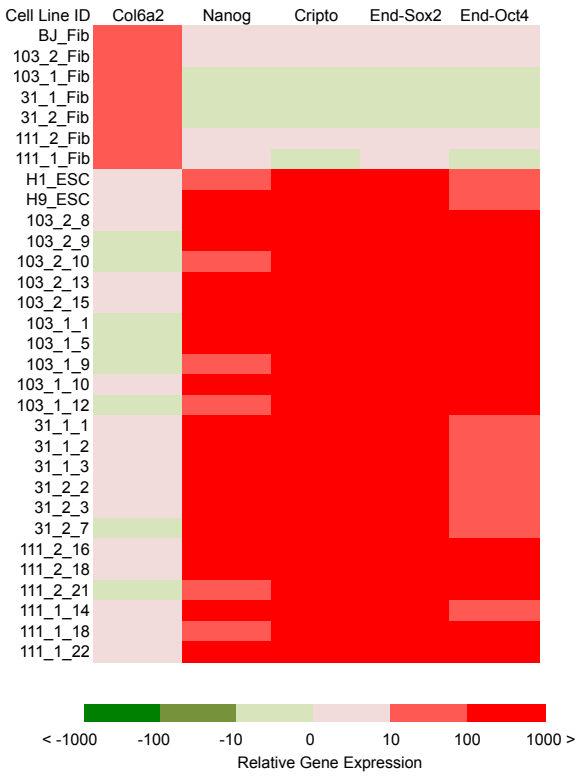
Figure S1

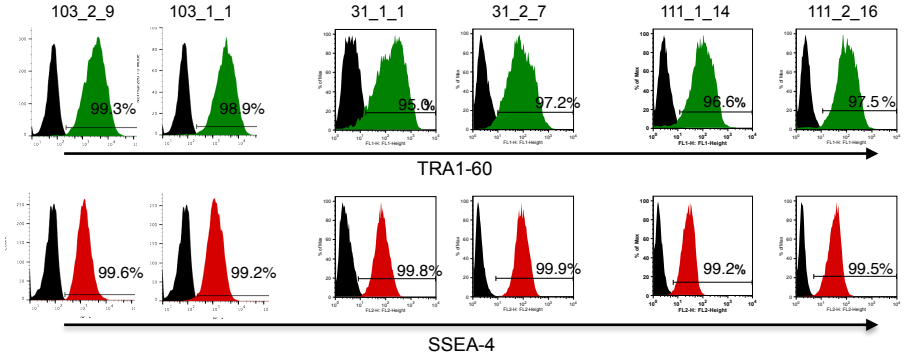
A



B

Cell Line ID	%SSEA-4	%TRA-1-60
103_2_8	99.4	99.6
103_2_9	99.6	99.3
103_2_10	99.0	99.4
103_2_13	99.3	97.9
103_2_15	98.0	98.6
103_1_1	99.2	98.9
103_1_5	98.2	97.9
103_1_9	99.2	99.6
103_1_10	99.3	99.5
103_1_12	98.4	99.6
31_1_1	99.8	95.0
31_1_2	99.9	97.6
31_1_3	99.2	96.4
31_2_2	99.9	95.9
31_2_3	99.9	95.3
31_2_7	99.9	97.2
111_2_16	99.5	97.5
111_2_18	99.8	96.7
111_2_21	99.8	98.7
111_1_14	99.2	96.6
111_1_18	98.9	95.0
111_1_22	99.1	96.7

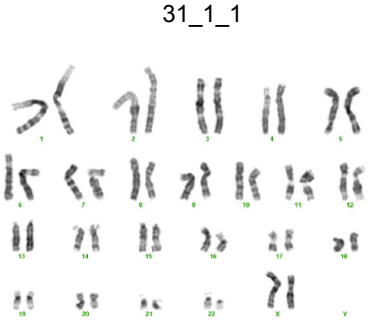
C



D

Cell Line ID	Chromosome Region	Event	Length	Cytoband
103_2_8	chr2:174,747,071-174,882,934	CN Loss	135864	q31.1
103_2_8	chr3:122,405,121-122,558,390	CN Loss	153270	q21.1
103_2_8	chr6:155,451,351-155,633,953	CN Loss	182603	q25.2 - q25.3
103_1_12	chr13:94,762,409-95,036,783	CN Loss	274375	q31.3 - q32.1
103_1_12	chr19:28,263,696-28,392,521	CN Loss	128826	q11
31_1_1	chr16:82,873,402-83,357,973	CN Loss	484572	q23.3
31_1_1	chr16:83,539,508-83,920,716	CN Loss	381209	q23.3
31_1_1	chr4:177,781,135-177,897,753	CN Loss	116619	q34.3
31_1_1	chr4:87,410,363-87,677,014	CN Loss	266652	q21.3
31_2_2	chr3:174,251,037-174,547,829	CN Loss	296793	q26.31
111_2_16	No CNVs detected			
111_1_14	No CNVs detected			

E



F

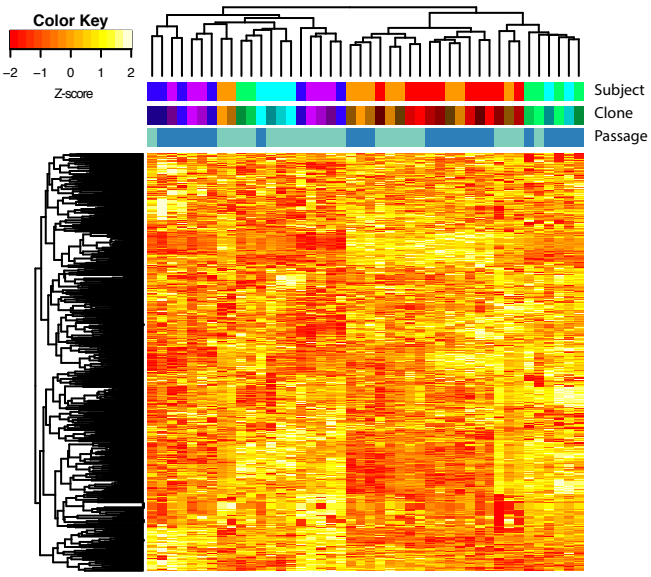


Figure S1: Characterization of iPSC lines (Related to Figure 1). A) Heat map showing expression levels as determined by real-time PCR of the pluripotent genes *NANOG*, *CRIPTO*, Endogenous SOX2 (*END-SOX2*) and Endogenous OCT4 (*END-OCT4*) relative to the fibroblast marker *COL6A2*, in controls (BJ_Fib, H1_ESC, H9_ESC), the twin somatic starting population fibroblasts as well as iPSC clonal lines derived from each individual. B) Flow cytometry analysis of the cell surface pluripotent markers SSEA-4 and TRA-1-60. All iPSC lines showed >95% expression for both markers. C) Example histograms of the flow analysis performed for iPSCs in each twin set as described in B. D) HumanCoreExome array analysis demonstrating that there are from zero to five genomic alterations in each of the iPSC lines examined at passage 20 and that the genomic alterations are relatively small (less than 0.5Mbp each). E) Standard G-banding karyotype analysis for iPSC line 31_1_1 was reported as normal. The HumanCoreExome array analysis of six iPSC lines indicated that 31_1_1 at passage 20 had the most genomic alterations (Table in panel D), so we examined this sample by standard G-banding karyotype analysis. These data demonstrate the genomic alterations identified by the array analysis are considerably smaller than those detected by standard G-banding karyotype analysis. Overall, the HumanCoreExome array and standard G-banding data confirm the normal genomic integrity of iPSC lines used in this study. F) Heatmap showing expression levels of 500 genes as determined by RNA-seq of the 42 samples at passage 9 and 20 shown in Figure 1A plus two additional samples for which we did not have methylation data. We identified the top 500 genes showing variable expression among the 44 samples and performed hierarchical clustering. Color- coding is as in Figure 1A.

Figure S2

A

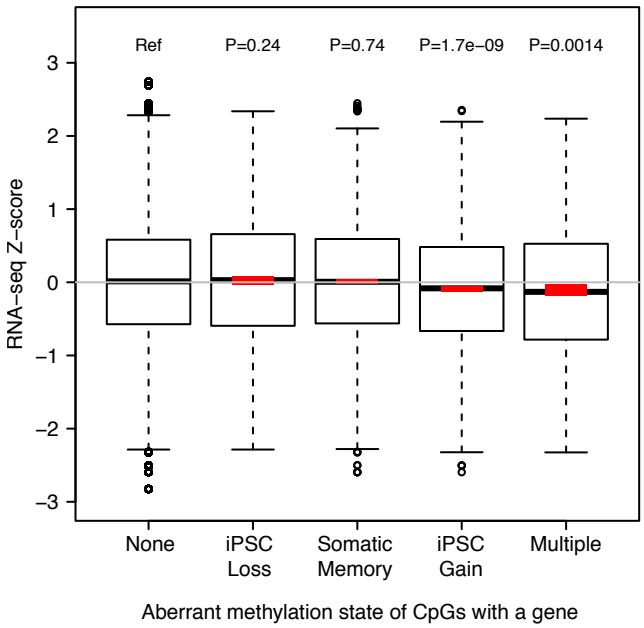


Figure S2: Impact of Aberrant methylation (Related to Figure 2). A) Boxplot showing the distributions of RNA-seq Z-scores for genes that carry at least one aberrant CpG. Each point that goes into the box plot represents a single gene expression level for a single sample. The open black boxes indicate median (thick central rectangle), and 25th and 75th percentiles (bottom and top of box). The red rectangles indicate the mean +/- 3 standard errors. A grey line is shown at 0. Multiple indicates that the gene had CpGs annotated to multiple classes of aberrancy. P-values result from the output of multiple regression of the aberrant class with the “None” class as reference on the residuals of RNA-seq Z-scores after adjusting for sample ID

Figure S3

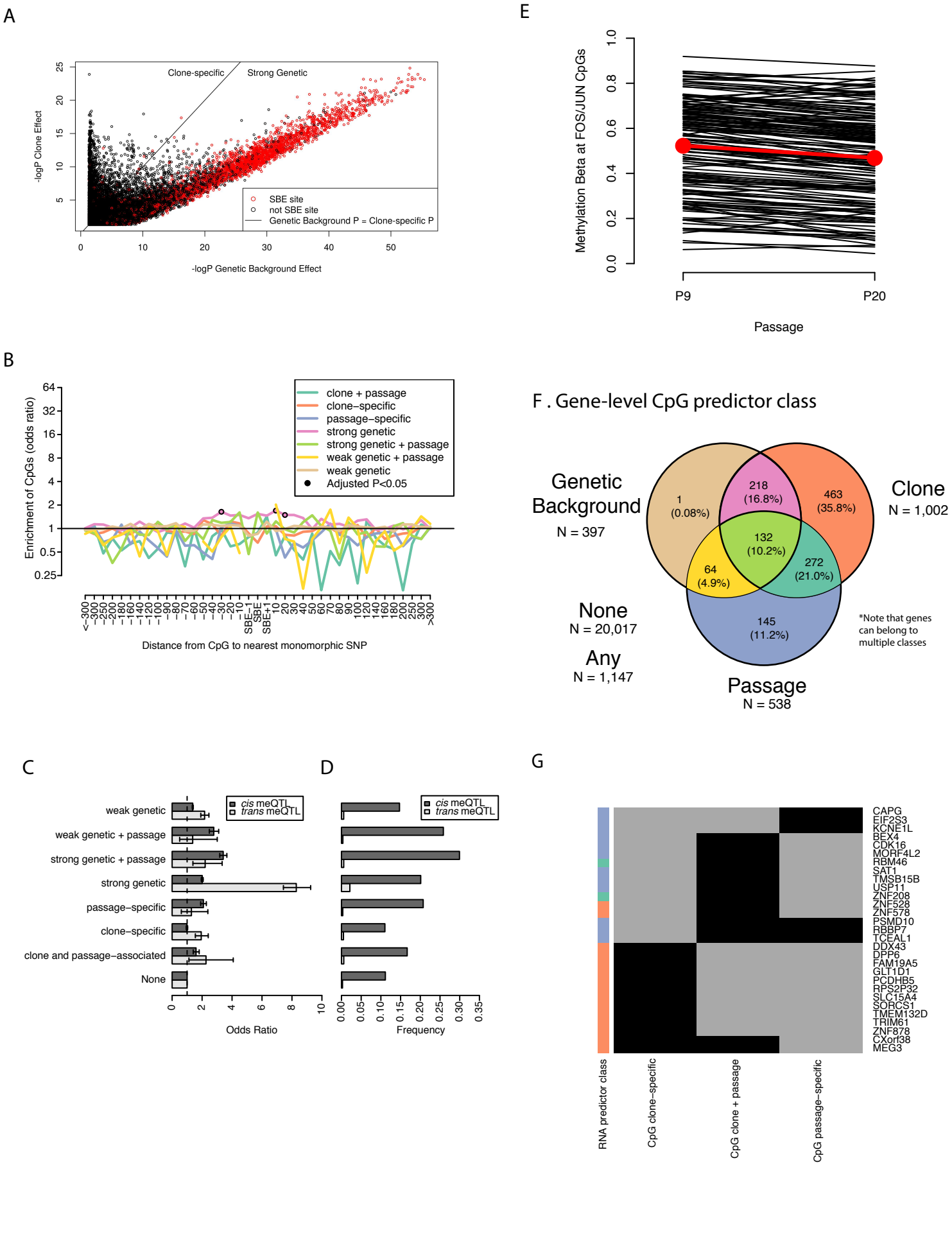
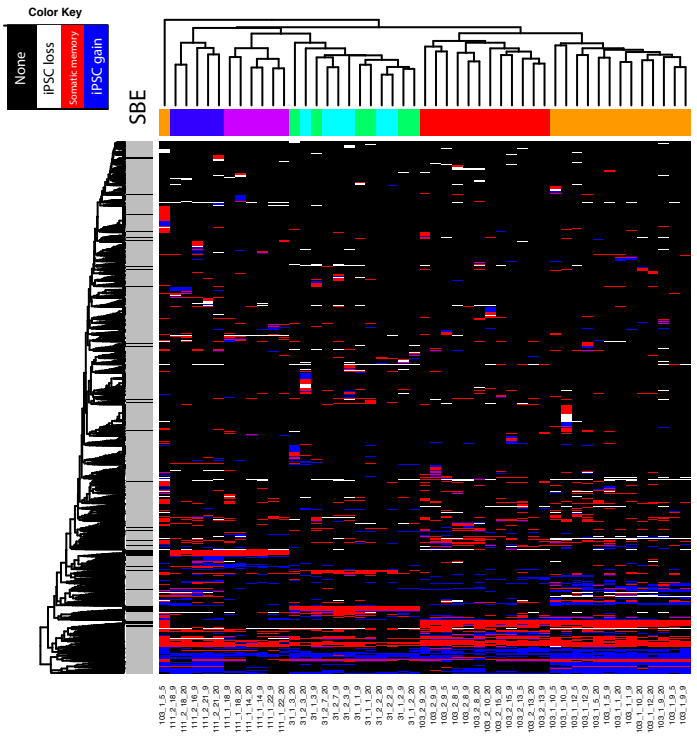

A
Clone-specific
Strong Genetic
-logP Clone Effect
-logP Genetic Background Effect
SBE site
not SBE site
Genetic Background P = Clone-specific P
E
Methylation Beta at FOS/JUN CpGs
P9
P20
Passage
B
Enrichment of CpGs (odds ratio)
clone + passage
clone−specific
passage−specific
strong genetic
strong genetic + passage
weak genetic + passage
weak genetic
Adjusted P<0.05
Distance from CpG to nearest monomorphic SNP
F . Gene-level CpG predictor class
Genetic Background
N = 397
None
N = 20,017
Any
N = 1,147
Clone
N = 1,002
Passage
N = 538
1 (0.08%)
218 (16.8%)
463 (35.8%)
132 (10.2%)
64 (4.9%)
272 (21.0%)
145 (11.2%)
*Note that genes can belong to multiple classes
C
weak genetic
weak genetic + passage
strong genetic + passage
strong genetic
passage−specific
clone−specific
clone and passage−associated
None
cis meQTL
trans meQTL
Odds Ratio
D
Frequency
G
CAPG
EIF2S3
KCNE1L
BEX4
CDK16
MORF4L2
RBM46
SAT1
TMSB15B
USP11
ZNF208
ZNF528
ZNF578
PSMD10
RBBP7
TCEAL1
DDX43
DPP6
FAM19A5
GLT1D1
PCDHB5
RPS2P32
SLC15A4
SORCS1
TMEM132D
TRIM61
ZNF878
CXorf38
MEG3
RNA predictor class
CpG clone−specific
CpG clone + passage
CpG passage−specific

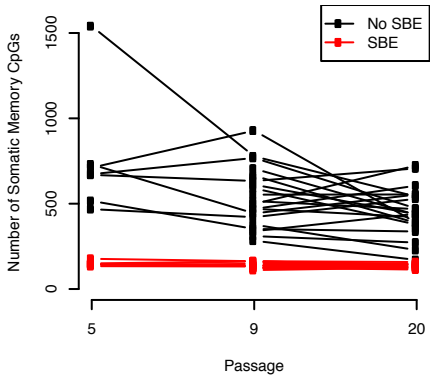
Figure S3: Characterization of predictor-associated methylation variation (Related to Figure 3). A) Relationship between -logP values for CpGs showing significant (FDR < 0.05) association with genetic background and clone. CpGs where the significance of the clone effect was greater than the significance of the genetic background effect (points above the y=x line) were called clone-specific, while those below the line were called strong genetic. CpGs with genetic variants at the single base extension (SBE) are shown in red and a show highly proportional relationship between clone and genetic background effects, consistent with the expectation of true genetic effects (see STAR Methods). B) Line plot showing odds ratios (OR) of the relationship between a CpG being associated with genetic background and harboring a monomorphic (i.e. all twin pairs carry the same genotype, either homozygous or heterozygous) genetic variant at a given distance from the probe for each CpG predictor class from Figure 2A. CpGs are grouped according to distance from SBE site (e.g. -10 includes -2 to -10 and 10 includes +2 to +10). Open black circles indicate that the association was significant at FDR < 0.05. X-axis indicates distance from SBE site. Y-axis is on a log scale. Black bars indicate the position of the assay probe or bases considered to be SBE variants. C) Barplot showing the enrichment odds ratio of *cis* and *trans* meQTLs (Lemire et al., 2015) in the seven CpG predictor classes, with variants in the None category (not associated with any of the seven CpG predictor classes) serving as a reference. Error bars indicate 95% confidence intervals. D) Frequency of meQTLs by predictor classification. Y-labels are as in D. E) Average methylation Beta values for CpGs that carry the JUN/FOS motif and show association with passage-specific effects. Each black line indicates an individual CpG and the red circles show the average for the 21 P9 iPSC samples and the 21 P20 iPSC samples (connected by a red line for ease of visualization). F) Venn diagram showing the number of genes that showed enrichment for each of the seven CpG predictor classes. Genes can show enrichment for multiple classes and thus the combinations of multiple groups may add up to less than sum of the individual cells. G) List of genes showing overlap between the gene-level CpG and RNA predictor classes for clone-specific, clone + passage, and passage classes (see Figure 3E). Only genes from significantly overlapping comparisons are shown. RNA predictor class is shown as row labels colored as in F. Black indicates a gene was associated with a category and grey indicates it was not.

Figure S4

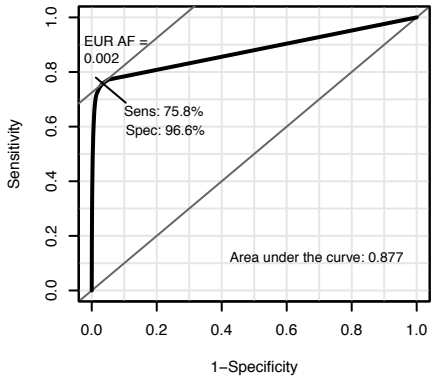
A



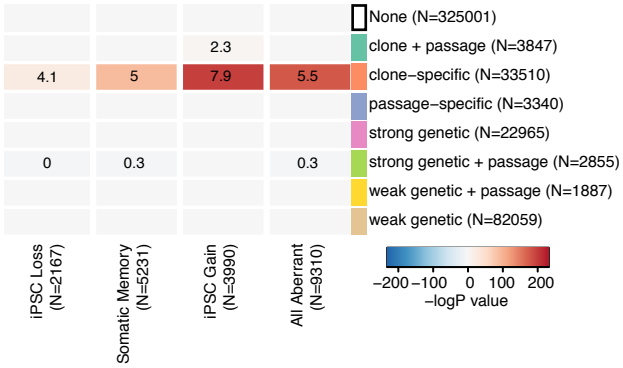
B



C



D



Example calculation for clone-specific CpGs

In aberrant region in Lister et al.

iPSC Gain this study	no	yes	Total
no	30212	1841	32053
yes	982	475	1457
Total	31194	2316	33510

OR = 7.9
P = 3.9 x 10^{-200}

E

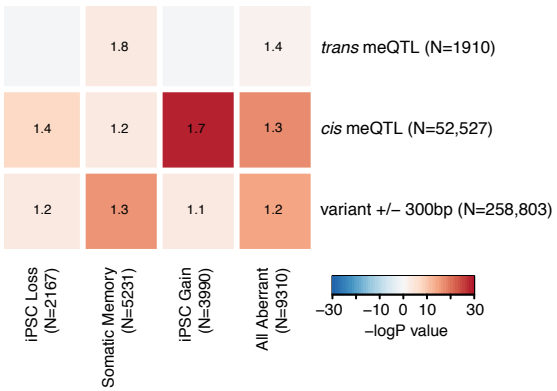


Figure S4: Overlap between Aberrant calls and CpG predictor classes (see also Figure 4). A) Heat map showing clustering of aberrantly methylated sites including sites associated with an SBE variant. Cells are colored according to whether they are not aberrant, iPSC loss, somatic memory, or iPSC gain in each of the samples. Columns are color coded by subject according to Figure 1A. The somatic memory aberrant sites associated with an SBE variant drive the clustering of samples by genetic background (twin pairs). B) Line plot showing the number of somatic memory sites called in each of the 49 samples (Figure 1A) according to whether or not the site was associated with an SBE. Each clone is shown as a line and each point corresponds to a sample. This shows that SBE sites, as opposed to other somatic memory sites, stay consistent though passages, while non-SBE somatic memory sites tend to decrease. C) Although we filtered SBE sites using WGS data, we examined the effectiveness of filtering based on reported European allele frequency if WGS was not available. Receiver operating characteristic curve showing the predictive ability of the European ancestry population-level frequency estimates of SBE variation on the presence of SBE as assayed by whole genome sequencing in a specific individual (103_2). D) Estimation of recurrently identified aberrant methylation across studies within aberrant and CpG predictor classes. For CpGs within each CpG predictor class and for each aberrant CpG class, we estimated the concordance between the CpG being in an aberrant region identified by Lister et al. (Lister et al., 2011) and the CpG being associated with aberrant methylation in this study. A Fisher's exact test was performed and the odds ratio is reported in the cell. The cells are color coded by the –log of the P-value. Only values where the significance exceeded FDR < 0.05 are shown. An example calculation showing the replication between Lister et al. sites and iPSC gain sites for clone-specific sites is shown to the right. E). Overlap between aberrant methylation and whether the CpG was previously associated with an meQTL or whether the CpG had a polymorphic genetic variant within +/-300bp of the probe. Values reflect OR's from a Fisher's exact test and the color reflects the –log of the P-value.

Table S4. Whole genome sequencing summary statistics (Related to Figure 3 and STAR Methods). Summary statistics (SNPs called, indels called, TiTv ratio, and SNP non-reference concordance with twin) are shown for whole genome sequencing of blood samples from 6 individuals.

Subject	SNPs called	Indels called	TiTv	SNP non-reference concordance with twin
31_1	3580838	512303	2.088	0.981
31_2	3590181	519148	2.088	0.978
111_1	3556408	495038	2.090	0.978
111_2	3565098	501552	2.090	0.976
103_1	3659059	544374	2.088	0.979
103_2	3647802	533869	2.089	0.982
Average	3603709.6	518796.2	2.089	0.979