Supplemental Material to:

Pan H, Chen L, Dogra S, The AL, Tan JH, Lim YI, et al. Measuring the methylome in clinical samples: Improved processing of the Infinium Human Methylation450 BeadChip Array. Epigenetics 2012; 7(10); http://dx.doi.org/10.4161/epi.22102
http://www.landesbioscience.com/journals/epigenetics/article/22102/

Supplementary table 1: Demographic characteristics of sample sets
Supplementary table 2: Schematic of experimental design

Supplementary table 3: Number of CpG % methylation values showing agreement within 5%, 10% and 20% ranges, between Infinium 450K and RRBS (Nreads ≥4) data, when Infinium data is subjected to individual steps processing, combined processing, GenomeStudio (GS) processing, GS and type II adjustment or SWAN. The greatest number of CpGs agreeing at every level, between processing methods are highlighted in yellow. Overall correlation statistics are also shown (n=454,660)

Supplementary table 4: Number of CpG % methylation values showing agreement within 5%, 10% and 20% ranges, between Infinium 450K and RRBS (Nreads ≥10) data, when Infinium data is subjected to individual steps processing, combined processing, GenomeStudio (GS) processing, GS and type II adjustment or SWAN. The greatest number of CpGs agreeing at every level, between processing methods are highlighted in yellow. Overall correlation statistics are also shown (n=387,789)

Supplementary table 5: Number of CpG % methylation values showing agreement or not (within 10%), between SWAN processed Infinium 450K and RRBS data (Nreads ≥4), at different % methylation value ranges.

Supplementary table 6: Number of probes associated with chip-batch and gender and gestational age phenotypes at ANOVA pvalue <0.01, in raw data, processed data and SWAN processed data, n=426,831

Supplementary table 7: Number of probes associated with chip-batch and gender and gestational age phenotypes at ANOVA FDR corrected pvalue ≤0.2 in raw data, processed data SWAN processed data, n=426,831

Supplemental Methods

```
############################################################################################
#The R script for signal correction from red and green color channels in Infinium 450k
############################################################################################
#The signal file from GenomeStudio should contain AVG_Beta, Intensity and etc as the below items
#for each sample.
######################
#TargetID
```

```
#6164655096_R05C02.AVG_Beta
#6164655096_R05C02.Intensity
#6164655096_R05C02.Avg_NBEADS_A
#6164655096_R05C02.Avg_NBEADS_B
#6164655096_R05C02.Signal_A
#6164655096_R05C02.Signal_B
#6164655096_R05C02.Detection Pval
######################

######################################################################################
#ControlProfile (one sample) from GenomeStudio
######################################################################################
#Index	TargetID	6164655096_R05C02.Signal_Grn	6164655096_R05C02.Signal_Red
#1	BISULFITE CONVERSION I	5769.917	9938
#2	BISULFITE CONVERSION II	2299	37534
#3	EXTENSION	23293.5	33670.75
#4	HYBRIDIZATION	26142.67	1692.333
#5	NEGATIVE	161.2217	282.0367
#6	NON-POLYMORPHIC	9678.25	16405.5
#7	NORM_A	1005.938	11977.75
#8	NORM_C	8361.771	652.1312
#9	NORM_G	9988.906	944.7188
#10	NORM_T	464.8524	10485.23
#11	SPECIFICITY I	3786.667	7669.75
#12	SPECIFICITY II	580.3333	29149.33
#13	STAINING	9124.75	14790.5
#14	TARGET REMOVAL	519.5	1783.5
######################################################################################
######################################Script Start
#####################Input and output file name
signal_file <- "U7noncontrolnorm_complete.txt"            #the signal files from GenomeStudio
color_ann_file <- "Illumina_450K_annotation_light.txt"   #Illumina450K annotation file
profile_file <- "U7ControlProfile.txt"                                                   #the controlprofile
from GeomeStudio
output_file <- "beta_color.txt"                                                                   #the result
file: color adjusted bata value

#####################Read in signal
signal <- read.table(signal_file, sep="\t", head=TRUE)
ann <- read.table(color_ann_file, sep="\t", quote = "", head=TRUE)
signal <- merge(signal, ann[,c(1:3)], by.x="TargetID", by.y="TargetID")

#######################InternalQC, removal of Detection pvalue>0.05 and
#######################the number of BEADS <3.
pvalue_col <- c(grep("Detection.Pval", colnames(signal)))
x <- apply(signal[,pvalue_col] <= 0.05, 1, sum)
idx1 <- which(x == length(pvalue_col))

NBEADSA <- c(grep("NBEADS_A", colnames(signal)))
```

```
y <- apply(signal[,NBEADSA] >= 3, 1, sum)
idx2 <- which(y == length(NBEADSA))

NBEADSB <- c(grep("NBEADS_B", colnames(signal)))
z <- apply(signal[,NBEADSB] >= 3, 1, sum)
idx3 <- which(z == length(NBEADSB))

all_good <- intersect(intersect(idx1, idx2), idx3)

##########################control profile
profile_data <- read.table(profile_file, sep="\t", head=TRUE)
grn_col <- c(grep("Signal_Grn", colnames(profile_data)))
red_col <- c(grep("Signal_Red", colnames(profile_data)))

control_grn <- as.matrix((profile_data[8, grn_col]+ profile_data[9, grn_col])/2)
# average (NORM_C+NORM_G)

control_red <- as.matrix((profile_data[7, red_col]+ profile_data[10, red_col])/2)
 #average(NORM_A+NORM_T)

negative_grn <-as.matrix(profile_data[5, grn_col])
negative_red <-as.matrix(profile_data[5, red_col])

###########################signal_A: red(unmethylated) and signal_B: green(methylated)
signalA_col<-grep("Signal_A", colnames(signal))
signalB_col<-grep("Signal_B", colnames(signal))

####typeII
idx1 <- intersect(which(signal$INFINIUM_DESIGN_TYPE == "II"), all_good)
adjSA1 <- as.character(signal[idx1,1])
adjSB1 <- as.character(signal[idx1,1])
n <- length(signalA_col)

for (i in 1:n) {
	adjSA <- signal[idx1,signalA_col[i]] - negative_red[i]
	adjSB <- signal[idx1,signalB_col[i]] * (control_red[i]/control_grn[i]) -
negative_grn[i]*(control_red[i]/control_grn[i])
	adjSA[adjSA <= 0] <- 0.1
	adjSB[adjSB <= 0] <- 0.1

	adjSA1 <- data.frame(adjSA1, adjSA)
	adjSB1 <- data.frame(adjSB1, adjSB)
}

#typeI_red
rm(idx1, adjSA, adjSB)
idx1 <- intersect(which(signal$INFINIUM_DESIGN_TYPE == "I"), which(signal$COLOR_CHANNEL=="Red"))
idx1 <- intersect(idx1, all_good)
adjSA2 <- as.character(signal[idx1,1])
```

```
adjSB2 <- as.character(signal[idx1,1])

for (i in 1:n) {
        adjSA <- signal[idx1,signalA_col[i]] - negative_red[i]
        adjSB <- signal[idx1,signalB_col[i]] - negative_red[i]
        adjSA[adjSA <= 0] <- 0.1
        adjSB[adjSB <= 0] <- 0.1

        adjSA2 <- data.frame(adjSA2, adjSA)
        adjSB2 <- data.frame(adjSB2, adjSB)
}

#idx_typeI_grn
rm(idx1, adjSA, adjSB)
idx1 <- intersect(which(signal$INFINIUM_DESIGN_TYPE == "I"), which(signal$COLOR_CHANNEL=="Grn"))
idx1 <- intersect(idx1, all_good)
adjSA3 <- as.character(signal[idx1,1])
adjSB3 <- as.character(signal[idx1,1])

for (i in 1:n) {
        adjSA <- signal[idx1,signalA_col[i]] * (control_red[i]/control_grn[i]) -
negative_grn[i]*(control_red[i]/control_grn[i])
        adjSB <- signal[idx1,signalB_col[i]] * (control_red[i]/control_grn[i]) -
negative_grn[i]*(control_red[i]/control_grn[i])
        adjSA[adjSA <= 0] <- 0.1
        adjSB[adjSB <= 0] <- 0.1

        adjSA3 <- data.frame(adjSA3, adjSA)
        adjSB3 <- data.frame(adjSB3, adjSB)
}

colnames(adjSA1)[1] <- "ProbeID"
colnames(adjSA2)[1] <- "ProbeID"
colnames(adjSA3)[1] <- "ProbeID"
adjSA<- rbind(adjSA1,adjSA2,adjSA3)

colnames(adjSB1)[1] <- "ProbeID"
colnames(adjSB2)[1] <- "ProbeID"
colnames(adjSB3)[1] <- "ProbeID"
adjSB<- rbind(adjSB1,adjSB2,adjSB3)

adjSB_order <- adjSB[order(adjSB[,1]),]
adjSA_order <- adjSA[order(adjSA[,1]),]

beta <- as.character(adjSB_order[,1])

for (i in 2:(n+1)) {
        tmp_beta <- adjSB_order[i]/(adjSB_order[i] + adjSA_order[i] + 100)
        beta <- data.frame(beta, tmp_beta)
}

colnames(beta) <- c("TargetID", gsub(".Signal_A", ".AVG_Beta", colnames(signal)[signalA_col]))
colnames(beta) <- gsub("X", "", colnames(beta))
write.table(beta, output_file, row.names = F, quote = F, sep = "\t")

#end of script.
```

```
##############################################################
#R script for typeII correction.
##############################################################
datafile <- "beta_color.txt"                                            #beta file
typefile <- "Illumina_450K_annotation_light.txt"#Infinium450k annotation file
outputfile1 <- "beta_color_typeII.txt"                         #output file.

#read data file
data1 <- read.table(datafile, sep="\t", header=TRUE)
type1 <- read.table(typefile, sep="\t", quote=””, header=TRUE)
name_x <- as.character(colnames(data1)[1])
name_y <- as.character(colnames(type1)[1])
data1 <- merge(data1, type1[,c(1:2)], by.x=name_x, by.y=name_y)
head(data1); dim(data1)

naidx <- apply(is.na(data1), 1, sum)
data1 <- data1[which(naidx == 0),]

row_size <- dim(data1)[1]
col_size <- dim(data1)[2] - 1

idx1 <- which(data1$INFINIUM_DESIGN_TYPE=='I' )
idx2 <- which(data1$INFINIUM_DESIGN_TYPE=='II')

beta1_typeI_probe <- as.character(data1[idx1,1])
beta1_typeII_probe <- as.character(data1[idx2,1])
len <- length(colnames(data1))

beta1_corrected <- c(as.character(beta1_typeI_probe), as.character(beta1_typeII_probe))
for (i in 2:col_size){
beta1 <- data1[,i] #one sample
cat(colnames(data1)[i], "; ")
beta_typeI <- beta1[idx1]
beta_typeII <- beta1[idx2]

m1 <- log2(beta_typeI/(1 - beta_typeI))
m2 <- log2(beta_typeII/(1 - beta_typeII))
probe_typeII <- length(m2)

#plot the density of typeI and typeII beta value.
#dev.new(1)
#par(mfrow=c(2,2))
#plot(density(beta_typeI, bw=0.05), main='Fig1: Beta value for typeI and typeII(red)')
#lines(density(beta_typeII, bw=0.05), col='red')
#plot(density(m2, bw=0.5, kernel="gaussian", n=200, na.rm=TRUE), col='red', main='Fig2: M value for typeI
and typeII(red)')
#lines(density(m1, bw=0.5, kernel="gaussian", n=200, na.rm=TRUE))
#typeI M-value as the base.
dm1 <- density(m1, bw=0.5, kernel="gaussian", n=200, na.rm=TRUE)
sigma_m1 <- dm1$x[which.max(dm1$y[dm1$x >= 0])+ length(dm1$x[dm1$x < 0])]
sigma_u1 <- dm1$x[which.max(dm1$y[dm1$x < 0])]
cat(sigma_m1, "; ", sigma_u1, "\n")

#######adjust typeII M value according to typeI
dm2 <- density(m2, bw=0.5, kernel="gaussian", n=200, na.rm=TRUE)
sigma_m2 <- dm2$x[which.max(dm2$y[dm2$x >= 0])+ length(dm2$x[dm2$x < 0])]
sigma_u2 <- dm2$x[which.max(dm2$y[dm2$x < 0])]
cat(sigma_m2, "; ", sigma_u2, "\n")
mm2 <- rep(0, probe_typeII)
idx11 <- which(m2 >= 0)
idx22 <- which(m2 < 0)
mm2[idx11]<- (m2[idx11]/sigma_m2)*sigma_m1
mm2[idx22]<- (m2[idx22]/sigma_u2)*sigma_u1
beta_typeII_corrected <- 2^mm2/(2^mm2+1)

#plot(density(mm2,bw=0.5, kernel="gaussian", n=200, na.rm=TRUE), col='red', main="Fig3. M value in typeI
and adjust typeII(red)")
#lines(density(m1, bw=0.5, kernel="gaussian", n=200, na.rm=TRUE))
#plot(density(beta_typeI, bw=0.05), main='Fig4: Beta value in typeI and corrected typeII(red)')
#lines(density(beta_typeII_corrected, bw=0.05), col='red')

temp <- c(beta_typeI, beta_typeII_corrected)
beta1_corrected <- data.frame(beta1_corrected, temp)
}
colnames(beta1_corrected) <- as.character(colnames(data1)[1:col_size])
colnames(beta1_corrected) <- gsub("X", "", colnames(beta1_corrected))
write.table(beta1_corrected, outputfile1, sep='\t', quote = F, row.names=FALSE)

#end of script.
```

Supplementary table 1: Demographic characteristics of sample sets

SampleSet	Expanded 72	7 compared to RRBS
Gender	32 female, 40 male	7 male
Median BW (g)	3038.50	3352.00
Mean BW (+/- sd) (g)	3070.88(+/- 576.40)	3325.72(+/- 329.33)
Median GA (w)	39.00	39.43
Mean GA (+/- sd) (w)	38.46(+/-1.71)	38.88 (+/- 1.20)
Ethnic Group	Chinese 65, Indian 2, Malay 5	Chinese 7

Supplementary table 2

Sample	1		2		3		4		5		6		7	
Technology	RRBS	Infinium 450K	RRBS	Infinium 450K	RRBS	Infinium 450K	RRBS	Infinium 450K	RRBS	Infinium 450K	RRBS	Infinium 450K	RRBS	Infinium 450K
Cord	■	■	□	□	■	■	□	□	■	■	□	□	■	■
DNA extraction	■	□	■	■	□	□	■	■	□	□	■	■	□	□
bisulphite conversion	■	□	■	□	■	□	■	□	■	□	■	□	■	□
Infinium 450K experiment		1		2		2		2		3		3		3
Infinium 450K array		1A		2A		2B		2A		3A		3A		3B

Supplementary table 3

Number of CpG % methylation values showing agreement within 5%, 10% and 20% ranges, between Infinium 450K and RRBS (Nreads ≥4) data, when Infinium data is subjected to individual steps processing, combined processing, GenomeStudio (GS) processing, GS and type II adjustment or SWAN. The greatest number of CpGs agreeing at every level, between processing methods are highlighted in yellow. Overall correlation statistics are also shown (n=454,660)

Difference Range (n=454,660)	within 20%		within 10%		within 5%		Spearman's Rank R	Pearson's R^2	Slope
Probe Type	Type I	Type II	Type I	Type II	Type I	Type II			
Raw Data	196,937 93%	217,721 90%	172,003 81%	161,520 67%	132,267 62%	75,144 31%	0.83 p<0.001	0.92 p<0.001	0.83
Raw + colour adjustment	197,089 93%	221,887 92%	173,872 82%	165,158 68%	139,888 66%	73,510 30%	0.83 p<0.001	0.92 p<0.001	0.87
Raw + Type II adjustment	196,937 93%	225,922 93%	172,003 81%	193,550 80%	132,267 62%	144,968 60%	0.83 p<0.001	0.93 p<0.001	0.9
Raw + QN	196,514 92%	221,171 91%	172,154 81%	165,537 68%	132,776 62%	77,541 32%	0.83 p<0.001	0.92 p<0.001	0.85
Combined processing	197,444 93%	226,386 94%	175,211 82%	195,114 81%	143,022 67%	149,216 62%	0.83 p<0.001	0.93 p<0.001	0.93
Raw + GS	197,000 93%	221,523 92%	173,078 81%	161,101 67%	136,568 64%	65,024 27%	0.83 p<0.001	0.92 p<0.001	0.86
GS + Type II	197,000 93%	225,810 93%	173,078 81%	191,340 79%	136,568 64%	141,469 58%	0.83 p<0.001	0.94 p<0.001	0.92
Raw + SWAN	198,347 93%	224,342 93%	172,472 81%	177,745 73%	127,145 60%	92,454 38%	0.83 p<0.001	0.93 p<0.001	0.87

Supplementary table 4

Number of CpG % methylation values showing agreement within 5%, 10% and 20% ranges, between Infinium 450K and RRBS (Nreads ≥10) data, when Infinium data is subjected to individual steps processing, combined processing, GenomeStudio (GS) processing, GS and type II adjustment or SWAN. The greatest number of CpGs agreeing at every level, between processing methods are highlighted in yellow. Overall correlation statistics are also shown (n=387,789)

Difference Range (n=387,789)	within 20%		within 10%		within 5%		Spearman's Rank R	Pearson's R^2	Slope
Probe Type	Type I	Type II	Type I	Type II	Type I	Type II			
Raw Data	166,565	192,635	147,047	145,040	114,131	69,159	0.83	0.93	0.84
	94%	92%	83%	69%	64%	33%	p<0.001	p<0.001	
Raw + colour adjustment	166,723	195,934	148,589	148,186	119,916	67,879	0.83	0.93	0.88
	94%	93%	84%	71%	67%	32%	p<0.001	p<0.001	
Raw + Type II adjustment	166,565	199,467	147,047	173,432	114,131	131,670	0.83	0.94	0.91
	94%	95%	83%	83%	64%	63%	p<0.001	p<0.001	
Raw + QN	166,113	195,750	147,106	148,832	114,507	71,431	0.83	0.93	0.86
	93%	93%	83%	71%	64%	34%	p<0.001	p<0.001	
Combined processing	167,026	199,786	149,690	174,764	122,500	135,190	0.83	0.94	0.95
	94%	95%	84%	83%	69%	64%	p<0.001	p<0.001	
Raw + GS	166,629	195,614	147,931	144,716	117,450	60,266	0.83	0.93	0.87
	94%	93%	83%	69%	66%	29%	p<0.001	p<0.001	
GS + Type II	166,629	199,306	147,931	171,494	117,450	128,378	0.83	0.94	0.92
	94%	95%	83%	82%	66%	61%	p<0.001	p<0.001	
Raw + SWAN	167,878	197,973	147,670	159,599	110,506	85,521	0.83	94	0.88
	94%	94%	83%	76%	62%	41%	p<0.001	p<0.001	

Supplementary table 5

Number of CpG % methylation values showing agreement or not (within 10%), between SWAN processed Infinium 450K and RRBS data (Nreads ≥4), at different % methylation value ranges.

Infinium β Value Range	Difference < 10%	Difference >10%	Total	Median Absolute Difference	Standard Deviation of Absolute Difference
x ≤ 20%	242,195 91%	23,823 9%	266,018	4.11	3.89
20% < x ≤ 80%	35,284 38%	58,032 62%	93,316	13.31	11.76
80% < x ≤ 100%	72,738 76%	22,588 24%	95,326	5.99	8.28

Supplementary table 6

Number of probes associated with chip-batch and gender and gestational age phenotypes at ANOVA pvalue <0.01, in raw data, processed data SWAN processed data, n=426,831

	ChipBatch	Gender	Gestational Age
Raw	163,438 (38.29%)	29,642 (6.94%)	3,842 (0.90%)
Combined processing	58,532(13.71%)	48,436 (11.35%)	8,454 (1.98%)
SWAN	83,139 (19.48%)	26,061 (6.11%)	7,766 (1.82%)

Supplementary table 7

Number of probes associated with chip-batch and gender and gestational age phenotypes at ANOVA FDR corrected pvalue ≤0.2 in raw data, processed data SWAN processed data, n=426,831

	ChipBatch	Gender	Gestational Age
Raw	233,650 (54.74%)	41,108 (9.63%)	104 (0.02%)
Combined processing	123,611 (28.96%)	73,595 (17.24%)	179 (0.04%)
SWAN	182,161(42.68%)	30, 684(7.19%)	41(<0.01%)