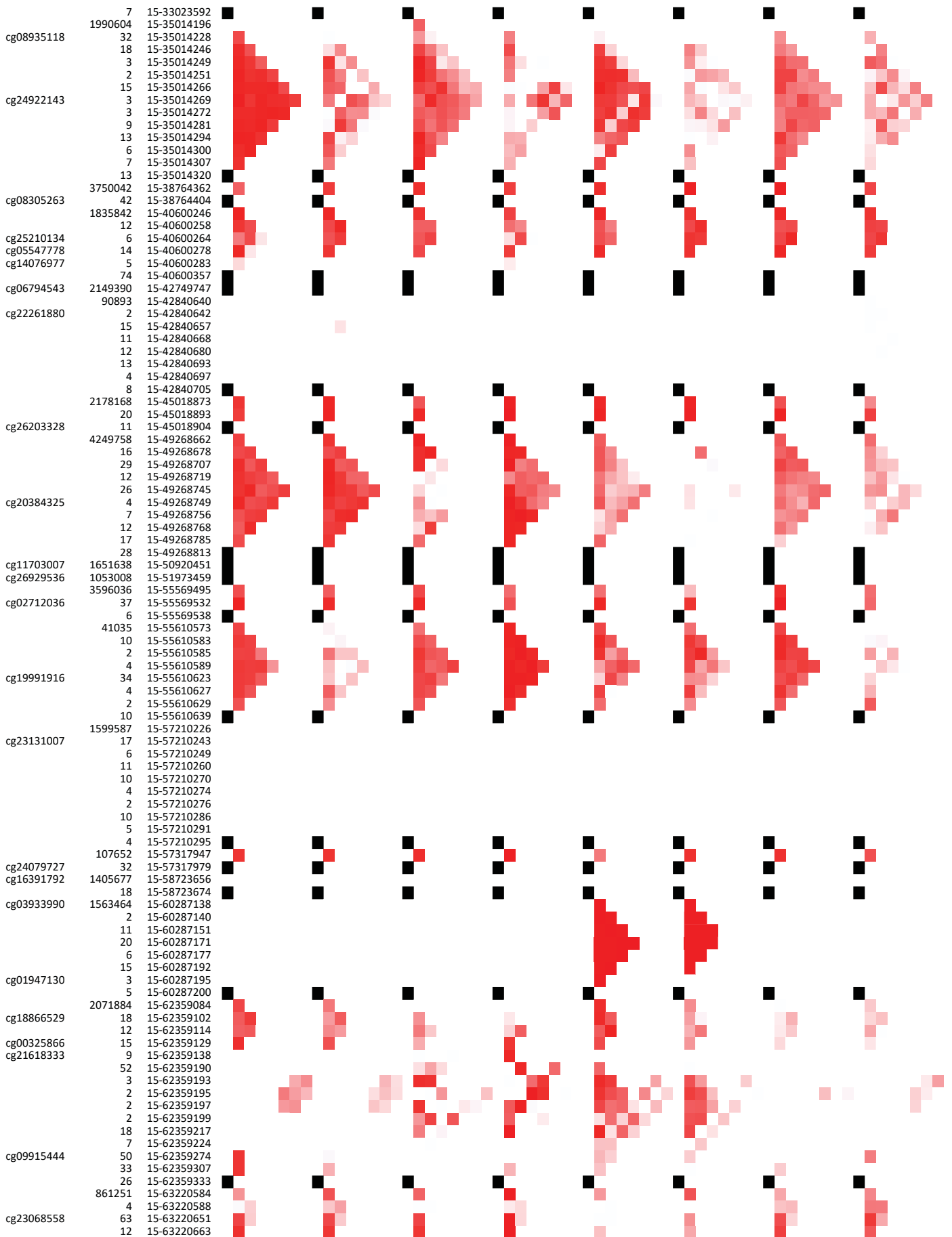


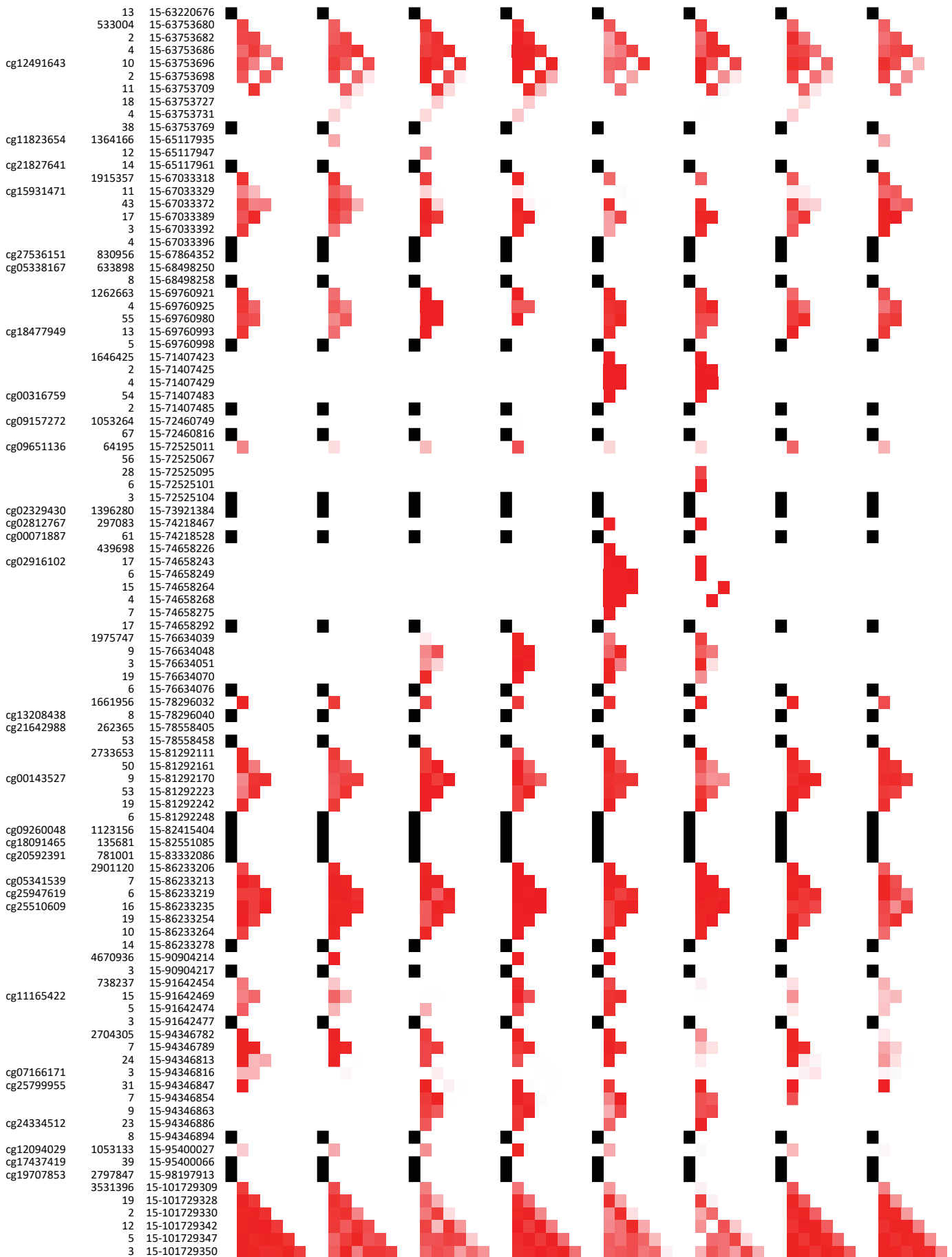


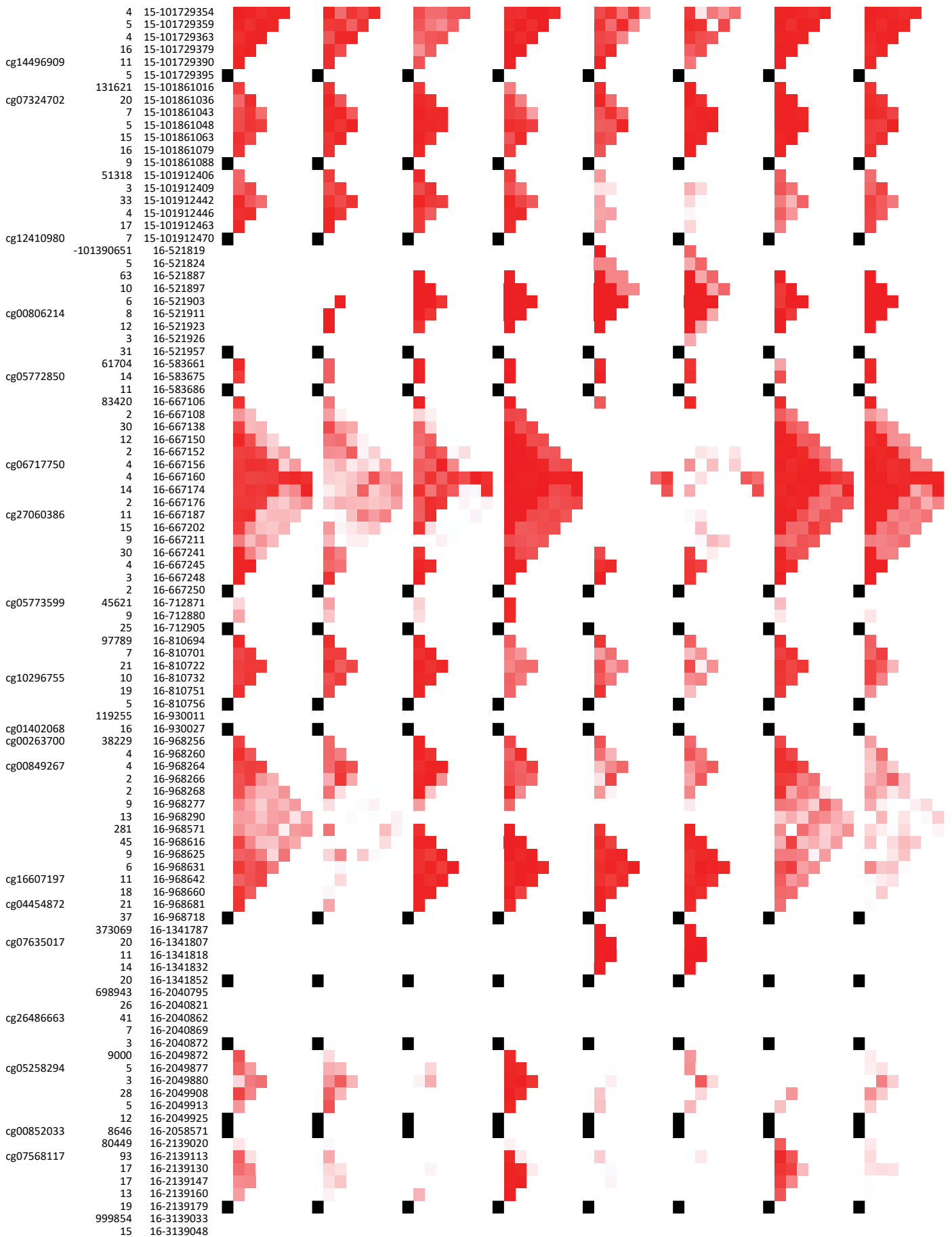


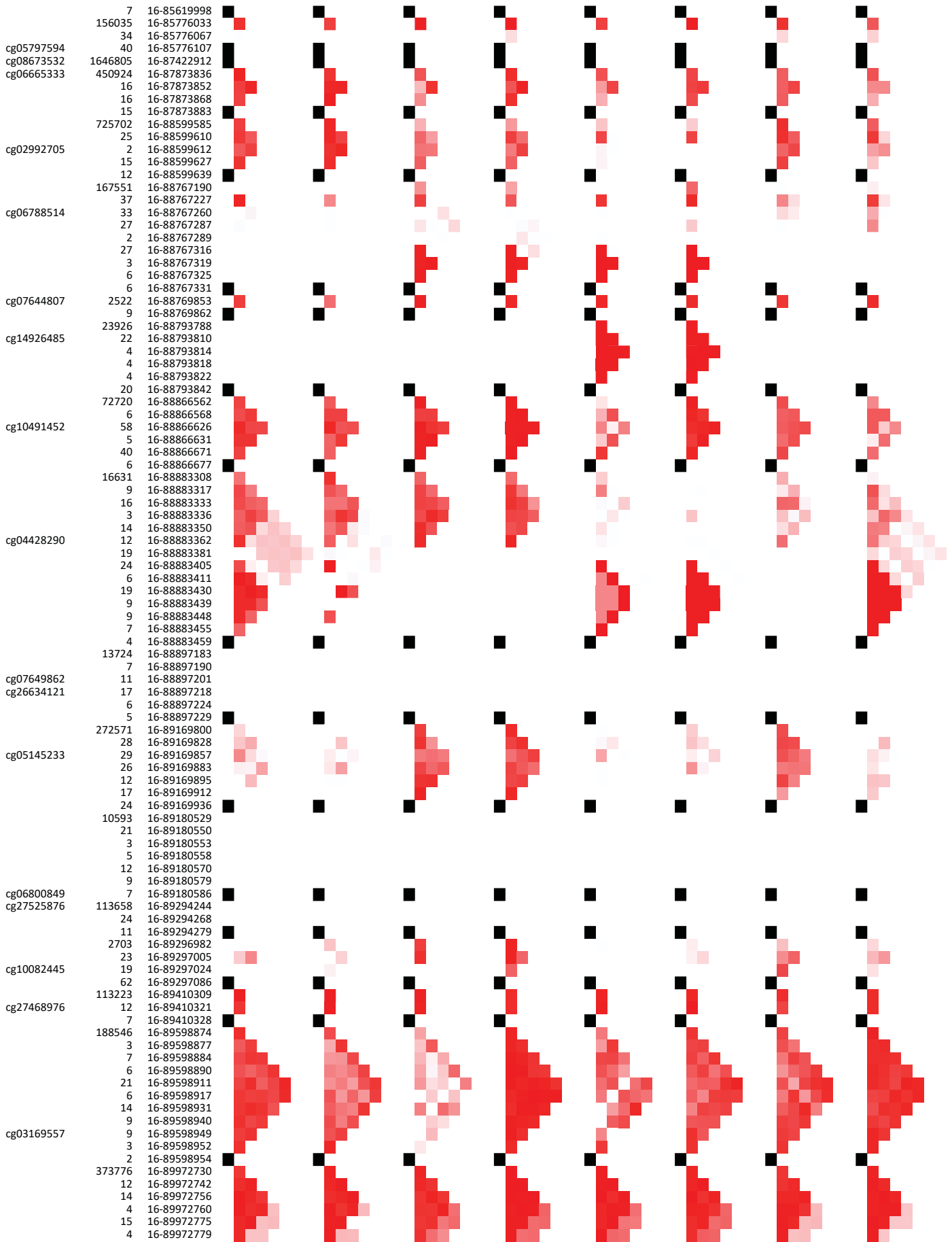


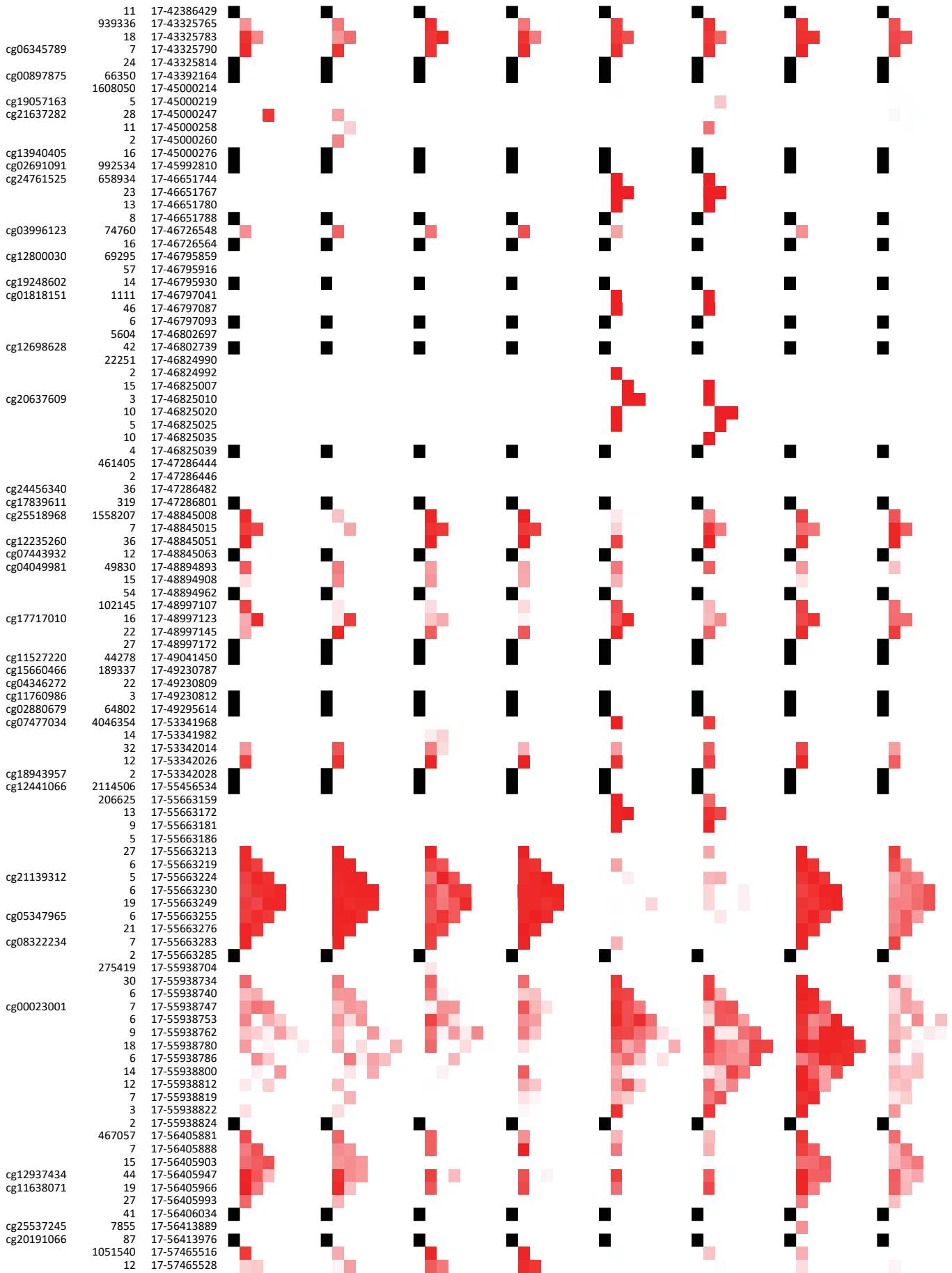


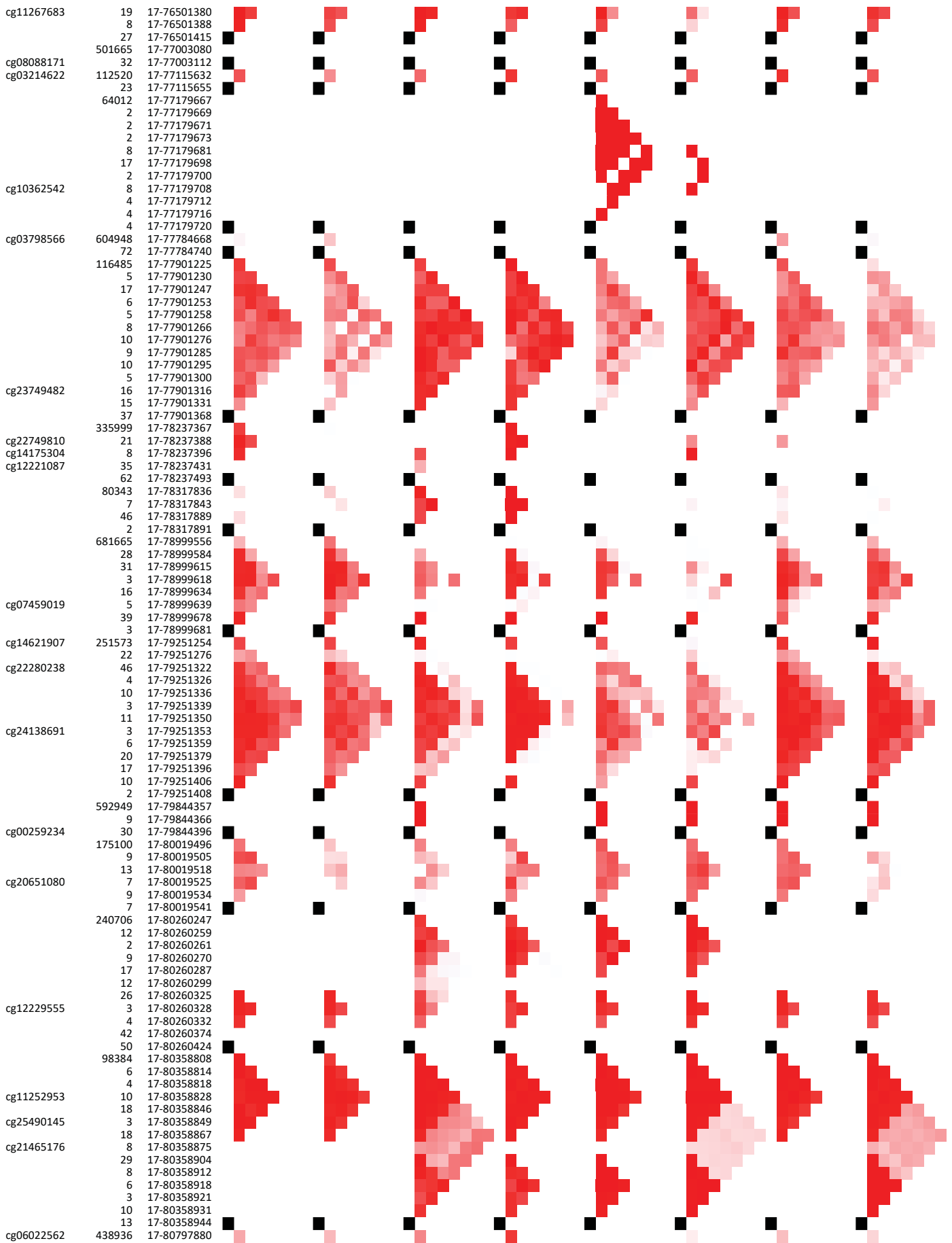


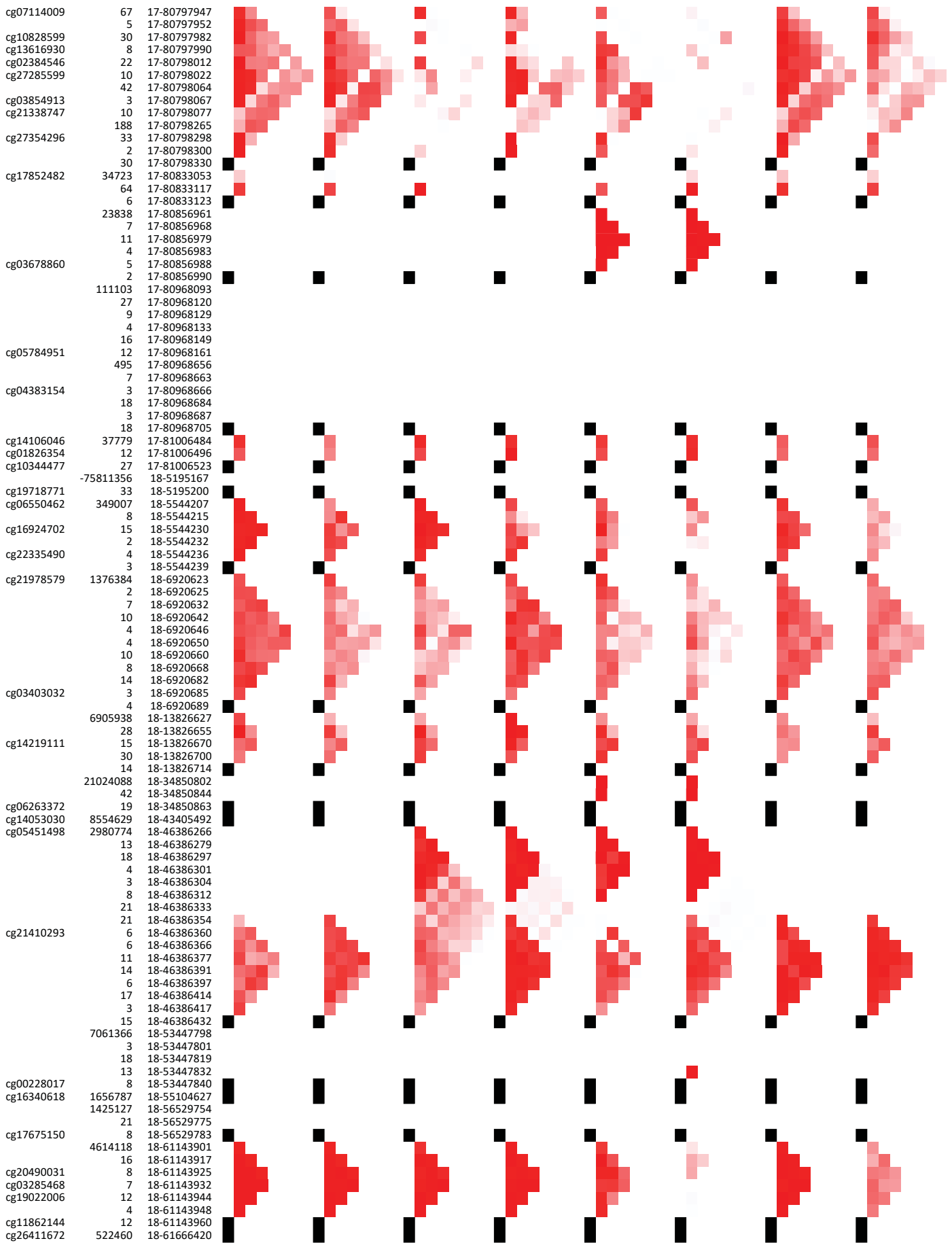


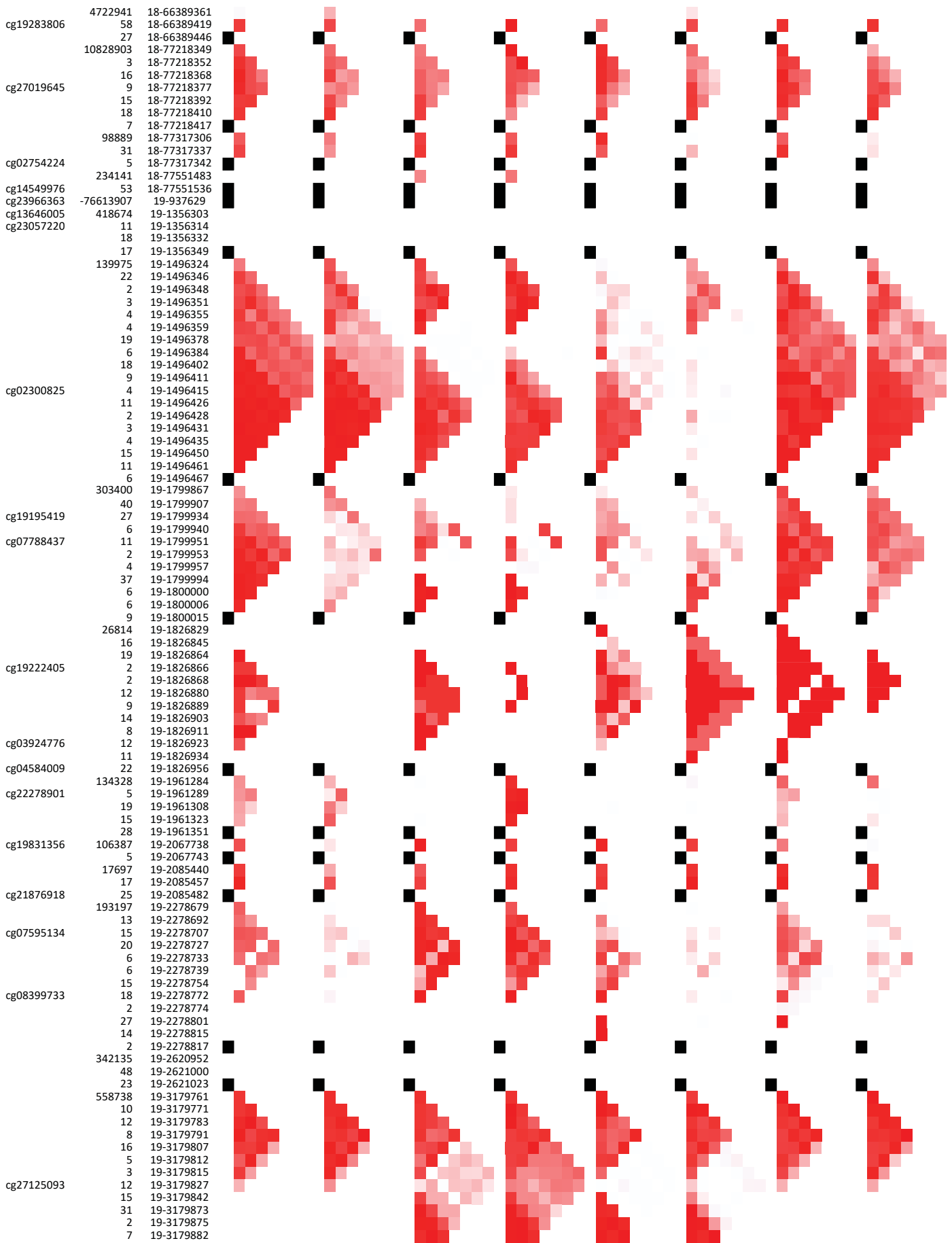


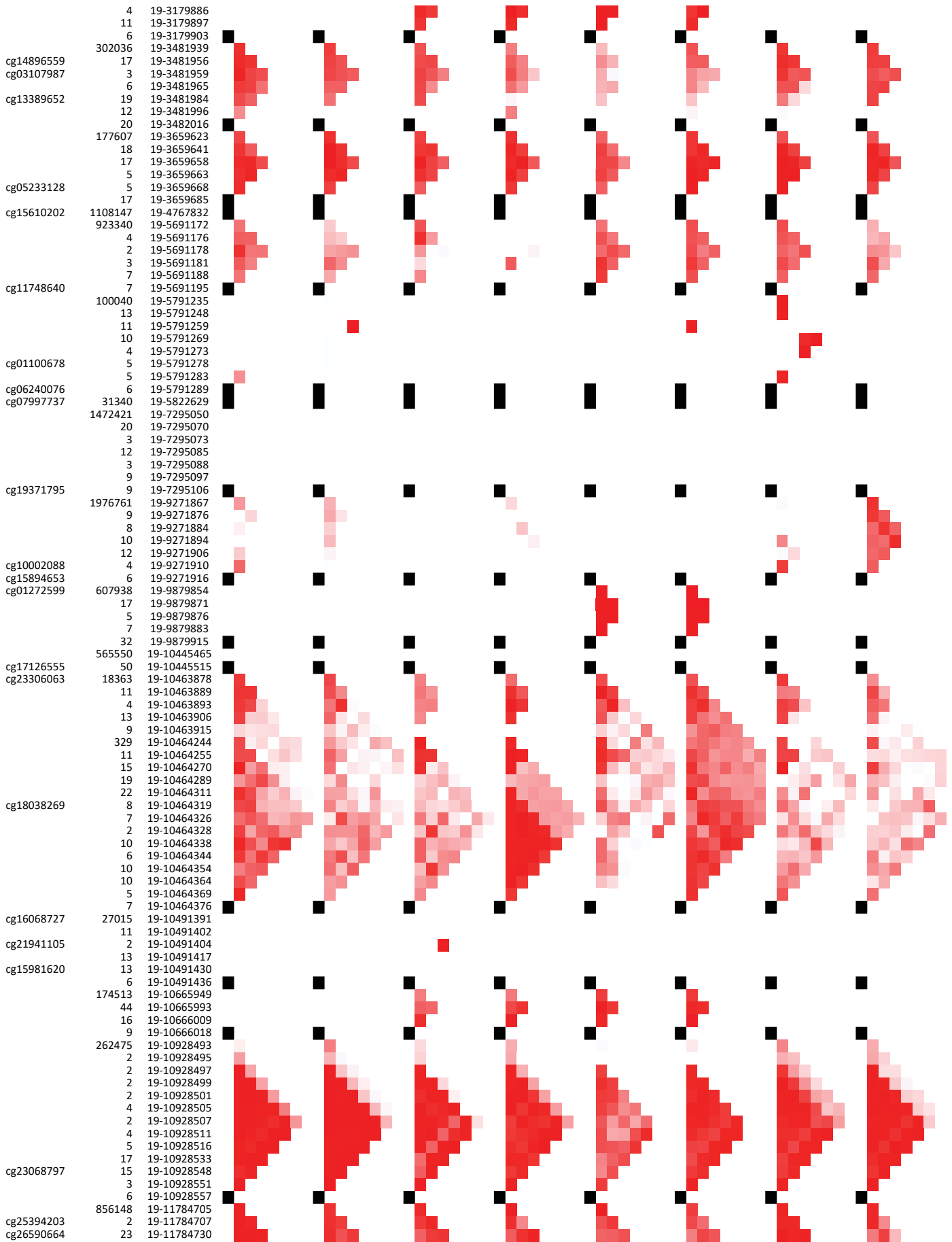


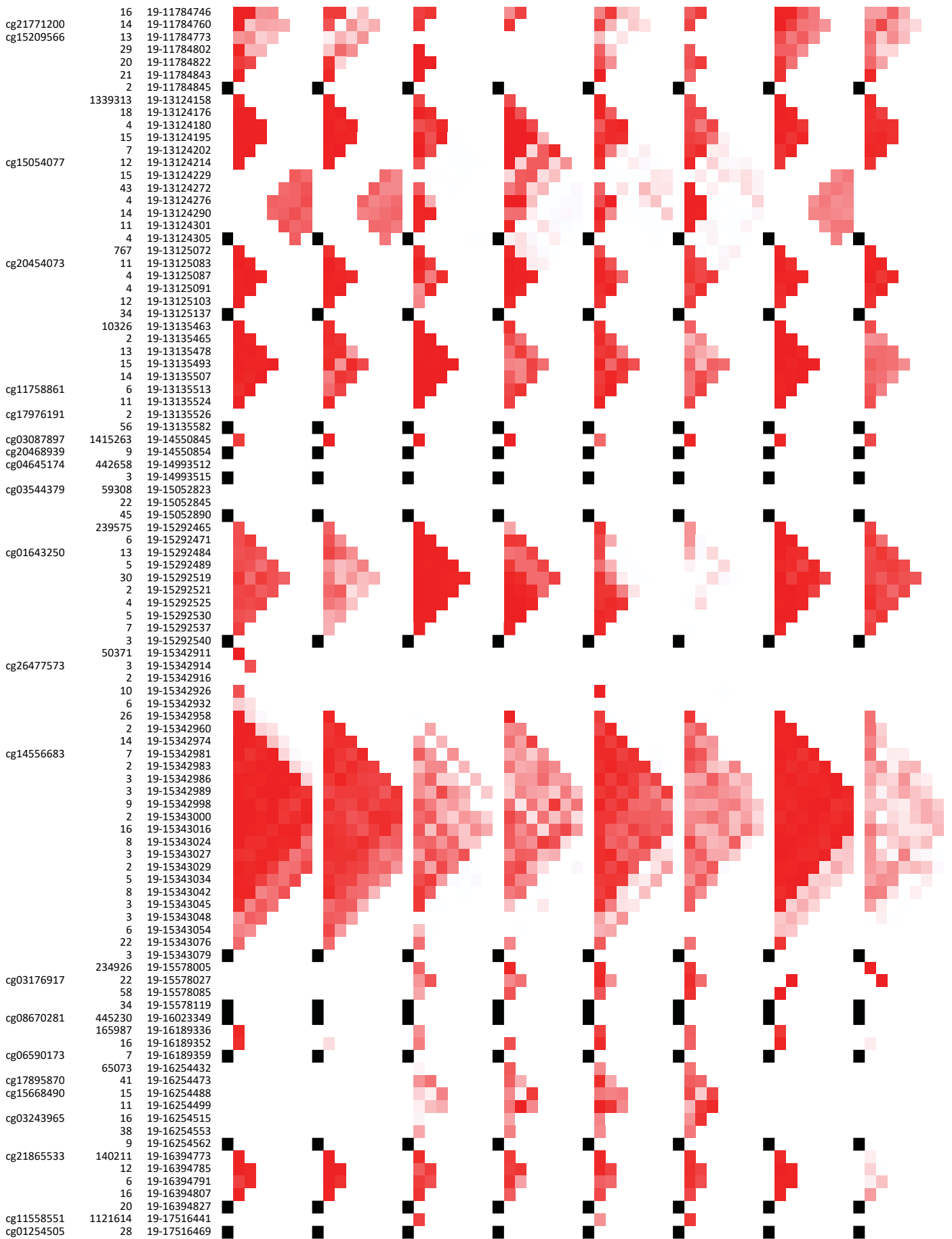


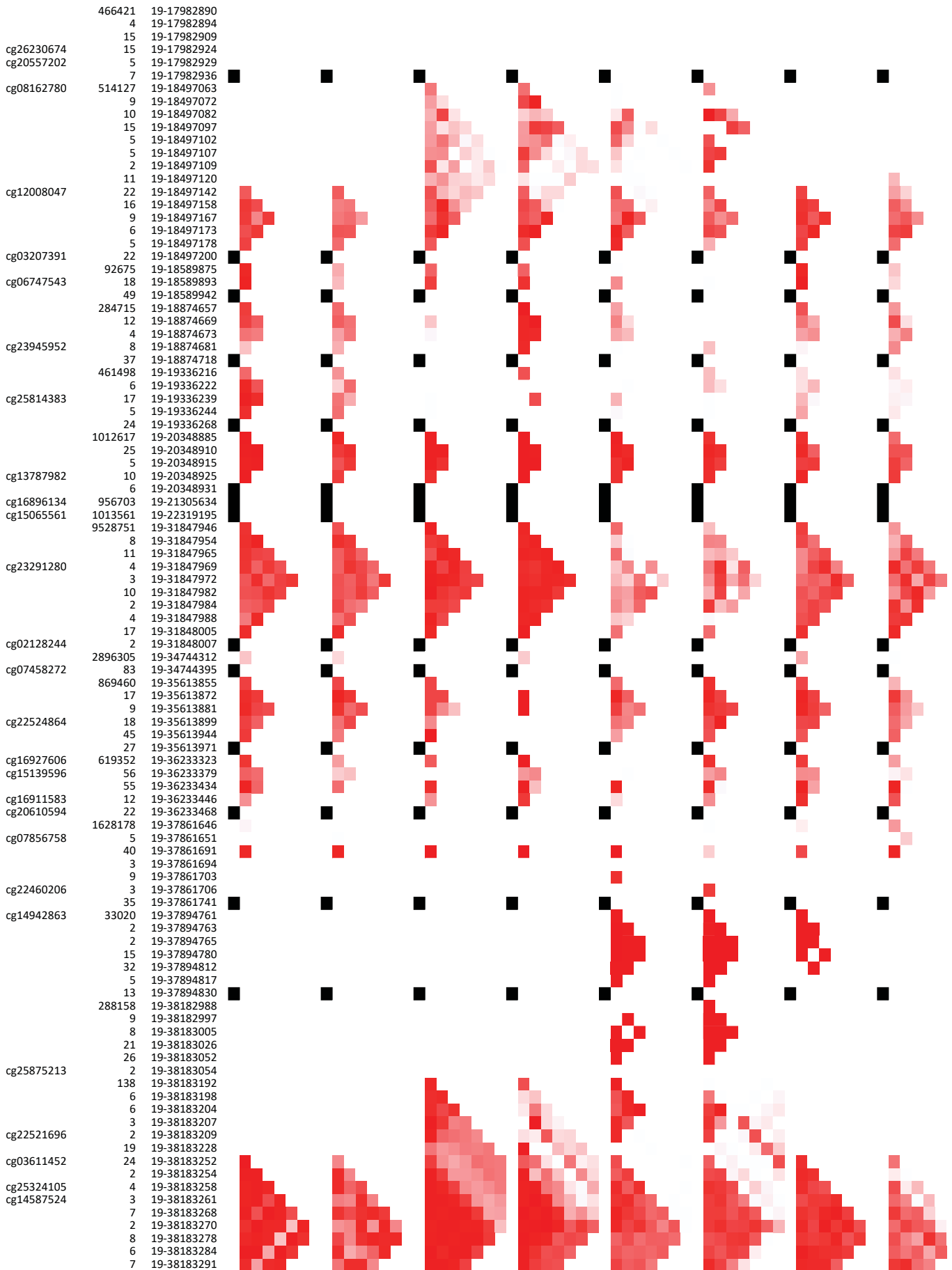


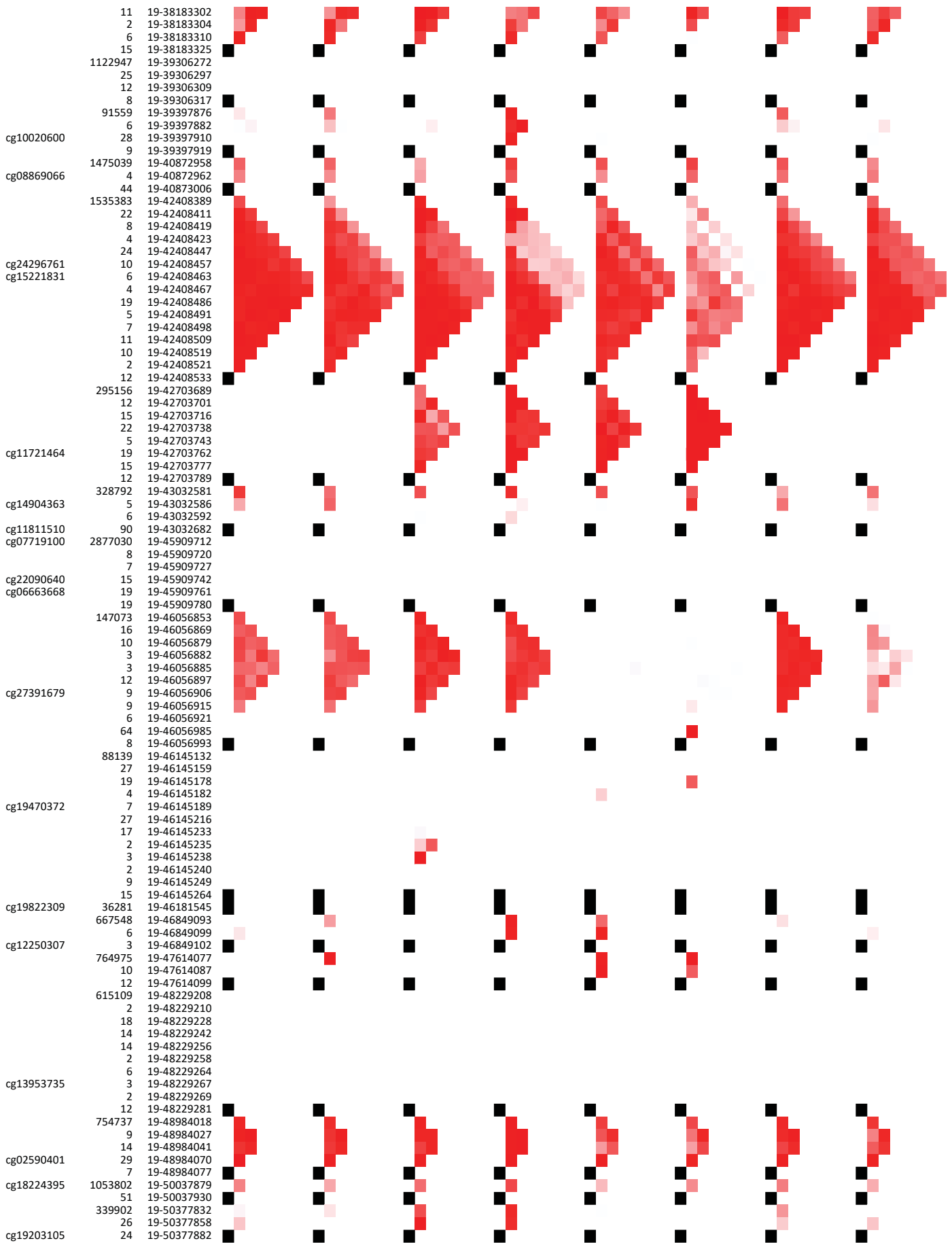


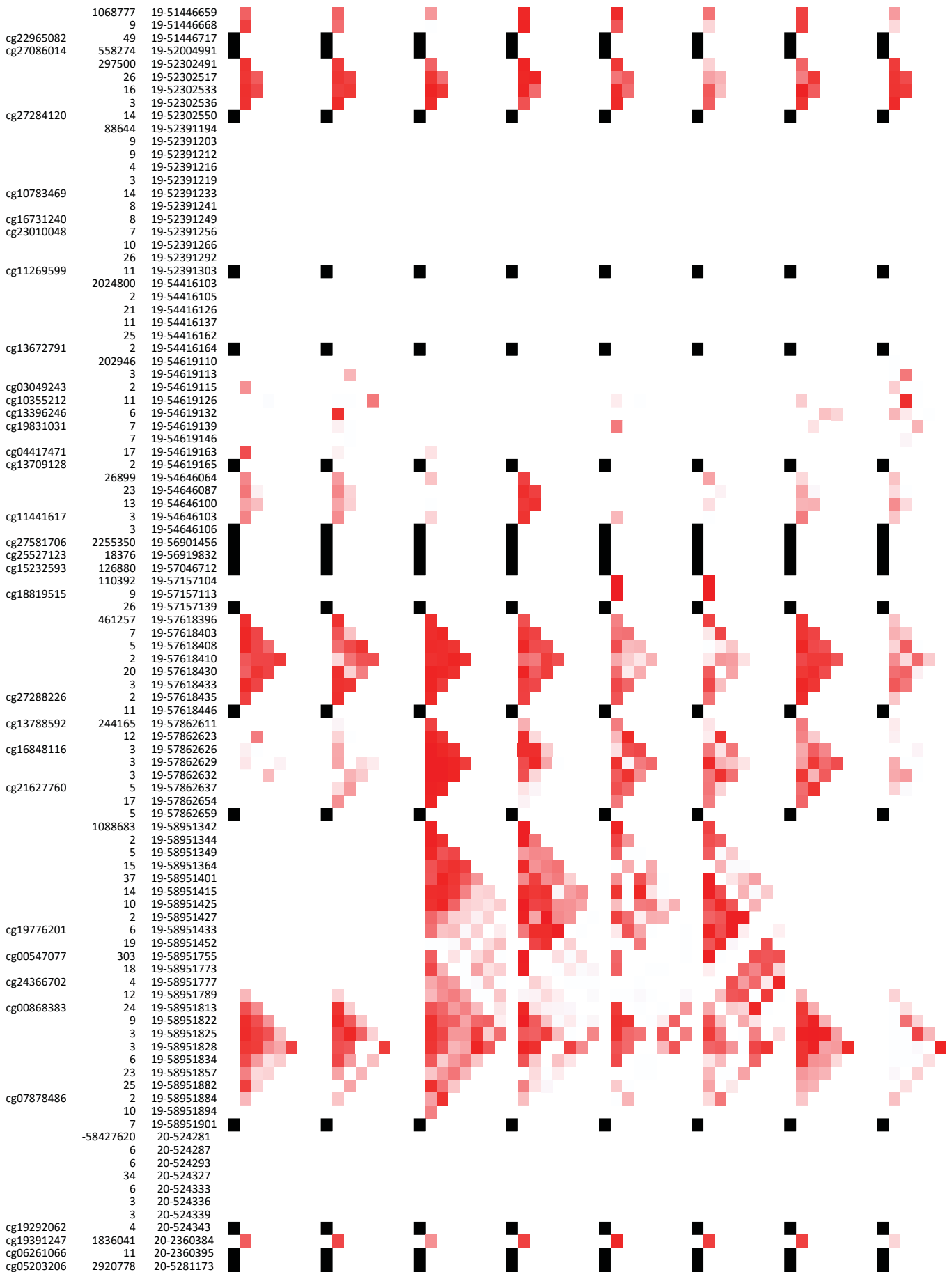


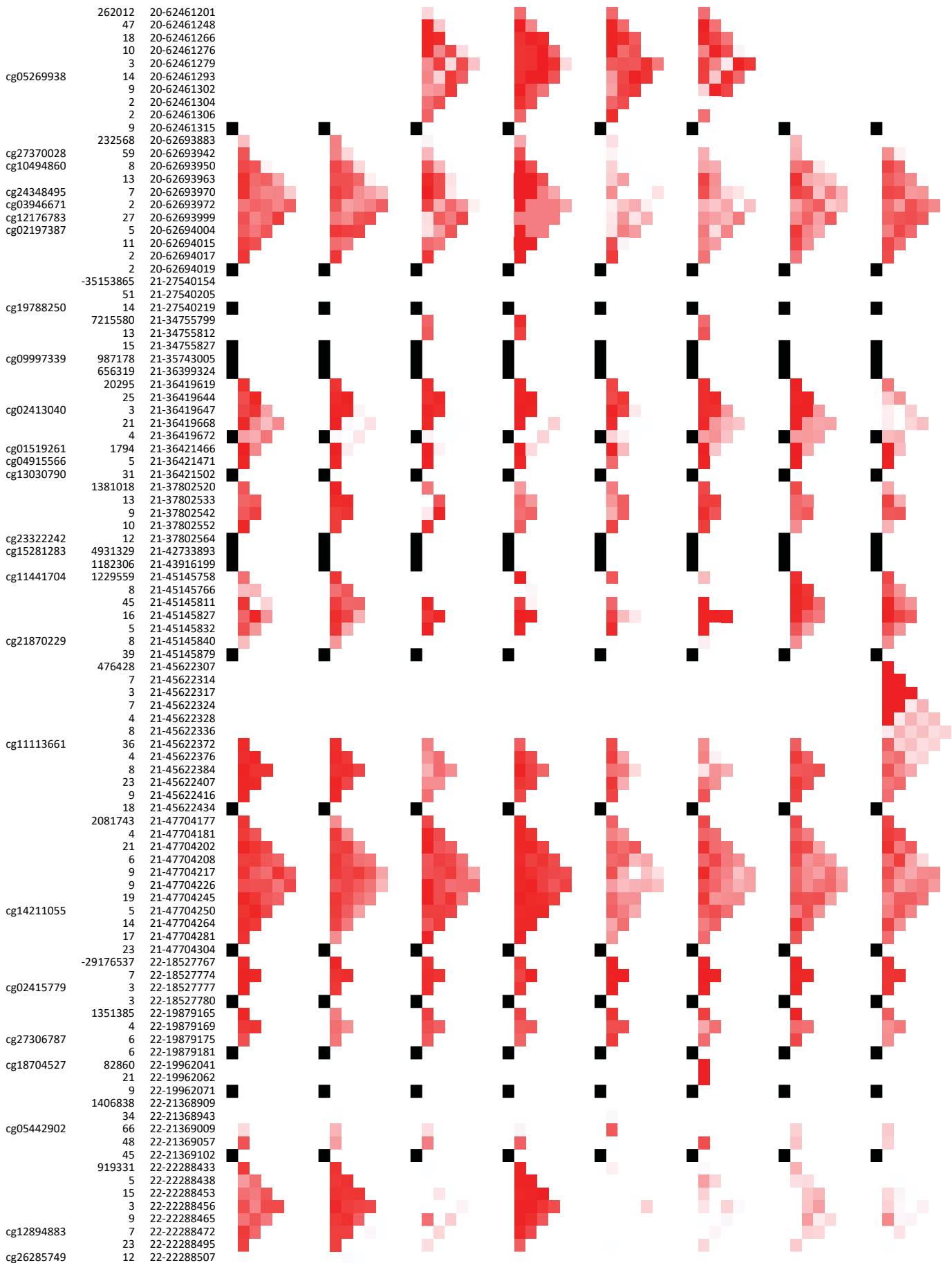








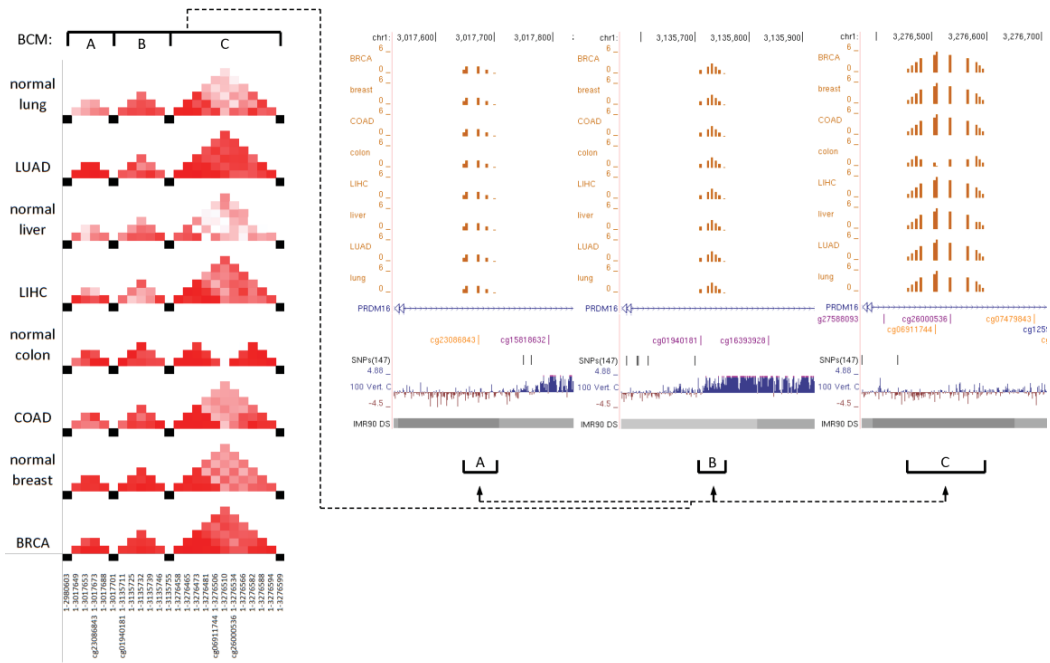




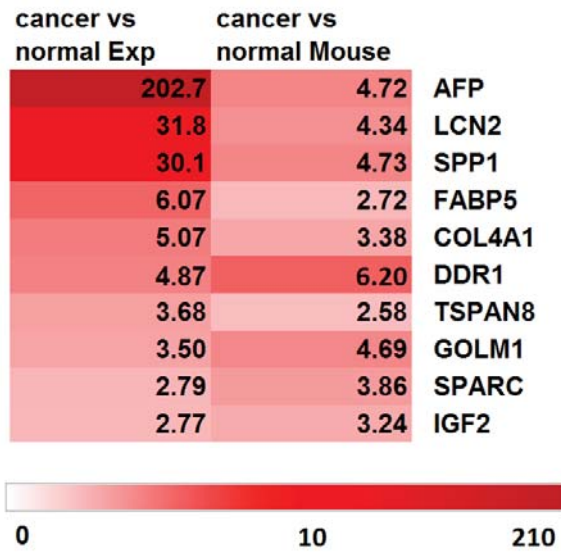
10 22-51136334
11 22-51136345
5 22-51136350



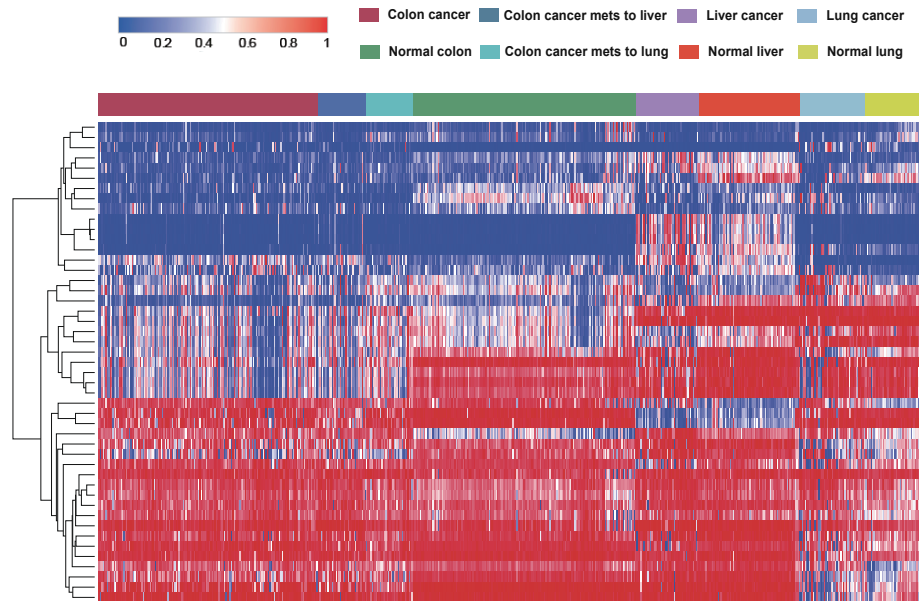
Suppl Fig 3



Suppl Fig4



Suppl Fig5



Suppl Table 1: Summary of study cohorts

	Training cohort	Validation cohort1	Validation cohort2	Total
Breast Cancer	520	270	73	863
Normal Breast	65	32	45	142
Colorectal Cancer	275	129	164	568
Colorectal Ca Mets to Liver	0	0	30	30
Colorectal Ca Mets to Lung	0	0	34	34
Normal Colon/rectal	27	18	161	206
Liver Cancer	239	138	46	423
Normal Liver	32	18	73	123
Lung Cancer	585	254	47	886
Normal Lung	49	25	45	119
Total	1792	884	718	3394

Suppl Table 2: Clinicopathological Characteristics of TCGA cohort

TCGA cohort				
Characteristic	Breast Ca	Colon Ca	Liver Ca	Lung Ca
Total (n)	790	404	377	839
Gender -no.(%)				
Female	774(98)	187(46)	123(33)	337(40)
Male	9(1)	217(54)	254(67)	491(59)
NA	7(1)	0(0)	0	11(1)
Age at diagnosis-yrs				
Mean	59	65	61	66
Range	26-90	31-90	16-90	33-90
Race -no.(%)				
White	571(72)	275(68)	187(50)	628(75)
Asian	38(5)	12(3)	161(42)	13(1)
Others	181(23)	117(29)	29(8)	198(24)
Smoking -no.(%)				
Non-smoker	NA	NA	NA	79(10)
Current reformed smoker	NA	NA	NA	507(60)
Current smoker	NA	NA	NA	220(26)
NA	NA	NA	NA	33(4)
Tumor status -no.(%)				
Tumor Free	620(75)	248(61)	236(63)	499(60)
With Tumor	65(82)	99(25)	114(30)	165(19)
NA	105(13)	57(14)	27(7)	175(21)
Vital status -no.(%)				
Alive	713(90)	355(88)	286(76)	607(73)
Dead	70(9)	49(12)	91(24)	221(26)
NA	7(1)	0(0)	0	11(1)
AICC stage -no.(%)				
Stage I	127(16)	64(16)	175(46)	420(50)
Stage II	438(56)	147(36)	87(23)	253(30)
Stage III	200(25)	122(30)	86(23)	127(15)
Stage IV	11(1)	57(14)	6(2)	25(3)
NA	14(2)	14(4)	23(6)	13(2)

Suppl Table 3: Clinicopathological Characteristics of Chinese cohort

Chinese cohort						
Characteristic	Breast Ca	Liver Ca	Lung Ca	Colorectal Ca	Colorectal Ca Mets to Liver	Colorectal Ca Mets to Lung
Total (n)	73	46	47	164	30	34
Gender -no.(%)						
Female	73(100)	8(17)	14(30)	52(32)	12(40))	9(26.5)
Male	0	38(83)	31(67)	74(45)	17(57)	25(73.5)
NA	0	0	2(4)	38(23)	1(3)	0
Age at diagnosis-yrs						
Mean	47	50	57	57	56	59
Range	29-63	21-60	37-83	27-73	30-77	32-81
Race -no.(%)						
White	0	0	0	0	0	0
Asian	73(100)	46(100)	47(100)	164(100)	30(100)	34(100)
Others	0	0	0	0	0	0
Smoking -no.(%)						
Non-smoker	NA	NA	NA	NA	NA	NA
Current reformed smoker	NA	NA	NA	NA	NA	NA
Current smoker	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA
Tumor status-no.(%)						
Tumor Free	NA	NA	NA	NA	NA	NA
With Tumor	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA
Vital status -no.(%)						
Alive	NA	NA	NA	NA	NA	NA
Dead	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA
AJCC stage -no.(%)						
Stage I	3(4)	22(48)	NA	4(2)	0	0
Stage II	13(18)	3(6)	NA	39(24)	0	0
Stage III	9(12)	9(20)	NA	23(14)	0	0
Stage IV	2(3)	0	NA	48(29)	30(100)	34(100)
NA	46(63)	12(26)	NA	50(31)	0	0

Supplemental Table 4: List of markers presented in at least 7 out of 10 random split analyses in a multinomial analysis

BRCA cancer	"cg01327147" "cg02680086" "cg04772948" "cg04917276" "cg05395187" "cg07493516" "cg08268679" "cg08549335" "cg09819083" "cg13976210" "cg14817783" "cg15412918" "cg18482112" "cg20069090" "cg24732563"
BRCA normal	"cg00886954" "cg23690893"
COAD cancer	"cg08088171" "cg13420112" "cg14642259" "cg20973720" "cg24583770" "cg27536151"
COAD normal	"cg24741563" "cg22979615"
LIHC cancer	"cg07360250" "cg08550839" "cg13499300"
LIHC normal	"ch.7.135065R" "cg14054357"
LUNG cancer	"cg01602690" "cg03993087" "cg04383154" "cg04933208" "cg05346286" "cg05784951" "cg06352912" "cg06800849" "cg07464206" "cg07903001" "cg08089301" "cg14419975" "cg15545942" "cg15963326" "cg17894293" "cg19924352" "cg20691722" "cg21845794" "cg24398479"
LUNG normal	"cg03169557" "cg04549287" "cg07649862" "cg08682723"

Supplemental Table 5. List of markers selected in 3 out of 10 training / test split analyses in survival analysis of breast cancer

LASSO	"cg01402068" "cg08858662" "cg10390979" "cg11145055" "cg14273607" "cg15145148" "cg23933602"
Boosting	"cg02837122" "cg08384322" "cg08858662" "cg11145055" "cg14273607" "cg18751588" "cg20661083" "cg25868675"
Overlapping markers	"cg08858662" "cg11145055" "cg14273607"

Supplemental Table 6. List of markers selected in 3 out of 10 training / test split analyses in survival analysis of lung cancer (LUAD/LUSC)

LASSO	"cg00221494" "cg00620629" "cg01043831" "cg01580888" "cg02504465" "cg02880679" "cg02909790" "cg03266453" "cg03316864" "cg03998173" "cg05216141" "cg05335315" "cg05338167" "cg05556202" "cg05589246" "cg05910970" "cg06462703" "cg06583518" "cg07034561" "cg07559730" "cg07997737" "cg08465774" "cg08980578" "cg09283635" "cg10003443" "cg10171125" "cg10729531" "cg10821722" "cg11132751" "cg11225410" "cg11340260" "cg11946503" "cg12291552" "cg13140267" "cg13618372" "cg13975369" "cg14047008" "cg14354749" "cg16042149" "cg16540704" "cg16744741" "cg16984812" "cg17029168" "cg17095731" "cg17606785" "cg18075299" "cg18275051" "cg19011603" "cg19107595" "cg19149785" "cg19221959" "cg19297232" "cg19308222" "cg19371795" "cg19427610" "cg19728382" "cg19928450" "cg20634573" "cg20895028" "cg21835643" "cg22791453" "cg22918700" "cg23131007" "cg23389061" "cg23412777" "cg23743114" "cg24335149" "cg24363955" "cg24402880" "cg24456340" "cg25526759" "cg25657700" "cg26133068" "cg27558666" "cg27651218"
Boosting	"cg00221494" "cg00620629" "cg01043831" "cg01580888" "cg02909790" "cg03266453" "cg03316864" "cg05335315" "cg05338167" "cg05556202" "cg05910970" "cg07997737" "cg08465774" "cg09283635" "cg10003443" "cg10171125" "cg10729531" "cg11132751" "cg11225410" "cg11340260" "cg11946503" "cg12291552" "cg12985418" "cg13140267" "cg13482233" "cg13539030" "cg13618372" "cg14354749" "cg14839257" "cg15522957" "cg16744741" "cg16984812" "cg17029168" "cg17095731" "cg17606785" "cg17965019" "cg19149785" "cg19221959" "cg19297232" "cg19371795" "cg19728382" "cg19928450" "cg20895028" "cg22791453"

"cg23389061" "cg23547073" "cg24335149" "cg24363955"
"cg24456340" "cg25526759" "cg25657700" "cg26133068"

Overlapping
markers

"cg00221494" "cg00620629" "cg01043831" "cg01580888"
"cg02909790" "cg03266453" "cg03316864" "cg05335315"
"cg05338167" "cg05556202" "cg05910970" "cg07997737"
"cg08465774" "cg09283635" "cg10003443" "cg10171125"
"cg10729531" "cg11132751" "cg11225410"
"cg11340260" "cg11946503" "cg12291552" "cg13140267"
"cg13618372" "cg14354749" "cg16744741" "cg16984812"
"cg17029168" "cg17095731" "cg17606785" "cg19149785"
"cg19221959" "cg19297232" "cg19371795"
"cg19728382" "cg19928450" "cg20895028" "cg22791453"
"cg23389061" "cg24335149" "cg24363955" "cg24456340"
"cg25526759" "cg25657700" "cg26133068"

Supplementary Table 7: Positive fold changes (FC) in expression of top ten genes of interest in human LIHC versus normal liver tissue and its correlation in mouse LIHC

gene	markerName	location	dm Ca vs No	liverCancer MethMeans	meanNormal LiverMeth	cancer vs normal Exp	cancer vs normal Mouse
AFP	cg10778295	promoter	-7%	81%	88%	202.68	4.7169
LCN2	cg13518265	promoter	-7%	70%	78%	31.85	4.3429
SPP1	cg00088885	promoter	-22%	49%	72%	30.08	4.7306
FABP5	cg01962077	promoter	-12%	74%	87%	6.07	2.7162
COL4A1	cg27546237	promoter	-1%	16%	17%	5.07	3.376
TSPAN8	cg27304204	promoter	-18%	46%	63%	4.87	6.2049
DDR1	cg00934322	promoter	-23%	43%	66%	3.68	2.5785
GOLM1	cg13925220	promoter	-9%	80%	88%	3.50	4.6902
SPARC	cg13319042	promoter	-22%	38%	60%	2.79	3.8594
IGF2	cg14890224	promoter	-21%	49%	69%	2.77	3.2437

Supplemental Table 8: Breast cancer: validation and prediction performance in survival analysis

LASSO	ρ^2 on training	0.86 (min: 0.59 - max: 0.98)
	c-index on test	0.63 (min: 0.56 - max: 0.71)
Boosting	ρ^2 on training	0.90 (min: 0.75 - max: 0.99)
	c-index on test	0.61 (min: 0.54 - max: 0.68)
log-rank LASSO	5/10 p-value < 0.05	
log-rank Boosting	3/10 p-value < 0.05	

Supplemental Table 9: Lung cancer: validation and prediction performance in survival analysis

LASSO	ρ^2 on training	0.91 (min: 0.80 - max: 0.96)
	c-index on test	0.57 (min: 0.53 - max: 0.63)
Boosting	ρ^2 on training	0.93 (min: 0.46 - max: 0.95)
	c-index on test	0.64 (min: 0.50 - max: 0.66)
log-rank LASSO	2/10 p-value < 0.05	
log-rank Boosting	5/10 p-value < 0.05	