

平成 28 年度  
修士論文

NGS データからのアレル特異的メチル化領域  
の同定

金沢大学大学院 自然科学研究科 電子情報科学専攻  
ゲノム情報工学研究室  
員 元奇

指導教員  
山田 洋一 准教授

平成 29 年 2 月 14 日

目次

1. 背景 .....	1
1.1 DNA メチル化 .....	1
1.2 一塩基多型 (SNP: Single Nucleotide Polymorphism) .....	2
1.3 アレル特異的メチル化 (ASM) .....	3
1.4 HM-PCR 法 .....	3
1.5 全ゲノムバイサルファイトシーケンシング (WGBS: whole-genome bisulfite sequencing) 法 .....	4
2. 目的 .....	5
3. 研究方法 .....	5
3.1 HM-PCR 法の結果 .....	5
3.2 使用した WGBS データ .....	5
3.3 Bismark によるバイサルファイトリードマッピング及び CpG 配列のメチル化状態の決定 .....	6
3.4 BisSNP によるヘテロ SNP サイトの抽出 .....	8

3.4.1	重複リードの除去と再アライメント .....	9
3.4.2	各バイサルファイトリードにおけるヘテロ SNP 塩基の抽出と各リード由来するアレルの識別 .....	10
3.4.3	Fisher's exact test によるアレル特異的メチル化領域候補の推定 .....	11
3.4.4	Benjamini & Hochberg 法(BH 法)による False Discovery Rate の調整 .....	12
3.5	メチル化エントロピーの計算 .....	12
3.5.1	プロモーターにおけるメチル化エントロピーの計算 .....	13
3.5.2	メチル化エントロピー平均値によるプロモーターの分類 .....	14
4.	研究結果 .....	15
4.1	Bismark によるバイサルファイトリードマッピング .....	15
4.2	BisSNP によるヘテロ SNP サイトの抽出 .....	15
4.3	Fisher's exact test によるアレル特異的メチル化領域候補の推定 .....	16
4.4	アレル特異的メチル化を受けたヒト転写因子遺伝子プロモーターの推定 .....	16
4.5	アレル特異的メチル化を受けたヘテロ SNP 周辺配列の特徴 .....	17
4.6	メチル化エントロピー平均値によるプロモーターの分類 .....	18
5.	まとめと考察 .....	18
	謝辞 .....	19
	参考文献 .....	19

# 1. 背景

## 1.1 DNA メチル化

DNA(deoxyribonucleic acid)メチル化は、シトシンの次にグアニンが現れる 2 塩基配列(ジヌクレオチド)である CpG 配列のシトシン (cytosine) のピリミジン環の 5 位炭素原子へのメチル基の付加反応である(図 1). DNA メチル化は正常な発生に必須であり, ゲノムインプリンティング, X 染色体の不活性化, 発癌など多くのキーステップと関係している. CpG サイトの出現頻度が, ゲノム中の他と比べ高い領域は CpG Island と言う. ヒトゲノムでは遺伝子プロモーターの約 70%が CpG Island を含むとされる. CpG Island は通常メチル化されていないが, がん細胞ではしばしば CpG Island のメチル化による遺伝子転写抑制がみられる(図 2).

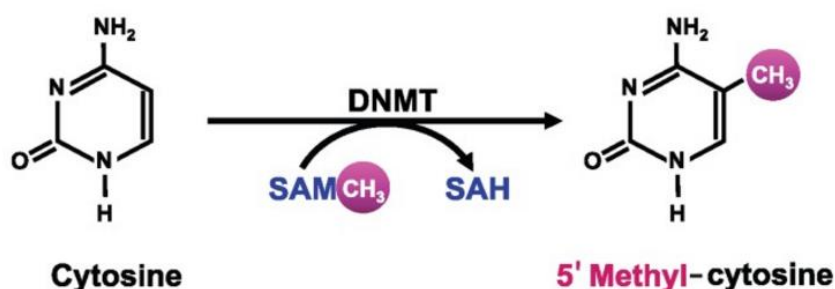


図 1 DNA メチル化



### 1.3 アレル特異的メチル化 (ASM)

ヒトの両親由来の DNA アレル間の SNP のなかには、近傍領域の化学修飾や遺伝子発現を変化させ、糖尿病への罹患率の違いを引き起こすものが既に報告されている。図 4 のように父親と母親からそれぞれ受け継いだアレルの間に SNP がある場合、これが近隣の DNA メチル化修飾の有無を変化させることがある。この結果、同様に近隣の遺伝子発現のオンオフがアレル間や個人間で変化する。このように SNP が DNA のメチル化修飾の有無を変える領域をアレル特異的メチル化領域と言い、これは同一遺伝子の発現をアレル間や個人間で変化させることで各個人の各種疾患への罹患のし易さを変化させる。このため、新たにヒトアレル特異的メチル化領域が同定できれば、個人の各種疾患への罹患のし易さを決める仕組みを網羅的に解明できることが期待できる。

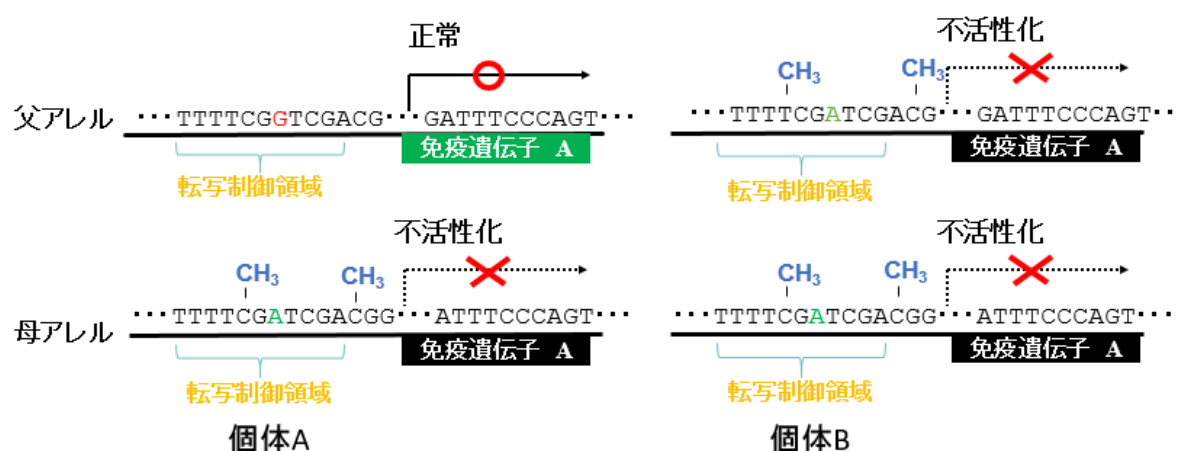


図 4 : アレル特異メチル化(左)と非アレル特異的メチル化(右)

### 1.4 HM-PCR 法

4 種類のアレル別メチル化状態を識別できる HpaII-McrBC PCR (HM-PCR) 法は、非メチル化 DNA のみを切断する HpaII とメチル化 DNA のみを切断する McrBC という対照的な制限酵素によりそれぞれ独立にゲノムを消化し、これを鋳型として標的 DNA 領域特異的なプライマー対を用いて PCR 増幅を行う。図 5 に、HM-PCR 法の原理を示す。図 5 の **m** は、標的部位 (PCR で増幅する領域) 中の HpaII および McrBC 認識部位であり、これがメチル化されていることを示す。標的部位が消化された場合は PCR の鋳型とならずバンドが得られないので、HpaII あるいは McrBC で消化した DNA からのバンドの出現パターンから、標的部位のアレル別メチル化状態が判定できる、というのが HM-PCR 法の原理である。HM-PCR によって得られた PCR 産物は電気泳動により分離同定され、各制限酵素消化ゲノムからのバンドの有無からアレル別のメチル化状態が判別される[2]。図 5 では、アレル特異的メチル化領域は、混合型メチル化として検出される。

HM-PCR 法は、試料として用いた全細胞のメチル化状態を調べることが可能であるが、制限酵素によるゲノムの不完全消化に由来するメチル化状態の誤判定がありうるため、より精度の高いバイサルファイトシーケンシング法により、そのメチル化状態を確認しなければならない。

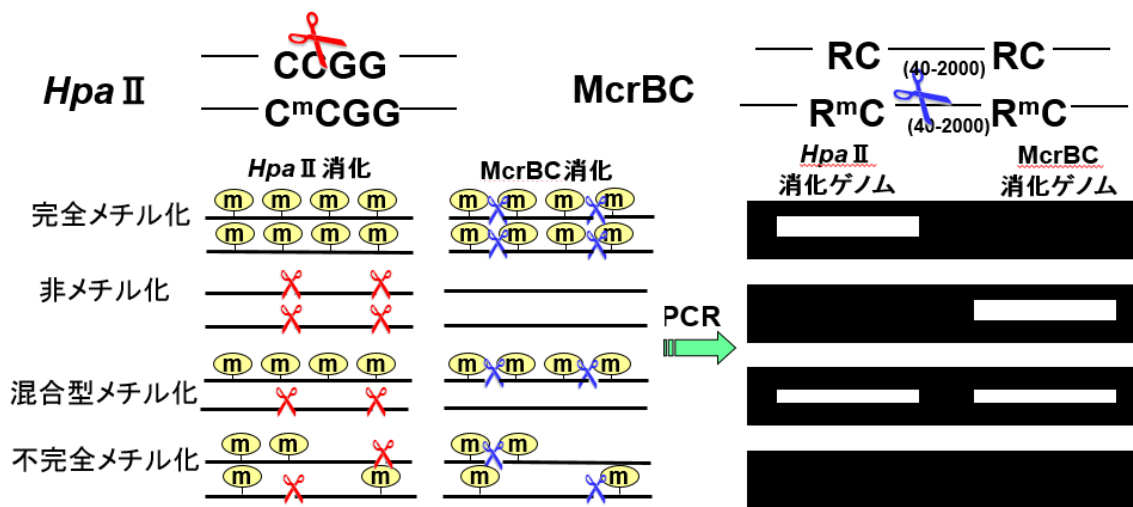



図5：HM-PCR法（はメチル化CG）

## 1.5 全ゲノムバイサルファイトシーケンシング（WGBS: whole-genome bisulfite sequencing）法

WGBS法は、様々な組織やがん細胞などでDNAのメチル化パターンを知るのに最も適した方法である。この手法により、比較的少量のDNAサンプルから、一塩基解像度で、5-メチルシトシンの出現パターンを全ゲノムレベルで調べることができる。

WGBS法は、早津らによって発見された原理を用いる。一本鎖DNAの断片化とバイサルファイト（亜硫酸水素）塩による処理を行うと、非メチル化シトシン環の6位の炭素にスルホン基が付き、更に5位の炭素の脱アミノ化が起こる。続くアルカリ処理による加水分解でスルホン基を除去してウラシル（uracil）に変換する。5-メチルシトシンはバイサルファイトへの反応性が極めて低いためそのまま残る。このDNAを鋳型として全ゲノム領域由来の全てのDNA断片をPCR増幅後、次世代シーケンサーを用いて配列決定する。この結果、5-メチルシトシンはシトシンとして、非メチル化シトシンはチミンとしてPCR増幅されるため、全ゲノム領域由来の各DNA断片配列（バイサルファイトリード）におけるメチル化状態を明らかにすることができる（図6）。

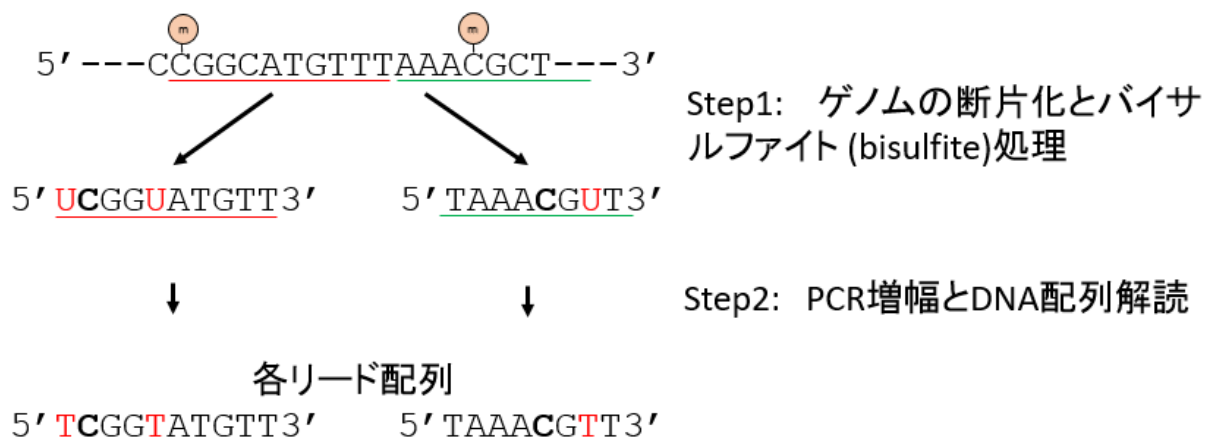


図6：WGBS法による一塩基解像度全ゲノムメチル化解析

## 2. 目的

本研究室で、HM-PCR 法により、末梢血細胞における 1068 個のヒト転写因子遺伝子プロモーター領域のメチル化状態が 4 つに分類された。しかしながら、HM-PCR 法は、試料として用いた全細胞のメチル化状態を調べることが可能であるが、制限酵素によるゲノムの不完全消化に由来するメチル化状態の誤判定がありうるため、より精度が高く、一塩基解像度のバイサルファイトシーケンシング法により、そのメチル化状態を確認する必要がある。そこで、HM-PCR 法から得られた結果を、一塩基解像度で DNA メチル化状態を調べることが可能な WGBS 法より得られた結果と比較し、ヒトアレル特異的メチル化領域候補を推定する。

## 3. 研究方法

### 3.1 HM-PCR 法の結果

本研究室では、HM-PCR 法を用いて 1,068 個のヒト転写因子遺伝子プロモーター領域に対し、4 つのメチル化パターンに分類した (表 1)。1,068 個のヒト転写因子遺伝子プロモーターの内、非メチル化を 776 個、完全メチル化を 76 個、混合型メチル化を 132 個、不完全メチル化を 84 個同定した。

表 1 : 転写因子遺伝子プロモーター領域(一部)

ヒト転写因子プロモーター領域	染色体	start	end	メチル化
0504141409_00186_0052_01F#A0001#TH1L#TH1-like	chr20	57556025	57556470	非メチル化
0504141409_00203_0012_01F#A0005#PFDN5#prefoldin	chr12	53688618	53689379	完全メチル化
0504141409_00048_0014_01F#A0002#SNAPC2#small	chr19	7984513	7984815	混合型メチル化
0504141409_00333_0002_01F#A0073#POLRMT#polymerase	chr19	633683	634502	不完全メチル化

### 3.2 使用した WGBS データ

表 2 の WGBS データ を Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) より入手した。図 7 は、入手した WGBS データ中のバイサルファイトリードの一例である。図 7 における最初の行は、リードの ID と長さを、二行目はバイサルファイトリード配列をそれぞれ示している。

表 2 : 使用した WGBS データ

サンプル ID	生物	組織	年齢	健康状況	リード数	リードの長さ
---------	----	----	----	------	------	--------

GSM848927	ヒト	末梢血	26	健康	約 2.8 億	90 塩基
GSM774850	ヒト	末梢血	新生児	健康	約 5.5 億	90 塩基
GSM1134680	ヒト	脳組織	不明	健康	約 5.4 億	90 塩基

```
@SRR389248.1 FCC02FUACXX:6:1101:1235:2196 length=90
TTCGGGTTGTTATGGAATGAGAGAATATCGTTTAAATTTATTAGTTATTAAGGAAATATAAATTGATGGAATAGAGAGATATTGTTTTGT
+SRR389248.1 FCC02FUACXX:6:1101:1235:2196 length=90
@CCFFDFHHDHIIIFHIGEH0EEHGGIIIGIIIIIFIIIHGIFIIIGIIIIIIIIHIIIIIIIBC GHHIGIEE?7=?=CDDDB@BCE##
```

図 7 : バイサルファイトリード

### 3.3 Bismark によるバイサルファイトリードマッピング 及び CpG 配列のメチル化状態の決定

各バイサルファイトリードの由来するゲノム上の位置を調べる(マッピング)ために、既に配列が決定されているヒトゲノム配列(リファレンスゲノム)の hg19 バージョンを The University of California Santa Cruz (UCSC) よりダウンロードした。

また、バイサルファイトリードのリファレンスゲノムへのマッピングには、Bismark フリーソフトウェアを用いた。Bismark は、図 8 に示すように、リファレンス配列のすべての C を T に変換にしたものと、G を A に変換したものをそれぞれ用意する。次に、バイサルファイトリード配列も同様に、C→T および G→A 変換した配列を用意する。最後に、これらの変換後のバイサルファイトリード配列を、同様に塩基変換後の各リファレンスゲノムと比較し、完全に一致するゲノム領域にマッピングを行う。得られたリファレンスゲノム配列とバイサルファイトリード配列間のアラインメントを基に、各 CpG 配列のメチル化状態を決定する[3]。



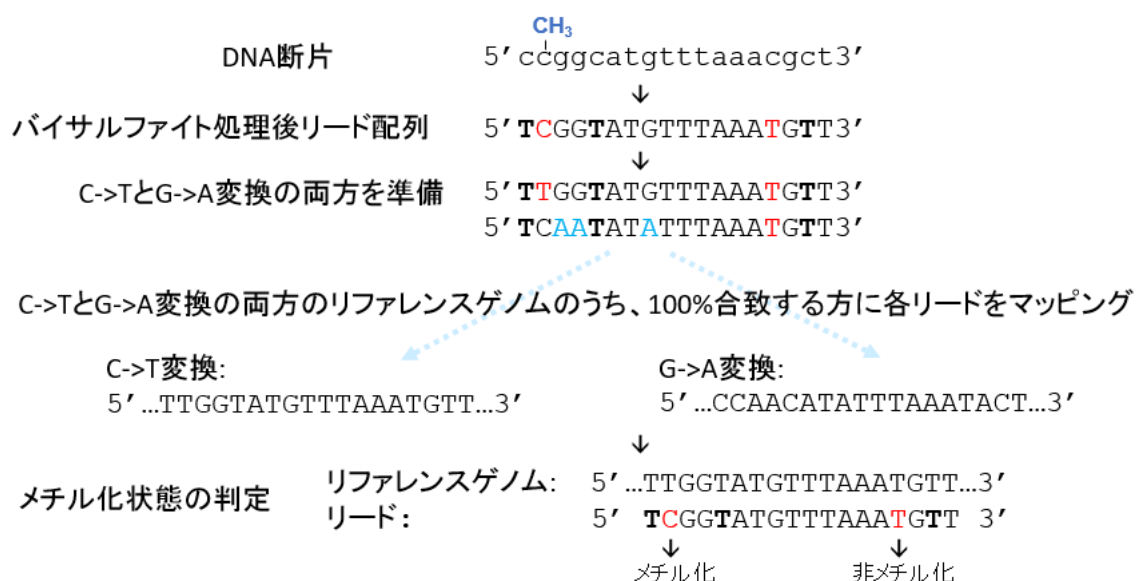


図 8 : Bismark によるリードマッピングの流れ

マッピング済みの各リード情報は.sam ファイルに格納される。バイサルファイトリードのマッピング結果の一部を図 9 に示す。また、マッピング結果における各属性値の詳細を表 3 に示す。更に、メチル化状態記号の詳細を表 4 に記載した。

```
SRR389248.1_FCC02FUACXX:6:1101:1235:2196_length=90      99      chr10      130472134
      42      90M      =      130472257      213      TTCGGGTTGTTATGGAATGAGAGAATA
TCGTTTAAATTTATTAGTTATTAAGGAAATATAAATTGATGGAATAGAGAGATATTGTTTGT      @CCFFDDFHHDH
IIFHIGEH@EEHGGIIGIIIFIIHGIIFIIIGIIIIIIIIHIIIIIBC GHHIGIEE?7=?=CDDDB@BCE##
      MD:Z:0C0C7C0C14C5C0C3C1C1C0C3C2C9C12C9C3C0C2C0      XG:Z:CT NM:i:19 XM:Z:hxZ...
..hh.....h..Z..hh...h.h.hx...h..h.....h.....x.....x...hx..
h XR:Z:CT
```

図 9 マッピング済みのリード

表 3 : マッピング結果における各属性値の詳細

列番号	意味	例
1	配列 ID	SRR389248.1_FCC02FUACXX:6:1101:1235:2196_length=90
3	染色体	chr10
4	リード開始座標	130472134
10	バイサルファイトリード配列	TTCGGGTTGTTATGGAATGAGAGAATA TCGTTTAAATTTATTAGTTATTAAGGA AATATAAATTGATGGAATAGAGAGATA TTGTTTGT
13	マッピング対象レファレンスゲノムにおける塩基変換状態	XG:Z:CT
15	メチル化されうる各シトシン配列のメチル化状態	XM:Z:hxZ.....hh.....h..Z..hh...h.h.hx.. .h..h.....h.....x.....x...hx..h
16	マッピングしたリード	XR:Z:CT



	における塩基変換状態	
--	------------	--

表 4 メチル化状況の記号の意味

記号	意味
z	CpG における非メチル化 C
Z	CpG におけるメチル化 C
x	CHG における非メチル化 C
X	CHG におけるメチル化 C
h	CHH における非メチル化 C
H	CHH におけるメチル化 C

### 3.4 BisSNP によるヘテロ SNP サイトの抽出

Bismark によるバイサルファイトリードマッピングの結果から、アレル特異的メチル化領域を直接推定するためには、両親由来のアレルを識別するためのヘテロ SNP をリード内に同定することが必要である。そこで、マッピングされたリード中に存在するヘテロ SNP を BisSNP フリーソフトウェアを用いて抽出した。BisSNP は、バイサルファイトリードマッピング結果が格納されたバイナリファイル (.bam) を入力とし、最終的な出力は vcf フォーマットで行う(図 10a)。

ヘテロ SNP の抽出は、SNP データベースに登録されているすべての SNP 位置ごとに行う。一つの SNP 位置に対し、アレルを示す 2 つの塩基の組み合わせにより、10 種類の 2 塩基の組み合わせ(G)が生じる(表 5)。リード(r)から読み取った塩基(D), 及び SNP Database における G の事前確率  $\pi$  (G) もベイズ推論に利用される(図 10b)。最大尤度 ( $Pr(G|D)$ ) は、以下の式 1 および 2 により計算される[4]。

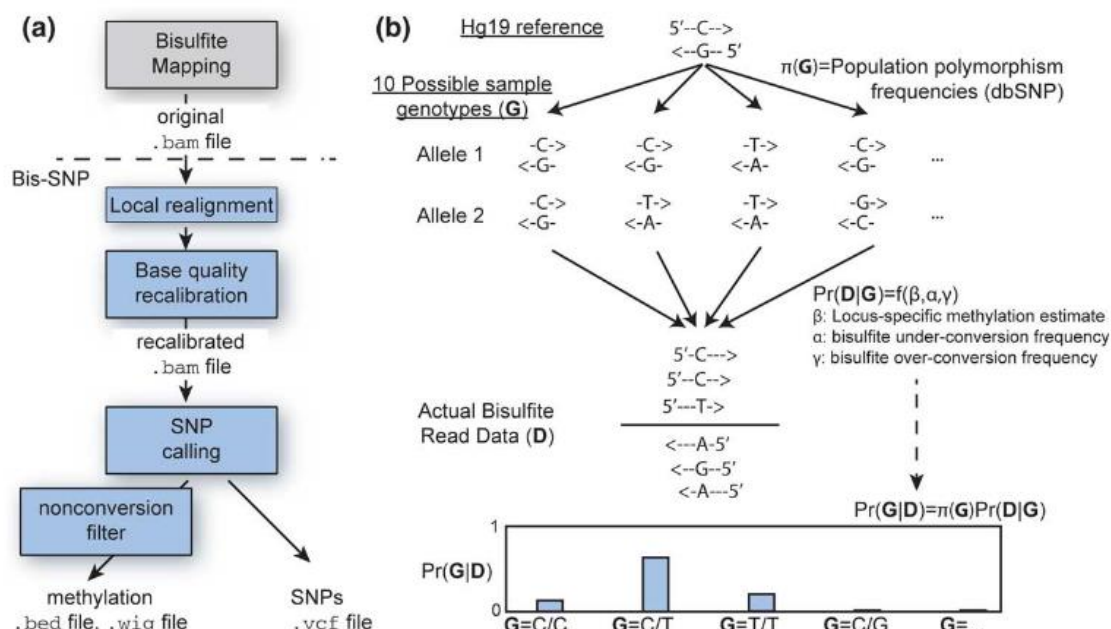


図 10 : BisSNP の手順及び SNP 抽出アルゴリズム (出典 : Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data, Figure 2, Genome Biology, 2012)

表 5 両アレルから可能な遺伝子型

遺伝子型(G)	A	T	C	G
A	AA	AT	AC	AG
T		TT	TC	TG
C			CC	CG
G				GG

$$Pr(\mathbf{D}|G) = \prod_{j=1}^r Pr(D_j|G) \dots\dots\dots (式1)$$

$$Pr(G|\mathbf{D}) = \frac{\pi(G)Pr(\mathbf{D}|G)}{Pr(\mathbf{D})} \dots\dots\dots (式2)$$

BisSNP によるヘテロ SNP 抽出結果の一部を表 6 に示す. 表 6 では, SNP のゲノム上の位置, SNP の ID, SNP 塩基, そしてバイサルファイトリードマッピングの結果から得られた SNP のタイプ(ホモ(1/1)またはヘテロ(0/1))の情報を知ることができる.

表 6 : 抽出された SNP(一部)

染色体番号	ポジション	ID	多型 1	多型 2	タイプ
chr1	34052305	rs12727575	C	T	0/1
chr1	34052605	rs7526990	A	G	0/1
chr1	34061741	rs72662021	G	C	0/1
chr1	34061826	rs67344823	A	T	0/1
chr1	34065436	rs10914750	C	T	0/1
chr1	34326409	rs1321626	T	C	1/1

### 3.4.1 重複リードの除去と再アライメント

WGBS データでは, PCR 増幅により, 同じ DNA 配列が偏って数多く読み取られていることがある. バイサルファイトリードから SNP を検出する際, このような重複リードは SNP の誤検出を引き起こすため, 複数の同一リードを一つに見なす(重複リードの除去) ことが必要である. 重複リードの除去は bismark 付属のプログラムを使って以下のコマンドで行った.

`deduplicate_bismark -p file_name.bam`

またリード DNA 配列に欠損や挿入があるとき、それらを考慮せずにバイサルファイトリードをレファレンスゲノムにマッピングしてしまうと、欠損や挿入のためにアラインメントがずれる塩基が生じてしまい、SNP の検出に影響を与える。そのため、欠損や挿入がある部分は再アラインメントを行う必要がある。リードの再アラインメントは以下の手順で行いた。

#### 1. Find indel region

既知の indel 多型データを download し、この indel 領域において、レファレンスゲノムとリードの間で多くのミスマッチを有する領域を BisSNP 付属のプログラムである BisulfiteRealignerTargetCreator を使用して抽出。

#### 2. Realign in the indel region

抽出した intervals ファイル(表 7)を使用し、BisSNP 付属のプログラムである BisulfiteIndelRealigner でレファレンスゲノムとリードの再アラインメントを行う。

表 7 抽出された indel 区間(一部)

染色体番号	区間(開始-終了)
chr1	10330-10440
chr1	13657-13678
chr1	13955

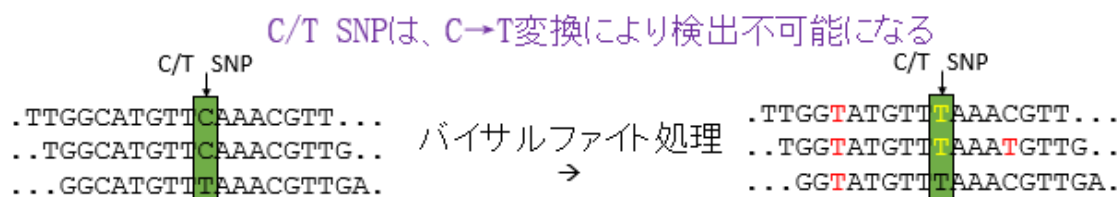
### 3.4.2 各バイサルファイトリードにおけるヘテロ SNP 塩基 の抽出と各リード由来するアレルの識別

抽出したヘテロ SNP を含むバイサルファイトリードを、同定した各ヘテロ SNP ごとにまとめた。次に、ヘテロ SNP を含むバイサルファイトリードごとに、ヘテロ SNP 位置の塩基を抽出し、アレルの識別を行った。この際、C/T の SNP 及び G/A の SNP は、図 11a に示すように、バイサルファイト処理によりヘテロ塩基の区別が不可能になる。また、C/T と G/A 以外の SNP では、G/C→G/T や G/T→A/T のように、バイサルファイト処理により、本来の SNP 塩基とは異なる二塩基の組み合わせになる。そこで、ヘテロ SNP を含むバイサルファイトリードからのヘテロ SNP 塩基の抽出とアレルの識別を以下のように行った。

1. データベースに登録されている SNP 塩基は、C/T (or T/C) ではないが、バイサルファイトリード中の当該ヘテロ SNP サイトに塩基 T が存在する場合は、その塩基 T をバイサルファイト変換前の塩基 C に戻してから、リード上の SNP 塩基として抽出し、各リードの由来するアレルの識別を行う。
2. データベースに登録されている SNP 塩基は、G/A (or A/G)ではないが、バイサルファイトリード中のヘテロ SNP サイトに塩基 A が存在する場合は、その塩基 A をバイサルファイト変換前の塩基 G に戻してから、リード上の SNP 塩基として抽出し、各リードの由来するアレルの識別を行う。

3. 図 11b に示すように、データベースに登録されている SNP が C/T (or T/C) の場合、G→A 変換リードのみから、ヘテロ SNP サイトの塩基を抽出し、各リードの由来するアレルの識別を行う。
4. 図 11b に示すように、データベースに登録されている SNP は G/A (or A/G) の場合、C→T 変換リードのみから、ヘテロ SNP サイトの塩基を抽出し、各リードの由来するアレルの識別を行う。

#### a) C/T SNP と C→T 変換



#### b) リードの選択

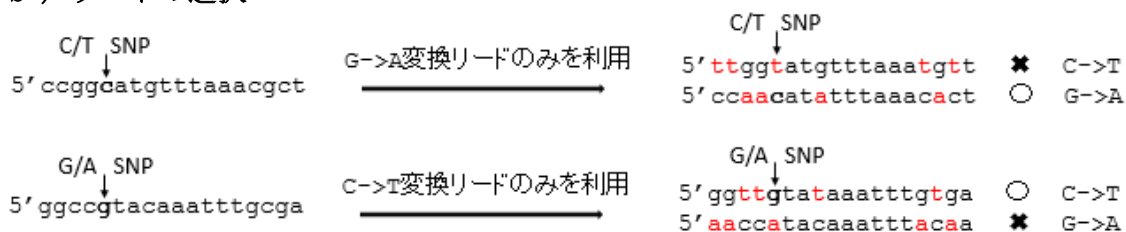


図 11: C/T SNP におけるバイサルファイト変換の修正

### 3.4.3 Fisher's exact test によるアレル特異的メチル化領域候補の推定

フィッシャーの正確確率検定 (Fisher's exact test) は、2×2 の分割表において行および列で示される 2 つの要因が互いに独立であるかどうかを判定するノンパラメトリックな検定法である。アレル特異的メチル化領域を推定するために、リードから同定した各ヘテロ SNP サイトにおいて、これを含むリードのアレル間でのメチル化割合の違いをフィッシャーの正確確率検定で検討した。フィッシャーの正確確率検定の帰無仮説、対立仮説及び有意水準は以下のとおりである。

帰無仮説 H0: アレル間で、SNP 近傍の CpG メチル化状態に違いがない。  
 対立仮説 H1: アレル間で、SNP 近傍の CpG メチル化状態に違いがある。  
 有意水準:  $\alpha = 0.05$

フィッシャーの正確確率検定の p 値は式 3 により求められる。なお、各変数の意味は表 8 の通りである。表 8 では、ヘテロ SNP サイトの SNP 塩基 1 を含むリード (アレル 1) におけるメチル化 CpG 数と非メチル化 CpG 数の合計数に占めるメチル化 CpG 数の割合と、SNP 塩基 2 を含むリード (アレル 2) におけるメチル化 CpG 数と非メチ

ル化 CpG 数の合計数に占めるメチル化 CpG 数の割合の違いを検定している。

$$p = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! n!}$$

・・・式 3

表 8：フィッシャーの正確確率検定の分割表

	メチル化 CpG	非メチル化 CpG	Total
SNP 塩基 1	a	b	a+b
SNP 塩基 2	c	d	c+d
Total	a+c	b+d	a+b+c+d=n

### 3.4.4 Benjamini & Hochberg 法(BH 法)による False Discovery Rate の調整

フィッシャーの正確確率検定は、リードから同定した各ヘテロ SNP サイトにおいて実施されるため、各多重検定となる。そこで、フィッシャーの正確確率検定から得られた p 値に対して FDR (False Discovery Rate)補正を行った。FDR は、検定によって棄却された帰無仮説の内、間違っ棄却された帰無仮説の割合のことであり、これを FDR q 値と呼ぶ。FDR q 値は 以下の手順により求められる。フィッシャーの正確確率検定により得られた p 値に補正を掛け、得られた FDR q 値が 0.05 以下となった場合、帰無仮説を棄却し対立仮説を採用した。対立仮説が成立したゲノム領域はアレル特異的メチル化領域の候補なため、HM-PCR 法により同定された 1,068 個のヒト転写因子遺伝子プロモーター領域のメチル化パターンと比較し、両者の結果が一致するかを調べた。

1. N 個の帰無仮説を、p 値の小さい順に並べる
2. p 値の小さい順に並べた帰無仮説において、各 p 値を  $q=p*N/m$  ( $m=1,2,...N$ ) により補正する
3.  $q<0.05$  を満たす帰無仮説を棄却する

## 3.5 メチル化エントロピーの計算

バイサルファイトリードを使ったメチル化分析では、一般に平均メチル化レベルより分析されるが、このような従来の方法では、DNA 鎖における連続する CpG サイトのメチル化状態の組み合わせである DNA メチル化パターンを分析することができない。またアレルを区別するためには、バイサルファイトリード中にヘテロ SNP を同定する必要がある、これがホモ SNP の場合は、アレル特異的メチル化領域を推定することが困

難である。

そこで本研究では「メチル化エントロピー」を利用し，DNA メチル化パターンの変動性を評価し，アレル特異的メチル化領域の強力な候補である「メチル化アレルと非メチル化アレルの混合型」の転写因子プロモーター領域をバイサルファイトリードを使って抽出することにした．メチル化エントロピー(Methylation entropy)は以下の式 4 により定義された[5]．

$$\text{Methylation Entropy} = \frac{1}{b} \sum \left( -\frac{n_i}{N} \log \frac{n_i}{N} \right)$$

$b$ : Number of cytosine sites

$n_i$ : Observed occurrence of methylation pattern  $i$

$N$ : Total number of sequence reads generated

・・・式 4

### 3.5.1 プロモーターにおけるメチル化エントロピーの計算

図 12 に示すように，メチル化 CpG を「1」，非メチル化 CpG を「0」，メチル化情報欠損 CpG を「-」とした，各プロモーター領域にマップされたバイサルファイトリード上の CpG メチル化情報を抽出した．

```

---00000000000000---
---01100011111011---
----00000000000000---
----00000000000000---
----01001110110111---
-----01010111111111-----
-----00000000000000-----
-----00000000000000-----
-----00000000000000-----
-----111011110111-----
-----111011110000-----
-----000000000000-----
-----1111111110-0---
-----1111101110-0---
-----000000000000---
-----000000000000---
-----11111100-00---
-----10001111110--
-----000000000000--
-----000000000000-
-----011111111000

```

図 12 : プロモーター領域におけるメチル化 CpG データ

図 13 に示すように、各プロモーター領域において、リードカバー数が 16 以上の連続した 4 つの CpG セグメントごとに、メチル化エントロピーを計算した。また各プロモーター領域において、各 CpG セグメントのメチル化エントロピー値の平均値と分散値を計算した。

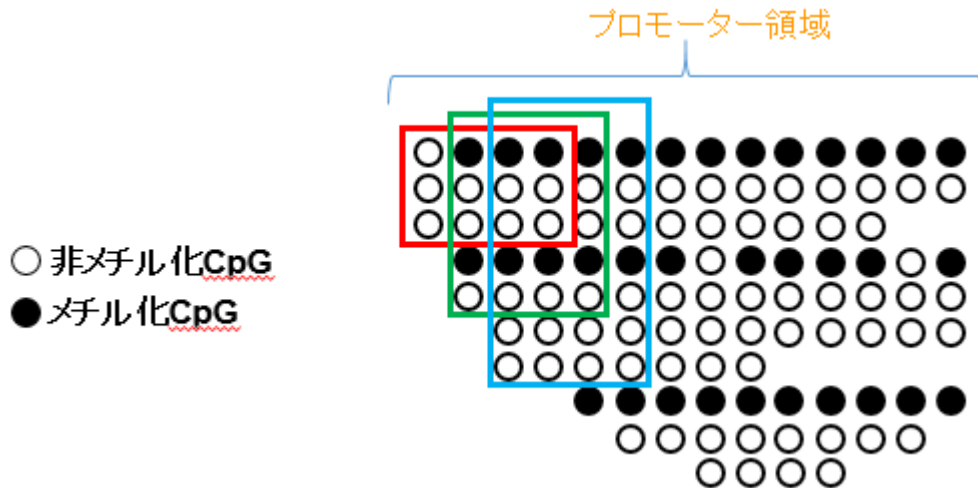


図 13 : プロモーター領域における連続した 4 つの CpG セグメントの検出

### 3.5.2 メチル化エントロピー平均値によるプロモーターの分類

図 14 に示すように、メチル化エントロピーにより、プロモーター領域を 4 つのメチル化パターンに分類することができる。そこでまず、各プロモーター領域における各 CpG セグメントのメチル化エントロピーの分散が 0.05 以下のものを選別した。次に、各プロモーター領域のメチル化エントロピーの平均値と、それぞれのプロモーター領域にマップされた全てのリード中におけるメチル化 CpG と非メチル化 CpG 中に占めるメチル化 CpG の割合（メチル化レベル）を加味することで、HM-PCR 法と同様にヒト転写因子プロモーター領域を 4 つのメチル化パターンに分類した。



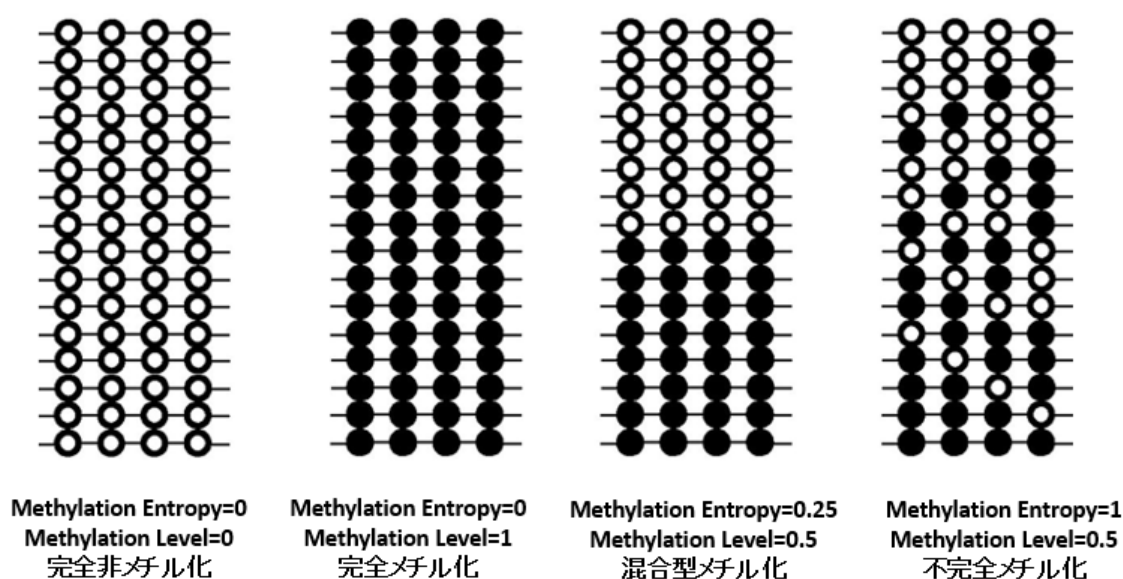


図 14：メチル化エントロピー及びメチル化レベルによる分類

## 4. 研究結果

### 4.1 Bismark によるバイサルファイトリードマッピング

Bismark による各サンプルにおけるバイサルファイトリードのレファレンスゲノムへのマッピング結果を表 9 に示す. すべてのサンプルにおいて 50%以上のリード配列が一意にレファレンスゲノム上にマッピングされた.

表 9：各サンプルにおけるバイサルファイトリードのマッピング結果

サンプル名	マッピング率	リード数	マッピングできたリード数
GSM848927 (末梢血 26 歳)	54%	約 2.8 億	約 1.5 億
GSM774850 (新生児)	80%	約 5.5 億	約 4.4 億
GSM1134680 (脳組織)	73%	約 5.4 億	約 4 億

### 4.2 BisSNP によるヘテロ SNP サイトの抽出

各サンプルにおいて, マップされたバイサルファイトリードから BisSNP により抽出されたヘテロ SNP の個数を表 10 に示す. すべてのサンプルにおいて 100 万~200 万程度のヘテロ SNP が同定された.

表 10：BisSNP により抽出したヘテロ SNP の数

サンプル名	同定した SNP の数	抽出したヘテロ SNP の数
GSM848927 (末梢血 26 歳)	約 101 万	約 91 万
GSM774850 (新生児)	約 128 万	約 109 万

GSM1134680 (脳組織)	約 309 万	約 209 万
---------------------	---------	---------

### 4.3 Fisher's exact test によるアレル特異的メチル化領域候補の推定

各サンプルにおいて、リードから同定した各ヘテロ SNP サイトにおいて、これを含むリードのアレル間でのメチル化割合の違いをフィッシャーの正確確率検定で検討した結果を表 11 に示す。また、フィッシャーの正確確率検定から得られた p 値を FDR 補正した結果を表 12 に示す。すべてのサンプルにおいて 5 千～1 万程度のアレル特異的メチル化領域候補が同定された。

表 11：フィッシャーの正確確率検定の結果

サンプル名	検定した回数	棄却された帰無仮説の数 (ASM 候補)
GSM848927 (末梢血 26 歳)	約 90.1 万	約 5.3 万
GSM774850 (新生児)	約 107.9 万	約 8.6 万
GSM1134680 (脳組織)	約 178 万	約 8.8 万

表 12：FDR 調整の結果

サンプル名	ASM 候補(FDR 調整前)	ASM 候補(FDR 調整後)
GSM848927 (末梢血 26 歳)	約 5.3 万個	4756 個
GSM774850 (新生児)	約 8.6 万個	15870 個
GSM1134680 (脳組織)	約 8.9 万個	9017 個

### 4.4 アレル特異的メチル化を受けたヒト転写因子遺伝子プロモーターの推定

4.3 で同定したアレル特異的メチル化領域候補のうち、1,068 個のヒト転写因子遺伝子プロモーター領域と重複するものを検索した結果、表 13 のように 8 個存在した。これら 8 個のうち 3 個が HM-PCR 法の結果においてもメチル化アレルと非メチル化アレルの混合型であった。

表 13 プロモーター領域にマッピングされたアレル特異的メチル化領域

chomos ome	SNP position	SNP name	promoter_name	promoter_ start	promoter_ end	sample_n ame
---------------	-----------------	-------------	---------------	--------------------	------------------	-----------------

chr4	53618	rs10000	0508240933_00840_0012_01F#B0141#ZNF	52789	53728	GSM8489
		110	595#zinc			27
chr11	47448547	rs61895	0504141409_00295_0002_01F#A0101#PSM	47447875	47448864	GSM7748
		077	C3#proteasome			501
chr16	3494207	rs27242	0504141409_00523_0004_01F#A0033#ZNF	3493783	3494347	GSM7748
			597#zinc			50
chr19	57351911	rs12973	0508240933_00623_0030_01F#B0149#PEG	57351628	57352407	GSM1134
		605	3#paternally			680
chrX	15015130	rs22937	0504141409_00489_0003_01F#A0084#HMG	150150952	15015160	GSM1134
	1	38	B3#high-mobility		6	680
chr19	57352051	rs23023	0504141409_00007_0001_01F#A0048#ZIM2	57351943	57352811	GSM1134
		76	#zinc			680
chr19	57352051	rs23023	0508240933_00623_0030_01F#B0149#PEG	57351628	57352407	GSM1134
		76	3#paternally			680
chr10	12758425	rs10794	0504141409_00563_0009_01F#A0046#FAN	12758422	12758507	GSM1134
	7	035	K1#fibronectin	6	4	680

表 13 のアレル特異的メチル化を有すると考えられるヒト転写因子遺伝子プロモーター領域のうちの 1 つを図 15 に示すプロモーター. 図 15 では, SNP サイトが C のアレルでは, その周辺 DNA 配列がメチル化を受けないが, T のアレルの周辺 DNA 配列では, 半数以上のアレルがメチル化を受けている. このことからこのプロモーター領域は, アレル特異的メチル化を受けていると考えられ, 表 13 の残りのヒト転写因子遺伝子プロモーターも, 同様にアレル特異的メチル化を受けていることが強く示唆された.

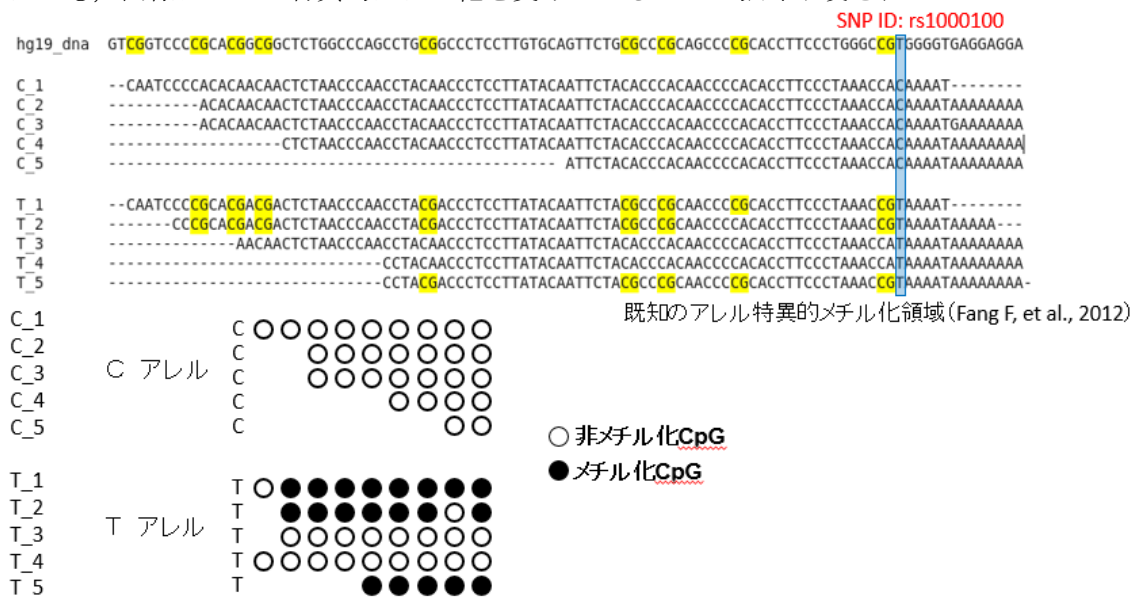


図 15 アレル特異的メチル化を有するヒト転写因子遺伝子プロモーター領域の例

## 4.5 アレル特異的メチル化を受けたヘテロ SNP 周辺配列の特徴

4.3 で同定した全てのアレル特異的メチル化領域候補において, メチル化の多いアレルと非メチル化が多いアレルに分け, それぞれのアレルにマップされたリードにおける SNP 配列とその隣接配列を調べた. 図 16 は, SNP 位置とその隣接位置における各塩基の出現頻度を Sequence LOGO[6] ソフトウェアを用いて視覚的に表示した結果である. この結果から, SNP によってアレル特異的メチル化 CpG が一つ作られ, 周辺の CpG にも

アレル特異的メチル化が波及する可能性が示唆された。

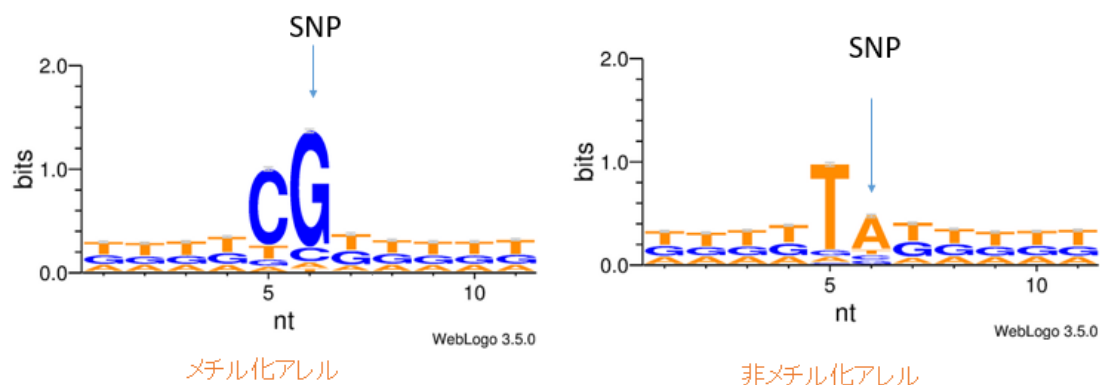


図 16 アレル特異的メチル化領域の配列構成

## 4.6 メチル化エントロピー平均値によるプロモーターの分類

プロモーターメチル化エントロピーの平均値により，HM-PCR 法で調べた 1,068 個のヒト転写因子遺伝子プロモーター領域のメチル化状態を 4 つに分類した結果を表 14 に示す．また，多くのプロモーターにおいて，HM-PCR 法のメチル化分類結果とメチル化エントロピーによるメチル化分類結果が一致した．

表 14 ヒト転写因子遺伝子プロモーター領域のメチル化エントロピー分類結果

	非メチル化	完全メチル化	混合型メチル化	不完全メチル化
GSM848927 (末梢血 26 歳)	112(77)	0(0)	4(2)	1(0)
GSM774850 (新生児)	52(33)	0(0)	5(3)	1(0)
GSM1134680 (脳組織)	65(43)	0(0)	7(2)	1(0)
HM-PCR の結果	776	76	132	84

\*0内の数字は，HM-PCR と一致した数．

## 5. まとめと考察

ヘテロ SNP を用いたアレル特異的メチル化領域の直接推定では，すべてのサンプルにおいて 5 千～1 万程度のアレル特異的メチル化領域候補が同定されたにも関わらず，ヒト転写因子遺伝子プロモーター領域と重複するものは 8 個しか存在しなかった．またこのうち 3 個は，HM-PCR 法の結果においてもメチル化アレルと非メチル化アレルの混合型であったため，これらは間違いなくアレル特異的メチル化領域であることが予想された．

また，ヒト転写因子遺伝子プロモーター領域における各 CpG セグメントのメチル化

エントロピーの値は、セグメント間で分散が高く、メチル化エントロピーの平均値によるプロモーターのメチル化分類が出来なかったものが多く存在した。しかしながら、多くのプロモーターにおいて、HM-PCR 法のメチル化分類結果とメチル化エントロピー平均値によるメチル化分類結果が一致した。

また今回の WGBS データを用いた各転写因子遺伝子プロモーター領域のメチル化分類結果と、HM-PCR 法から得られた分類結果の違いは、HM-PCR 法における制限酵素によるゲノムの不完全消化に由来するメチル化状態の誤判定によることが考えられる。

## 謝辞

本論文を終えるにあたり、日ごろから有益なご指導を賜りました山田洋一准教授に心から感謝いたします。一部のプログラムの作成支援を行って頂いた科学技術短期留学生の VEHVILAINEN TIMO TANELI さんに感謝いたします。また、共に研究を行った博士前期課程 2 年の佐々木将さん、辻郁奈さん、学部 4 年の豊田峰都さん、松川昌生さん、原泰知郎さんに感謝いたします。

## 参考文献

- [1] Clare Stirzaker, Phillippa C. Taberlay, Aaron L. Statham and Susan J. Clark. Mining cancer methylomes: prospects and challenges. Trends in genetics. 2014.
- [2] Yoichi Yamada, Hidemi Watanabe, Fumihito Miura, Hidenobu Soejima, Michiko Uchiyama, Tsuyoshi Iwasaka, Tsunehiro Mukai, Yoshiyuki Sakaki, and Takashi Ito. A Comprehensive Analysis of Allelic Methylation Status of CpG Islands on Human Chromosome 21q. Genome Res. 2004.
- [3] Felix Krueger and Simon R. Andrews; Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011.
- [4] Liu Y, Siegmund KD, Laird PW and Berman BP. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biology. 2012.
- [5] Hehuang Xie, Min Wang, Alexandre de Andrade, Maria de F. Bonaldo, Vasil Galat, Kelly Arndt, Veena Rajaram, Stewart Goldman, Tadanori Tomita and Marcelo B. Soares; Genome-wide quantitative assessment of variation in DNA methylation patterns. Nucl Acids Res. 2011.
- [6] T D Schneider and R M Stephens. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990.