TruDiagnostic

DNA Methylation Data Pre-Processing

Natàlia Carreras Gallo and Juan Ramón González

TruDiagnostic™ The Epigenetic Company

ISGlobal Barcelona Institute for Global Health

Table of Contents

| General information

Array: IlluminaHumanMethylationEPIC
Quality control software: *meffil*
N (initial): 5,816 → N (final): 3,424
Probes (initial): 865,859 → Probes (end): 740,023

| SampleSheet from idat files

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/A_Create_SampleSheet.R
  Time ~ 2 min

We first separated the initial idat files (6,187) in two folders:

- **idat_remove** (387 idat files)
  This folder contains the negligible data (most of them are fictitious data). It consists of 6 smaller folders: 08142020 Pre Open DMAPs, Duplicates, 205735180078, iScan Comp Comparison, Redo, and Test Run 02162022. We avoided these idat files.
- **idat_use** (5,816 idat files)
  Here, we collected all the idat files that should be considered in the following steps.

We created a SampleSheet using the idat files from the idat_use folder with the *meffil.create.sample.sheet* function. This function generated a *data.frame* of 5,816 rows (IDs).

Since we needed the sex annotation of these individuals for the sample Quality Control (QC), we compared the IDs from the SampleSheet with the ones in the Patient Metadata file (after removing those individuals with a BMI out of the range 10-60 and with intersex sex).

At the end, we obtained 3,599 individuals matching between the SampleSheet and the Metadata with Male/Female sex. However, we removed three individuals (205772280052_R06C01, 205772280052_R07C01, and 205772280052_R08C01) that showed troubles in the QC step and 6 individuals that were duplicated (205772280137_R08C01, 205772280146_R01C01, 205772290045_R01C01, 205828610080_R05C01, 205832310130_R08C01, and 205832310143_R07C01), leading to a final SampleSheet of 3,590 individuals with biological sex annotated (**Figure 1**).

| Sample Quality Control

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/B_Sample_QC.R
- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/B_call_SampleQC.sh
  Time ~ 4h 30 min / Mem ~ 23 Gb

Using the SampleSheet previously mentioned with 3,590 IDs, we performed the sample QC using the *meffil.qc* function with the "blood gse35069 complete" as reference.

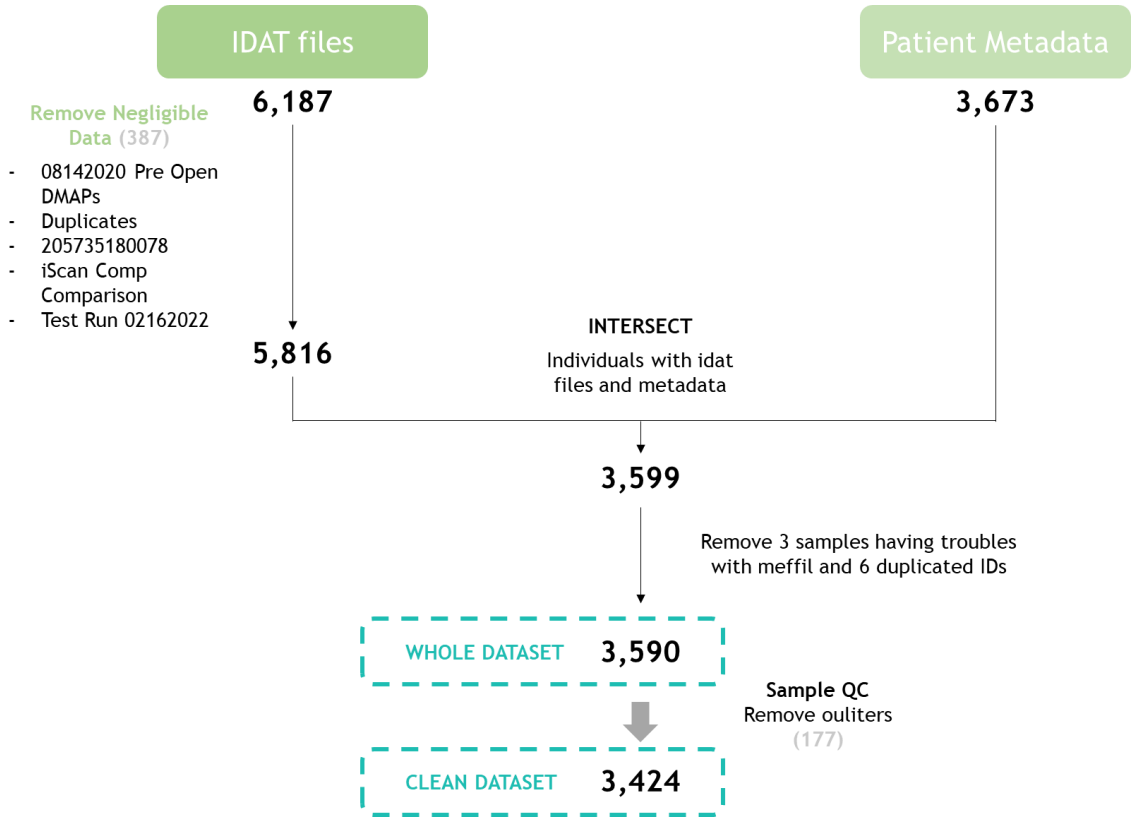


Figure 1 | Selection of the individuals based on the available metadata and the sample Quality Control (QC).

We used the default parameters for the QC report:

- detection.threshold = 0.01,
- bead.threshold = 3
- beadnum.samples.threshold = 0.05
- detectionp.samples.threshold = 0.05
- detectionp.cpgs.threshold = 0.05
- beadnum.cpgs.threshold = 0.05
- sex.outlier.sd = 3

The QC report file can be find at the following path:

/PROJECTES/GENOMICS/TruDiagnostic/prepro_files/QC/qc-report_whole.html

The report showed outliers for a lot of conditions but we only selected the outliers based on:

- Control probe (dye.bias) – 7
- Methylated vs Unmethylated - 74
- X-Y ratio outlier - 55
- Low bead numbers - 1
- Detection p-value - 1
- Sex mismatch - 36
- Control probe (bisulfite1) - 0
- Control probe (bisulfite2) - 0

Among them, we found 7 samples with more than one issue:

- 205676380102_R02C01
  - ❖ Sex mismatch
  - ❖ X-Y Ratio Outlier
- 205676390016_R08C01
  - ❖ Sex mismatch
  - ❖ Detection p-value
  - ❖ X-Y Ratio Outlier
- 205676390106_R03C01
  - ❖ Methylated vs Unmethylated
  - ❖ X-Y Ratio Outlier
- 205772280075_R02C01
  - ❖ Methylated vs Unmethylated
  - ❖ X-Y Ratio Outlier
- 205772280075_R03C01
  - ❖ Methylated vs Unmethylated
  - ❖ X-Y Ratio Outlier
- 205772280091_R04C01
  - ❖ Sex mismatch
  - ❖ X-Y Ratio Outlier
- 205832330027_R08C01
  - ❖ Methylated vs Unmethylated
  - ❖ X-Y Ratio Outlier

In total, the outliers represented 166 samples. We decided to remove all of them and continue the analysis with 3,424 individuals (**Figure 1**). We estimated the cellular composition based on methylation levels and we generated another QC report with these selected individuals:

/PROJECTES/GENOMICS/TruDiagnostic/prepro_files/QC/qc-report_clean.html

| Functional Normalization

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/C_Functional_Normalization.R
- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/C_call_FunctNorm.sh
  Time ~ 11h 30min / Mem ~ 100 Gb

The next step is to normalize the CpG methylation values. To this end, we first estimate the number of principal components to use based on the methylation levels of the control probes (**Figure 2**). Looking at the plot, we considered that 10 PCs was a good approximation for the normalization.

Previously to generate the beta values, we set poorly detected methylation values to missing. Poor signal was identified during QC as signal that failed to pass the detection p-value threshold (0.01) or bead threshold (3). Moreover, we removed probes that have more than 5% of poorly detected values. In total, we removed 28,117 probes. Among them, 3,297 had poor detection p-value, 24,365 failed the bead threshold, and 455 failed both thresholds (**Figure 3**).

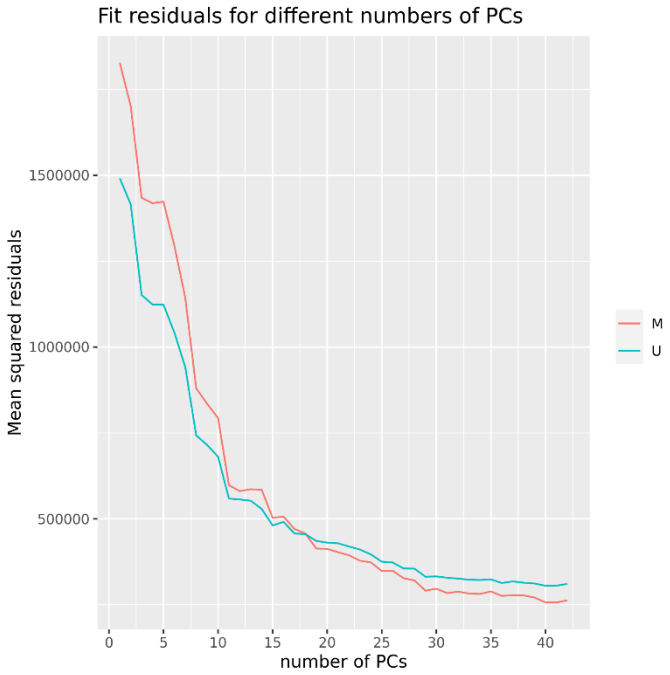


Figure 2 | Plot showing the fit of the residuals for different numbers of principal components (PCs). M: methylated; U: unmethylated.

During the normalization process, we decided to save the output to a GDS (Genomic Data Structure) because of the magnitude of the output and the high memory demand.

Once we got the normalized betas, we removed the CpG sites that had accumulated more than 5% of missing (1,110 probes), leading to a total of 836,632 CpG sites. In the case of the IDs, there were not individuals with more than 5% of missing. In the final norm.beta object we had 0.34% of missing.

Finally, we calculated the principal components of normalized betas based on the 50,000 most variable CpG sites (this is the value by default).

We created a normalization report using 4 variables as batches (slide, sex, Sentrix_Row, and Sentrix_Col) and the default parameters:

- control.pcs=1:5
- batch.pcs=1:5
- batch.threshold=0.01

The report can be accessible at:

/PROJECTES/GENOMICS/TruDiagnostic/prepro_files/Functional_Normalization/normalization-report_clean.html

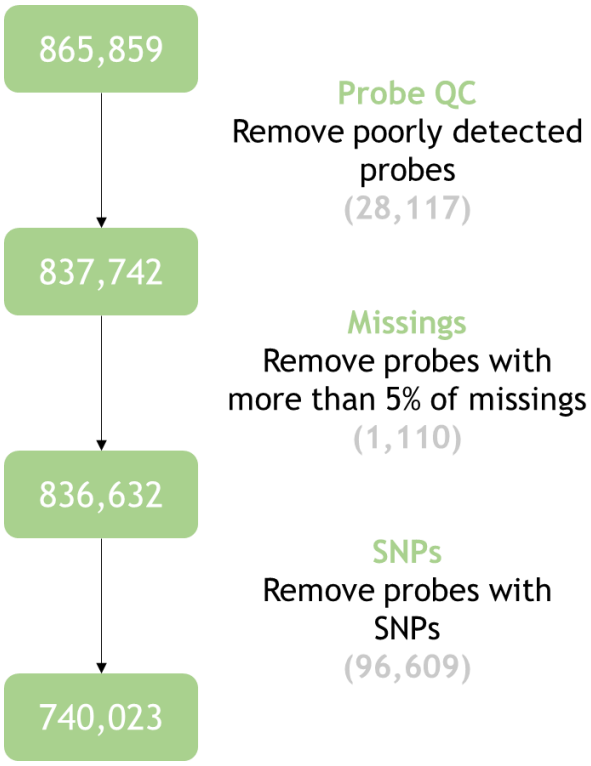


Figure 3 | Selection of the CpG sites based on the Quality Control (QC), the number of missing, and the probes with SNPs. The removal of probes with SNPs is based on the InfiniumAnnotation from https://zwdzwd.github.io/InfiniumAnnotation.

| Create GenomicRatioSet

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/D_Create_GenomicRatioSet.R
- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/D_call_GRset.sh
  Time ~ 2h 15min / Mem ~ 185Gb

The GenomicRatioSet (GRset) is an object that can group different type of data:

- Beta values mapped to a genomic location
- Metadata → accessible with the pData() function
- Annotation data → accessible with the getAnnotation() function or with the rowData() in case you have included it there

We created a GRSet for our data using the normalized beta values from the norm.beta_clean object. For the metadata, we recodified some variables to simplify the further analyses (See "Descriptive_Analysis_metadata.html"). We joined the 120 metadata variable, the estimated cellular composition (7 variables), and the Slide variable and we included it in the GRset metadata. For the annotation information, we used the default EPIC annotation by Illumina (ilm10b4.hg19).

We created the GRSet object using the *makeGenomicRatioSetFromMatrix* function from the *minfi* package.

After creating the GRset object, we used InfiniumAnnotation from https://zwdzwd.github.io/InfiniumAnnotation to filter probes where 30bp 3'-subsequence of the probe is non-unique, probes with INDELs, probes with extension base inconsistent with specified color channel (type-I) or CpG (type-II) based on mapping, probes with a SNP in the extension base that causes a color channel switch from the official annotation, and probes where 5bp 3'-subsequence overlap with any of the SNPs with global population frequency higher than 1%.

Finally, we checked the last version in HGNC of the gene symbols and we changed the ones that were annotated using previous versions.

Our final GenomicRatioSet object contains:

- 3,424 columns (IDs)
- 740,423 rows (CpG sites)
- 128 columns in the colData (metadata + cellular composition + Slide)
- 6 columns in the rowData
- 46 columns in the annotation information

This object can be found at:

/PROJECTES/GENOMICS/TruDiagnostic/Final_datasets/GRset_clean.Rdata

| Principal Component Analysis

In the normalization-report, we observed that there was a slightly division of the normalized betas into two clusters when comparing PC1 and PC2 (**Figure 4**). We tested whether the first 3 PCs were associated with any of the variables from the metadata that were potential variables to show big differences in methylation: sex, ethnicity, slide, age, cell type and collection date (**Table 2**). In **Figure 4**, we can see graphically the association between sex, ethnicity, and collection year with the first 3 PCs. In **Table 2**, we can see the most significant associations with their effect and significance. Among these, the different cell types, the slide, and the collection date are very associated with the first 3 components. The age, sex, and ethnicity are also associated with the PC1, but the significance and the correlation are lower.

Although we have found different variables that are associated with the PCs, we performed a Surrogate Variable Analysis (SVA) to detect batch variables that were unknown. To this end, we first imputed missing data.

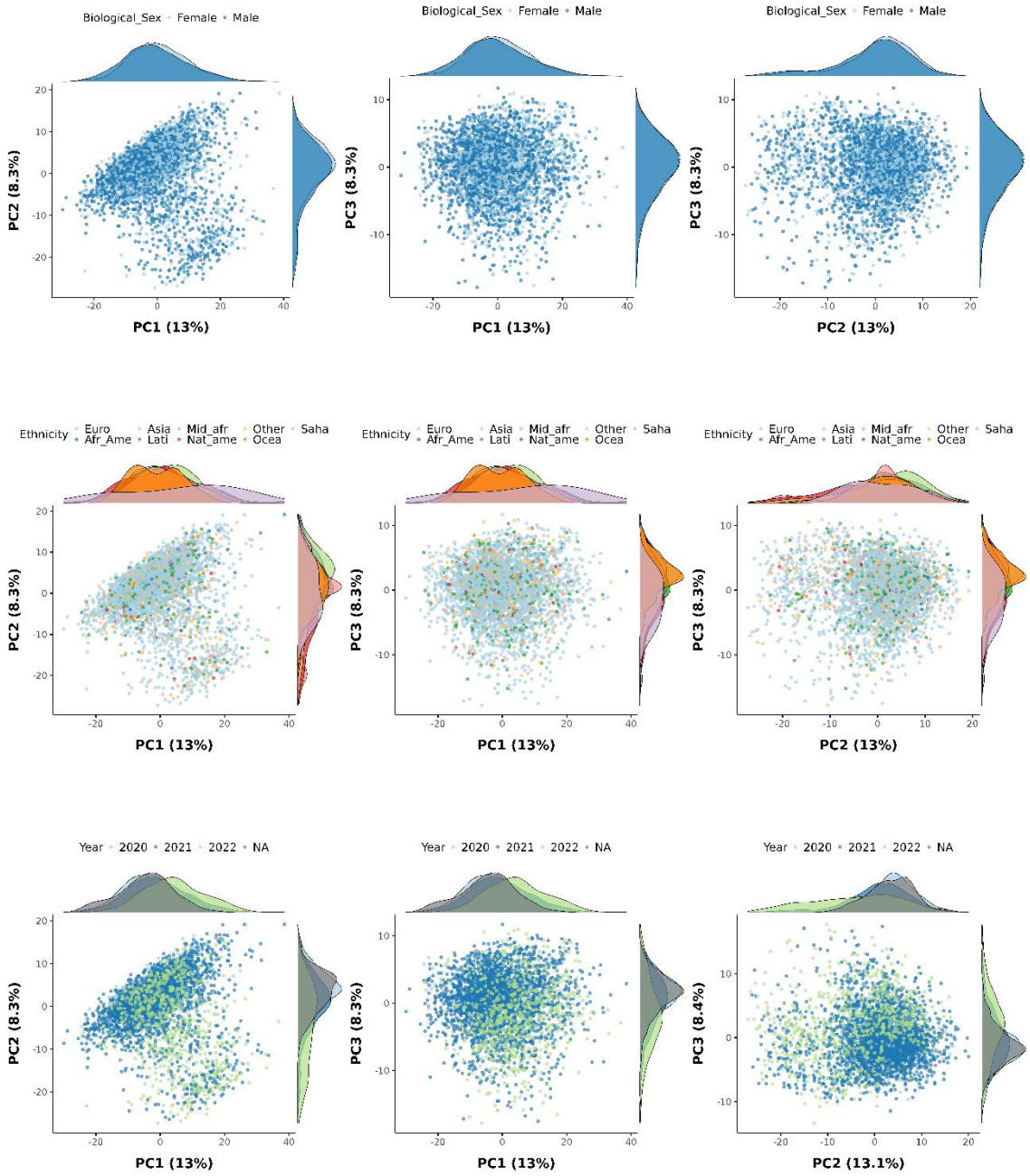


Figure 4 | Principal Component Analysis comparing the first 3 Principal Components (PCs) with Biological sex, ethnicity, and collection year in normalized beta.

Table 2 | Significant pairwise associations (p<0.000001) between variables (sex, ethnicity, slide, age, cell type, and collection date) and the first 3 Principal Components (PCs) in normalized beta. "Year" means the year when the sample was collected. "Year_Month" is calculated by multiplying the year by 12 plus the month.

x	y	test	p.value	estimate	r2
CD4T	PC1	F-test	0	2721.623083	0.442999684
Neu	PC1	F-test	0	10322.13186	0.751021015
NK	PC1	F-test	0	2323.246347	0.404377151
Bcell	PC1	F-test	1.1581E-273	1508.132326	0.305900983
Neu	PC2	F-test	1.0991E-138	689.9388481	0.16778918
CD4T	PC2	F-test	1.7935E-121	595.9287929	0.14831741
CD8T	PC2	F-test	1.43432E-94	453.3301882	0.116978468
CD8T	PC1	F-test	2.82849E-91	436.2226842	0.113063117
Year_Month	PC2	F-test	4.44468E-86	410.1926453	0.1102315
Slide	PC2	F-test	3.60693E-80	379.1273236	0.099740759
Year	PC2	F-test	1.21324E-69	326.3142668	0.08971297
Year_Month	PC1	F-test	4.74343E-52	238.7017295	0.067245574
Slide	PC1	F-test	2.0153E-48	220.6752853	0.060580554
NK	PC3	F-test	4.18784E-41	185.1520449	0.051329149
Year	PC1	F-test	5.52047E-40	179.9051618	0.051535391
CD8T	PC3	F-test	2.68267E-36	162.0260174	0.045207824
Year	PC3	F-test	3.12093E-36	161.837986	0.046601076
Year_Month	PC3	F-test	7.38517E-34	150.4612995	0.043467567
Mono	PC3	F-test	1.41682E-28	125.2231654	0.035301744
Slide	PC3	F-test	3.31958E-26	114.0206628	0.032245474
NK	PC2	F-test	1.78607E-19	82.45216775	0.023527834
Mono	PC2	F-test	1.03948E-18	78.88967139	0.022534178
Bcell	PC3	F-test	1.53303E-15	64.19268692	0.018413408
Biological_Sex	PC2	t-test	1.90839E-12	-1.82941353	0.014601133
Bcell	PC2	F-test	7.12737E-11	42.75478946	0.012339918
age	PC1	F-test	1.29471E-10	41.57303771	0.012002934
Mono	PC1	F-test	2.37329E-09	35.83178487	0.010362501
Ethnicity	PC2	t-test	3.44457E-07	-1.552481302	0.007647631

| Impute data

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/F_Impute_Data.R
  Time ~ 25 min / Mem ~ 80Gb

To perform SVA, we need data without missing. Then, we have created another GRset with imputed betas based on the median of each CpG site:

/PROJECTES/GENOMICS/TruDiagnostic/Final_datasets/GRset_clean_imp.Rdata

Again, we have tested pairwise associations between the previous variables and the first 3 PCs and the results are very similar compared with the non-imputed normalized betas (**Figure 5** and **Table 3**). Therefore, we can assume that the imputation is not altering our data.

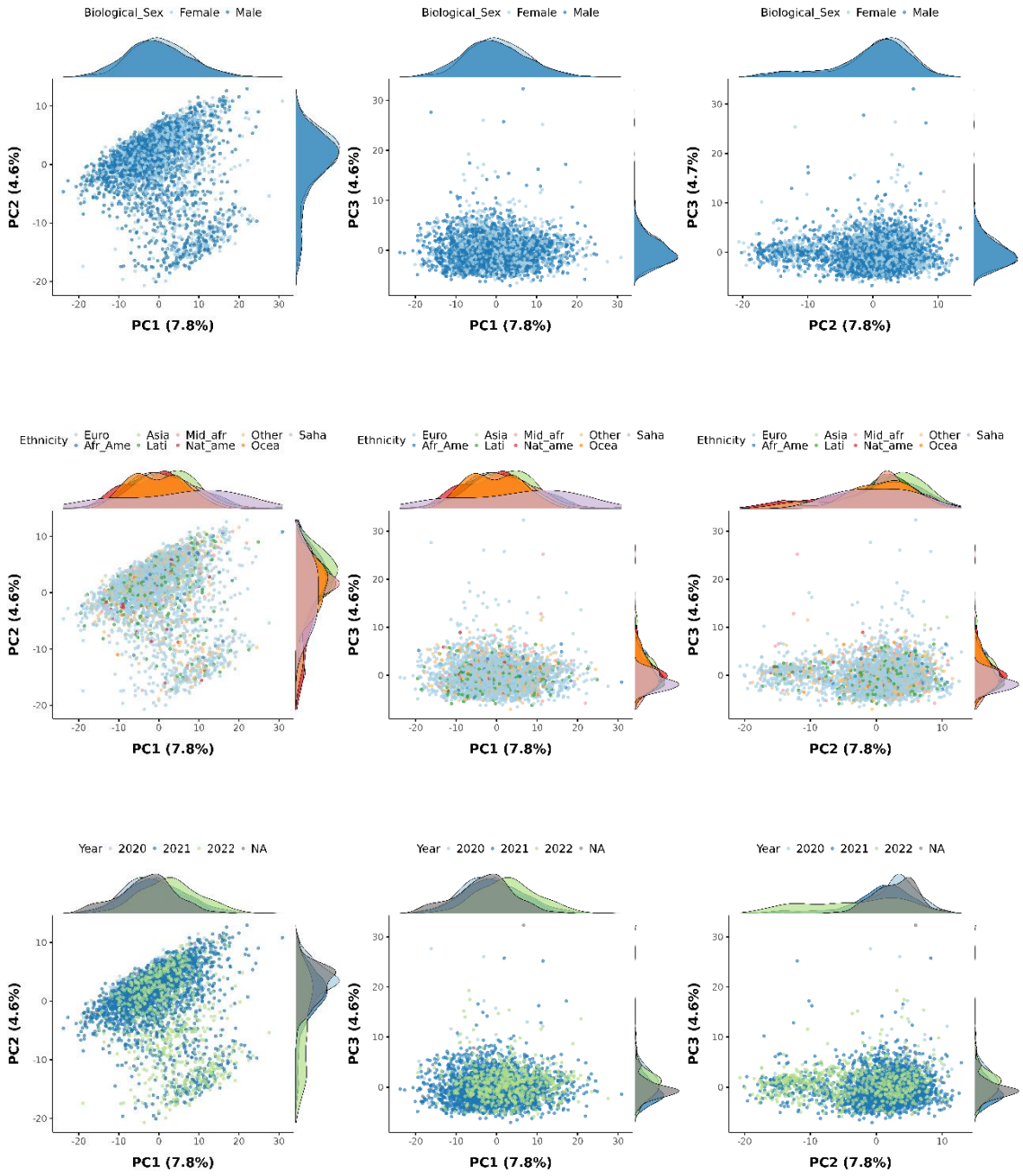


Figure 5 | Principal Component Analysis comparing the first 3 Principal Components (PCs) with Biological sex, ethnicity, and collection date in imputed normalized beta.

Table 3 | Significant pairwise associations (p<0.000001) between variables (sex, ethnicity, slide, age, cell type, and collection date) and the first 3 Principal Components (PCs) in imputed normalized beta. "Year" means the year when the sample was collected. "Year_Month" is calculated by multiplying the year by 12 plus the month.

x	y	test	p.value	estimate	r2
CD4T	PC1	F-test	0	3298.833	0.490837
Neu	PC1	F-test	0	14642.29	0.810565
NK	PC1	F-test	0	2408.706	0.413107
Bcell	PC1	F-test	1.6E-276	1527.071	0.308557
CD8T	PC1	F-test	9.3E-113	549.1642	0.138288
Year_Month	PC2	F-test	4.7E-89	425.5886	0.113898
Neu	PC2	F-test	4.92E-85	404.0305	0.1056
Slide	PC2	F-test	1.22E-84	402.0029	0.105126
CD4T	PC2	F-test	9.79E-79	371.8198	0.098007
Year	PC2	F-test	7.02E-73	342.6801	0.09379
CD8T	PC2	F-test	7.93E-66	306.6752	0.082248
Year_Month	PC1	F-test	8.34E-36	159.7895	0.046038
Mono	PC3	F-test	1.62E-35	158.2805	0.044209
Slide	PC1	F-test	2.67E-33	147.6898	0.041373
Year	PC1	F-test	2.06E-27	119.7947	0.034917
CD8T	PC3	F-test	7.9E-19	79.44316	0.022689
Biological_Sex	PC2	t-test	1.2E-18	-1.7044	0.02303
NK	PC3	F-test	1.2E-17	73.95152	0.021153
Mono	PC2	F-test	1.48E-12	50.44953	0.014529
Mono	PC1	F-test	2.82E-12	49.16834	0.014165
age	PC1	F-test	9.3E-12	46.79456	0.01349
Ethnicity	PC2	t-test	3.58E-10	-1.44119	0.011619
NK	PC2	F-test	5.33E-10	38.77541	0.011204
Ethnicity	PC2	t-test	3.05E-08	-1.52286	0.009507
Biological_Sex	PC3	t-test	9.09E-08	0.535045	0.008388

| Surrogate Variable Analysis

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/G_SVA.R
- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/G_call_SVA.sh
  Time ~ 1 day / Mem ~ 150Gb

In the SVA analysis, we included some covariates in the model in order to keep their effect on DNA methylation: marijuana, biological sex, ethnicity, age, neuropsychological, cardiovascular, respiratory, and endocrine diseases, tobacco, alcohol, amphetamines, benzodiazepines, hallucinogens, and MDMA use, and drug or alcohol addiction for mother or father. We did not include cell type nor slide nor collection date because they were not variables of our interest, and we want to remove their effect on DNA methylation.

First, we estimated the number of surrogate variables (SVs) using isva::EstDimRMT and it was 127. Since it was a huge number of SVs, we decided to follow the guidelines from GTEX where they recommend using 60 SVs when N>350. (https://gtexportal.org/home/documentationPage#staticTextAnalysisMethods).

Second, we calculated the 60 SVs from our data, and we saved the object:

/PROJECTES/GENOMICS/TruDiagnostic/prepro_files/batch/sv.obj.Rdata

Third, we adjusted the beta values by these SVs and we created a new GRset with the residuals:

/PROJECTES/GENOMICS/TruDiagnostic/Final_datasets/GRset_SVA.Rdata

Finally, we calculated the first PCs from these residuals, and we evaluated again the association of these PCs with the previously variables mentioned (**Figure 6** and **Table 4**).

Table 4 | Significant pairwise associations (p<0.00001) between variables (sex, ethnicity, slide, age, cell type, and collection date) and the first 3 Principal Components (PCs) in SVA.

x	l	y	test	p.value	estimate	r2
Ethnicity		PC2	F-test	1.9412E-220	152.7985525	0.263594142
Biological_Sex		PC1	F-test	1.0874E-124	613.3341816	0.151990927
Biological_Sex	Male	PC1	t-test	2.9545E-123	-0.922146543	0.151087284
Biological_Sex	Female	PC1	t-test	9.3239E-123	0.91750348	0.150596286
Ethnicity	Other	PC2	t-test	1.8096E-114	1.285252055	0.142778418
Ethnicity	Ocea	PC2	t-test	5.2228E-100	3.736507287	0.127807482
Ethnicity	Asia	PC2	t-test	5.43227E-70	2.313987409	0.090544531
Ethnicity	Euro	PC2	t-test	8.12027E-52	-0.53141297	0.067179053
Ethnicity	Afr_Ame	PC3	t-test	1.19913E-51	-1.741490472	0.064962381
Ethnicity		PC3	F-test	8.30806E-49	31.95998289	0.069654634
Biological_Sex	Male	PC2	t-test	1.26926E-47	0.489724572	0.060903123
Ethnicity	Afr_Ame	PC2	t-test	3.12632E-25	1.003281787	0.0323071
Ethnicity		PC1	F-test	1.36925E-18	12.98417007	0.029518925
Ethnicity	Afr_Ame	PC1	t-test	3.32232E-17	-1.160572825	0.020770515
Biological_Sex		PC2	F-test	7.53485E-17	70.24919389	0.020115745
Ethnicity	Lati	PC2	t-test	3.56142E-14	0.383422047	0.017339014
Biological_Sex	Female	PC2	t-test	1.53287E-09	-0.207886878	0.010819591
Ethnicity	Euro	PC3	t-test	4.7569E-09	0.228699768	0.009996995

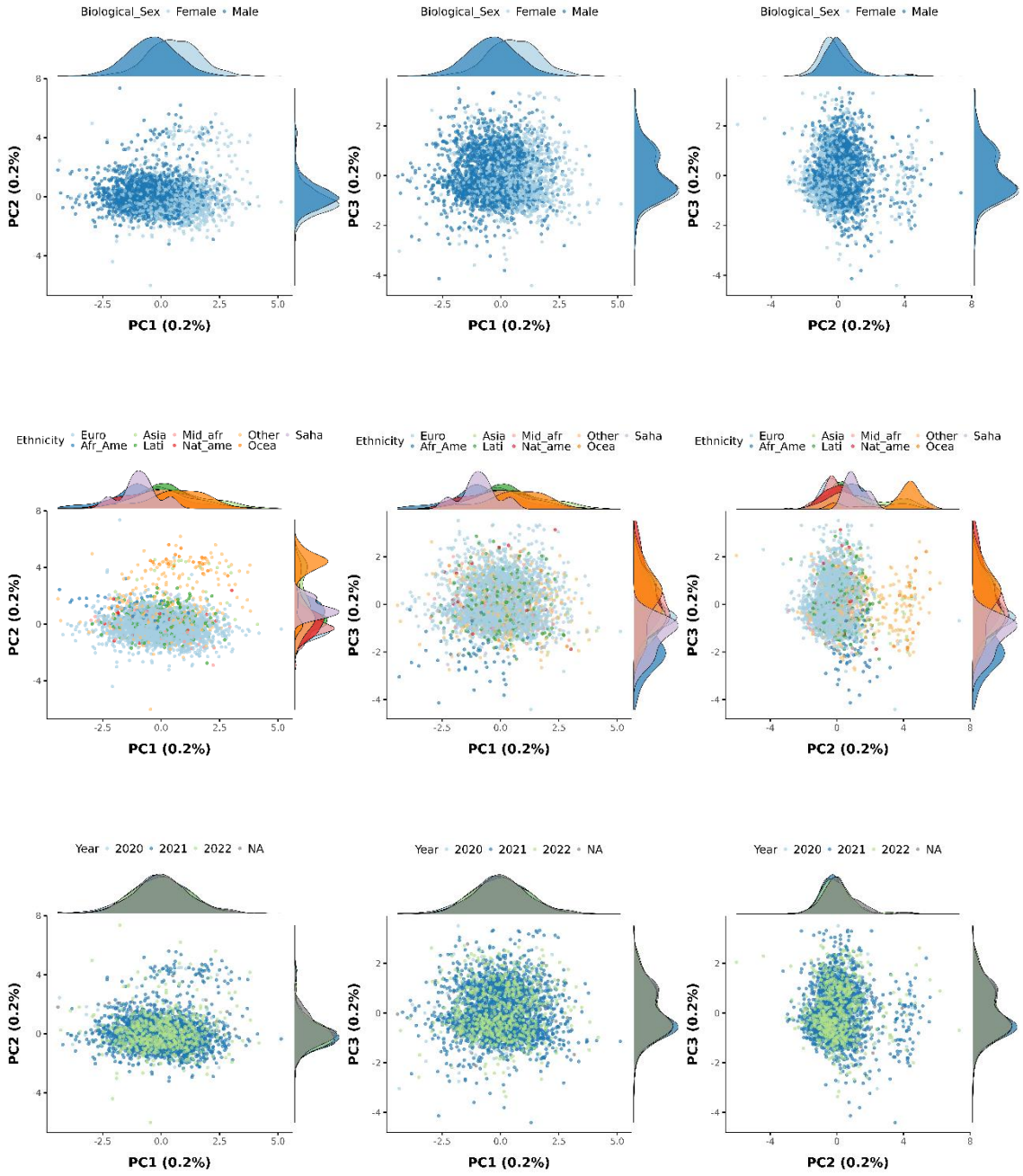


Figure 6 | Principal Component Analysis comparing the first 3 Principal Components (PCs) with biological sex, ethnicity, and collection date in SVA residuals.

We can see that the associations with cell type, slide, and collection date have disappeared, most likely because the SVs estimated explain most of their variability. To prove this, we have tested correlation between SVs and these variables (**Figure 7**). First, it is worth mentioning that slide is highly associated with Year_Month variable (r^2=0.837). This did not surprise us because the different slides have been used in different days or months. In addition, the different cell types are also correlated. Second, we can see that SV1 and SV2 are the two surrogate variables that are more correlated with covariates.

To see these correlations more in detail, we evaluated the pair-wise associations between a lot of variables (including drugs consumption and some diseases) with SVs (**Table 5**). Among them, we can see again slide, cell type, and collection date altogether with sex and age. Sex is mainly correlated with SV19 and age is mainly correlated with SV17.

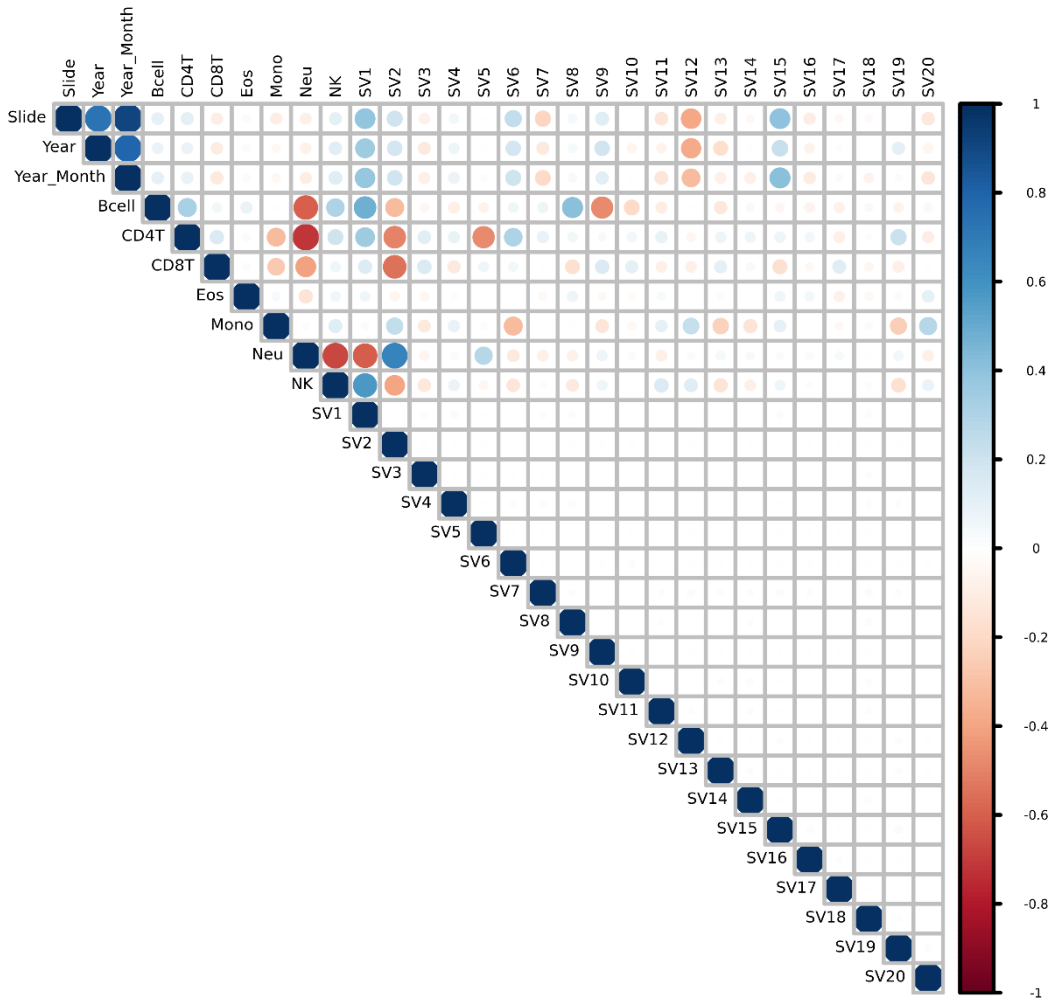
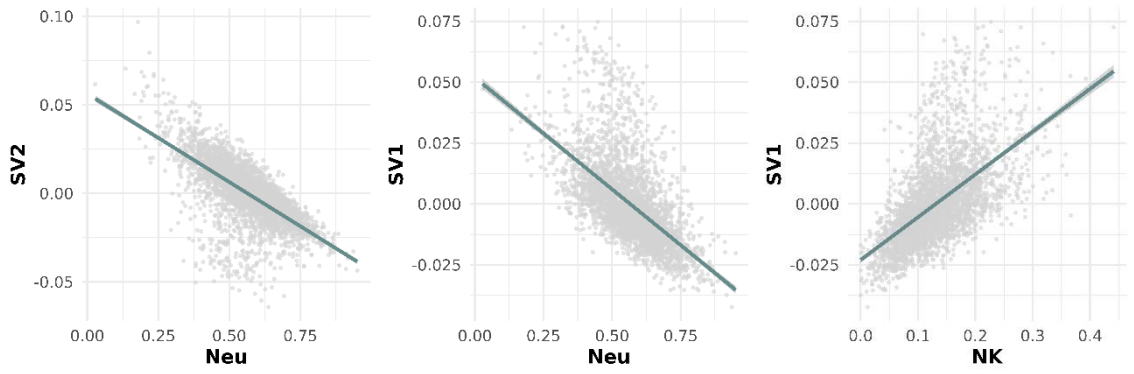
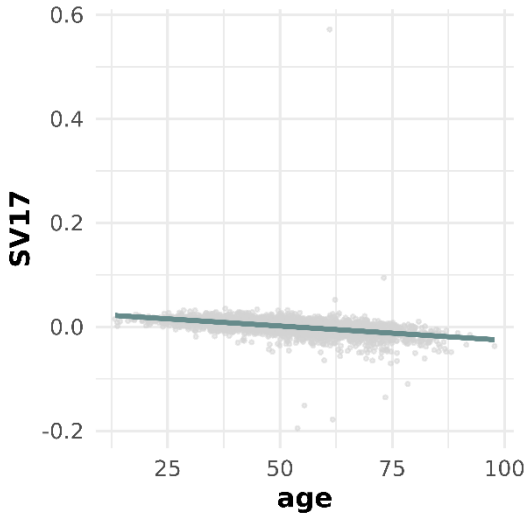


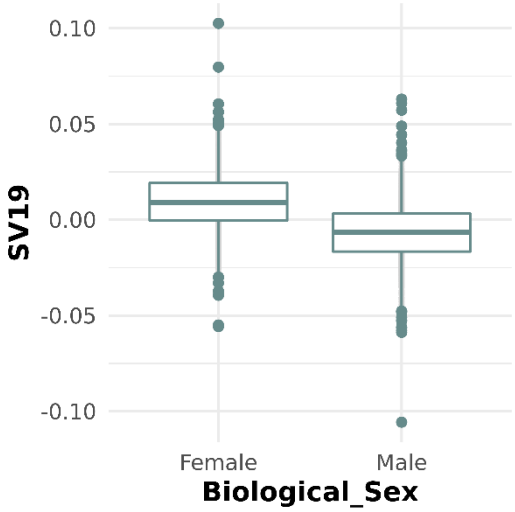
Figure 7 | Correlation plot between slide, collection date (Year and Year_Month), cell types (Bcell, CD4T, CD8T, Eos, Mono, Neu, and NK) with the first 20 surrogate variables (SVs).

Table 5 | Significant pairwise associations (r^2<0.1) between variables (sex, ethnicity, slide, age, cell type, collection date, drugs, alcohol, and neurological, cardiovascular, respiratory, and endocrine diseases) and surrogate variables (SVs).

x	l	y	test	p.value	estimate	r2
Neu		SV2	F-test	0	2808.473	0.450764031
Neu		SV1	F-test	0	1971.284	0.365507137
NK		SV1	F-test	0	1751.664	0.338573201
CD8T		SV2	F-test	2.8E-269	1479.149	0.30179633
Bcell		SV1	F-test	5.4E-212	1115.784	0.245887396
CD4T		SV2	F-test	2.3E-211	1111.932	0.245246688
CD4T		SV5	F-test	8.6E-195	1012.049	0.228244849
Bcell		SV9	F-test	7E-186	959.2464	0.218943723
Biological_Sex	Female	SV19	t-test	6E-176		0.209365268
Biological_Sex		SV19	F-test	5.6E-171	872.3528	0.203139519
Biological_Sex	Male	SV19	t-test	9.7E-170		0.202708486
age		SV17	F-test	2.7E-145	726.6454	0.175152449
Bcell		SV8	F-test	1.5E-143	716.8678	0.173203836
Slide		SV15	F-test	2E-140	699.5484	0.169729512
Slide		SV1	F-test	5.3E-140	697.2175	0.169259686
Year_Month		SV15	F-test	1.2E-134	669.6864	0.168233907
NK		SV2	F-test	1E-127	629.8141	0.155440029
Year_Month		SV1	F-test	1.6E-119	587.0054	0.150591228
Year		SV12	F-test	6.4E-114	556.7952	0.14395674
Slide		SV12	F-test	3.6E-117	572.7905	0.143384361
Year		SV1	F-test	4.1E-100	483.3929	0.12739663
CD4T		SV1	F-test	4.9E-101	487.1228	0.124611782
age		SV16	F-test	5.6E-88	419.1817	0.109128321
Bcell		SV2	F-test	1.57E-81	386.0765	0.10138361







| Summary GRsets

	Description GenomicRatioSet
GRset_clean	Normalized beta with missing values after QC sample
GRset_clean_imp	Normalized beta with imputed values using median method
GRset_ SVA	Residuals after adjusting the normalized and imputed betas by the 60 SVs