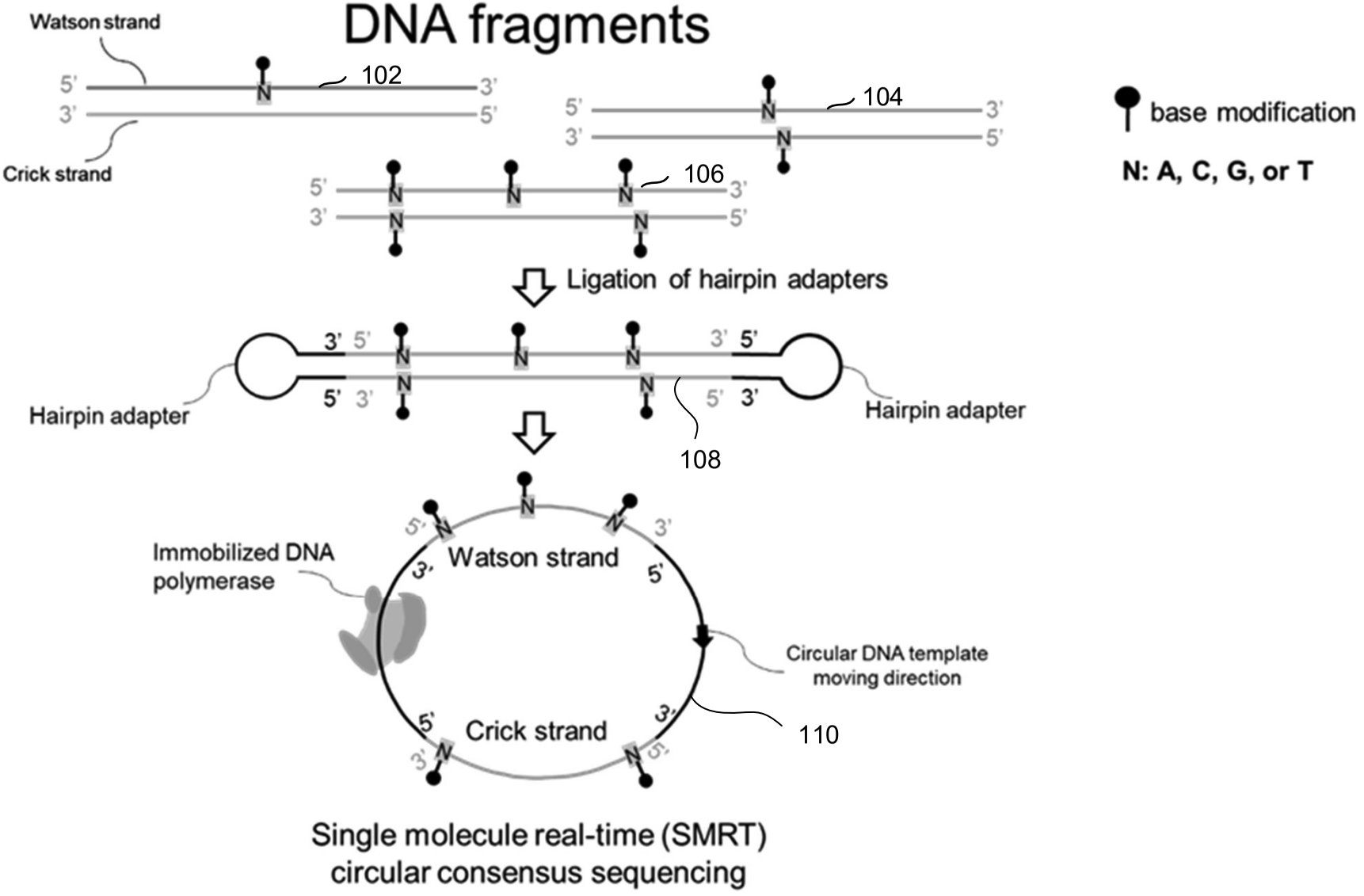

DNA fragments
Watson strand
Crick strand
102
104
106
base modification
N: A, C, G, or T
Ligation of hairpin adapters
Hairpin adapter
Hairpin adapter
108
Immobilized DNA polymerase
Watson strand
Crick strand
Circular DNA template moving direction
110
Single molecule real-time (SMRT) circular consensus sequencing

FIG. 1

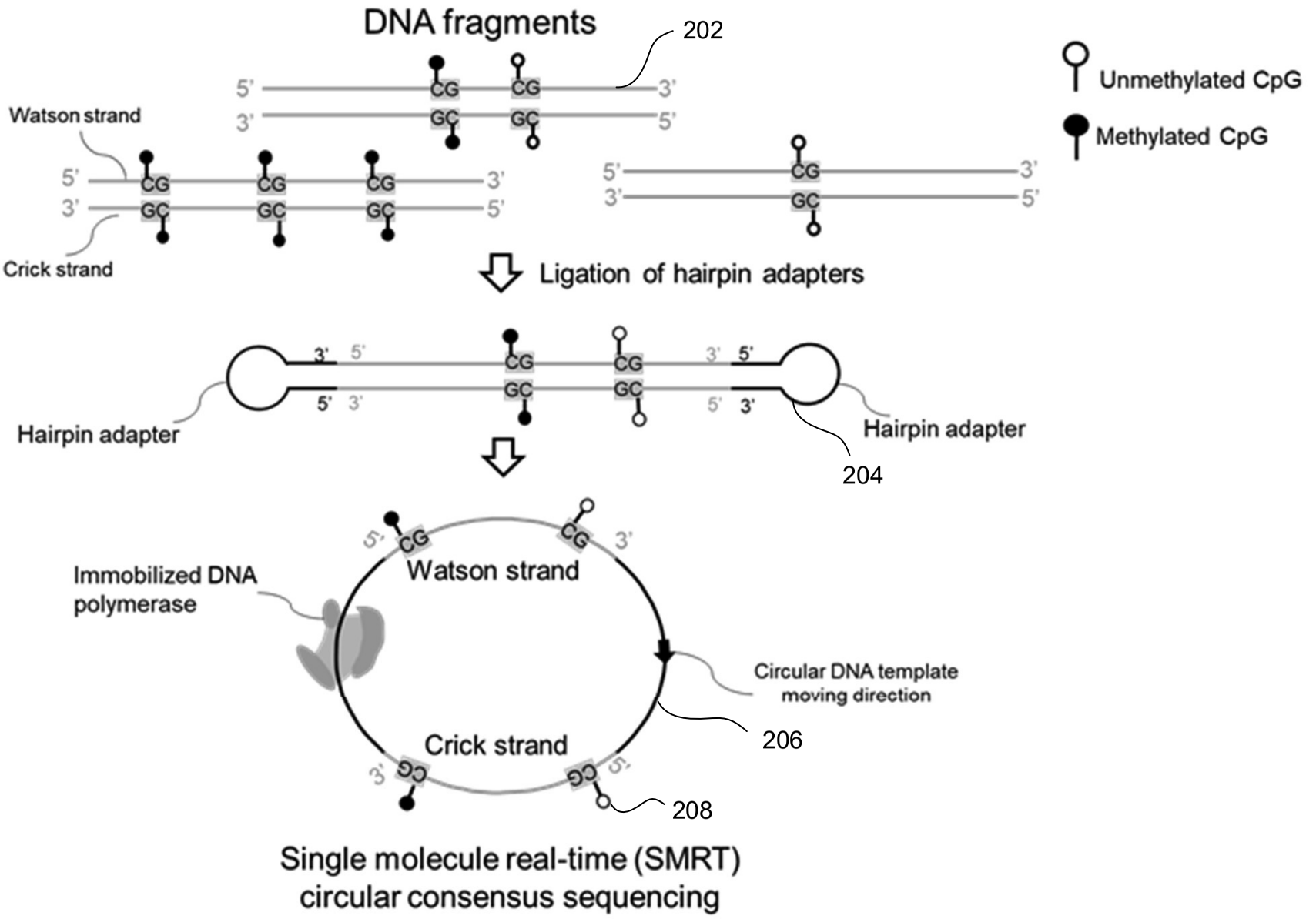

DNA fragments
202
Unmethylated CpG
Methylated CpG
Watson strand
Crick strand
Ligation of hairpin adapters
Hairpin adapter
Hairpin adapter
204
Immobilized DNA polymerase
Watson strand
Circular DNA template moving direction
206
Crick strand
208
Single molecule real-time (SMRT) circular consensus sequencing

FIG. 2

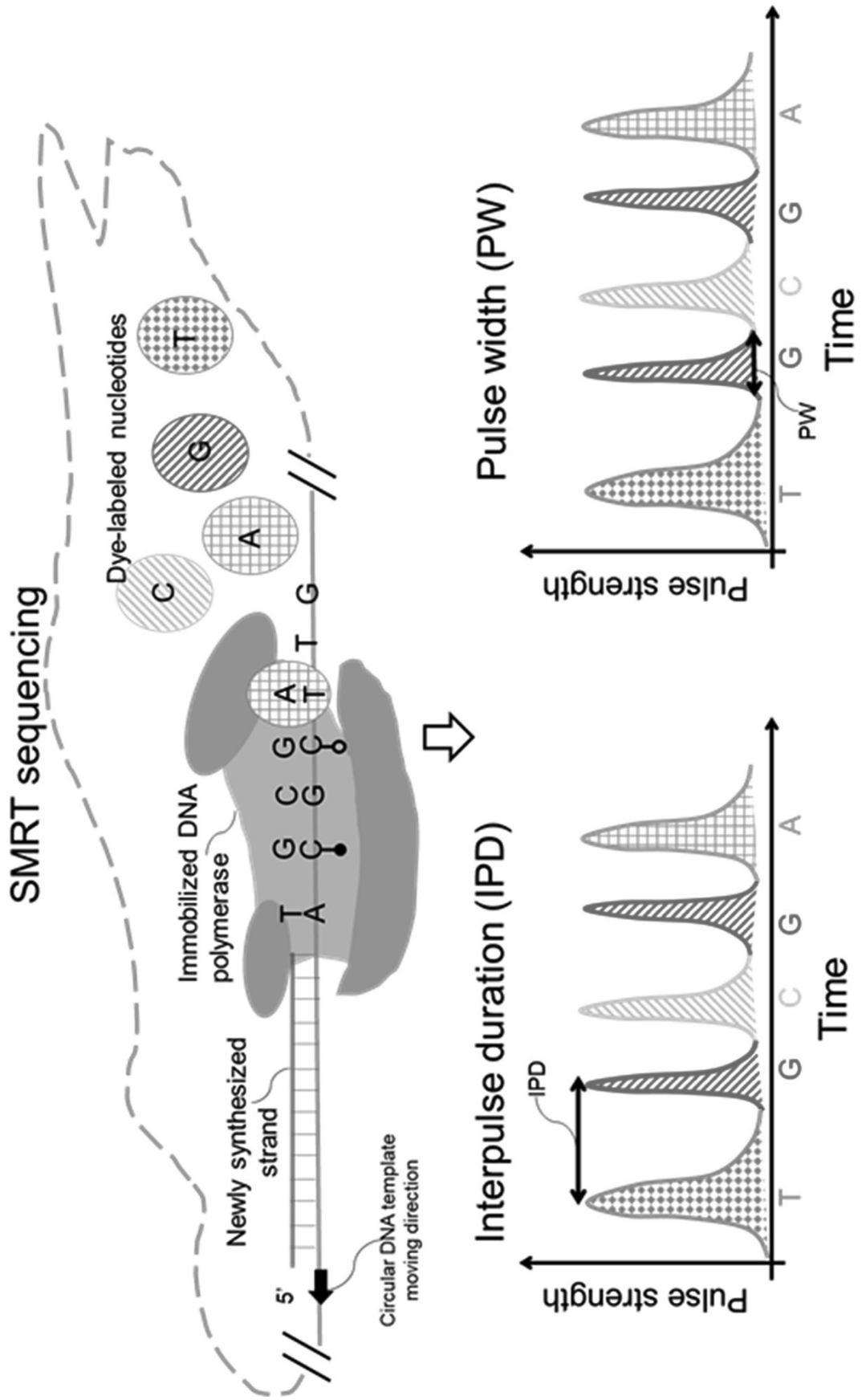
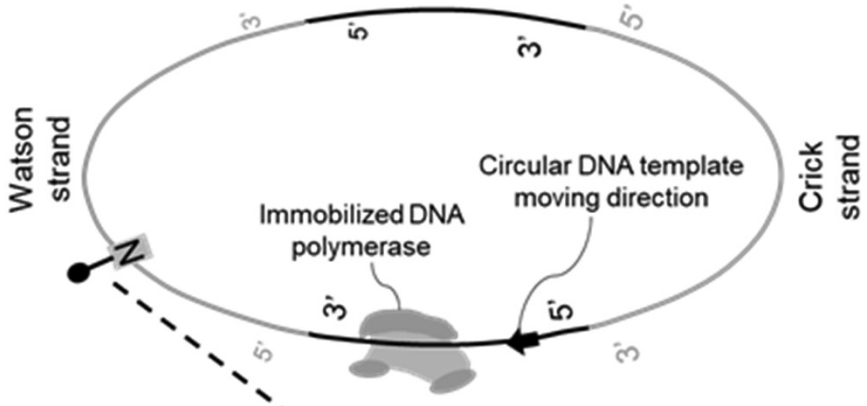

SMRT sequencing
Dye-labeled nucleotides
C
A
G
T
Immobilized DNA polymerase
Newly synthesized strand
5'
Circular DNA template moving direction
Pulse width (PW)
Pulse strength
Time
PW
Interpulse duration (IPD)
IPD

FIG. 3



N: represents any base.

base modification



Measurement window 400

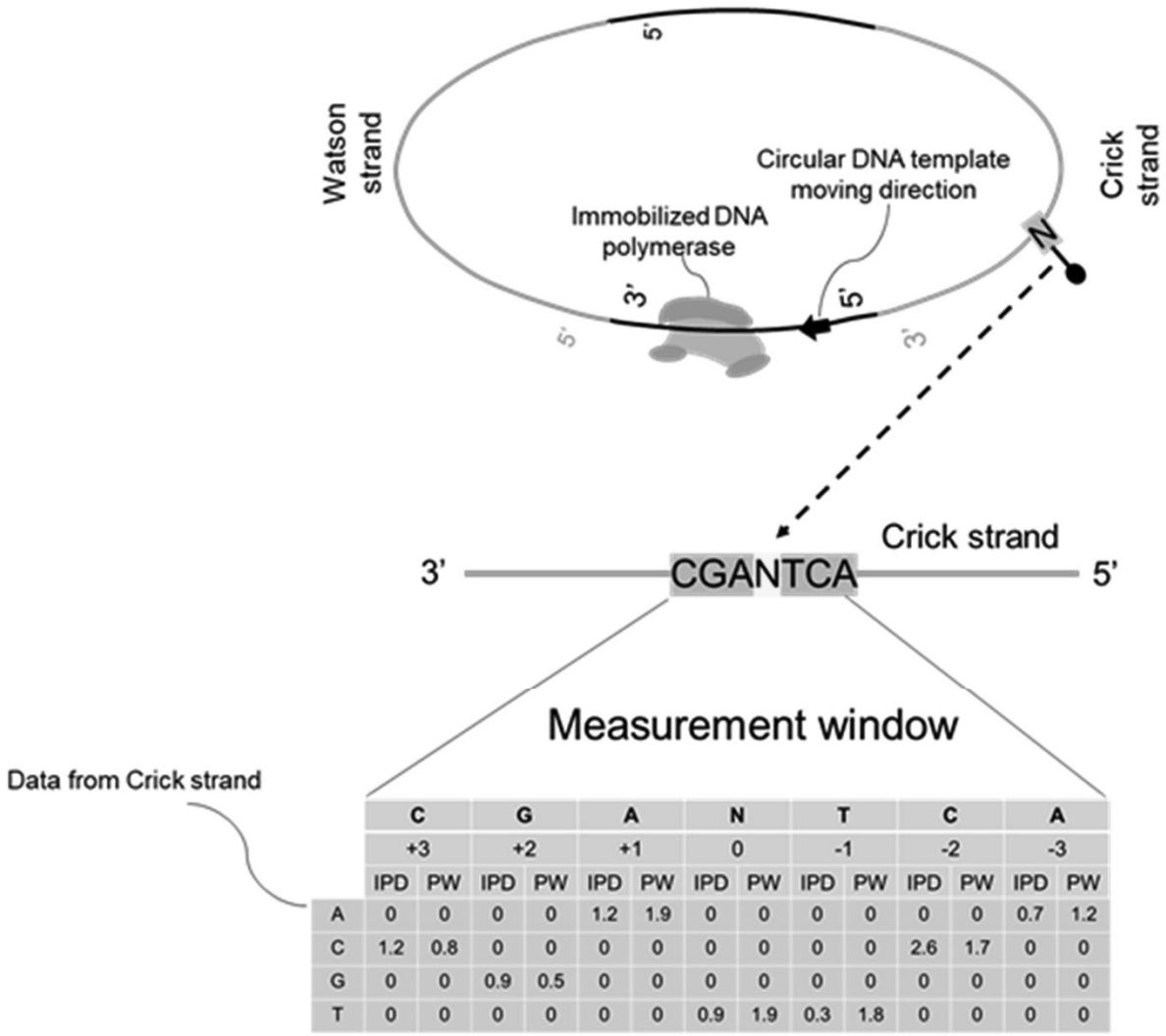
Data from Watson strand

	G		A		T		N		A		C		T	
	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	1.0	0.3	0	0	0	0	1.1	0.8	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	2.3	0.8	0	0
G	1.5	1.7	0	0	0	0	0.9	1.2	0	0	0	0	0	0
T	0	0	0	0	1.2	0.2	0	0	0	0	0	0	1.5	1.7

402
404
408
412
416
420

Measurement window: e.g. 3 bases from each side of base that is interrogated (position of '0')

FIG. 4



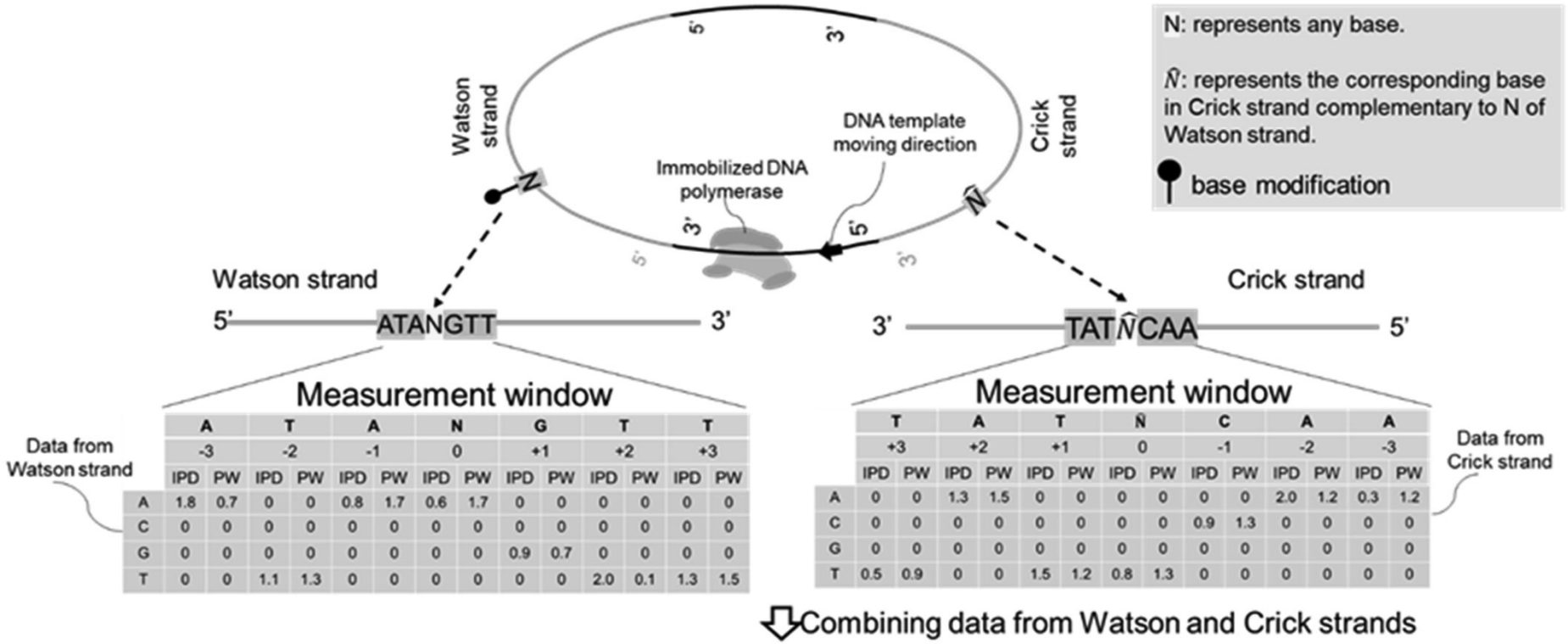
	C		G		A		N		T		C		A	
	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0.7	1.2
C	1.2	0.8	0	0	0	0	0	0	0	0	2.6	1.7	0	0
G	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0.9	1.9	0.3	1.8	0	0	0	0

N: represents any base.

base modification

Measurement window: e.g. 3 bases from each side of base that is interrogated (position of '0')

FIG. 5



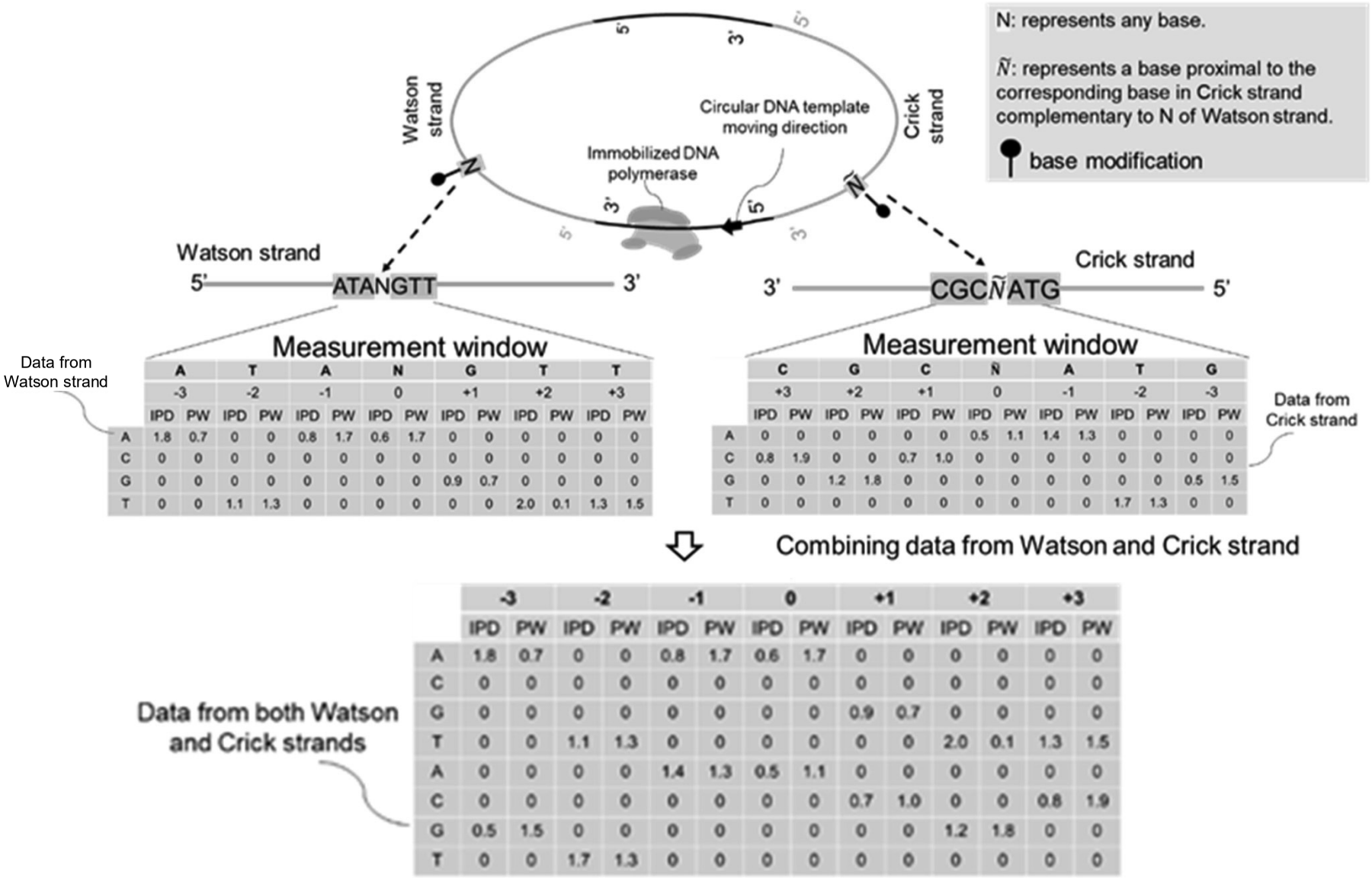
	A		T		A		N		G		T		T	
	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5

	T		A		T		$\hat{N}$		C		A		A	
	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	1.3	1.5	0	0	0	0	0	0	2.0	1.2	0.3	1.2
C	0	0	0	0	0	0	0	0	0.9	1.3	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0.5	0.9	0	0	1.5	1.2	0.8	1.3	0	0	0	0	0	0

Data from both Watson and Crick strands

	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5
A	0.3	1.2	2.0	1.2	0	0	0	0	0	0	1.3	1.5	0	0
C	0	0	0	0	0.9	1.3	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0.8	1.3	1.5	1.2	0	0	0.5	0.9

FIG. 6

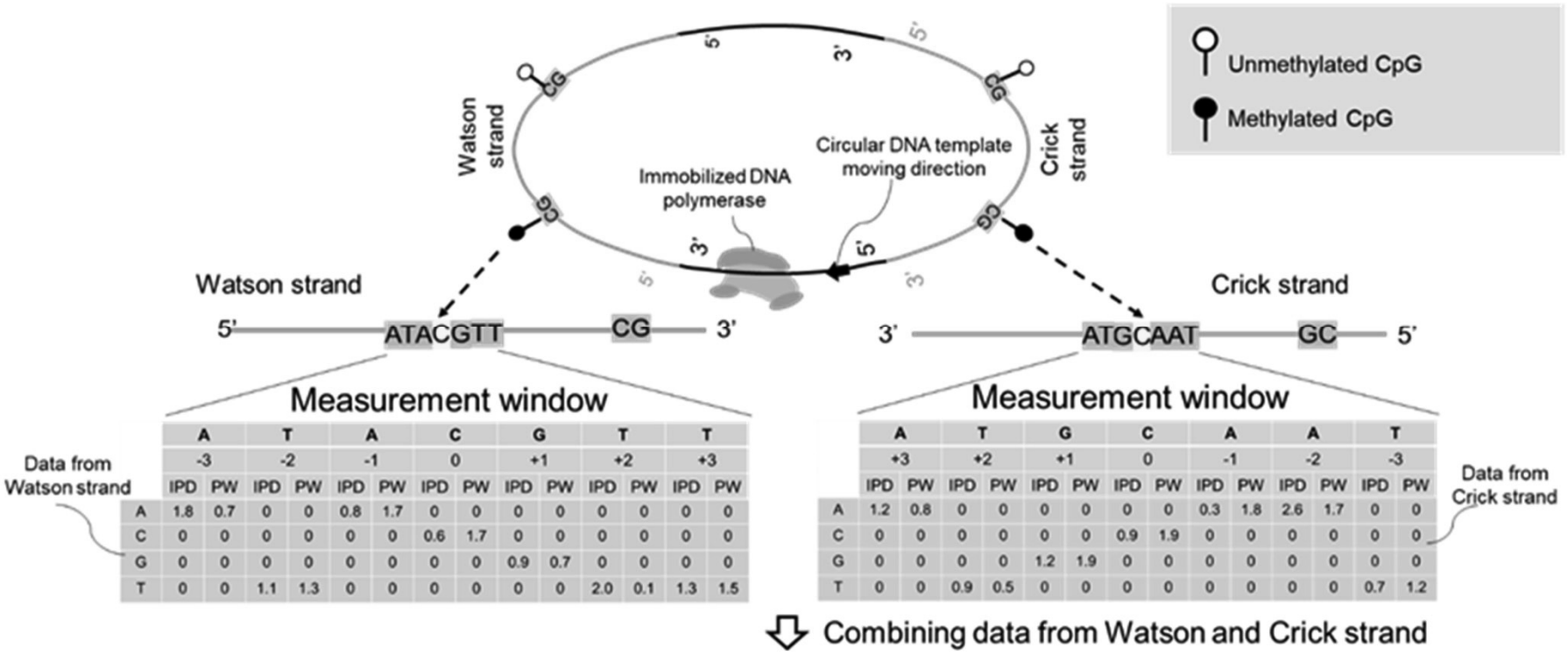


	A		T		A		N		G		T		T	
	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5

	C		G		C		Ñ		A		T		G	
	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	0	0	0	0	0.5	1.1	1.4	1.3	0	0	0	0
C	0.8	1.9	0	0	0.7	1.0	0	0	0	0	0	0	0	0
G	0	0	1.2	1.8	0	0	0	0	0	0	0	0	0.5	1.5
T	0	0	0	0	0	0	0	0	0	0	1.7	1.3	0	0

	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5
A	0	0	0	0	1.4	1.3	0.5	1.1	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0.7	1.0	0	0	0.8	1.9
G	0.5	1.5	0	0	0	0	0	0	0	0	1.2	1.8	0	0
T	0	0	1.7	1.3	0	0	0	0	0	0	0	0	0	0

FIG. 7



	A		T		A		C		G		T		T	
	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0.6	1.7	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5

	A		T		G		C		A		A		T	
	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.2	0.8	0	0	0	0	0	0	0.3	1.8	2.6	1.7	0	0
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0	0
T	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0.7	1.2

Data from both Watson and Crick strands

	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0.6	1.7	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5
A	0	0	2.6	1.7	0.3	1.8	0	0	0	0	0	0	1.2	0.8
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	1.2	1.9	0	0	0	0
T	0.7	1.2	0	0	0	0	0	0	0	0	0.9	0.5	0	0

FIG. 8

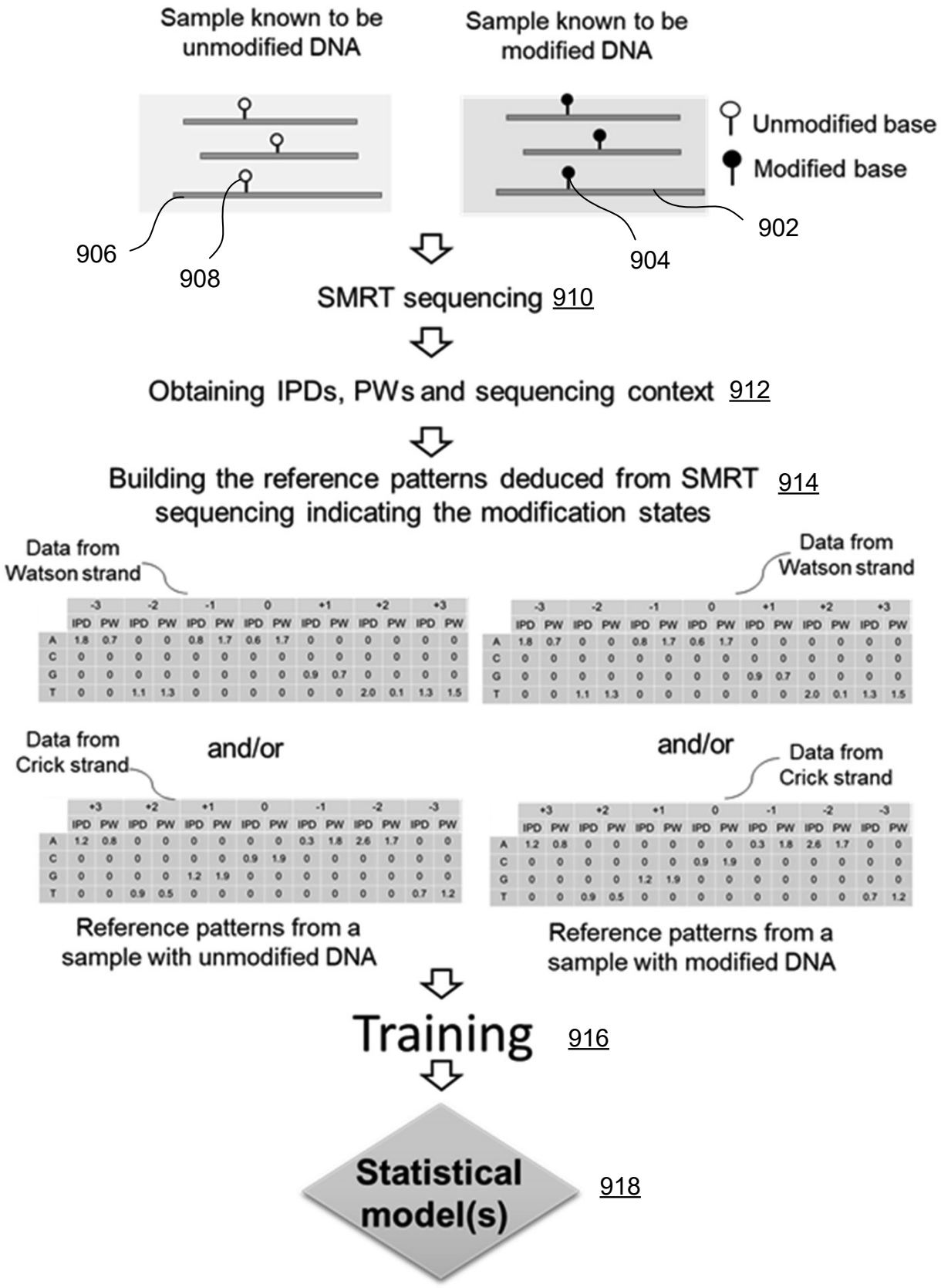
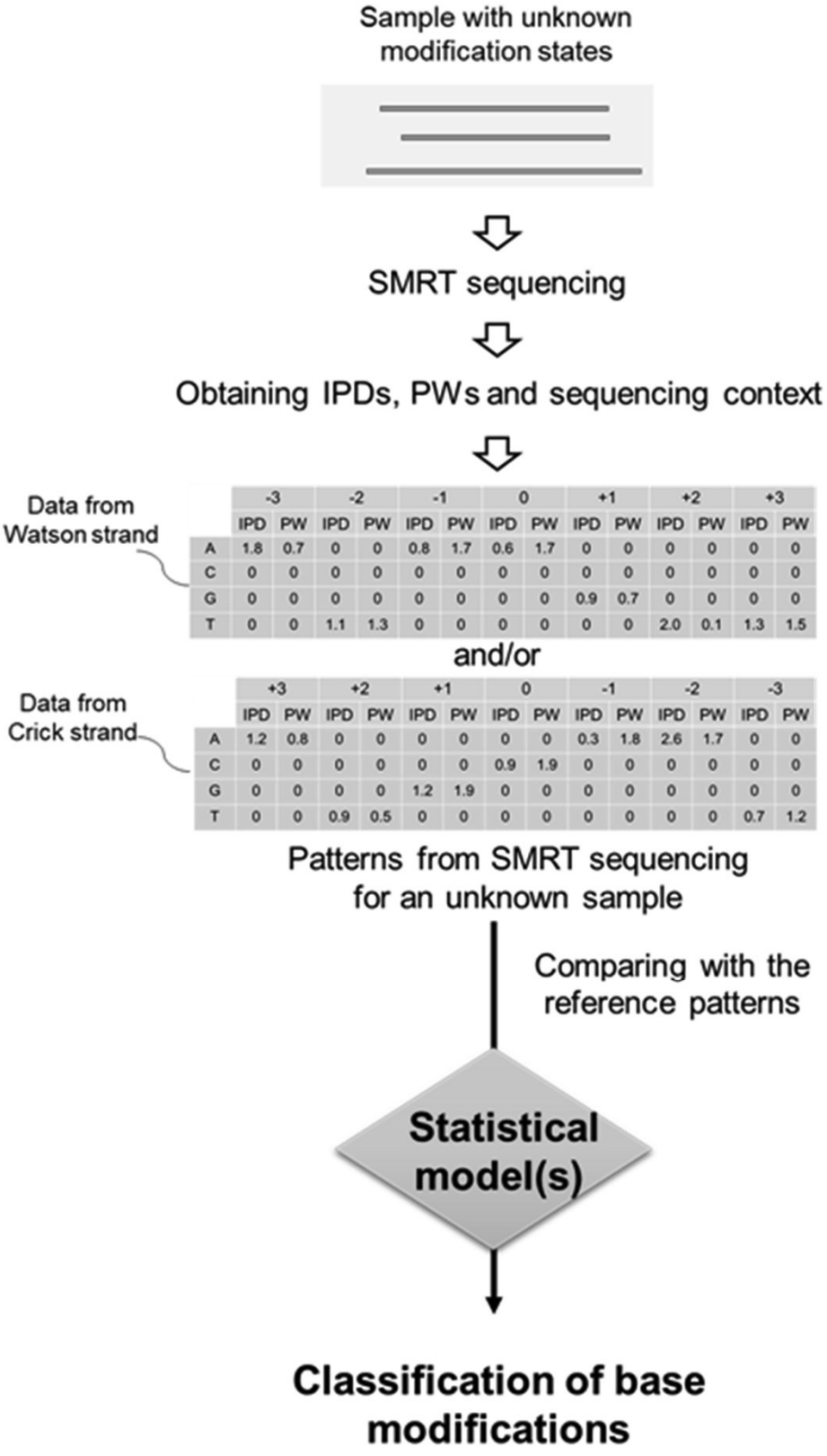

Sample known to be unmodified DNA
Sample known to be modified DNA
Unmodified base
Modified base
902
904
906
908
SMRT sequencing 910
Obtaining IPDs, PWs and sequencing context 912
Building the reference patterns deduced from SMRT sequencing indicating the modification states 914
Data from Watson strand
Data from Watson strand
Data from Crick strand
and/or
and/or
Data from Crick strand
Reference patterns from a sample with unmodified DNA
Reference patterns from a sample with modified DNA
Training 916
Statistical model(s) 918

FIG. 9



	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5

	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.2	0.8	0	0	0	0	0	0	0.3	1.8	2.6	1.7	0	0
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0	0
T	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0.7	1.2

FIG. 10

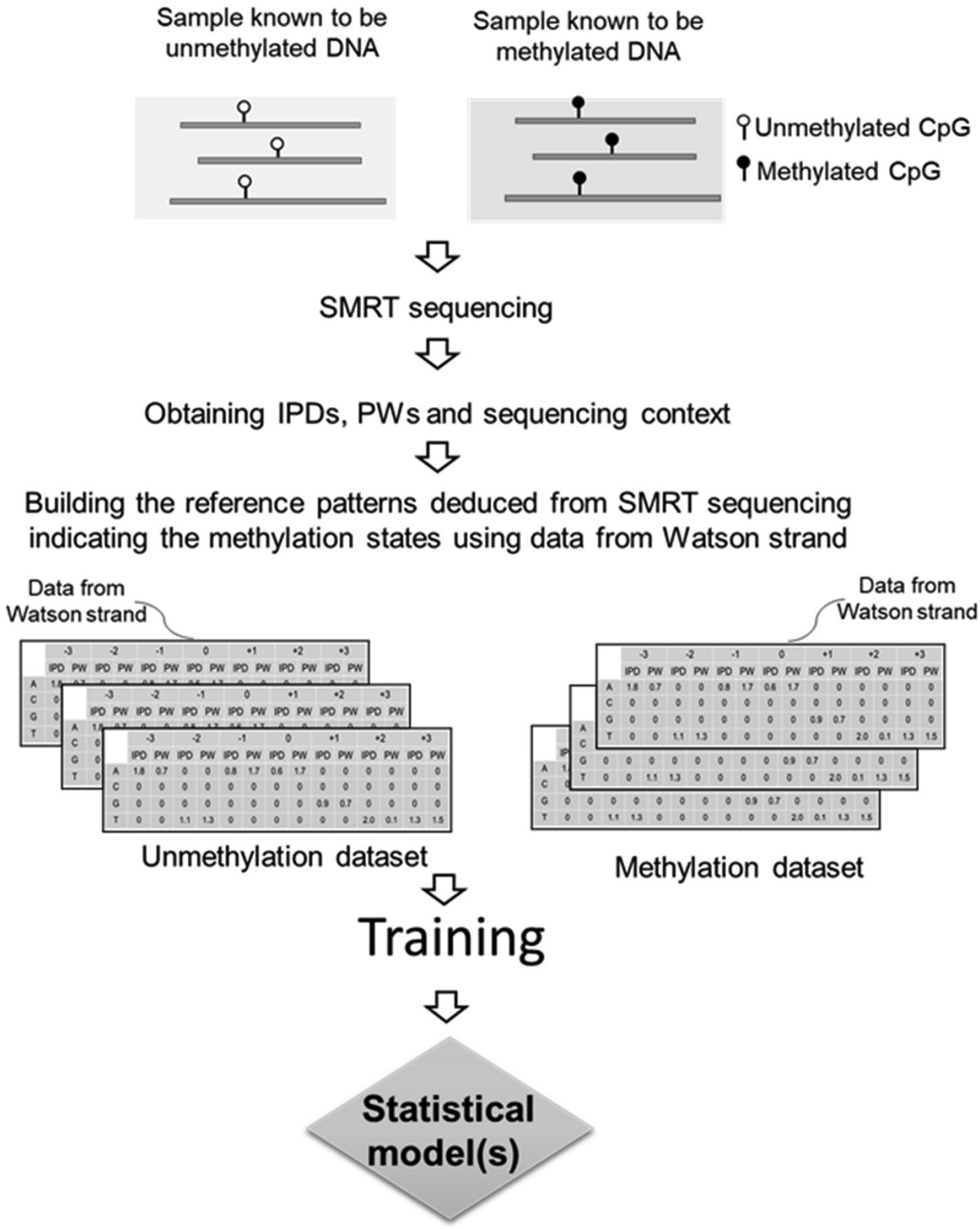
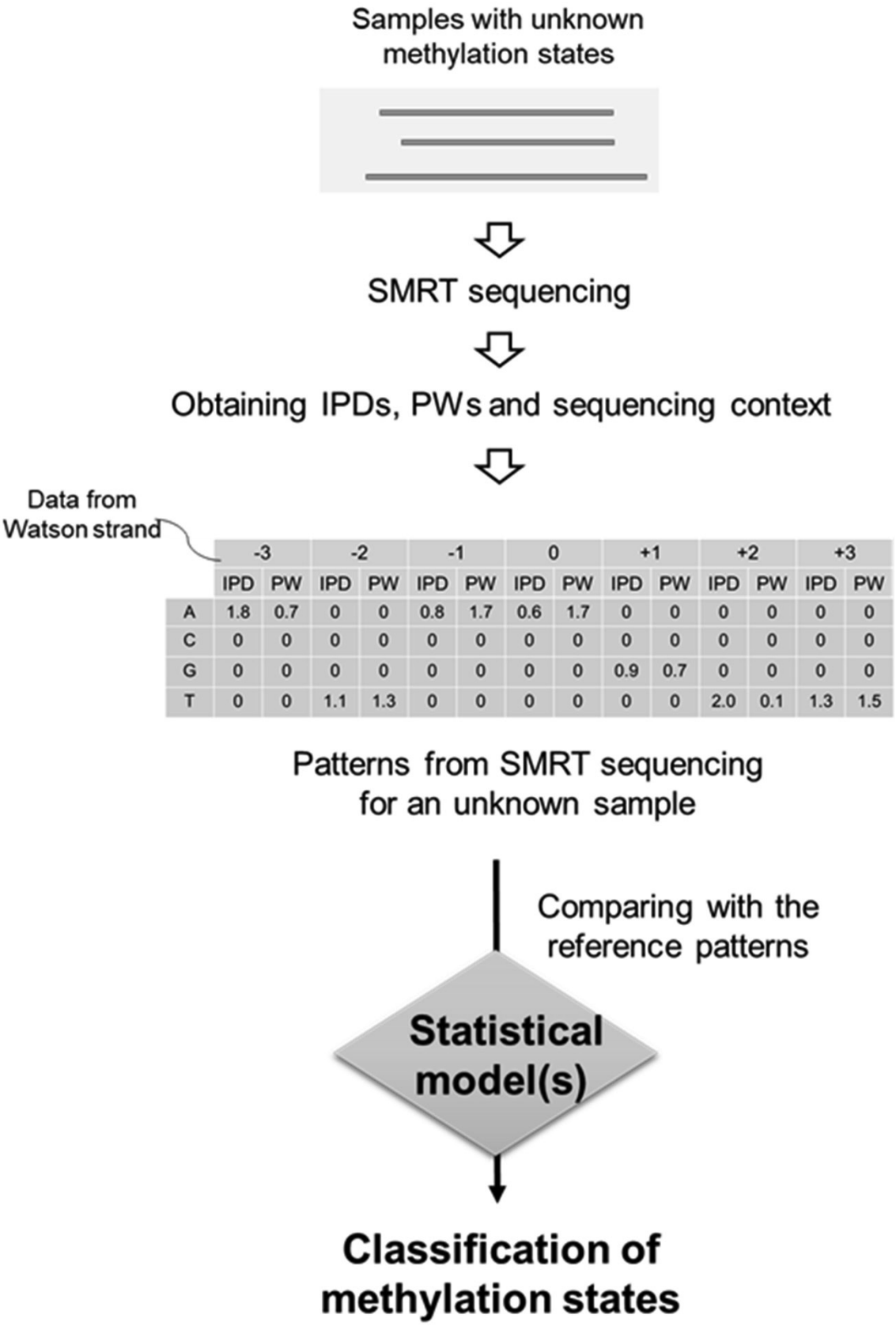

Sample known to be unmethylated DNA
Sample known to be methylated DNA
Unmethylated CpG
Methylated CpG
SMRT sequencing
Obtaining IPDs, PWs and sequencing context
Building the reference patterns deduced from SMRT sequencing indicating the methylation states using data from Watson strand
Data from Watson strand
Data from Watson strand
Unmethylation dataset
Methylation dataset
Training
Statistical model(s)

FIG. 11



	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5

FIG. 12

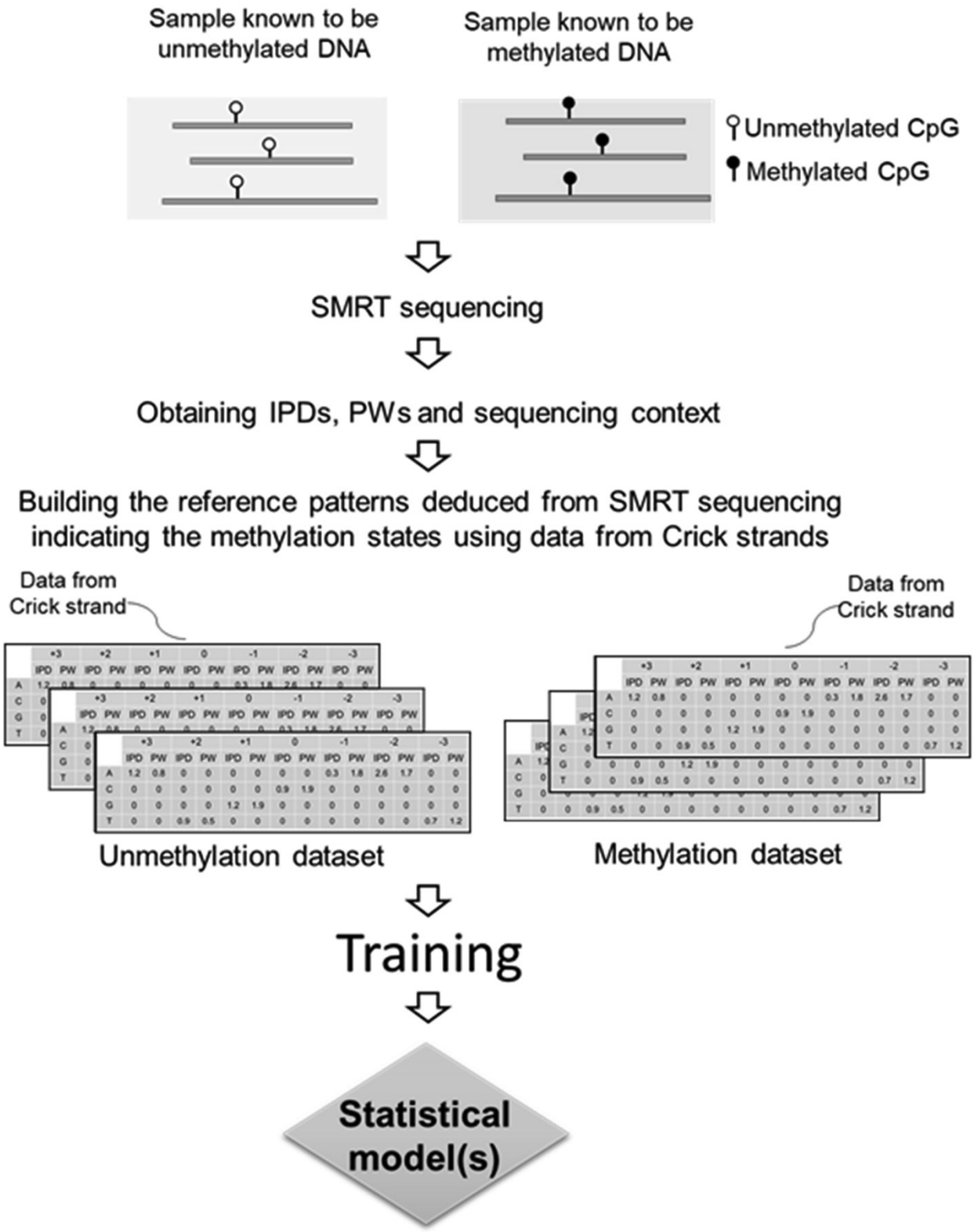
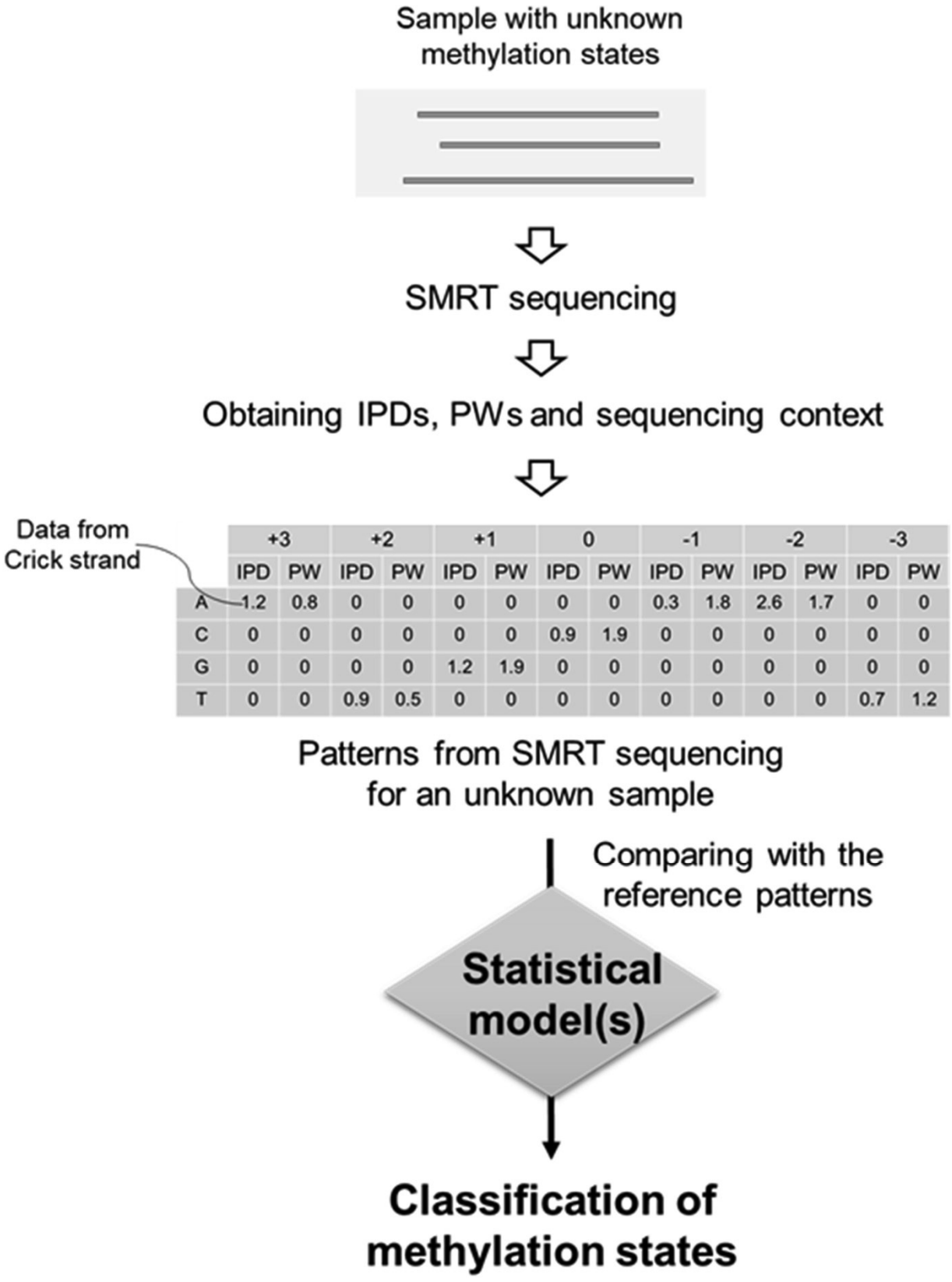

Sample known to be unmethylated DNA
Sample known to be methylated DNA
Unmethylated CpG
Methylated CpG
SMRT sequencing
Obtaining IPDs, PWs and sequencing context
Building the reference patterns deduced from SMRT sequencing indicating the methylation states using data from Crick strands
Data from Crick strand
Data from Crick strand
Unmethylation dataset
Methylation dataset
Training
Statistical model(s)

FIG. 13



	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.2	0.8	0	0	0	0	0	0	0.3	1.8	2.6	1.7	0	0
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0	0
T	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0.7	1.2

FIG.14

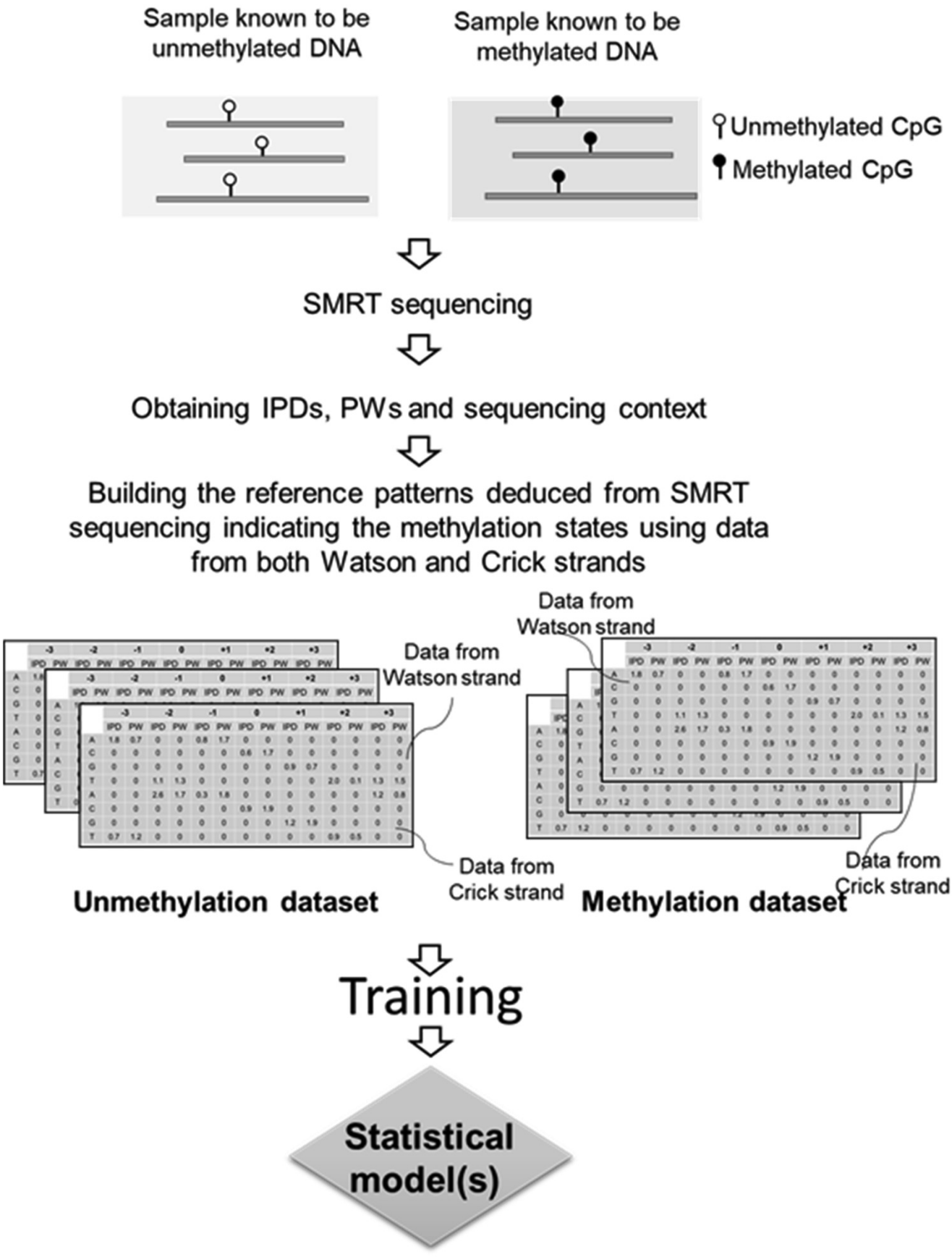
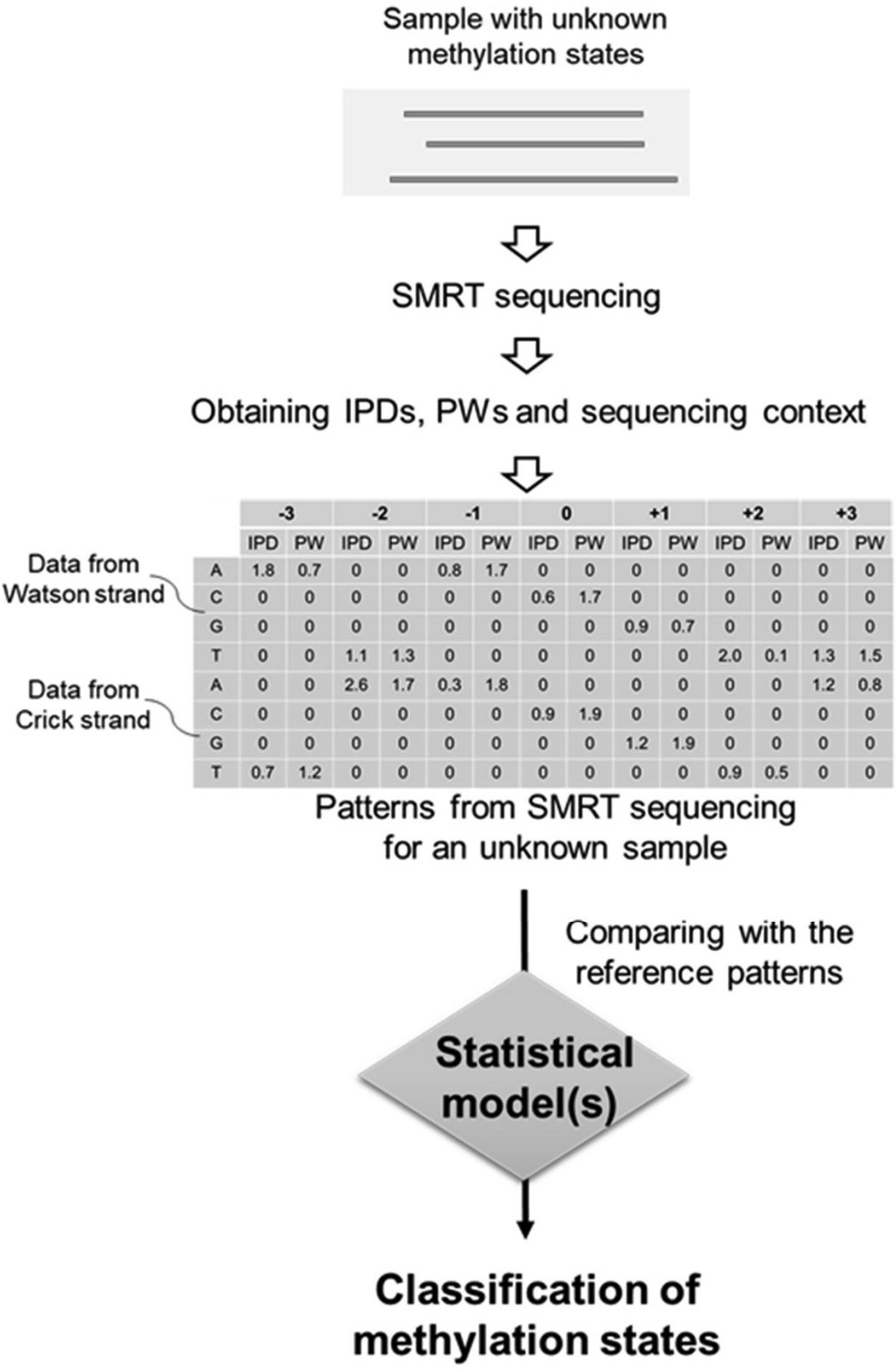

Sample known to be unmethylated DNA
Sample known to be methylated DNA
Unmethylated CpG
Methylated CpG
SMRT sequencing
Obtaining IPDs, PWs and sequencing context
Building the reference patterns deduced from SMRT sequencing indicating the methylation states using data from both Watson and Crick strands
Data from Watson strand
Data from Watson strand
Data from Crick strand
Data from Crick strand
Unmethylation dataset
Methylation dataset
Training
Statistical model(s)

FIG. 15



	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0.6	1.7	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5
A	0	0	2.6	1.7	0.3	1.8	0	0	0	0	0	0	1.2	0.8
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	1.2	1.9	0	0	0	0
T	0.7	1.2	0	0	0	0	0	0	0	0	0.9	0.5	0	0

FIG. 16

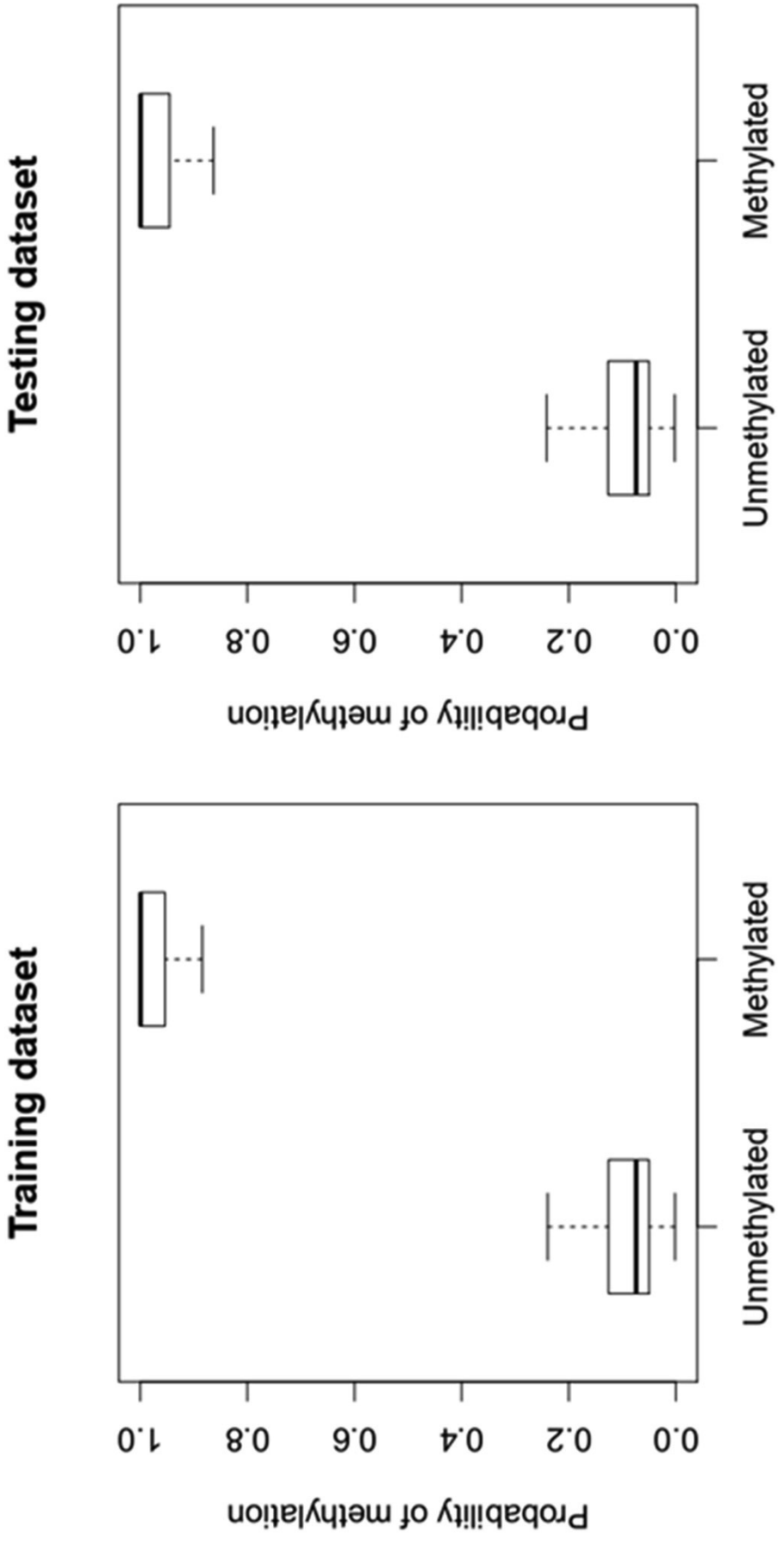


FIG. 17A

FIG. 17B

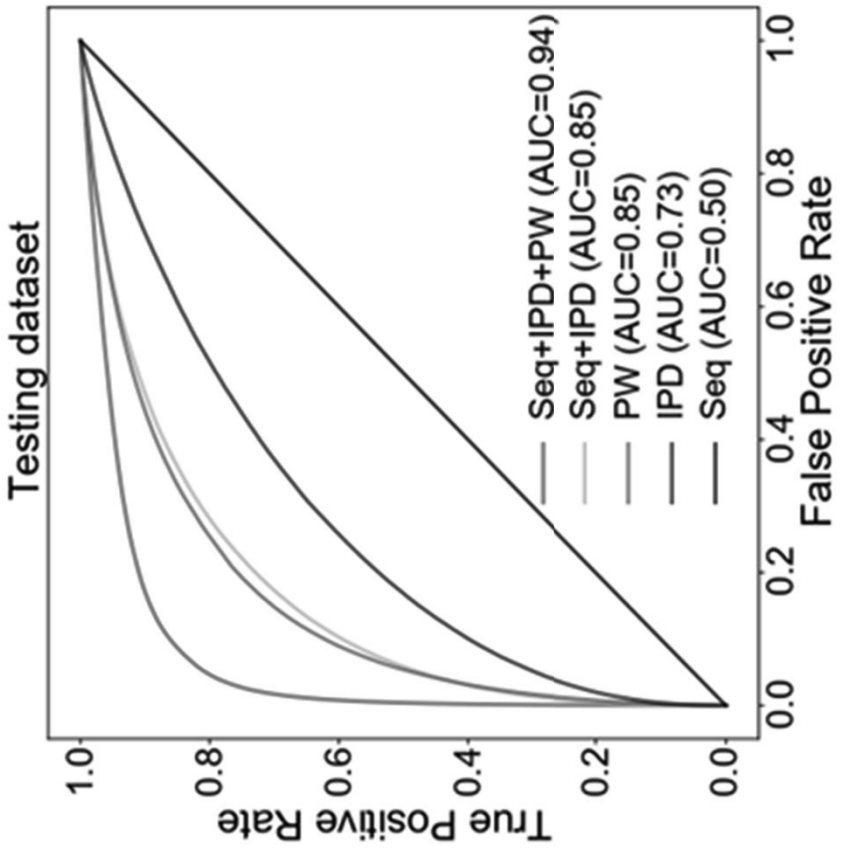


FIG. 18B

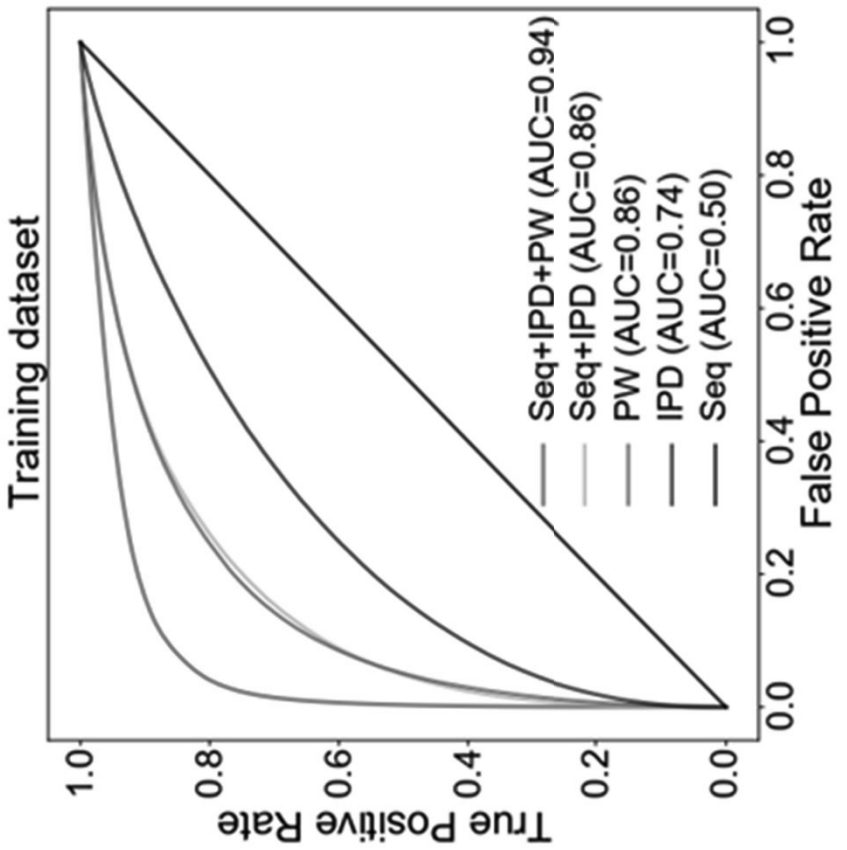


FIG. 18A

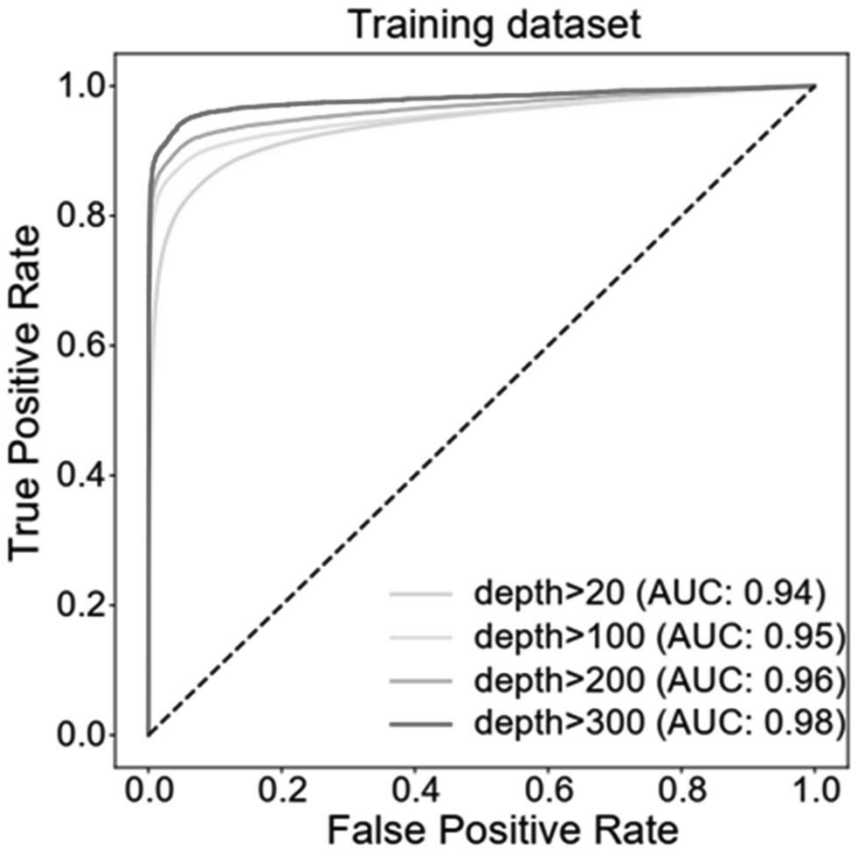


FIG. 19A

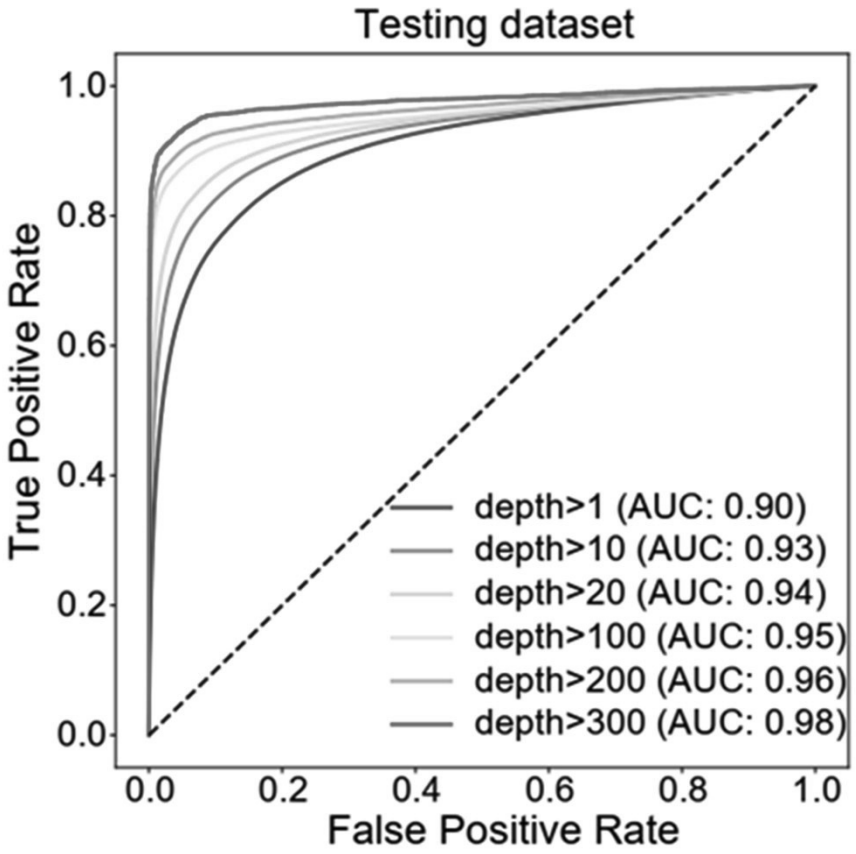


FIG. 19B

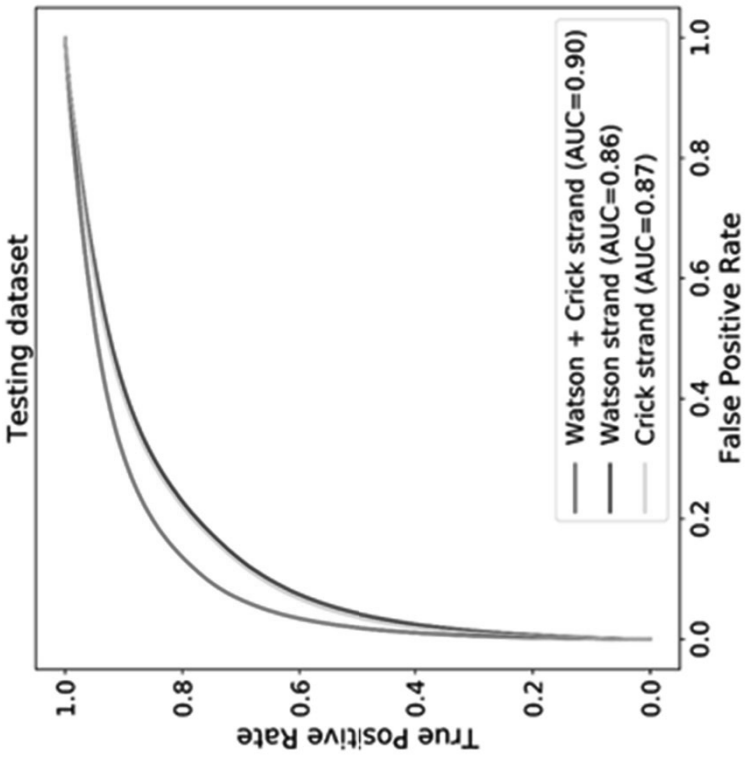


FIG. 20B

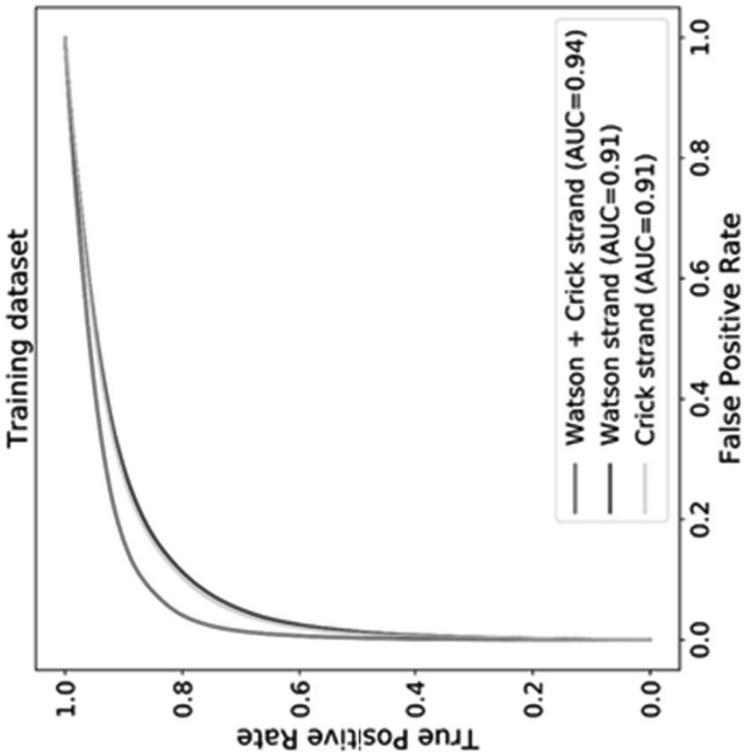


FIG. 20A

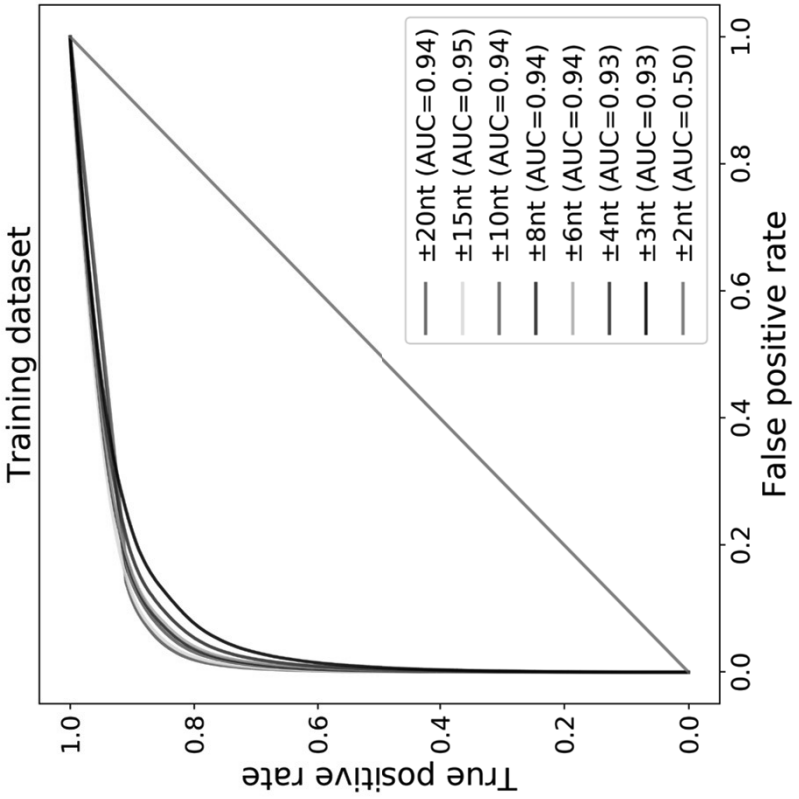


FIG. 21A

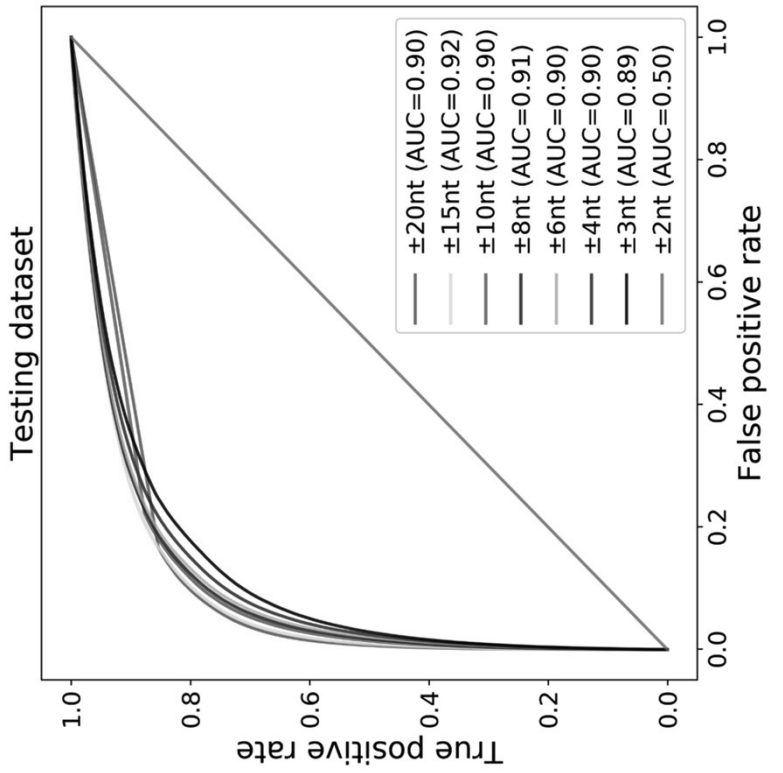


FIG. 21B

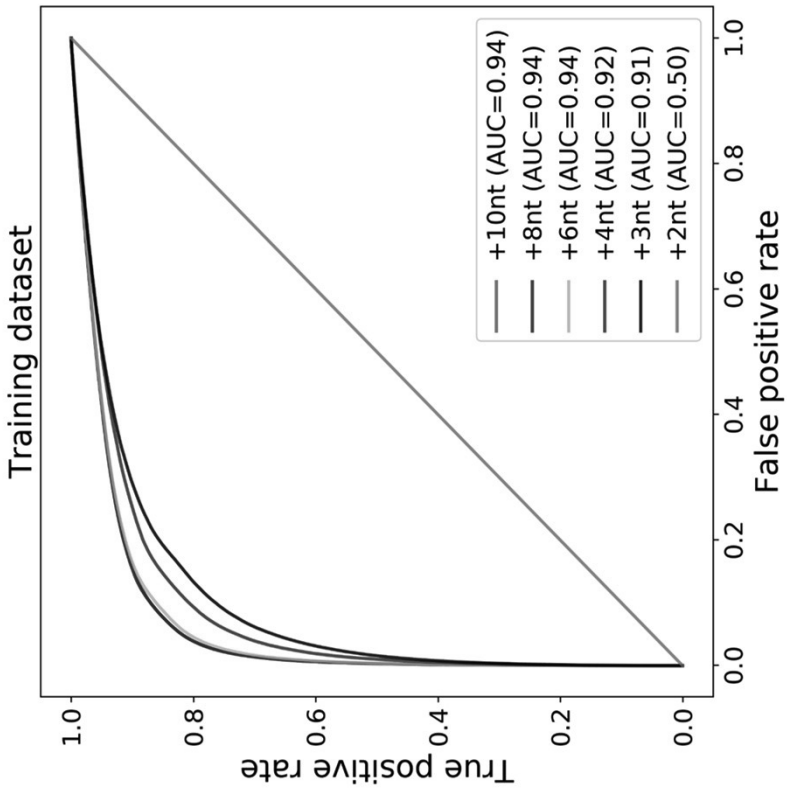


FIG. 22A

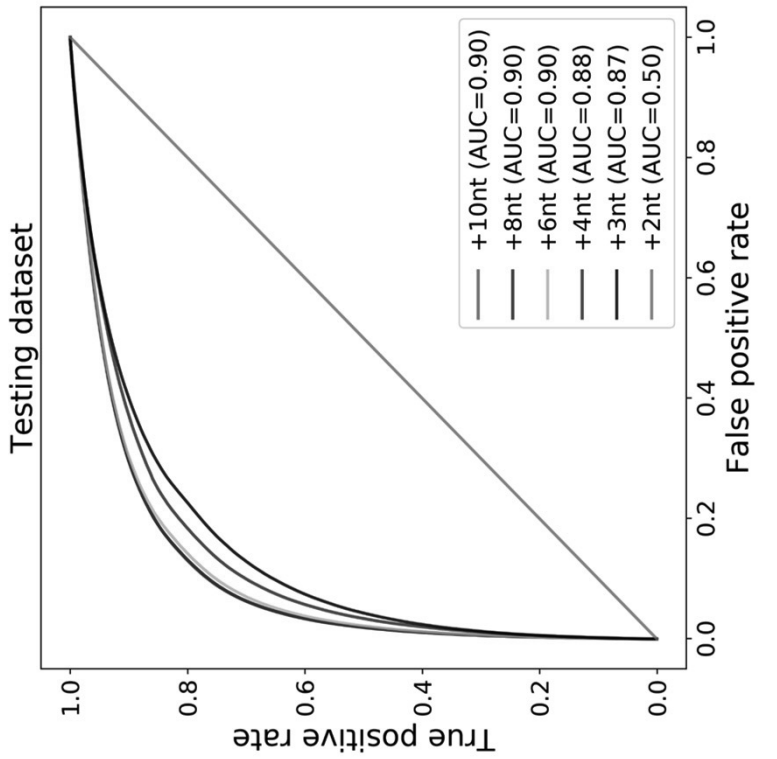


FIG. 22B

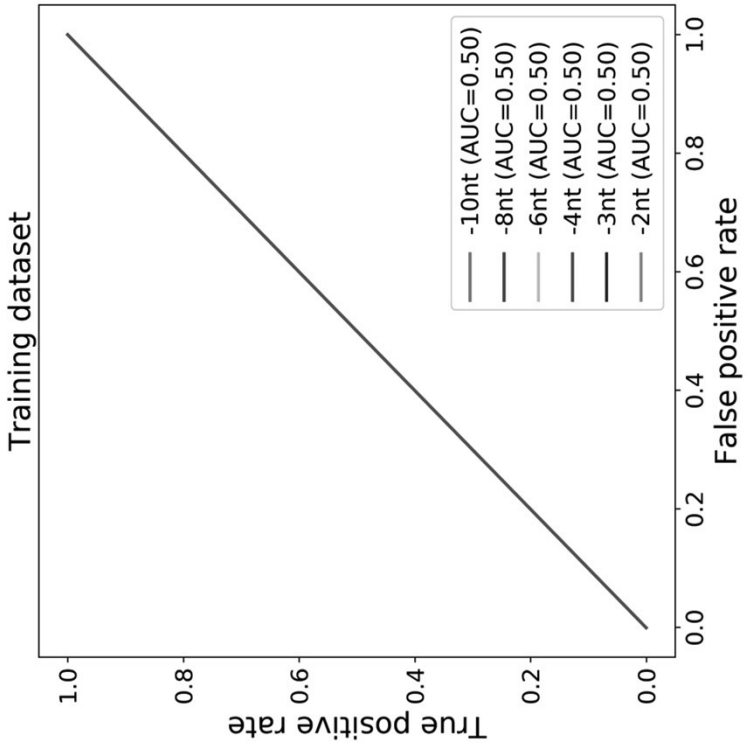


FIG. 23A

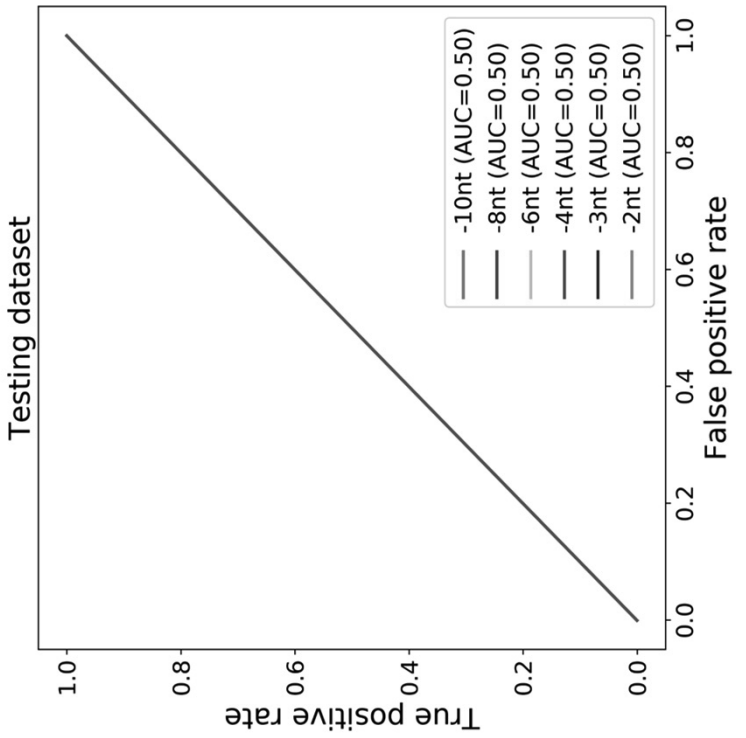


FIG. 23B

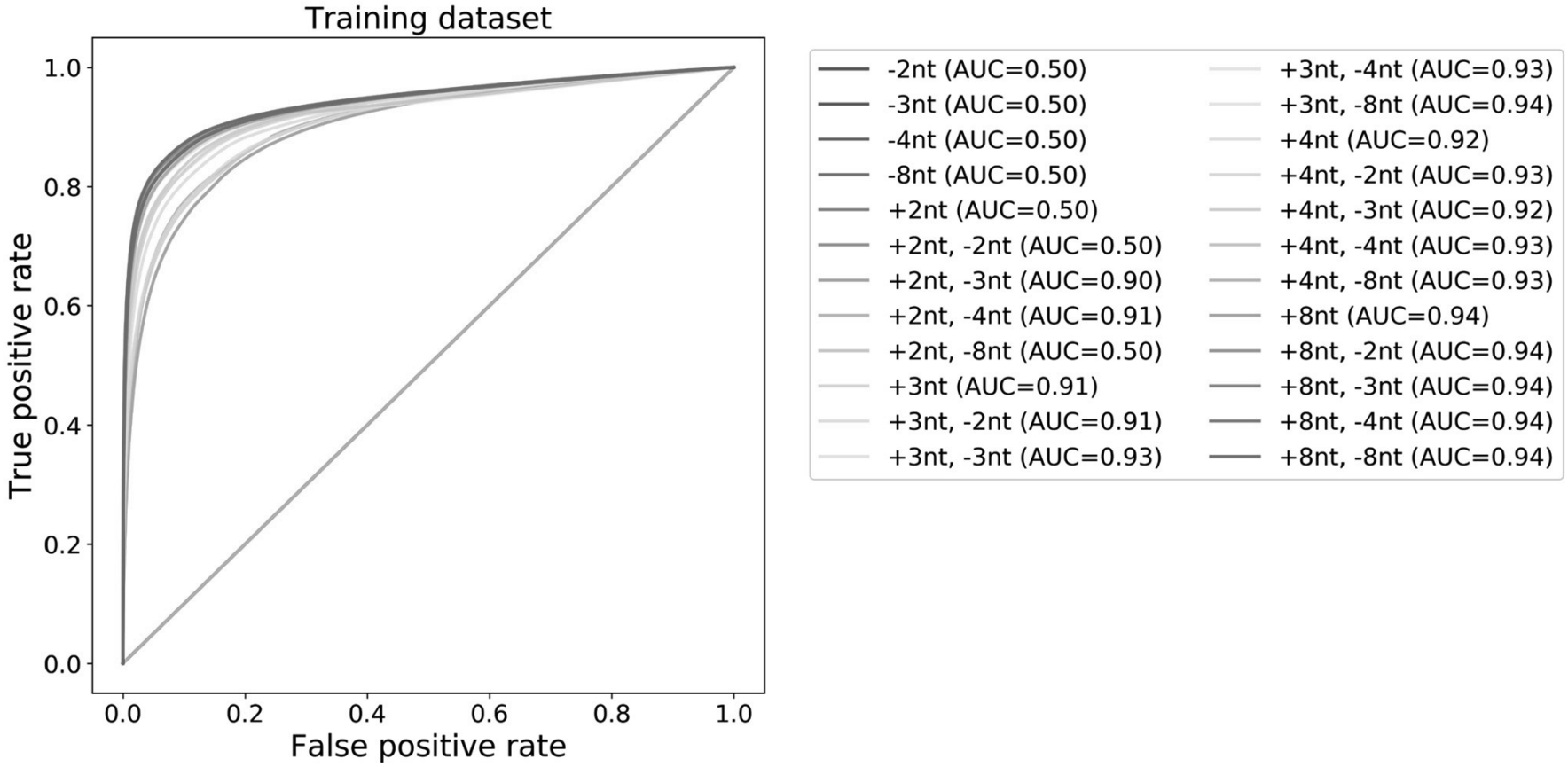

Training dataset
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
-2nt (AUC=0.50)
-3nt (AUC=0.50)
-4nt (AUC=0.50)
-8nt (AUC=0.50)
+2nt (AUC=0.50)
+2nt, -2nt (AUC=0.50)
+2nt, -3nt (AUC=0.90)
+2nt, -4nt (AUC=0.91)
+2nt, -8nt (AUC=0.50)
+3nt (AUC=0.91)
+3nt, -2nt (AUC=0.91)
+3nt, -3nt (AUC=0.93)
+3nt, -4nt (AUC=0.93)
+3nt, -8nt (AUC=0.94)
+4nt (AUC=0.92)
+4nt, -2nt (AUC=0.93)
+4nt, -3nt (AUC=0.92)
+4nt, -4nt (AUC=0.93)
+4nt, -8nt (AUC=0.93)
+8nt (AUC=0.94)
+8nt, -2nt (AUC=0.94)
+8nt, -3nt (AUC=0.94)
+8nt, -4nt (AUC=0.94)
+8nt, -8nt (AUC=0.94)

FIG. 24

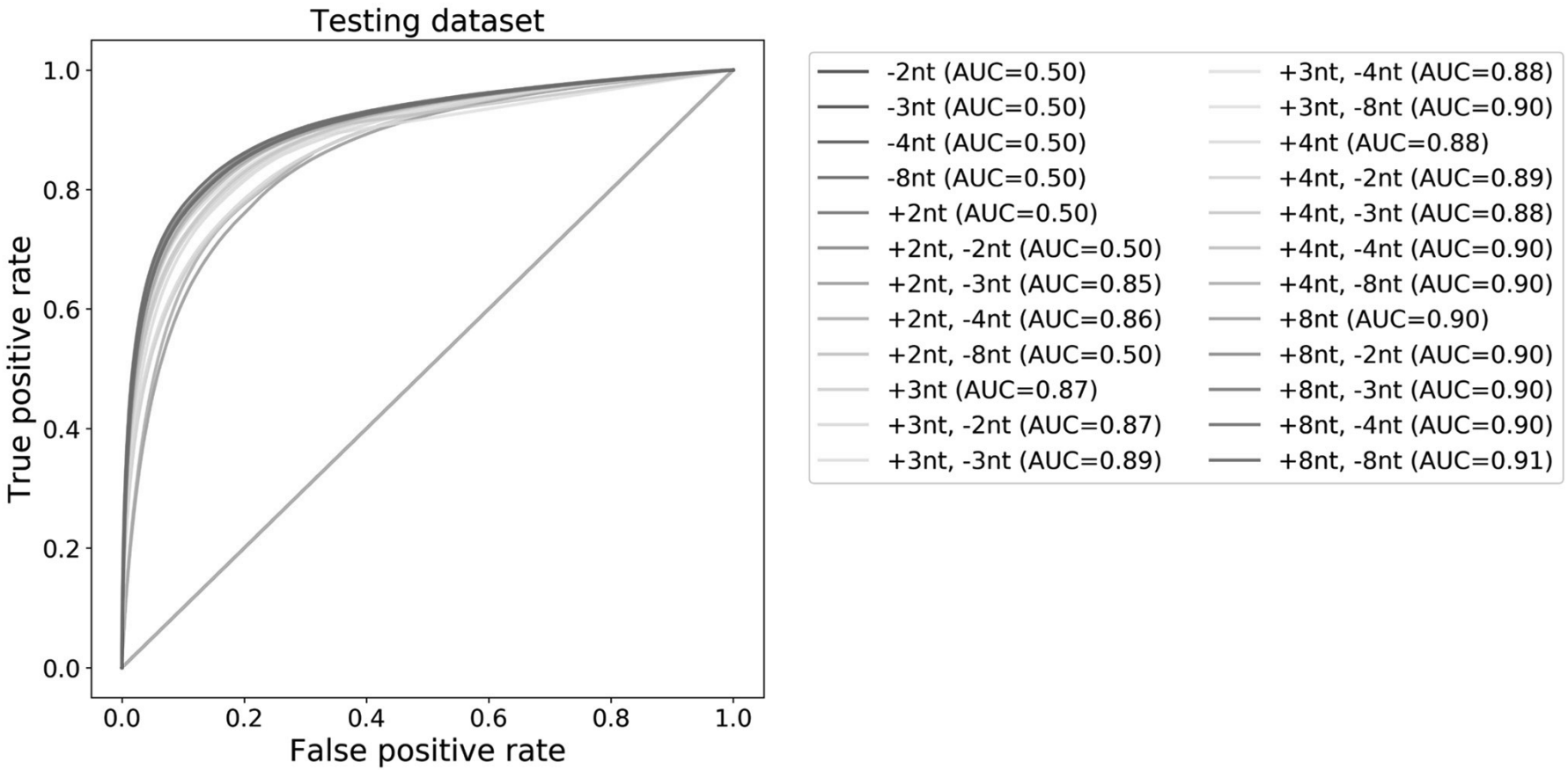

Testing dataset
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
-2nt (AUC=0.50)
-3nt (AUC=0.50)
-4nt (AUC=0.50)
-8nt (AUC=0.50)
+2nt (AUC=0.50)
+2nt, -2nt (AUC=0.50)
+2nt, -3nt (AUC=0.85)
+2nt, -4nt (AUC=0.86)
+2nt, -8nt (AUC=0.50)
+3nt (AUC=0.87)
+3nt, -2nt (AUC=0.87)
+3nt, -3nt (AUC=0.89)
+3nt, -4nt (AUC=0.88)
+3nt, -8nt (AUC=0.90)
+4nt (AUC=0.88)
+4nt, -2nt (AUC=0.89)
+4nt, -3nt (AUC=0.88)
+4nt, -4nt (AUC=0.90)
+4nt, -8nt (AUC=0.90)
+8nt (AUC=0.90)
+8nt, -2nt (AUC=0.90)
+8nt, -3nt (AUC=0.90)
+8nt, -4nt (AUC=0.90)
+8nt, -8nt (AUC=0.91)

FIG. 25

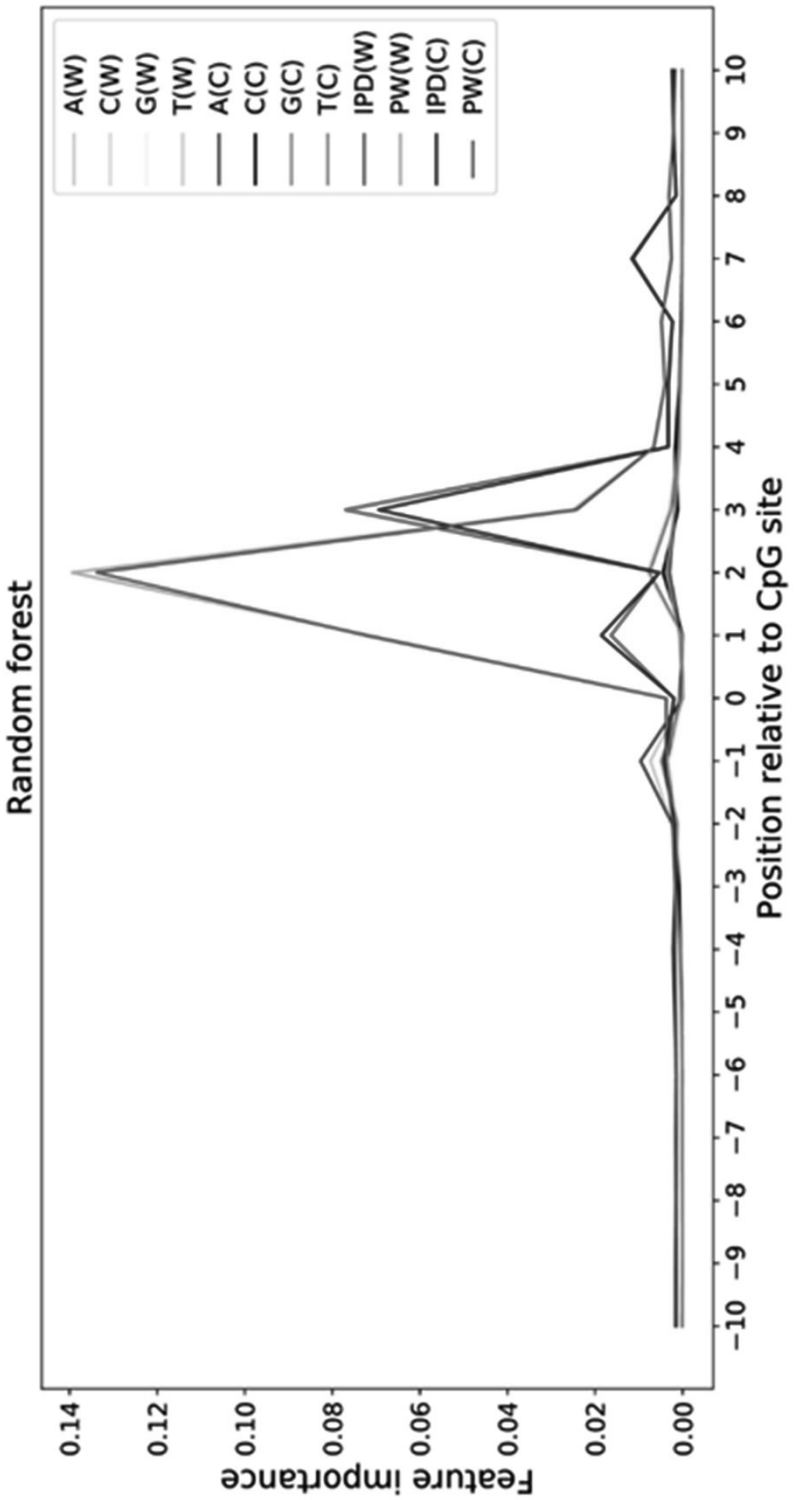

Random forest
Feature importance
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−10 −9 −8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7 8 9 10
Position relative to CpG site
A(W)
C(W)
G(W)
T(W)
A(C)
C(C)
G(C)
T(C)
IPD(W)
PW(W)
IPD(C)
PW(C)

FIG. 26

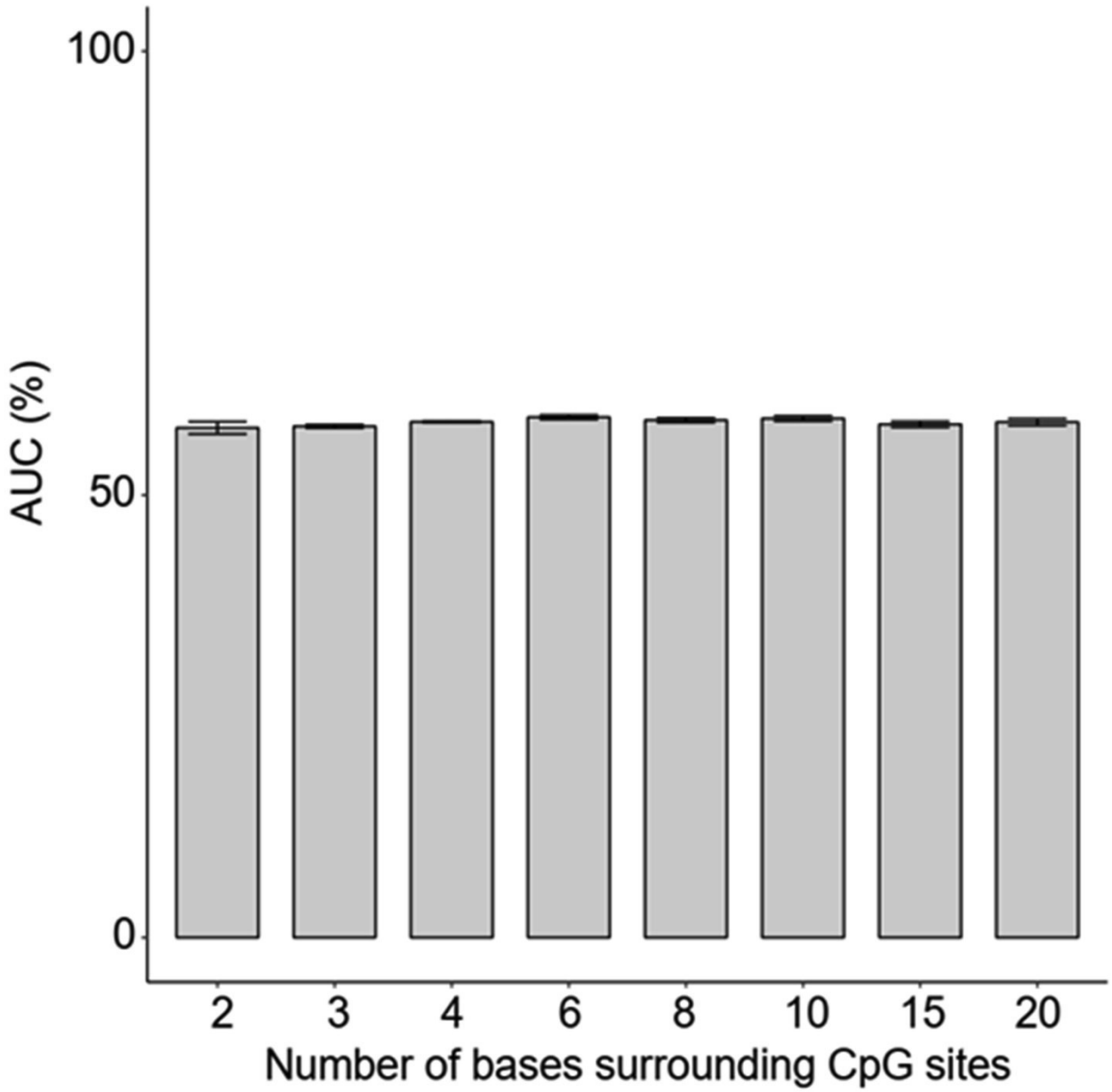

100
50
0
AUC (%)
2
3
4
6
8
10
15
20
Number of bases surrounding CpG sites

FIG. 27

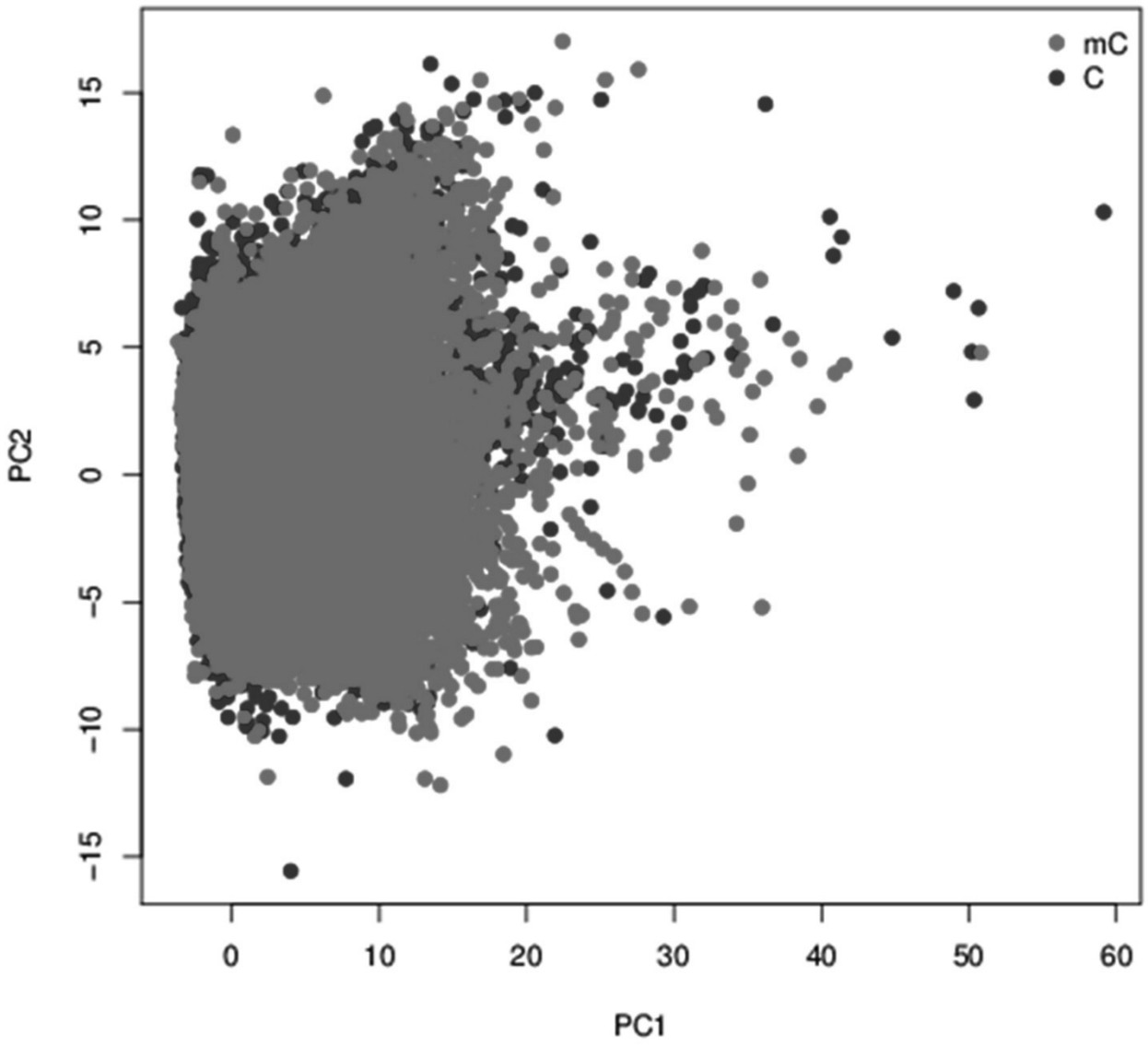


FIG. 28

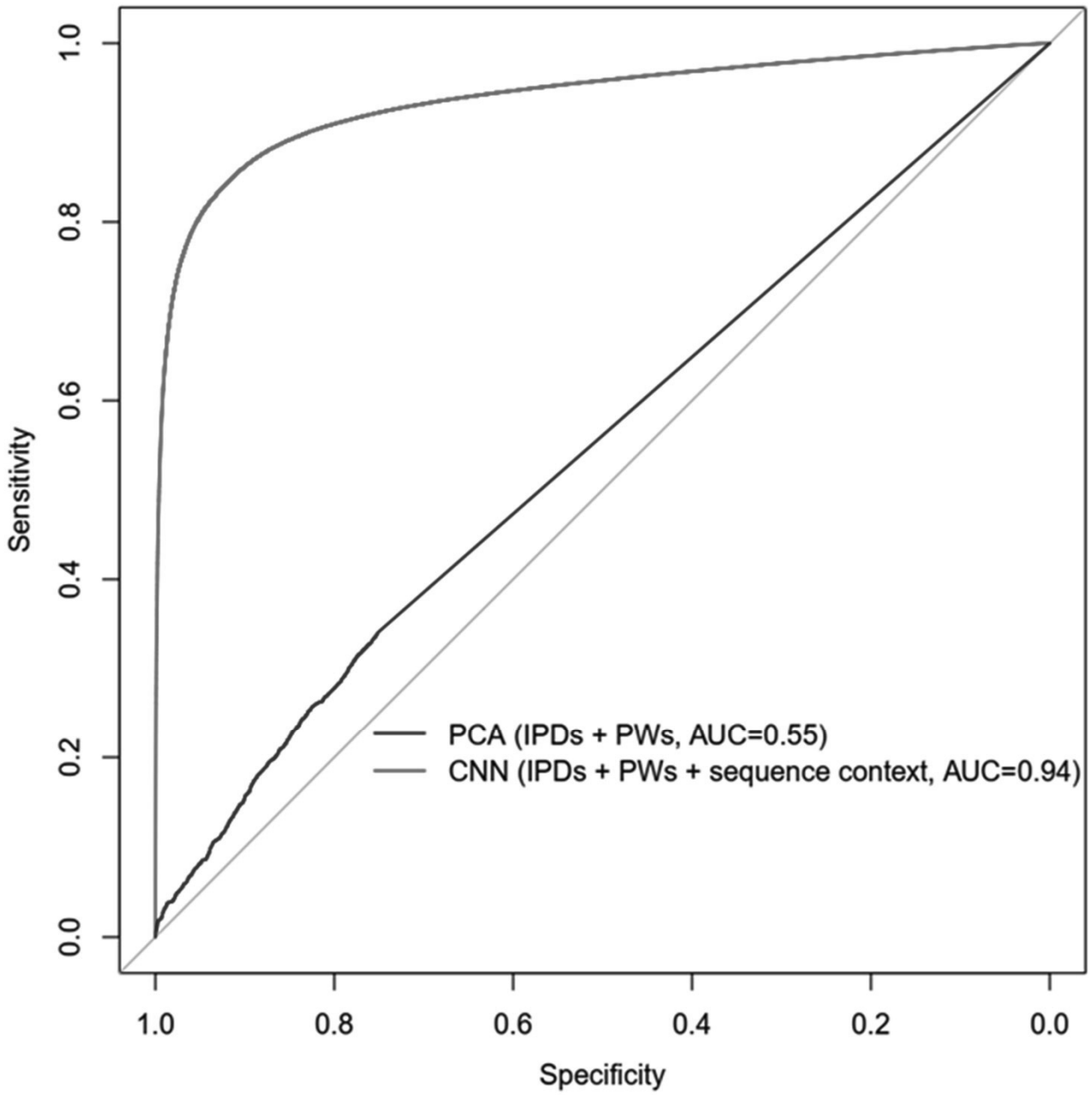

Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
PCA (IPDs + PWs, AUC=0.55)
CNN (IPDs + PWs + sequence context, AUC=0.94)

FIG. 29

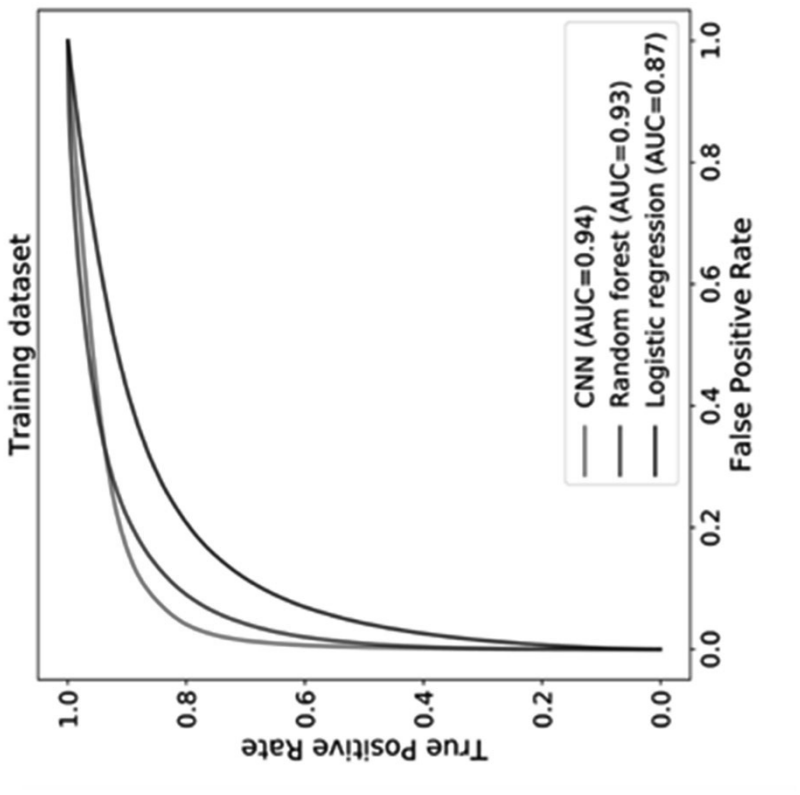


FIG. 30A

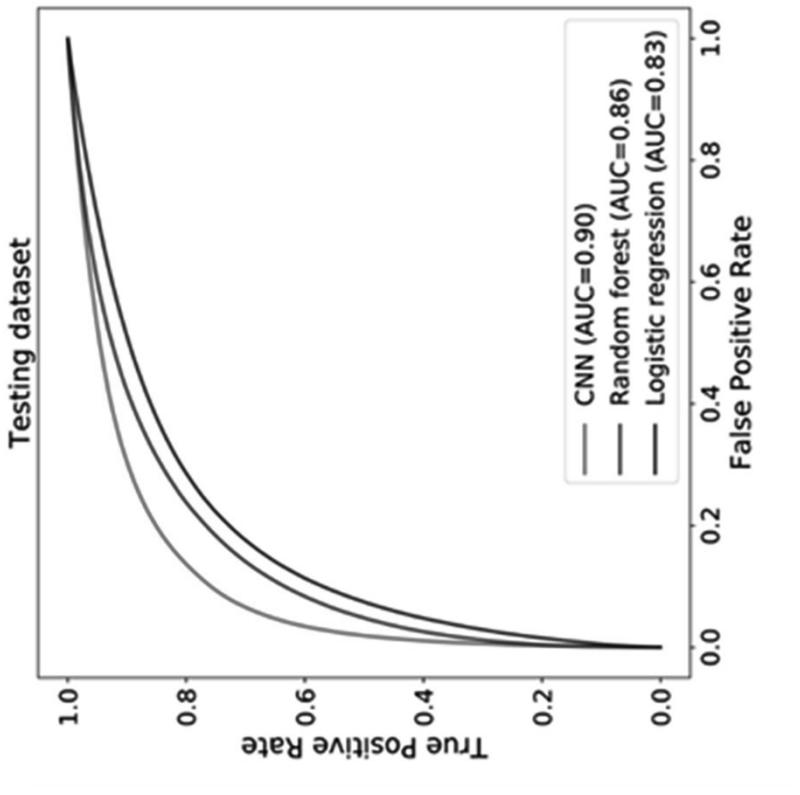
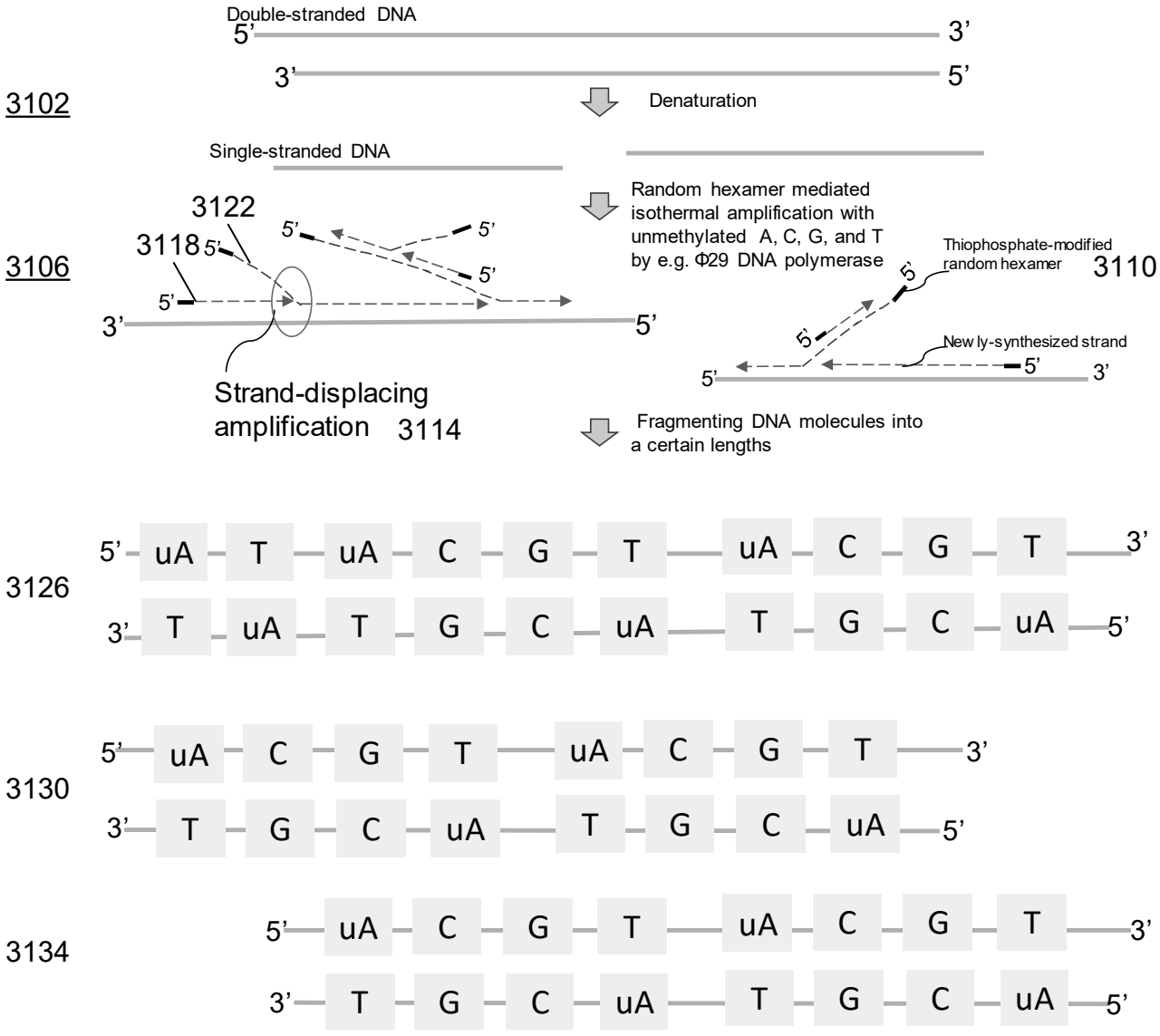
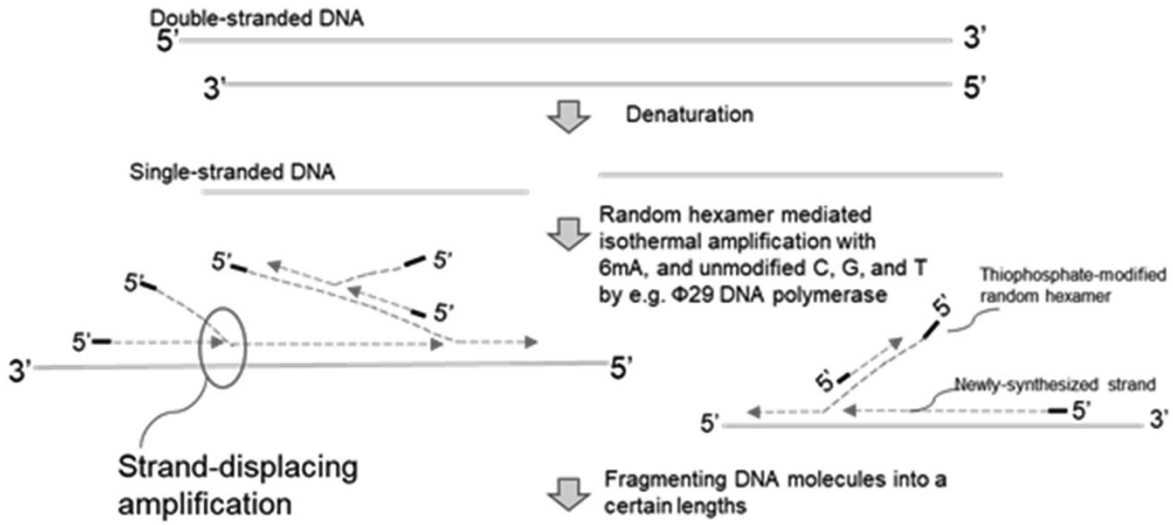


FIG. 30B



Whole-genome amplified DNA products

FIG. 31A



(I)

5' mA T mA C G T mA C G T 3'

3' T mA T G C mA T G C mA 5'

(II)

5' uA C G T mA C G T 3'

3' T G C mA T G C mA 5'

(III)

5' uA C G T uA C G T 3'

3' T G C mA T G C mA 5'

(IV)

5' uA C G T uA C G T 3'

3' T G C uA T G C uA 5'

Whole-genome amplified DNA products

FIG. 31B

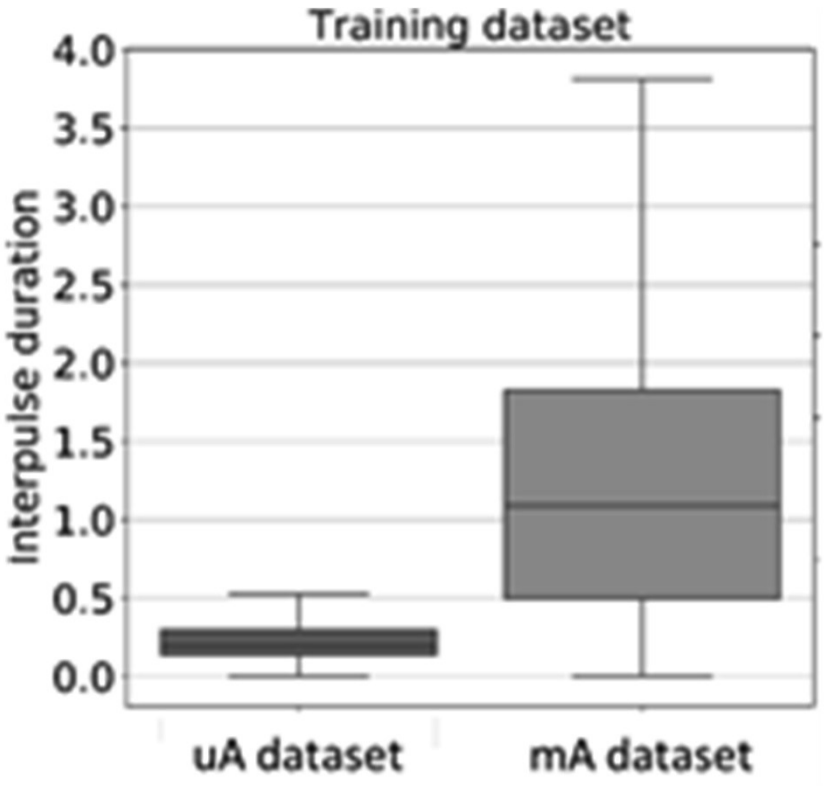


FIG. 32A

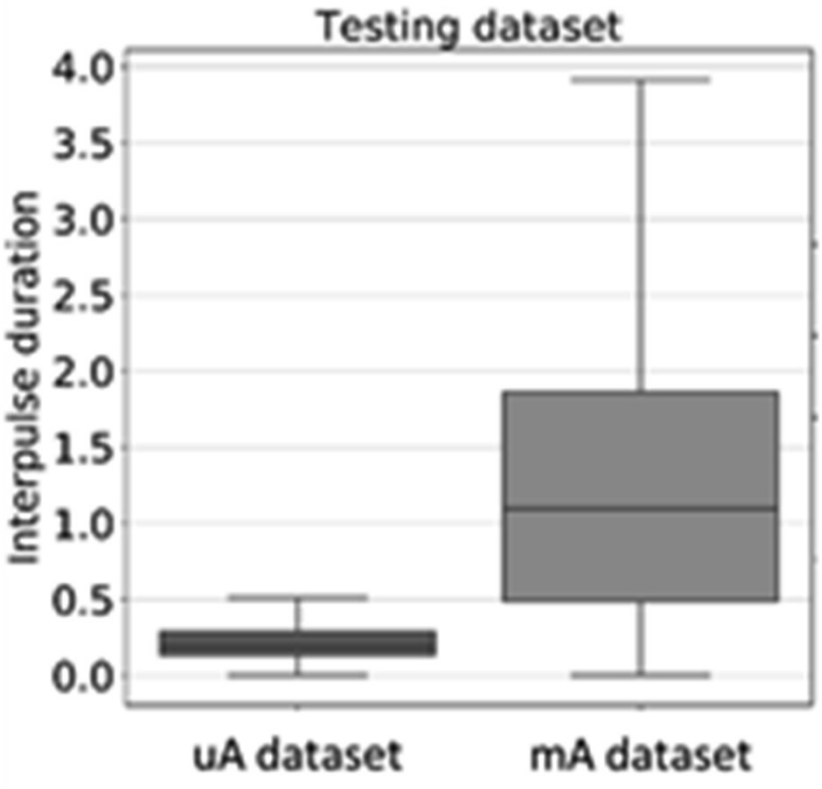


FIG. 32B

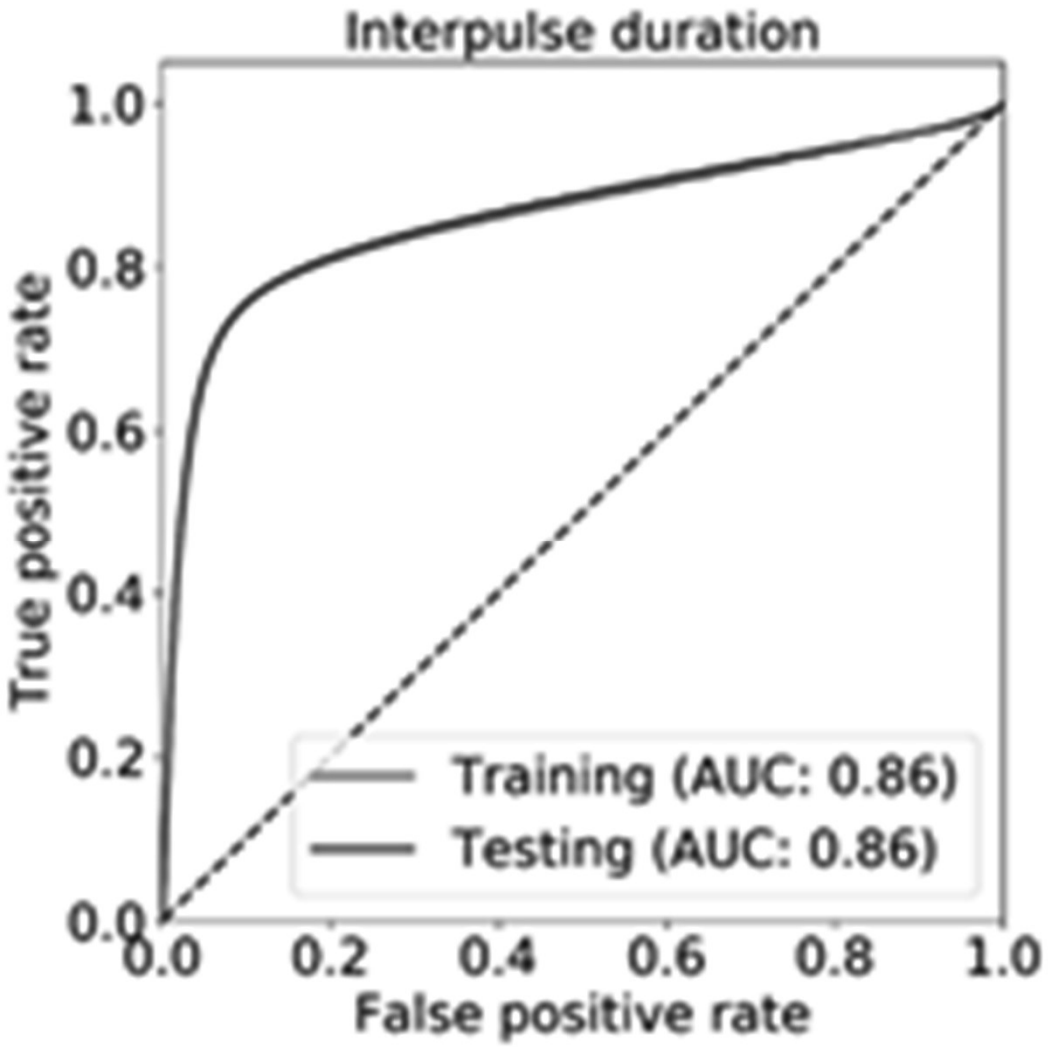

Interpulse duration
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Training (AUC: 0.86)
Testing (AUC: 0.86)

FIG. 32C

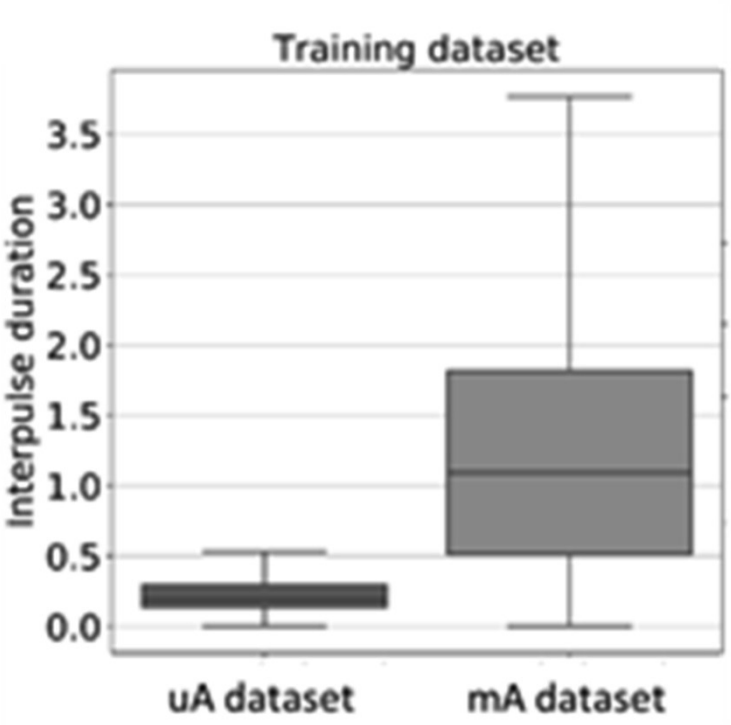


FIG. 33A

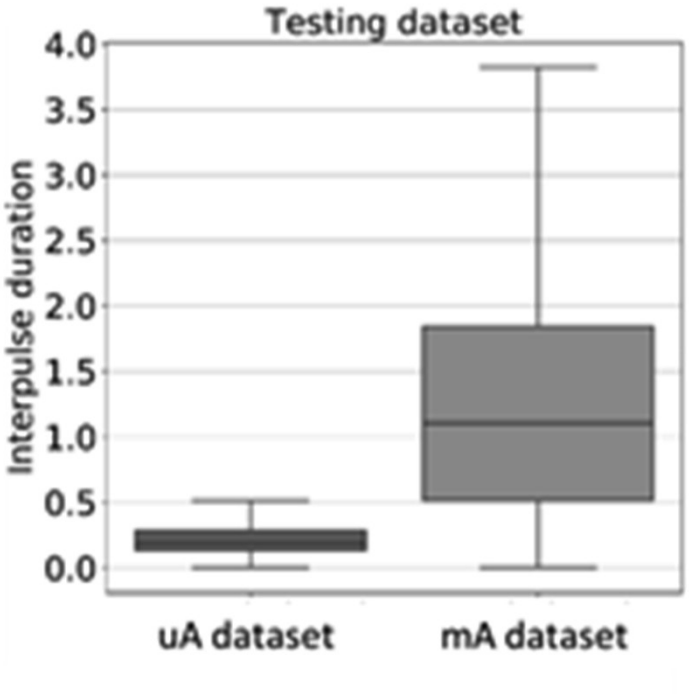


FIG. 33B

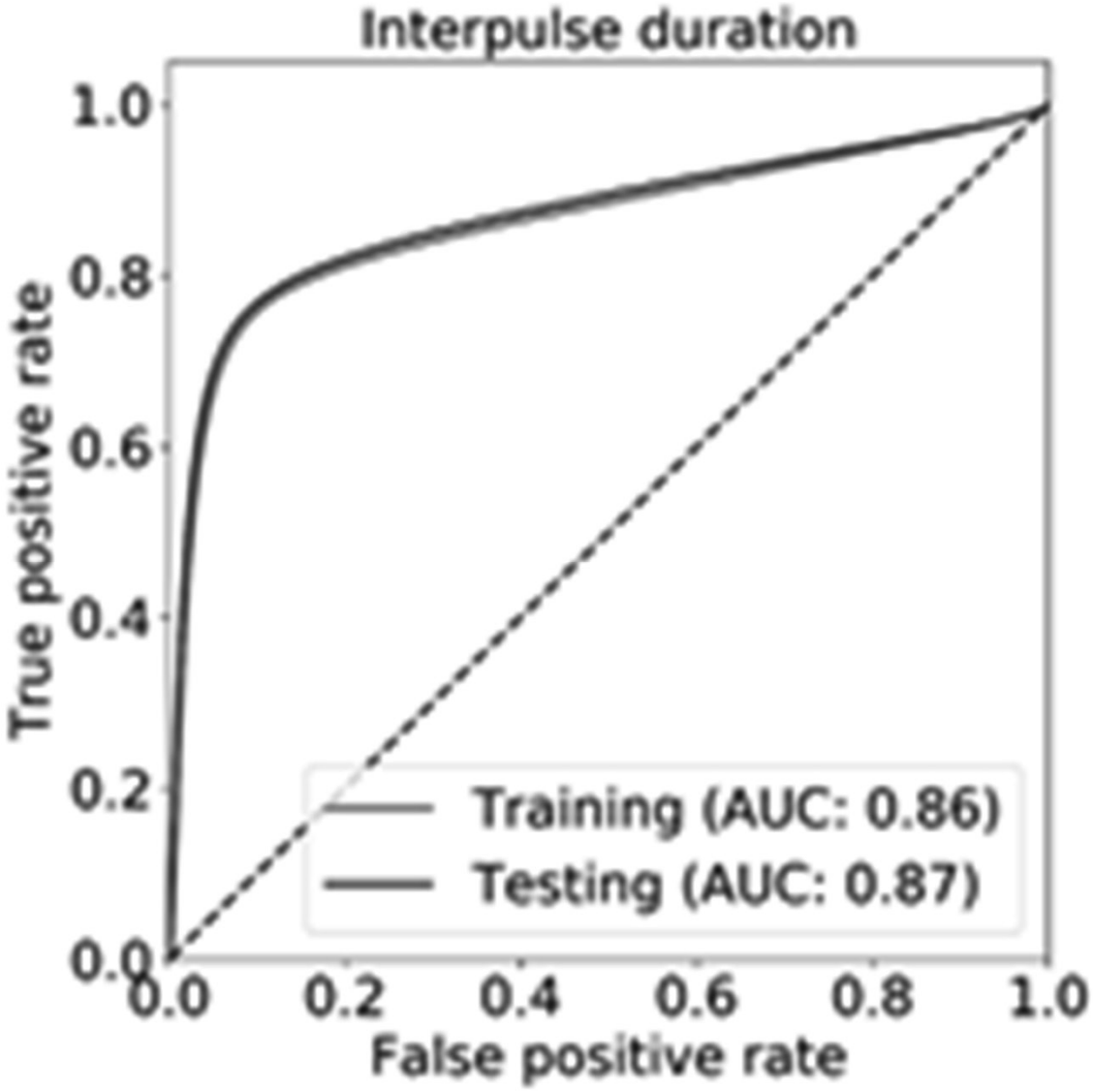
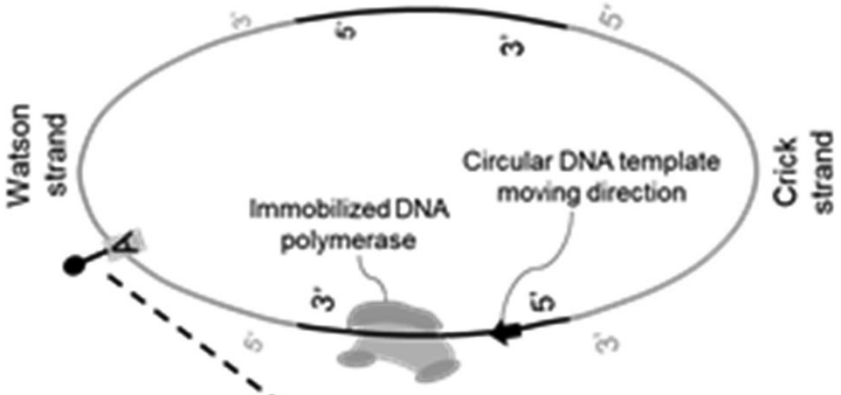

Interpulse duration
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Training (AUC: 0.86)
Testing (AUC: 0.87)

FIG. 33C







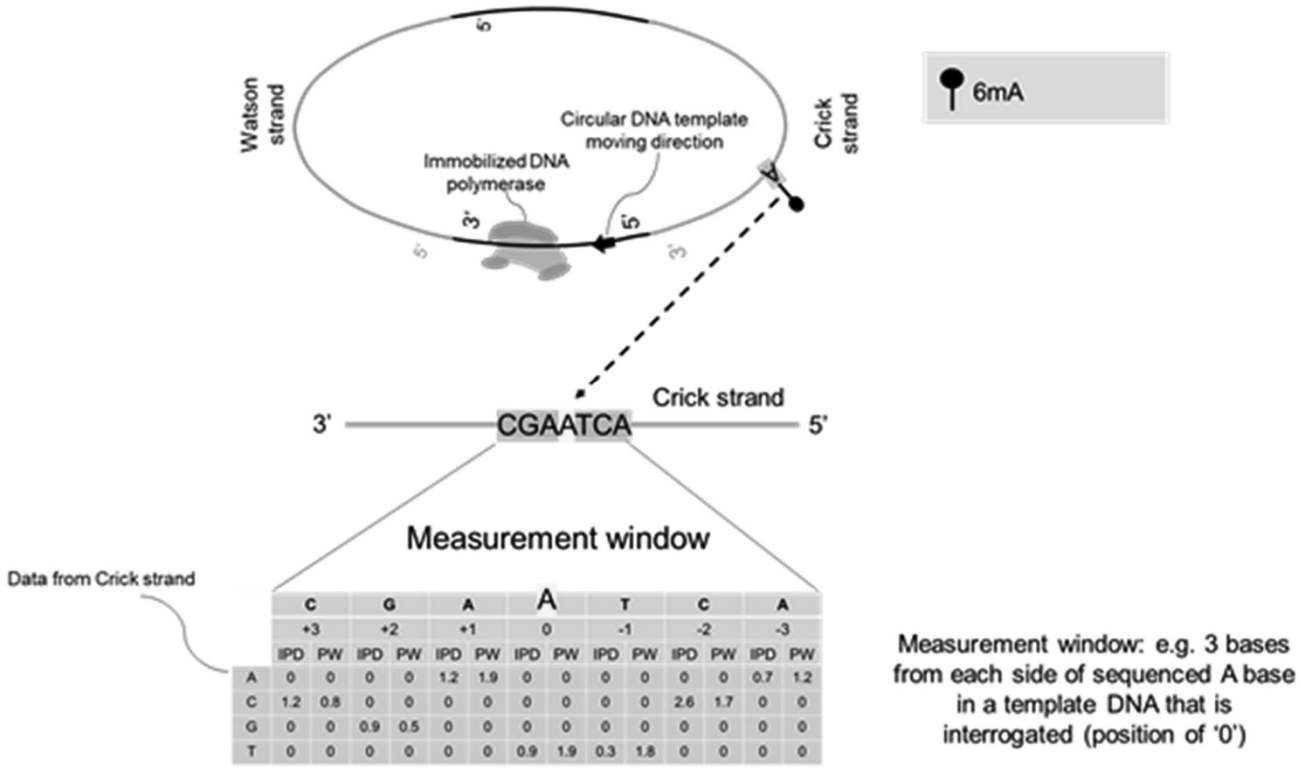
Measurement window

Data from Watson strand

	G		A		T		A		A		C		T	
	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	1.0	0.3	0	0	0	0	1.1	0.8	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	2.3	0.8	0	0
G	1.5	1.7	0	0	0	0	0.9	1.2	0	0	0	0	0	0
T	0	0	0	0	1.2	0.2	0	0	0	0	0	0	1.5	1.7

Measurement window: e.g. 3 bases from each side of sequenced A base in a template DNA that is interrogated (position of '0')

FIG. 34



	C		G		A		A		T		C		A	
	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0.7	1.2
C	1.2	0.8	0	0	0	0	0	0	0	0	2.6	1.7	0	0
G	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0.9	1.9	0.3	1.8	0	0	0	0

FIG.35

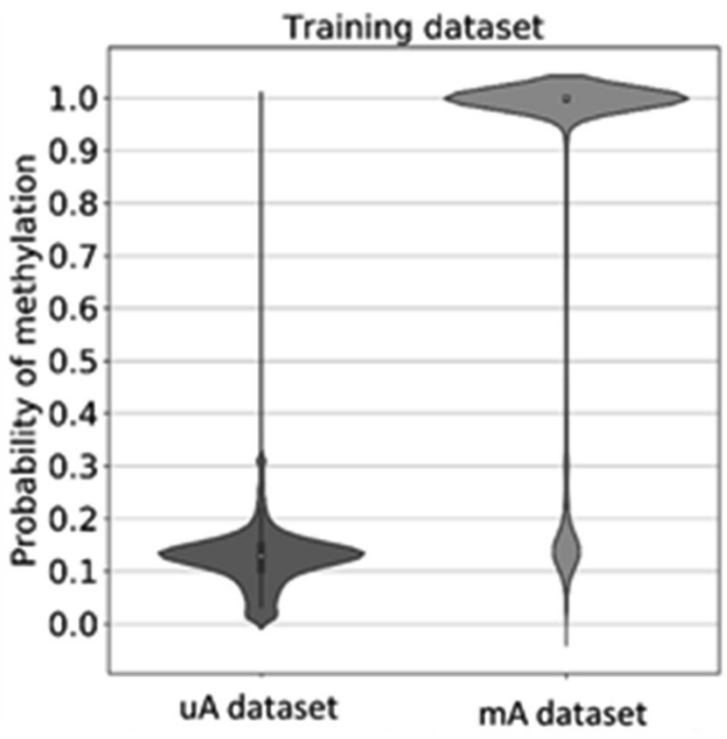


FIG. 36A

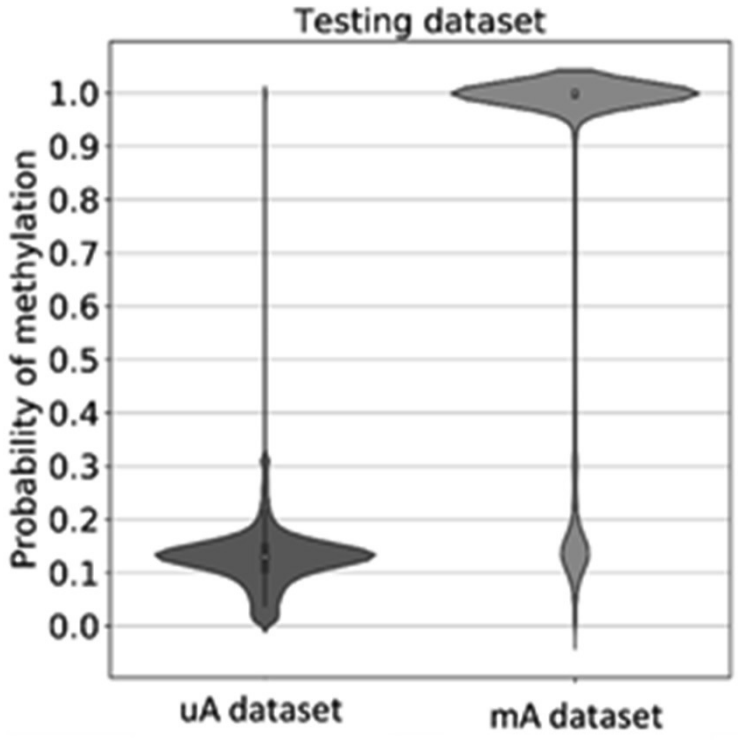


FIG. 36B

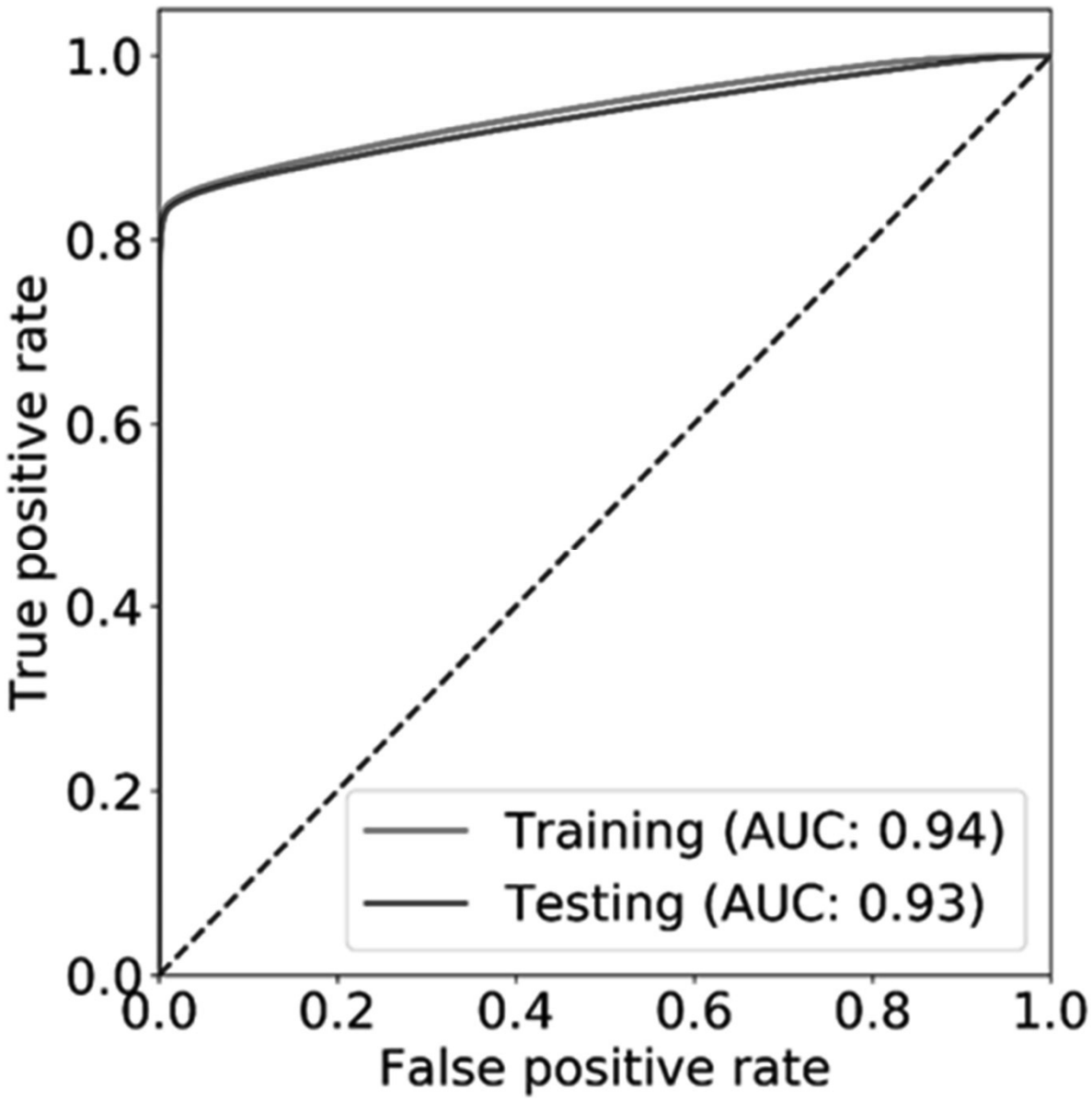

1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Training (AUC: 0.94)
Testing (AUC: 0.93)

FIG. 37

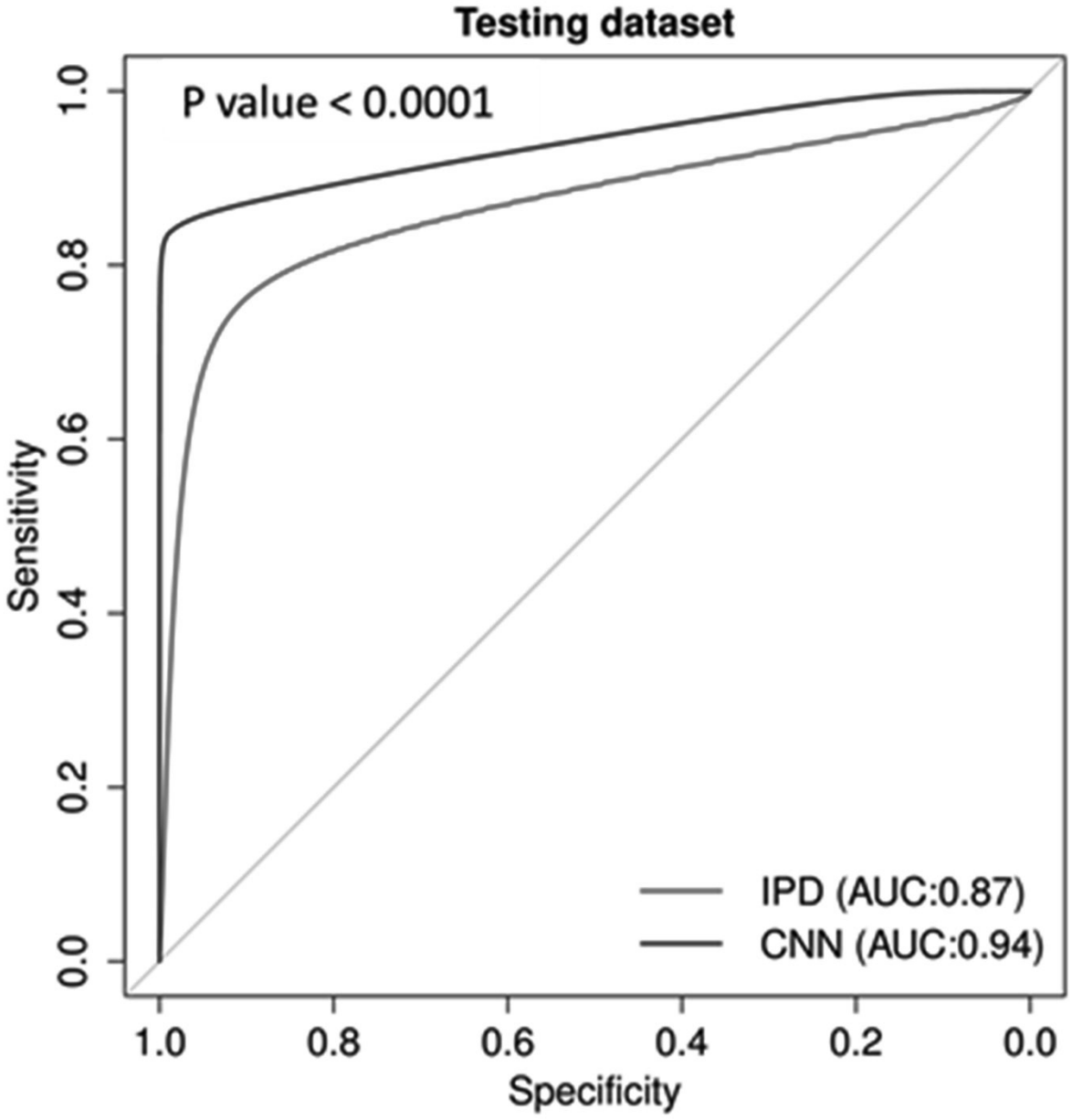

Testing dataset
P value < 0.0001
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
IPD (AUC:0.87)
CNN (AUC:0.94)

FIG. 38

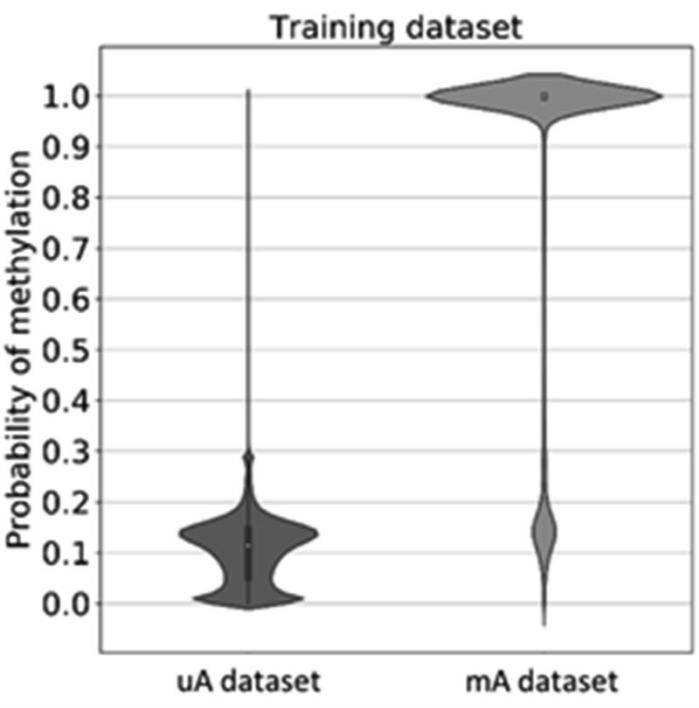


FIG. 39A

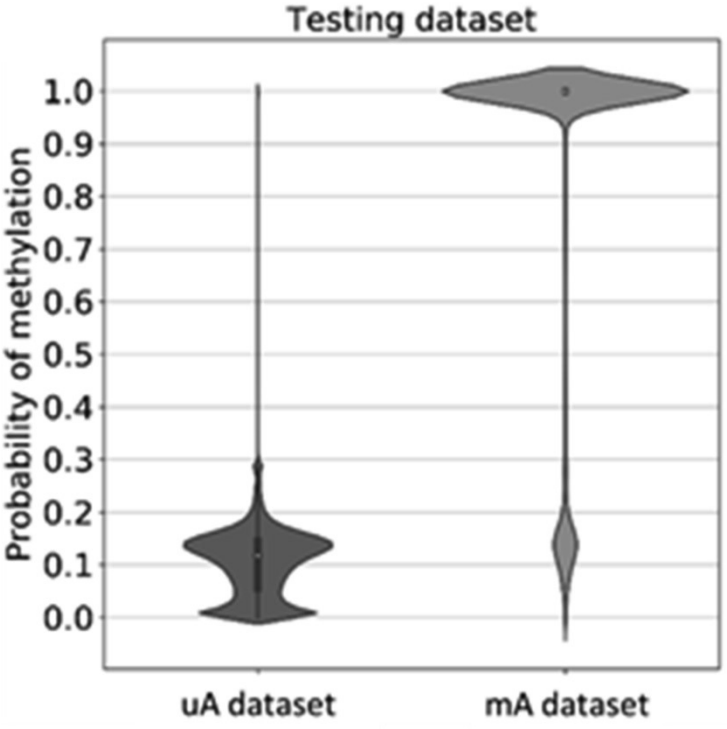


FIG. 39B

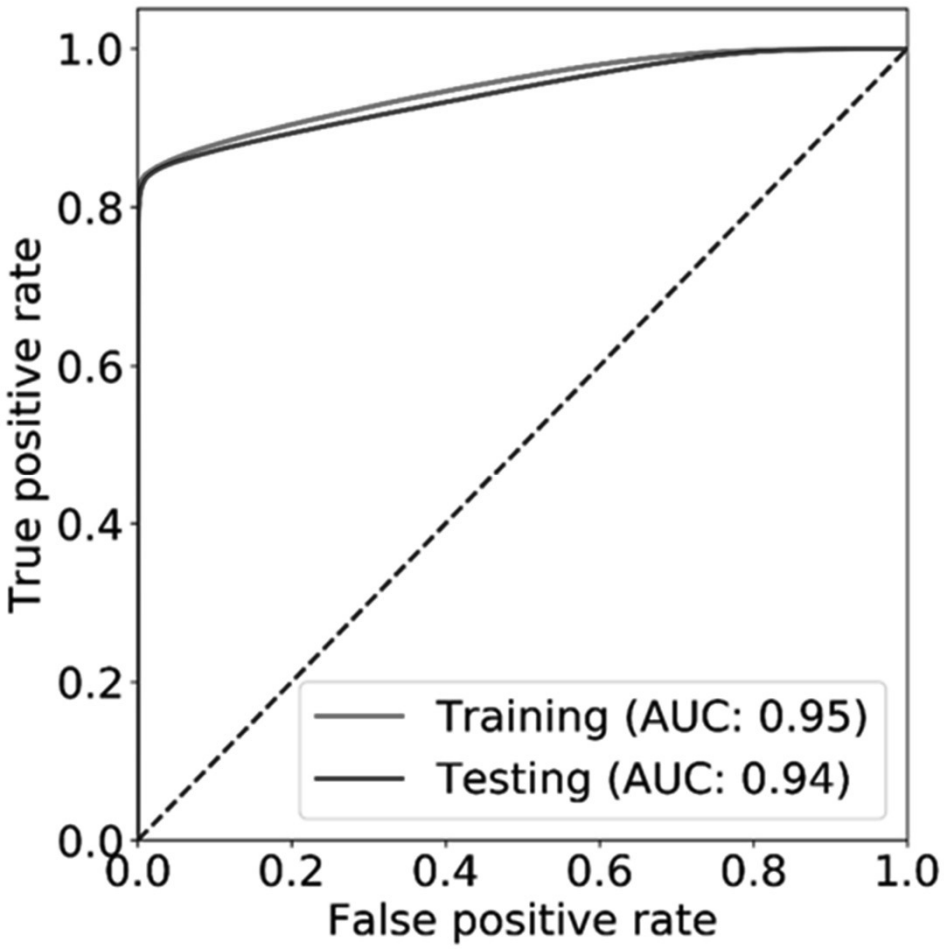

1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Training (AUC: 0.95)
Testing (AUC: 0.94)

FIG. 40

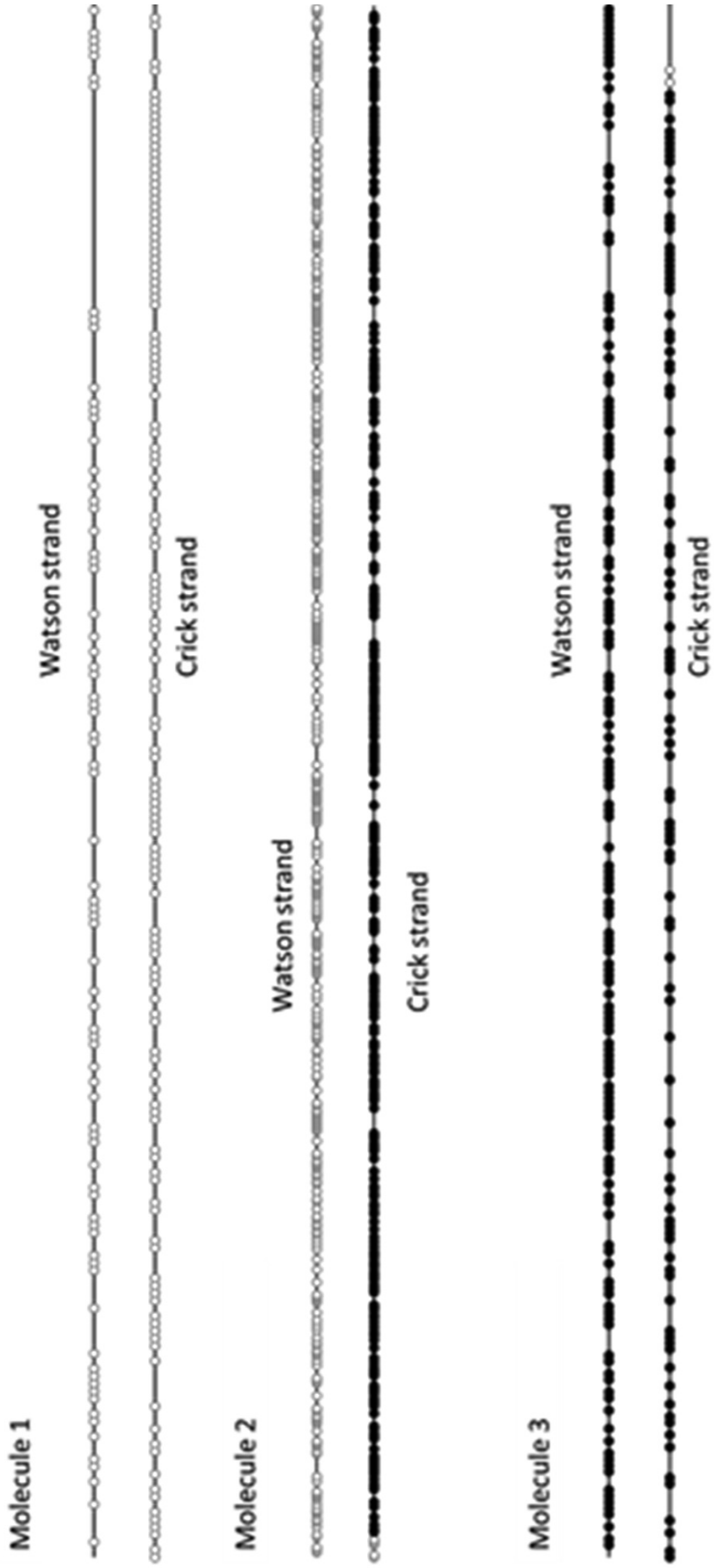

Molecule 1
Watson strand
Crick strand
Molecule 2
Watson strand
Crick strand
Molecule 3
Watson strand
Crick strand

FIG. 41

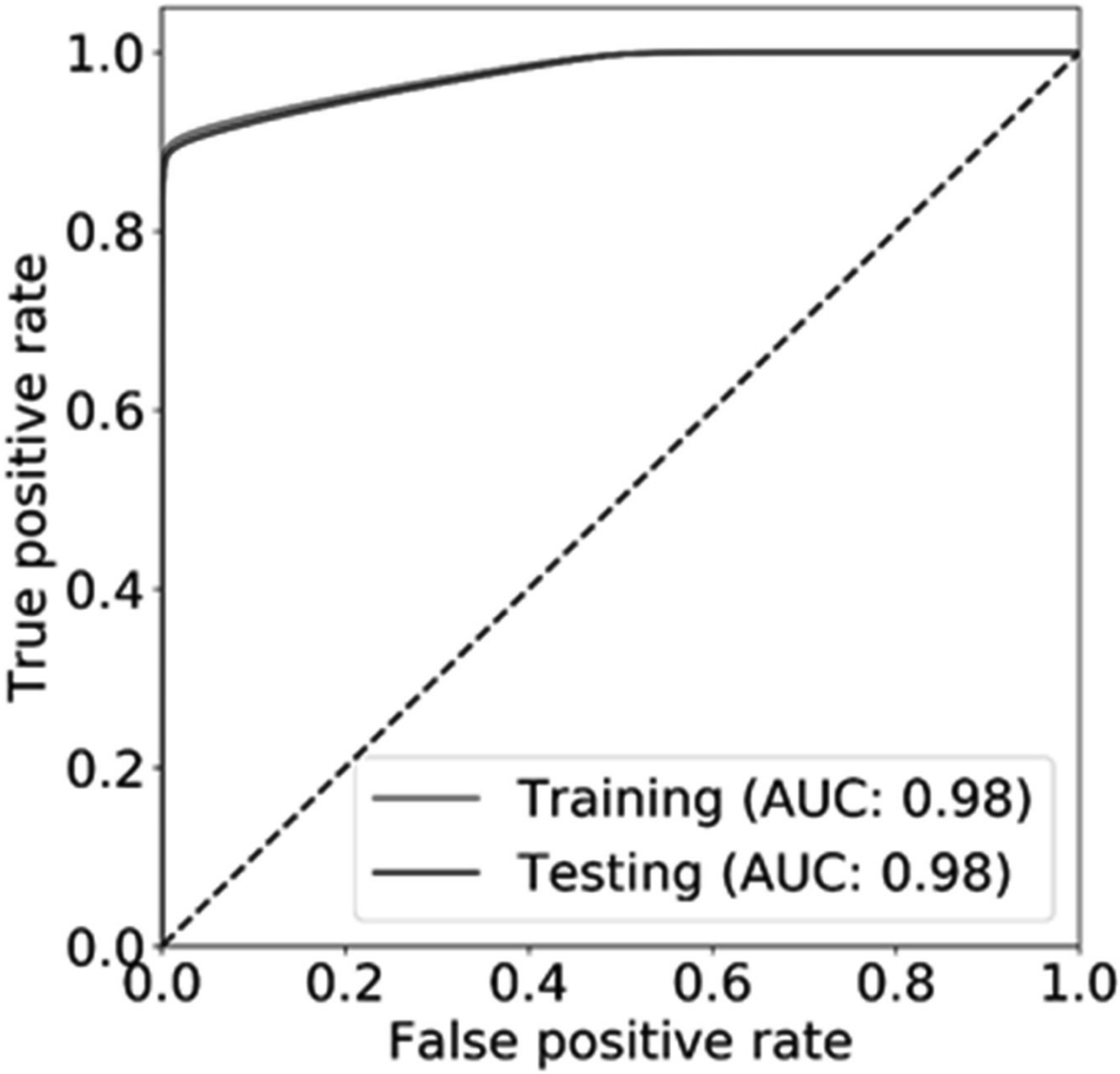

True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Training (AUC: 0.98)
Testing (AUC: 0.98)

FIG. 42

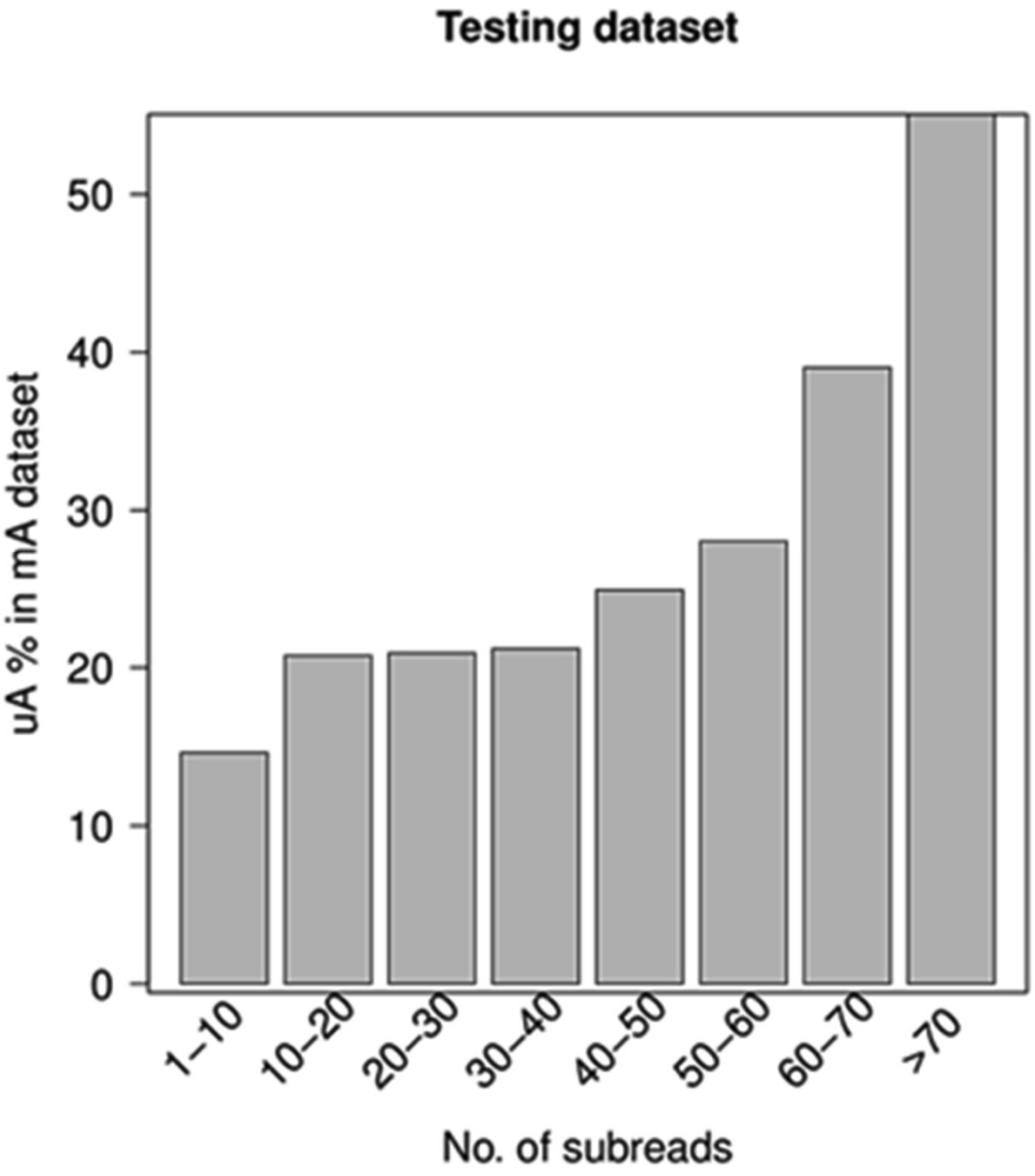

Testing dataset
uA % in mA dataset
0
10
20
30
40
50
1–10
10–20
20–30
30–40
40–50
50–60
60–70
>70
No. of subreads

FIG. 43

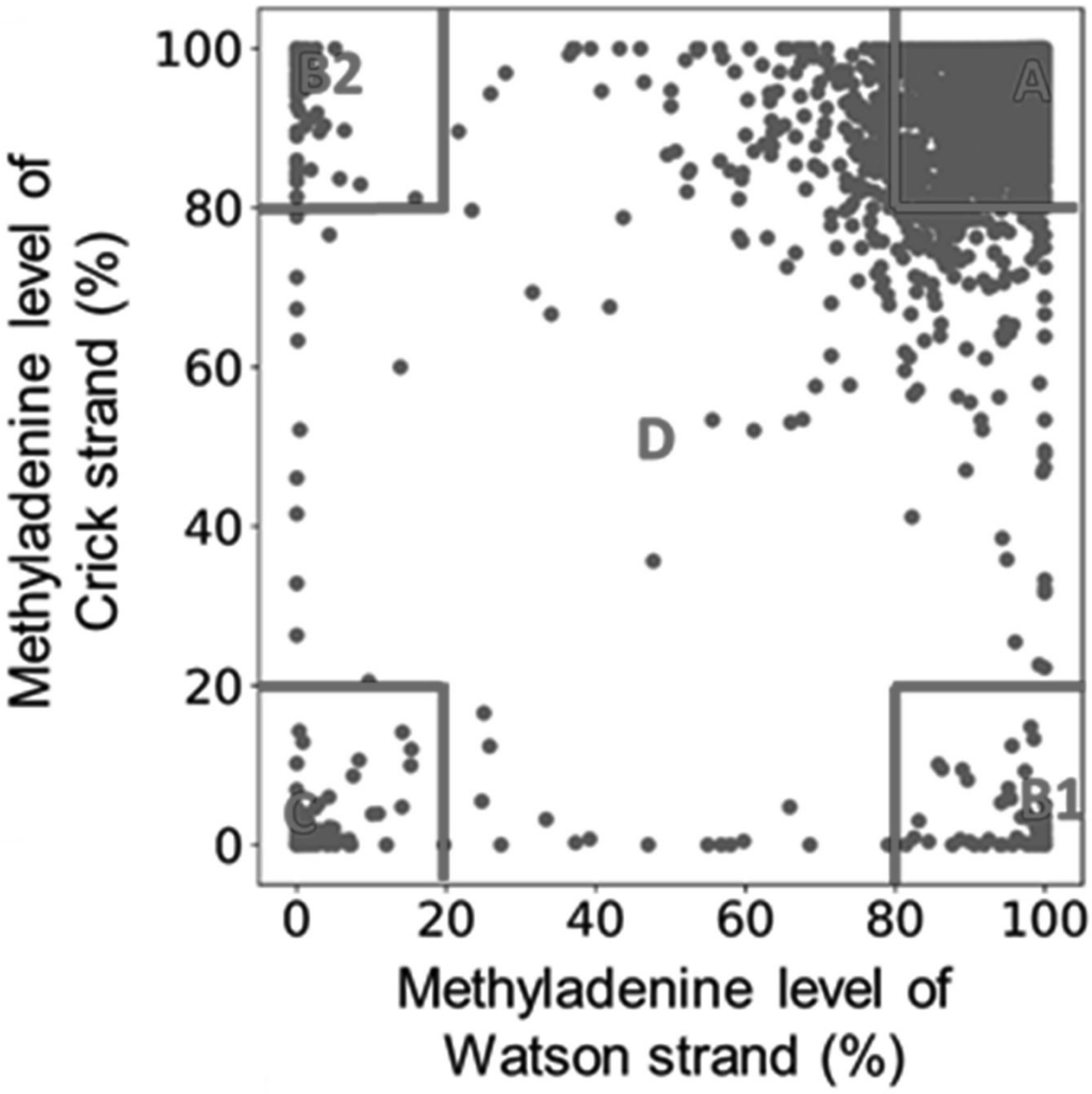

Methyladenine level of Crick strand (%)
Methyladenine level of Watson strand (%)
100
80
60
40
20
0
0
20
40
60
80
100
A
B1
B2
C
D

FIG.44

Categories	Training dataset	Testing dataset
Fully-unmethylated	283 (7.0%)	276 (7.0%)
Hemi-methylated	401 (10.0%)	389 (9.8%)
Fully-methylated	3194 (79.4%)	3142 (79.4%)
Interlacing methylation patterns	145 (3.6%)	148 (3.7%)

FIG. 45

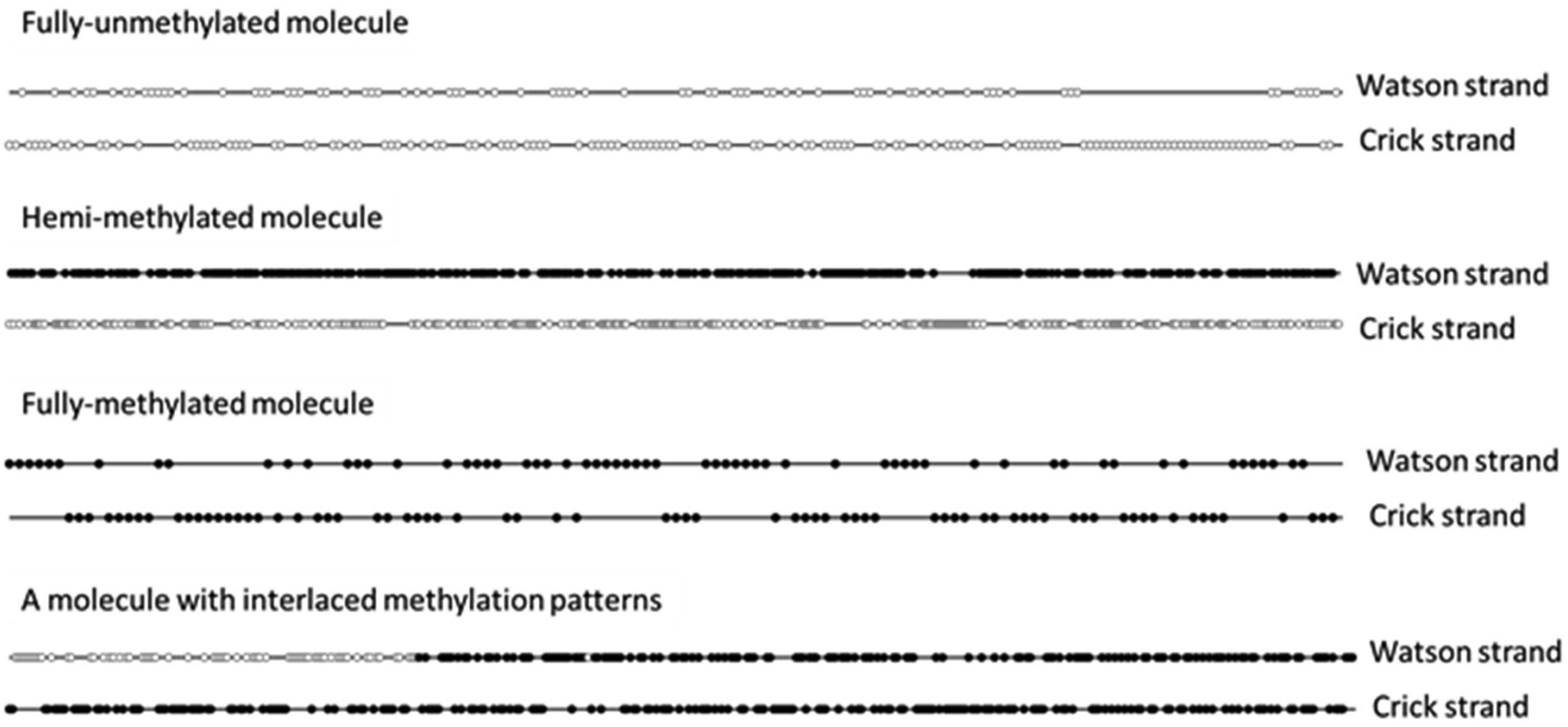

Fully-unmethylated molecule
Watson strand
Crick strand
Hemi-methylated molecule
Watson strand
Crick strand
Fully-methylated molecule
Watson strand
Crick strand
A molecule with interlaced methylation patterns
Watson strand
Crick strand

FIG. 46

ZMW hole number: m54276_180626_162240/40763503
Mapped location: chr1:113246546-113252811
Size: 6265

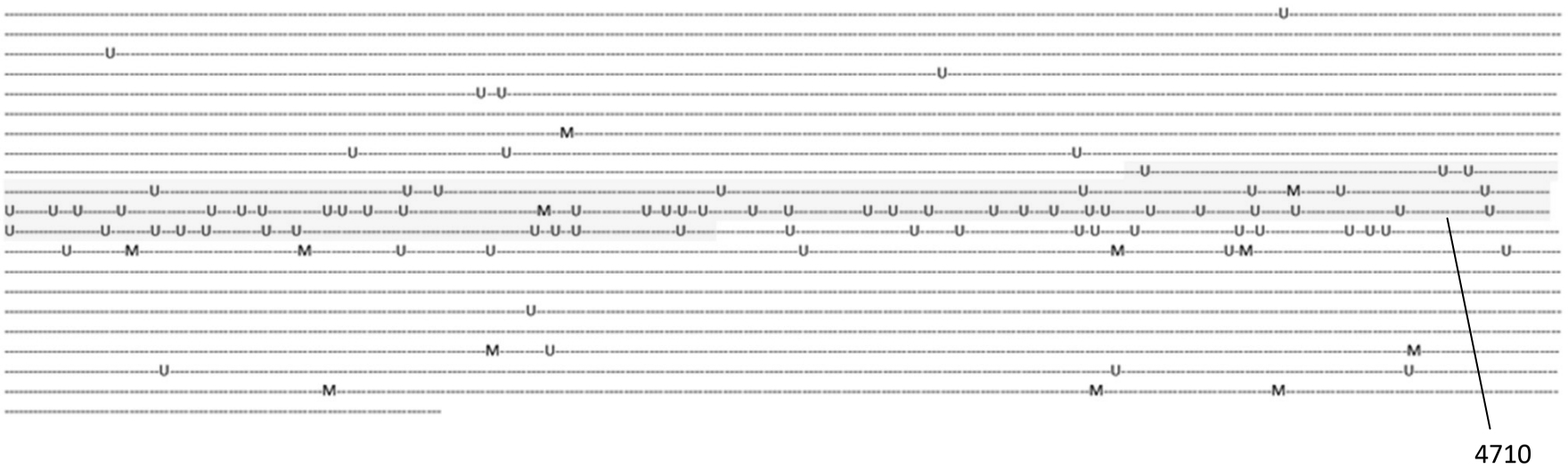


FIG. 47

Chromosomes	Start	End	Imprinted gene names	Length of CpG islands	Molecules sequenced by PacBio SMRT sequencing and methylation states determined according to embodiments present in this disclosure	Methylation call for a molecule
chr11	2013333	2013617	H19	284	-U----------M----------M-M---------U----U----------M----------M----- --U----------M----------M----------M----- M----------M-M------U---------[T]----------M----------M----- --M---	Methylated
chr11	2019565	2019863	H19	298	-M-M----------M----------M----------[C]-----M-M------M---------- ----------M----------M---------M-M----------M---------M---------- M----M-M--M----------M------M----------M-M----------M----- ----------	Methylated
chr11	32460586	32461004	WT1-AS/WT1	418	-U--U----------U-M-------[C]----------U----------U---U--------- U----------U---U-U----------U----M---M---------U---------- -----U-U-U------U------U----------U-----U----------U-U----------U-- ----M	Unmethylated
chr14	101192851	101193499	DLK1	648	-U-------U----------U----------U----------M---M----------U----------U--------- -U--------U----------U-U--U---U-U----------U----------U- --U-U----------U----------U-----U-U---U----------U--U--U-----U----- M----	Unmethylated
chr14	101201559	101201763	DLK1	204	-M--M----------U----------M-------M----M----------[T]--------- ---------M---------M---M----------M----M----------M--------- ----------M----------M---------M-	Methylated
chr14	101292863	101293101	MEG3	238	----------M----------U----------U----------M- ----M----------M-M---------M-	Methylated
chr15	25981176	25981392	ATP10A	216	-----*---------M----------M----------M----------M---------- M--M----------M----------[T]-----M-M----------U-	Methylated
chr2	80531367	80531719	LRRTM1	352	---*----[G]----------U---U-------M-------M----------M----------U----U-- ----U------U----------U----------U---------M----------U---- --U-U----------M----------U----------U--U--U----------U-U--- --	Unmethylated
chr7	79082174	79082427	MAGI2	253	---------*-U----------[A]-M-U----------U----U---U--------U--- U----------U----------*--------	Unmethylated

FIG. 48

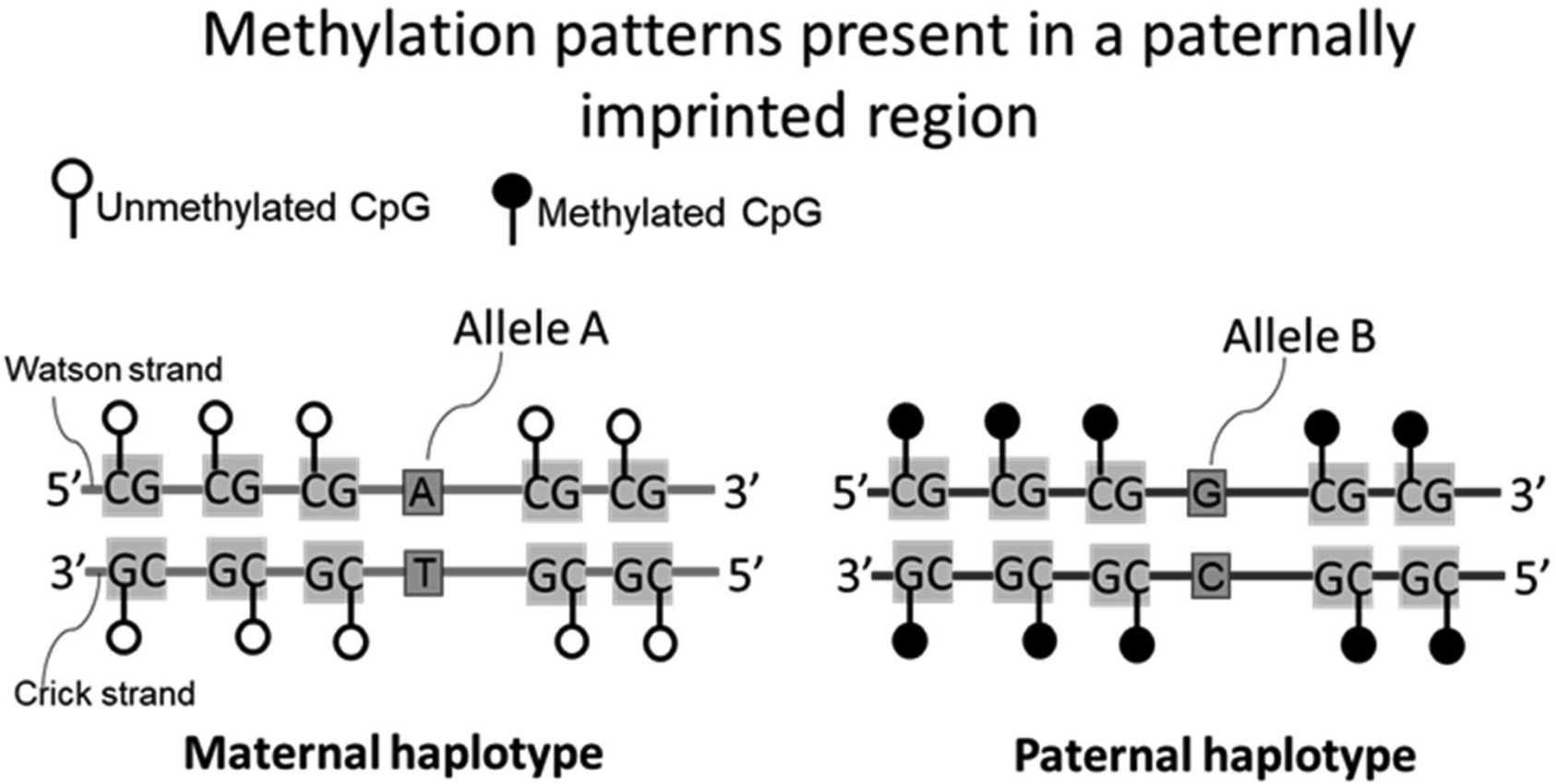

Methylation patterns present in a paternally imprinted region
Unmethylated CpG
Methylated CpG
Allele A
Watson strand
5′ CG CG CG A CG CG 3′
3′ GC GC GC T GC GC 5′
Crick strand
Maternal haplotype
Allele B
5′ CG CG CG G CG CG 3′
3′ GC GC GC C GC GC 5′
Paternal haplotype

FIG. 49

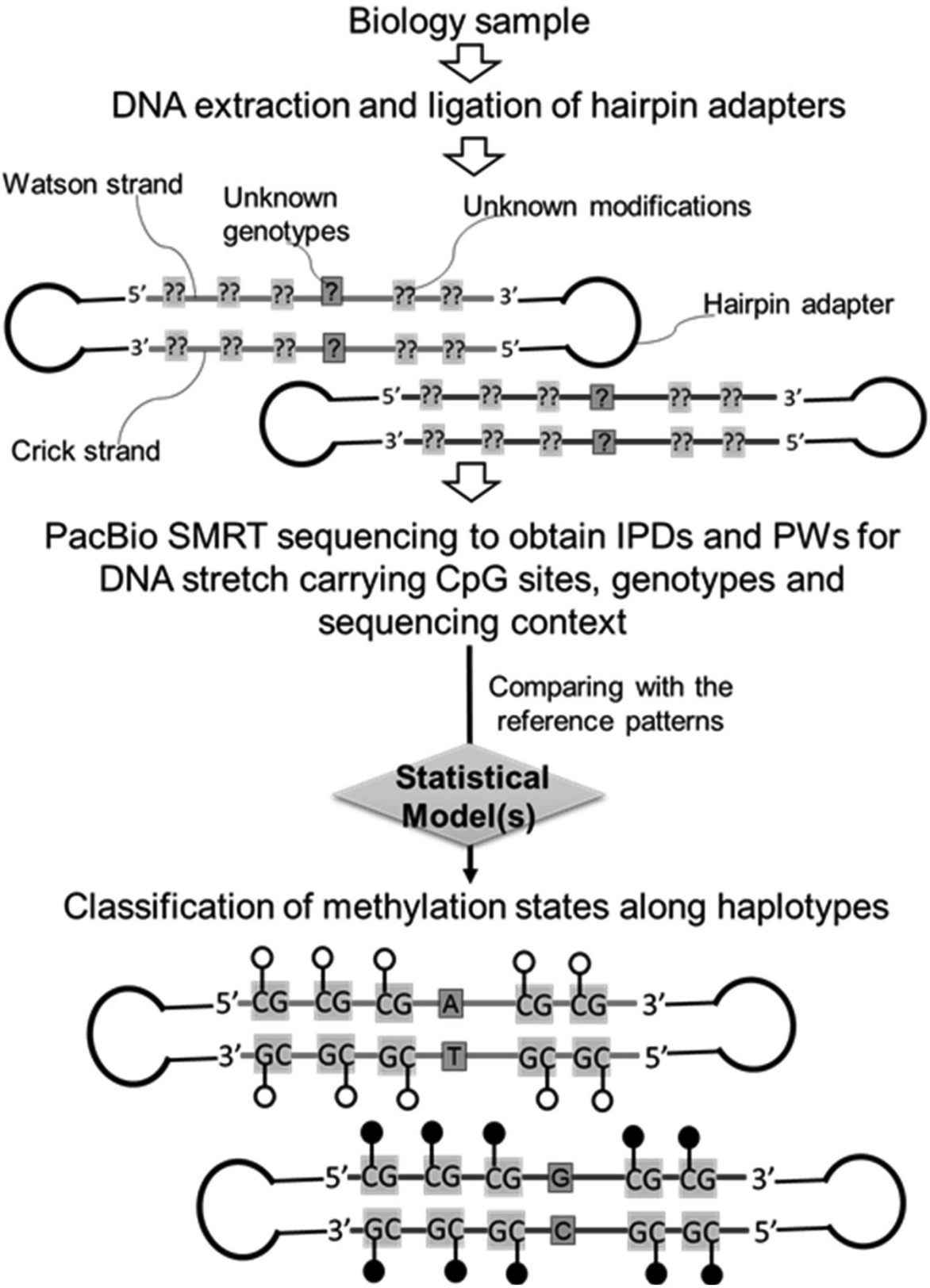

Biology sample
DNA extraction and ligation of hairpin adapters
Watson strand
Unknown genotypes
Unknown modifications
Hairpin adapter
Crick strand
PacBio SMRT sequencing to obtain IPDs and PWs for DNA stretch carrying CpG sites, genotypes and sequencing context
Comparing with the reference patterns
Statistical Model(s)
Classification of methylation states along haplotypes

FIG. 50

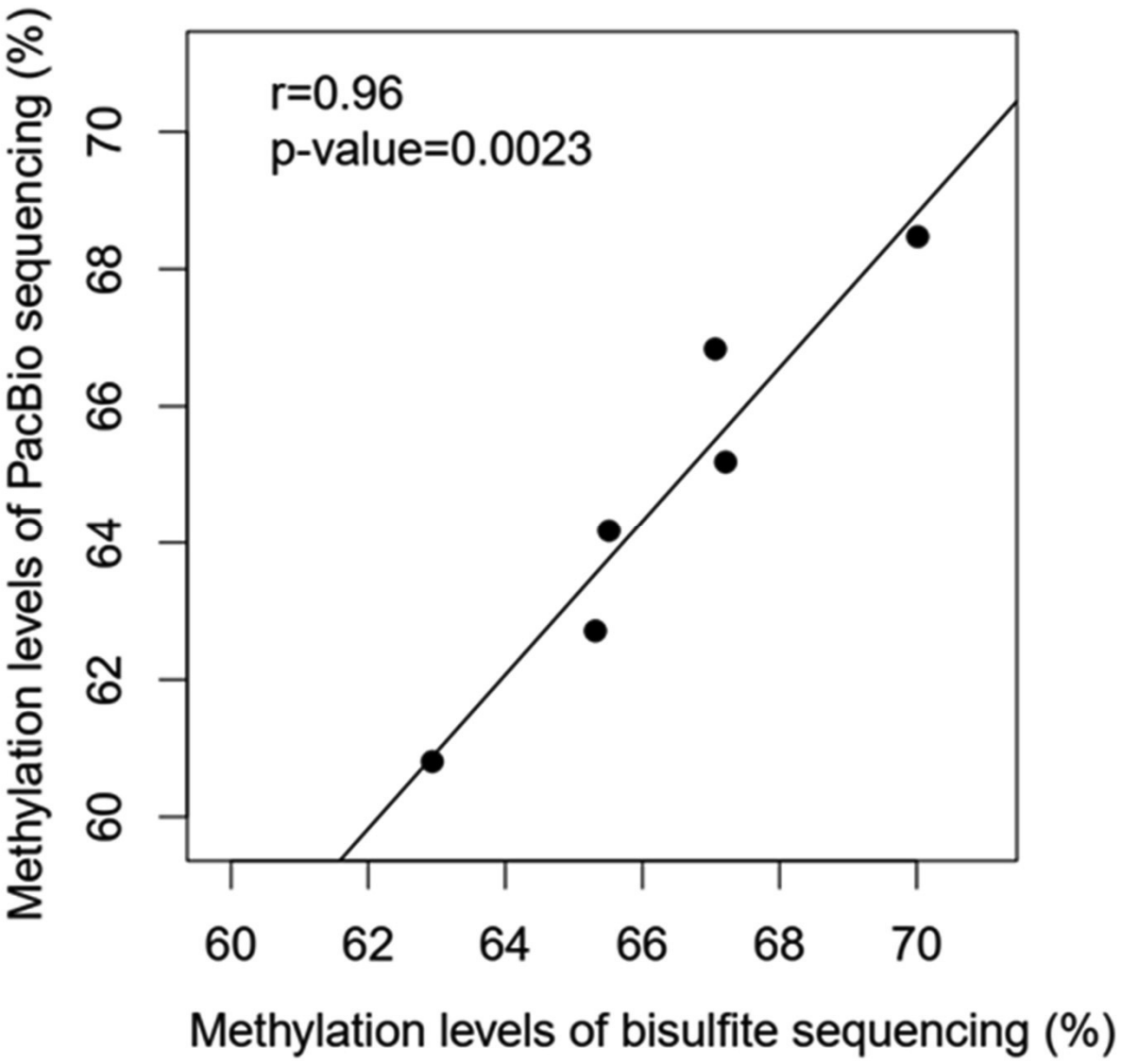

r=0.96
p-value=0.0023
Methylation levels of PacBio sequencing (%)
60
62
64
66
68
70
Methylation levels of bisulfite sequencing (%)

FIG. 51

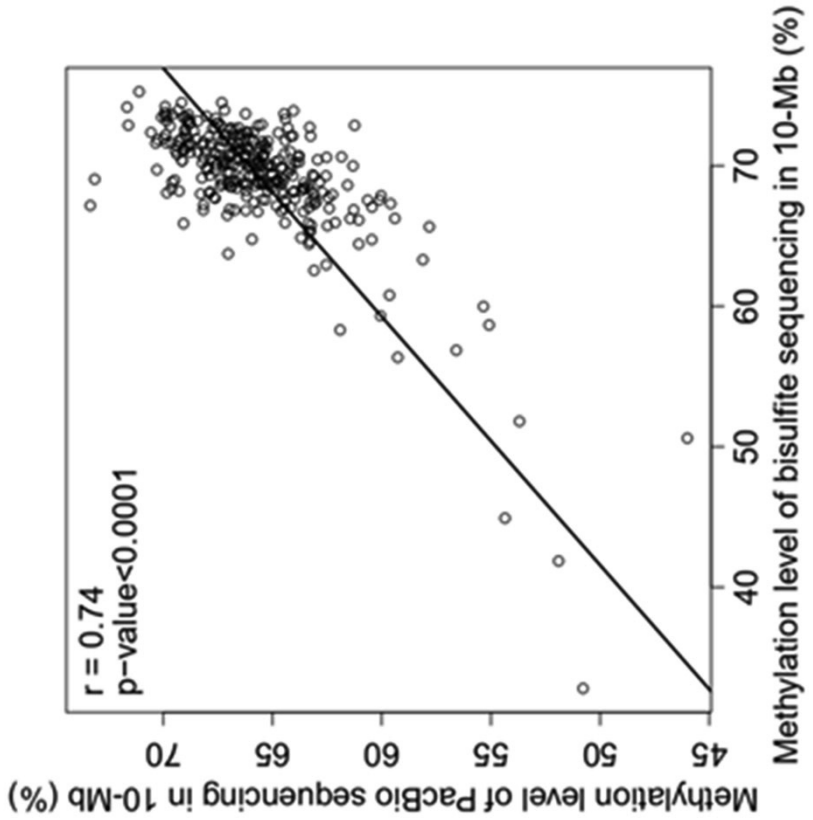


FIG. 52B

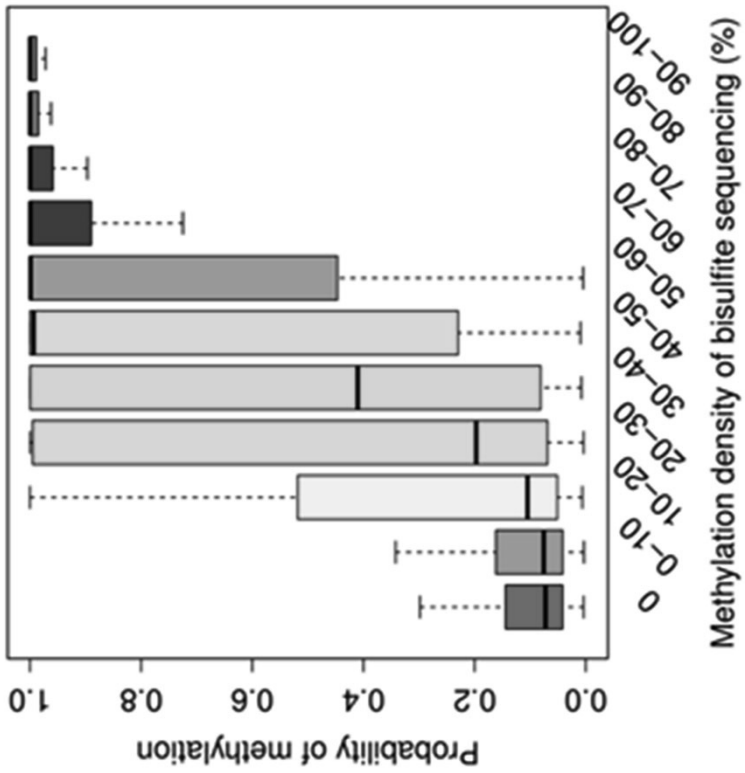


FIG. 52A

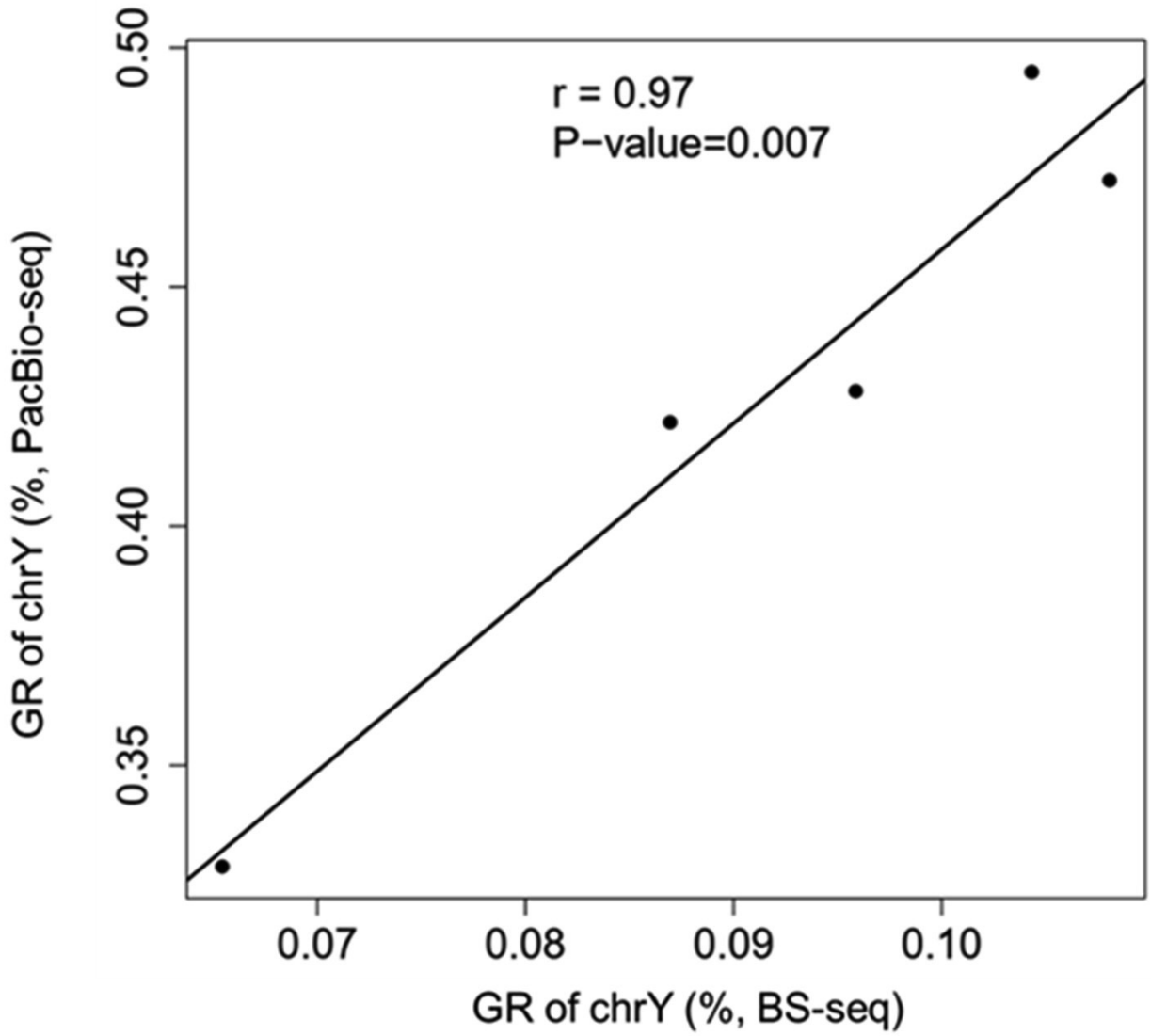

r = 0.97
P−value=0.007
GR of chrY (%, PacBio-seq)
0.35
0.40
0.45
0.50
0.07
0.08
0.09
0.10
GR of chrY (%, BS-seq)

FIG. 53

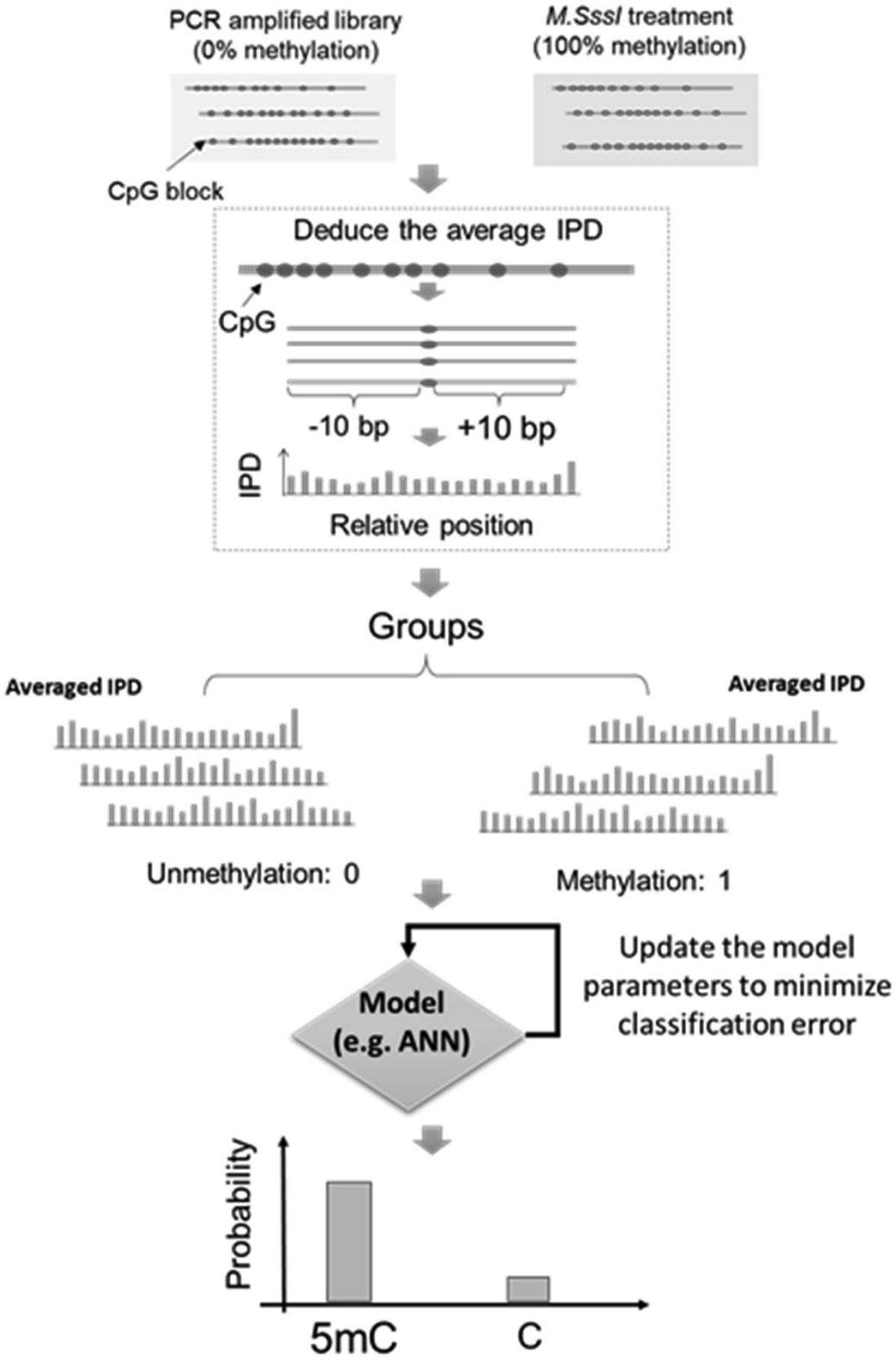

PCR amplified library
(0% methylation)
M.SssI treatment
(100% methylation)
CpG block
Deduce the average IPD
CpG
-10 bp
+10 bp
IPD
Relative position
Groups
Averaged IPD
Averaged IPD
Unmethylation: 0
Methylation: 1
Model
(e.g. ANN)
Update the model
parameters to minimize
classification error
Probability
5mC
C

FIG. 54

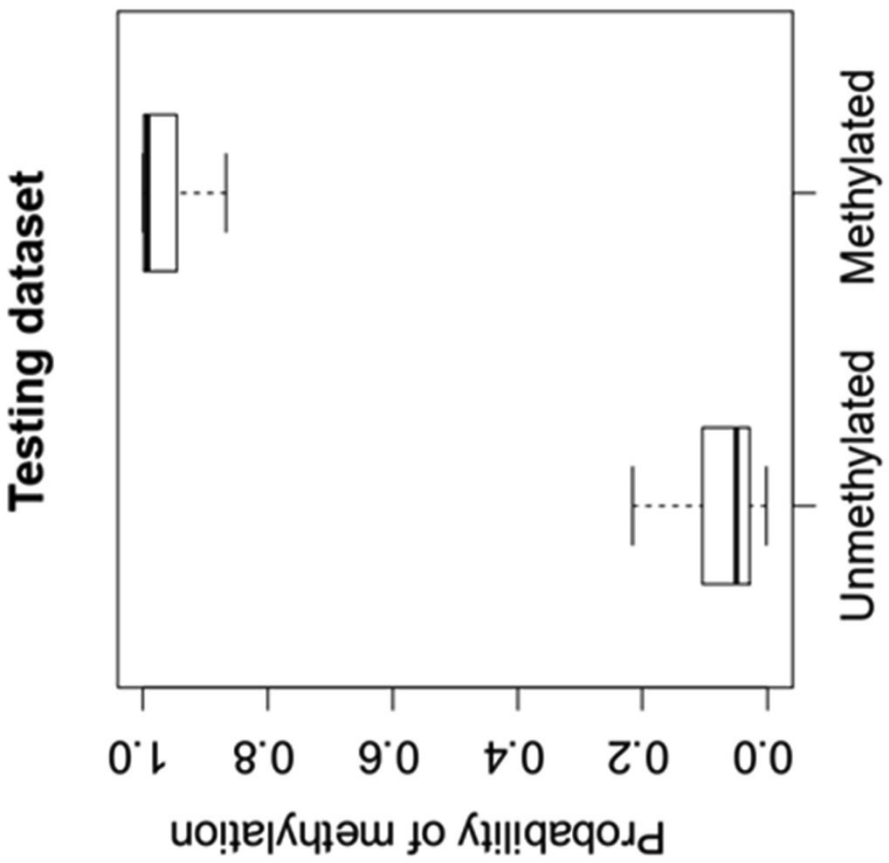


FIG. 55A

FIG. 55B

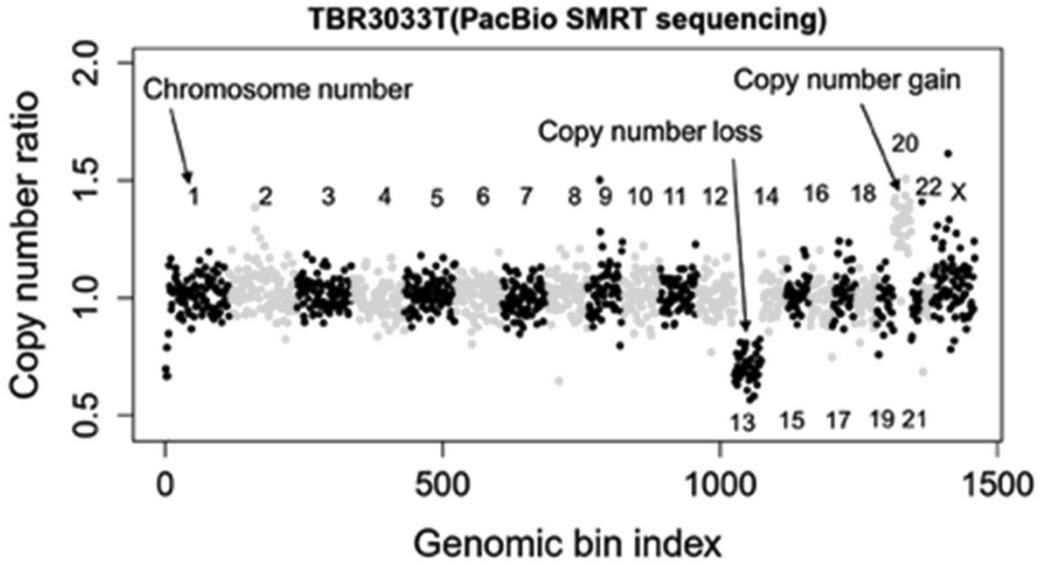


FIG. 56A

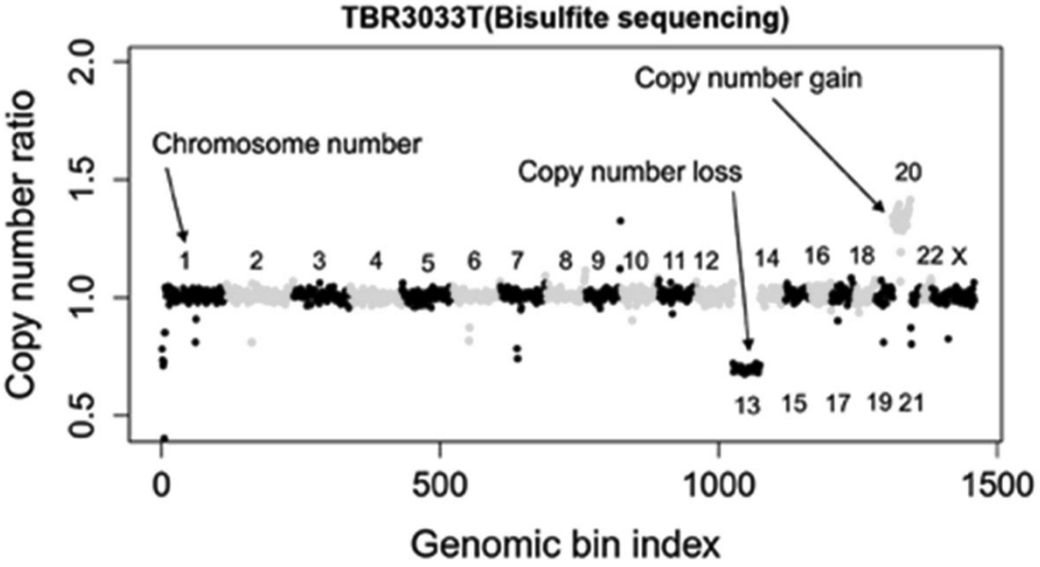


FIG. 56B

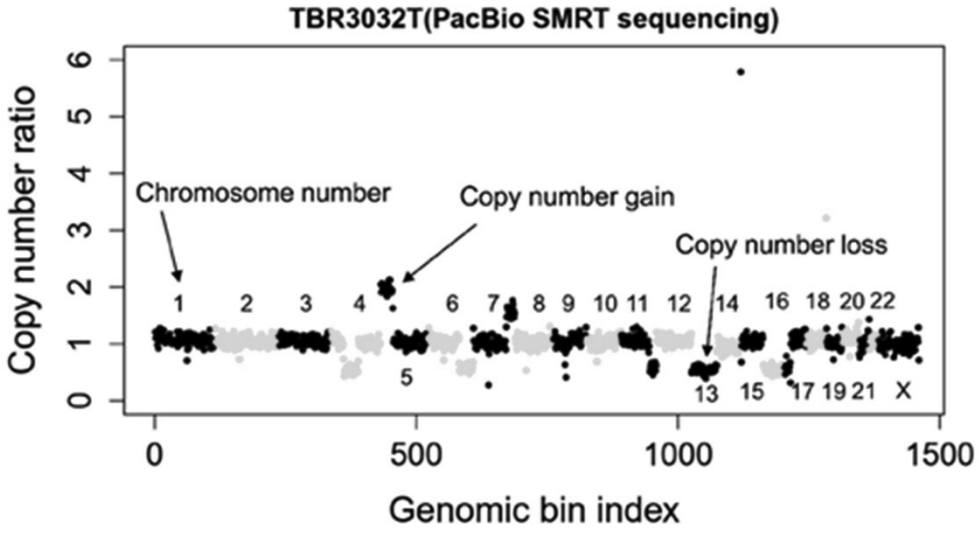


FIG. 57A

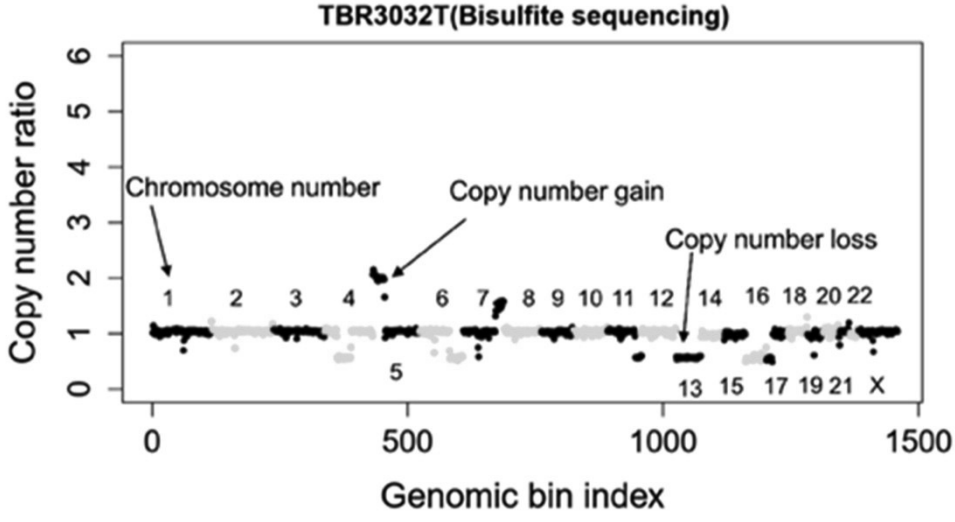


FIG. 57B

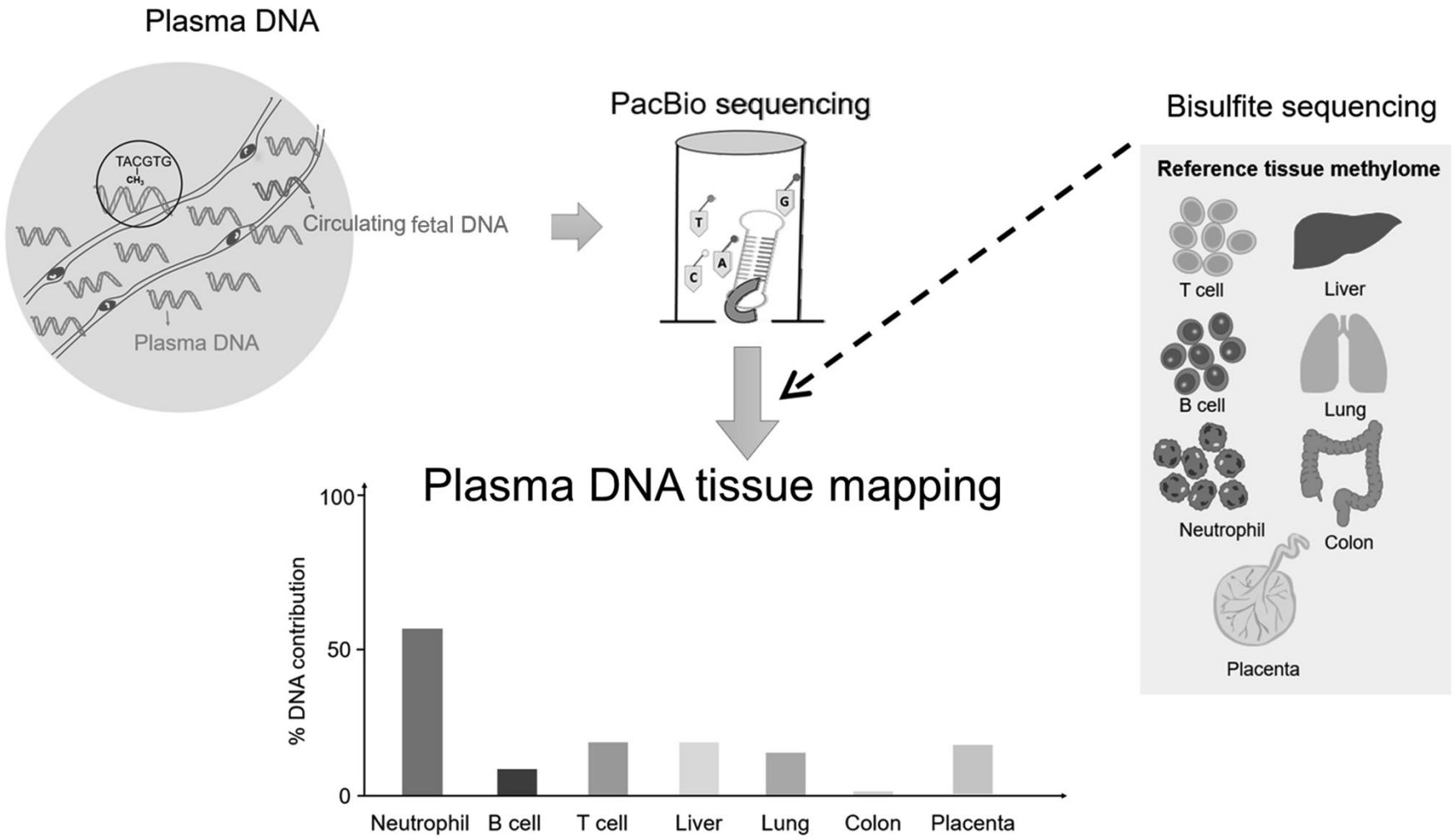

Plasma DNA
TACGTG
CH3
Circulating fetal DNA
Plasma DNA
PacBio sequencing
T
G
A
C
Bisulfite sequencing
Reference tissue methylome
T cell
Liver
B cell
Lung
Neutrophil
Colon
Placenta
Plasma DNA tissue mapping
100
50
0
% DNA contribution
Neutrophil
B cell
T cell
Liver
Lung
Colon
Placenta

FIG. 58

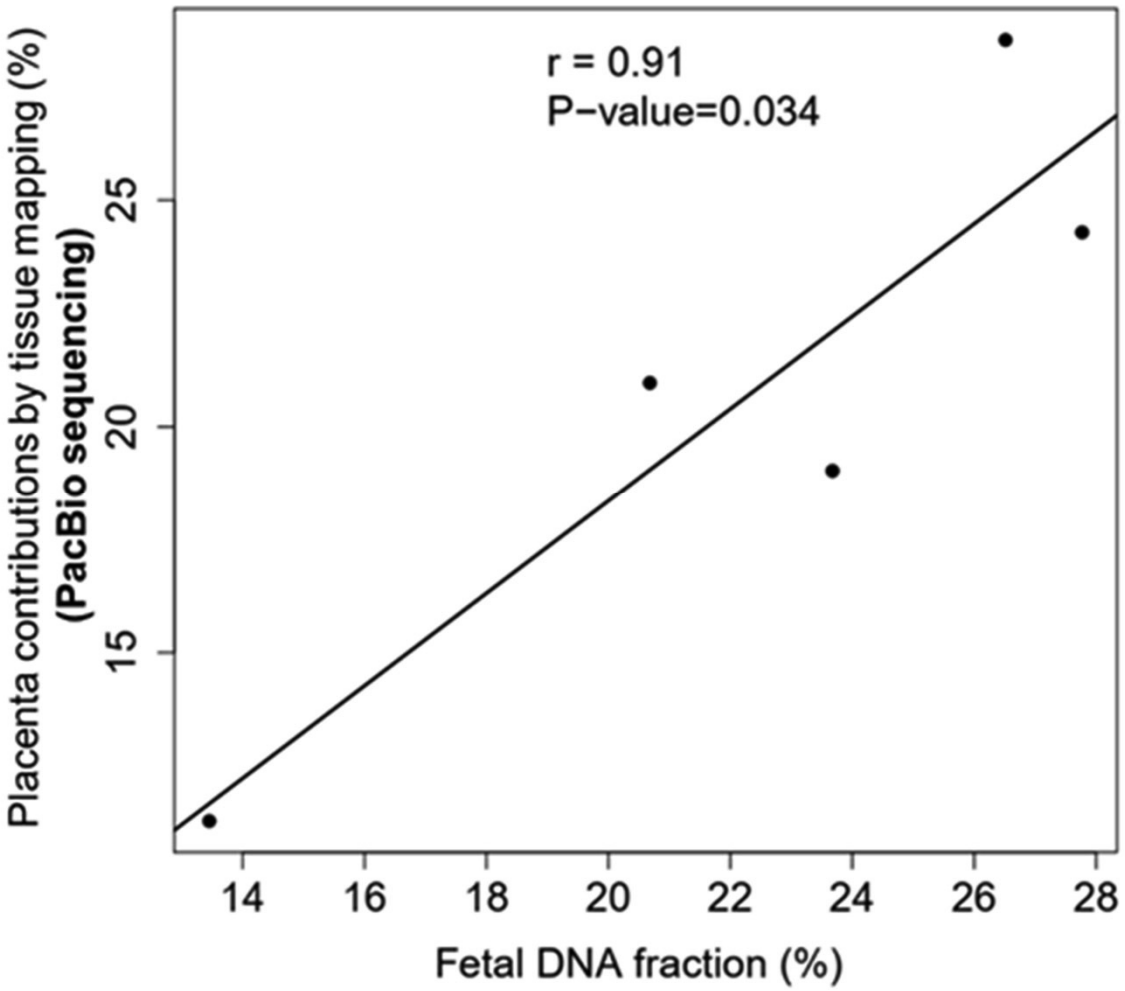

r = 0.91
P−value=0.034
Placenta contributions by tissue mapping (%)
(PacBio sequencing)
15
20
25
14
16
18
20
22
24
26
28
Fetal DNA fraction (%)

FIG. 59

Groups	Samples	Total subreads	Mapped subreads	Subread mappability (%)	Mean subread depth per SMRT well (x)	No. of SMRT wells	Mappable wells	Mappable well rate (%)
Maternal buffy coat	M13153w	39,006,460	30,673,525	78.6	13.4	3,157,310	2,295,002	72.7
Placenta	N13153	23,013,428	16,374,758	71.2	10.4	2,393,400	1,573,540	65.7
HCC tissues	TBR3032T	20,164,513	15,232,744	75.5	13.1	1,742,990	1,147,985	64.8
	TBR3033T	22,639,692	17,479,024	77.2	8.1	2,832,627	2,157,196	76.2
Adjacent normal tissues	TBR3033N	73,118,110	56,446,202	77.2	12.6	6,881,142	4,471,370	65.0
	TBR3032N	76,852,680	60,145,452	78.3	12.8	6,000,227	4,702,130	78.4
Buffy coat (healthy control subjects)	M1	44,777,423	28,325,587	63.3	7.7	7,316,000	3,659,996	50.0
	F2	49,840,758	32,994,645	66.2	8.6	7,215,112	3,823,329	53.0
	F1	40,012,804	24,717,289	61.8	6.5	7,301,768	3,800,392	52.0
	M2	152,530,411	88,596,520	58.1	7.7	21,794,606	11,563,500	53.1
HCC cell line	HepG2	47,308,982	34,581,721	73.1	7.3	6,220,000	4,750,581	76.4

FIG. 60

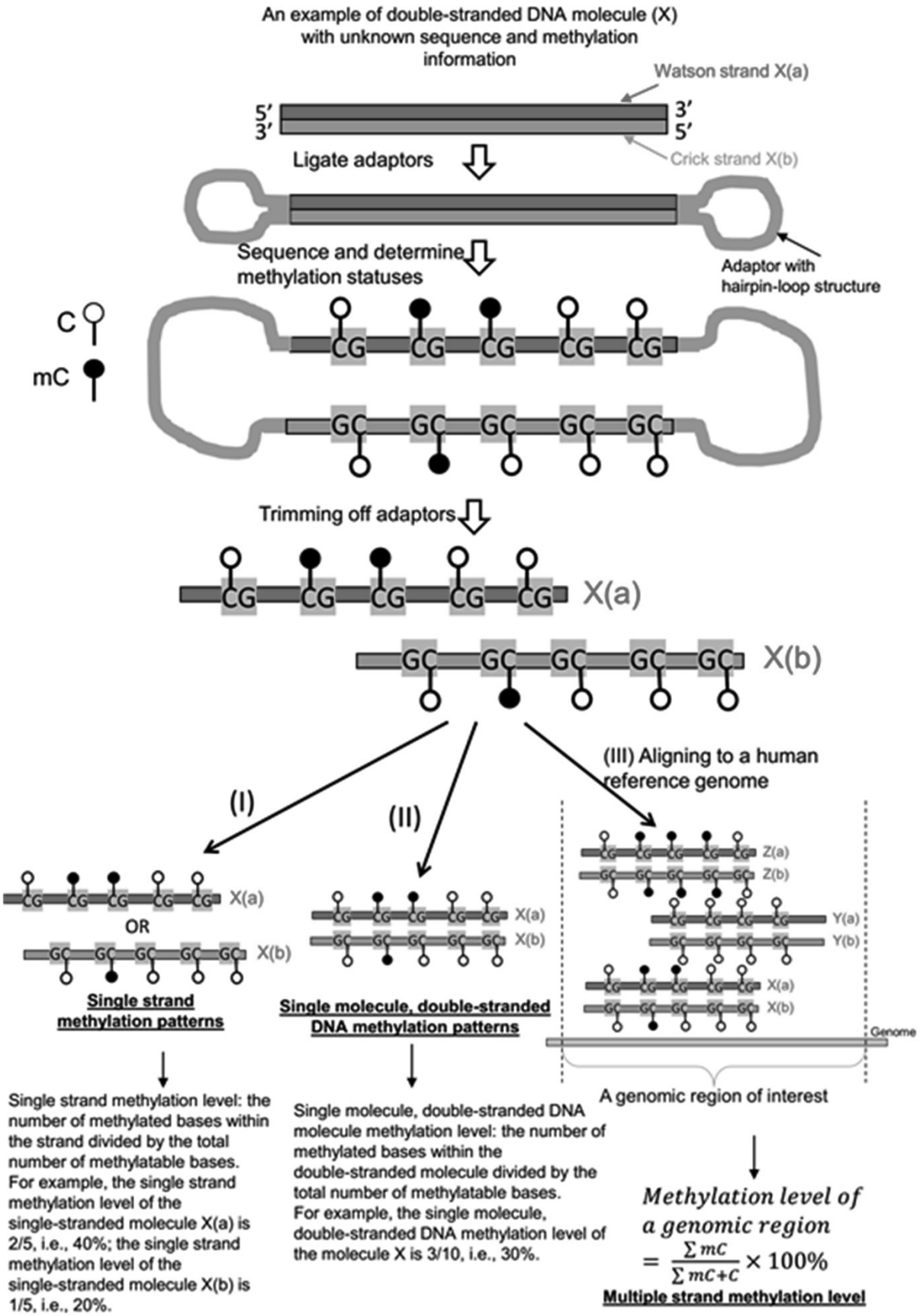

An example of double-stranded DNA molecule (X) with unknown sequence and methylation information
Watson strand X(a)
5′
3′
3′
5′
Crick strand X(b)
Ligate adaptors
Sequence and determine methylation statuses
Adaptor with hairpin-loop structure
C
mC
CG
GC
Trimming off adaptors
X(a)
X(b)
(I)
(II)
(III) Aligning to a human reference genome
OR
Single strand methylation patterns
Single molecule, double-stranded DNA methylation patterns
Z(a)
Z(b)
Y(a)
Y(b)
Genome
A genomic region of interest
Single strand methylation level: the number of methylated bases within the strand divided by the total number of methylatable bases. For example, the single strand methylation level of the single-stranded molecule X(a) is 2/5, i.e., 40%; the single strand methylation level of the single-stranded molecule X(b) is 1/5, i.e., 20%.
Single molecule, double-stranded DNA molecule methylation level: the number of methylated bases within the double-stranded molecule divided by the total number of methylatable bases. For example, the single molecule, double-stranded DNA methylation level of the molecule X is 3/10, i.e., 30%.
Methylation level of a genomic region = ∑mC / ∑mC+C × 100%
Multiple strand methylation level

FIG. 61

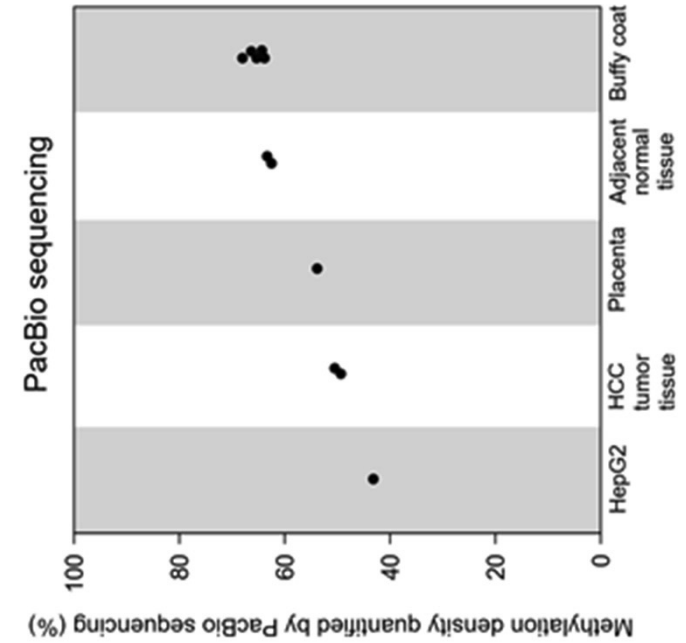


FIG. 62B

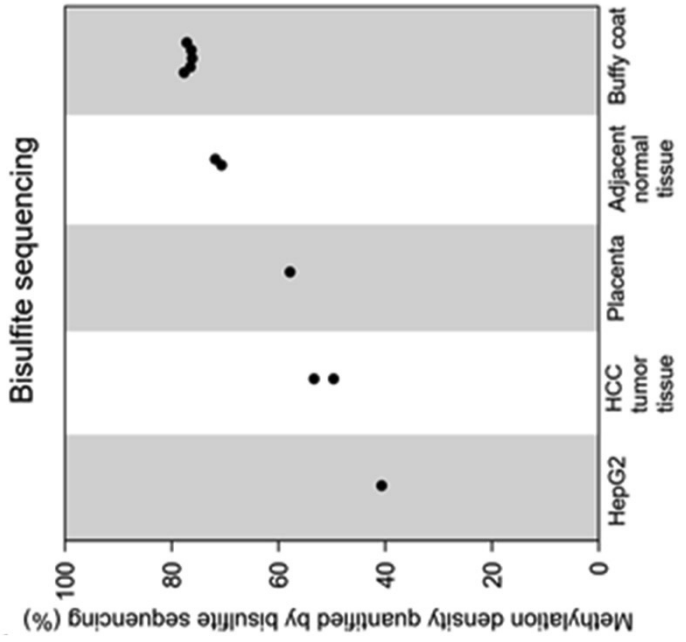


FIG. 62A

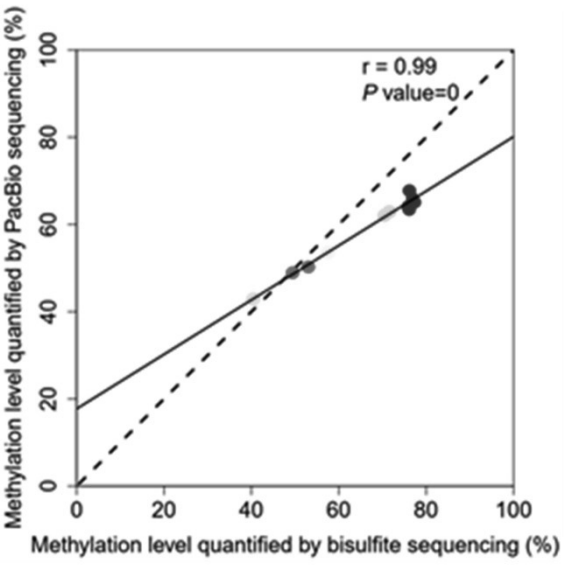


FIG. 63A

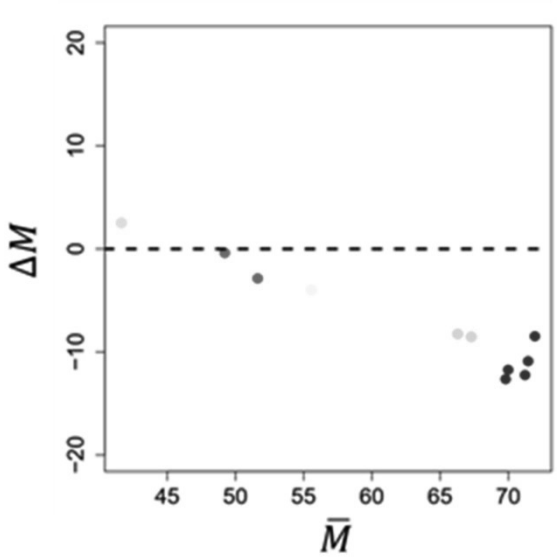


FIG. 63B

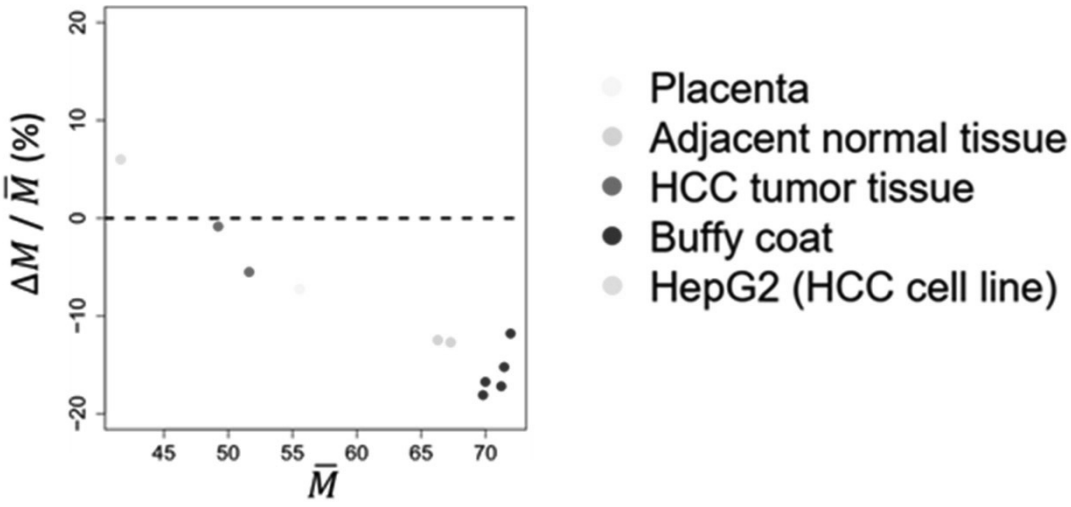


FIG. 63C

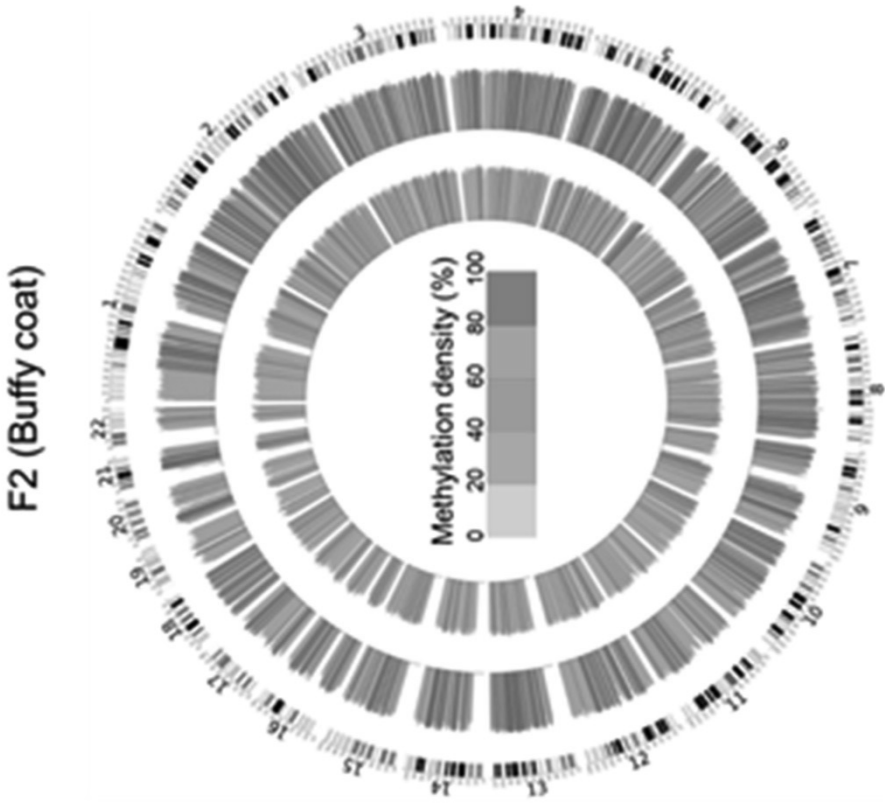


FIG. 64B

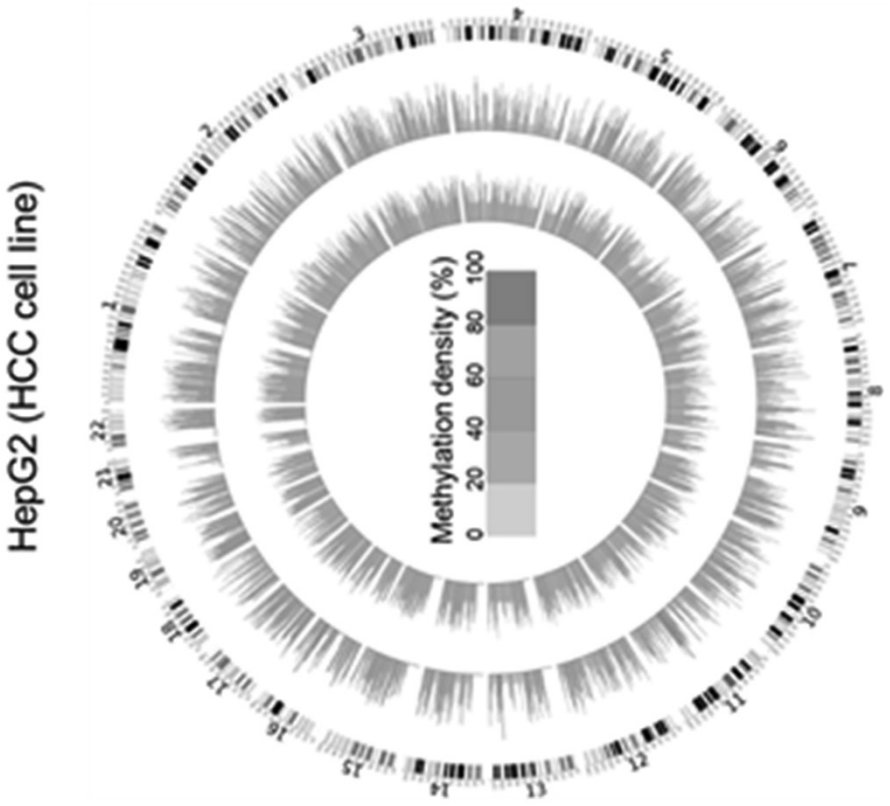


FIG. 64A

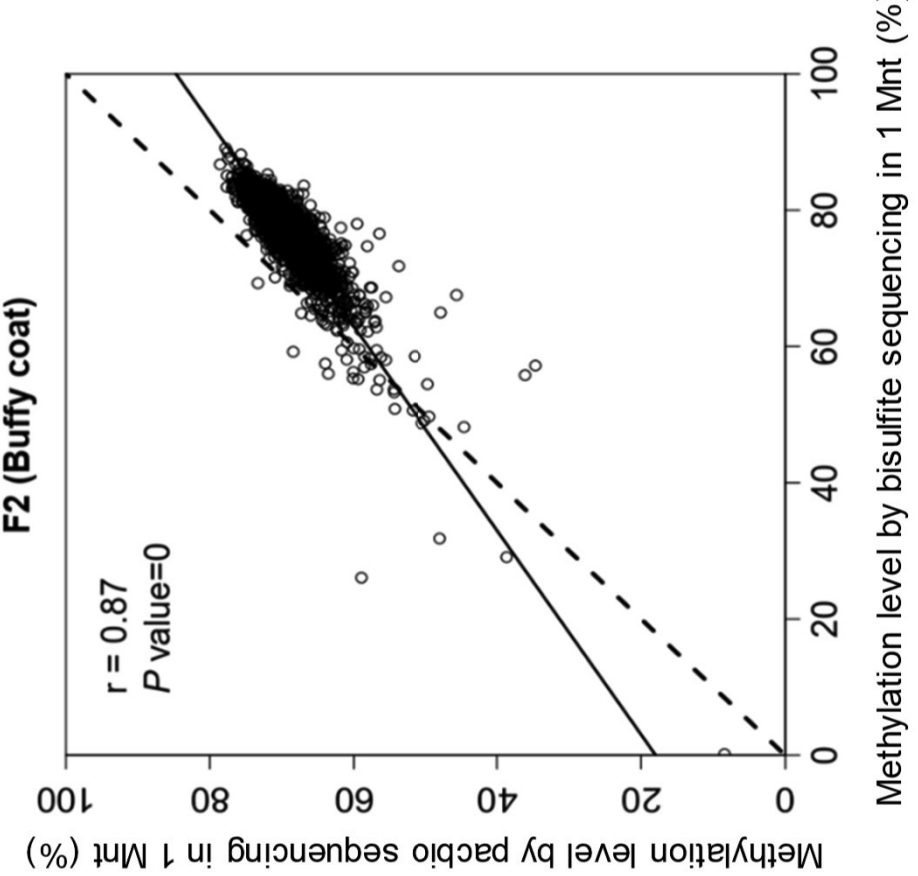


FIG. 65B

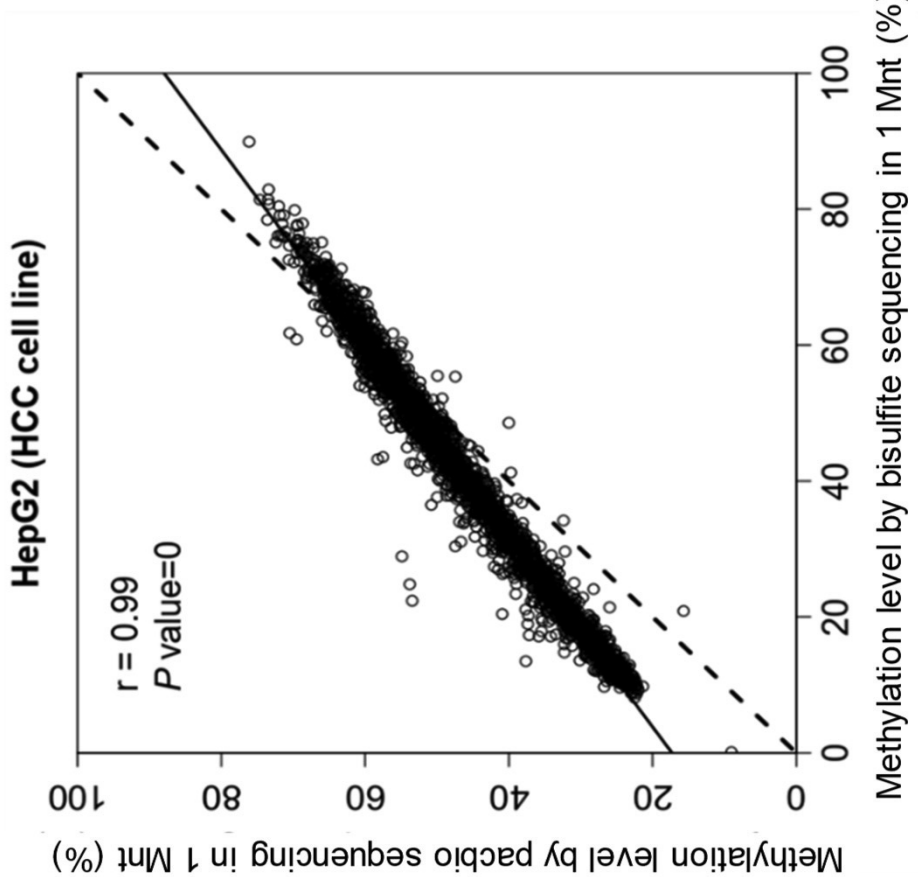


FIG. 65A

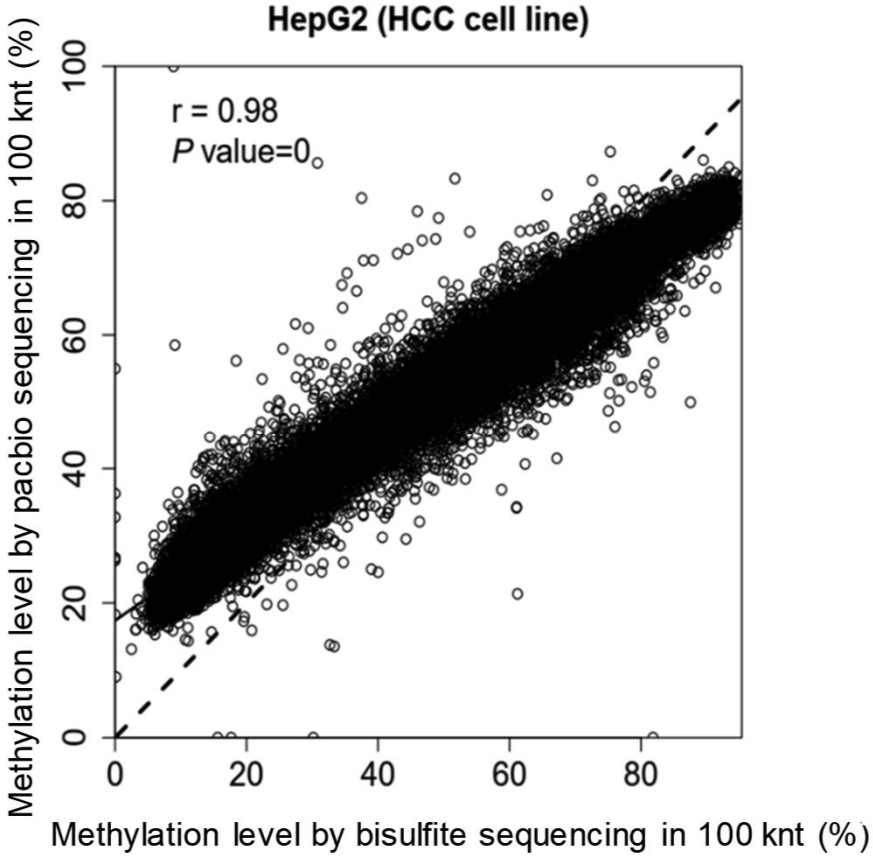


FIG. 66A

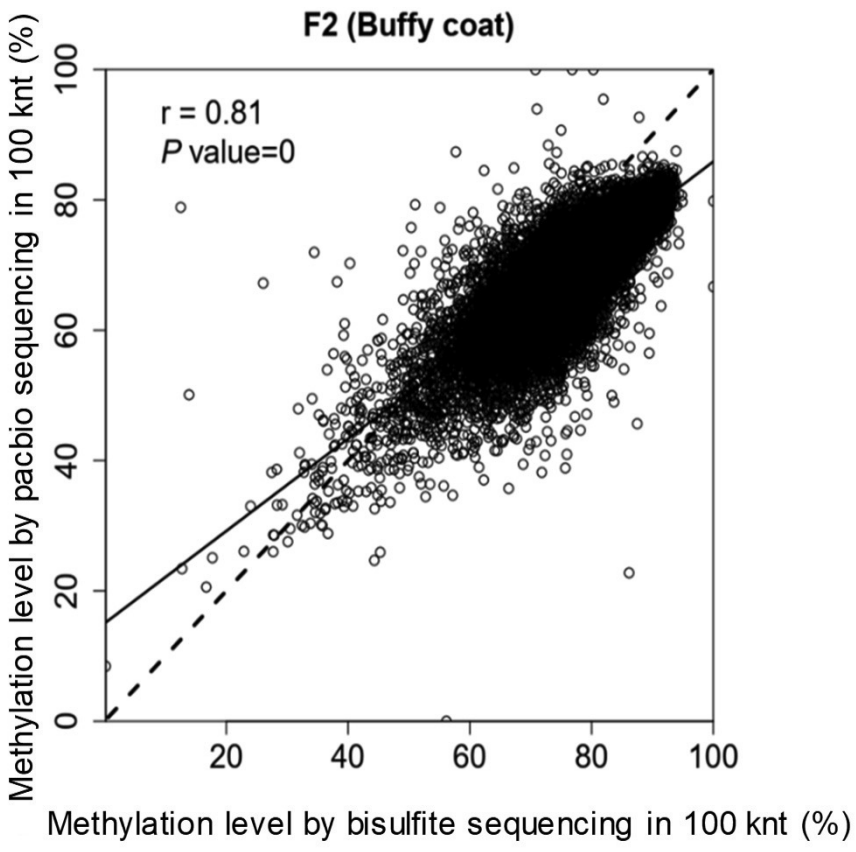


FIG. 66B

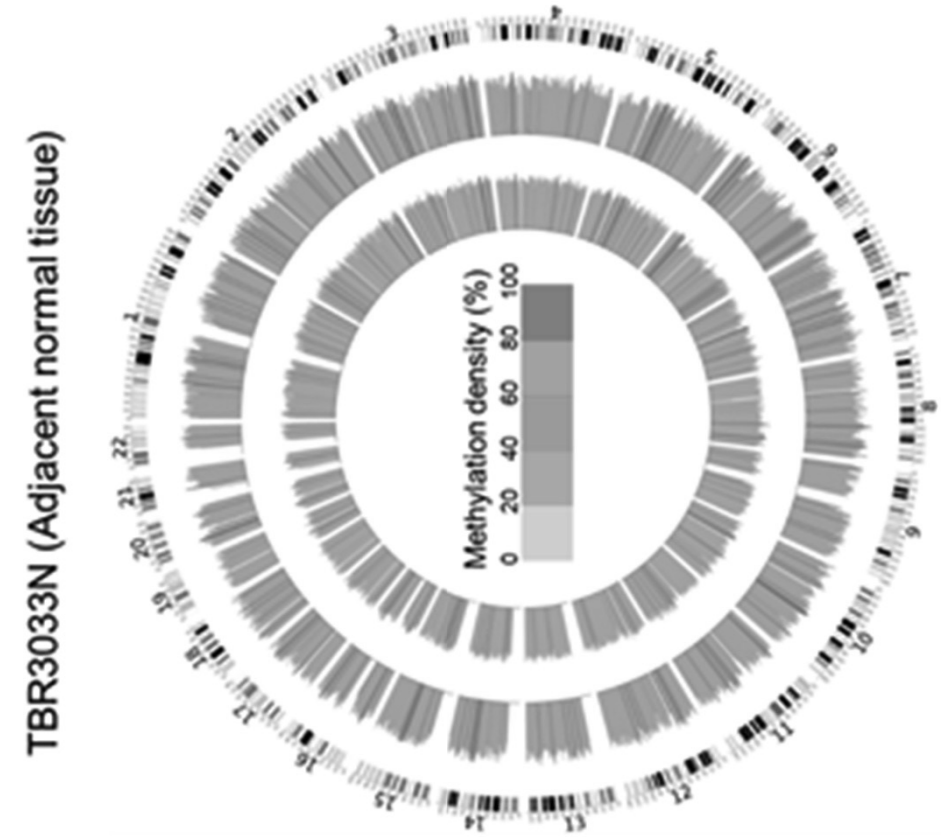


FIG. 67B

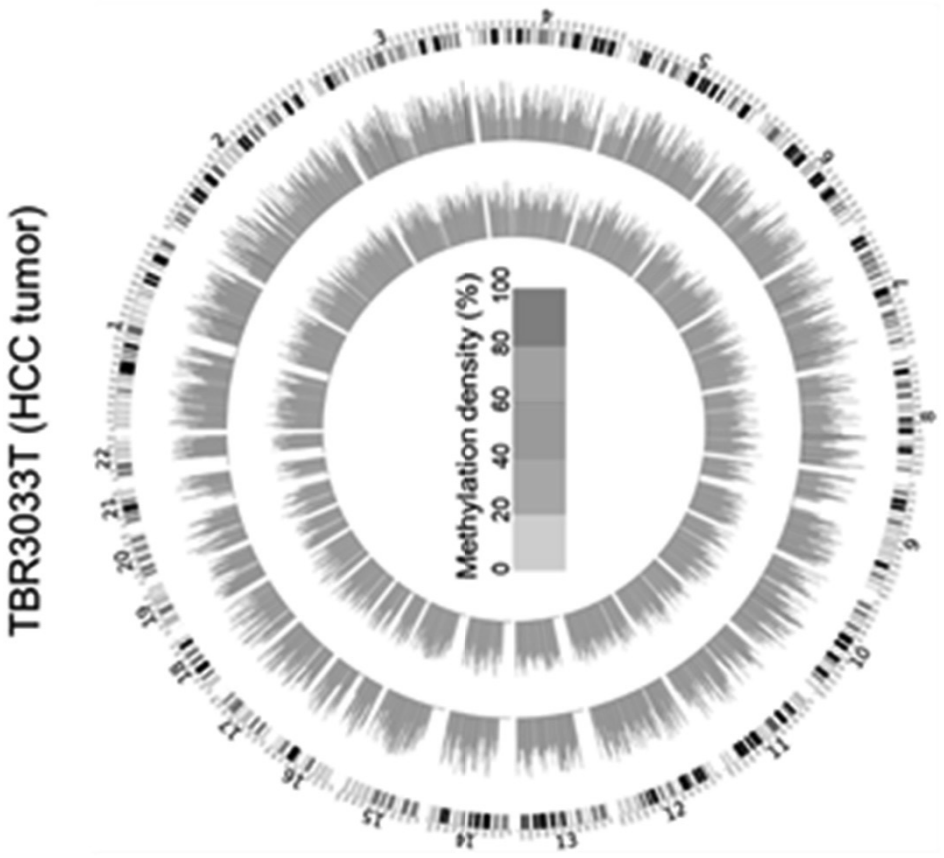


FIG. 67A

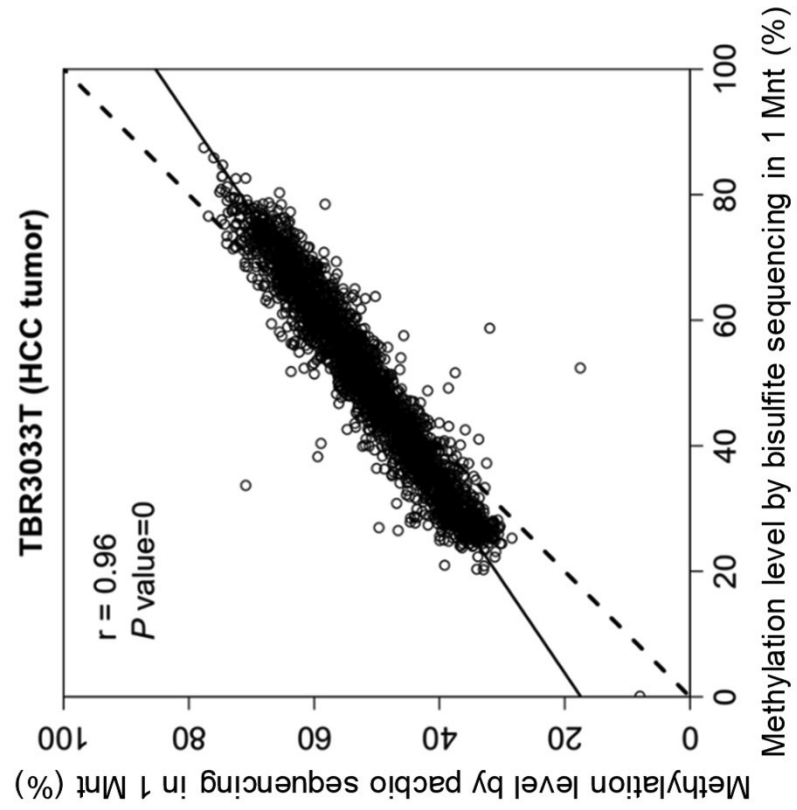


FIG. 68A

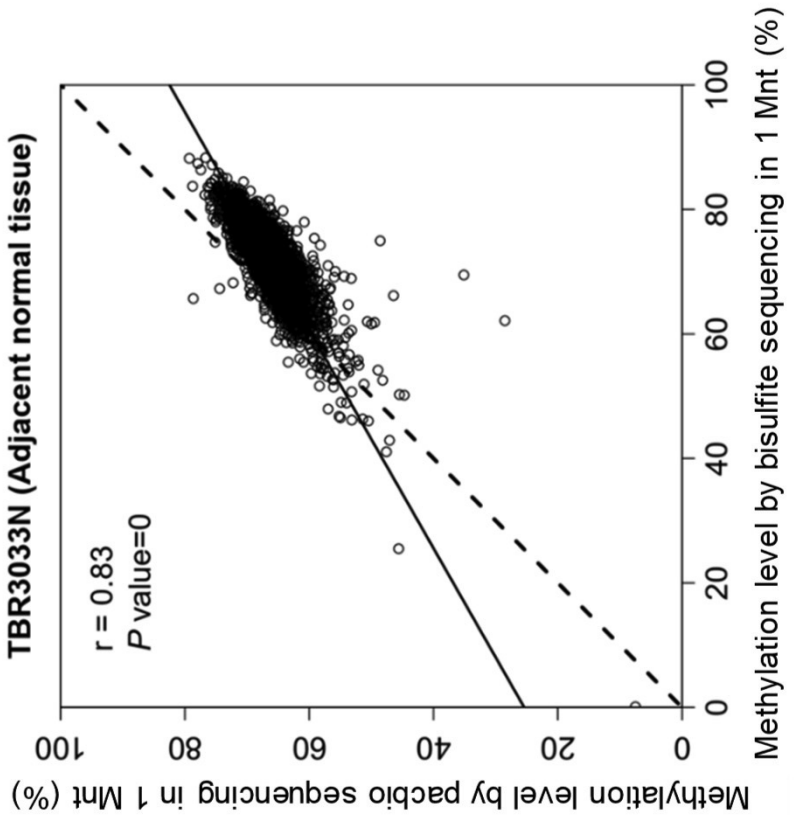


FIG. 68B

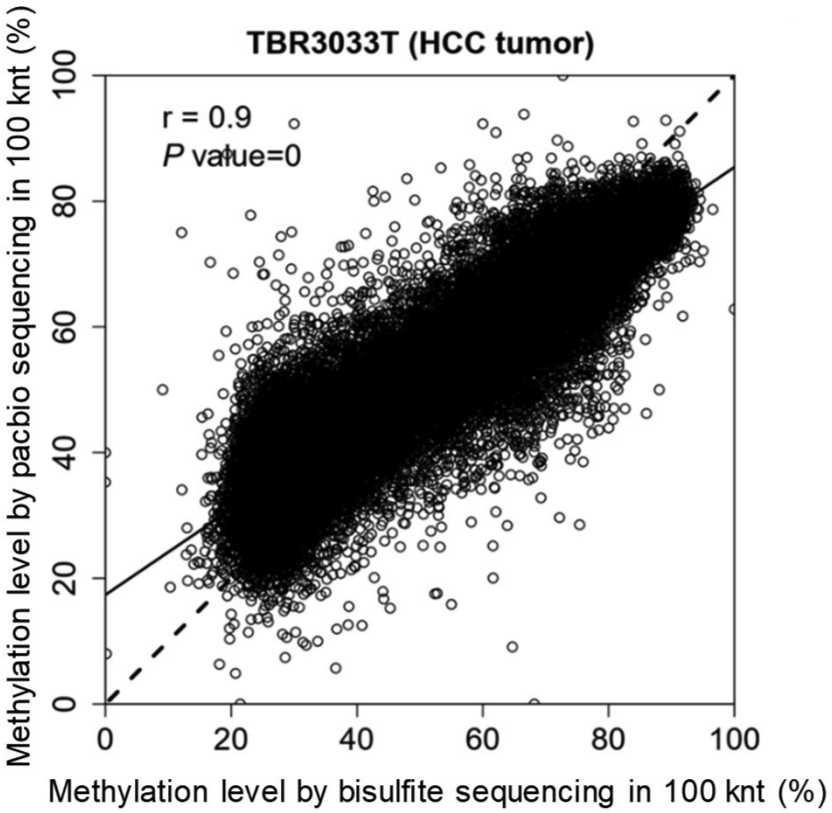


FIG. 69A

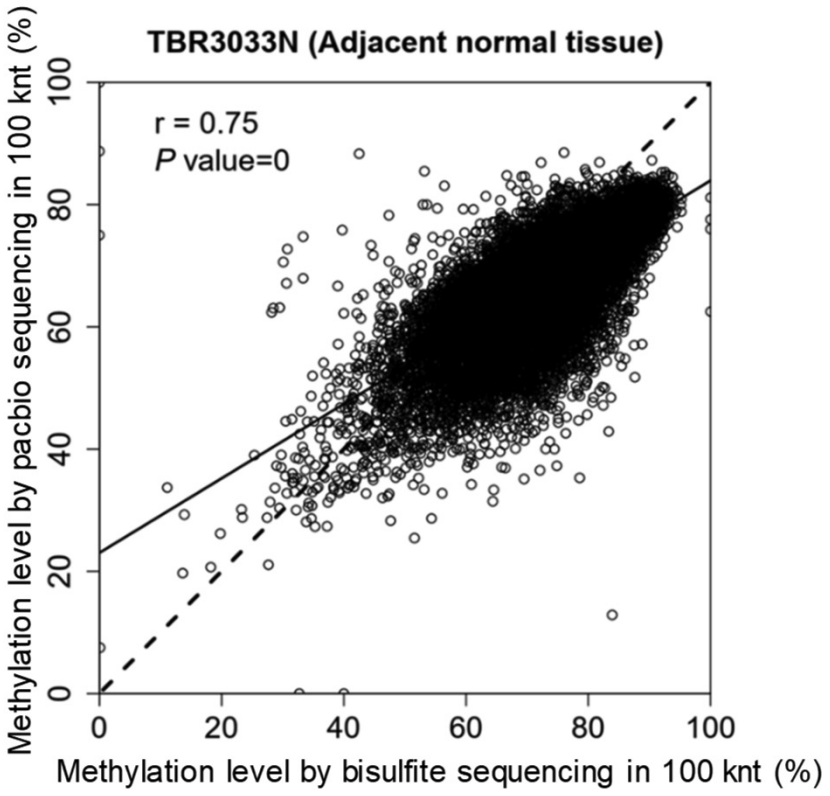


FIG. 69B

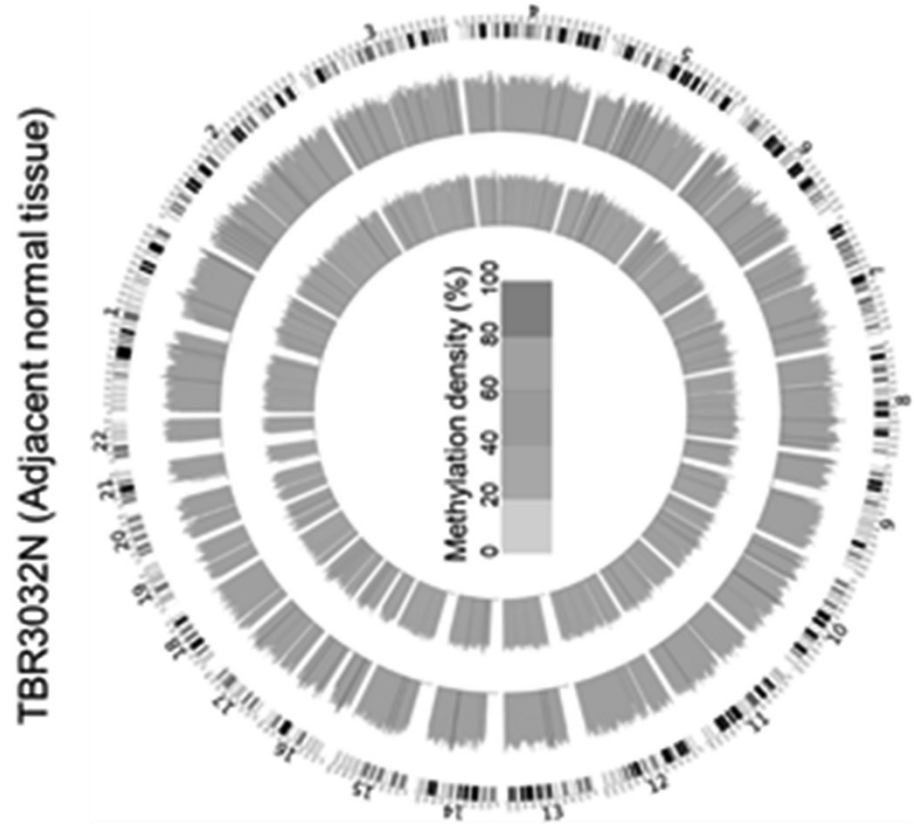


FIG. 70B

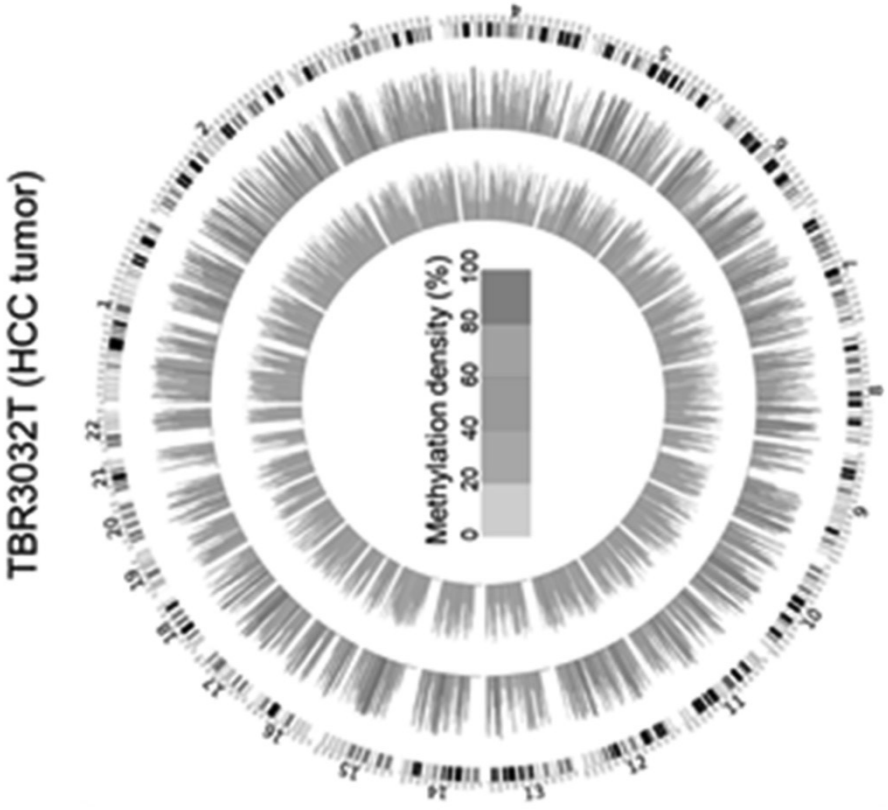


FIG. 70A

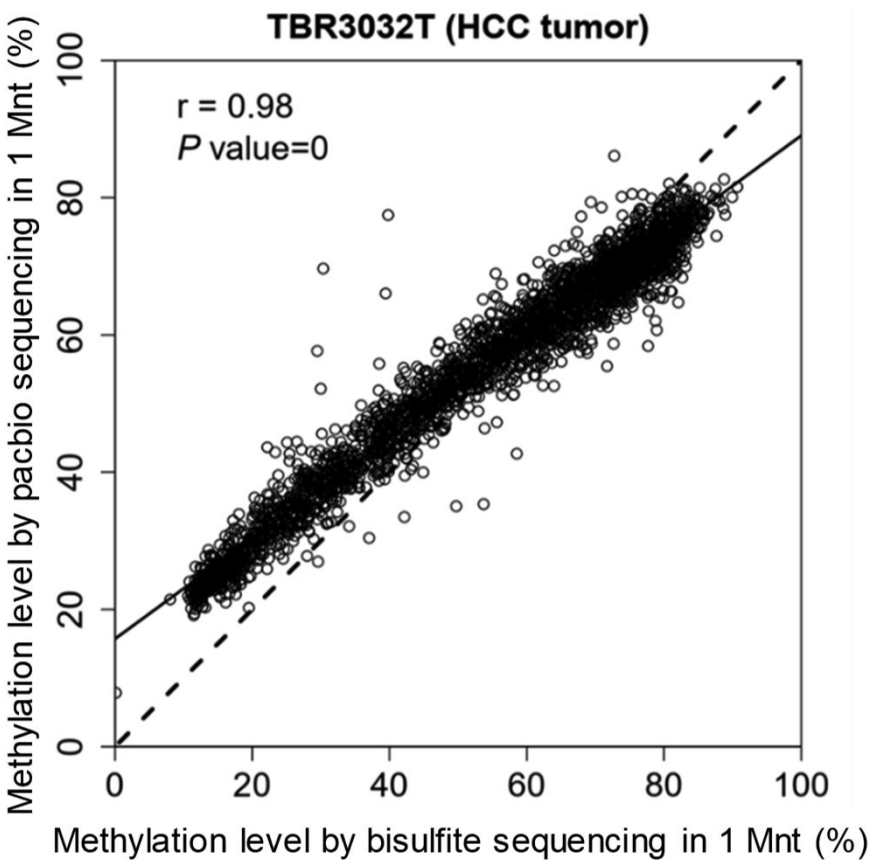


FIG. 71A

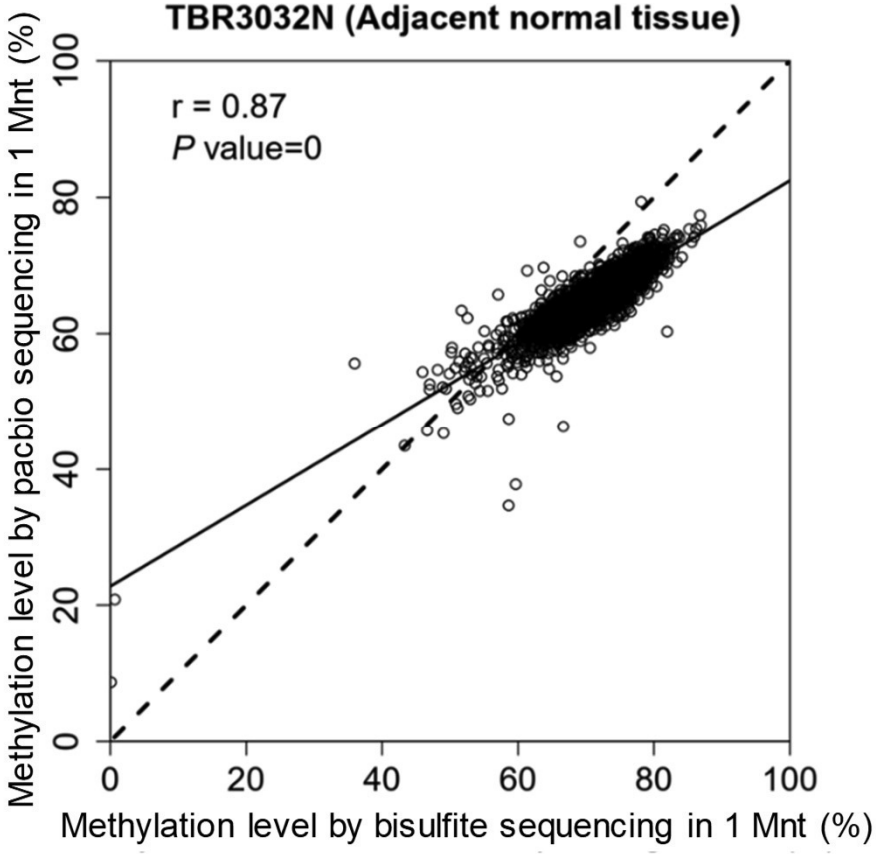


FIG. 71B

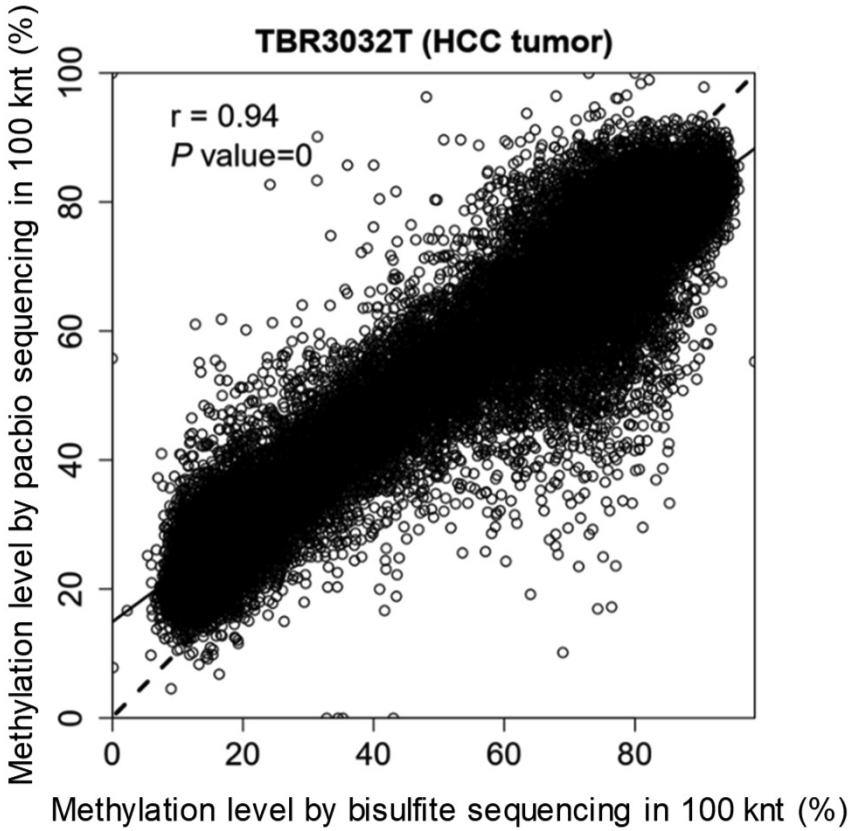


FIG. 72A

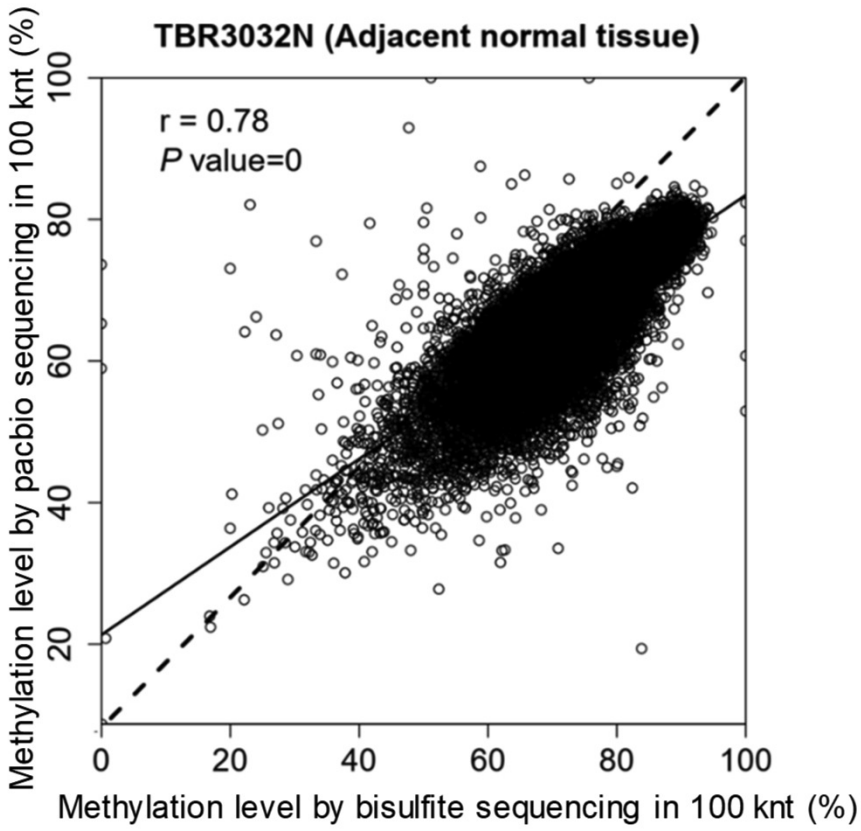


FIG. 72B

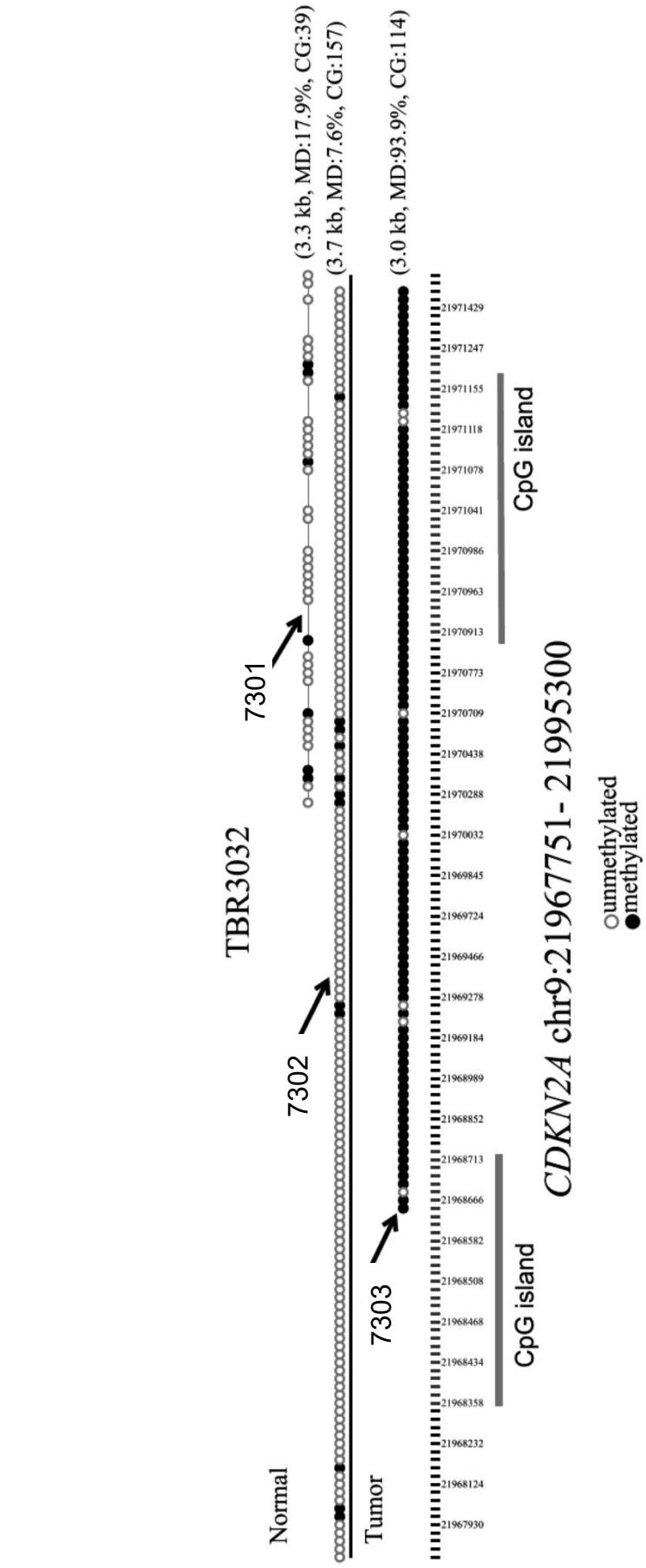

TBR3032
7301
7302
7303
(3.3 kb, MD:17.9%, CG:39)
(3.7 kb, MD:7.6%, CG:157)
(3.0 kb, MD:93.9%, CG:114)
Normal
Tumor
CpG island
CpG island
CDKN2A chr9:21967751- 21995300
unmethylated
methylated
21971429
21971247
21971155
21971118
21971078
21971041
21970986
21970963
21970913
21970773
21970709
21970438
21970288
21970032
21969845
21969724
21969466
21969278
21969184
21968989
21968852
21968713
21968666
21968582
21968508
21968468
21968434
21968358
21968232
21968124
21967930

FIG. 73

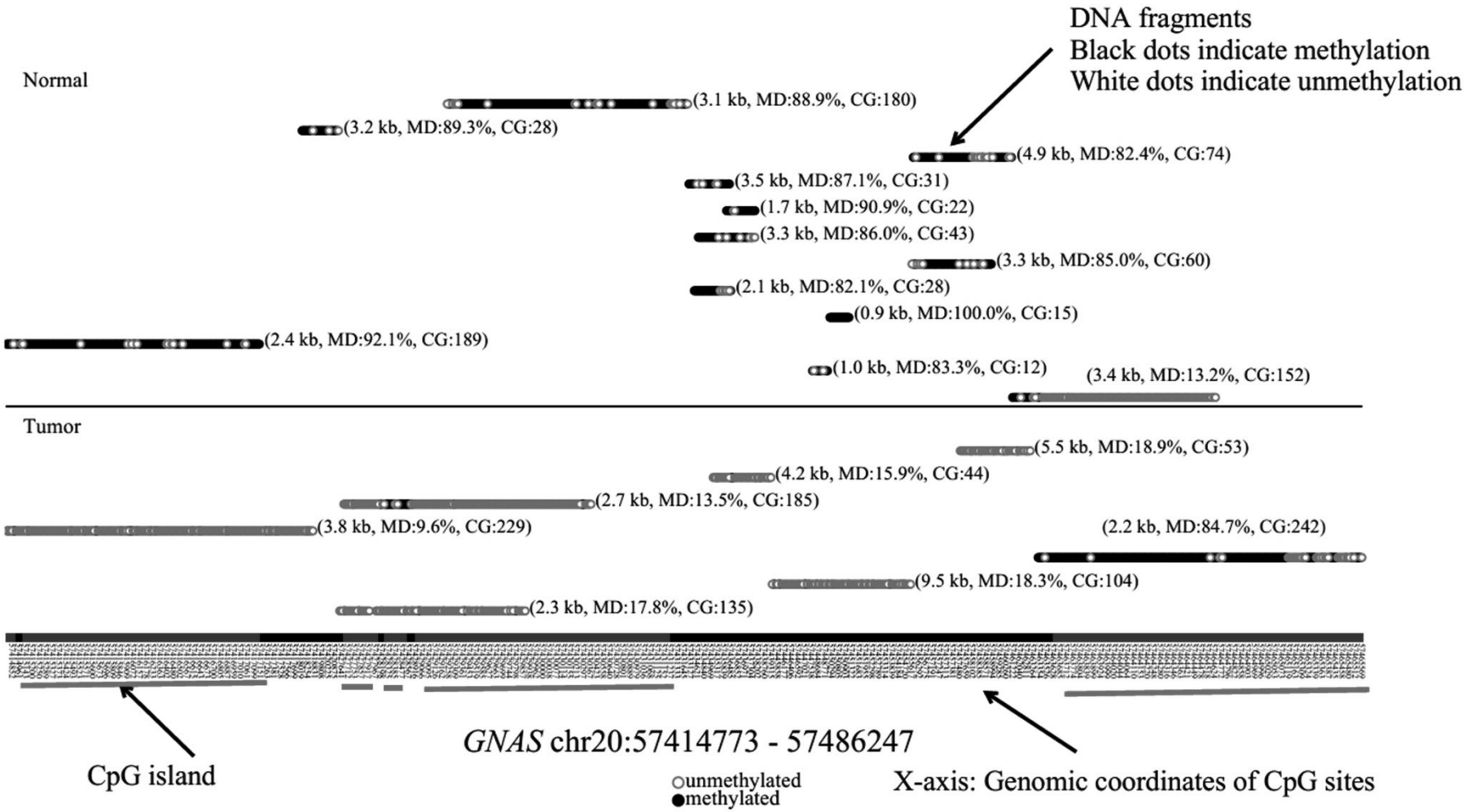

DNA fragments
Black dots indicate methylation
White dots indicate unmethylation
Normal
(3.1 kb, MD:88.9%, CG:180)
(3.2 kb, MD:89.3%, CG:28)
(4.9 kb, MD:82.4%, CG:74)
(3.5 kb, MD:87.1%, CG:31)
(1.7 kb, MD:90.9%, CG:22)
(3.3 kb, MD:86.0%, CG:43)
(3.3 kb, MD:85.0%, CG:60)
(2.1 kb, MD:82.1%, CG:28)
(0.9 kb, MD:100.0%, CG:15)
(2.4 kb, MD:92.1%, CG:189)
(1.0 kb, MD:83.3%, CG:12)
(3.4 kb, MD:13.2%, CG:152)
Tumor
(5.5 kb, MD:18.9%, CG:53)
(4.2 kb, MD:15.9%, CG:44)
(2.7 kb, MD:13.5%, CG:185)
(3.8 kb, MD:9.6%, CG:229)
(2.2 kb, MD:84.7%, CG:242)
(9.5 kb, MD:18.3%, CG:104)
(2.3 kb, MD:17.8%, CG:135)
GNAS chr20:57414773 - 57486247
CpG island
unmethylated
methylated
X-axis: Genomic coordinates of CpG sites

FIG. 74A

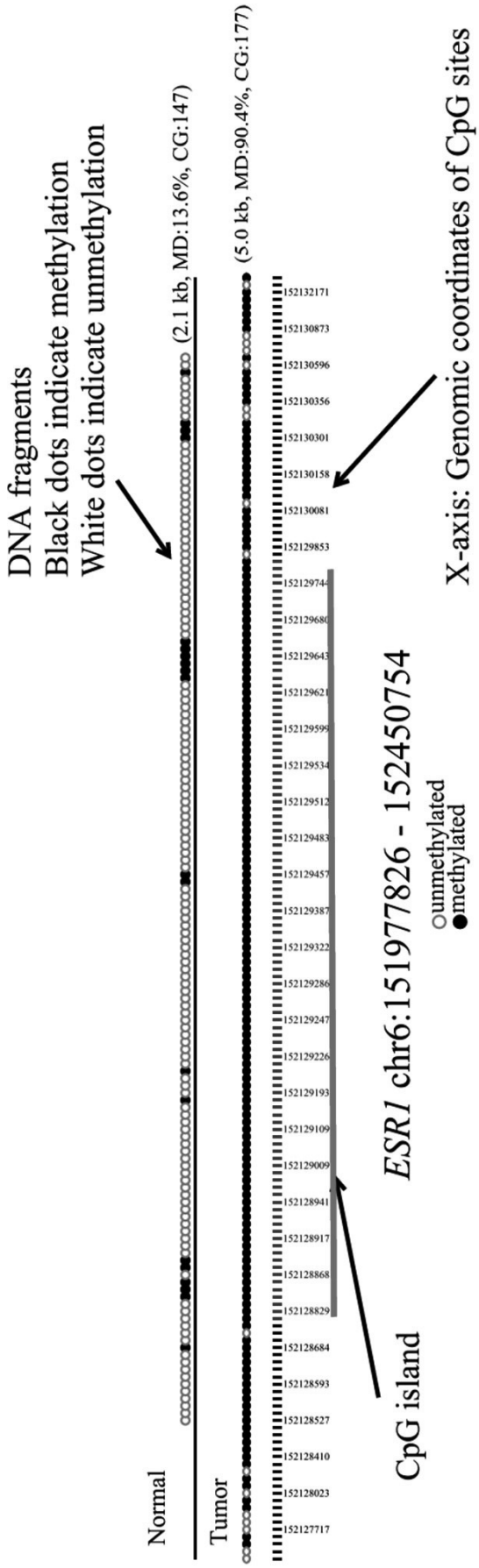

DNA fragments
Black dots indicate methylation
White dots indicate unmethylation
(2.1 kb, MD:13.6%, CG:147)
(5.0 kb, MD:90.4%, CG:177)
Normal
Tumor
CpG island
ESR1 chr6:151977826 - 152450754
unmethylated
methylated
X-axis: Genomic coordinates of CpG sites
152132171
152130873
152130596
152130356
152130301
152130158
152130081
152129853
152129744
152129680
152129643
152129621
152129599
152129534
152129512
152129483
152129457
152129387
152129322
152129286
152129247
152129226
152129193
152129109
152129009
152128941
152128917
152128868
152128829
152128684
152128593
152128527
152128410
152128023
152127717

FIG. 74B

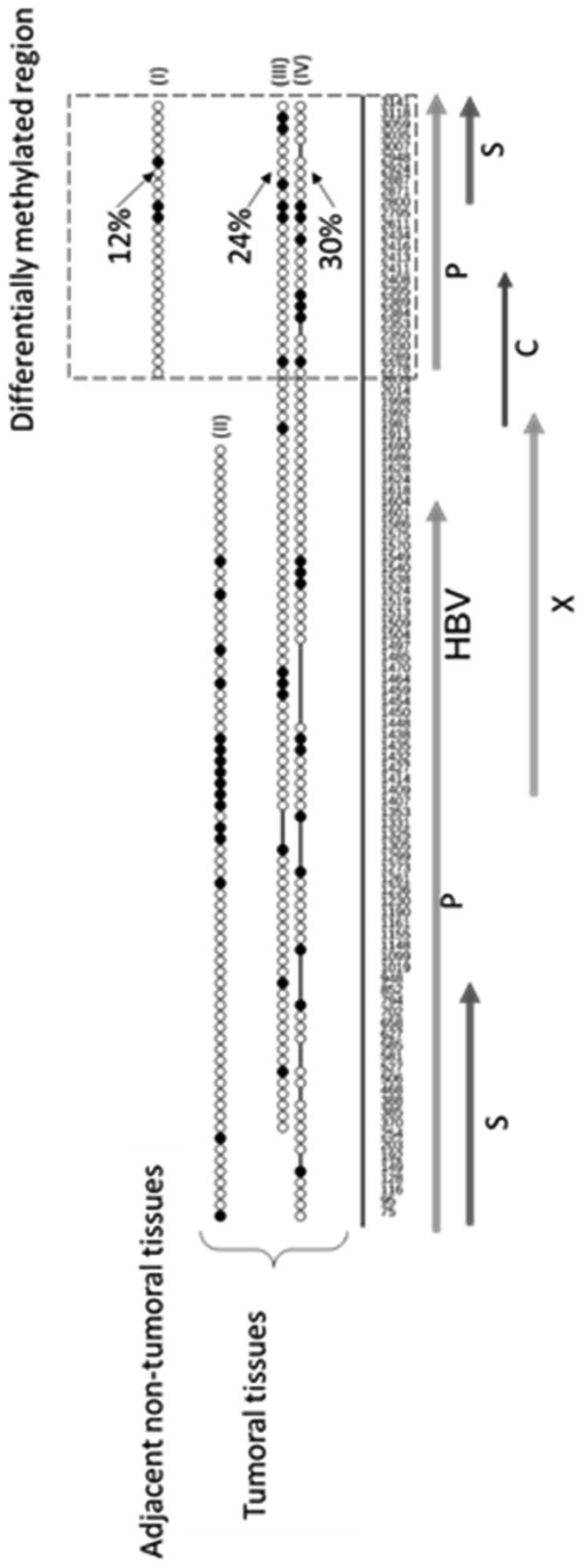
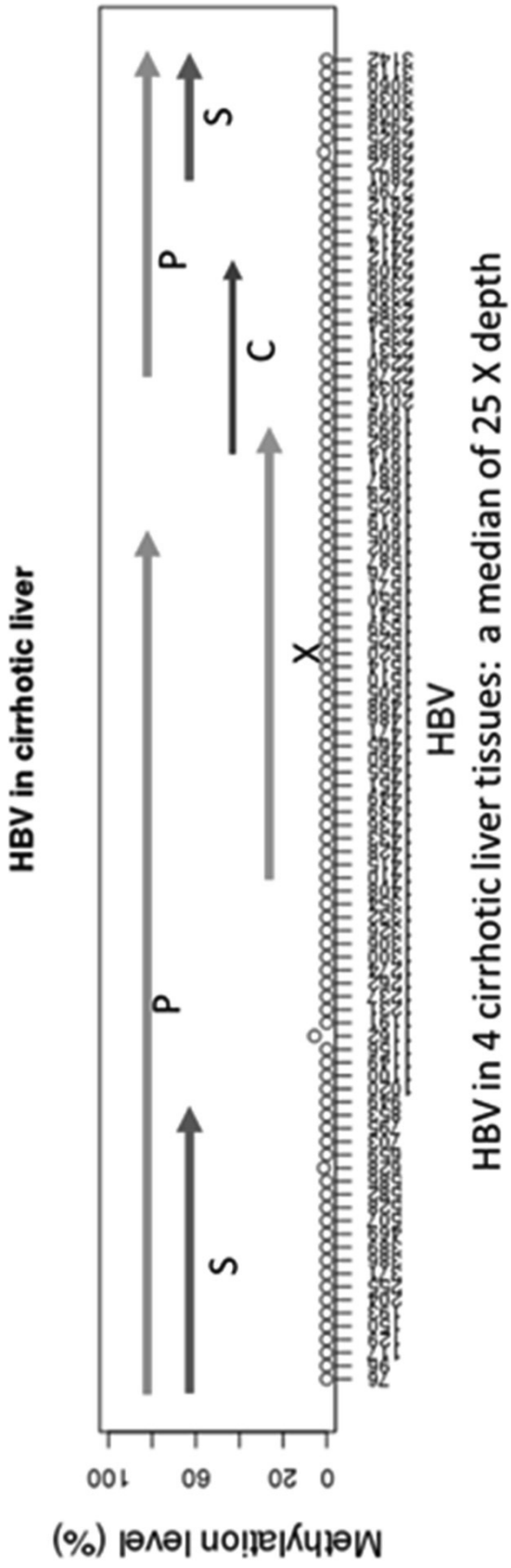
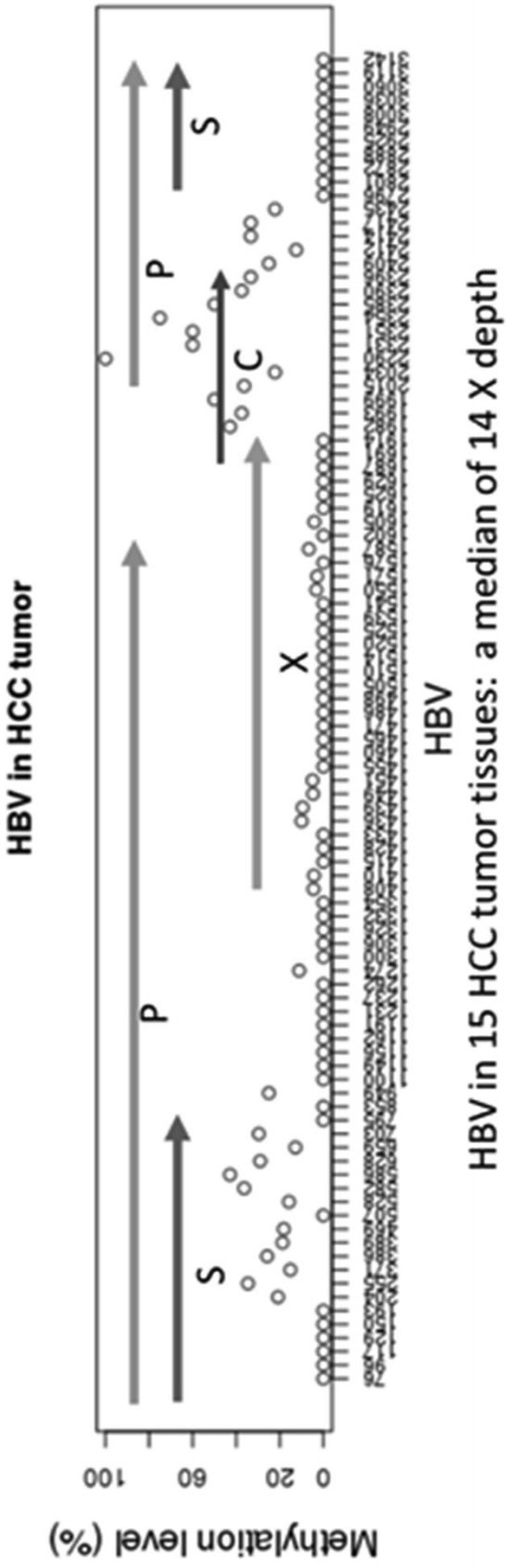

Differentially methylated region
(I)
(II)
(III)
(IV)
12%
24%
30%
Adjacent non-tumoral tissues
Tumoral tissues
HBV
S
P
X
P
C
S

FIG. 75



HBV in 4 cirrhotic liver tissues: a median of 25 X depth

FIG. 767A



HBV in 15 HCC tumor tissues: a median of 14 X depth

FIG. 76B

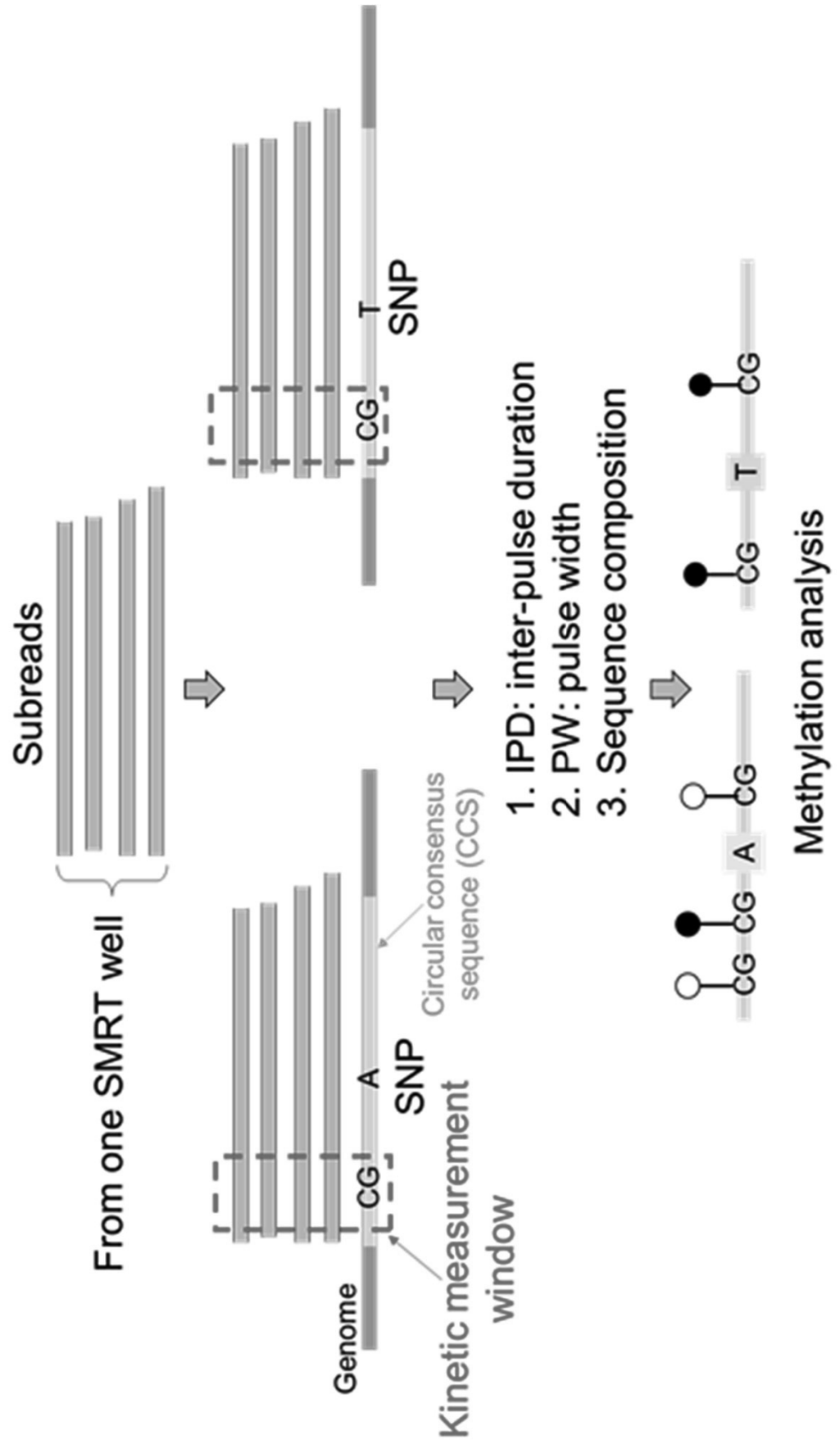

Subreads
From one SMRT well
Genome
CG
A
SNP
Kinetic measurement window
Circular consensus sequence (CCS)
CG
T
SNP
1. IPD: inter-pulse duration
2. PW: pulse width
3. Sequence composition
CG CG
A
CG
CG
T
CG
Methylation analysis

FIG. 77

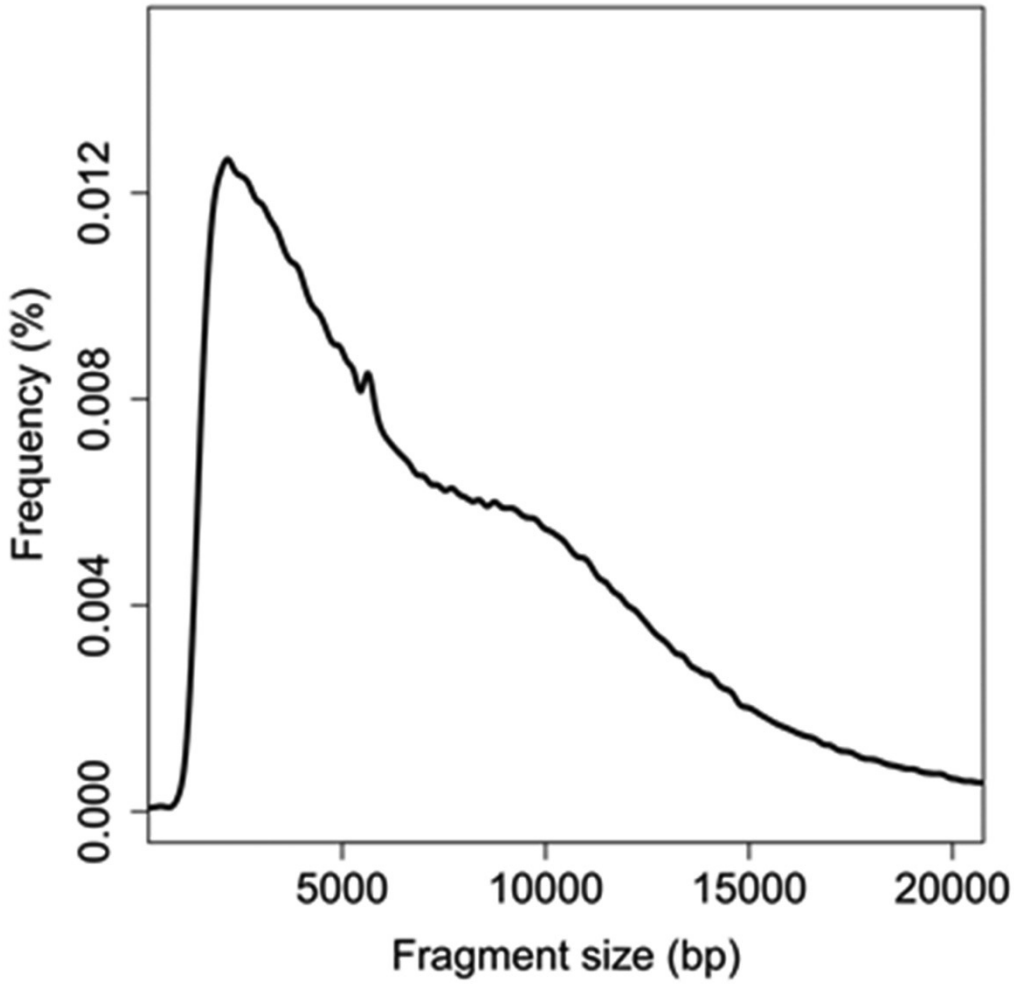

Frequency (%)
0.000
0.004
0.008
0.012
5000
10000
15000
20000
Fragment size (bp)

FIG. 78

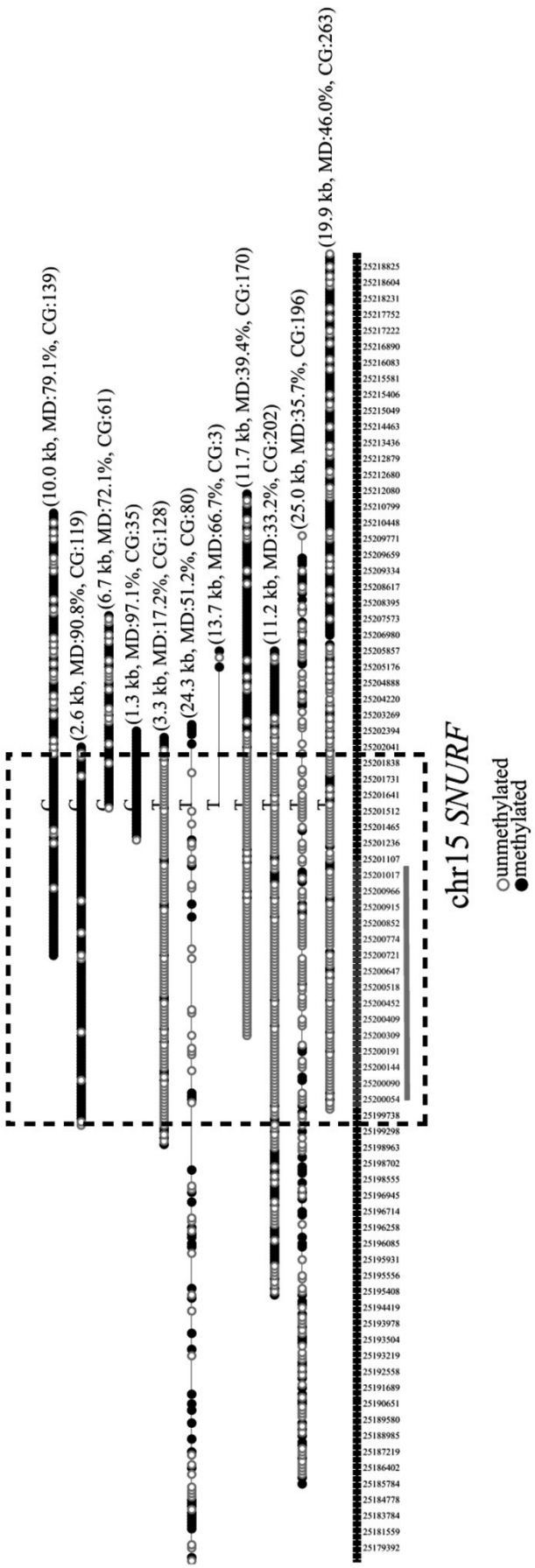

(10.0 kb, MD:79.1%, CG:139)
(2.6 kb, MD:90.8%, CG:119)
(6.7 kb, MD:72.1%, CG:61)
(1.3 kb, MD:97.1%, CG:35)
(3.3 kb, MD:17.2%, CG:128)
(24.3 kb, MD:51.2%, CG:80)
(13.7 kb, MD:66.7%, CG:3)
(11.7 kb, MD:39.4%, CG:170)
(11.2 kb, MD:33.2%, CG:202)
(25.0 kb, MD:35.7%, CG:196)
(19.9 kb, MD:46.0%, CG:263)
chr15 SNURF
unmethylated
methylated

FIG. 79A

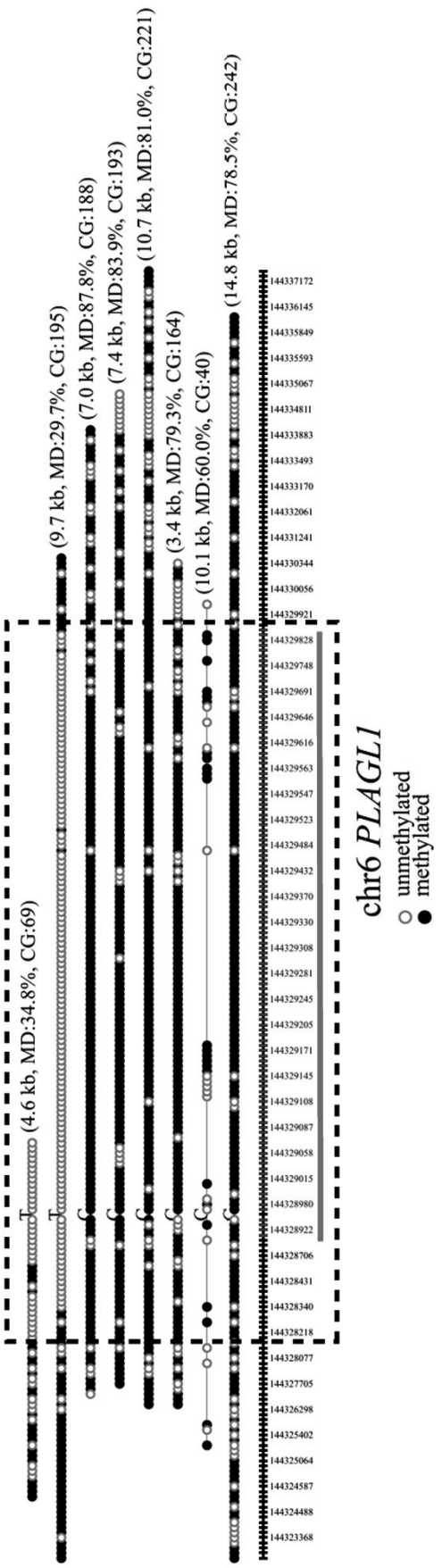

(4.6 kb, MD:34.8%, CG:69)
(9.7 kb, MD:29.7%, CG:195)
(7.0 kb, MD:87.8%, CG:188)
(7.4 kb, MD:83.9%, CG:193)
(10.7 kb, MD:81.0%, CG:221)
(3.4 kb, MD:79.3%, CG:164)
(10.1 kb, MD:60.0%, CG:40)
(14.8 kb, MD:78.5%, CG:242)
144337172
144336145
144335849
144335593
144335067
144334811
144333883
144333493
144333170
144332061
144331241
144330344
144330056
144329921
144329828
144329748
144329691
144329646
144329616
144329563
144329547
144329523
144329484
144329432
144329370
144329330
144329308
144329281
144329245
144329205
144329171
144329145
144329108
144329087
144329058
144329015
144328980
144328922
144328706
144328431
144328340
144328218
144328077
144327705
144326298
144325402
144325064
144324587
144324488
144323368
chr6 PLAGL1
unmethylated
methylated

FIG. 79B

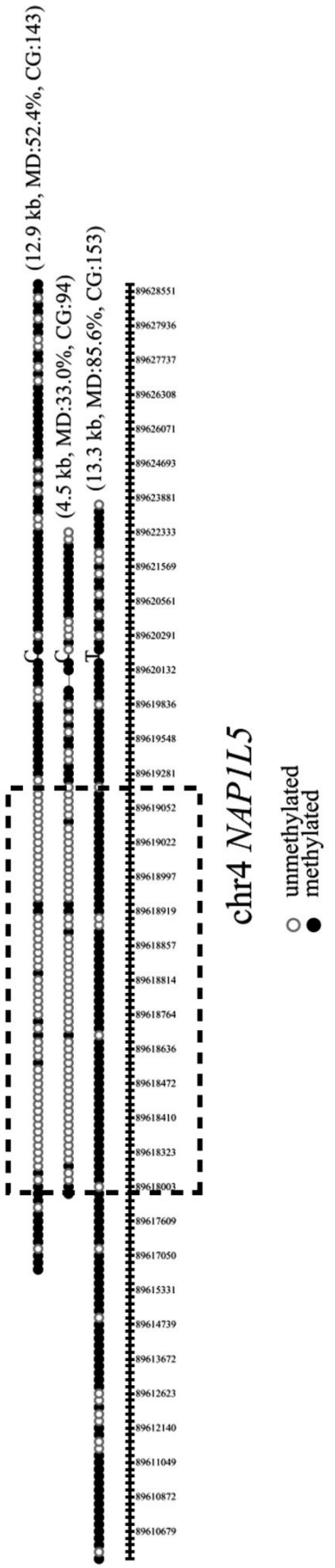

(12.9 kb, MD:52.4%, CG:143)
(4.5 kb, MD:33.0%, CG:94)
(13.3 kb, MD:85.6%, CG:153)
89628551
89627936
89627737
89626308
89626071
89624693
89623881
89622333
89621569
89620561
89620291
89620132
89619836
89619548
89619281
89619052
89619022
89618997
89618919
89618857
89618814
89618764
89618636
89618472
89618410
89618323
89618003
89617609
89617050
89615331
89614739
89613672
89612623
89612140
89611049
89610872
89610679
chr4 NAP1L5
unmethylated
methylated

FIG. 79C

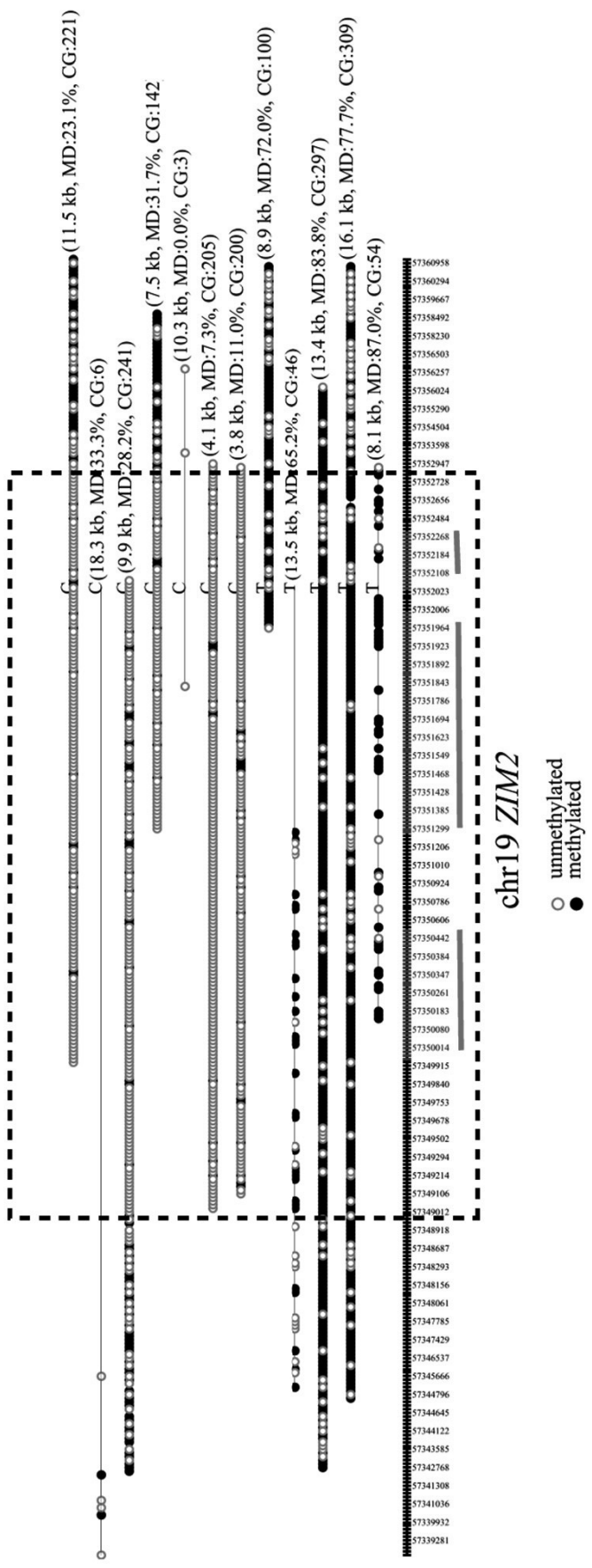

(11.5 kb, MD:23.1%, CG:221)
C(18.3 kb, MD:33.3%, CG:6)
C(9.9 kb, MD:28.2%, CG:241)
(7.5 kb, MD:31.7%, CG:142
C(10.3 kb, MD:0.0%, CG:3)
C(4.1 kb, MD:7.3%, CG:205)
C(3.8 kb, MD:11.0%, CG:200)
T(8.9 kb, MD:72.0%, CG:100)
T(13.5 kb, MD:65.2%, CG:46)
T(13.4 kb, MD:83.8%, CG:297)
T(16.1 kb, MD:77.7%, CG:309)
T(8.1 kb, MD:87.0%, CG:54)
chr19 ZIM2
unmethylated
methylated

FIG. 79D

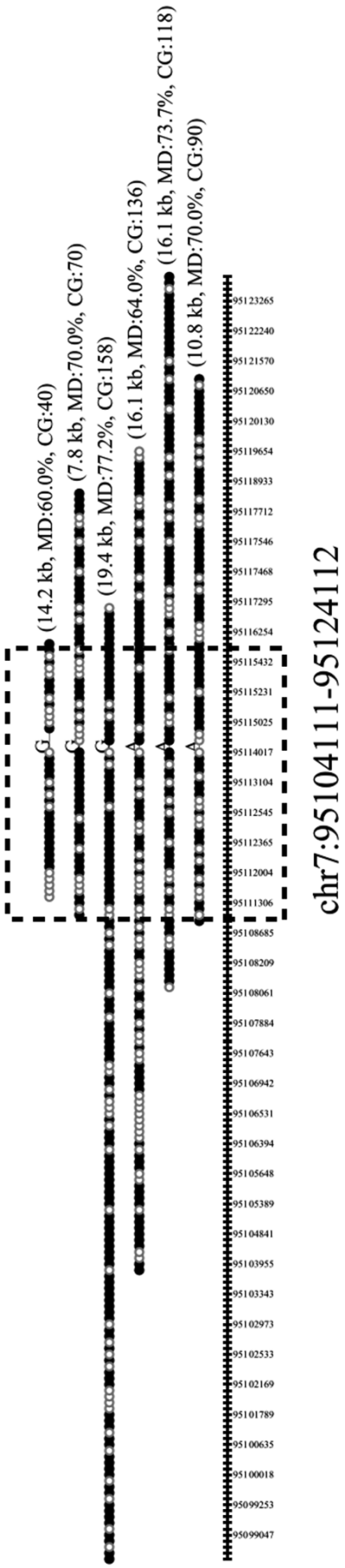

(14.2 kb, MD:60.0%, CG:40)
(7.8 kb, MD:70.0%, CG:70)
(19.4 kb, MD:77.2%, CG:158)
(16.1 kb, MD:64.0%, CG:136)
(16.1 kb, MD:73.7%, CG:118)
(10.8 kb, MD:70.0%, CG:90)
chr7:95104111-95124112
unmethylated
methylated

FIG. 80A

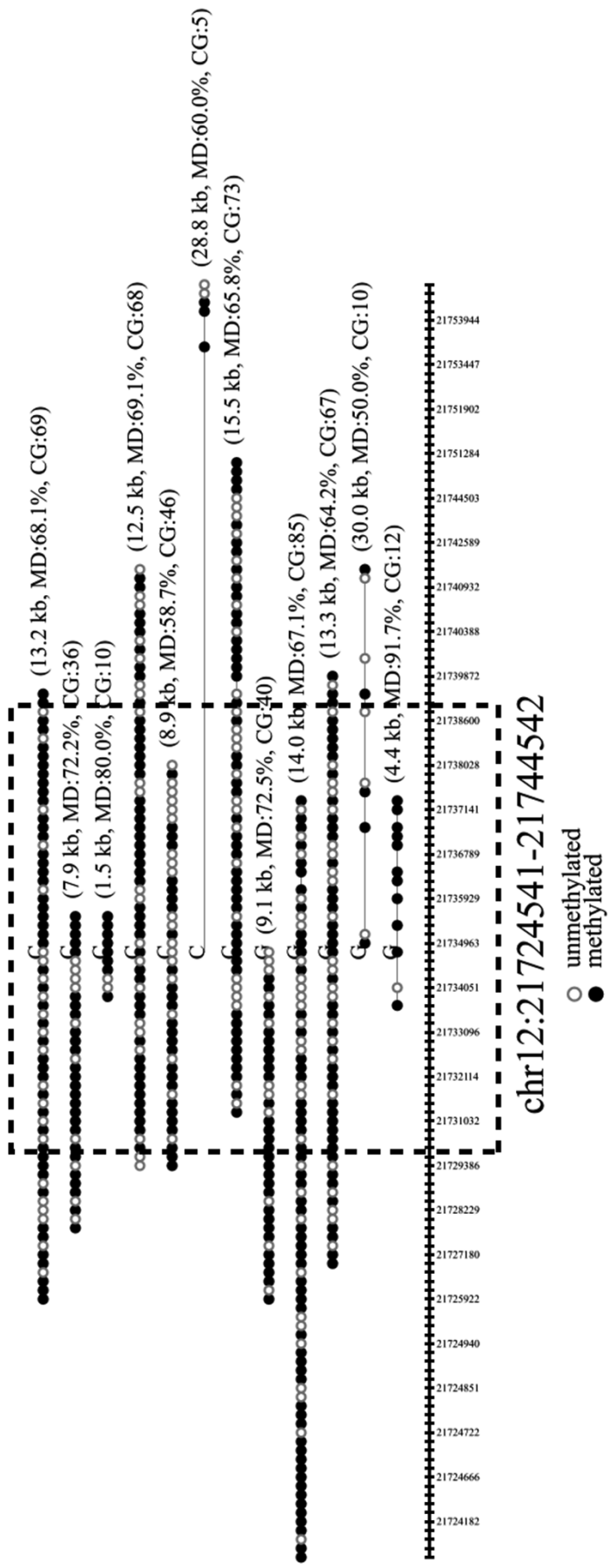

(13.2 kb, MD:68.1%, CG:69)
(7.9 kb, MD:72.2%, CG:36)
(1.5 kb, MD:80.0%, CG:10)
(12.5 kb, MD:69.1%, CG:68)
(8.9 kb, MD:58.7%, CG:46)
(28.8 kb, MD:60.0%, CG:5)
(15.5 kb, MD:65.8%, CG:73)
(9.1 kb, MD:72.5%, CG:40)
(14.0 kb, MD:67.1%, CG:85)
(13.3 kb, MD:64.2%, CG:67)
(30.0 kb, MD:50.0%, CG:10)
(4.4 kb, MD:91.7%, CG:12)
chr12:21724541-21744542
unmethylated
methylated
21753944
21753447
21751902
21751284
21744503
21742589
21740932
21740388
21739872
21738600
21738028
21737141
21736789
21735929
21734963
21734051
21733096
21732114
21731032
21729386
21728229
21727180
21725922
21724940
21724851
21724722
21724666
21724182

FIG. 80B

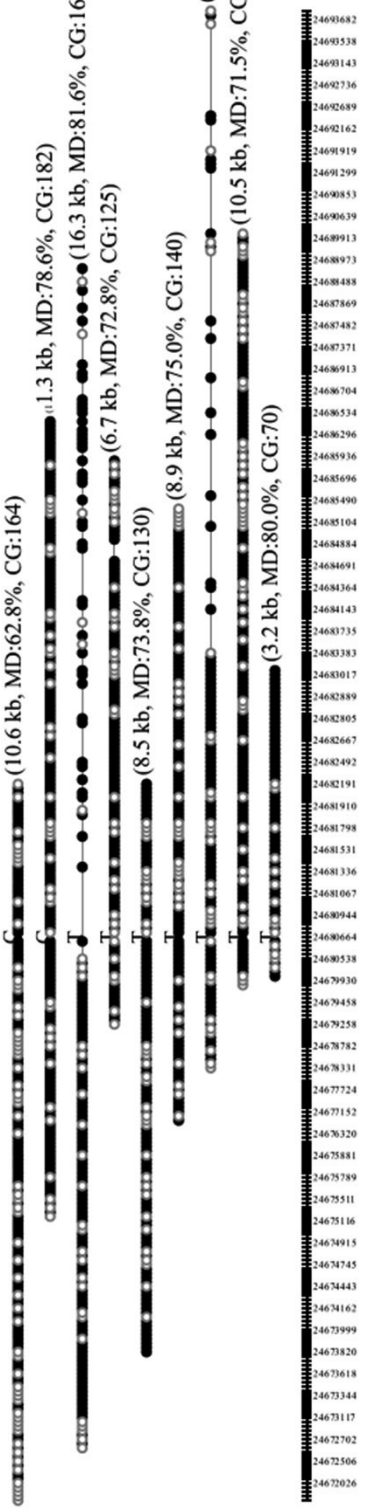

(10.6 kb, MD:62.8%, CG:164)
(1.3 kb, MD:78.6%, CG:182)
(16.3 kb, MD:81.6%, CG:163)
(6.7 kb, MD:72.8%, CG:125)
(8.5 kb, MD:73.8%, CG:130)
(8.9 kb, MD:75.0%, CG:140)
(16.3 kb, MD:75.4%, CG:118)
(10.5 kb, MD:71.5%, CG:172)
(3.2 kb, MD:80.0%, CG:70)
chr1:24670664-24690665
unmethylated
methylated

FIG. 80C

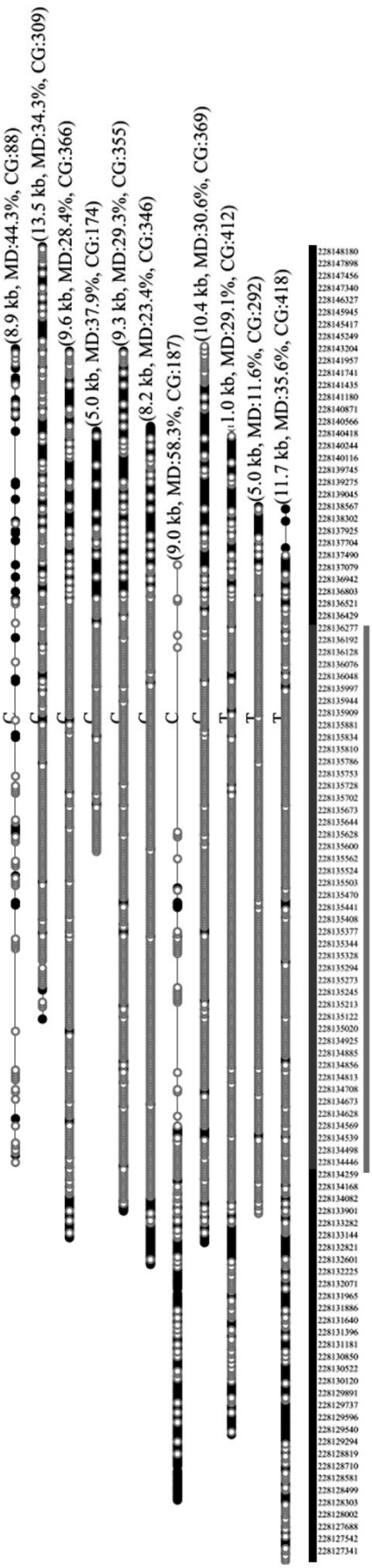

(8.9 kb, MD:44.3%, CG:88)
(13.5 kb, MD:34.3%, CG:309)
(9.6 kb, MD:28.4%, CG:366)
(5.0 kb, MD:37.9%, CG:174)
(9.3 kb, MD:29.3%, CG:355)
(8.2 kb, MD:23.4%, CG:346)
(9.0 kb, MD:58.3%, CG:187)
(10.4 kb, MD:30.6%, CG:369)
(1.0 kb, MD:29.1%, CG:412)
(5.0 kb, MD:11.6%, CG:292)
(11.7 kb, MD:35.6%, CG:418)
chr1:228125898-228145899
unmethylated
methylated

FIG. 80D

	Gene	Allele 1	Allele 2	Methylation level (%)	
				Allele 1	Allele 2
Imprinted genes	*SNURF*	T	C	15.73	89.37
	PLAGL1	T	C	7.56	89.41
	NAP1L5	C	T	12.5	91.07
	ZIM2	C	T	13	84.64
Randomly selected regions	Region 01	G	A	71.79	69.17
	Region 02	T	G	63.22	65.22
	Region 03	C	T	73.33	74.9
	Region 04	C	T	10.83	8.51

FIG. 81

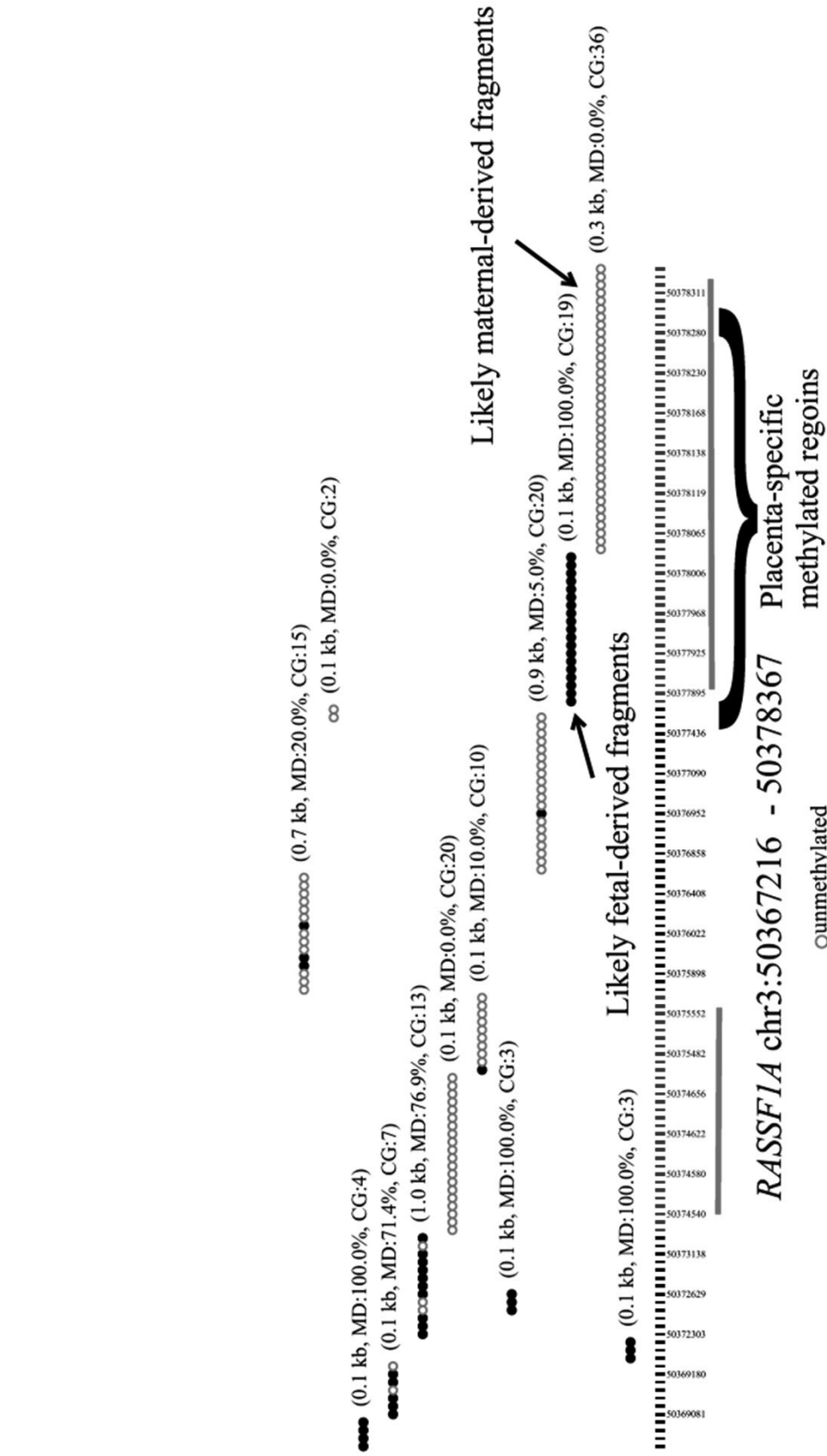

(0.7 kb, MD:20.0%, CG:15)
(0.1 kb, MD:0.0%, CG:2)
(0.1 kb, MD:100.0%, CG:4)
(0.1 kb, MD:71.4%, CG:7)
(1.0 kb, MD:76.9%, CG:13)
(0.1 kb, MD:0.0%, CG:20)
(0.1 kb, MD:10.0%, CG:10)
(0.1 kb, MD:100.0%, CG:3)
(0.9 kb, MD:5.0%, CG:20)
(0.1 kb, MD:100.0%, CG:19)
(0.3 kb, MD:0.0%, CG:36)
(0.1 kb, MD:100.0%, CG:3)
Likely maternal-derived fragments
Likely fetal-derived fragments
50378311
50378280
50378230
50378168
50378138
50378119
50378065
50378006
50377968
50377925
50377895
50377436
50377090
50376952
50376858
50376408
50376022
50375898
50375552
50375482
50374656
50374622
50374580
50374540
50373138
50372629
50372303
50369180
50369081
RASSF1A chr3:50367216 - 50378367
Placenta-specific methylated regoins
unmethylated
methylated

FIG. 82

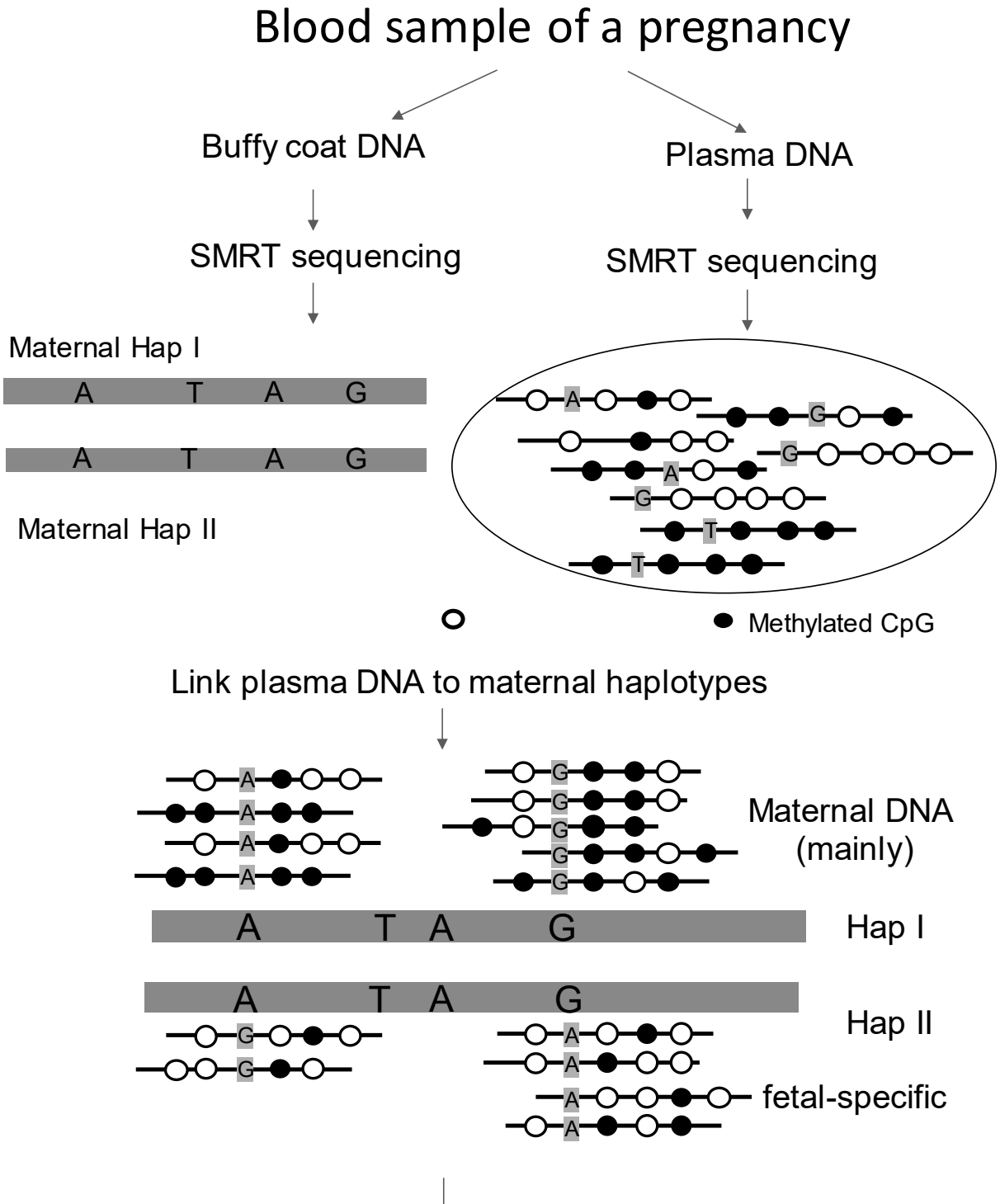

Blood sample of a pregnancy
Buffy coat DNA
Plasma DNA
SMRT sequencing
SMRT sequencing
Maternal Hap I
A T A G
A T A G
Maternal Hap II
Methylated CpG
Link plasma DNA to maternal haplotypes
Maternal DNA (mainly)
A T A G
Hap I
A T A G
Hap II
fetal-specific
The methylation level of fetal-specific DNA would be lower compared with that of maternal fragments

FIG. 83

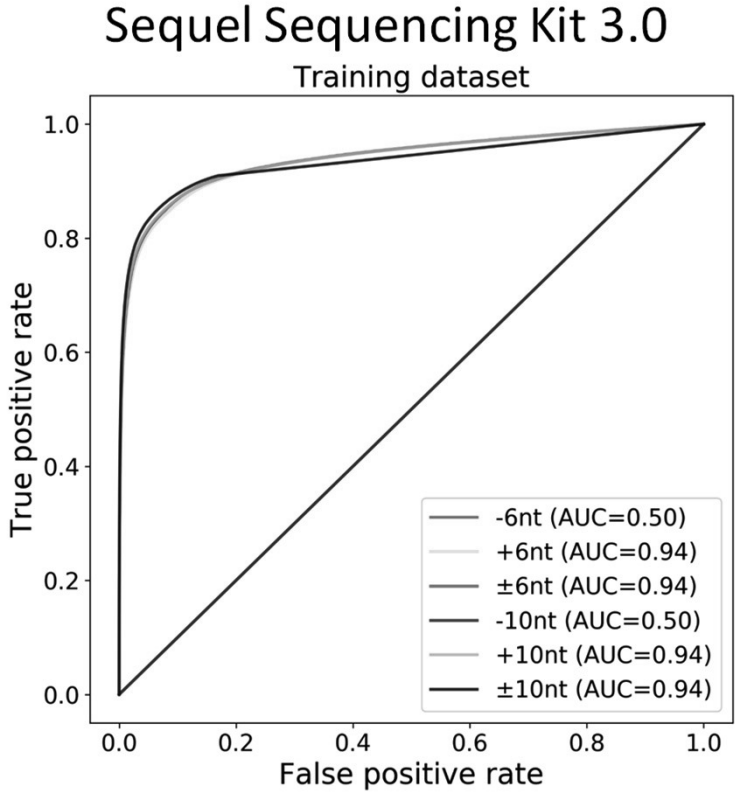


FIG. 84A

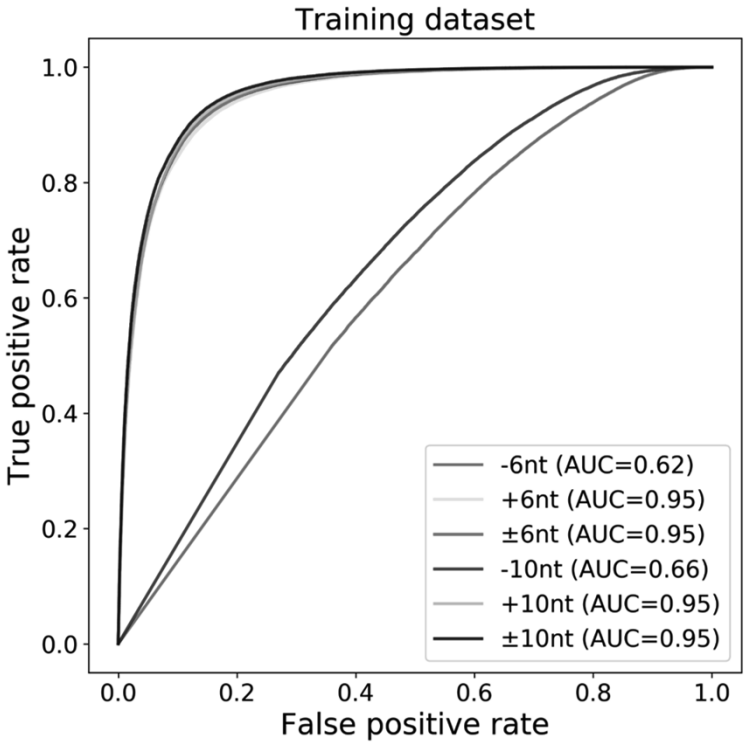


FIG. 84B

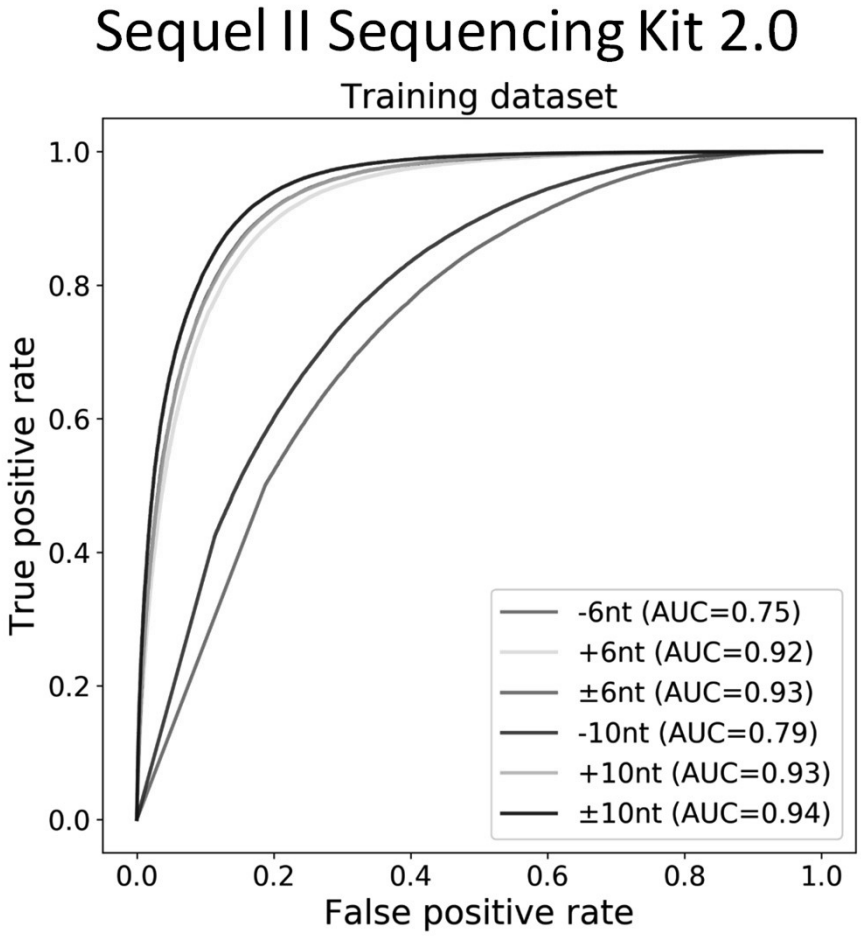

Sequel II Sequencing Kit 2.0
Training dataset
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.75)
+6nt (AUC=0.92)
±6nt (AUC=0.93)
-10nt (AUC=0.79)
+10nt (AUC=0.93)
±10nt (AUC=0.94)

FIG. 84C

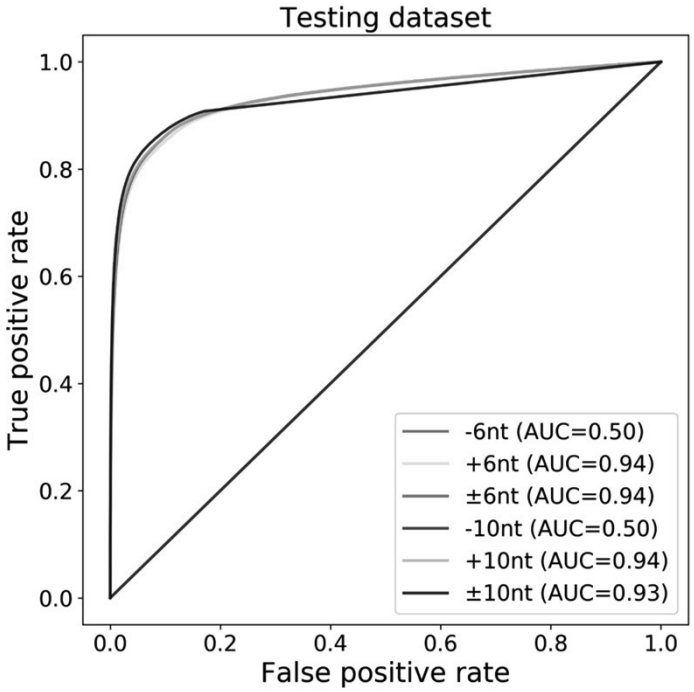


FIG. 85A

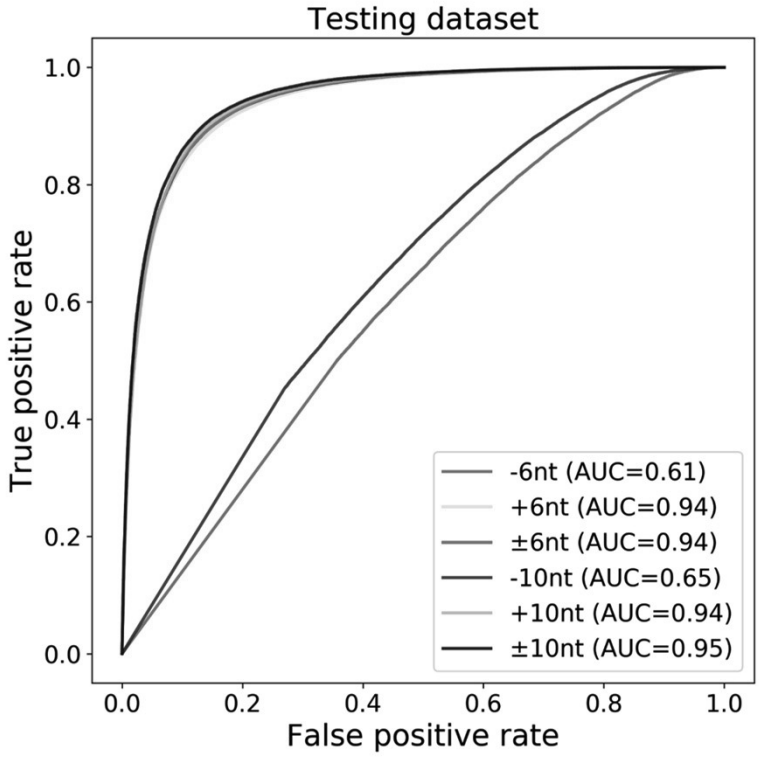


FIG. 85B

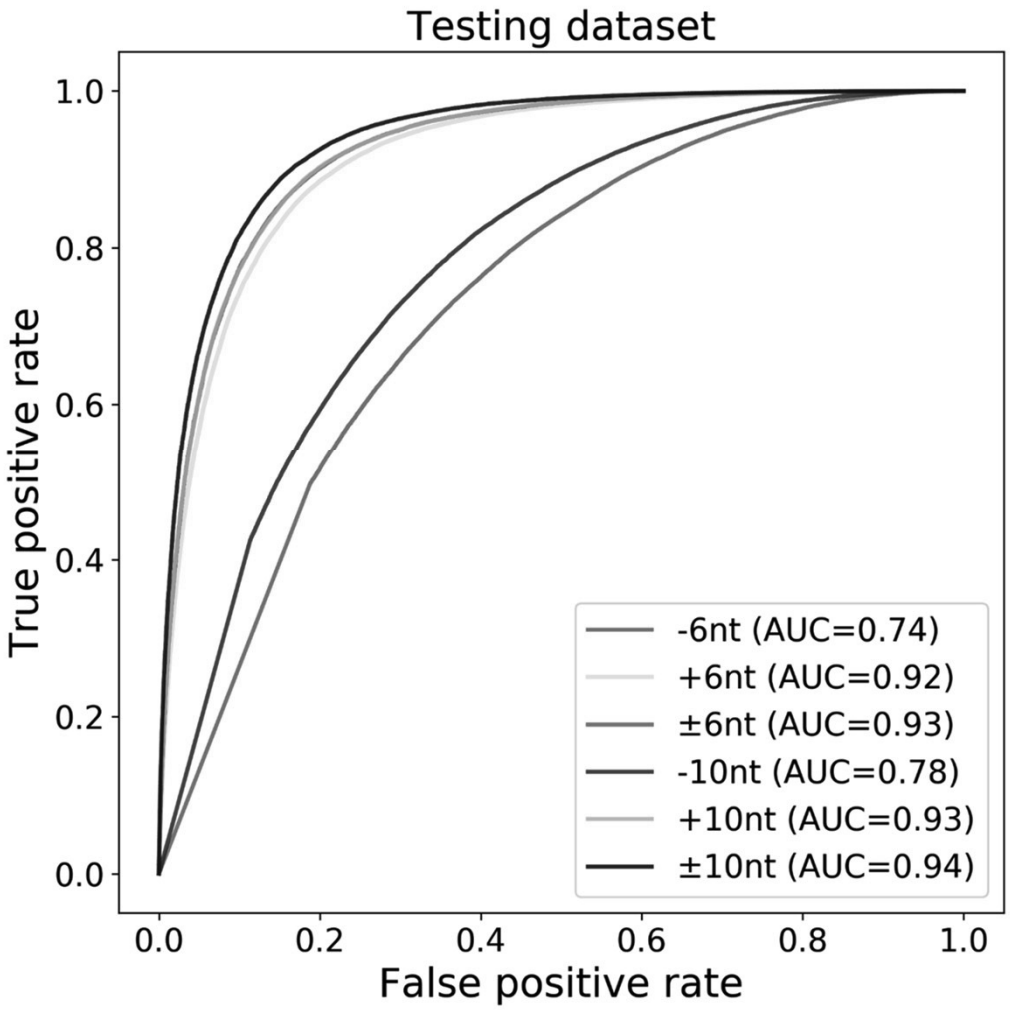

Sequel II Sequencing Kit 2.0
Testing dataset
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.74)
+6nt (AUC=0.92)
±6nt (AUC=0.93)
-10nt (AUC=0.78)
+10nt (AUC=0.93)
±10nt (AUC=0.94)

FIG. 85C

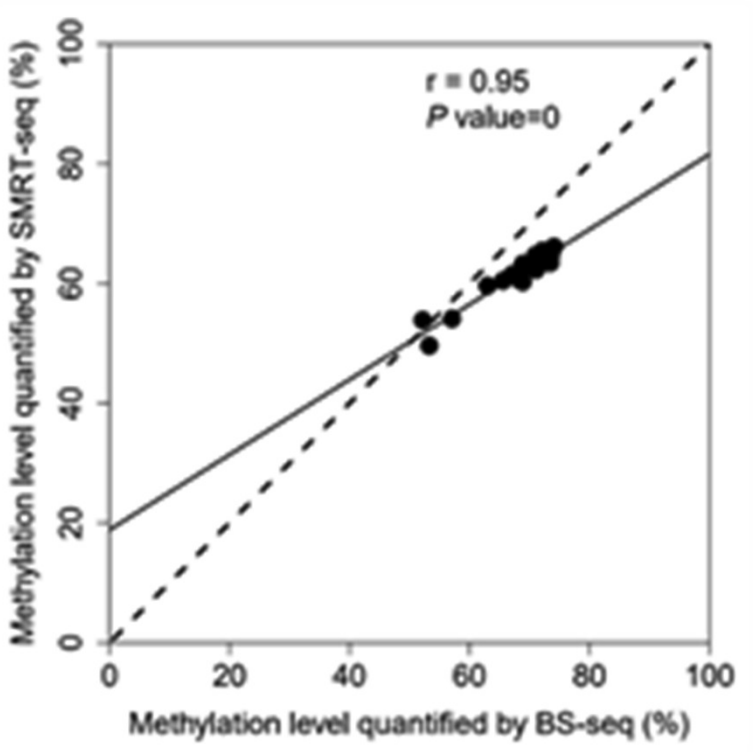


FIG. 86A

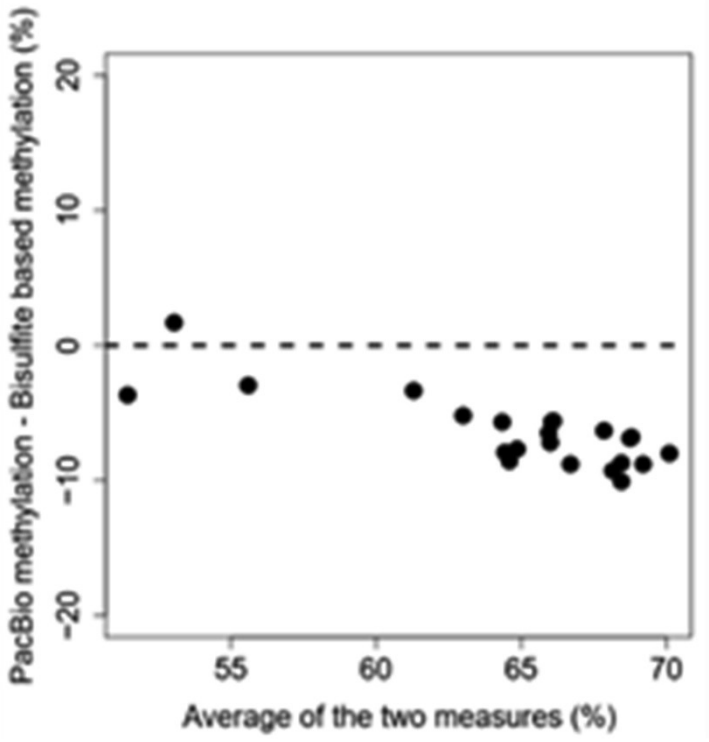


FIG. 86B

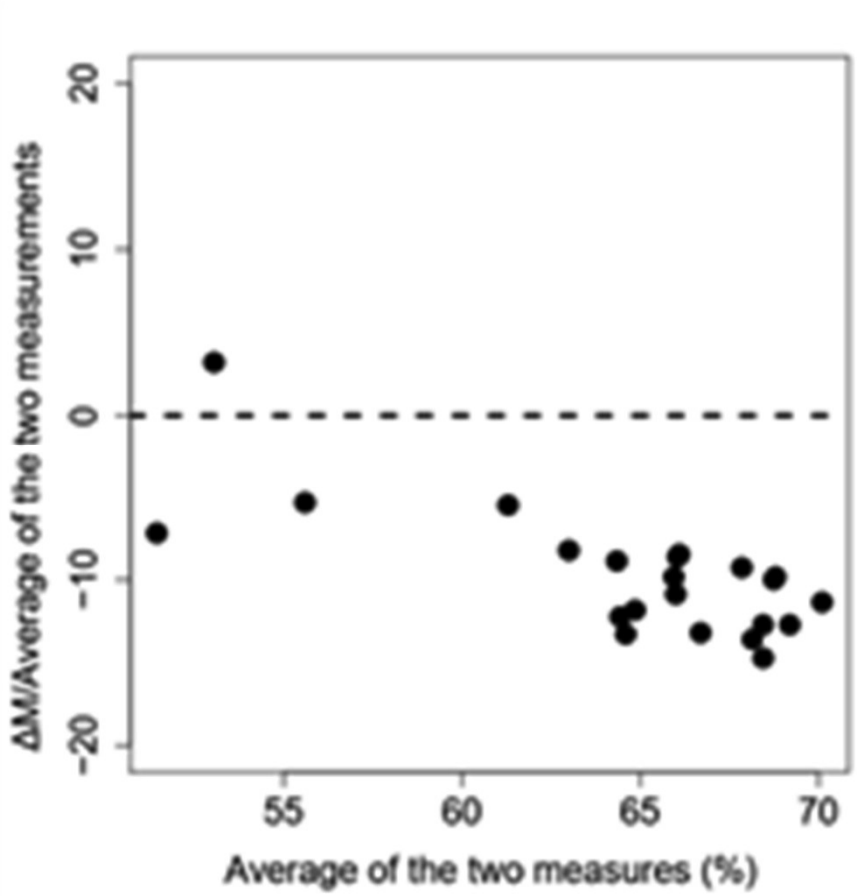

ΔM/Average of the two measurements
20
10
0
−10
−20
55
60
65
70
Average of the two measures (%)

FIG. 86C

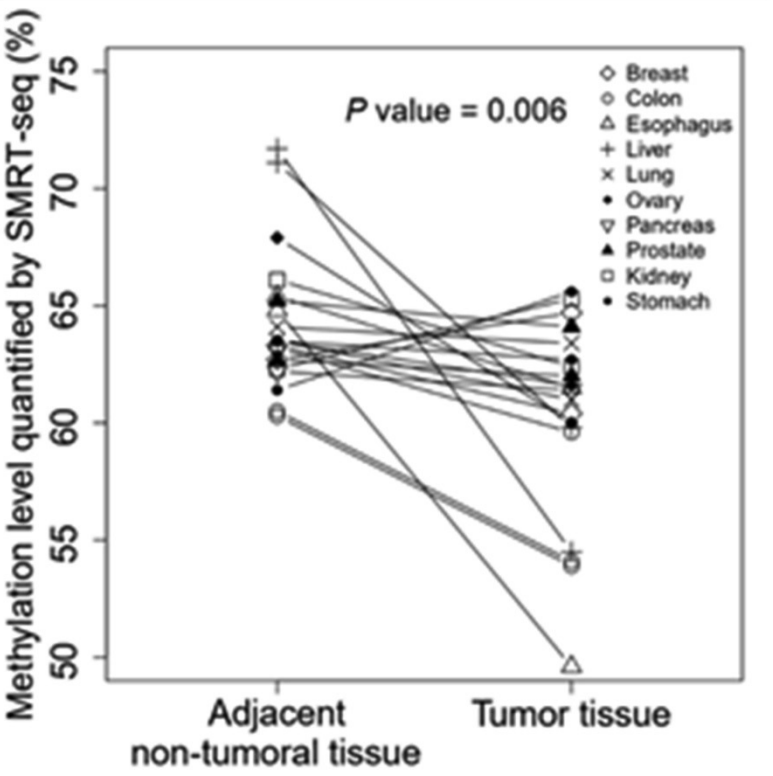


FIG. 87A

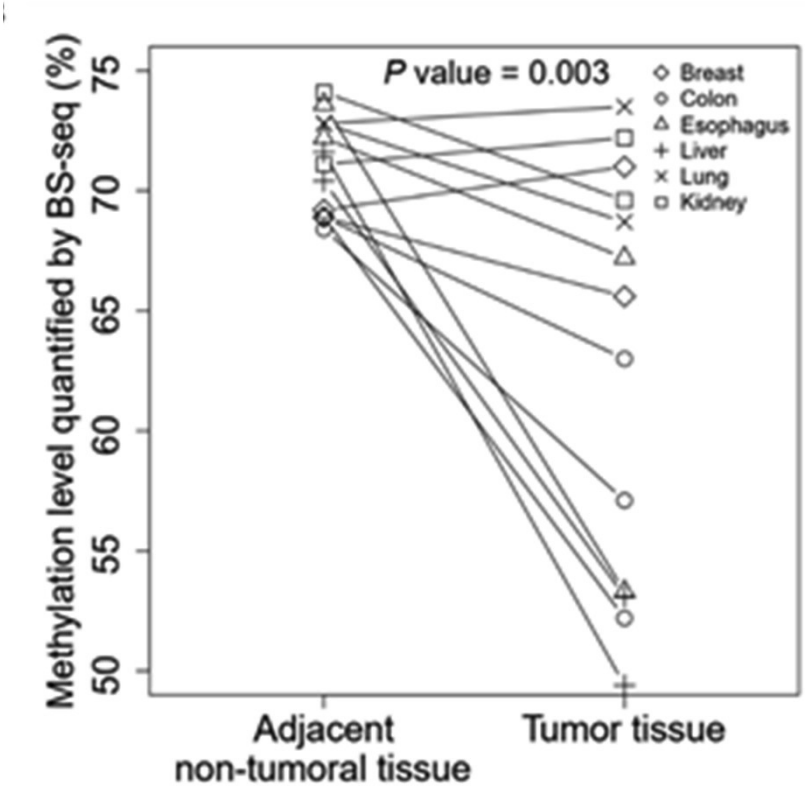


FIG. 87B

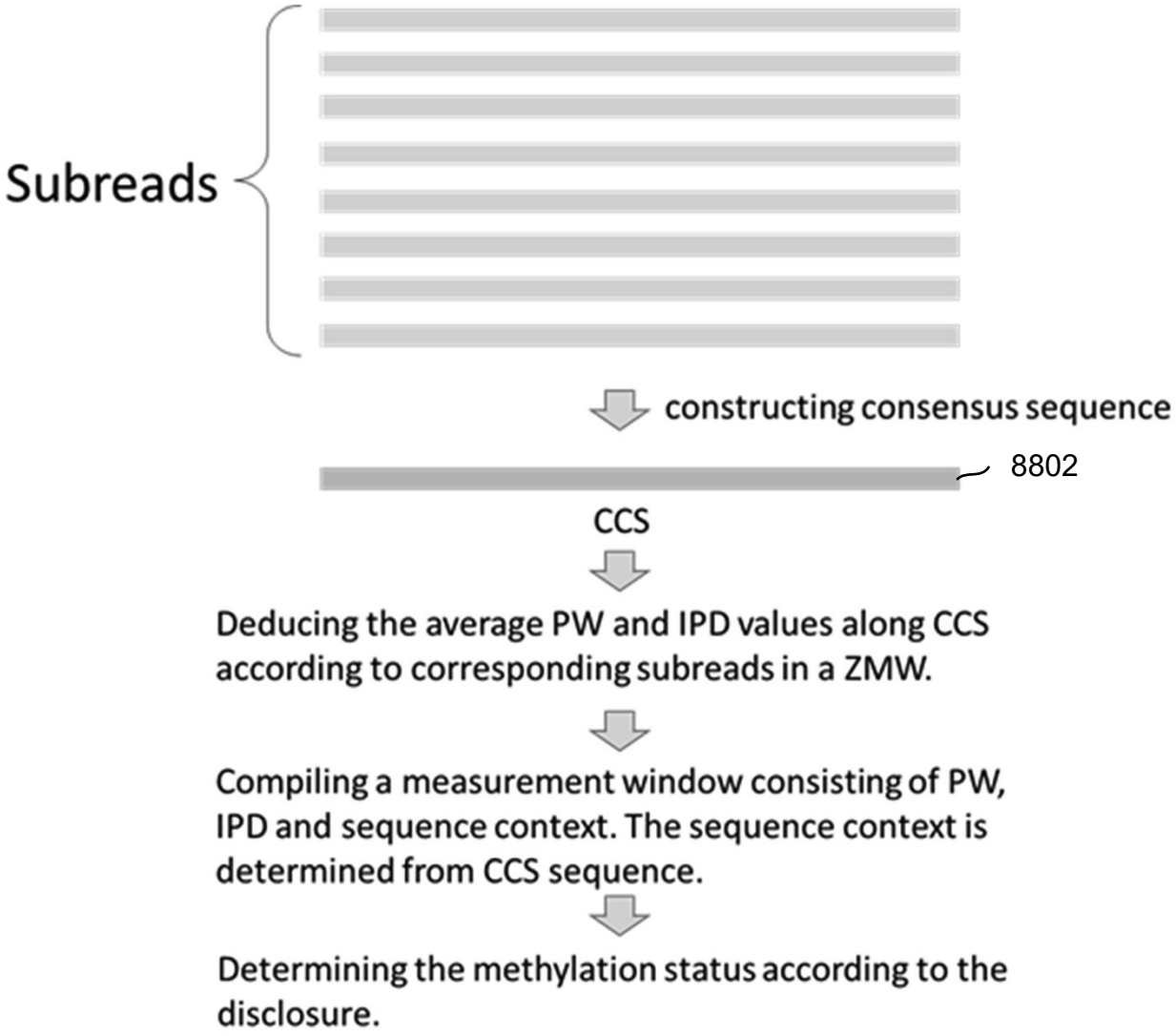

Subreads
constructing consensus sequence
8802
CCS
Deducing the average PW and IPD values along CCS according to corresponding subreads in a ZMW.
Compiling a measurement window consisting of PW, IPD and sequence context. The sequence context is determined from CCS sequence.
Determining the methylation status according to the disclosure.

FIG. 88

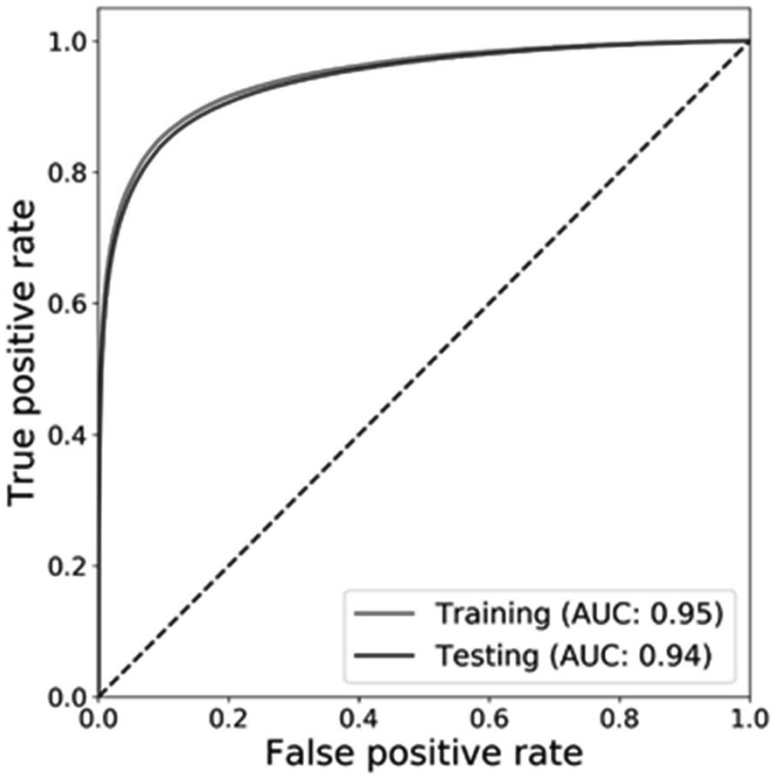

True positive rate
False positive rate
Training (AUC: 0.95)
Testing (AUC: 0.94)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 89

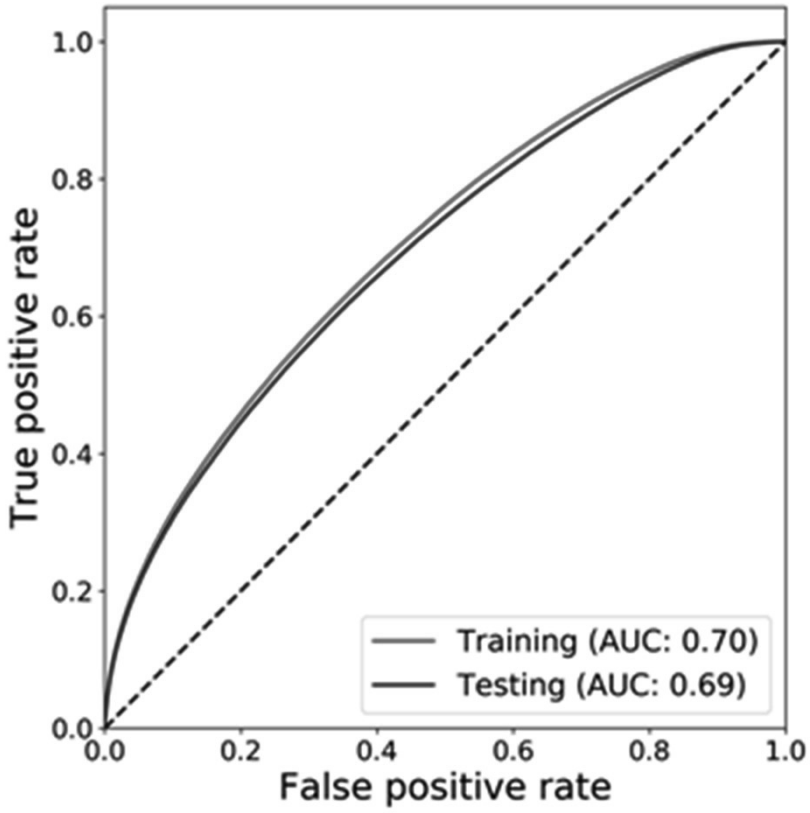

True positive rate
False positive rate
Training (AUC: 0.70)
Testing (AUC: 0.69)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 90

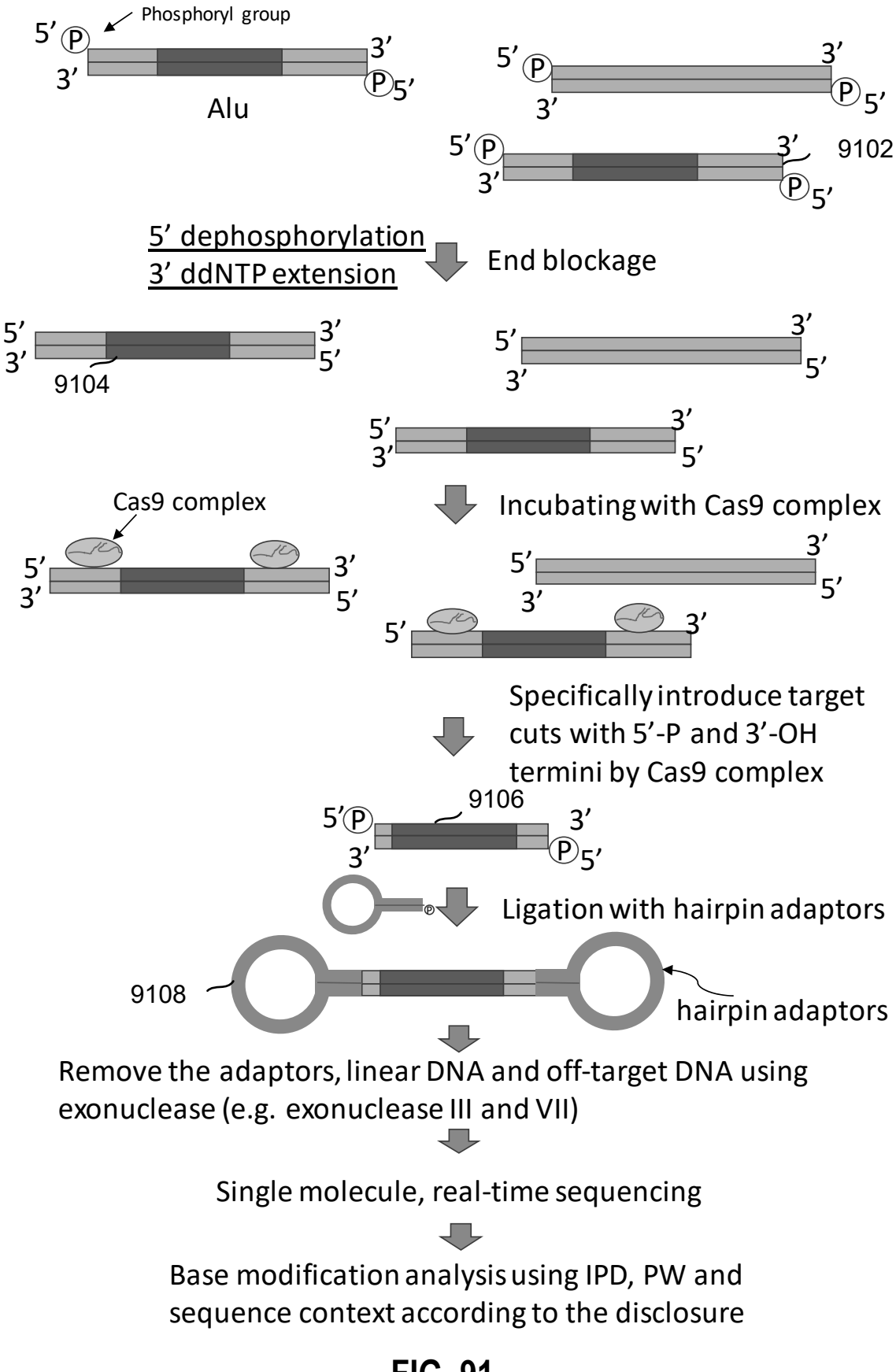

Phosphoryl group
5'
P
3'
Alu
9102
5' dephosphorylation
3' ddNTP extension
End blockage
9104
Cas9 complex
Incubating with Cas9 complex
Specifically introduce target cuts with 5'-P and 3'-OH termini by Cas9 complex
9106
Ligation with hairpin adaptors
9108
hairpin adaptors
Remove the adaptors, linear DNA and off-target DNA using exonuclease (e.g. exonuclease III and VII)
Single molecule, real-time sequencing
Base modification analysis using IPD, PW and sequence context according to the disclosure

FIG. 91

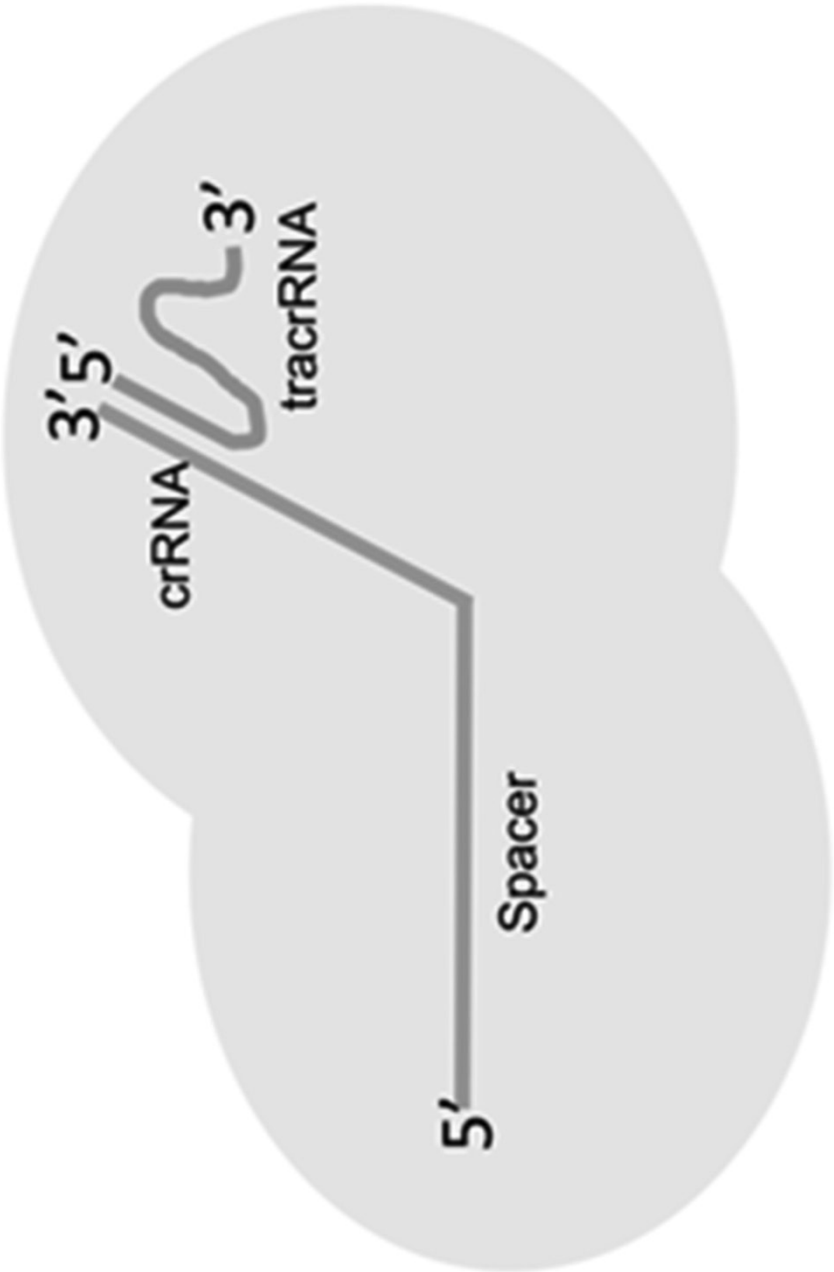
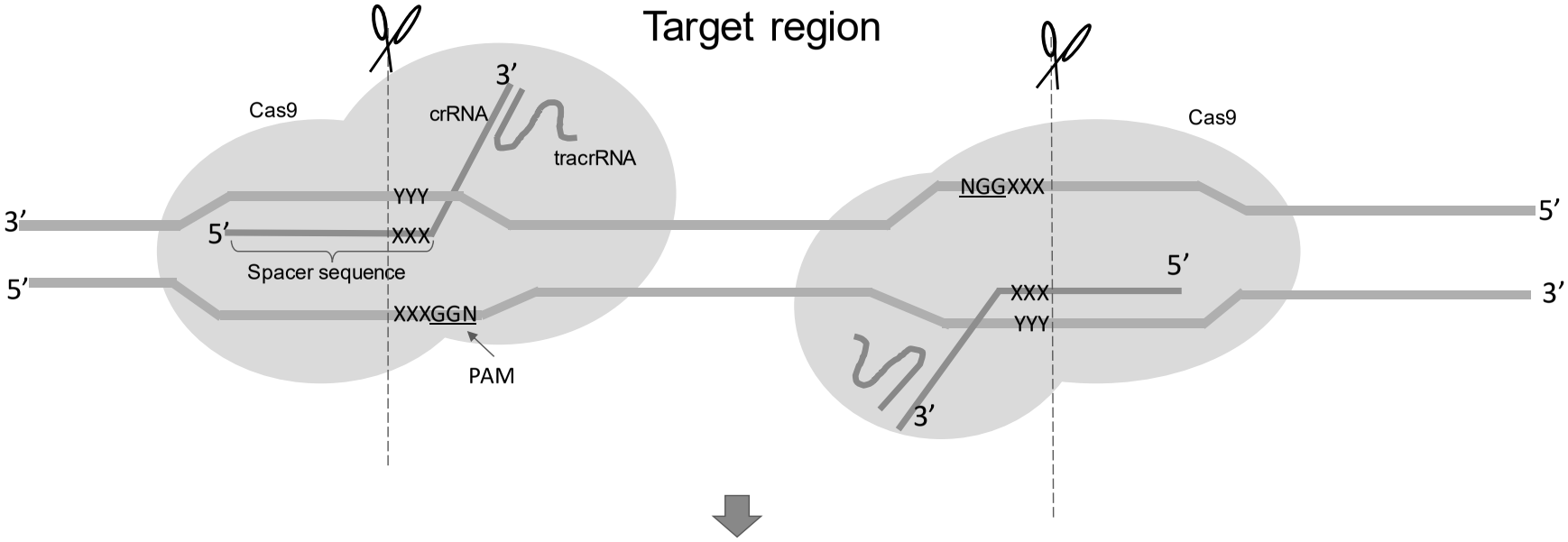


FIG. 92



Obtaining target sequences between two cas9 nuclease cutting sites



FIG. 93

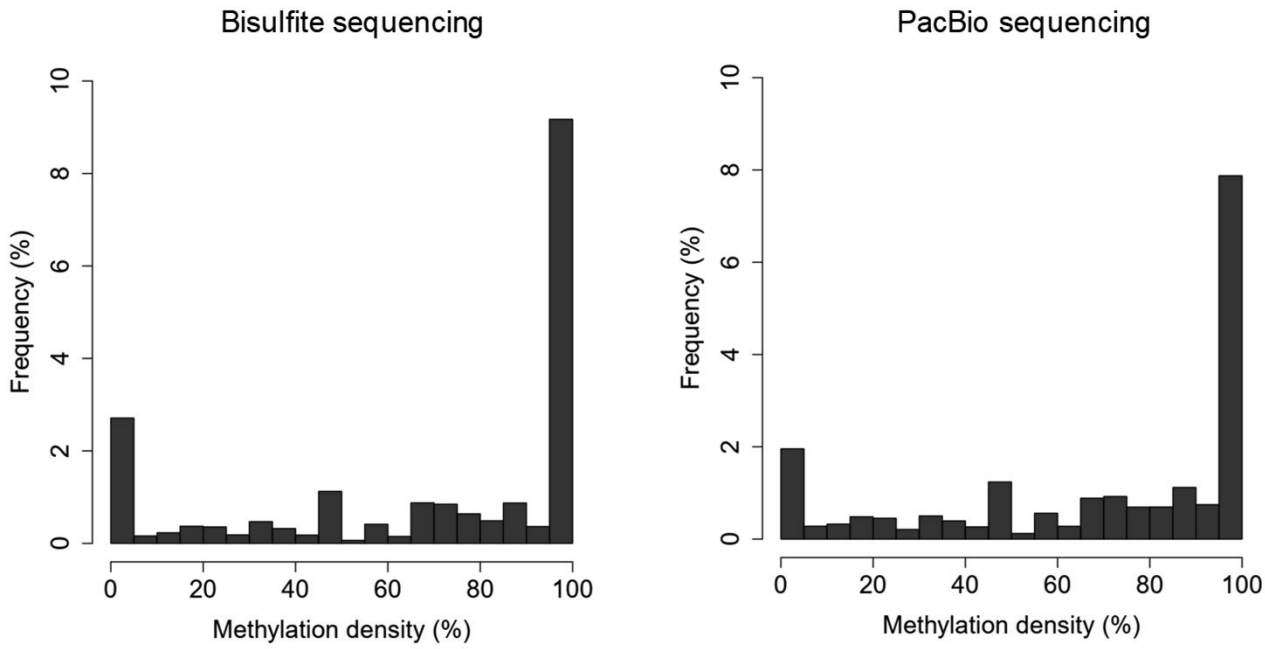

Bisulfite sequencing
PacBio sequencing
Frequency (%)
Methylation density (%)
0
2
4
6
8
10
0
20
40
60
80
100

FIG. 94

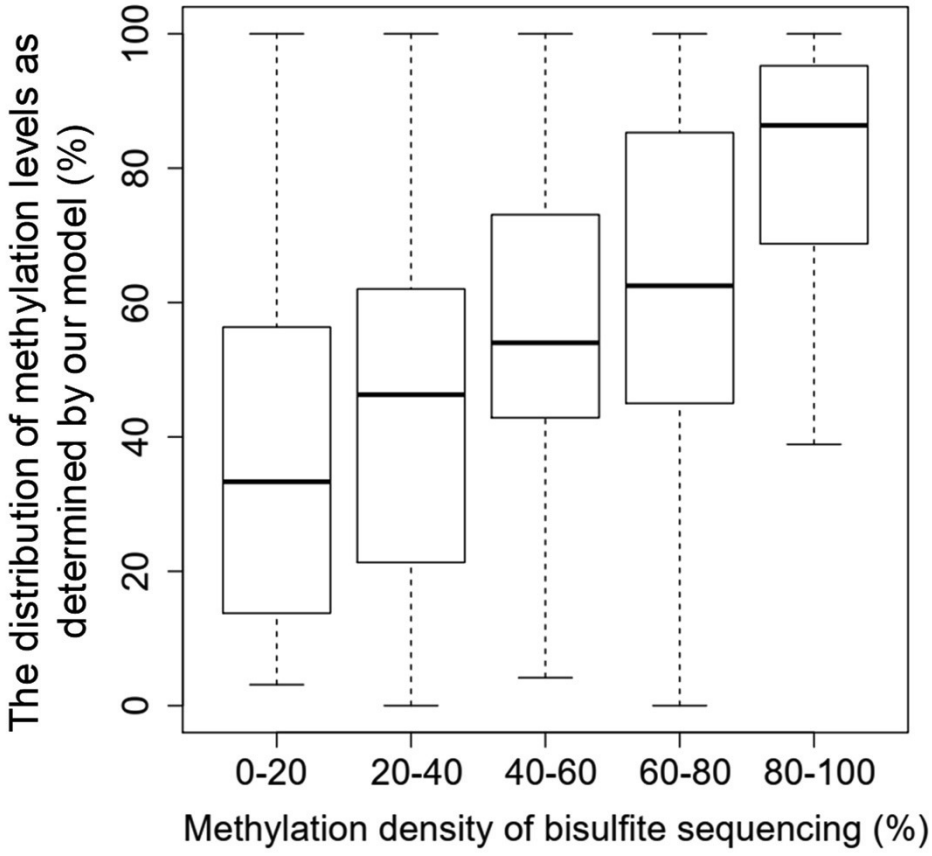

The distribution of methylation levels as determined by our model (%)
Methylation density of bisulfite sequencing (%)
0
20
40
60
80
100
0-20
20-40
40-60
60-80
80-100

FIG. 95

Tissues	Methylation level of Alu (%)
Buffy coat	89.54
Liver	88.18
Colon	89.56
Lung	91.52
Small intestine	86.56
Adrenal gland	89.07
Adipose	91.44
Pancreas	85.82
Brain	91.79
HCC	76.74
Placenta	73.04

FIG. 96

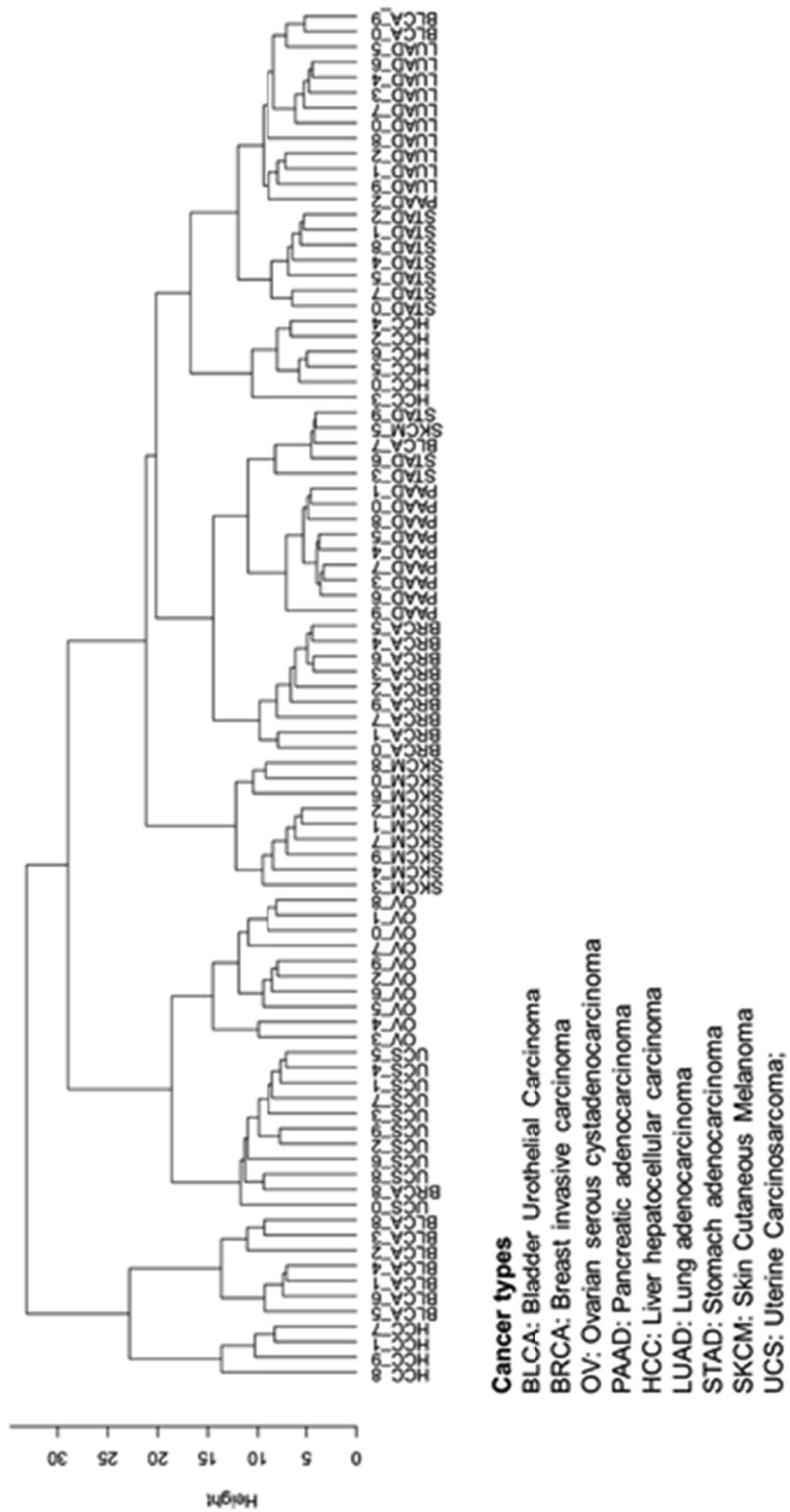

Height
0
5
10
15
20
25
30
Cancer types
BLCA: Bladder Urothelial Carcinoma
BRCA: Breast invasive carcinoma
OV: Ovarian serous cystadenocarcinoma
PAAD: Pancreatic adenocarcinoma
HCC: Liver hepatocellular carcinoma
LUAD: Lung adenocarcinoma
STAD: Stomach adenocarcinoma
SKCM: Skin Cutaneous Melanoma
UCS: Uterine Carcinosarcoma;

FIG. 97

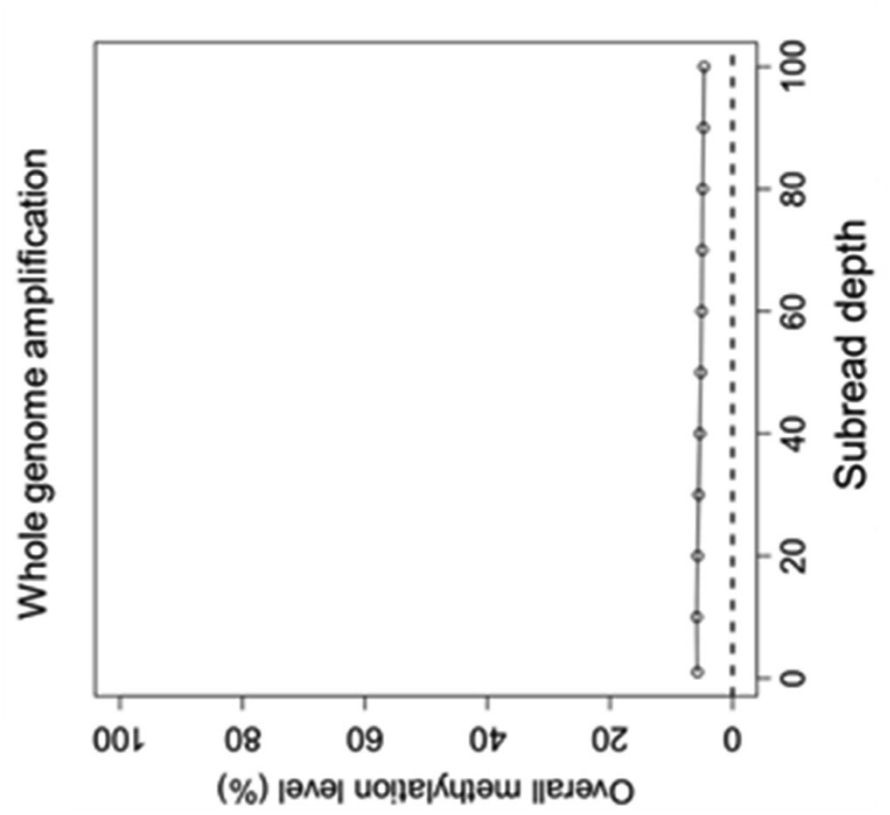


FIG. 98A

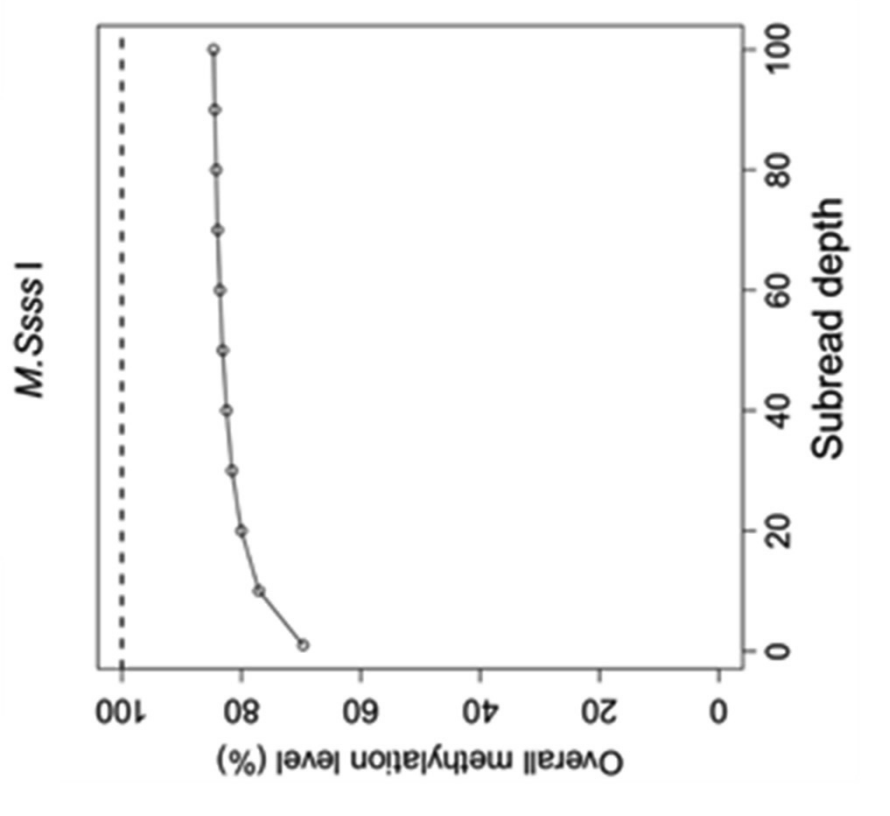


FIG. 98B

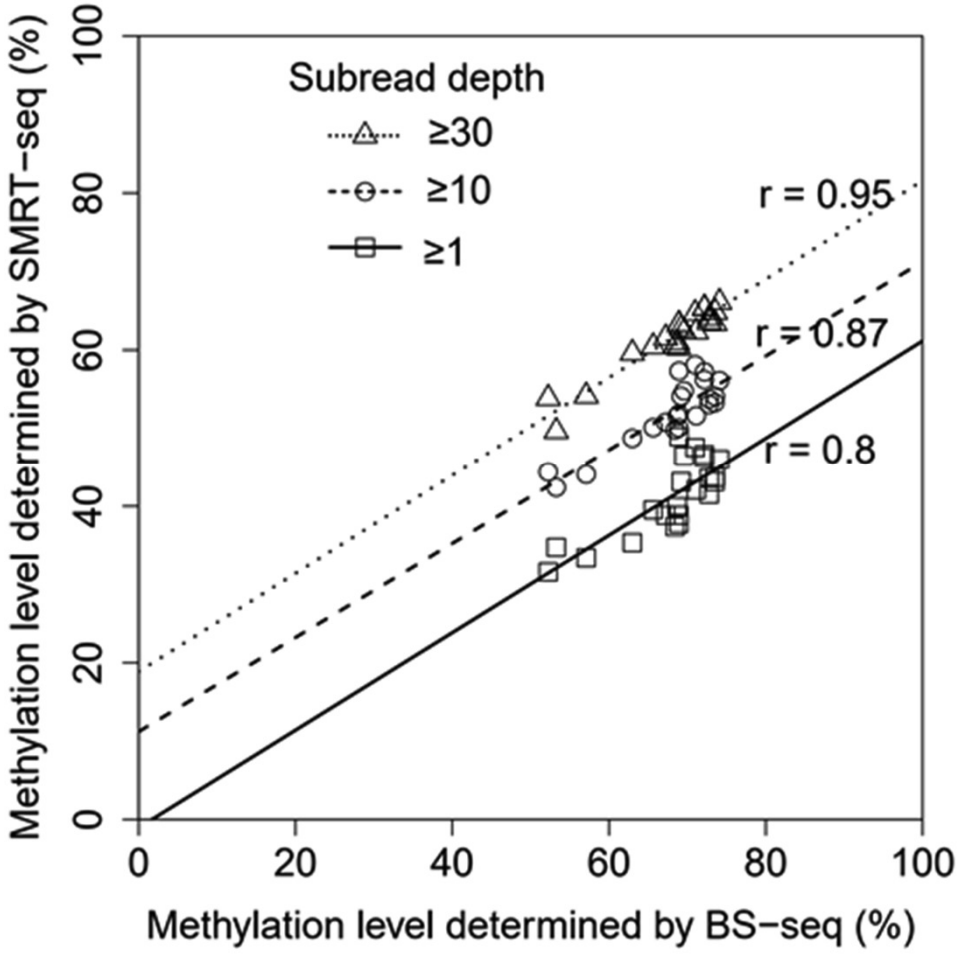

Subread depth
≥30
≥10
≥1
r = 0.95
r = 0.87
r = 0.8
Methylation level determined by SMRT−seq (%)
Methylation level determined by BS−seq (%)
0
20
40
60
80
100

FIG. 99

Subread depth cutoffs ≥	Pearson's r (SMRT-seq vs BS-seq)	No. of CpG sites
1	0.797	25,606,068 (23,949,832-27,008,582)
10	0.873	21,668,418 (18,263,886-23,515,147)
20	0.933	14,276,212 (10,526,406-16,736,887)
30	0.952	6,736,890 (4,255,452-10,449,814)
40	0.948	3,420,790 (2,232,511-5,792,825)
50	0.941	1,684,871 (1,278,475-3,055,876)
60	0.929	911,961 (707,295-1,581,313)
70	0.917	532,422 (350,001-866,045)
80	0.907	284,375 (177,698-534,540)
90	0.906	150,974 (98,000-333,933)
100	0.875	89,788 (58,552-182,861)

FIG. 100

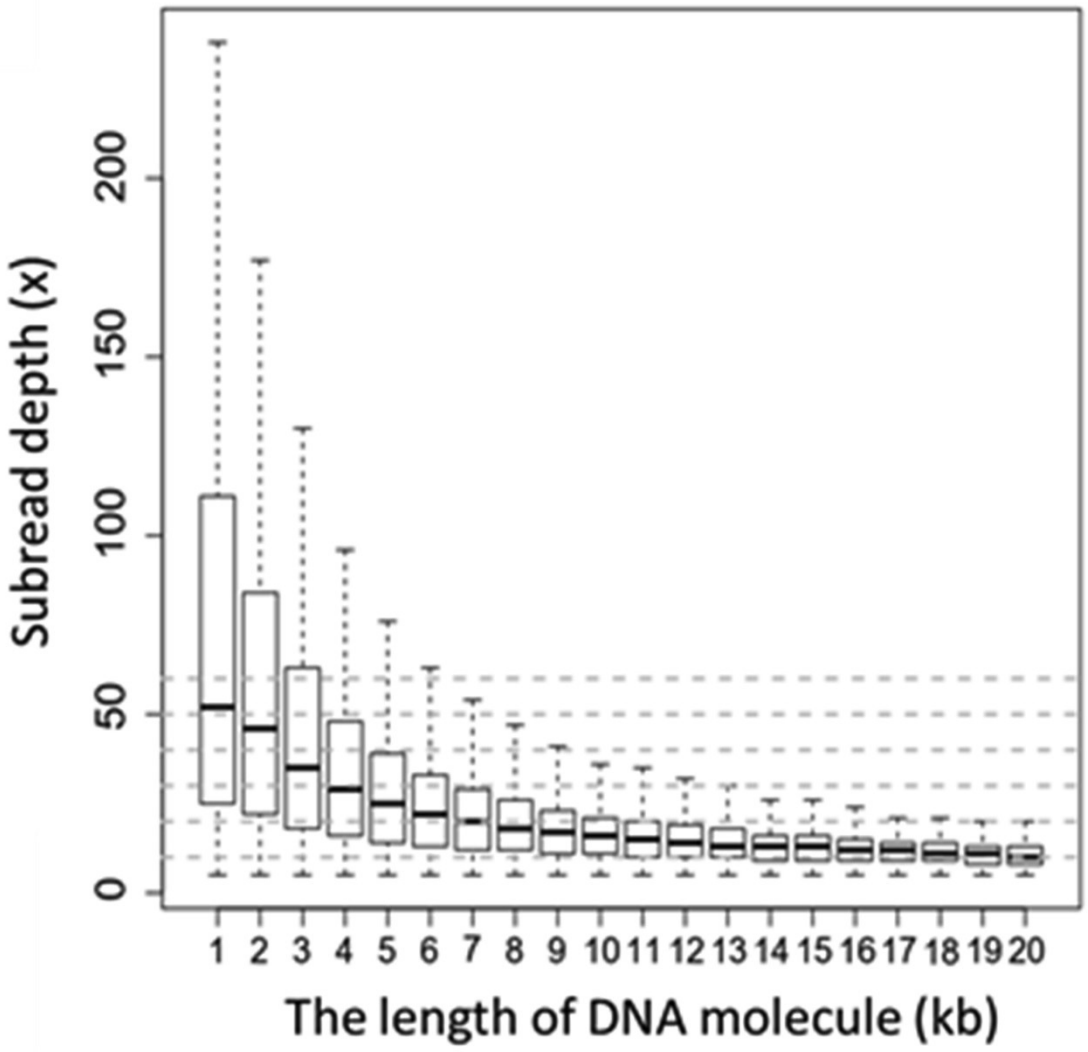

Subread depth (x)
0
50
100
150
200
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
The length of DNA molecule (kb)

FIG. 101

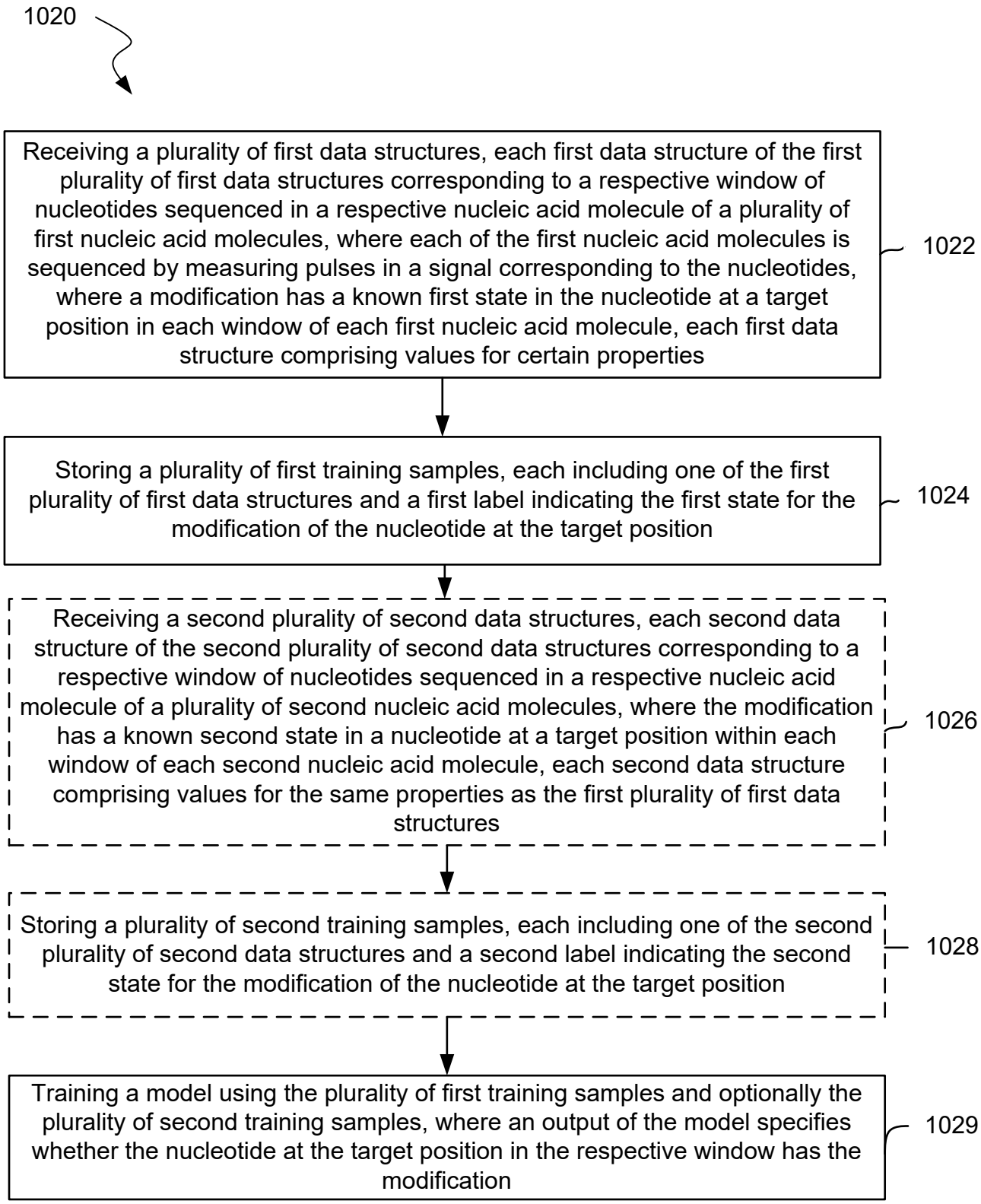

1020
Receiving a plurality of first data structures, each first data structure of the first plurality of first data structures corresponding to a respective window of nucleotides sequenced in a respective nucleic acid molecule of a plurality of first nucleic acid molecules, where each of the first nucleic acid molecules is sequenced by measuring pulses in a signal corresponding to the nucleotides, where a modification has a known first state in the nucleotide at a target position in each window of each first nucleic acid molecule, each first data structure comprising values for certain properties
1022
Storing a plurality of first training samples, each including one of the first plurality of first data structures and a first label indicating the first state for the modification of the nucleotide at the target position
1024
Receiving a second plurality of second data structures, each second data structure of the second plurality of second data structures corresponding to a respective window of nucleotides sequenced in a respective nucleic acid molecule of a plurality of second nucleic acid molecules, where the modification has a known second state in a nucleotide at a target position within each window of each second nucleic acid molecule, each second data structure comprising values for the same properties as the first plurality of first data structures
1026
Storing a plurality of second training samples, each including one of the second plurality of second data structures and a second label indicating the second state for the modification of the nucleotide at the target position
1028
Training a model using the plurality of first training samples and optionally the plurality of second training samples, where an output of the model specifies whether the nucleotide at the target position in the respective window has the modification
1029

FIG. 102

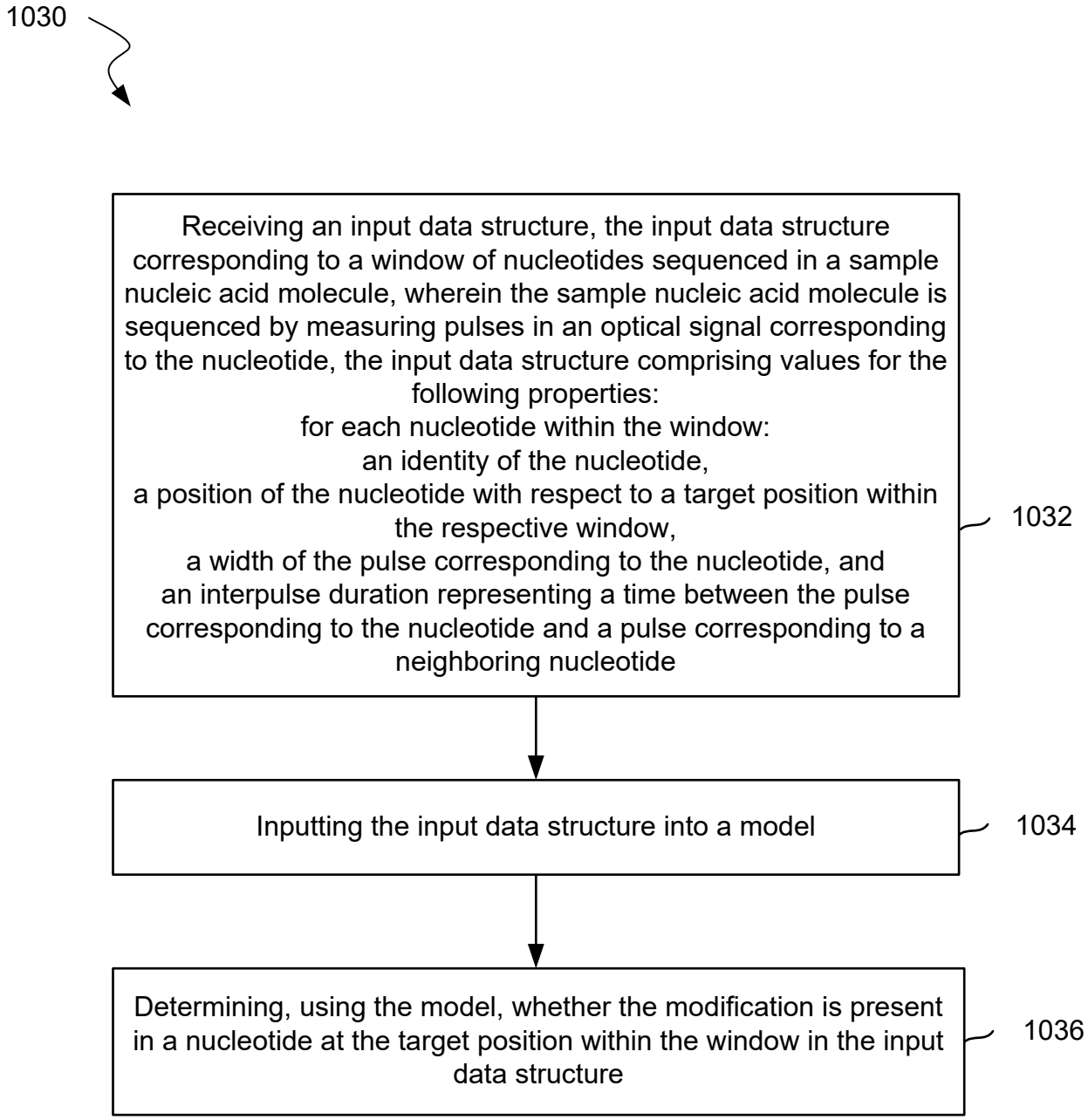
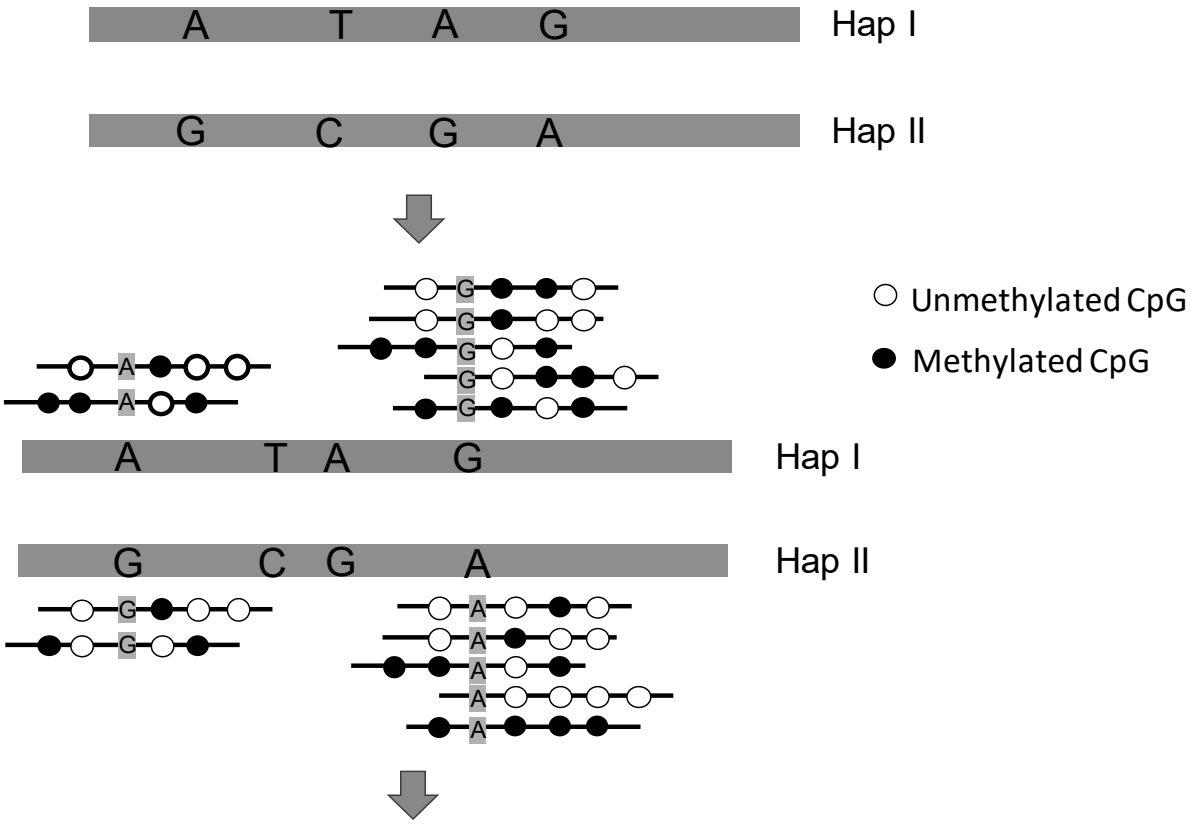

1030
Receiving an input data structure, the input data structure corresponding to a window of nucleotides sequenced in a sample nucleic acid molecule, wherein the sample nucleic acid molecule is sequenced by measuring pulses in an optical signal corresponding to the nucleotide, the input data structure comprising values for the following properties:
for each nucleotide within the window:
an identity of the nucleotide,
a position of the nucleotide with respect to a target position within the respective window,
a width of the pulse corresponding to the nucleotide, and
an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide
1032
Inputting the input data structure into a model
1034
Determining, using the model, whether the modification is present in a nucleotide at the target position within the window in the input data structure
1036

FIG. 103



Relative haplotype-based methylation imbalance index

e.g. Methylation level of Hap I – Methylation level of Hap II

FIG. 104

Chr	start	end	Length	Haplotype block id	PacBio sequencing			
					Methylation level in adjacent non-tumoral tissue		Methylation level in tumor tissue	
					Hap I	Hap II	Hap I	Hap II
chr1	56312395	56347696	35301	hap1927	68.2	67.4	60.3	23.5
chr1	194413819	194424806	10987	hap5953	52.8	49.5	48.8	9.3
chr1	220674478	220699011	24533	hap6863	63.0	64.5	50.4	17.3
chr10	113088792	113124248	35456	hap11838	62.7	63.4	38.1	5.7
chr11	5482746	5498801	16055	hap12904	70.3	75.0	16.3	51.7
chr11	42819351	42852772	33421	hap14385	54.6	54.9	65.3	17.8
chr11	57983961	58051078	67117	hap14930	67.3	66.4	58.2	18.6
chr11	60174708	60204209	29501	hap14990	58.4	59.8	49.6	10.8
chr12	128079419	128114656	35237	hap22249	60.0	58.3	12.1	45.2
chr15	20480575	20533464	52889	hap29631	64.7	69.1	27.7	59.3
chr15	94902853	94946231	43378	hap32161	74.1	74.5	74.9	15.8
chr15	96526684	96549225	22541	hap32221	70.8	68.8	28.9	64.4
chr16	31595372	31613277	17905	hap33499	55.9	59.3	46.3	14.4
chr16	80151778	80182097	30319	hap34821	71.1	71.0	11.5	51.8
chr16	82519715	82554191	34476	hap34920	71.3	66.5	47.4	13.0
chr17	21668593	21685572	16979	hap36049	50.3	47.8	67.4	19.6
chr17	44999177	45012087	12910	hap36640	47.1	45.2	81.6	35.1
chr17	69911623	69926625	15002	hap37435	67.3	63.0	37.8	5.2
chr18	11441122	11458521	17399	hap38335	65.5	66.8	65.9	22.4
chr18	23405569	23423387	17818	hap38673	66.3	61.7	3.3	48.1
chr18	68887284	68925031	37747	hap40390	63.0	61.0	22.0	53.4
chr18	69487809	69505470	17661	hap40414	74.5	74.1	33.3	72.2
chr2	41480394	41514135	33741	hap43972	54.0	54.0	14.9	77.8
chr2	114171214	114182880	11666	hap46226	72.4	68.8	79.7	16.7
chr2	123762541	123797629	35088	hap46589	66.7	68.1	24.0	54.5
chr2	125236882	125241950	5068	hap46673	58.9	59.2	10.7	46.4
chr2	130016110	130040331	24221	hap46835	54.6	50.8	5.6	41.6
chr2	137757638	137783716	26078	hap47090	61.8	61.4	13.5	69.2
chr2	144128597	144160845	32248	hap47343	65.8	66.6	9.3	50.3
chr20	15736792	15753459	16667	hap51505	78.9	74.3	45.8	77.3
chr20	26167979	26177235	9256	hap51868	55.0	52.2	38.5	68.6
chr20	44255808	44264190	8382	hap52246	57.4	56.1	9.7	50.6
chr20	59518410	59559273	40863	hap52761	61.0	62.4	30.0	72.8
chr21	21402034	21424129	22095	hap53197	63.5	67.3	25.0	75.5
chr21	24750027	24768793	18766	hap53333	68.2	64.6	3.4	38.9
chr21	26666833	26701575	34742	hap53418	62.1	66.5	47.6	16.7
chr3	2364024	2387896	23872	hap55539	67.4	67.8	54.9	10.9
chr3	21036965	21049451	12486	hap56223	54.8	51.4	53.1	21.1
chr3	56011690	56046642	34952	hap57346	64.2	61.2	71.2	22.6

FIG. 105A

chr3	73330942	73371216	40274	hap57939	60.9	62.9	9.4	42.9
chr3	106372440	106401301	28861	hap59077	67.8	67.9	13.8	53.2
chr3	107772994	107807482	34488	hap59122	69.6	73.5	30.4	66.4
chr3	116742501	116776747	34246	hap59493	64.3	69.1	14.1	51.6
chr3	171076306	171100102	23796	hap61495	68.0	66.0	80.6	48.8
chr3	193058272	193080344	22072	hap62231	65.5	64.7	54.6	20.0
chr4	30411613	30432317	20704	hap63589	59.3	60.6	53.4	14.6
chr4	31304718	31338193	33475	hap63633	60.2	60.0	7.2	55.0
chr4	92003467	92030505	27038	hap65794	65.3	65.1	54.1	21.7
chr4	155224697	155250915	26218	hap68104	60.5	57.5	57.3	25.0
chr5	2281802	2299281	17479	hap69632	71.5	66.9	69.9	6.6
chr5	4624948	4664704	39756	hap69739	62.8	61.0	14.0	52.0
chr5	89593236	89606080	12844	hap72628	76.6	74.0	20.3	78.4
chr5	119214026	119233058	19032	hap73698	62.8	61.2	57.6	13.1
chr5	119940397	119972658	32261	hap73720	59.1	54.7	53.8	12.2
chr5	132859668	132877415	17747	hap74150	62.5	66.6	59.5	28.3
chr6	26914610	26936918	22308	hap76887	41.9	40.9	71.9	32.6
chr6	66879106	66957243	78137	hap78266	61.6	59.6	25.4	62.0
chr6	77349083	77377529	28446	hap78674	64.5	66.4	27.0	62.9
chr6	159738794	159751033	12239	hap81616	79.6	79.0	21.2	59.8
chr7	26585255	26641907	56652	hap83161	66.2	64.7	49.4	13.3
chr7	48214640	48248036	33396	hap84003	76.0	76.7	78.0	32.3
chr7	88558182	88575482	17300	hap85335	63.8	59.6	63.8	22.9
chr7	96588562	96607580	19018	hap85620	60.4	63.1	19.7	50.0
chr7	122942180	122956897	14717	hap86454	42.3	39.0	19.2	50.0
chr7	132321970	132344802	22832	hap86807	61.4	60.7	52.5	11.5
chr7	153296219	153302441	6222	hap87487	48.7	53.7	64.4	19.3
chr7	156356247	156371897	15650	hap87631	74.9	71.6	87.5	56.6
chr7	159091986	159119486	27500	hap87738	54.0	49.1	52.0	13.2
chr8	51530582	51550889	20307	hap89477	66.4	65.7	68.0	19.9
chr8	63513932	63537543	23611	hap89942	62.0	63.3	11.6	48.4
chr8	72373321	72398122	24801	hap90226	58.0	54.9	71.6	32.0
chr8	94100451	94141855	41404	hap90991	65.2	65.7	36.2	68.7
chr8	109300499	109326404	25905	hap91510	63.6	67.7	29.5	65.8

FIG. 105B

Chr	start	end	Length	Haplotype block id	PacBio sequencing			
					Methylation level in adjacent non-tumoral tissue		Methylation level in tumor tissue	
					Hap I	Hap II	Hap I	Hap II
chr9	27803548	27888202	84654	hap58508	64.2	60.9	20.6	75.4
chr6	242149	386636	144487	hap47880	62.3	63.3	77.4	32.2
chr5	28219159	28302858	83699	hap44666	59.3	58.0	16.8	58.2
chr5	18119943	18153743	33800	hap44475	61.6	65.0	53.2	21.7
chr7	24906307	25046195	139888	hap52069	69.3	68.7	44.0	76.2
chr15	27689897	27752573	62676	hap18337	65.9	61.9	64.8	20.5
chr12	42183870	42212433	28563	hap12045	63.5	68.4	19.4	51.2
chr21	9825597	9935752	110155	hap34175	54.3	53.5	60.9	29.1
chr2	118813055	118893366	80311	hap30060	62.6	62.3	77.0	38.6
chr6	90307702	90344869	37167	hap49779	69.1	66.4	84.7	53.9
chr7	107932914	108049376	116462	hap53838	67.2	62.9	43.8	76.4
chr7	137039327	137160933	121606	hap54447	59.5	60.9	22.9	72.0
chr17	21193754	21254930	61176	hap22633	59.2	54.3	69.7	31.6
chr12	11473697	11644714	171017	hap11451	62.8	66.4	35.5	75.9
chr5	129212299	129353349	141050	hap46632	50.9	54.5	45.5	14.0
chr11	93910738	94028887	118149	hap10288	67.6	63.6	36.6	74.2
chr3	131707434	132003636	296202	hap38642	57.8	55.9	17.9	60.2
chr3	43024004	43161785	137781	hap36769	69.1	66.5	46.1	80.2
chr3	190403156	190606658	203502	hap39947	60.9	61.6	36.9	72.7
chr15	40218970	40279780	60810	hap18606	53.4	57.5	79.1	47.4

FIG. 106

Tissue types	No. of haplotype blocks showing methylation imbalance between two haplotypes in tumor tissues	No. of haplotype blocks showing methylation imbalance between two haplotypes in paired adjacent non-tumoral tissues
Colon	92	47
Breast	57	13
Kidney	68	18
Lung	31	21
Prostate	26	19
Stomach	2	0

FIG. 107A

Tissue types	No. of haplotype blocks showing methylation imbalance between two haplotypes in tumor tissues	Tumor Staging information (TNM) available
Breast	18	T2
	57	T3
Kidney	68	T3a
	0	T2

FIG. 107B

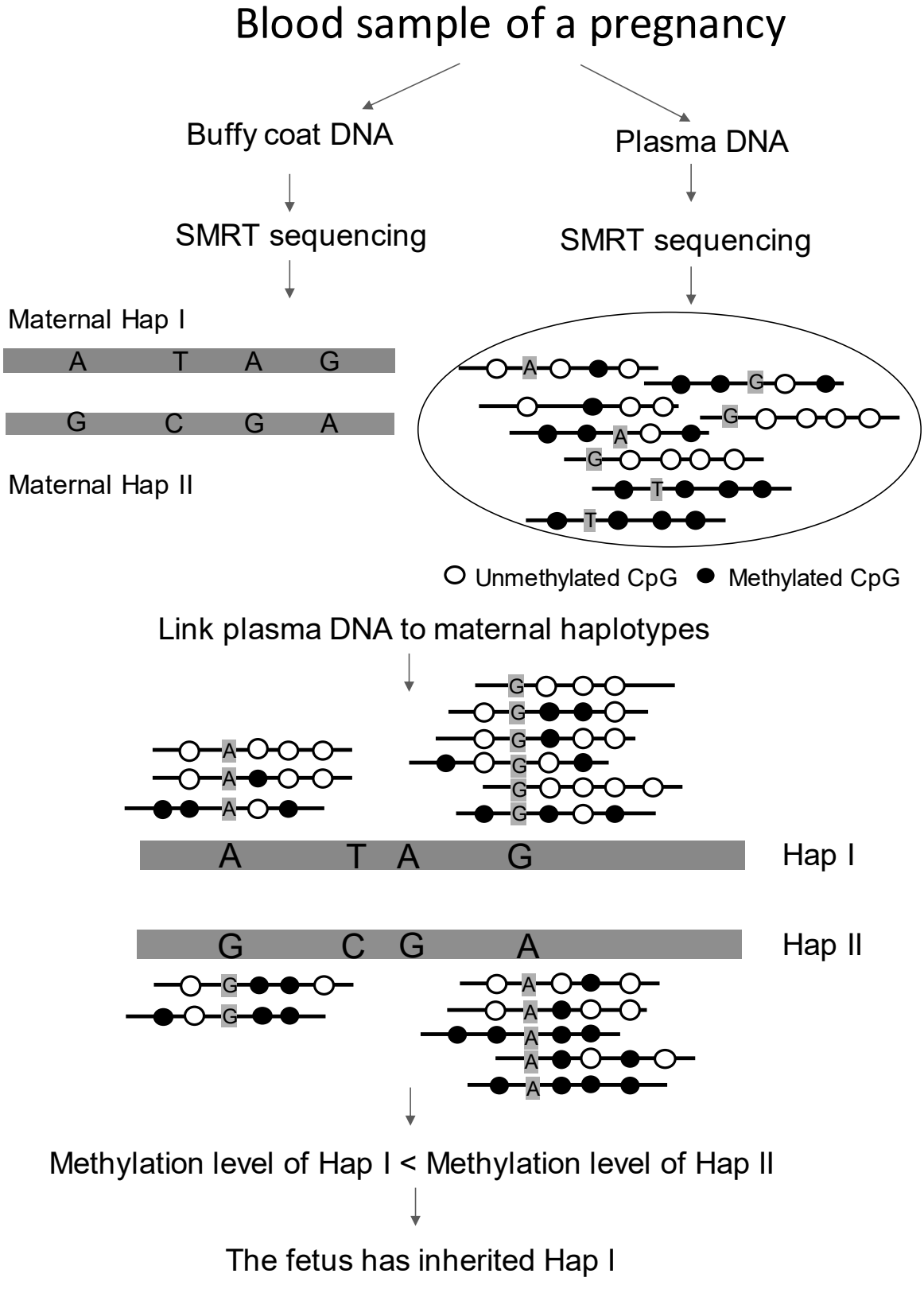

Blood sample of a pregnancy
Buffy coat DNA
Plasma DNA
SMRT sequencing
SMRT sequencing
Maternal Hap I
A T A G
G C G A
Maternal Hap II
Unmethylated CpG
Methylated CpG
Link plasma DNA to maternal haplotypes
A T A G
Hap I
G C G A
Hap II
Methylation level of Hap I < Methylation level of Hap II
The fetus has inherited Hap I

FIG. 108

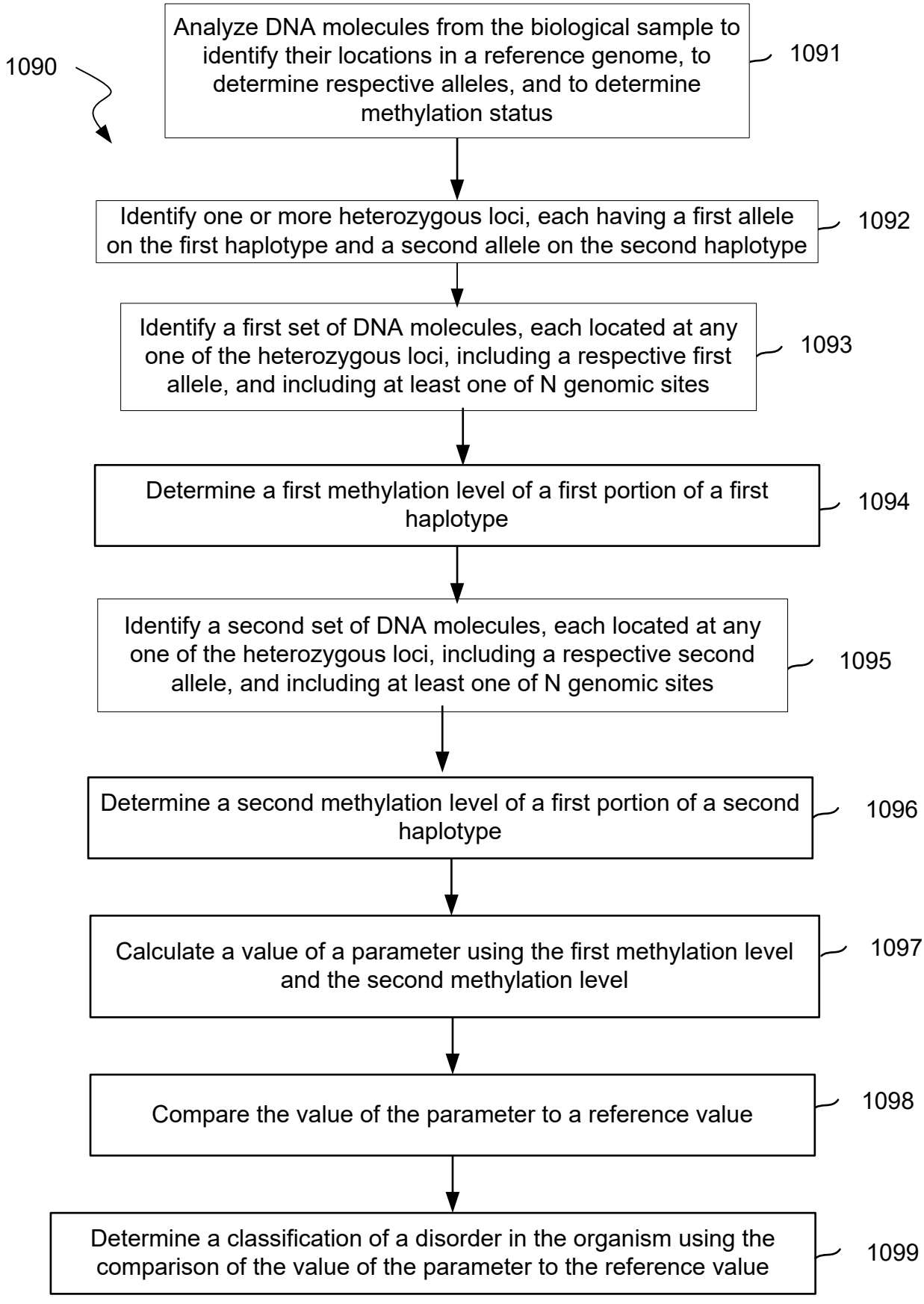

1090
Analyze DNA molecules from the biological sample to identify their locations in a reference genome, to determine respective alleles, and to determine methylation status
1091
Identify one or more heterozygous loci, each having a first allele on the first haplotype and a second allele on the second haplotype
1092
Identify a first set of DNA molecules, each located at any one of the heterozygous loci, including a respective first allele, and including at least one of N genomic sites
1093
Determine a first methylation level of a first portion of a first haplotype
1094
Identify a second set of DNA molecules, each located at any one of the heterozygous loci, including a respective second allele, and including at least one of N genomic sites
1095
Determine a second methylation level of a first portion of a second haplotype
1096
Calculate a value of a parameter using the first methylation level and the second methylation level
1097
Compare the value of the parameter to a reference value
1098
Determine a classification of a disorder in the organism using the comparison of the value of the parameter to the reference value
1099

FIG. 109

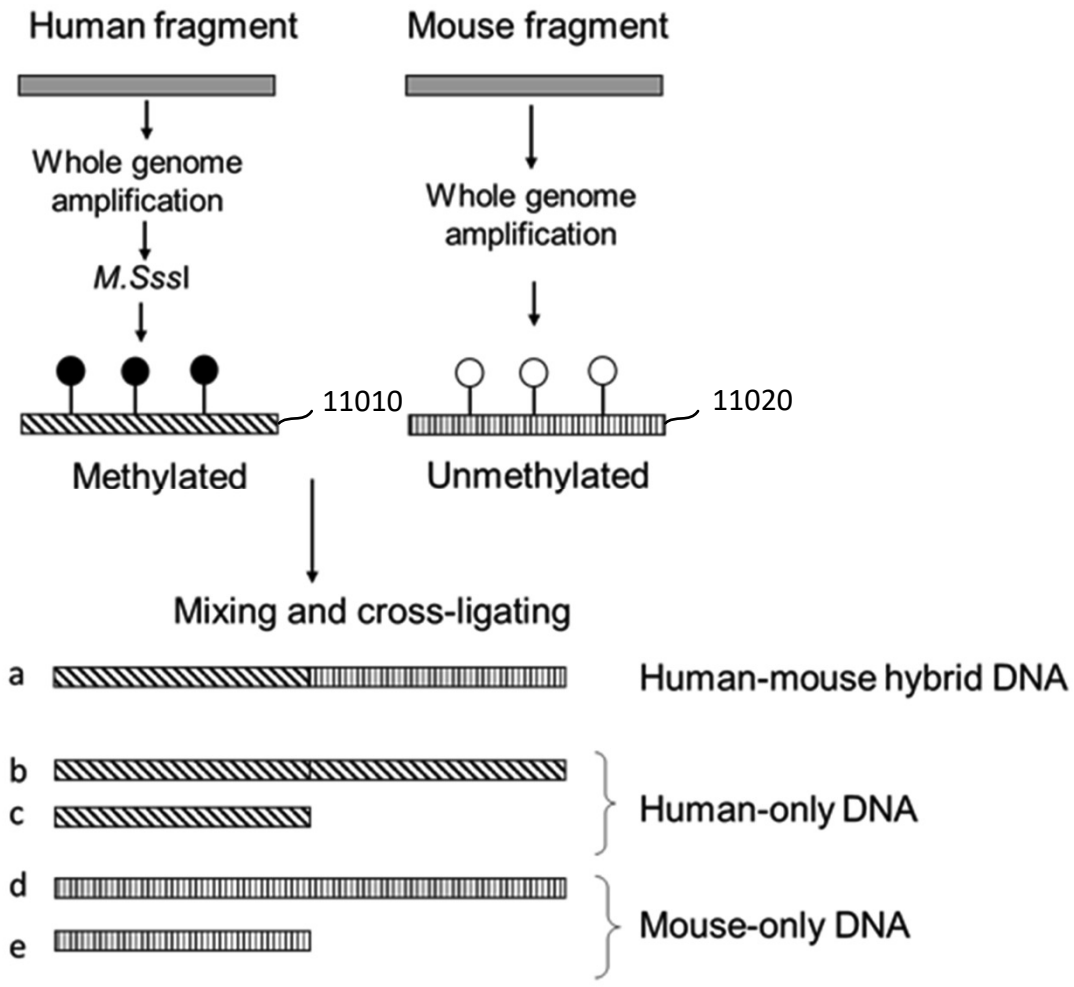

Human fragment
Mouse fragment
Whole genome amplification
M.SssI
Whole genome amplification
11010
11020
Methylated
Unmethylated
Mixing and cross-ligating
a
Human-mouse hybrid DNA
b
c
Human-only DNA
d
e
Mouse-only DNA

FIG. 110

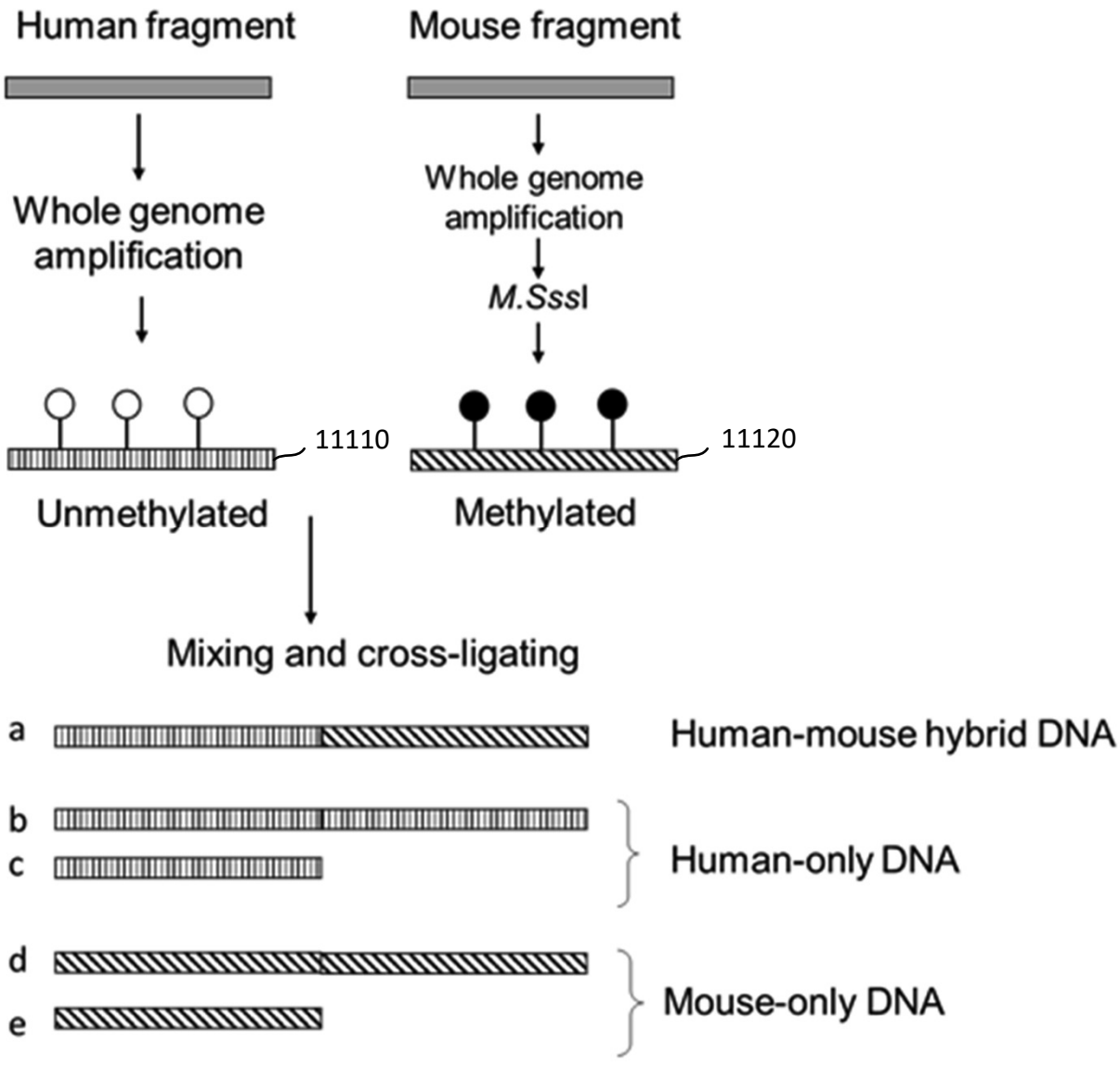

Human fragment
Mouse fragment
Whole genome amplification
Whole genome amplification
M.SssI
11110
11120
Unmethylated
Methylated
Mixing and cross-ligating
a
Human-mouse hybrid DNA
b
c
Human-only DNA
d
e
Mouse-only DNA

FIG. 111

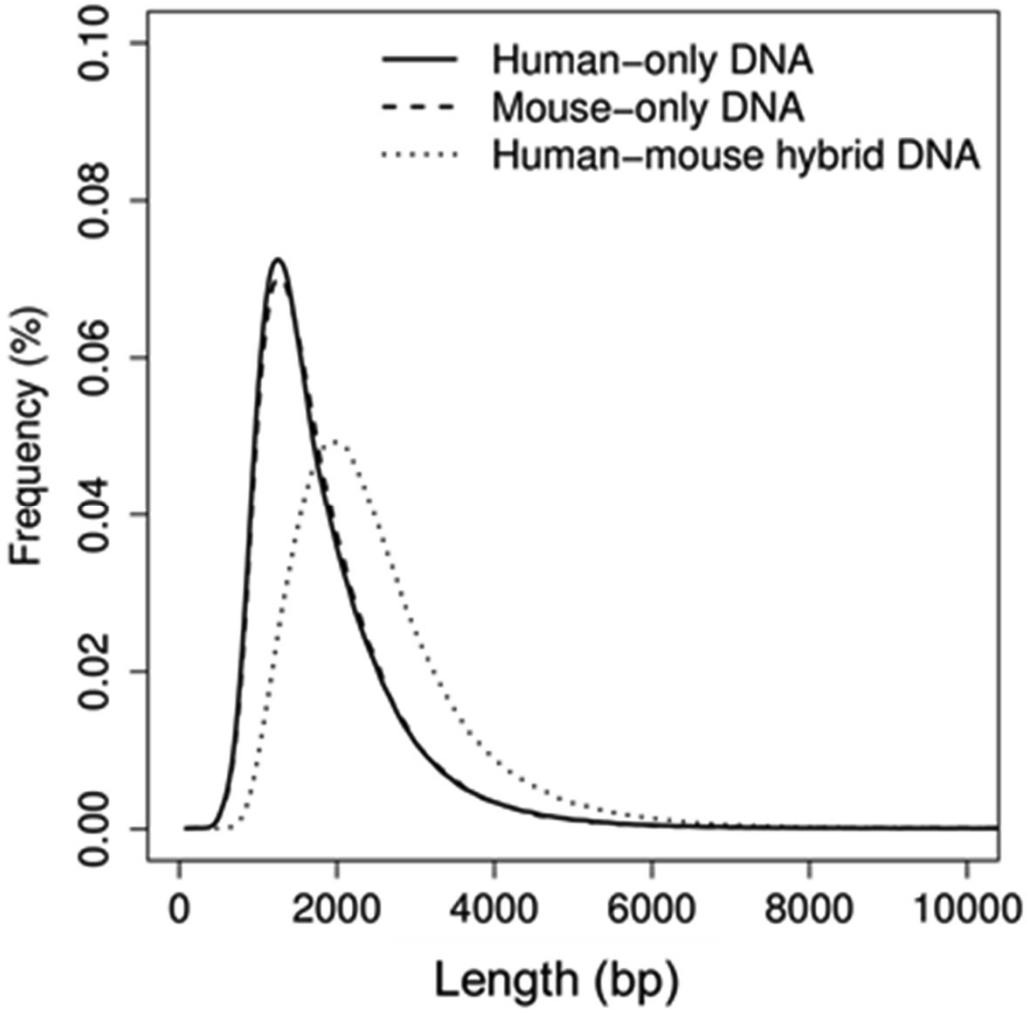

Human–only DNA
Mouse–only DNA
Human–mouse hybrid DNA
Frequency (%)
0.00
0.02
0.04
0.06
0.08
0.10
0
2000
4000
6000
8000
10000
Length (bp)

FIG. 112

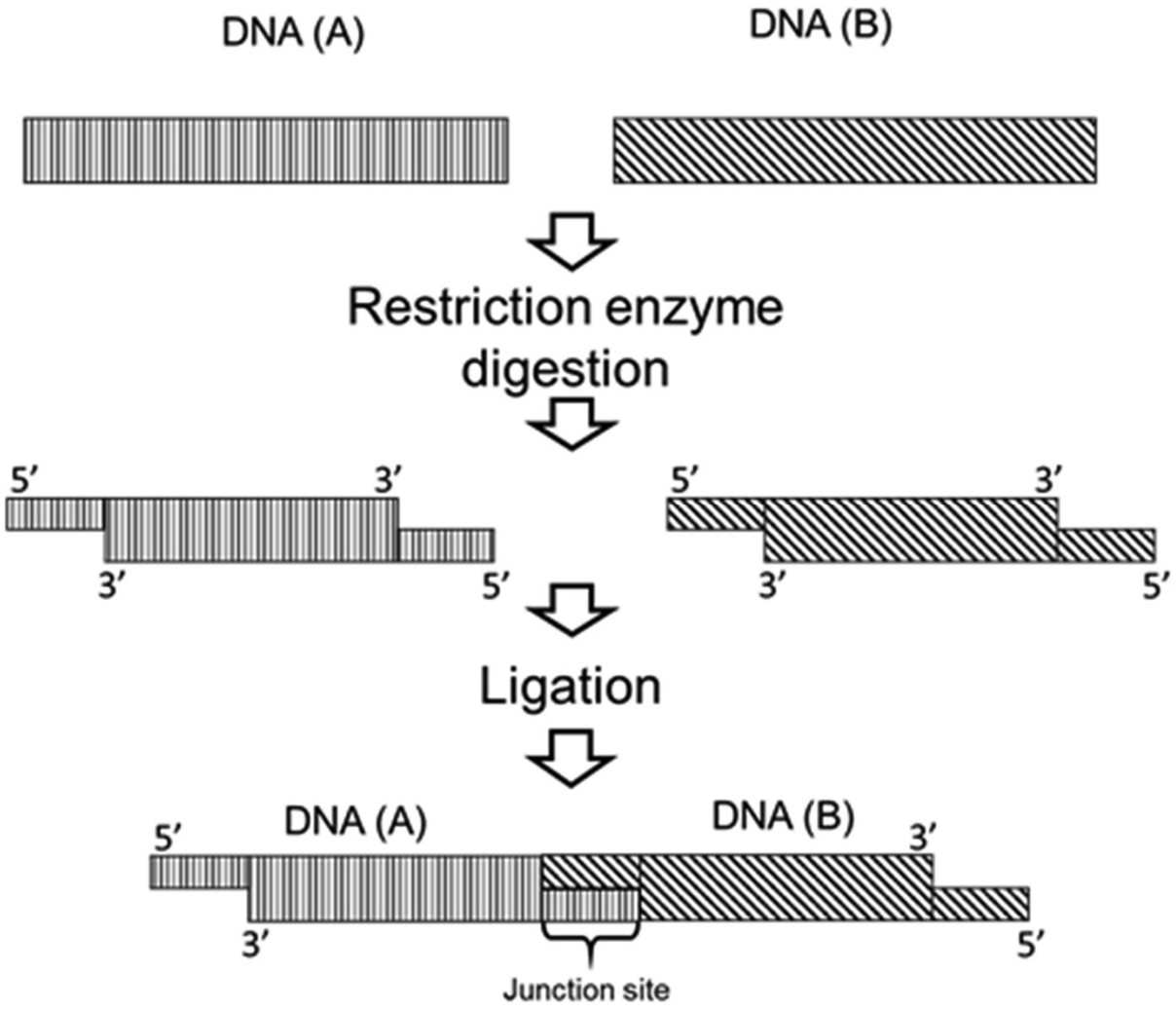

DNA (A)
DNA (B)
Restriction enzyme digestion
5′
3′
3′
5′
5′
3′
3′
5′
Ligation
5′
DNA (A)
DNA (B)
3′
3′
5′
Junction site

FIG. 113

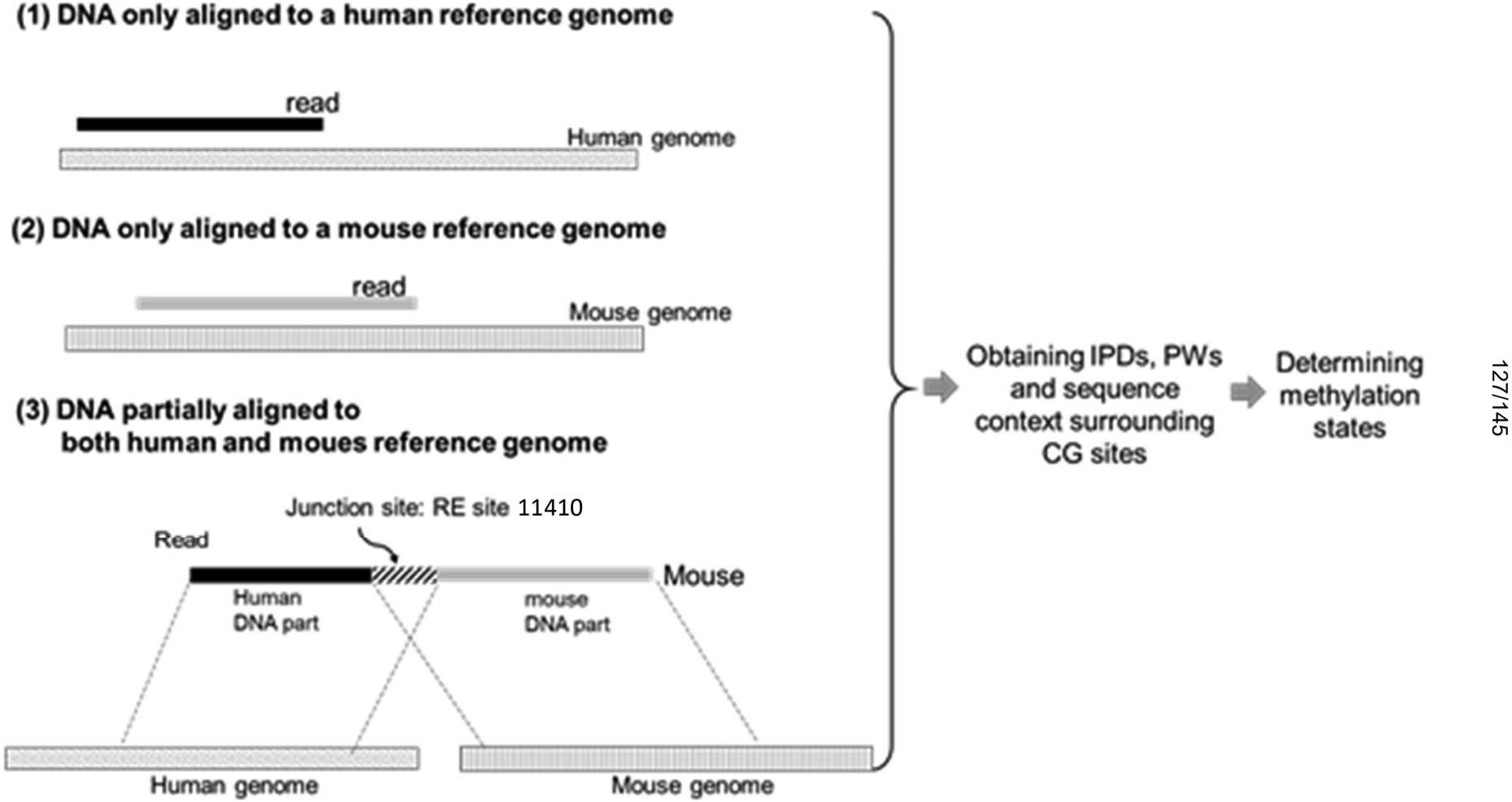

(1) DNA only aligned to a human reference genome
read
Human genome
(2) DNA only aligned to a mouse reference genome
read
Mouse genome
(3) DNA partially aligned to
both human and moues reference genome
Junction site: RE site 11410
Read
Mouse
Human
DNA part
mouse
DNA part
Human genome
Mouse genome
Obtaining IPDs, PWs
and sequence
context surrounding
CG sites
Determining
methylation
states

FIG. 114

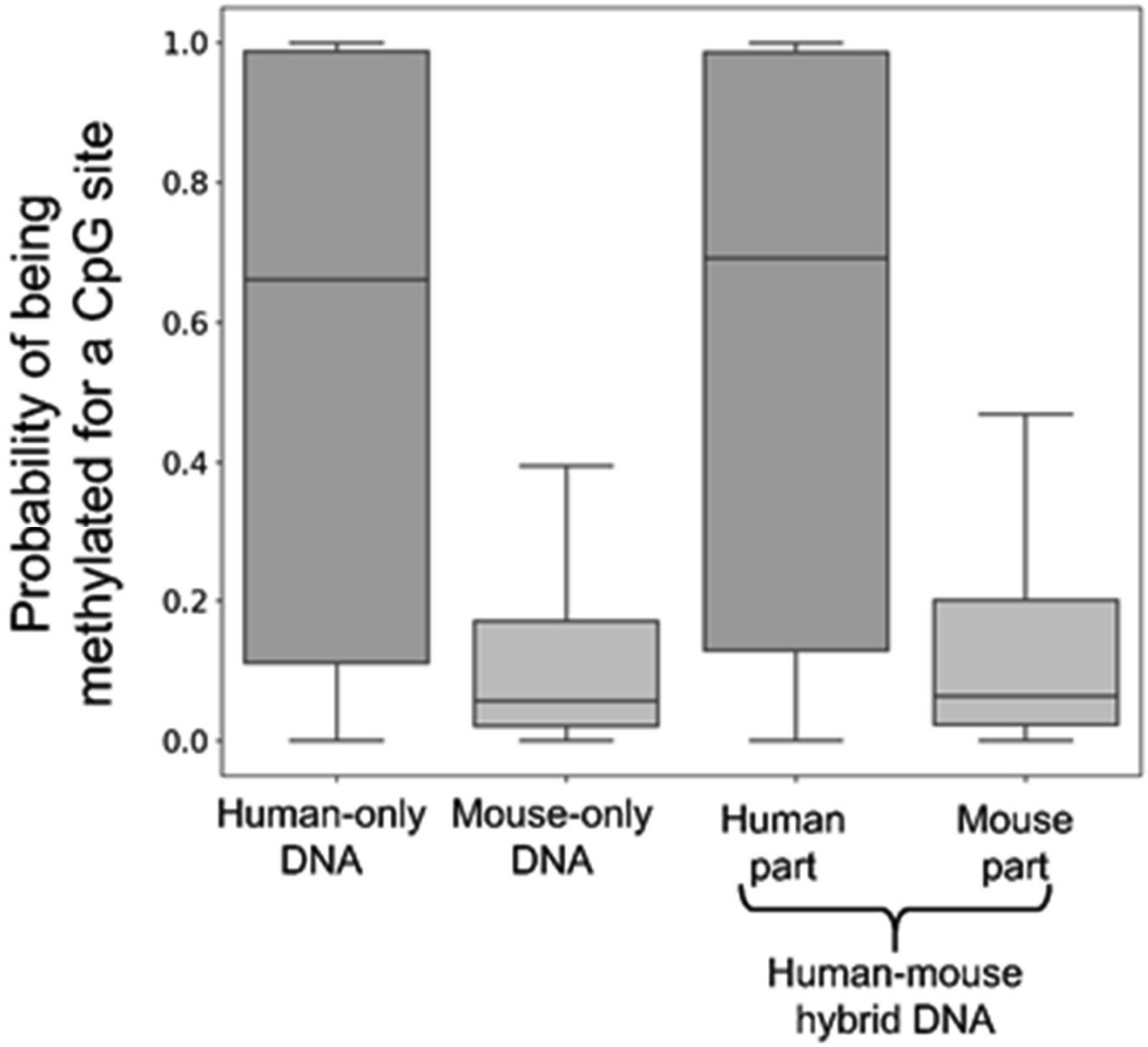

Probability of being methylated for a CpG site
1.0
0.8
0.6
0.4
0.2
0.0
Human-only DNA
Mouse-only DNA
Human part
Mouse part
Human-mouse hybrid DNA

FIG. 115

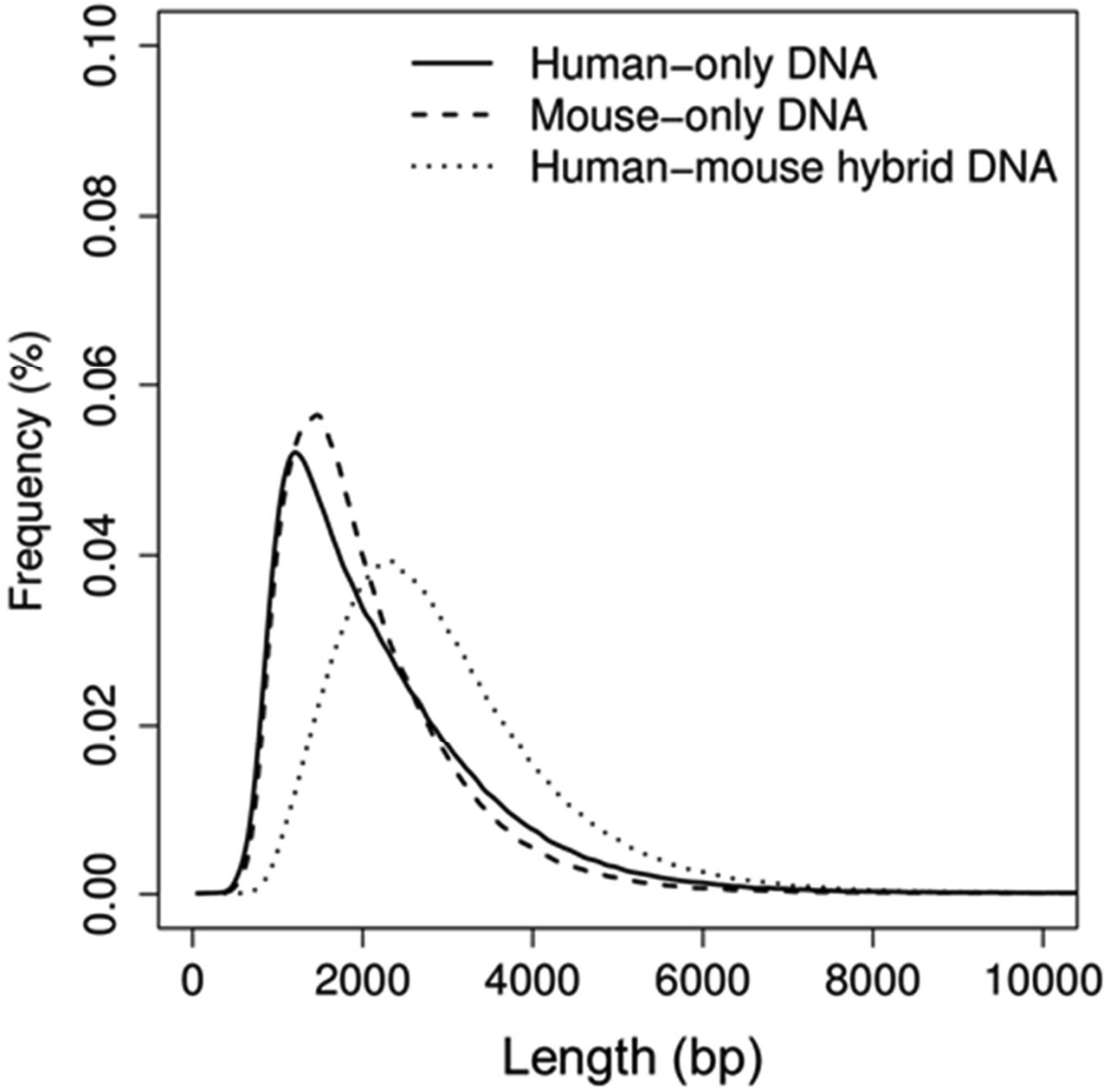

Human–only DNA
Mouse–only DNA
Human–mouse hybrid DNA
Frequency (%)
0.00
0.02
0.04
0.06
0.08
0.10
0
2000
4000
6000
8000
10000
Length (bp)

FIG. 116

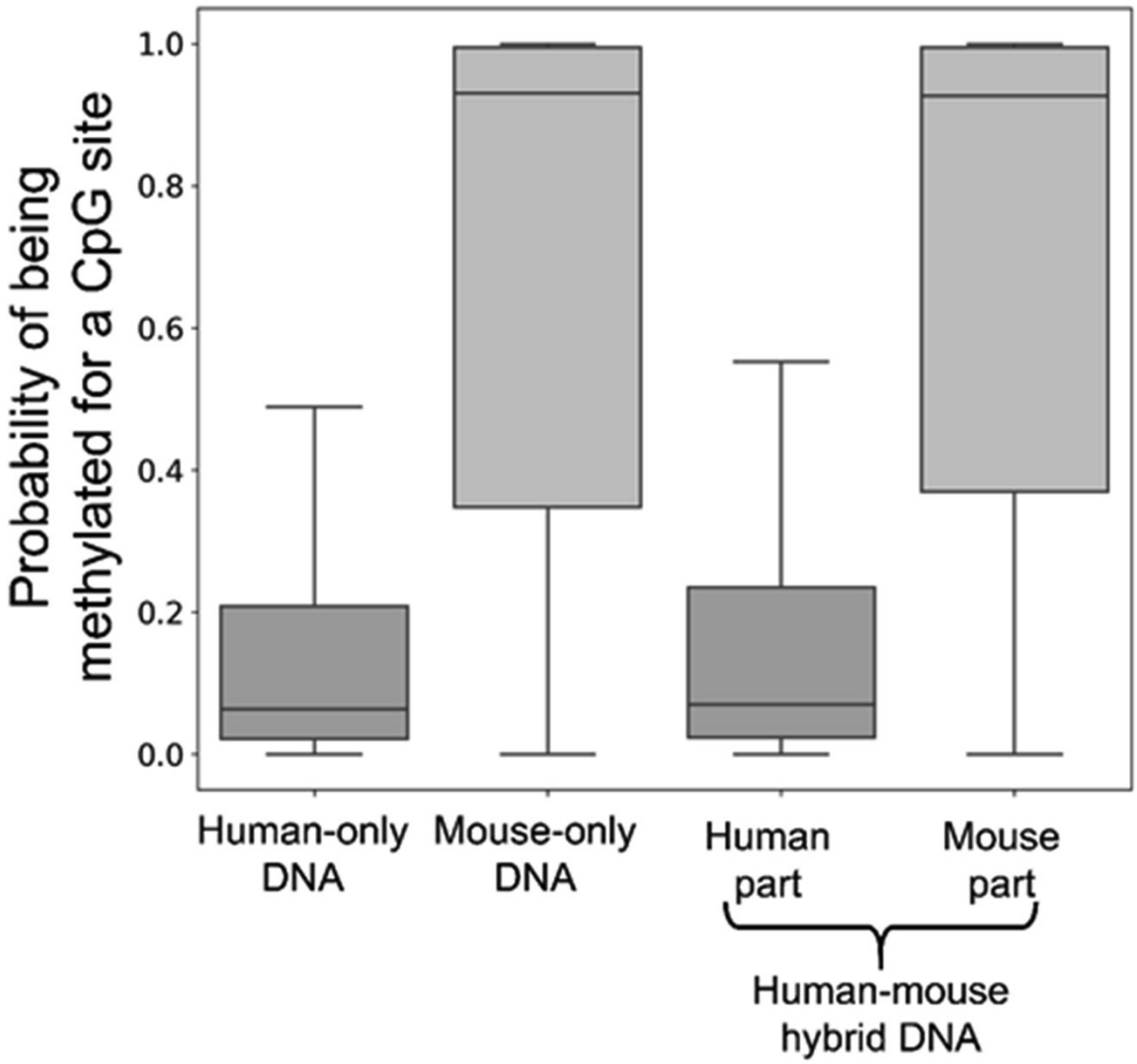

Probability of being methylated for a CpG site
1.0
0.8
0.6
0.4
0.2
0.0
Human-only DNA
Mouse-only DNA
Human part
Mouse part
Human-mouse hybrid DNA

FIG. 117

		Bisulfite sequencing		PacBio sequencing	
		No. of CG sites	Methylation density (%)	No. of CG sites	Methylation density (%)
1) Human-only		2,230,407	41.4	16,226,014	56.0
2) Mouse-only		2,726,499	1.6	9,398,340	10.7
3) Human-mouse hybrid DNA	Human part	73,780	46.8	4,838,454	57.4
	Mouse part	76,312	2.3	4,385,046	12.1

FIG. 118

		Bisulfite sequencing		PacBio sequencing	
		No. of CG sites	Methylation density (%)	No. of CG sites	Methylation density (%)
1) Human-only		2,938,088	1.6	14,503,548	11.6
2) Mouse-only		1,513,971	62.4	11,348,555	71.5
3) Human-mouse hybrid DNA	Human part	67,371	1.8	5,824,379	13.1
	Mouse part	58,242	67.4	5,093,097	72.2

FIG. 119

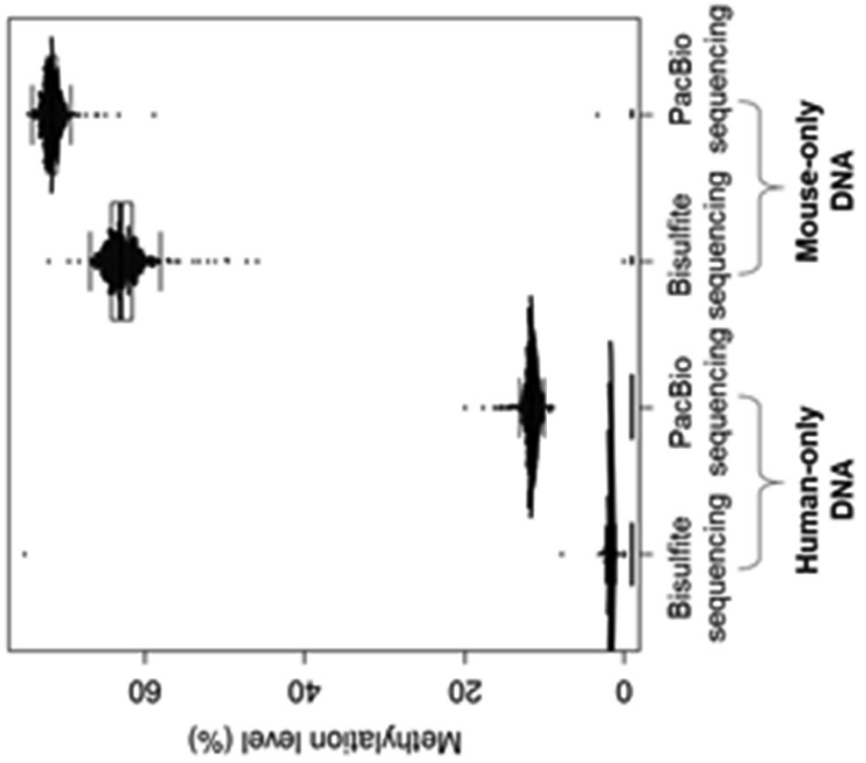


FIG. 120B

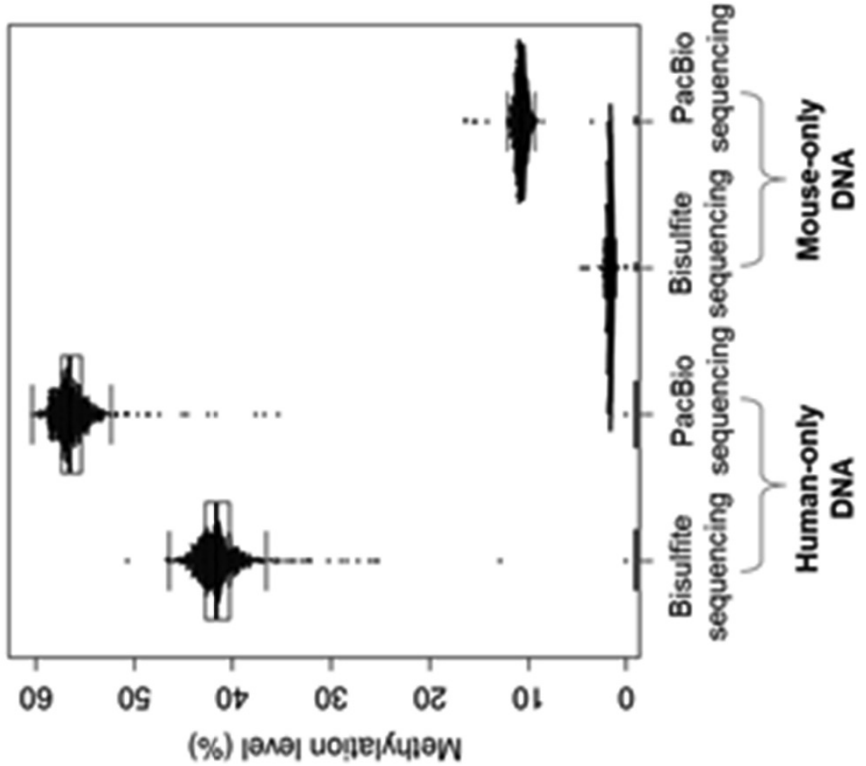


FIG. 120A

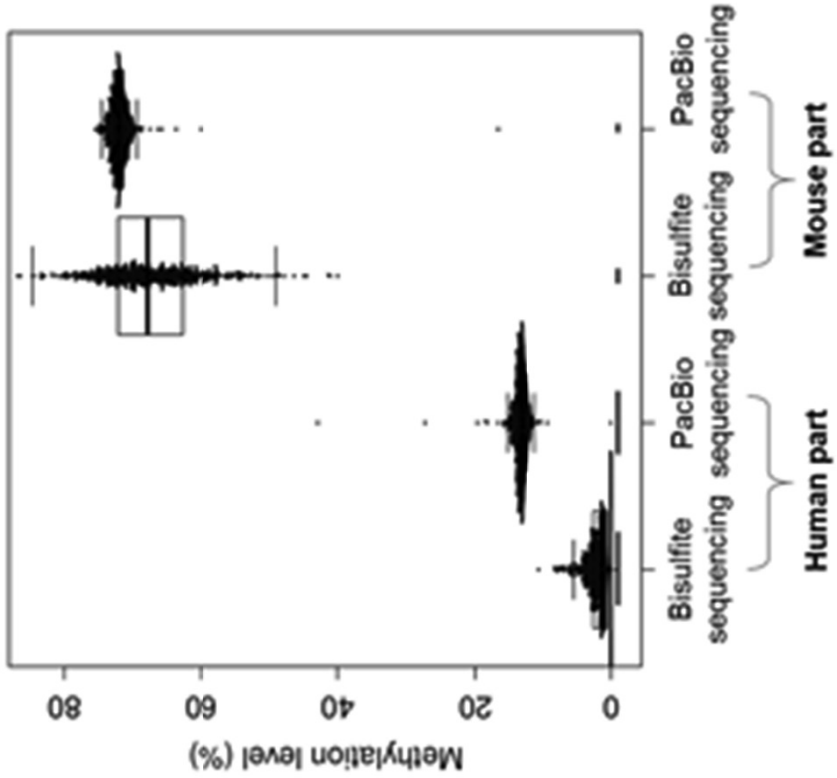


FIG. 121B

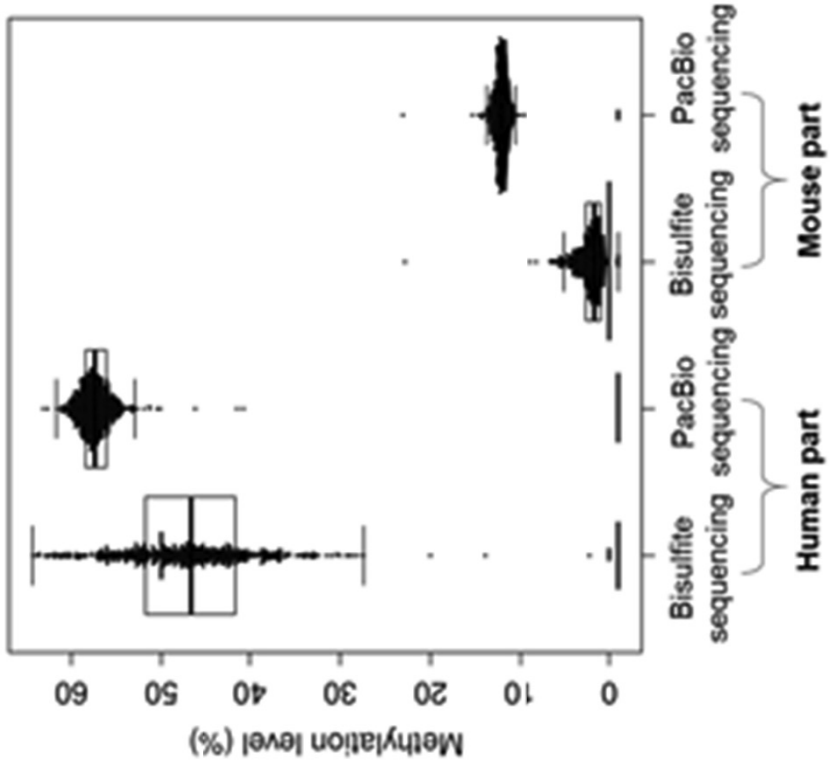


FIG. 121A

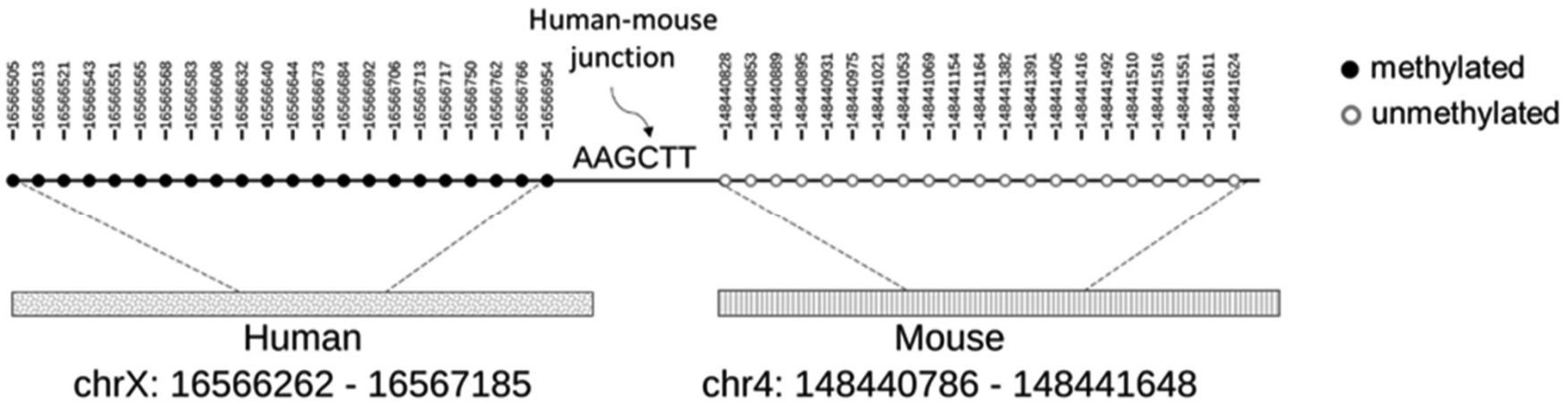


FIG. 122A

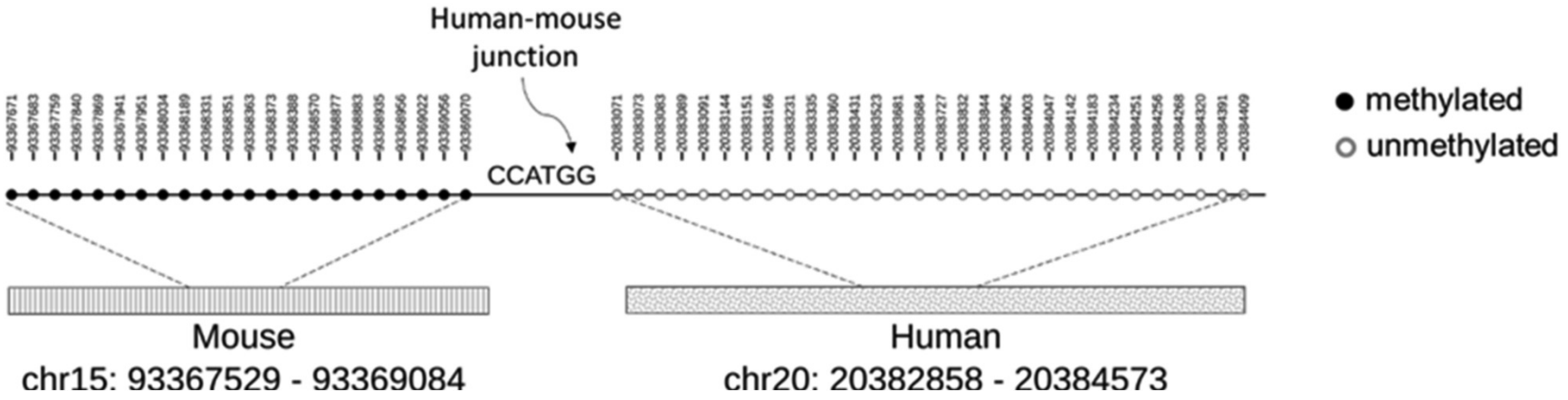


FIG. 122B

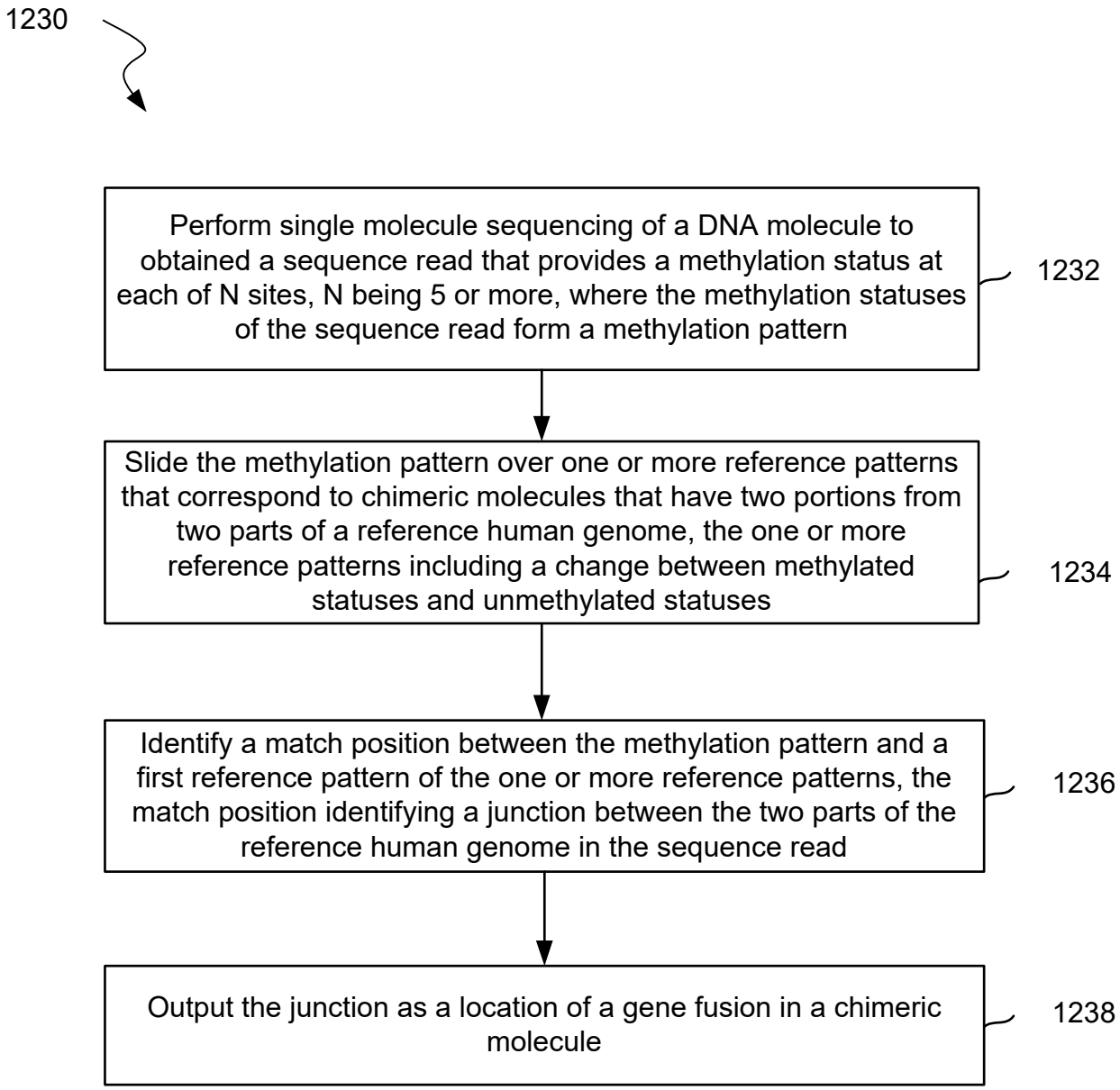

1230
Perform single molecule sequencing of a DNA molecule to obtained a sequence read that provides a methylation status at each of N sites, N being 5 or more, where the methylation statuses of the sequence read form a methylation pattern
1232
Slide the methylation pattern over one or more reference patterns that correspond to chimeric molecules that have two portions from two parts of a reference human genome, the one or more reference patterns including a change between methylated statuses and unmethylated statuses
1234
Identify a match position between the methylation pattern and a first reference pattern of the one or more reference patterns, the match position identifying a junction between the two parts of the reference human genome in the sequence read
1236
Output the junction as a location of a gene fusion in a chimeric molecule
1238

FIG. 123

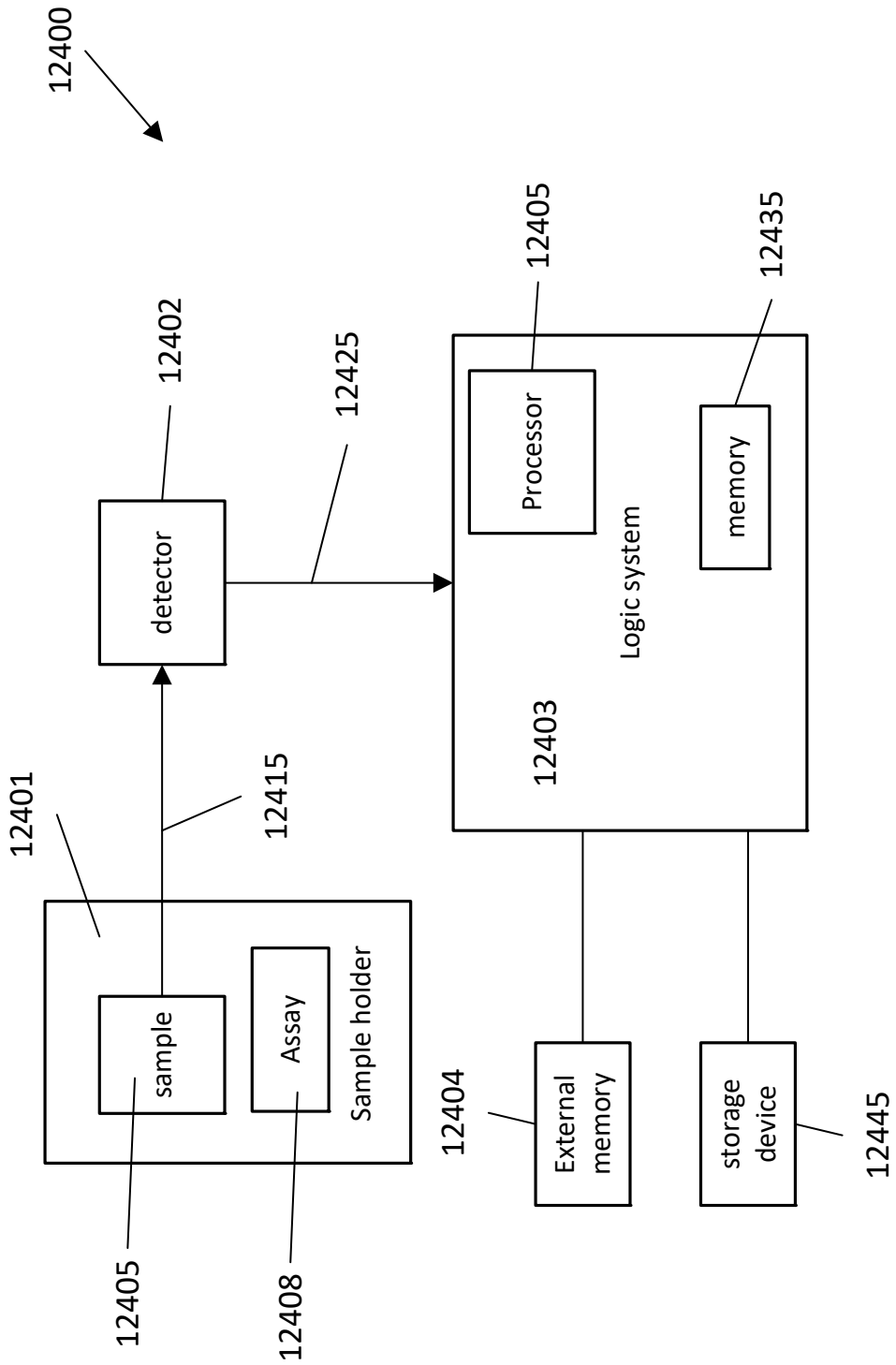

12400
12402
12425
12405
12435
detector
Processor
memory
Logic system
12403
12401
12415
Sample holder
sample
Assay
12405
12408
12404
External memory
storage device
12445

FIG. 124

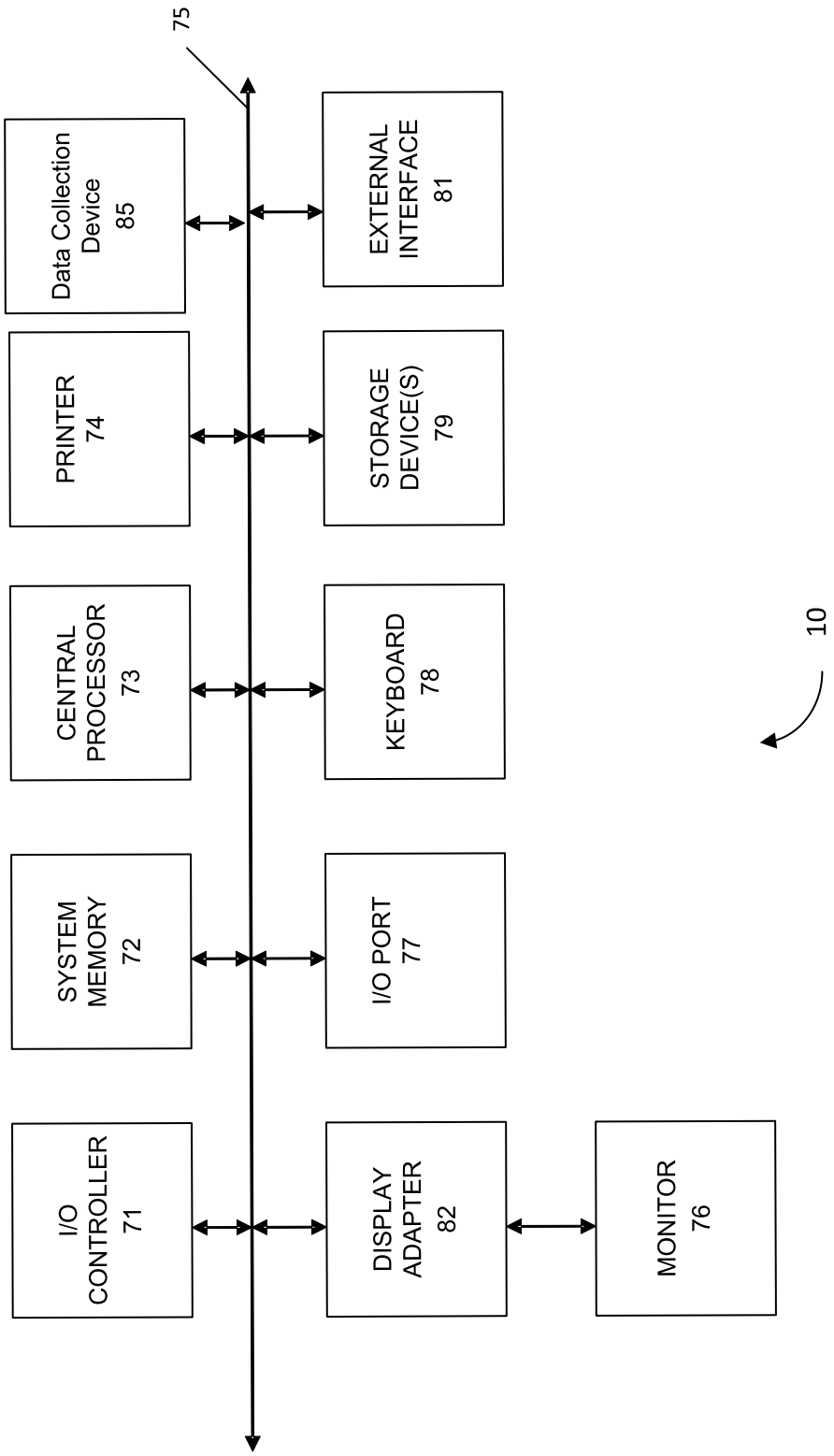

75
Data Collection Device 85
EXTERNAL INTERFACE 81
PRINTER 74
STORAGE DEVICE(S) 79
CENTRAL PROCESSOR 73
KEYBOARD 78
SYSTEM MEMORY 72
I/O PORT 77
I/O CONTROLLER 71
DISPLAY ADAPTER 82
MONITOR 76
10

FIG. 125

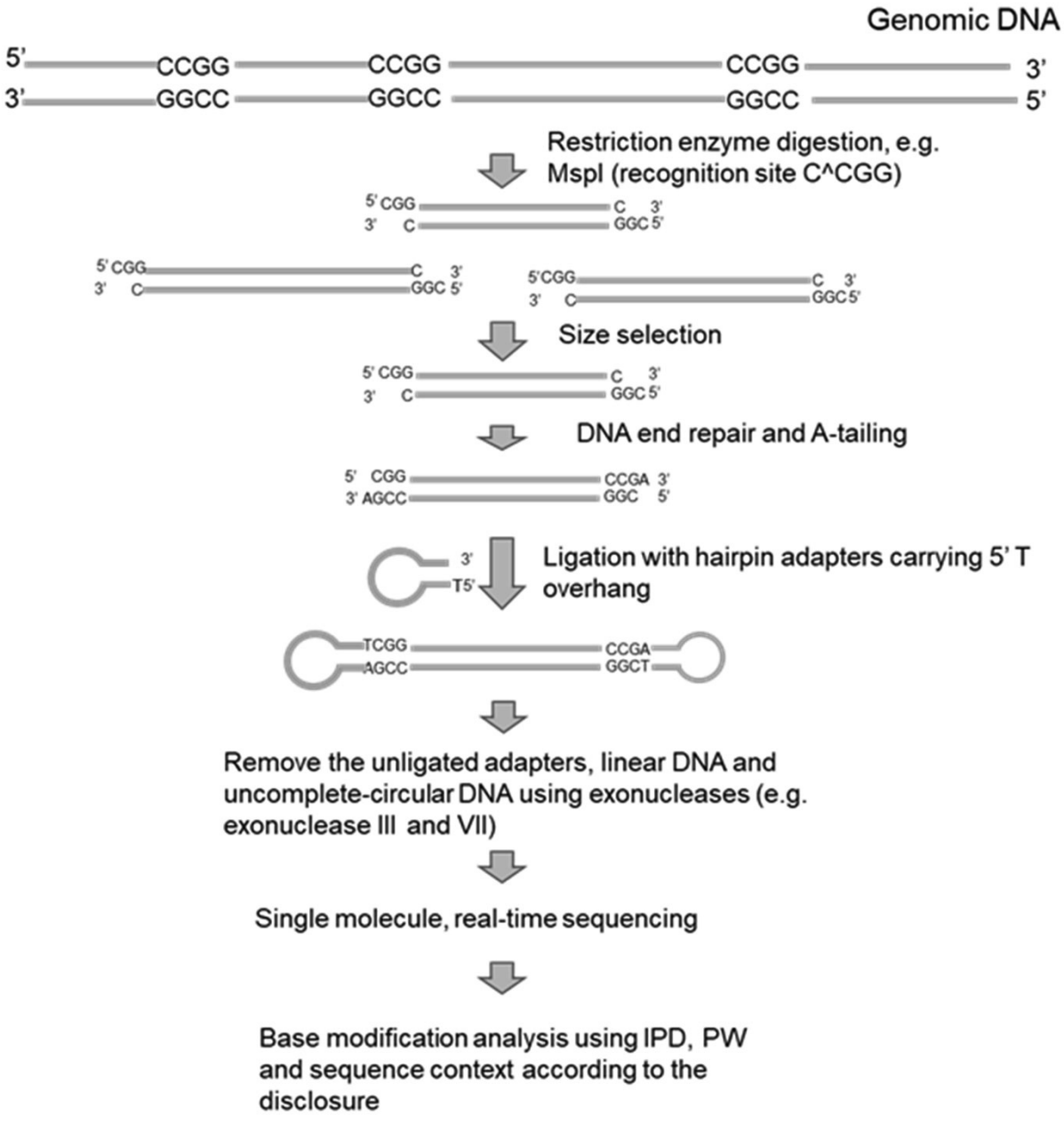

Genomic DNA
5'
CCGG
CCGG
CCGG
3'
3'
GGCC
GGCC
GGCC
5'
Restriction enzyme digestion, e.g. MspI (recognition site C^CGG)
Size selection
DNA end repair and A-tailing
Ligation with hairpin adapters carrying 5' T overhang
Remove the unligated adapters, linear DNA and uncomplete-circular DNA using exonucleases (e.g. exonuclease III and VII)
Single molecule, real-time sequencing
Base modification analysis using IPD, PW and sequence context according to the disclosure

FIG. 126

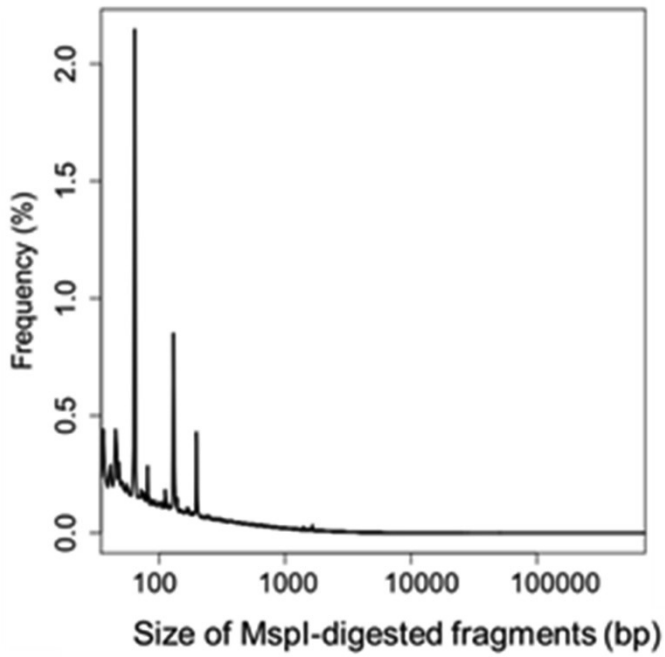


FIG. 127A

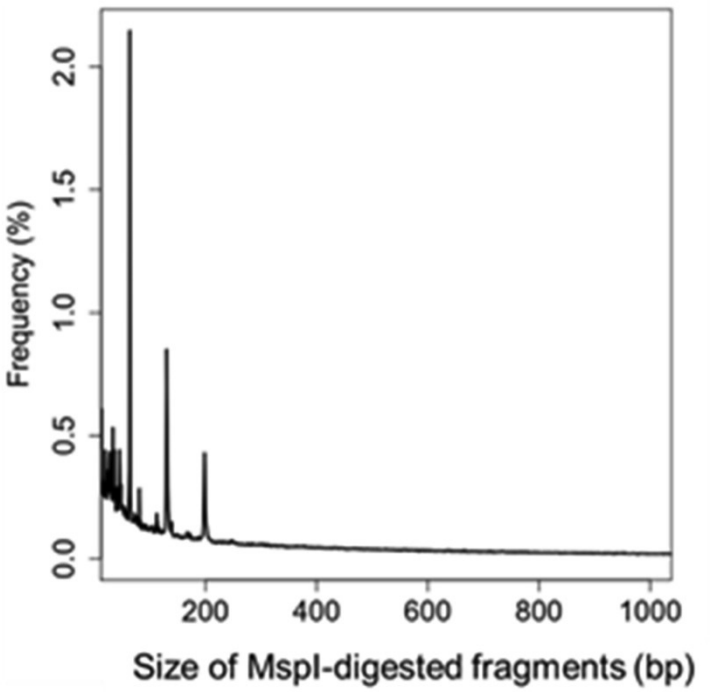


FIG. 127B

Size ranges (bp)	No. of molecules	Percentage of molecules within a size range relative to total fragments (%)	No. of molecules within a size range overlapping CpG islands	Percentage of molecules within a size range overlapping CpG islands (%)	No. of CpG sites being sequenced	No. of CpG sites falling within CpG islands	Percentage of CpG sites targeted by size selection and falling within CpG islands (%)
50-200	526,543	23.03	104,059	19.76	2,358,020	885,041	37.53
200-400	269,562	11.79	23,927	8.88	1,781,556	353,087	19.82
400-600	177,776	7.77	7,369	4.15	1,468,561	107,130	7.29
600-800	133,927	5.86	3,673	2.74	1,326,544	48,851	3.68
800-1000	104,976	4.59	2,168	2.07	1,193,233	25,821	2.16
1000-2000	311,596	13.63	4,596	1.47	4,610,504	58,288	1.26
2000-3000	149,468	6.54	1,771	1.18	3,036,951	25,106	0.83
3000-4000	86,760	3.79	809	0.93	2,165,171	10,785	0.50
5000-6000	36,931	1.62	266	0.72	1,242,712	3,412	0.27
6000-7000	25,027	1.09	202	0.81	947,874	3,354	0.35
7000-8000	17,597	0.77	86	0.49	736,830	791	0.11
8000-9000	12,658	0.55	76	0.60	583,680	993	0.17
9000-10000	9,184	0.40	48	0.52	461,935	591	0.13
10000-15000	20,790	0.91	97	0.47	1,255,731	2,003	0.16
15000-20000	5,111	0.22	16	0.31	414,400	163	0.04
20000-25000	1,441	0.06	6	0.42	147,731	34	0.02

FIG. 128

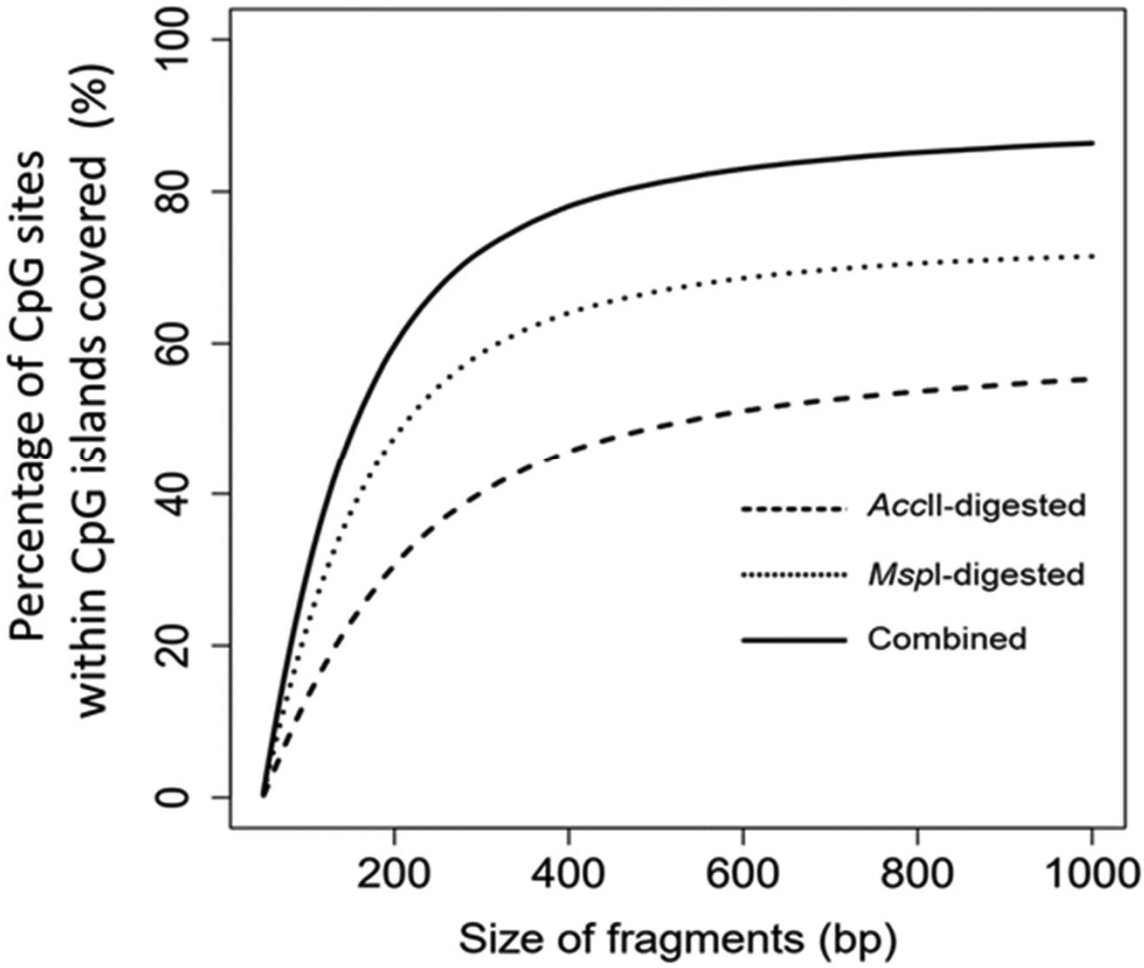

Percentage of CpG sites within CpG islands covered (%)
0
20
40
60
80
100
200
400
600
800
1000
Size of fragments (bp)
AccII-digested
MspI-digested
Combined

FIG. 129

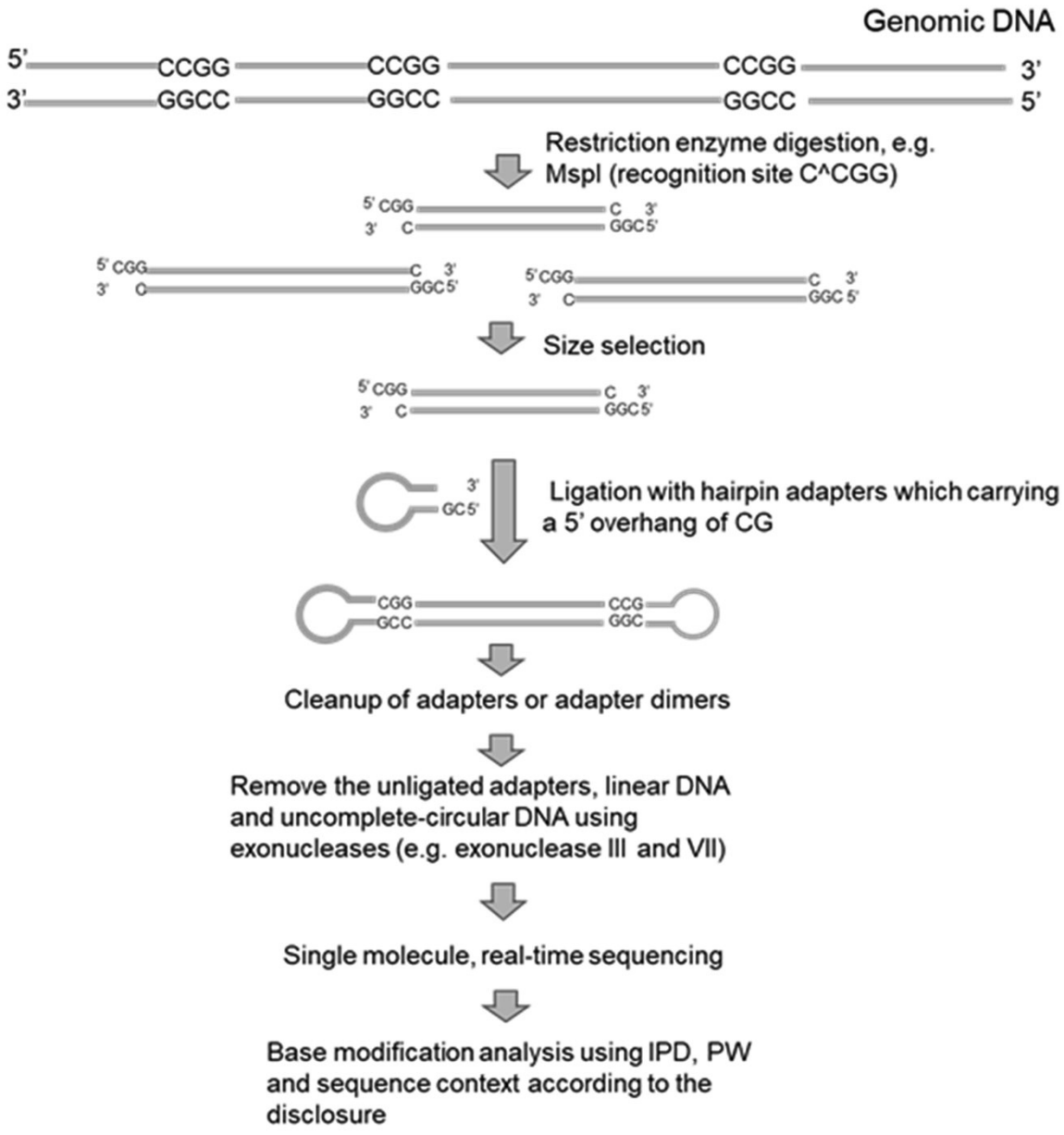

Genomic DNA
5'
CCGG
CCGG
CCGG
3'
3'
GGCC
GGCC
GGCC
5'
Restriction enzyme digestion, e.g.
MspI (recognition site C^CGG)
Size selection
Ligation with hairpin adapters which carrying
a 5' overhang of CG
Cleanup of adapters or adapter dimers
Remove the unligated adapters, linear DNA
and uncomplete-circular DNA using
exonucleases (e.g. exonuclease III and VII)
Single molecule, real-time sequencing
Base modification analysis using IPD, PW
and sequence context according to the
disclosure

FIG. 130

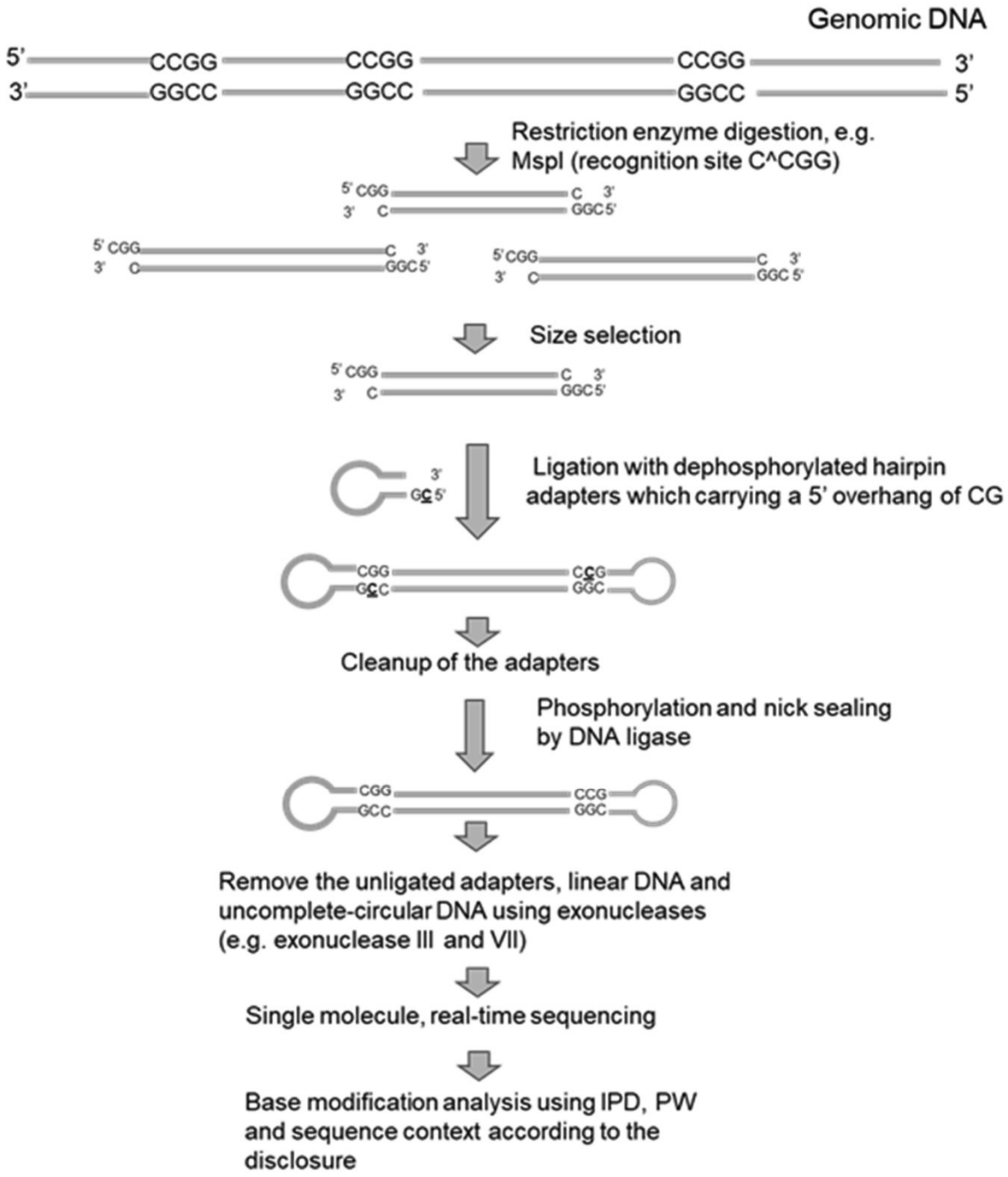

Genomic DNA
5'
CCGG
CCGG
CCGG
3'
3'
GGCC
GGCC
GGCC
5'
Restriction enzyme digestion, e.g.
MspI (recognition site C^CGG)
Size selection
Ligation with dephosphorylated hairpin
adapters which carrying a 5' overhang of CG
Cleanup of the adapters
Phosphorylation and nick sealing
by DNA ligase
Remove the unligated adapters, linear DNA and
uncomplete-circular DNA using exonucleases
(e.g. exonuclease III and VII)
Single molecule, real-time sequencing
Base modification analysis using IPD, PW
and sequence context according to the
disclosure

FIG. 131

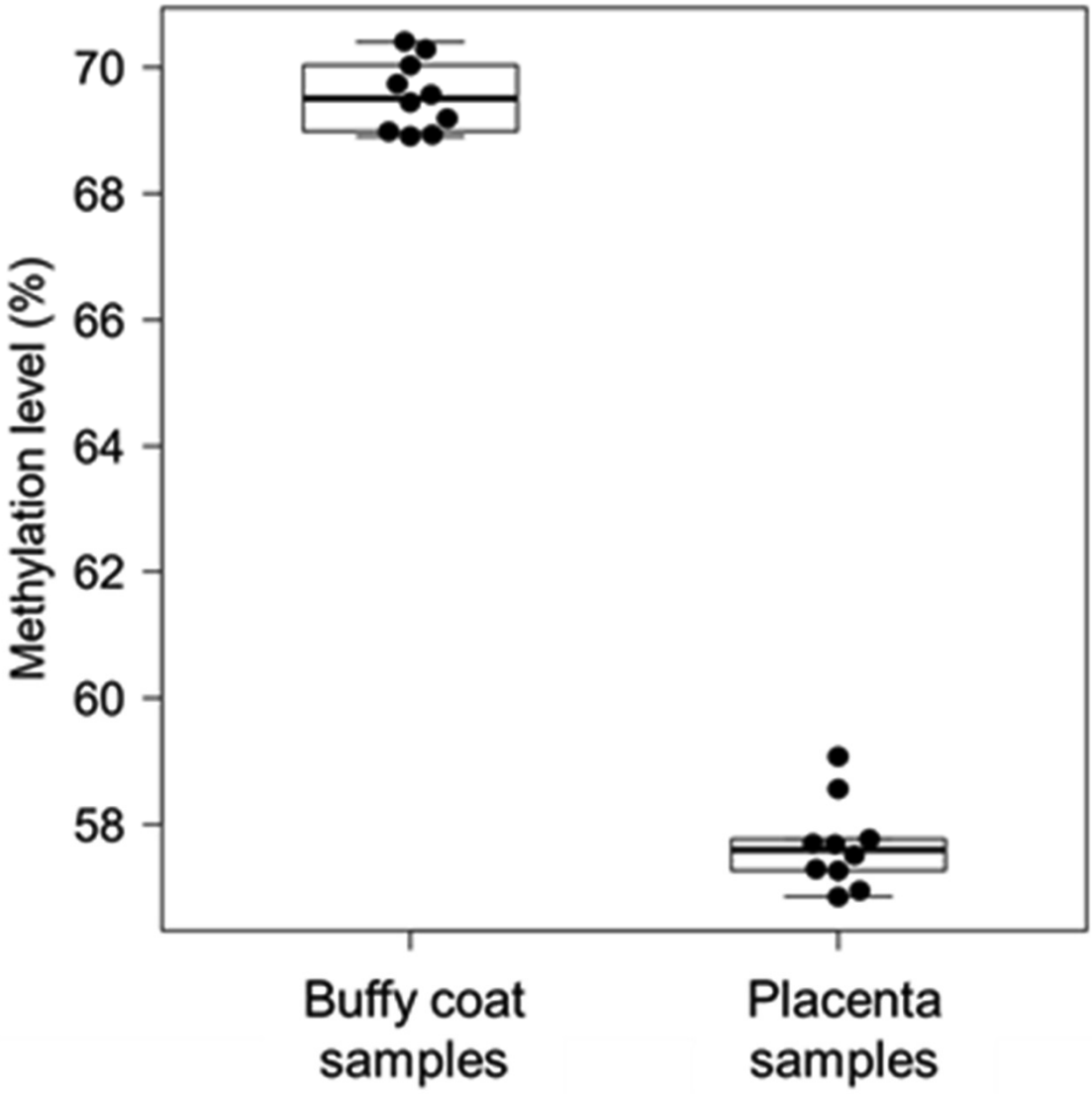

70
68
66
64
62
60
58
Methylation level (%)
Buffy coat samples
Placenta samples

FIG. 132

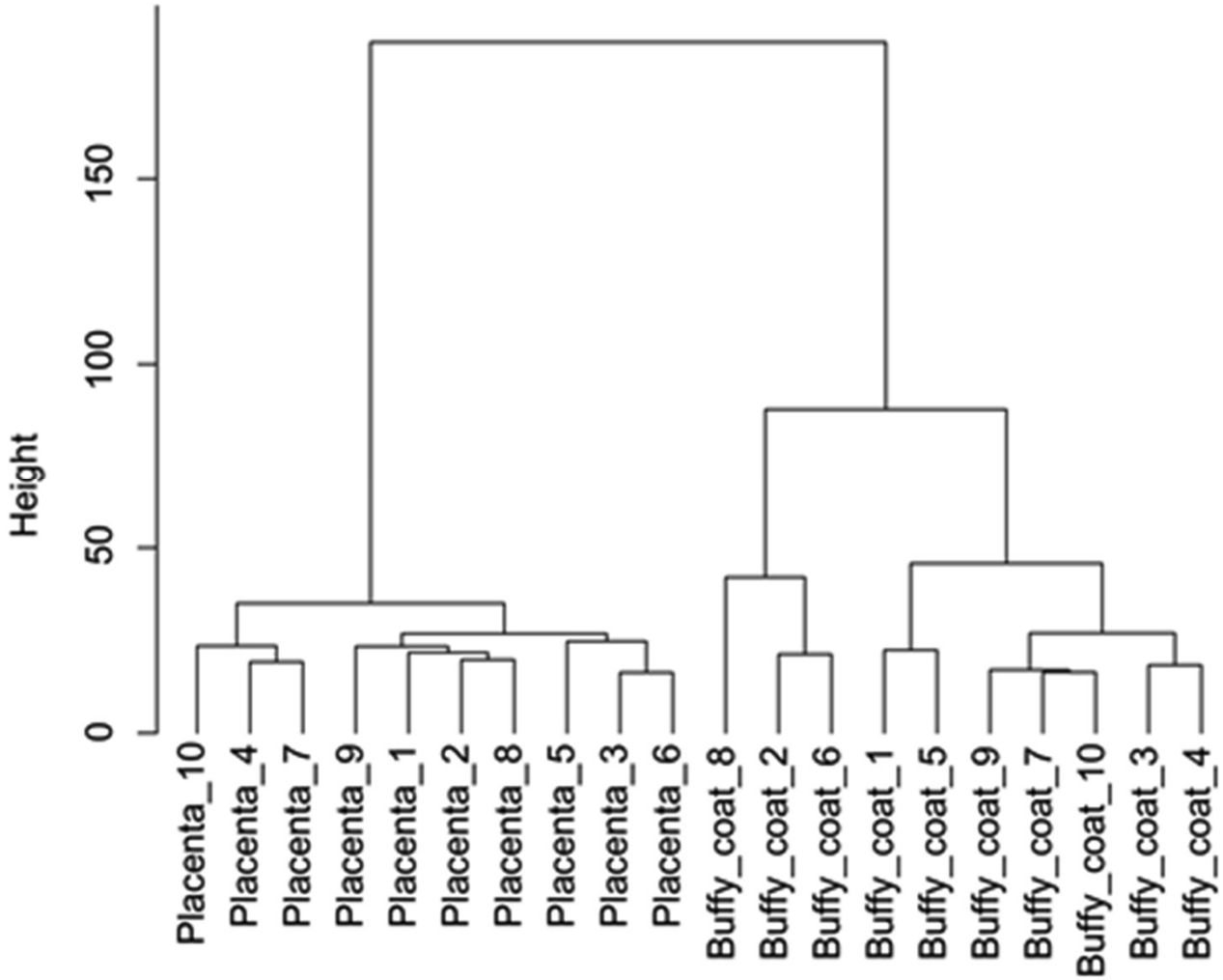

Height
0
50
100
150
Placenta_10
Placenta_4
Placenta_7
Placenta_9
Placenta_1
Placenta_2
Placenta_8
Placenta_5
Placenta_3
Placenta_6
Buffy_coat_8
Buffy_coat_2
Buffy_coat_6
Buffy_coat_1
Buffy_coat_5
Buffy_coat_9
Buffy_coat_7
Buffy_coat_10
Buffy_coat_3
Buffy_coat_4

FIG. 133