博士論文

ヒト ES 細胞の性質に関わる
新規エピジェネティックコードの探索

石川　靖久

目次

第1章　序論

1.1　本研究の背景・目的

DNA メチル化やヒストン修飾などのエピジェネティックな制御因子は遺伝子発現変化をもたらし、細胞分化の過程において多大な影響を与え、また生物の一生においてその存在部位と量は大きく変化する。さらに、これらのエピジェネティックファクターはメモリーとしての役割を持ち、分化後もその情報は引き継がれ保存される。そして生殖細胞形成時、胚発生、さらには人工多能性幹細胞（induced pluripotent stem cell : iPSC）生成時の細胞初期化（リプログラミング）等において、このエピジェネティックメモリーにはリセット等の大幅な変化が生じ、それは細胞が多能性を獲得する上で不可欠なものと考えられている[1]。中でも 5-hydroxymethylcytosine（5hmC）は、DNA 脱メチル化過程において 5-methylcytosine（5mC）が酸化されることにより生成する分子であり、特に ES 細胞（ESC）において多く存在し、多能性幹細胞の特性獲得におけるその重要性が報告されている[2, 3]。

また近年、“epigenetic code”という概念が提唱されるようになってきており、これは genetic code に対応した呼称であり、異なる細胞において異なる表現型を作り出す一連のエピジェネティックな機能を意味する。これを示す例として、ESC のプロモーター領域において H3K4me3 と H3K27me3 が共起する bivalent 領域が報告されている。この領域を持つ、development に関わる遺伝子群は発現が poised な状態に保たれ、それにより ESC 特有の性質をもたらす[4, 5]。他にも DNA メチル化（ヒドロキシメチル化）とヒストン修飾間の共起が、哺乳類において実験を基に報告されている。例えば Poly-comb Repressive Complex 2（PRC2）複合体は、bivalent promoter 領域へ 5mC 酸化酵素である TET タンパクを呼び寄せ、ESC の bivalent プロモーター領域を低メチル化状態に保つ[6, 7]。他にもエンハンサー領域において、パイオニア・トランスクリプション・ファクターは TET タンパクと協働して DNA 脱メチル化を行い、その部位にエピジェネティック修飾酵素が結合して H3K4me1 や H3K27ac を生成することによりクロマチンアクセシビリティを高める[8, 9]。これらは一種のエピジェネティックコードと考えることができる。しかし現在までに報告されているものはエピジェネティックファクター間の相互作用機構の一部に過ぎず、全体の解明には程遠いのが現状である。

そこで今回、ESC などの多能性幹細胞の性質に影響を与える可能性を持った、新たなエピジェネティックコードの発見を目的として研究を行った。特に DNA 脱メチル化機構の中間生成物であり ESC に特異的に多く存在する 5hmC に焦点を当て、共にコードを形成する可能性を持ったヒストン修飾群をヒト ESC（hESC）に関する公共データを活用して包括的に探索した。

まず始めに、この研究においてエピジェネティックコードを特定する上での条件を設定した。第一に遺伝子領域において共起が見られるエピジェネティックファクター群であること、第二にこの共起が ESC 特異的な遺伝子発現に影響を及ぼすこと。第三にこれらの結果として ESC 特有の性質をもたらすこと、以上の条件を満たすものとした。

本研究では、まず ESC における DNA・ヒストン修飾に関する 19 種類のエピジェネティックファクター群に関して、お互いに遺伝子領域における分布状況が類似しているペアを探索した。次にその共起が見られるファクター群に注目し、体細胞との比較による ESC 特異的遺伝子発現へ

の影響、及びこれに対する共起の重要性について調べた。さらにこの共起を持つ遺伝子群の ESC 生物学的性質に関わる特徴、及びこれら共起ファクター群の相互作用機序について調べた。最後に、生物学的分野におけるこのエピジェネティックコード候補の持つ意義及び可能性についての考察を行った。

1.2　本論文の構成

本論文は以下のように構成される。

第２章では、本研究の背景について述べる。

第３章では、本研究で用いた材料と方法について述べる。

第４章では、本研究で得られた結果を述べる。

第５章では、結果に対する考察を述べる。

第2章　背景

2.1　エピジェネティックコード

ヒストンコード仮説は、「ヒストン修飾が特定の組み合わせを形成し、暗号（コード）的な意味を持つことにより統合的にクロマチン機能の制御を行う。」というものである。この組み合わせは、お互いが協調的もしくは対立的に作用し、またその存在条件によっても遺伝子発現に異なる影響を及ぼすことが知られている。

一方先行研究における定義では、「エピジェネティックコードは、真核細胞において DNA の化学的変化、クロマチン修飾因子、ノンコーディング RNA を含む、基礎となる DNA 配列を変更しない、特定の細胞タイプにおけるすべてのエピゲノム修飾と制御の合計であり、主に DNA メチル化とヒストン修飾によって定義される。」となっている[10-12]。つまりエピジェネティックコードとは、DNA やヒストン修飾をはじめとしたエピジェネティック関連因子を包括的に含んだ組み合わせ及びその作用効果を指すと言える。個々の研究において対象となるエピジェネティック因子の多様性もあり、コードの構成要素及び定義は研究毎に必ずしも一定ではない。しかしヒストンコード仮説を基にすると、その必要条件には同じ領域における各種因子の共起、およびその協働作用がもたらすクロマチンアクセシビリティの変化などによる細胞特異的遺伝子発現の誘因、その結果としての細胞特有の性質の獲得であると考えられる。

2.2　DNA メチル化

エピジェネティックファクターの内、DNA メチル化は原核生物と真核生物とでは修飾様式が異なっており、原核生物においてはアデニンの４位窒素、シトシンの５位炭素や４位アミノ基にメチル基が付加されることにより、DNA 複製と修復、外来 DNA からの防護などが制御される[13] 。真核生物では DNA シトシンのピリミジン環の 5 位炭素原子へのメチル基の付加反応であり、その結果 5mC という分子が生成する。この修飾は細胞分裂後も娘細胞に継承され、また生殖・発生過程においては大規模なリプログラミングなどのダイナミックな変化を受ける。したがって DNA メチル化は発生・分化における重要な役割を持っており、遺伝子発現の調節、X 染色体の不活性化、ゲノムインプリンティング、がんの発生など様々な生命現象にも重要な影響を及ぼす[14]。体細胞組織では、DNA メチル化は通常 CpG ジヌクレオチド部位（シトシン-ホスホジエステル結合-グアニン）で起こる。ほ乳類においては全 CpG 部位の 60～90%がメチル化されている[15]。DNA のメチル化を触媒するものとして DNA メチル基転移酵素があり、哺乳類では Dnmt1、Dnmt3a、及び Dnmt3b [16]が同定されている。この内 Dnmt1 は DNA 複製時の維持型メチル化を、Dnmt3a と Dnmt3b は新たに DNA をメチル化する de novo 型メチル化を担っている。

また、5mC が遺伝子プロモーター部位に存在することにより転写が抑制されると考えられている。その機構は 2 つ存在し、1 つは転写因子の認識配列がメチル化されることで転写因子の結合が阻害され、その結果転写調節に影響を及ぼすというものであり、もう 1 つは 5mC を特異的に認識するタンパク質が結合し、その複合体が転写抑制因子として働くというものである。また 5mC はプロモーター領域だけでなく、遺伝子内領域（Gene body）や遺伝子間領域（Intergenic）にも多く存在し、これらの場合は逆に転写が活性化される傾向が見られる[17]。

2.3　DNA 脱メチル化機構

5mC は DNA メチル化により常に増え続けるものではなく、特に生殖細胞形成時や胚発生、さらには分化細胞の初期化の際に大規模な DNA の脱メチル化が起こり、5mC の量は著しく変化する[18]。この DNA 脱メチル化の際に、5mC が酸化されることにより生成する分子が 5hmC である。5mC は様々な組織、細胞で比較的一定して存在するのに対して、5hmC は神経細胞[19-21]や ESC のように強く認められる細胞がある一方で、ほとんど検出できない細胞もあり、その存在量は多様である。

DNA 上のメチル基は、受動的機構または能動的機構により取り除かれると考えられている。受動的脱メチル化機構は、DNA 複製時に鋳型鎖のメチル基を新生鎖にコピーする“維持メチル化”が起こらないことによるもので、これにより片側の鎖のみのメチル化（ヘミメチル化）状態となり、新生鎖 DNA にシトシンが取り込まれて、その結果 5mC が減少するものである[22]。一方能動的脱メチル化機構は、5mC が TET（Ten-Eleven Translocation）という酵素により酸化されることにより 5hmC に変換され、その後さらなる酸化を受けて 5fC（5-formylcytosine)、5caC（5-carboxylcytosine）へと変化していき、最終的にはチミン DNA グリコシラーゼ（Thymine DNA Glycosylase：TDG）という酵素により DNA 上のメチル基が除去されるというものである[23, 24]（図 1)。この能動的脱メチル化は、始原生殖細胞や一部の体細胞などにおいて観察されている。5hmC は古くからその存在が知られていたが[25] 、本格的に注目を浴びるようになったのは 5mC 酸化酵素である TET タンパクが発見されてからである[24]。現在、３種類の TET タンパク（TET1、TET2、TET3）が発見されており、このうち TET1 は ESC や始原生殖細胞、TET2 は ESC の他に様々な組織で、TET3 は卵細胞において高い発現が確認されている。

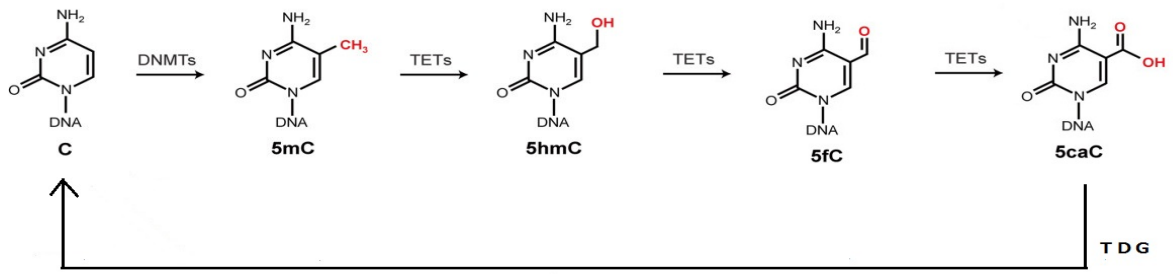


図 1：能動的 DNA 脱メチル化機構：シトシンのピリミジン環の 5 位炭素原子へメチル基が付加することにより 5mC（5-methylcytosine）が生成し、この分子が TET タンパクによる酸化を受けることにより、5hmC（5-hydroxymethylcytosine)、5fC（5-formylcytosine)、5caC（5-carboxylcytosine）へと変化していき、最後には TDG（Thymine DNA Glycosylase）により、シトシンからメチル基が除去される。

2.4　5hmC 検出法

2.4.1　hMeDIP 法

5mC に対する抗体を用いたメチル化シトシンの検出法を MeDIP 法（Methylated DNA Immuno Precipitation）といい、ヒストン修飾や転写因子結合領域を対象とするクロマチン免疫沈降法をメチル化シトシンに応用した手法である[26]。すなわち、抗体を用いてメチル化シトシンを含む DNA 断片を回収し、この断片の分布を次世代シークエンサーで解析することで、メチル化シトシンの多い領域をゲノムワイドに同定する方法である（図 2）。この方法のメリットは、CpG アイランドなど CpG 密度の高い（CpG 配列の多い）領域に対して検出感度が高い点である。また他のバイサルファイト処理を用いた方法などと比べると、ヒドロキシメチル化シトシンは認識していないことからメチル化シトシン特異的な検出も可能である。欠点としては、捕捉している断片の大きさが数百塩基であり、断片中のどの CpG がメチル化しているかは特定できないため、特定の領域の詳細なメチル化率の定量は困難である。この手法を 5hmC に対して適用したものが hMeDIP 法である[27]。

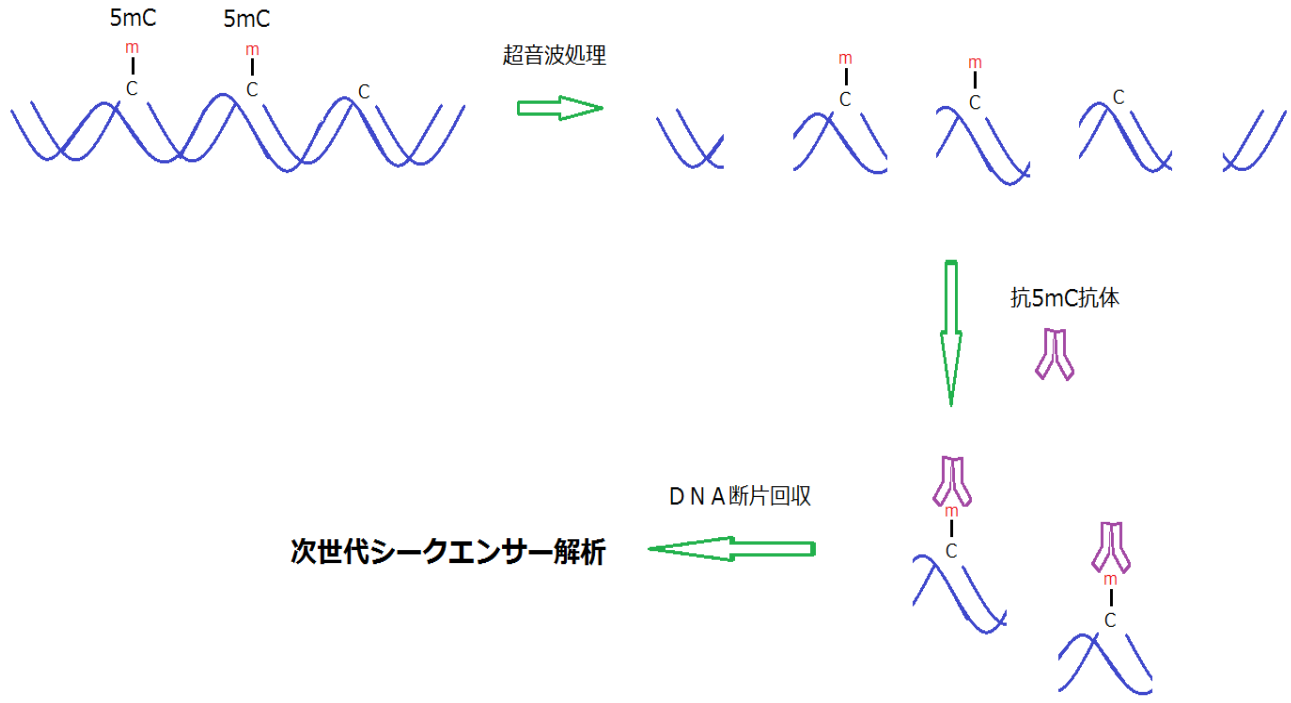


図 2：MeDIP 法

2.4.2　TAB-seq 法

β-GT（β-グルコシルトランスフェラーゼ）を利用する方法で、5hmC をあらかじめβ-GT によりグリコシル化して 5gmC に変換する。その後 TET タンパクで酸化させると、5gmC は酸化を受けずそのまま残るが、5mC は酸化を受けて 5fC、5caC へと変換される。その後 Bisulfite 処理をすることにより、全てのシトシン（C）と 5caC はウラシル（U）や 5caU へと変換される。これを Bisulfite-seq にかけると、5hmC を C として一塩基精度で検出することが可能になる[28, 29]（図 3）。

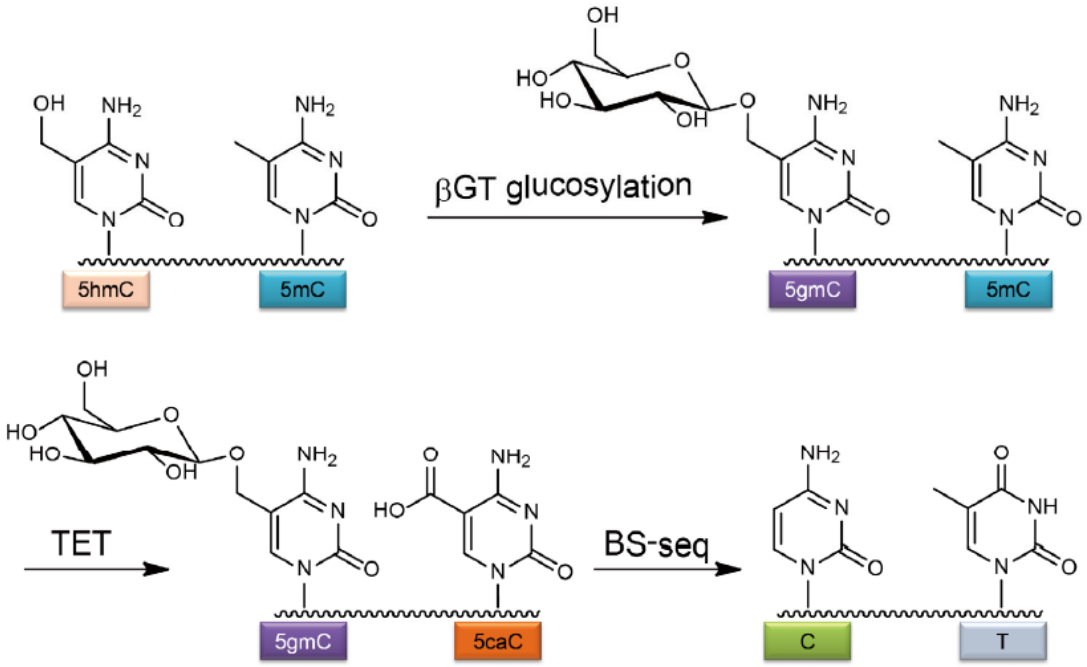


図 3：TAB-seq 法（出典：Single-base resolution analysis of DNA epigenome via high-throughput sequencing. Sci China Life Sci, 2016. 59(3): p. 219-26、Figure5 から引用）

2.5　非負値行列因子分解アルゴリズム

データ解析において、多数のサンプル間の類似度を調べる方法として階層的クラスタリングや主成分分析などがあるが、非負値行列因子分解（NMF：Non-negative Matrix Factorization）はこれら多変量解析に属するアルゴリズムの一つで、観測データ行列 X を 2 つのより低次元の基底行列 W および係数行列 H に因数分解するものであり[30-32]、次の式で表される。

$$\boldsymbol{X} \simeq \boldsymbol{WH}$$

X：観測データ行列（入力行列）　$X \in \mathbb{R}^{(N \times M)}$

W：基底行列　$W \in \mathbb{R}^{(N \times k)}$

H：係数行列　$H \in \mathbb{R}^{(k \times M)}$

これはさらに下記の形で表される。

$$x_m \simeq \sum_{k=1}^{K} w_k \, h_{k,m} \; (m = 1,2,\dots,M)$$

$$X = [x_1,\dots,x_M] = \left(x_{n,m}\right)_{N \times M}$$

$$W = [w_1,\dots,w_K] = \left(w_{n,k}\right)_{N \times K}$$

$$H = [h_1,\dots,h_M] = \left(h_{k,m}\right)_{K \times M}$$

非負値行列の名称は、各々の行列が負の要素を持たないことに由来する。NMF は機械学習法に分類され、信号処理から生物医学系データ解析まで幅広い応用が可能である。なお、基底行列 W は N 行 k 列、係数行列 H は k 行 M 列からなり、k のことを基底数（ランク）と呼ぶ。この式から $||X - WH||^2$ が最小となるように計算を行っていく。まず非負値からなる行列 W と H を用いて、H を固定して W の最適化を行う。次に W を固定して H を最適化するということを順に繰り返し、最終的に２つの行列WとHを求める。

重要な問題は、特定のランク k がサンプルを「意味のある」クラスターに分解するかどうかを判断することであり、そのためこのランクは NMF における重要なパラメーターとなる。この値を決定する一般的な方法は、異なる値を試してそれぞれの結果の品質尺度を計算し、それにより最適な値を選択することである。そのために様々な方法が提案されており、コーフェン相関係数が減少し始める最初の値や[33]、RSS 曲線が変曲点を示す最初の値[34]、さらには RSS の減少がランダムデータから取得した RSS の減少よりも小さい最小値[35]を採用するなどの方法がある。

第3章　材料と方法

3.1　データセット

表 1：本研究に使用した公共データ

Type	Cell line	Modification	Accession Number	NGS protocol
Fibroblast	IMR90	5hmC	GSM909339 [36]	hMeDIP-seq
	CRL2097	5hmC	GSM909335 [36]	hMe-Seal
	GM0011	5hmC	GSM909337 [36]	
	IMR90	Input	GSM909321 [36]	
hESC	H1	5hmC	GSM747152 [3]	
			SRR299100 [3]	
		Input	GSM747151 [3]	
	HUES48	5hmC	GSM909322 [36]	
	HUES49	5hmC	GSM909323 [36]	
	HUES53	5hmC	GSM909324 [36]	
	H1	5hmC	GSM882245 [29]	TAB-seq
Mouse ESC		5hmC	GSM711882 [2]	hMeDIP-seq
		Input	GSM711884 [2]	
Fibroblast		5mC	GSM1462769 [37]	MeDIP-seq
hESC	H1	5mC	SRR042409	MBD-seq
		Input	SRR042411	
hESC	H1	H2BK5ac	GSM605302 ※	ChIP-seq
		Input	GSM605333 ※	
		H2BK12ac	GSM605296 ※	
		Input	GSM605333 ※	
		H2BK15ac	GSM605298 ※	
		Input	GSM605333 ※	
		H3K4ac	GSM667624 ※	
		Input	GSM605334 ※	
		H3K9ac	GSM605323 ※	
		Input	GSM605333 ※	
		H3K14ac	GSM667615 ※	
		Input	GSM605334 ※	
		H3K4ac	GSM667624 ※	
		Input	GSM605334 ※	
		H3K18ac	GSM605304 ※	

		Input	GSM605333 ※	
		H3K23ac	GSM667618 ※	
		Input	GSM605334 ※	
		H3K27ac	GSM466732 ※	
		Input	GSM605333 ※	
Fibroblast	IMR90	H3K4me1	GSM1418961 [38]	
			SRR037547	
			SRR037548	
			SRR037549	
			SRR037552	
			SRR037597	
			SRR037551	
			SRR037556	
		Input	GSM1418971 [38]	
			SRR037634	
			SRR037635	
			SRR037636	
			SRR037637	
hESC	H1	H3K4me1	SRR020519	
			SRR067947	
			SRR029619	
			GSM433177 [39]	
			GSM605312 ※	
			SRR020519	
Mouse ESC		H3K4me1	GSM970225 [40]	
		Input	GSM970219 [40]	
hESC	H1	Input	SRR020520	
			SRR067973	
			SRR067970	
			GSM605335 ※	
		H3K4me2	GSM602260 ※	
		Input	GSM605333 ※	
		H3K4me3	GSM469971 ※	
		Input	GSM605333 ※	
Fibroblast	IMR90	H4K8ac	GSM521919 ※	
			GSM521921 ※	
			GSM521922 ※	

			GSM521923 ※	
		Input	GSM521926 ※	
			GSM521927 ※	
			GSM521928 ※	
			GSM521929 ※	
hESC	H1	H4K8ac	GSM896166 ※	
			SRR445380	
		Input	GSM605333 ※	
Mouse ESC		H4K8ac	DRR022261	
		Input	DRR022259	
hESC	H9	H4K8ac	GSM667638 ※	
		Input	GSM706081 ※	
Fibroblast	HDF		GSM1282330 [41]	RNA-seq
			GSM1378026 [41]	
hESC	hESO7		GSM1282324 [41]	
	hESO8		GSM1282325 [41]	
	H1		GSM1888669 [42]	
hESC	HUES8	TET	GSM2642522	ChIP-seq
		Input	GSM2642525	

※　UCSD Human Reference Epigenome Mapping Project から引用

3.2　データの取得と解析前処理

3.2.1　公共データの取得

本研究でのデータ解析の流れを図 5 に纏めた。NGS データは Gene Expression Omnibus (GEO、https://www.ncbi.nlm.nih.gov/geo/) 及び DNA Data Bank of Japan （ DDBJ 、http://www.ddbj.nig.ac.jp）データベースから Fastq 及び BED ファイル形式のデータをダウンロードした（表 1)。Fastq 形式は配列 ID、リードの塩基配列、リードのクオリティ情報などから構成される、次世代シークエンサーから出力されたリード情報である。本研究で使用した 5mC、5hmC、ヒストン修飾、及び TET に関するデータは、免疫沈降法を利用または応用した検出法によるものである。これに加えて 5hmC については Bisulfite 法を応用した一塩基検出精度を持つ TAB-seq 法によるデータを使用し、また遺伝子発現データとして RNA-seq データを使用した。

3.2.2　クオリティコントロール

塩基配列をシークエンシングするときにエラーが発生する。そのエラーの生じる確率を P-err とすると、クオリティスコア Q は次の式で表される。

$$Q = -10 \log_{10} P\text{-}err$$

クオリティスコア Q は次世代シークエンサーにより読み取られた塩基の信頼度を表し、クオリティスコアが 20 の場合、シークエンシングエラーの発生確率は 1%となり、塩基の信頼度は 99%である。

取得したデータに対し、PRINSEQ（http://prinseq.sourceforge.net/）ソフトウェアによるクオリティコントロールを行った。一例として、まずリードに含まれる全塩基の 80%以上がクオリティ 20 未満のリードを除去し、次に 3 末端からクオリティ 20 未満の塩基をトリミングした。その結果、長さが 30 塩基未満になったリードを除去した。さらに tagcleaner [43]ソフトウェアを使用して、5 末端からアダプター配列を取り除いた（図 4）。

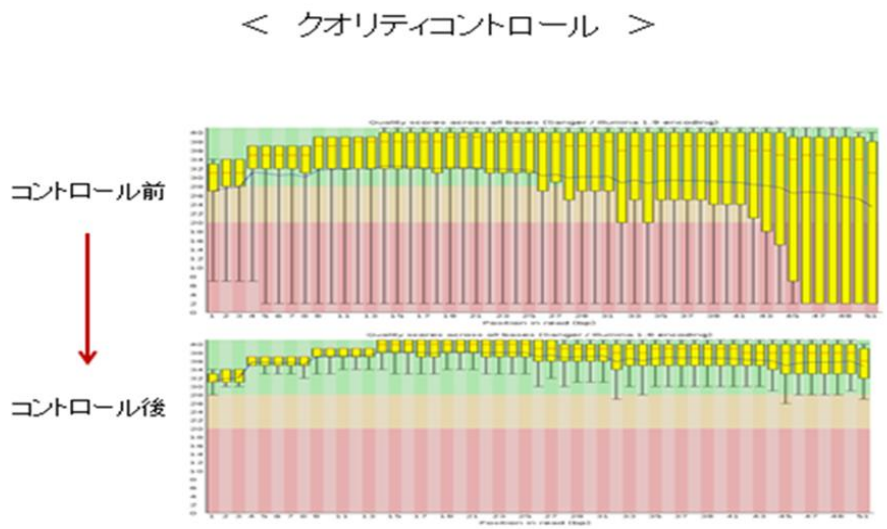


図 4：クオリティコントロールによるリードクオリティの変化：横軸はリードの塩基位置（左が 5 末端側、右が 3 末端側）、縦軸はクオリティ値を表している。赤色の領域はクオリティが 20 未満であることを表し、緑色の領域は 28 以上であることを表しており、緑色の領域の方が読み取られた塩基の信頼度が高い。

3.2.3　マッピング

hMeDIP-seq データについては bowtie2 [44]を使用してヒト・リファレンスゲノム（hg19）に対しマッピングを行い、遺伝子発現データ（RNA-seq）については tophat2 [45]を使用してマッピングを行った。tophat2 はリードが 2 つのエクソンにまたがっている場合も、ギャップを考慮して対応できる利点がある。TAB-seq データについては Bismark（https://www.bioinformatics.babraham.ac.uk/projects/bismark/）ソフトウェアを使用した。複数個所にマップされたリードは除去した（図 5）。

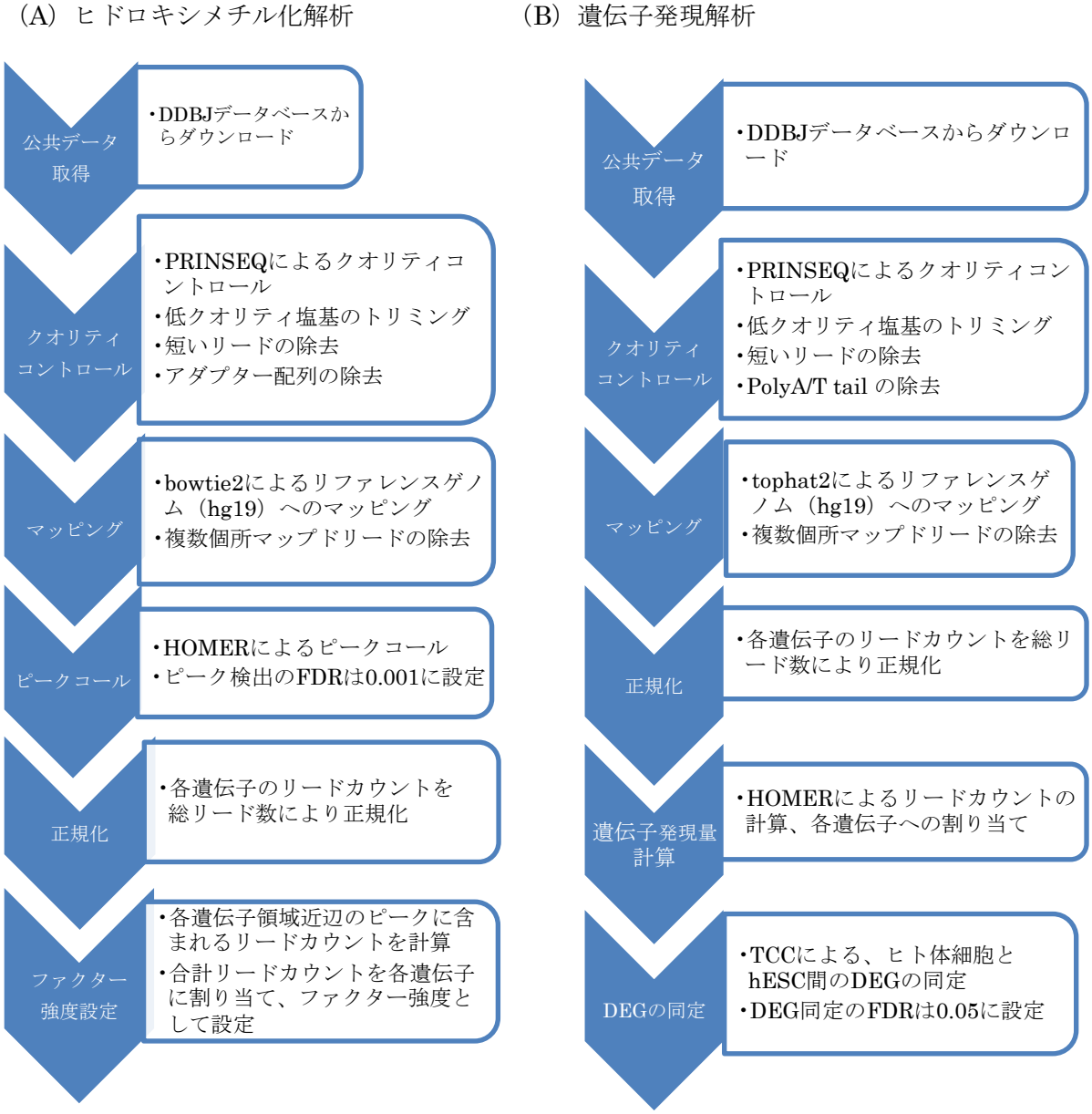


図 5：本研究でのデータ処理の流れ

3.2.4　ピークコーリング、正規化、及び遺伝子ファクター強度の設定

5hmC、5mC、及びヒストン修飾に関するマップドデータに対しては、HOMER ソフトウェア（http://homer.ucsd.edu/homer/index.html）を使用してピークコールを行った。ピーク検出に伴う FDR（False Discovery Rate）は 0.001 に設定した。次に各ピークに含まれる 5hmC リード数を総リード数で正規化した。これは次の式で定義される。

$$\textit{正規化後のリード数} \quad = \quad \frac{\textit{マップされたリード数}}{\textit{総リード数}} \quad \times \quad K$$

定数 K は 1,000,000

正規化後の遺伝子領域近辺のリードカウントを合計し、各遺伝子に割り当てることにより、遺伝子の持つエピジェネティックファクター強度として設定した。TAB-seq データに関しては、遺伝子領域に含まれる 5hmC の数を強度として設定した。RNA-seq データに関しても、マップドリードを HOMER による正規化及び遺伝子への割り当てを行い、各遺伝子の発現量を求めた。

3.3　クラスター解析における距離及び方法の設定

R の hclust 関数を用いて遺伝子発現量、5mC、及び 5hmC 強度に基づくクラスタリングを行った。距離にはユークリッド距離を使用し、方法として最短距離法と最遠距離法の中間的方法で、外れ値に強い群平均法を採用した（図 7-9）。

3.4　DEG の同定

遺伝子発現変化と各エピジェネティックファクター分布変化の関連性について調べるため、3.2 節で得られた各遺伝子の発現量を基に cufflinks（http://cole-trapnell-lab.github.io/cufflinks/）ソフトウェアを使用してヒト体細胞と hESC 間での発現量有意変動遺伝子（Deferentially Expressed Gene : DEG）を同定した。DEG の同定には R パッケージの TCC（http://bioconductor.org/packages/release/bioc/html/TCC.html）を使用し、同定基準として Fibroblast と hESC 間で FPKM の Fold Change（FC）の絶対値が 2 倍以上でかつ、DEG 判定の q-value が 0.05 未満の遺伝子群とした（図 6、青色点）。その結果 5029 個の DEGs が同定された。さらにこの DEGs を、ESC における発現量が体細胞の 2 倍以上の遺伝子群（DEG_ES_up）、半分以下の遺伝子群（DEG_ES_down）に分類し、その他の遺伝子群を non_DEG とした。

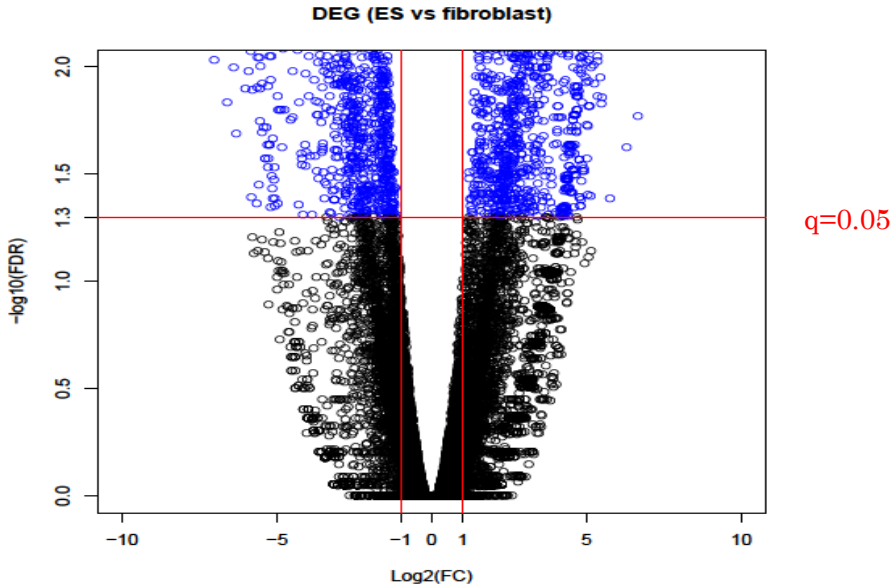


図 6：DEG の定義。青い点が DEG を表している。

3.5　ファクター有意変動遺伝子群に基づいた全遺伝子のグルーピング

体細胞と hESC 間で、有意（FDR < 0.01）に H3K4me1 の強度が異なる遺伝子群を同定し、これをもとに全遺伝子を次の 3 個のグループに分類した。

H3K4me1_up ・・・hESC において H3K4me1 が有意に増加している遺伝子群

H3K4me1_down ・・・ hESC において H3K4me1 が有意に減少している遺伝子群

H3K4me1_NSC（No Significant Change） ・・・ ①と②のどちらにも含まれない遺伝子群

上記の手順を H4K8ac についても実施した。次にそれぞれの各ファクターに関する 3 個ずつのグループを組み合わせて、合計 9 個の遺伝子グループを作成した。

① H3K4me1_up and H4K8ac_up (879 genes)
② H3K4me1_up and H4K8ac_down (0 gene)
③ H3K4me1_down and H4K8ac_up （1 gene)
④ H3K4me1_down and H4K8ac_down (423 genes)
⑤ H3K4me1_up and H4K8ac_NSC (432 genes)
⑥ H3K4me1_down and H4K8ac_NSC (224 genes)
⑦ H3K4me1_NSC and H4K8ac_up (2456 genes)
⑧ H3K4me1_NSC and H4K8ac_down (1099 genes)
⑨ H3K4me1_NSC and H4K8ac_NSC (15318 genes)

第4章　結果

4.1　5hmC の多能性幹細胞における特異的分布

4.1.1　体細胞・多能性幹細胞間における分布比較

先行研究において ESC に DNA ヒドロキシメチル化の中間産物である 5hmC が多く存在することが報告されているが[2, 3]、この 5hmC 分布の、hESC や hiPSC などの多能性幹細胞における特異性を調べるためクラスター解析を実施した。これに先立ち、ヒト多能性幹細胞及び体細胞（線維芽、腎臓、肝臓、肺）に関する各遺伝子の発現量データを用いてクラスター解析を行ったところ、多能性幹細胞群は体細胞群とは明らかに離れたクラスターを形成した（図 7）。

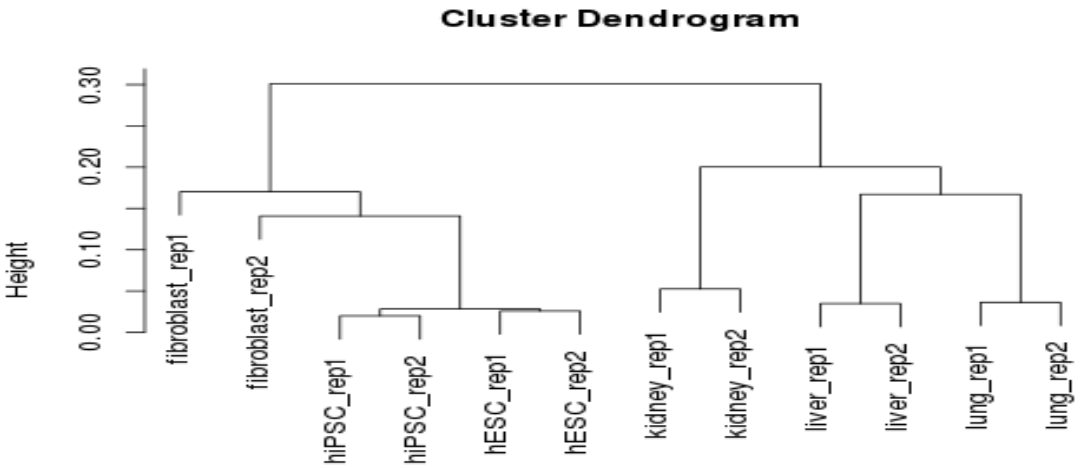


図 7：遺伝子発現に基づくクラスター解析。体細胞群は hESC や hiPSC とは明らかに離れたクラスターを形成している。

同様の方法で、各遺伝子の 5mC 及び 5hmC 強度に基づいたクラスター解析を行ったところ、やはり多能性幹細胞群のクラスターは、体細胞群とは遠く離れたクラスターを形成した（図 8-9）。

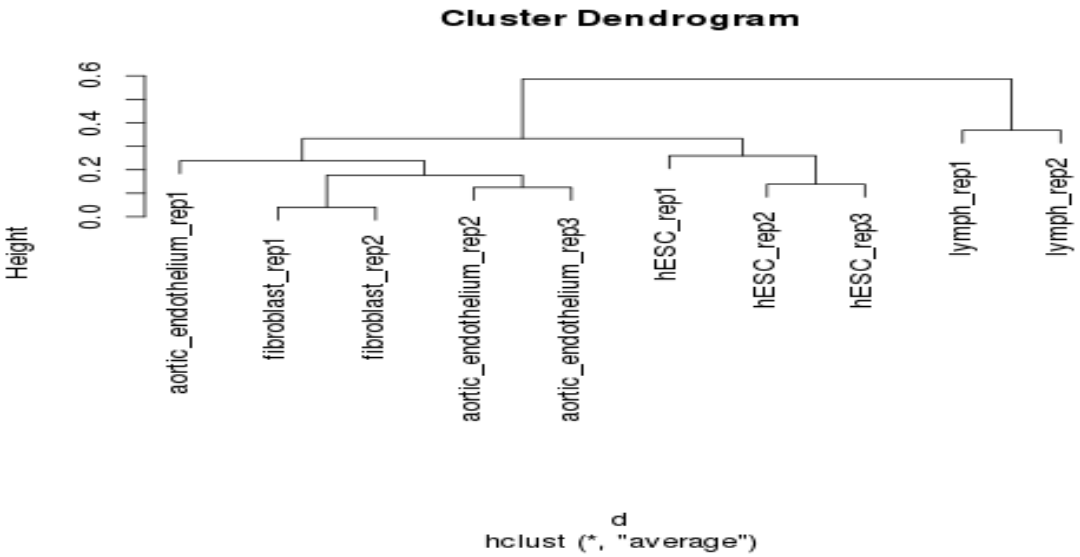


図 8：5mC 分布に基づくクラスター解析。hESC は体細胞群（線維芽、内皮、白血球）と離れたクラスターを形成しており、5mC 分布が両者の間で顕著に異なることを示している。

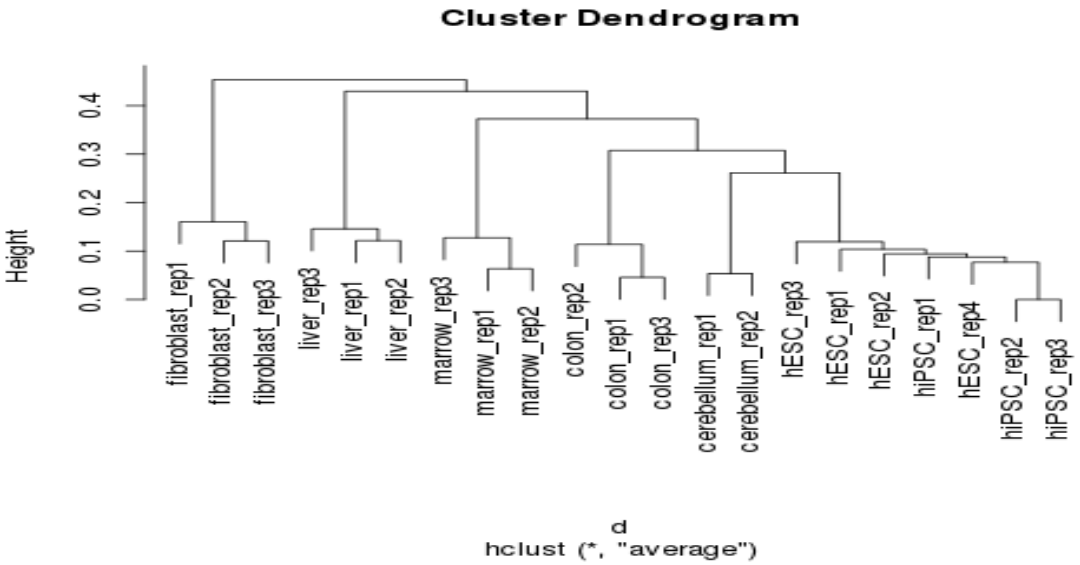


図 9：5hmC 分布に基づくクラスター解析。hESC と hiPSC は同じクラスターを形成しており、5mC と同様に 5hmC の分布が多能性幹細胞・体細胞間で顕著に異なることを示している。

4.1.2　ゲノムアノテーションに基づいた 5hmC 分布状況

マップドリードをゲノムアノテーションに基づいた各領域に割り当て、5hmC のヒト体細胞、hESC、及び hiPSC の DNA 各領域での存在傾向を調べたところ、ヒト線維芽細胞では遺伝子間領域以上に Intron に分布する傾向が強く見られた。また hESC や hiPSC では、Intergenic、Exon、TSS 等の Intron 以外の領域における分布割合が高く、お互いによく似た分布傾向が見られた（図 10）。

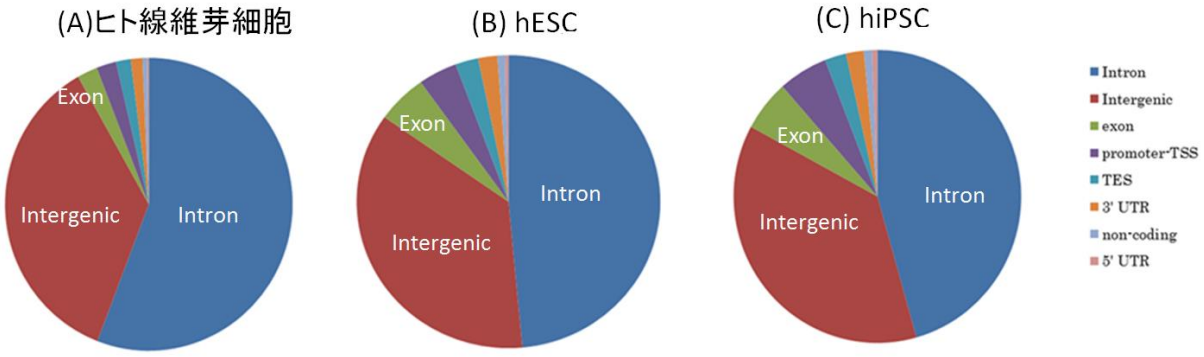


図 10：ヒト体細胞（線維芽細胞）と多能性幹細胞における 5hmC の分布傾向。多能性幹細胞においては、5hmC は Intergenic（Enhancer 含む）、Exon、TSS 近辺のプロモーター領域に多く分布する傾向が見られる。

4.1.3　遺伝子領域における 5hmC 分布状況

遺伝子領域における 5hmC 分布状況を、全遺伝子を平均化して調べたところ、転写開始点（TSS）近辺（TSS 前後 2000bp、図 11 における赤色円）において、ヒト体細胞よりも hESC の方の 5hmC マップドリード数が最大で 30%増加しており、転写終了点（TES）においても hESC の方に若干の増加傾向が見られた。もう一つ特徴的な点としては、5hmC リードのピークは正確には転写開始点ではなく、転写開始点の上流 300～1000bp の位置に集中していた。

一方、hiPSC の場合は、hESC よりもさらに全体的に 5hmC リード数が増加する傾向が見られた。ヒト体細胞と比較して、hiPSC では TSS 近辺で最大で 90%、その他の部位で約 20%、5hmC リード数が増加していた。この分布の違いから、特に TSS 近辺において 5hmC リード数が顕著に増加している hESC と比べると、hiPSC では DNA ヒドロキシメチル化について hESC ほどの位置特異性はないものと思われる。この他にも TES 近辺に関しては hiPSC と hESC の 5hmC リード数が同じレベルとなっている（図 11）。

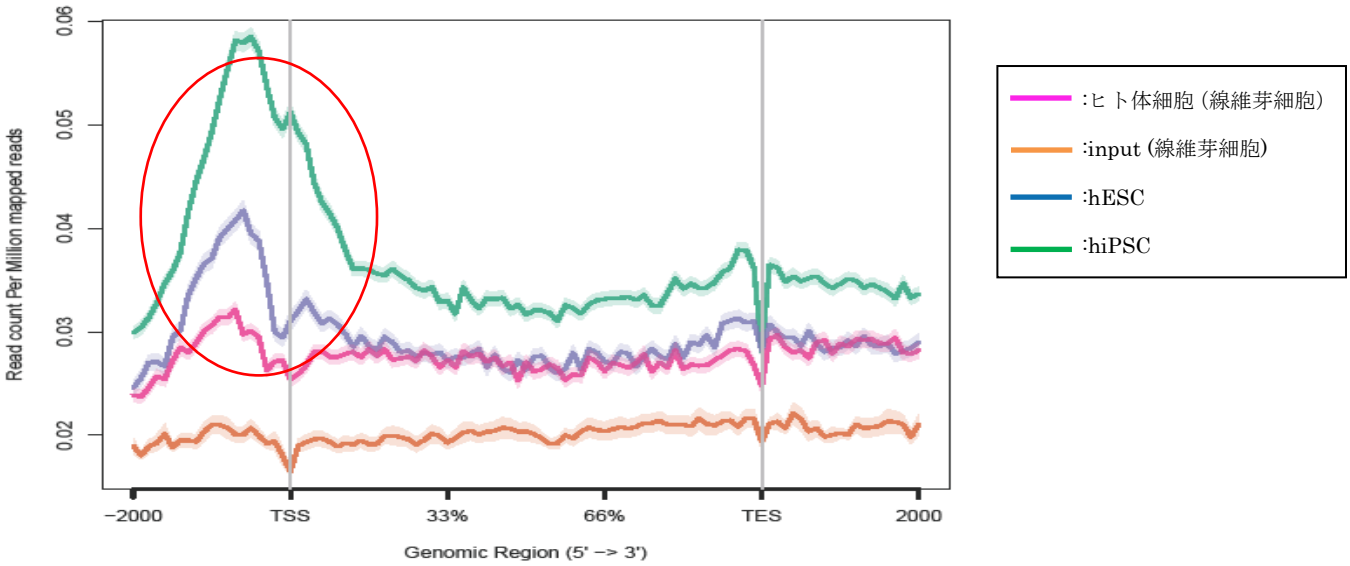


図 11：ヒト体細胞と多能性幹細胞の遺伝子領域における 5hmC 分布比較。横軸は遺伝子領域における位置を示し、縦軸は総リード数で正規化された 5hmC リードカウントを表している。グラフは全遺伝子の平均を表している。Input として、ヒト体細胞（線維芽細胞）DNA を使用している。

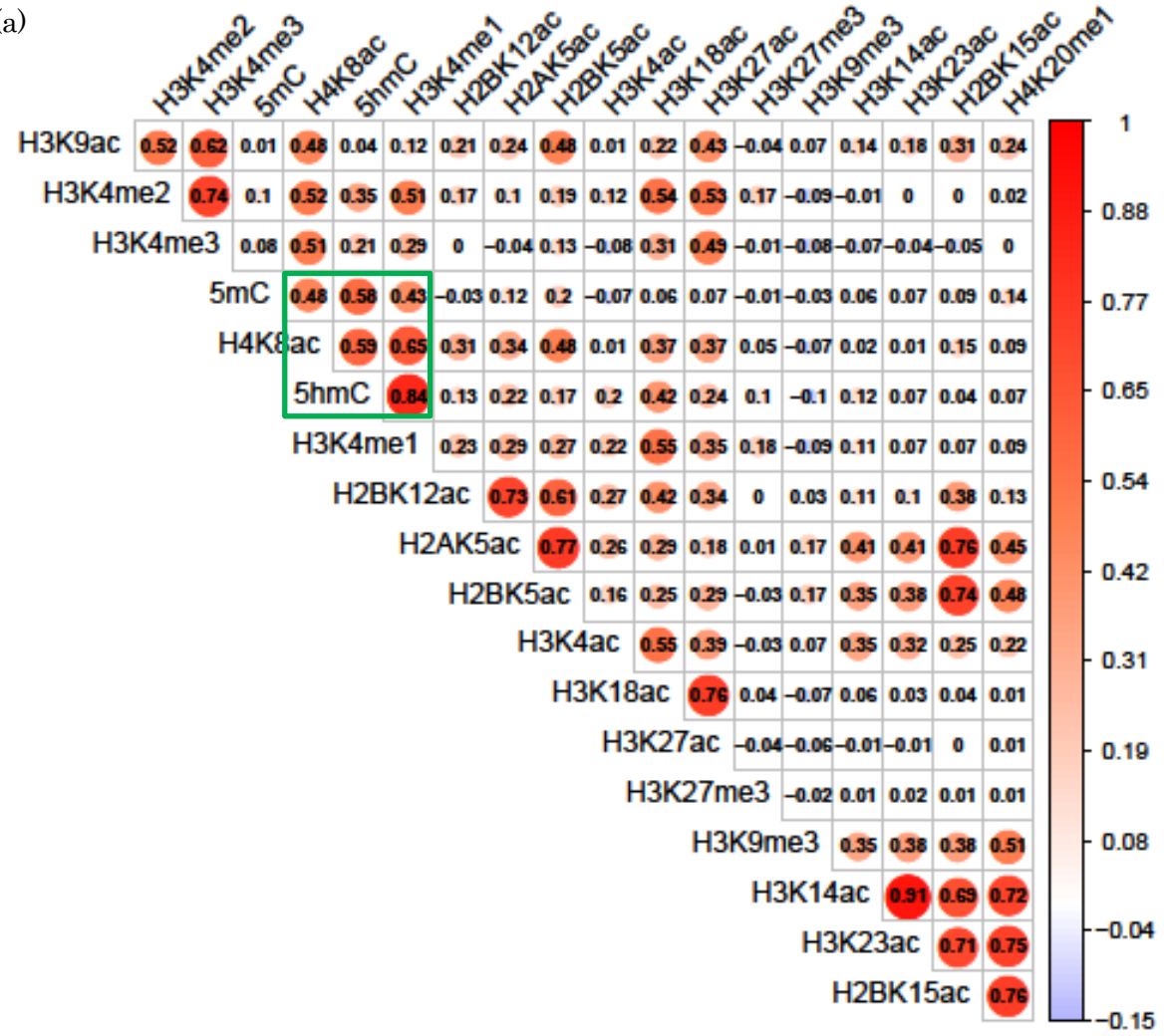
4.2　共起エピジェネティックファクターの探索

4.2.1　エピジェネティックファクター間の相関

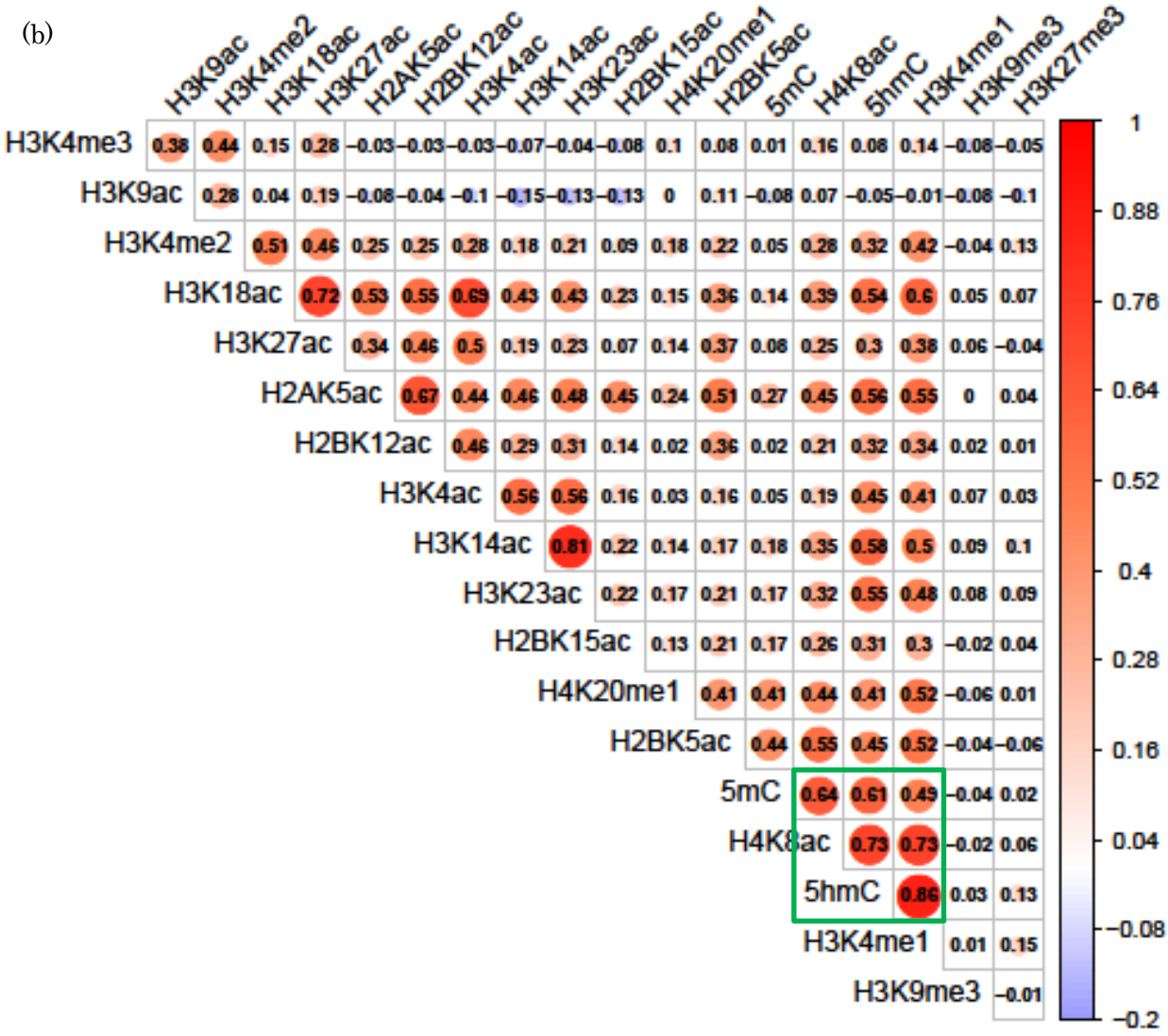
前節において確認された、多能性幹細胞に特異的な分布を示す 5hmC を含む、DNA 及びヒストン修飾群（全 19 種類）に関する hESC の公共 NGS データを用いて、ゲノム上の分布状況が類似しているものを探索した。

始めにゲノムを 50K-bp 毎の window に区切り、それぞれの window に含まれる各ファクター強度に基づいて、ファクター間 Pearson 相関係数（Peason’s Correlation Coefficeient：PCC）を求めたところ、5hmC と比較的高い相関を持つ（分布の類似度が高い）ヒストン修飾として、H3K4me1 や H4K8ac が存在することがわかった（PCC：5hmC vs H4K8ac=0.59、5hmC vs H3K4me1=0.84）（図 12a、緑色枠内）。次に遺伝子領域に注目して同様に相関係数を求めたところ、gene body 領域においてこれらの値はさらに高いものとなった（PCC : 5hmC vs H3K4me1 = 0.86、5hmC vs H4K8ac = 0.73）（図 12b、緑色枠内）。一方で、Promoter 領域における PCC 値は、gene body の場合よりも低い値（5hmC vs H3K4me1 = 0.65、5hmC vs H4K8ac = 0.18）となった（図 12c）。ESC において 5hmC と H3K4me1 の共起が観察されることは先行研究において既に報告されているが[46]、5hmC と H4K8ac の共起は今回の研究により新たに発見したものである。

(a)


H3K4me2
H3K4me3
5mC
H4K8ac
5hmC
H3K4me1
H2BK12ac
H2AK5ac
H2BK5ac
H3K4ac
H3K18ac
H3K27ac
H3K27me3
H3K9me3
H3K14ac
H3K23ac
H2BK15ac
H4K20me1
H3K9ac 0.52 0.62 0.01 0.48 0.04 0.12 0.21 0.24 0.48 0.01 0.22 0.43 −0.04 0.07 0.14 0.18 0.31 0.24
H3K4me2 0.74 0.1 0.52 0.35 0.51 0.17 0.1 0.19 0.12 0.54 0.53 0.17 −0.09 −0.01 0 0 0.02
H3K4me3 0.08 0.51 0.21 0.29 0 −0.04 0.13 −0.08 0.31 0.49 −0.01 −0.08 −0.07 −0.04 −0.05 0
5mC 0.48 0.58 0.43 −0.03 0.12 0.2 −0.07 0.06 0.07 −0.01 −0.03 0.06 0.07 0.09 0.14
H4K8ac 0.59 0.65 0.31 0.34 0.48 0.01 0.37 0.37 0.05 −0.07 0.02 0.01 0.15 0.09
5hmC 0.84 0.13 0.22 0.17 0.2 0.42 0.24 0.1 −0.1 0.12 0.07 0.04 0.07
H3K4me1 0.23 0.29 0.27 0.22 0.55 0.35 0.18 −0.09 0.11 0.07 0.07 0.09
H2BK12ac 0.73 0.61 0.27 0.42 0.34 0 0.03 0.11 0.1 0.38 0.13
H2AK5ac 0.77 0.26 0.29 0.18 0.01 0.17 0.41 0.41 0.76 0.45
H2BK5ac 0.16 0.25 0.29 −0.03 0.17 0.35 0.38 0.74 0.48
H3K4ac 0.55 0.39 −0.03 0.07 0.35 0.32 0.25 0.22
H3K18ac 0.76 0.04 −0.07 0.06 0.03 0.04 0.01
H3K27ac −0.04 −0.06 −0.01 −0.01 0 0.01
H3K27me3 −0.02 0.01 0.02 0.01 0.01
H3K9me3 0.35 0.38 0.38 0.51
H3K14ac 0.91 0.69 0.72
H3K23ac 0.71 0.75
H2BK15ac 0.76
1
0.88
0.77
0.65
0.54
0.42
0.31
0.19
0.08
−0.04
−0.15

(b)


H3K9ac
H3K4me2
H3K18ac
H3K27ac
H2AK5ac
H2BK12ac
H3K4ac
H3K14ac
H3K23ac
H2BK15ac
H4K20me1
H2BK5ac
5mC
H4K8ac
5hmC
H3K4me1
H3K9me3
H3K27me3
H3K4me3 0.38 0.44 0.15 0.28 −0.03 −0.03 −0.03 −0.07 −0.04 −0.08 0.1 0.08 0.01 0.16 0.08 0.14 −0.08 −0.05
H3K9ac 0.28 0.04 0.19 −0.08 −0.04 −0.1 −0.15 −0.13 −0.13 0 0.11 −0.08 0.07 −0.05 −0.01 −0.08 −0.1
H3K4me2 0.51 0.46 0.25 0.25 0.28 0.18 0.21 0.09 0.18 0.22 0.05 0.28 0.32 0.42 −0.04 0.13
H3K18ac 0.72 0.53 0.55 0.69 0.43 0.43 0.23 0.15 0.36 0.14 0.39 0.54 0.6 0.05 0.07
H3K27ac 0.34 0.46 0.5 0.19 0.23 0.07 0.14 0.37 0.08 0.25 0.3 0.38 0.06 −0.04
H2AK5ac 0.67 0.44 0.46 0.48 0.45 0.24 0.51 0.27 0.45 0.56 0.55 0 0.04
H2BK12ac 0.46 0.29 0.31 0.14 0.02 0.36 0.02 0.21 0.32 0.34 0.02 0.01
H3K4ac 0.56 0.56 0.16 0.03 0.16 0.05 0.19 0.45 0.41 0.07 0.03
H3K14ac 0.81 0.22 0.14 0.17 0.18 0.35 0.58 0.5 0.09 0.1
H3K23ac 0.22 0.17 0.21 0.17 0.32 0.55 0.48 0.08 0.09
H2BK15ac 0.13 0.21 0.17 0.26 0.31 0.3 −0.02 0.04
H4K20me1 0.41 0.41 0.44 0.41 0.52 −0.06 0.01
H2BK5ac 0.44 0.55 0.45 0.52 −0.04 −0.06
5mC 0.64 0.61 0.49 −0.04 0.02
H4K8ac 0.73 0.73 −0.02 0.06
5hmC 0.86 0.03 0.13
H3K4me1 0.01 0.15
H3K9me3 −0.01
1
0.88
0.76
0.64
0.52
0.4
0.28
0.16
0.04
−0.08
−0.2

(c)

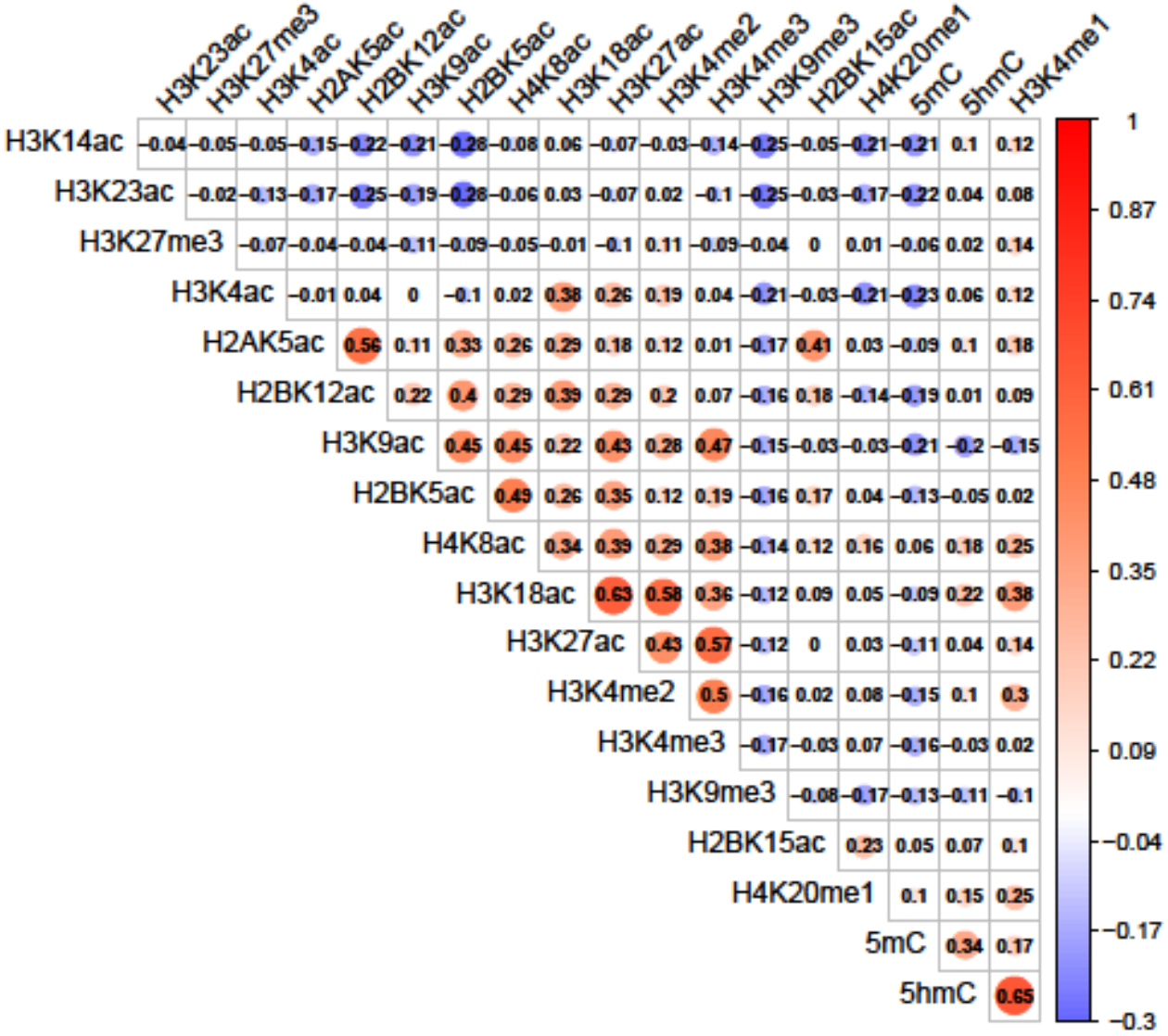


図 12：エピジェネティックファクター間のピアソン相関係数。（a）50K-bp 毎の Window に区切ったもの（b）Gene body 領域毎（c）Promoter 領域毎

また、他の生物種または検出法によるデータを用いて再度相関を調べ、この結果が人為的ミスや偶然によるものでないことを確認した。TAB-seq 法により得られた 5hmC データを用いて相関係数を調べたところ、免疫沈降法（hMeDIP）によるものと同様に高い値が得られた（図 13a）。次にマウスのデータを用いて調べたところ、ヒトデータより相関係数が若干低くなったが、依然として高い値が得られた（PCC : 5hmC vs H4K8ac=0.54、5hmC vs H3K4me1=0.65）（図 13b）。このことから今回の結果が実験方法に依存しないこと、他の哺乳類においても当てはまることが確認できた。

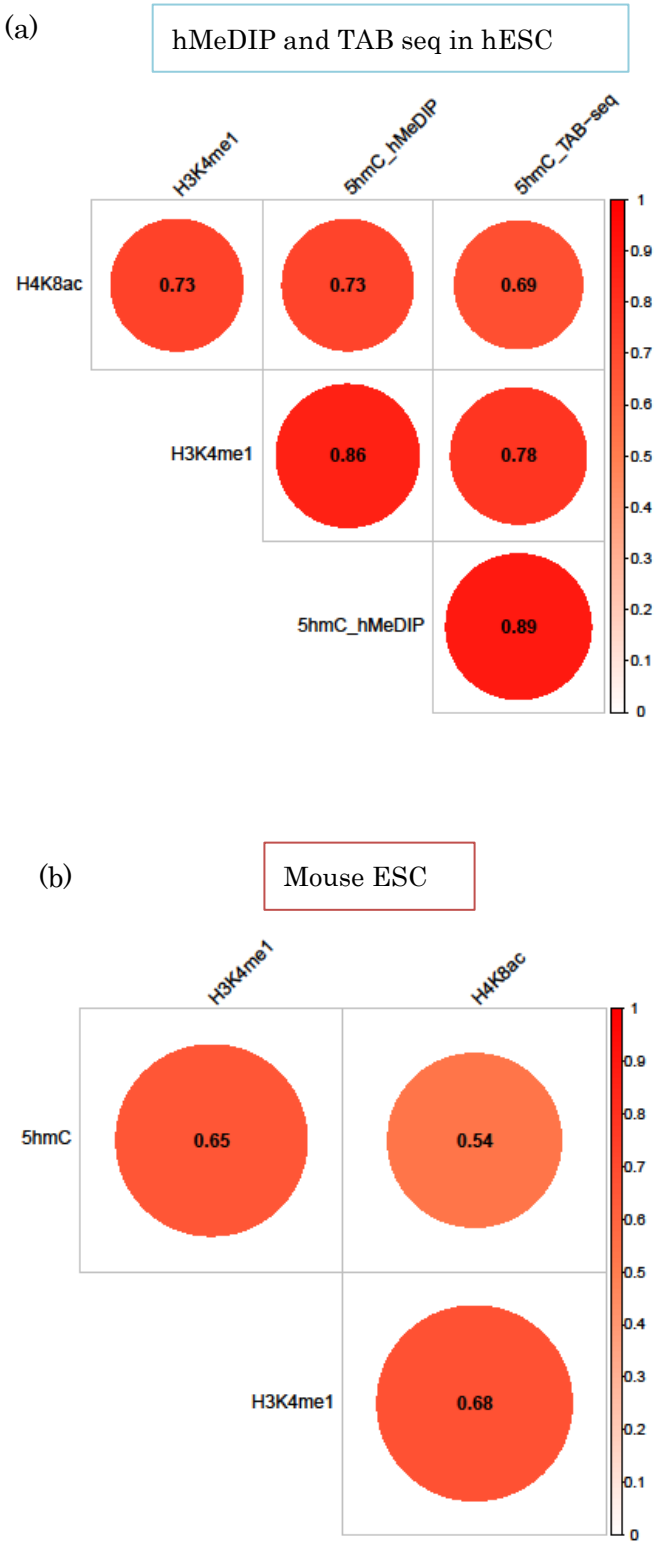


図 13：他の実験方法によるデータ及びマウスデータを用いた、エピジェネティックファクター間のピアソン相関係数。（a）TAB-seq によるデータと免疫沈降法（hMeDIP）によるデータ間の相関係数。(b) マウスデータを用いた相関係数。

4.2.2　多能性幹細胞特異的遺伝子における各ファクターの分布状況

個々の遺伝子の内、多能性幹細胞特異的遺伝子に注目して、その遺伝子領域における 5hmC、H3K4me1、及び H4K8ac の分布状況を調べた。代表的な多能性幹細胞マーカー遺伝子である POU5F1（Oct3/4）について調べたところ、特に TSS から+500～+2500bp の領域や TES の下流領域において、これら修飾の強い共起が表れた（図 14、緑色枠内）。他の多能性幹細胞特異的遺伝子については KLF4 には同様に強い共起があり、NANOG には中程度の共起が観察された。一方 SOX2 においては共起関係が見られなかった。

図 14：多能性幹細胞マーカー遺伝子 POU5F1（Oct3/4）における、5hmC、H3K4me1、及び H4K8ac の分布状況。

4.3　エピジェネティックファクター共起と遺伝子発現変化の関連性

4.3.1　有意発現変動遺伝子における各ファクター分布の特徴

DEG を基準として全遺伝子を 3 つのグループ（DEG_ES_up、DEG_ES_down、及び non_DEG）に分類し、遺伝子領域における 5hmC、H3K4me1、及び H4K8ac 各ファクターの分布状況を体細胞（fibroblast）と hESC 間で比較した。DEGs の内、体細胞と較べて hESC において顕著に発現量が増加している遺伝子群（DEG_ES_up、2888 genes）に注目すると（図 15a-c、左図）、hESC で顕著に発現量が減少している遺伝子群（DEG_ES_down、2141 genes）及びその他の遺伝子群（non_DEG、20660 genes）（図 15a-c、右図）と較べて、hESC の方が各修飾の分布が enrich になっていた。このことは、hESC における遺伝子発現の上昇とこれらファクター分布の増加に関連性があることを示している。

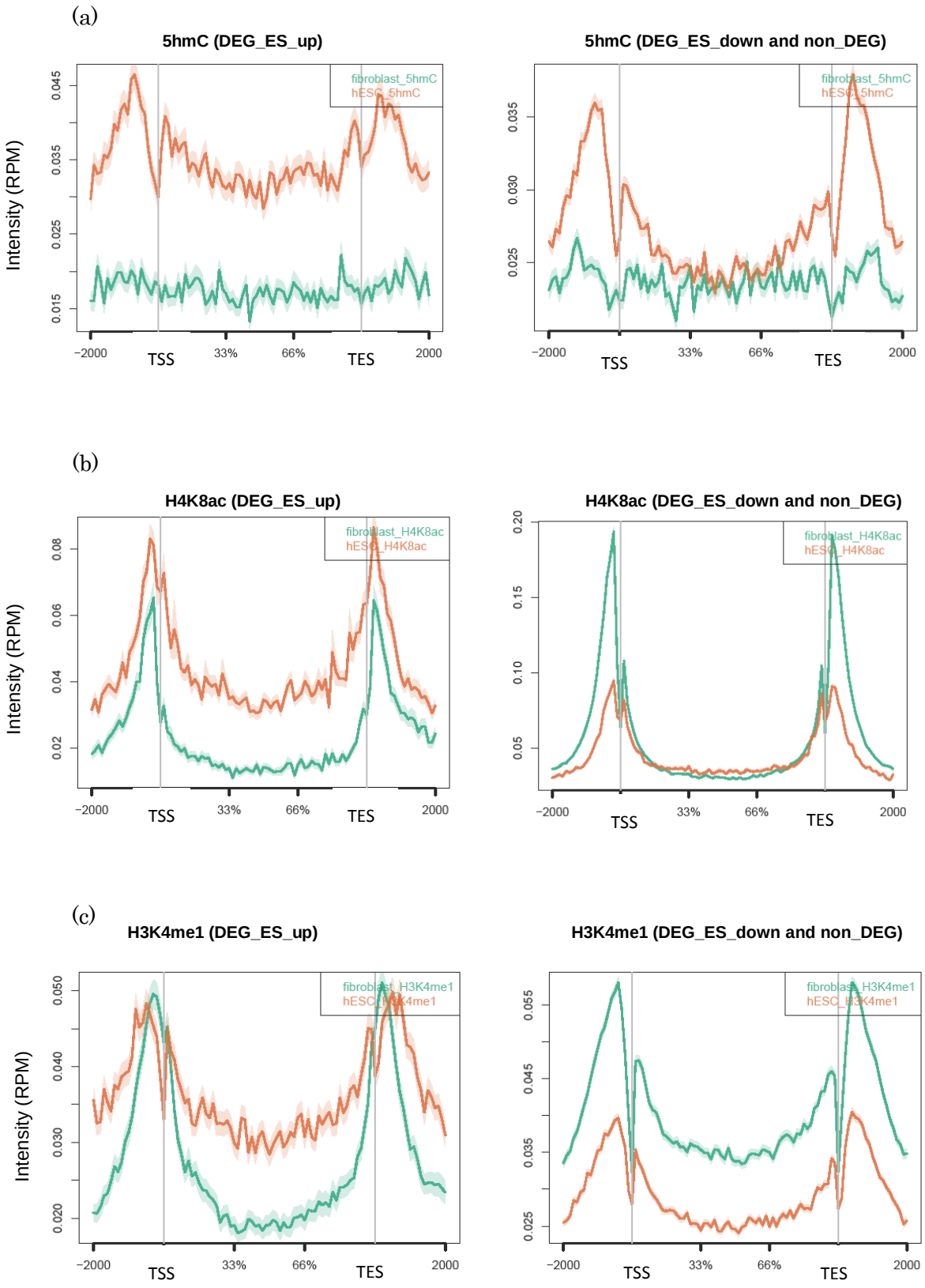


図 15：DEGs に基づいた、各エピジェネティックファクター分布の fibroblast（緑色）と hESC（橙色）における比較。（a）5hmC（b）H4K8ac（c）H3K4me1

4.3.2　有意発現変動遺伝子における各ファクター強度の変化

4.3.1 節で分類した 3 つの遺伝子グループ（DEG_ES_up、DEG_ES_down、及び non_DEG）について、体細胞と hESC における各ファクター強度を比較した（図 16）。その結果、各ファクターに関して DEG_ES_up 遺伝子群において、fibroblast（緑色）よりも hESC（赤色）の方が 5hmC、H3K4me1、及び H4K8ac の各強度が強くなった。逆に DEG_ES_down 遺伝子群においては、fibroblast（緑色）の方が hESC（赤色）よりも強度が強くなった。このことから、体細胞・hESC 間の各ファクター強度変化と遺伝子発現変化は、正の相関関係を示していると言える。

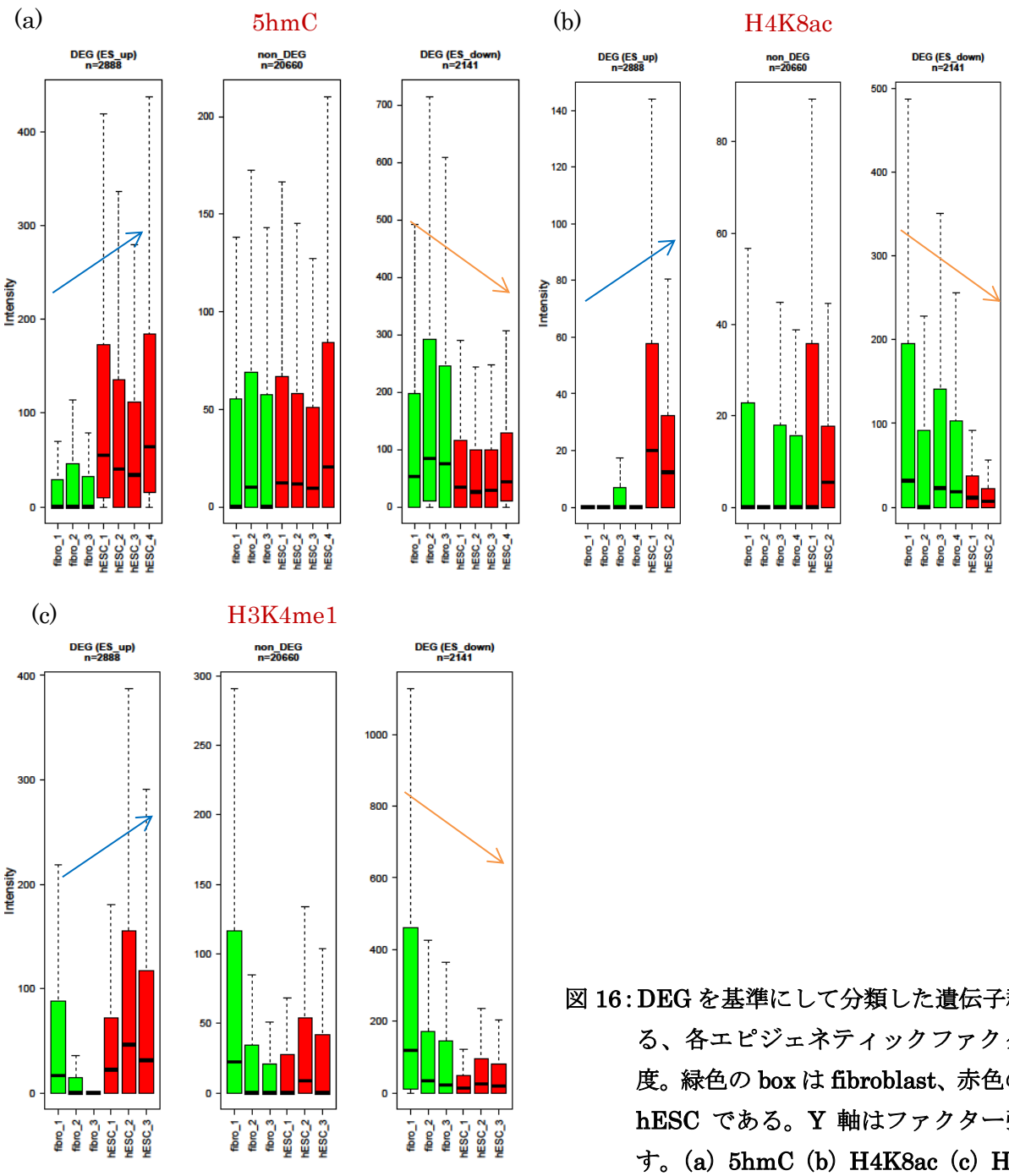


図 16：DEG を基準にして分類した遺伝子群における、各エピジェネティックファクターの強度。緑色の box は fibroblast、赤色の box は hESC である。Y 軸はファクター強度を表す。(a) 5hmC (b) H4K8ac (c) H3K4me1

4.3.3　エピジェネティックファクター共起による遺伝子発現への影響

ここまではこれら 3 個のエピジェネティックファクター（5hmC、H3K4me1、及び H4K8ac）の個々に注目し、体細胞と hESC 間における各ファクター分布・強度変化と遺伝子発現変化（DEG）との関連性を調べてきた。次にこれら 3 個のファクターに加えて、5hmC の前駆体である 5mC を組み合わせて一つのコードとして捉え、遺伝子発現との関連性を調べてみた。まず gene body 領域における hESC・体細胞間の各ファクター変化を組み合わせてパターン化する階層的クラスタリングを行い、全遺伝子を 19 個の遺伝子クラスターに分類した（図 17）。

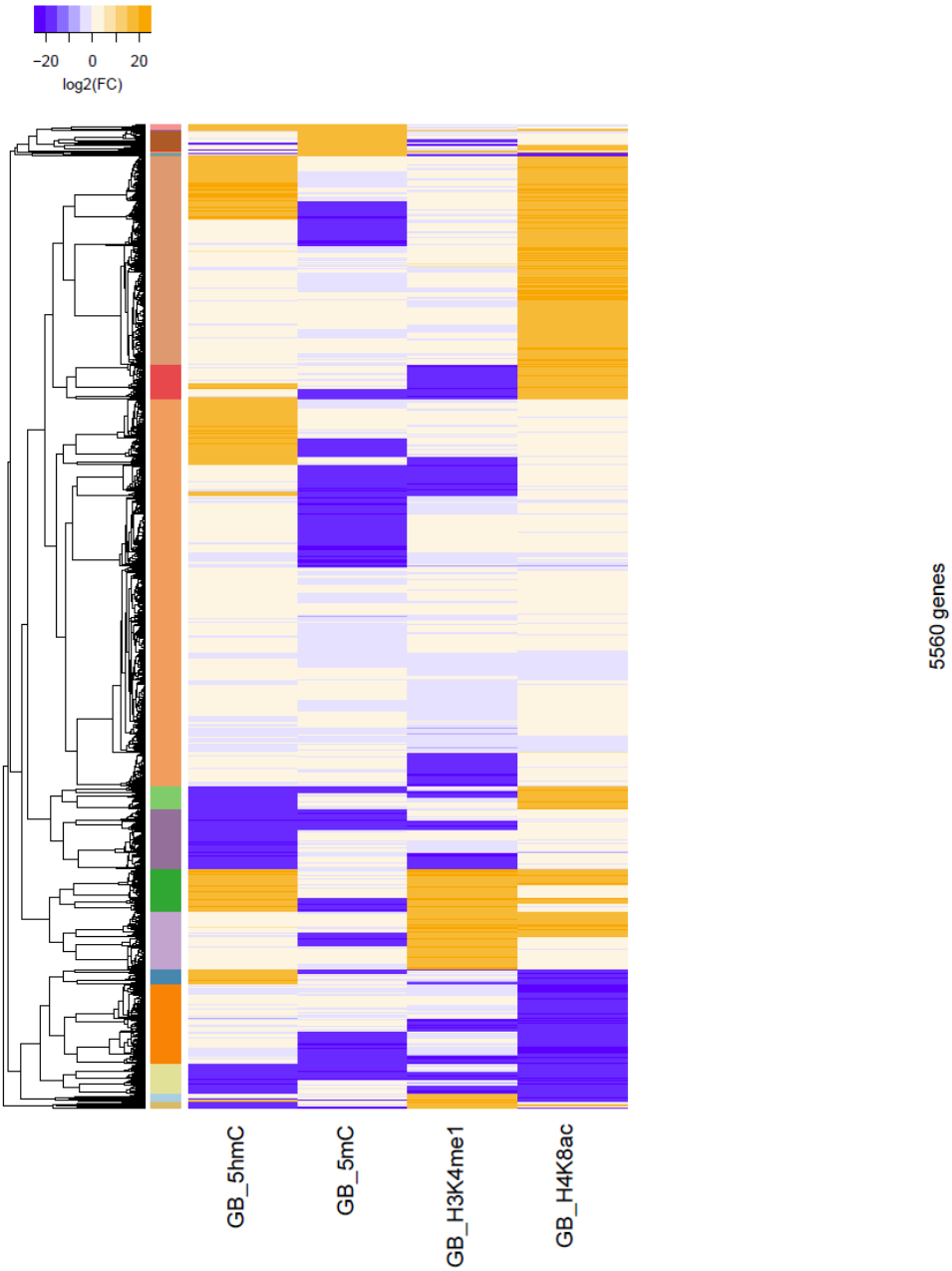


図 17：各エピジェネティックファクターの、gene body 領域における体細胞・hESC 間変動パターンに基づいた全遺伝子の階層的クラスタリング。GB は gene body を意味する。いずれのファクターにも変化が見られなかった遺伝子群を除外した結果、5560 個の遺伝子群に 19 種類の特徴的なエピジェネティックファクター変動パターンが見られた。

階層的クラスタリングにより得られた 19 個の遺伝子クラスターについて、それぞれに含まれる遺伝子群の発現量分布を調べた（図 18）。縦軸は体細胞・hESC 間の発現量比（Fold Change）を log で表したもので、ボックスが図の上側に行くほど、体細胞と比較して hESC の発現が高くなることを示している。この図から、特にクラスター6 と 18 が、他のクラスターと比べて特徴的な発現変動傾向があることが見て取れる。

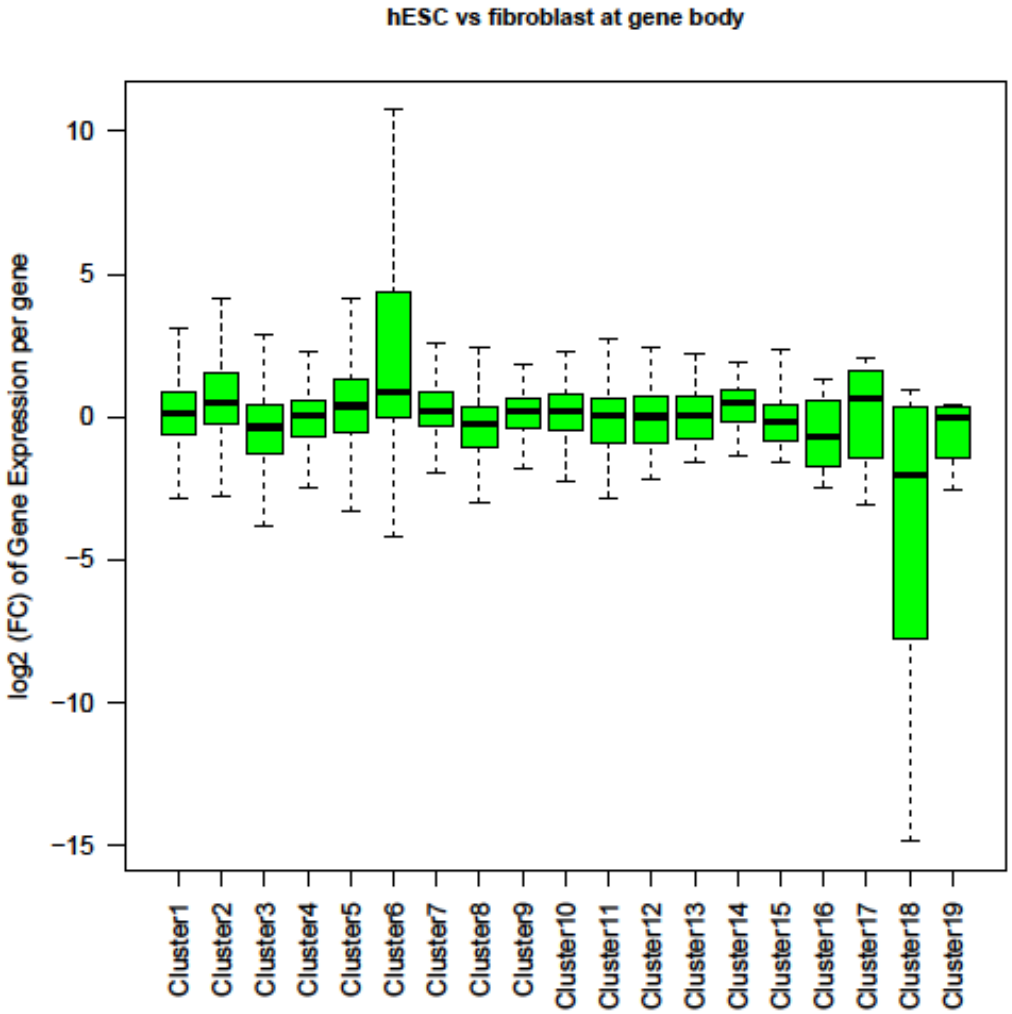


図 18 : Gene body 領域を対象とした、階層的クラスタリングにより得られた 19 個の遺伝子群の発現量分布。縦軸は体細胞・hESC 間の発現比（Fold Change）を log で表している。

また、図 19 及び図 S1 は 19 個のクラスターが、それぞれどのようなファクター変動パターンを持っている遺伝子群かを調べたものである。縦軸は各ファクター強度の体細胞・hESC 間変化比を log で表している。図の上側に行くほど、体細胞と比較して hESC のファクター強度が高くなることを示している。例えばクラスター2 に属する 1177 個の遺伝子群は、体細胞（fibroblast）と比較して hESC の方が、5hmC 及び H4K8ac の強度が高くなっている。また、H3K4me1 には違いが見られないが、5mC に関しては fibroblast の方が若干高い強度となっている。

特徴的な発現変動傾向を示したクラスター6 に含まれる 241 個の遺伝子群は、hESC の 5hmC、H3K4me1、及び H4K8ac が顕著に高い強度になっている。一方でクラスター2 はこれら 3 つのファクターの内、H3K4me1 に変化がなく、クラスター5 は 5hmC に変化が見られなかった。そして、これらのクラスターに含まれる遺伝子群は顕著な発現変動を示さなかった（図 18）。さらに、代表的な多能性幹細胞マーカー遺伝子である NANOG はクラスター6 に含まれていた。

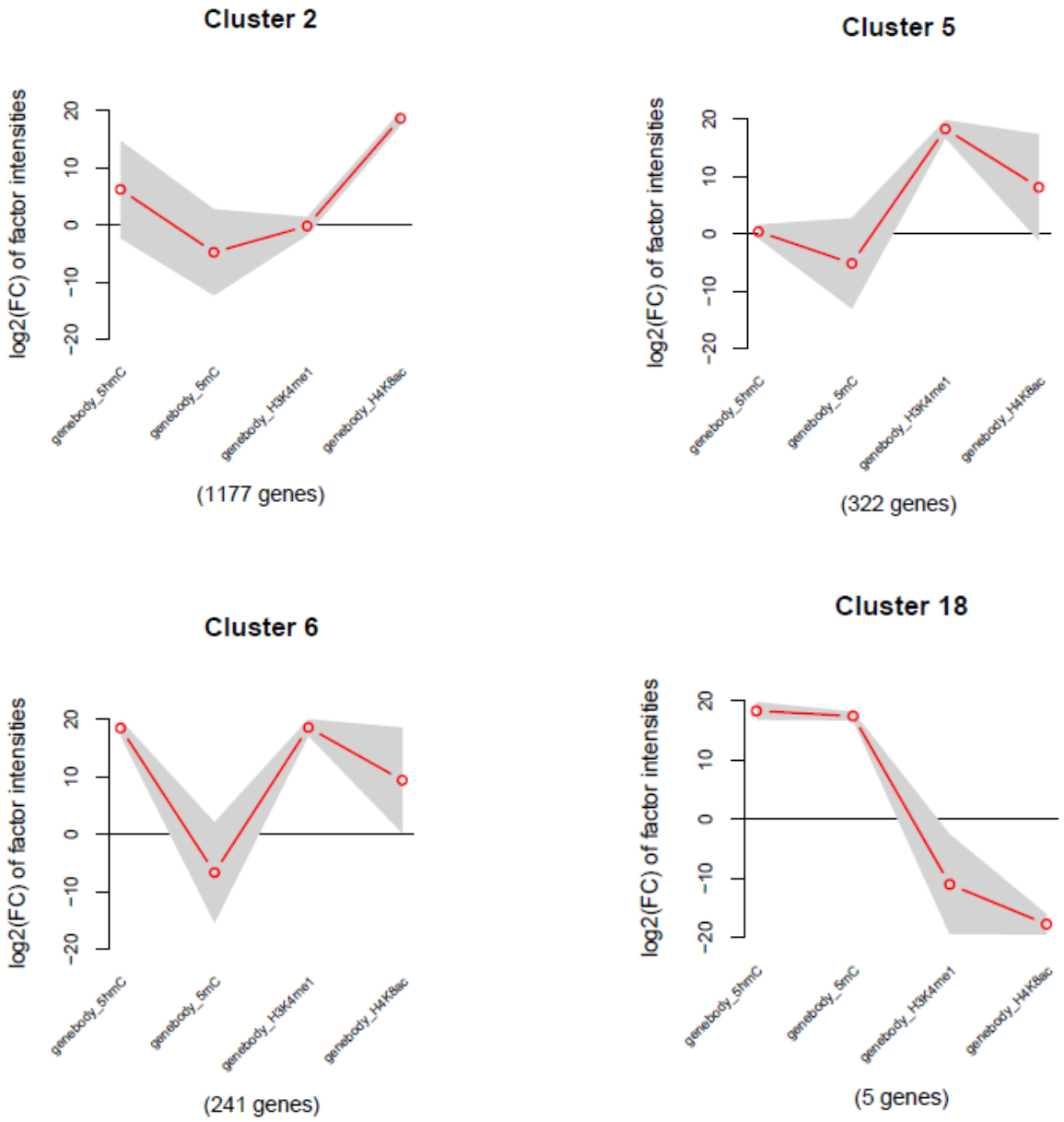


図 19: Gene body 領域を対象とした、19 個のクラスター遺伝子群が持つ 5hmC、5mC、H3K4me1、及び H4K8ac 変動パターン（抜粋）。縦軸は体細胞・hESC 間の各ファクター強度変化比を log で表している。

Gene body 領域における 5hmC、H3K4me1、及び H4K8ac 共起の傾向は、promoter 領域においても確認された（図 20-21）。クラスター13 及び 14 において特徴的な発現変動傾向が見られ、hESC における 5hmC、H3K4me1、及び H4K8ac の強度は高くなっていた。さらに、代表的な多能性幹細胞マーカー遺伝子である NANOG はこのクラスター13 に、POU5F1（Oct3/4）はクラスター14 に含まれていた。

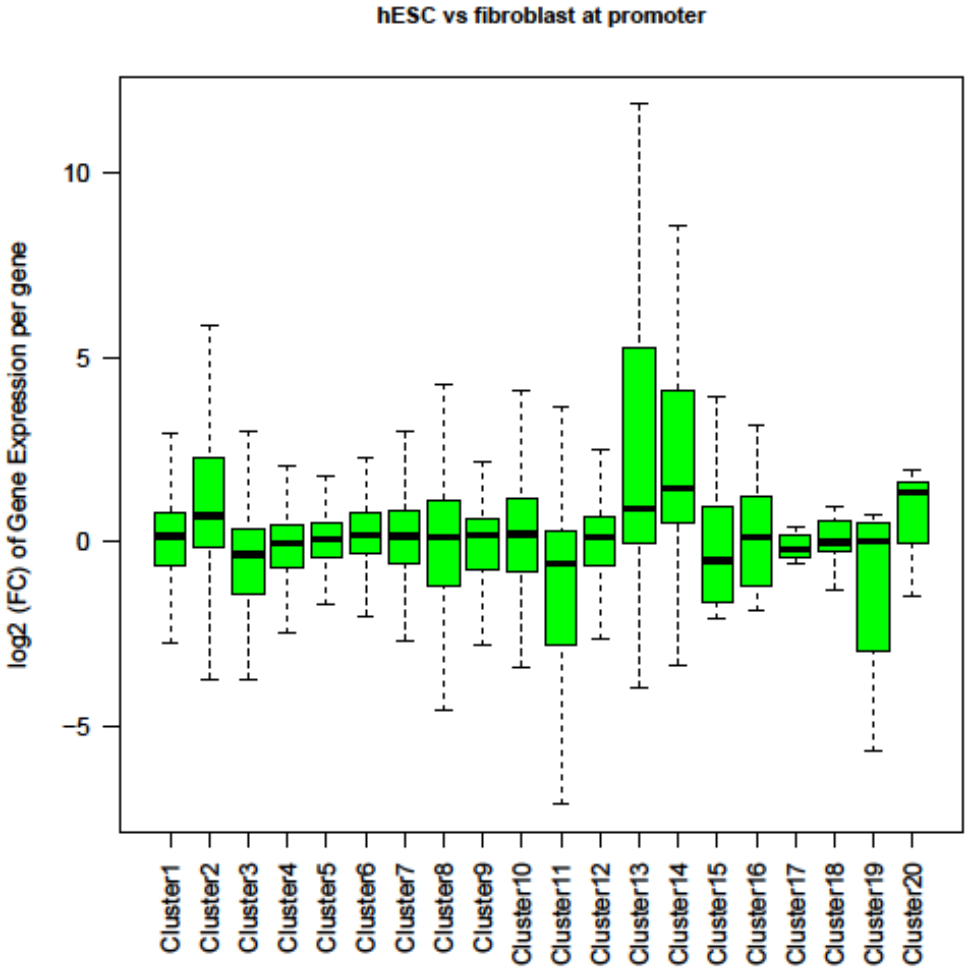


図 20：Promoter 領域を対象とした、階層的クラスタリングにより得られた 20 個の遺伝子群の発現量分布。縦軸は体細胞・hESC 間の発現比（Fold Change）を log で表している。

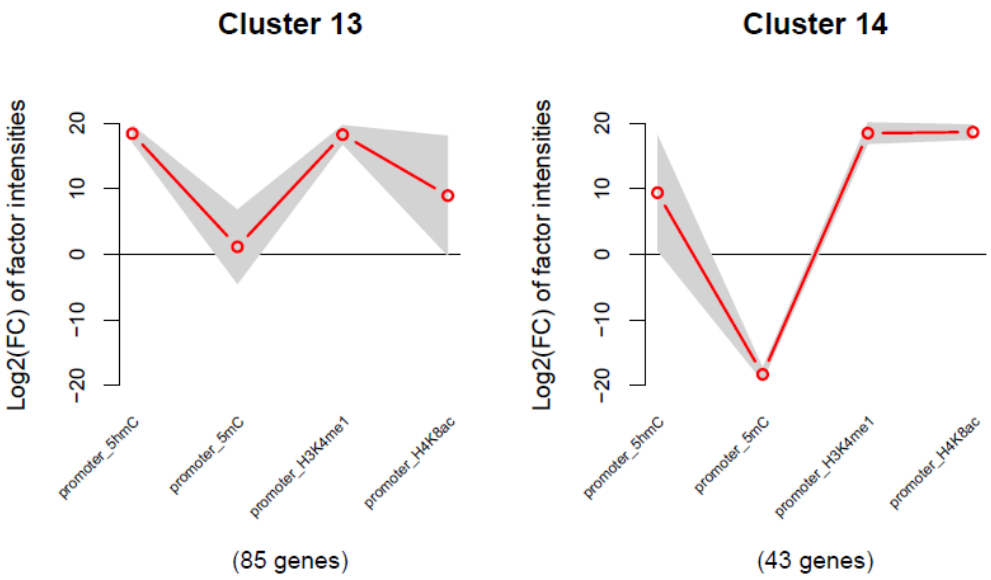


図 21：Promoter 領域を対象とした、20 個のクラスター遺伝子群が持つ 5hmC、5mC、H3K4me1、及び H4K8ac 変動パターン（抜粋）。縦軸は体細胞・hESC 間の各ファクター強度変化比を log で表している。

4.4　NMF アルゴリズムの活用による共起の持つ生物学的意義の探求

4.4.1　NMF アルゴリズムに基づく全遺伝子のグルーピング

共起が 5hmC・H3K4me1・H4K8ac 以外にも存在する可能性及びこの共起が持つ生物学的意義について調べるため、NMF アルゴリズムを適用して解析を行った。入力行列として全遺伝子（24087 genes）を縦軸、gene body と promoter 領域における DNA・ヒストン修飾に関する 19 種類のエピジェネティックファクター群を横軸とする行列を設定した。最初に分解のためのランクを設定するに際して、cophenetic 値が減少し始めるランクを採用した。今回の場合は、ランク 8 を最適なランクとして設定した（図 22、赤色点線部分）。

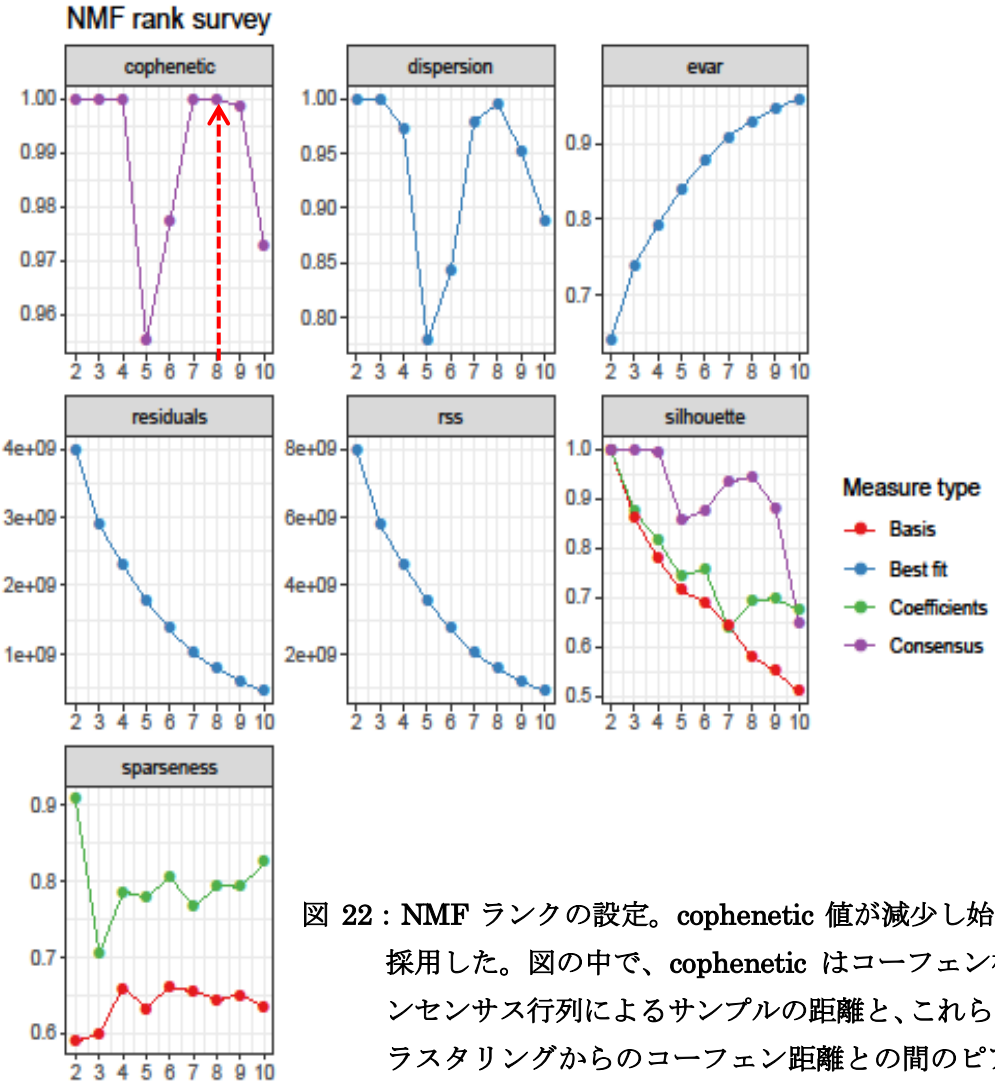


図 22：NMF ランクの設定。cophenetic 値が減少し始める値を最適ランクとして採用した。図の中で、cophenetic はコーフェン相関係数を表す。これはコンセンサス行列によるサンプルの距離と、これらの距離に基づいた階層的クラスタリングからのコーフェン距離との間のピアソン相関として定義される。これにより、それぞれのランクを適用した場合の分析結果の妥当性が判断できる。Dispersion はコンセンサス行列に基づいた分散係数を表し、得られたクラスターの再現性を測定したものである。Evar は各ランクを適用したモデルがデータセットの変動（分散）を占める割合を測定したもので、rss はモデルから得られる残差平方和を表す。

このランクを基に入力行列を分解すると、全遺伝子に関する基底行列 W とエピジェネティックファクター群に関する係数行列 H が得られ、それぞれ 8 個のグループに分類された（図 23）。

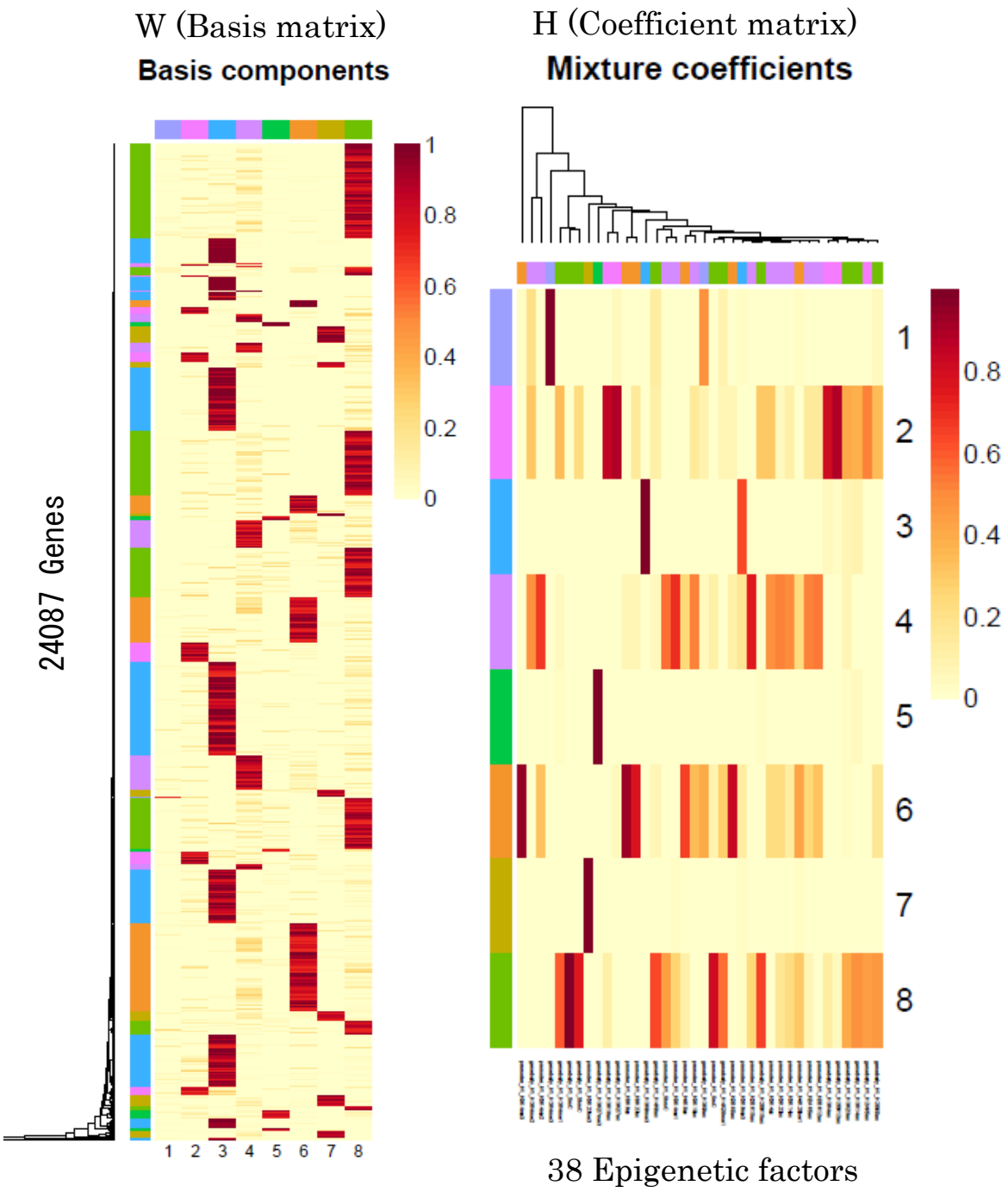


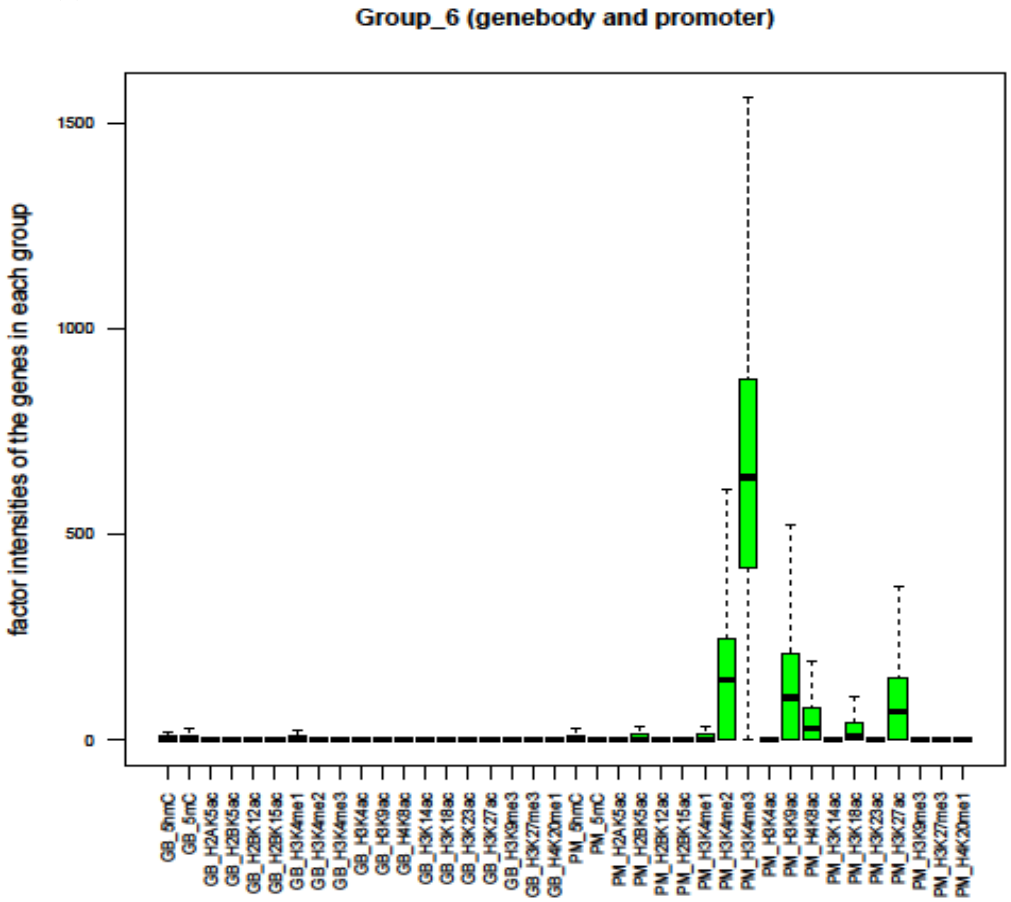
図 23：NMF アルゴリズムによる全遺伝子及びエピジェネティックファクター群のグルーピング。左図の縦軸は全遺伝子を表し、右図の横軸は gene body と promoter 領域を対象とする 38 種類（19 種類×2）のエピジェネティックファクター群を表す。それぞれが 8 個のグループ（ランク）に分類されている。

4.4.2　各遺伝子グループにおけるファクター共起状況

分類された各グループの内、グループ 6 遺伝子群の promoter 領域には、転写活性化に寄与する H3K4me2、H3K4me3、H3K9ac、及び H3K27ac が含まれており（図 24a）、他のグループと較べて高い発現が見られた（図 25a）。興味深い点として、グループ 7 の遺伝子群は promoter 領域において活性化マーク H3K4me3 と抑制マーク H3K27me3 の両方の enrichment が見られ（図 24b）、その発現は低く抑制されていた。

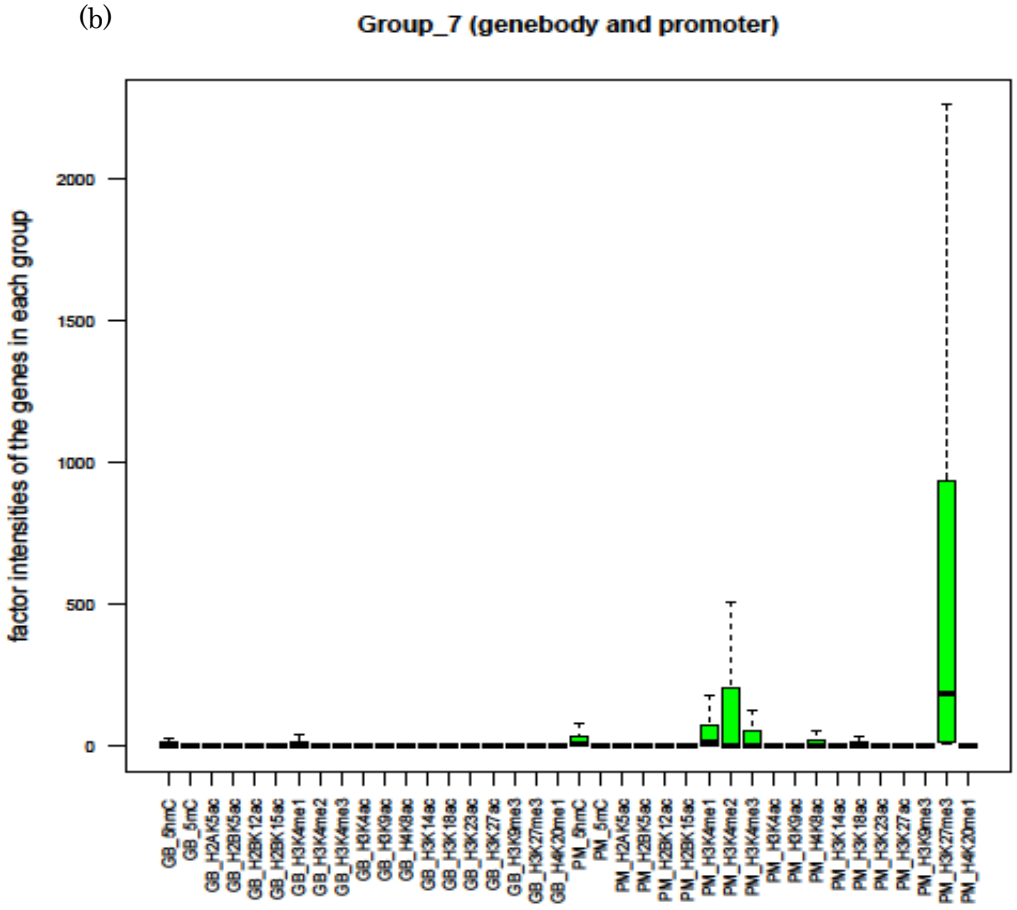
一方、グループ 8 に属する遺伝子群は、gene body と promoter 領域に 5hmC、5mC、H3K4me1、及び H4K8ac が enrich となっていた（図 24c）。これはこれまでの相関解析や階層的クラスタリングで共起傾向を示したエピジェネティックファクター群と一致しており、またその遺伝子群の発現も確認された（図 25a）。

(a)



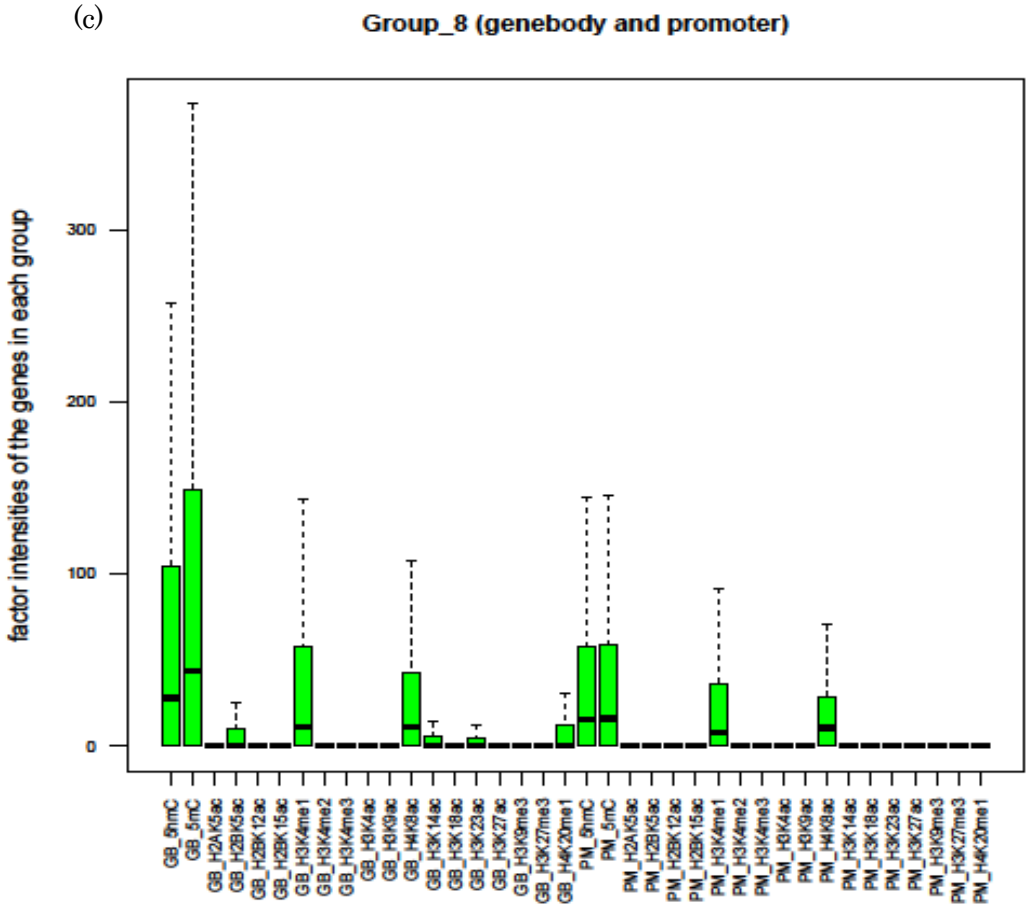
(962 genes)

(b)


Group_7 (genebody and promoter)
factor intensities of the genes in each group
2000
1500
1000
500
0
GB_5hmC
GB_5mC
GB_H2AK5ac
GB_H2BK5ac
GB_H2BK12ac
GB_H2BK15ac
GB_H3K4me1
GB_H3K4me2
GB_H3K4me3
GB_H3K4ac
GB_H3K9ac
GB_H4K8ac
GB_H3K14ac
GB_H3K18ac
GB_H3K23ac
GB_H3K27ac
GB_H3K9me3
GB_H3K27me3
GB_H4K20me1
PM_5hmC
PM_5mC
PM_H2AK5ac
PM_H2BK5ac
PM_H2BK12ac
PM_H2BK15ac
PM_H3K4me1
PM_H3K4me2
PM_H3K4me3
PM_H3K4ac
PM_H3K9ac
PM_H4K8ac
PM_H3K14ac
PM_H3K18ac
PM_H3K23ac
PM_H3K27ac
PM_H3K9me3
PM_H3K27me3
PM_H4K20me1

(367 genes)

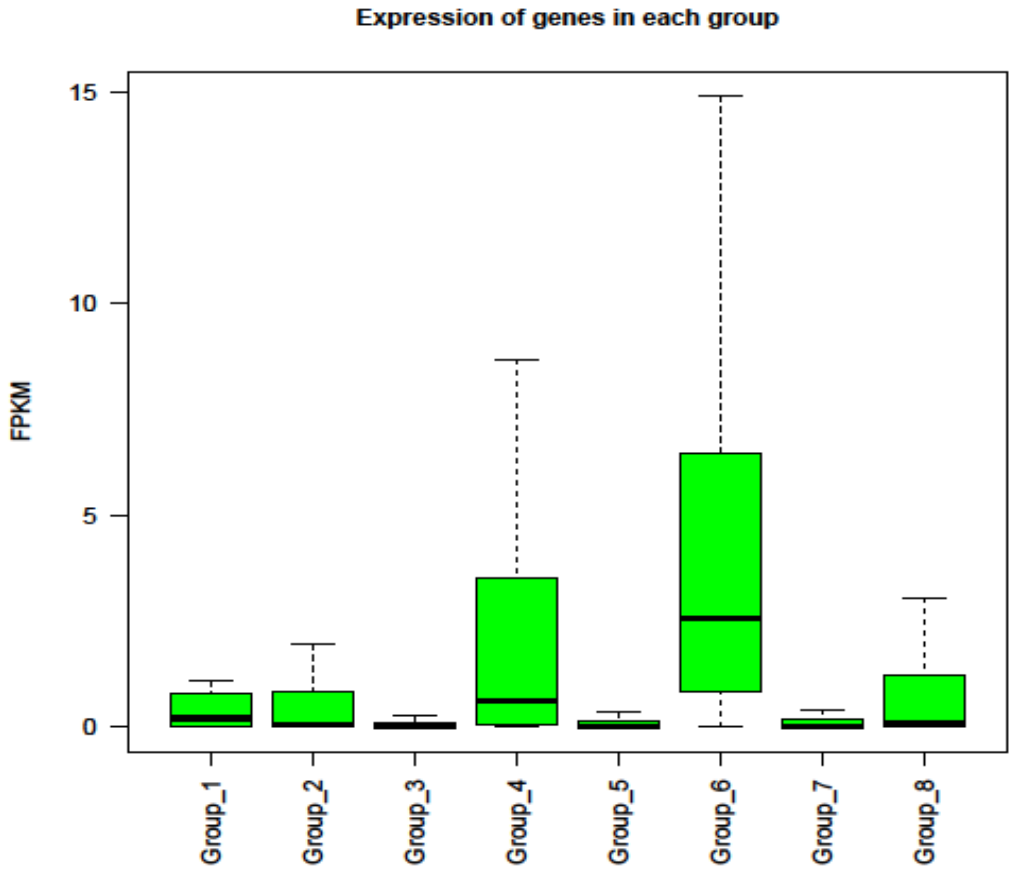
(c)



(1719 genes)

図 24：各グループに含まれる遺伝子群のエピジェネティックファクター強度分布。ボックスプロットの縦軸は各エピジェネティックファクターの強度を表しており、横軸の GB は gene body、PM は promoter を意味する。(a) グループ 6 遺伝子群 (b) グループ 7 遺伝子群 (c) グループ 8 遺伝子群

(a)



(b)

	Group_1	Group_2	Group_3	Group_4	Group_5	Group_6	Group_7	Group_8
GB_5hmC	–	@	@	-	@	-	@	@
GB_5mC	-	@	-	-	@	@	-	@
GB_H2AK5ac	-	@	-	-	-	-	-	-
GB_H2BK5ac	-	@	-	-	-	-	-	-
GB_H2BK12ac	-	@	-	-	@	-	-	-
GB_H2BK15ac	-	-	-	-	-	-	-	-
GB_H3K4me1	-	@	-	-	@	-	@	@
GB_H3K4me2	@	@	@	@	@	-	@	-
GB_H3K4me3	@	@	-	-	@	-	@	-
GB_H3K4ac	-	@	-	-	-	-	-	-
GB_H3K9ac	@	@	-	-	-	-	-	-
GB_H4K8ac	-	@	-	-	@	-	-	@
GB_H3K14ac	-	-	-	-	@	-	-	-

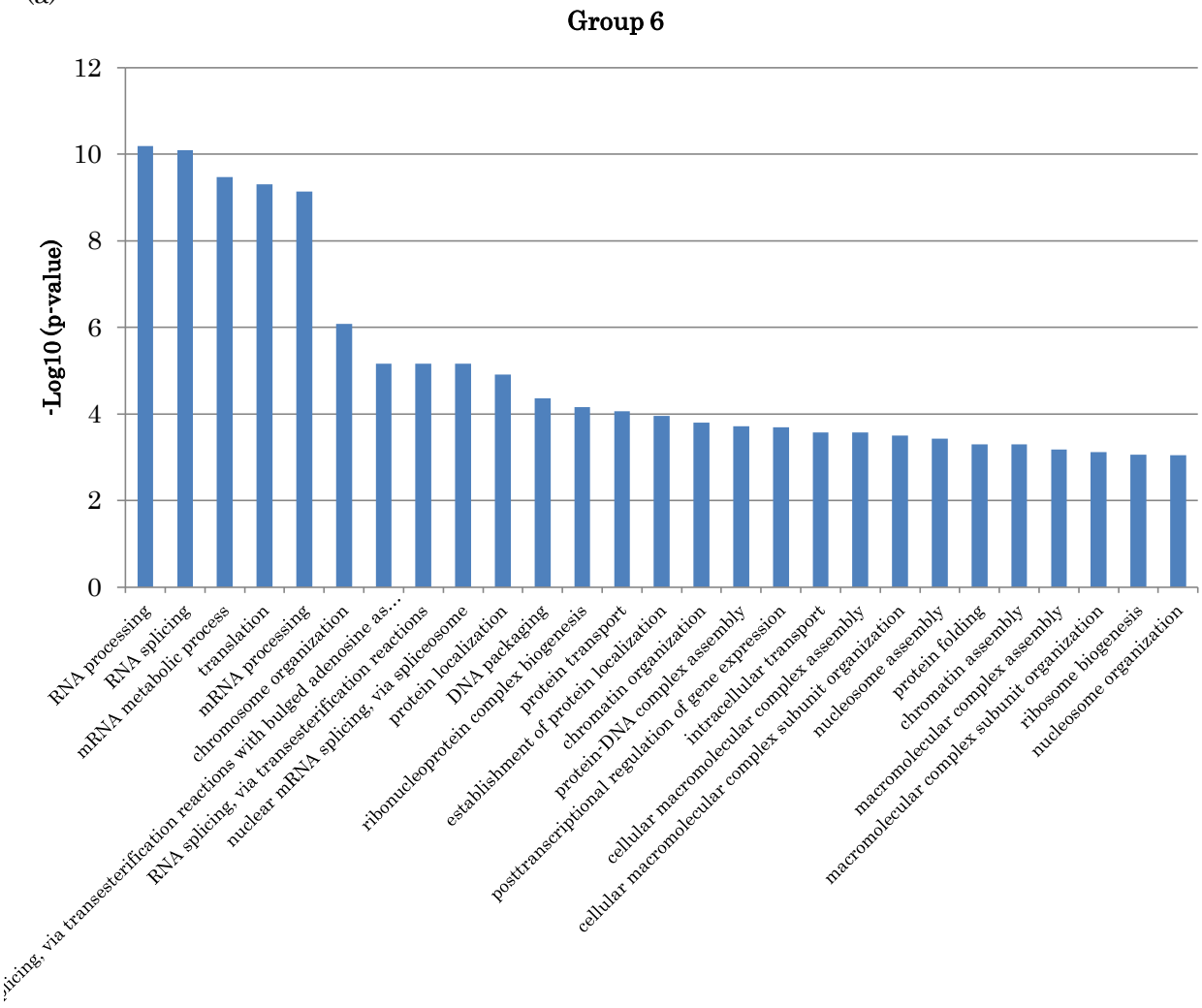
GB_H3K18ac	-	@	-	-	@	-	-	-
GB_H3K23ac	-	-	-	-	-	-	-	-
GB_H3K27ac	@	@	@	-	-	-	-	-
GB_H3K9me3	-	@	@	-	@	-	-	-
GB_H3K27me3	-	-	-	-	@	-	-	-
GB_H4K20me1	-	-	-	-	@	-	-	@
PM_5hmC	-	@	-	@	@	-	@	@
PM_5mC	-	-	-	-	-	-	-	@
PM_H2AK5ac	-	-	-	-	-	-	-	-
PM_H2BK5ac	-	-	-	-	-	@	-	-
PM_H2BK12ac	-	-	-	-	-	-	-	-
PM_H2BK15ac	-	-	-	-	-	-	-	-
PM_H3K4me1	-	@	-	@	@	@	@	@
PM_H3K4me2	-	@	@	@	@	@	@	-
PM_H3K4me3	-	@	@	@	@	@	@	-
PM_H3K4ac	-	-	-	-	-	-	-	-
PM_H3K9ac	-	-	-	@	-	@	-	-
PM_H4K8ac	-	-	-	@	-	@	@	@
PM_H3K14ac	-	-	-	-	-	-	-	-
PM_H3K18ac	-	@	-	@	@	@	@	-
PM_H3K23ac	-	-	-	-	-	-	-	-
PM_H3K27ac	-	@	-	@	-	@	-	-
PM_H3K9me3	-	-	@	-	-	-	-	-
PM_H3K27me3	-	-	-	-	@	-	@	-
PM_H4K20me1	-	-	-	-	-	-	-	-

図 25：(a) 各グループに属する遺伝子群の発現量分布 (b) 各グループに属する遺伝子群の持つ、エピジェネティックファクターの状況。各々のファクターにおける@マークは、強度値が 10 以上であることを表す。

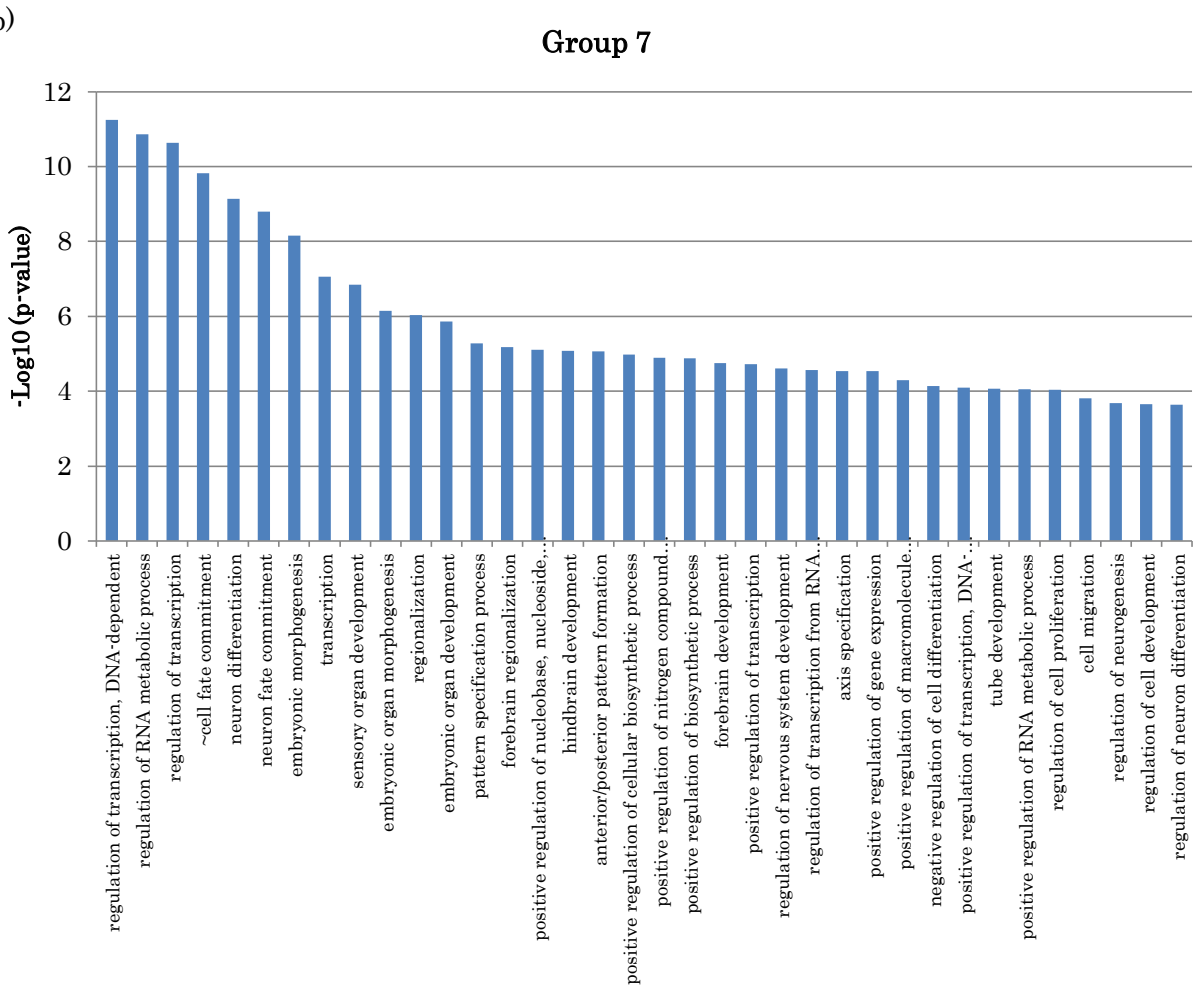
4.4.3　ファクター共起遺伝子群の生物学的特徴

NMF により分類された各グループに含まれる遺伝子群の生物学的特徴を調べるため、DAVID ソフトウェア（https://david.ncifcrf.gov/）を用いて gene ontology 解析を実施した。グループ 6 に含まれる遺伝子群は解析の結果、” RNA processing” や” chromosome organization” などの基本的細胞活動に関連していた（図 26a）。グループ 7 は”transcription”や”development”に関連した項目が該当していた（図 26b）。一方、グループ 8 の遺伝子群は“immune response”や“regulation of cytokine production”などの「免疫」への関連性が見られた（図 26c）。その内、“immune response”に関連する遺伝子群には自然免疫及び獲得免疫に関するものが両方とも含まれていた（表 S1）。例えば自然免疫の例として、TLR ファミリーに属し病原体の認識と自然免疫の活性化における基本的な役割を持っている TLR7、8、9 が含まれていた。獲得免疫の例としては、B 細胞活性化と免疫グロブリン合成の調節に重要な役割を果たしている CD27、蛋白質チロシンキナーゼファミリーに属する酵素をコードし T 細胞発生とリンパ球活性化に役割を果たす ZAP70、T 細胞抗原受容体（TCR）シグナル伝達経路の活性化後に ZAP70 タンパク質チロシンキナーゼによってリン酸化されるタンパク質をコードする LAT などが含まれていた。

(a)



(b)


Group 7
-Log10 (p-value)
12
10
8
6
4
2
0
regulation of transcription, DNA-dependent
regulation of RNA metabolic process
regulation of transcription
~cell fate commitment
neuron differentiation
neuron fate commitment
embryonic morphogenesis
transcription
sensory organ development
embryonic organ morphogenesis
regionalization
embryonic organ development
pattern specification process
forebrain regionalization
positive regulation of nucleobase, nucleoside,...
hindbrain development
anterior/posterior pattern formation
positive regulation of cellular biosynthetic process
positive regulation of nitrogen compound...
positive regulation of biosynthetic process
forebrain development
positive regulation of transcription
regulation of nervous system development
regulation of transcription from RNA...
axis specification
positive regulation of gene expression
positive regulation of macromolecule...
negative regulation of cell differentiation
positive regulation of transcription, DNA-...
tube development
positive regulation of RNA metabolic process
regulation of cell proliferation
cell migration
regulation of neurogenesis
regulation of cell development
regulation of neuron differentiation

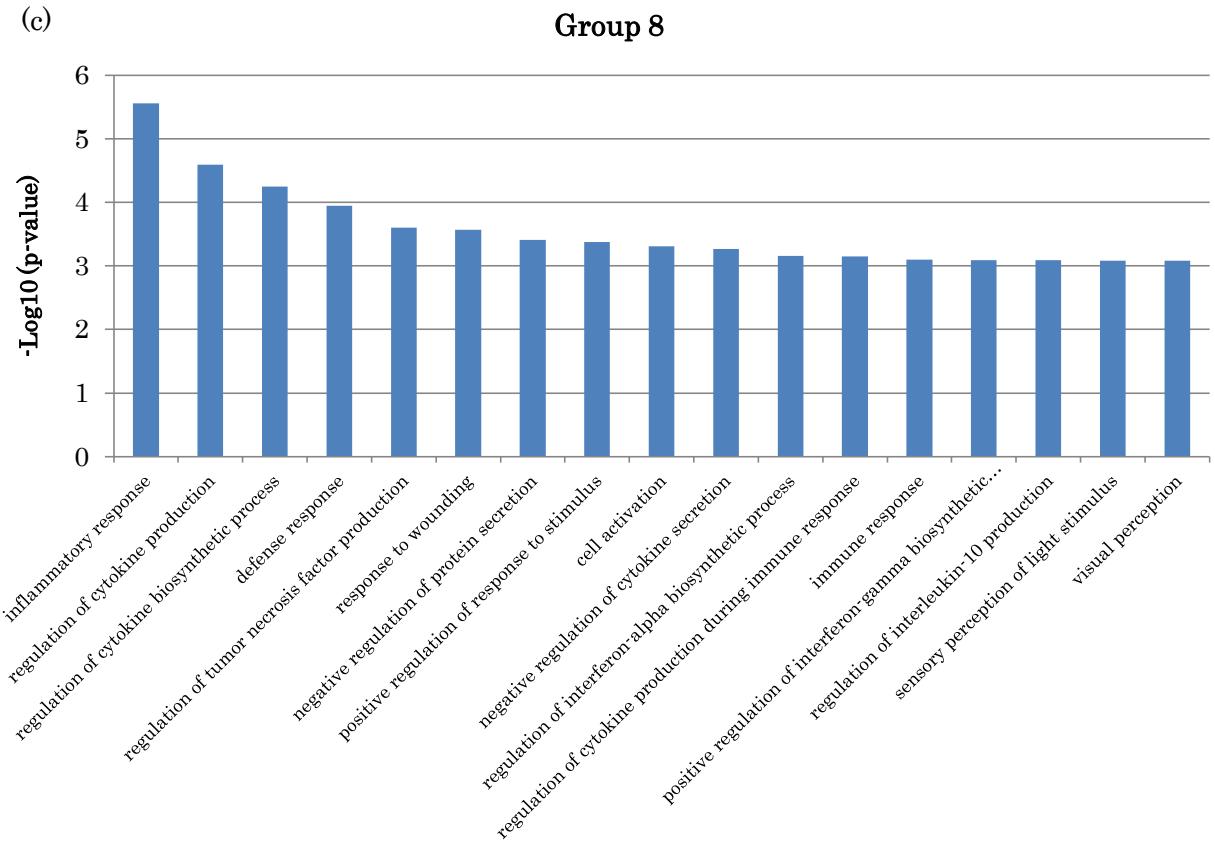
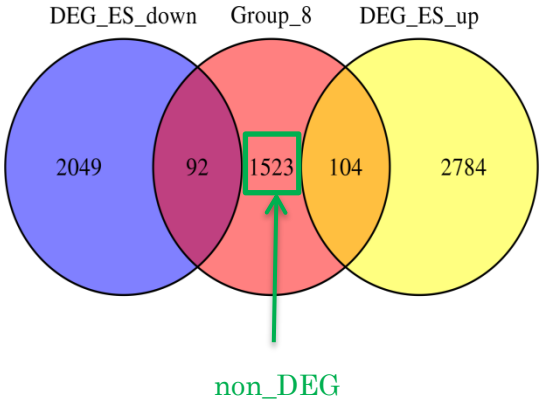


図 26：各遺伝子グループに関する Gene ontology 解析結果。Y 軸は解析により得られた p-value 値を対数で表している。（a）グループ 6（b）グループ 7（c）グループ 8

さらにグループ 8 遺伝子群を、DEG を基準として 3 つのグループ（DEG_ES_up、DEG_ES_down、及び non_DEG）に分類し、それぞれに gene ontology 解析を行った。その結果、DEG_ES_up 及び DEG_ES_down の遺伝子群には特徴が見られなかったが、non_DEG 遺伝子群（1523 genes）が主に「免疫」との関連性を示した（図 27a-b）。

(a)



(b)

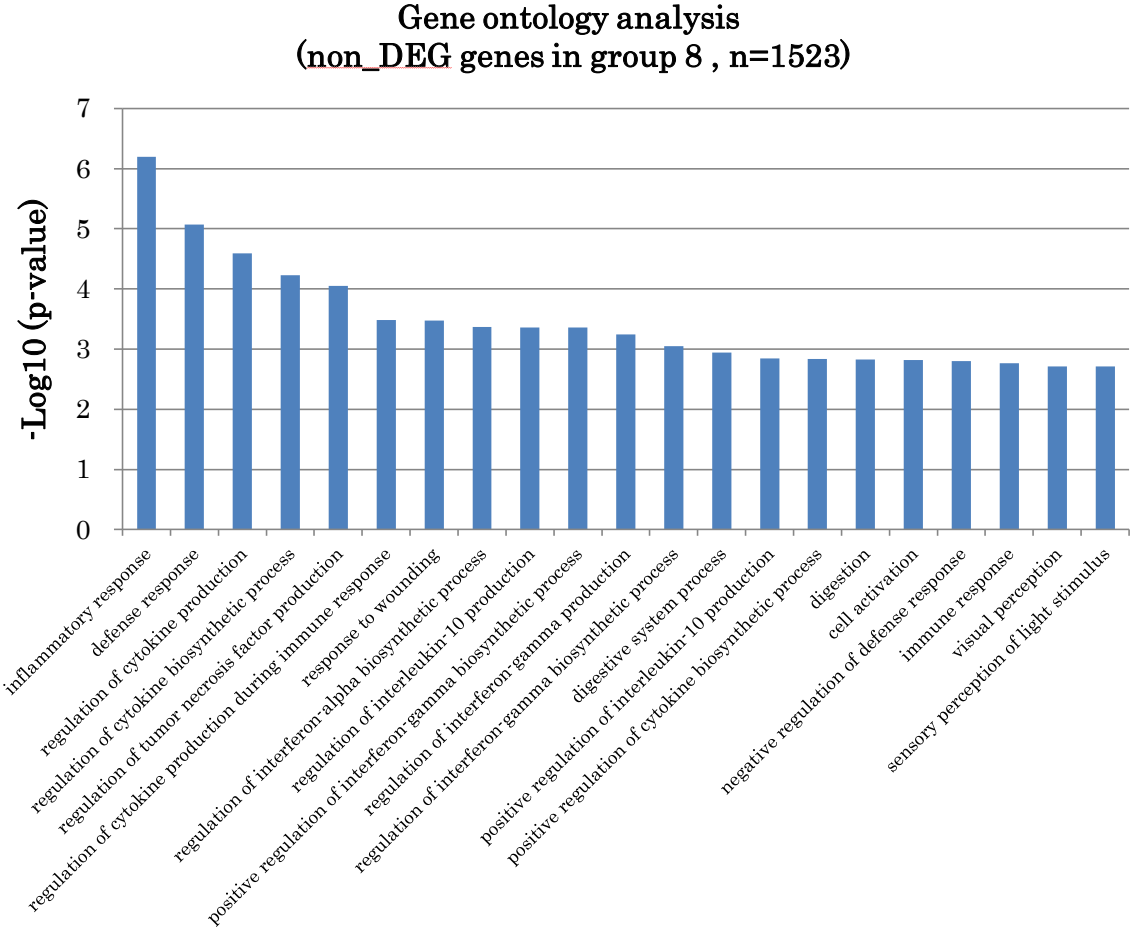


図 27：グループ 8 に含まれる遺伝子群のうち、体細胞・hESC 間で顕著な発現変化を示さなかった遺伝子群（non-DEG）に対する gene ontology 解析結果。（a）DEG_ES_up、DEG_ES_down、及び group_8 遺伝子間のベン図。（b）gene ontology 解析結果。Y 軸は解析により得られた p-value 値を対数で表したものである。”immune response”や”regulation of cytokine production”などの、免疫に関連した特徴を示している。

4.5　共起ファクター間の相互作用機序

3.5 節において作成した 9 個の遺伝子グループについて、それぞれのグループの体細胞・hESC 間 5hmC 変化を調べた（図 28）。この図とそれぞれのグループに含まれる遺伝子数（3.5 節）から、いくつかの傾向を見て取ることができる。一つ目は、②と③の群（H3K4me1_up and H4K8ac_down 、H3K4me1_down and H4K8ac_up）に含まれる遺伝子は 1 個のみで、ほとんど存在しないこと。二つ目は、⑤と⑦の群（H3K4me1_up and H4K8ac_NSC、H3K4me1_NSC and H4K8ac_up）に含まれる遺伝子群の大多数は 5hmC が増加しているが、増加していない遺伝子（図 28、赤色中括弧部分）も存在すること。三つめは、①の群に含まれる遺伝子群（H3K4me1_up and H4K8ac_up）は全て、5hmC が増加していることである。

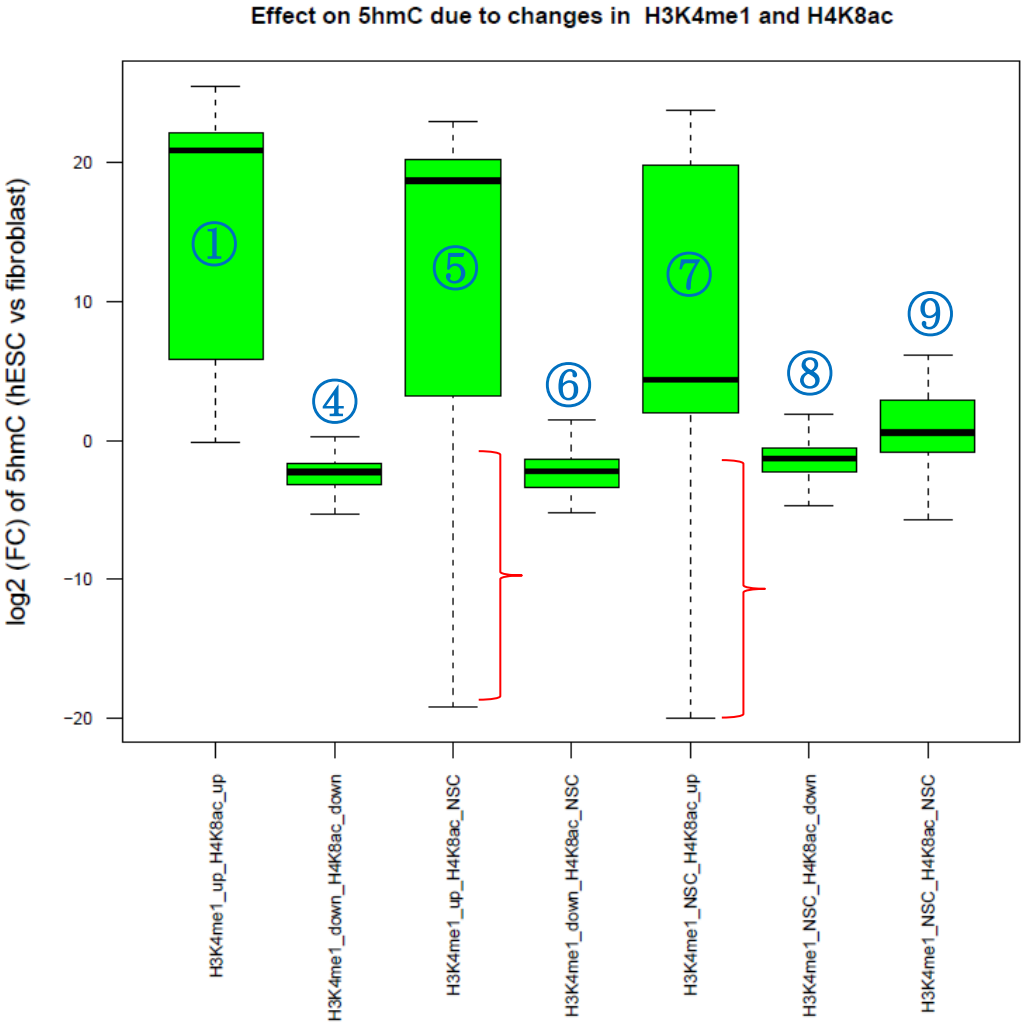


図 28：H3K4me1 と H4K8ac 強度の、体細胞・hESC 間有意変化に基づく 5hmC 強度変化。縦軸は体細胞・hESC 間の 5hmC 強度変化比を log であらわしたものであり、正の値（図の上方向）は体細胞と比較して hESC の方に 5hmC が多く存在することを表す。

第5章 考察

DNA 修飾とヒストン修飾間の相互作用により遺伝子発現に影響を与えることが、先行研究においていくつも報告されている。ESC においてバイバレント領域を有するプロモーターによる、遺伝子発現制御がその一例である。今回 ESC に代表される多能性幹細胞の持つ生物学的特徴に関わる、DNA・ヒストン修飾間の相互作用が他にも存在すると考え、新規の組み合わせ（エピジェネティックコード）の発見を目指した。コード形成の候補となるエピジェネティックファクターを調べていく上で、ESC に特異的に豊富に存在し、多能性幹細胞の性質に深く関わっていると考えられている 5hmC を中心として、これと高い共起を示すヒストン修飾の探索を行った。

始めに 5hmC の ESC における分布の特異性について調べた。クラスター解析の結果から ESC や iPSC における 5hmC 分布は、体細胞と較べて明らかに異なることが確認された（図 7-9）。ゲノムアノテーションに基づく分布の違いや（図 10）、遺伝子近辺における分布状況に関しても同様の傾向が確認された（図 11）。これらの結果は先行研究で述べられている、ESC において 5hmC が特異的分布を示す事実を裏付けるものである[2, 3]。

この 5hmC と共起するヒストン修飾を探すために、遺伝子発現に影響を及ぼす 17 個のヒストン修飾に 5hmC の前駆体である 5mC を加えて、それぞれのファクター強度に基づくお互いの相関を調べた。その結果、数種類のファクターに関して強い相関が見られた。とりわけ 5hmC、5mC、H3K4me1、及び H4K8ac からなる組み合わせは高い相関を示した（図 12）。5hmC と H3K4me1 が ESC において共起する傾向が強いことは既に報告されている[46] 。しかしながら、5hmC と H4K8ac の共起は今回新たに発見したものである。H4K8 はアセチル化だけを受け、転写領域において観察されることから、転写活性化に寄与するものと考えられている。また、H4K8ac は initiation よりも elongation に関わっていることが示唆されている[47]。全遺伝子を対象として見たとき、promoter 領域（図 12c）よりも gene body 領域（図 12b）の方がより高い相関が見られた。さらに図 11 において 5hmC はプロモーター領域に多くなっていた。これらのことから全遺伝子で見たときに、H3K4me1 や H4K8ac に対する 5hmC のプロモーター領域における強度が、gene body 領域の場合と較べて差があることを示していると考えられる。

さらに今回発見した 5hmC、H3K4m1、及び H4K8ac の共起は、50k-bp window 単位で見たときよりも、遺伝子レベルで見たときにより高いものとなった。このことは上記の因子群が互いに遺伝子レベルで共起しており、遺伝子発現に影響を与える可能性を示している。多能性幹細胞マーカー遺伝子である POU5F1（Oct3/4）においても共起が確認されたことは、この共起が多能性幹細胞の性質に影響を与えている可能性を示唆するものである（図 14）。これらファクター（5hmC, H3K4me1, 及び H4K8ac）個々が遺伝子発現に与える影響を調べた結果、各ファクター強度と遺伝子発現量には正の相関関係が見られた（図 15-16）。この内、図 15a では DEG_ES_up 遺伝子群の gene body 領域において 5hmC の enrichment が見られた。このことは gene body 領域の DNA メチル化が発現上昇につながる事実[17]に反しているように見えるが、5hmC は 5mC と較べて量的にかなり少ないこと[24]、さらには DNA メチル化が刻々と動的に変化する性質のものであることから、5hmC の enrichment が必ずしも 5mC の明確な減少として現れるとは限らないためであるのかもしれない。また図 15c において、DEG_ES__up 遺伝子群の TSS 及び TES 領域近

辺に関して、H3K4me1 の enrichment が ESC と fibroblast で同程度となっているが、その他の遺伝子群（DEG_ES_down 及び non_DEG）と比較した場合にはその違いが明確に現れている。このことから H3K4me1 においても、gene body と TSS・TES 近辺の領域における enrichment と遺伝子発現との関連性が見られたと言える。次に ESC 特異的遺伝子発現における、これらのファクター（5hmC、H3K4me1、及び H4K8ac）共起の必要性について調べた。各ファクターが個々に遺伝子発現に影響を及ぼすことは確認することができたが（図 15）、共起そのものが明確な ESC 特異的発現変化に必要であるかどうかは、ここまでの解析では不明である。言い換えれば遺伝子発現変化に影響を与えるのは、これら 3 個のファクターの内の一つだけかもしれない。そこで体細胞・hESC 間の各ファクター変化を組み合わせることによる、遺伝子発現変化への影響について調べた（図 18-21）。

クラスター6 において、hESC と体細胞間の 5hmC、H3K4me1、及び H4K8ac の強度比は顕著に高くなった（図 19）。そしてこの組み合わせを持つ遺伝子群は他の遺伝子群と比較して、顕著な遺伝子発現変化を示した（図 18）。この傾向は gene body 領域だけでなく promoter 領域においても同様に確認された（図 21、クラスター13 及び 14）。重要な点はクラスター2 において 5hmC と H3K4me1 の強度比は大きいが H3K4me1 に関しては変化がなかったこと、さらにはクラスター5 において H3K4me1 及び H4K8ac の強度比は大きいが 5hmC に関しては変化がなかったことであり、さらにこの両者共に顕著な遺伝子発現変化が見られなかったことである。このことから、これらの修飾が単独ではなくお互いが共起することが重要であり、それにより ESC 特異的な遺伝子発現変化がもたらされていると考えられる。クラスター18 において、H3K4me1 と H4K8ac 強度低下に伴い遺伝子発現低下が見られたことは（図 18-19）、H3K4me1 と H4K8ac の重要性を示しているように思われるが、対象遺伝子の数が少ないため（5 genes）結論付けるには不十分と考えられる。さらに相関解析のときと同様に、この階層的クラスタリングにおいても多能性幹細胞マーカー遺伝子の NANOG や POU5F1（Oct3/4）が、5hmC、H3K4me1、及び H4K8ac の共起を示すクラスター（図 19 のクラスター6、図 21 のクラスター13 及び 14）に含まれていたことは、この共起が ESC 特異的遺伝子の発現及び性質に影響を与えている可能性を裏付けるものである。

次に DNA 及び各種ヒストン修飾に関する強度を統合的に用いて NMF アルゴリズムを活用した解析を行い、潜在的に存在するエピジェネティックファクター共起を探索し、その生物学的特徴を調べた。グループ 6 の遺伝子群にはプロモーター領域における H3K4me2、H3K4me3、H3K9ac、及び H3K27ac といった転写活性化に関するヒストン修飾の enrichment が見られ、発現上昇が確認された（図 24a、図 25a）。また、その gene ontology（GO）解析の結果から、これらの遺伝子群には基本的細胞活動との関連性が見られた（RNA processing、chromosome organization 等）（図 26a）。さらにグループ 5 の遺伝子群は gene body 領域において H3K27me3 の enrichment が見られたが、H3K4me3 には enrichment が見られず、その発現は抑制傾向にあり、GO 解析の結果からこれらの遺伝子群は development process に関係していると考えられる（図 S4-5）。興味深いことに、グループ 7 に含まれる遺伝子群は、プロモーター領域に抑制マーク H3K27me3 と活性化マーク H3K4me3 の両方を含むという bivalent な特徴を有しており、その発現は抑制されていた（図 24b、図 25a）。さらには gene ontology 解析の結果から、これらの遺伝子群は、development に関連していた（図 26b）。この一連の結果は ESC などの多能性幹細

胞の性質と一致しており、NMF 解析により得られた 8 つの遺伝子グループはそれぞれ ESC 特有の性質を表していると考えられる。また同時に、この解析の妥当性をも示すものである。

さらに加えてグループ 8 に含まれる遺伝子群は、gene body と promoter 領域において 5hmC、5mC、H3K4me1、及び H4K8ac の高い共起を伴っていた（図 24c）。この組み合わせは、先の相関解析や階層的クラスタリングで新たに発見した共起と一致している。ここで図 24c では 5mC 強度も高くなっており、一方階層的クラスタリングにおける図 19 のクラスター6 では 5mC が低くなっている。これは階層的クラスタリングが体細胞と比較した ESC における強度比を示しているのに対し、NMF 解析では ESC のみを対象として強度を求めているためであるのかもしれない。

これらのことから、グループ 8 遺伝子群が bivalent genes のように何らかの ESC 特有の生物学的過程に影響を与えている可能性が考えられる。グループ 8 に含まれる遺伝子群の内、non_DEG の遺伝子群のみに特徴が見られ、主に免疫との関連性を示した（図 27）。階層的クラスタリングの結果からも、NANOG や Oct3/4 といった多能性幹細胞特有の発現を示す遺伝子群において、5hmC、H3K4me1、及び H4K8ac の共起が存在することは間違いない。この研究の始めにおいて、エピジェネティックファクターの共起が ESC 特異的発現に影響を及ぼすことをエピジェネティックコードの条件の一つとして設定した。しかしながら bivalent genes のように、エピジェネティックファクターの共起が必ずしも有意な発現変化をもたらすとは限らない。つまり、このエピジェネティックコード候補の持つ生物学的意義は、ESC において有意発現変化を示す多能性幹細胞マーカー遺伝子群の他に、有意な発現変化の見られない non_DEG 遺伝子群にも存在するのかもしれない。ホルモンによるシグナル伝達のように、少量の発現であっても生理活性に顕著な影響を及ぼす可能性は十分に考えられる。

先行研究では ESC において、体細胞と較べて弱い免疫拒絶が存在することが報告されている[48]。これは同じ多能性幹細胞である iPSC においても同様である[49, 50]。今回発見したエピジェネティックファクター群の共起は、この多能性幹細胞の持つ免疫性と関連している可能性が考えられる。例えば表 S1 に含まれる遺伝子群の内、LIR ファミリーに属する LILRB2 は、免疫提示細胞の MHC クラス I 分子に結合し、免疫応答の刺激を妨げる負のシグナルを伝達する。つまり、ESC における免疫性はこの種のシグナルにより阻害されるのかもしれない。そしてこの LILRB2 の発現は、今回発見した新規エピジェネティックコード候補により制御されている可能性が考えられる。

最後に、共起ファクター間の作用機序（各ファクターの生成順序）について調べた。4.5 節において見つかった 3 つの傾向から、次の可能性が考えられる。一つ目の傾向からは、H3K4me1 と H4K8ac の片方が有意に増加し、もう一方が逆に有意に減少する遺伝子はほとんど存在しないことが分かり、これは H3K4me1 と H4K8ac が単独で作用するのではなくお互いに協働していることを示していると考えられる。また二つ目と三つめの傾向からは、H3K4me1 と H4K8ac が共に有意に増加すること、つまりこれらの修飾が揃った領域において DNA ヒドロキシメチル化が生じると考えられる。すなわちこれは、H3K4me1 と H4K8ac の後に 5hmC が生じるという、ファクター間の生成順序を示唆していると考えられる。この結果に加えて、1.1 節で示した bivalent promoter 領域における PRC2 複合体と TET タンパクの相互作用の事例から[6, 7]、次のような hESC における DNA ヒドロキシメチル化に関する機構が予想される（図 29）。

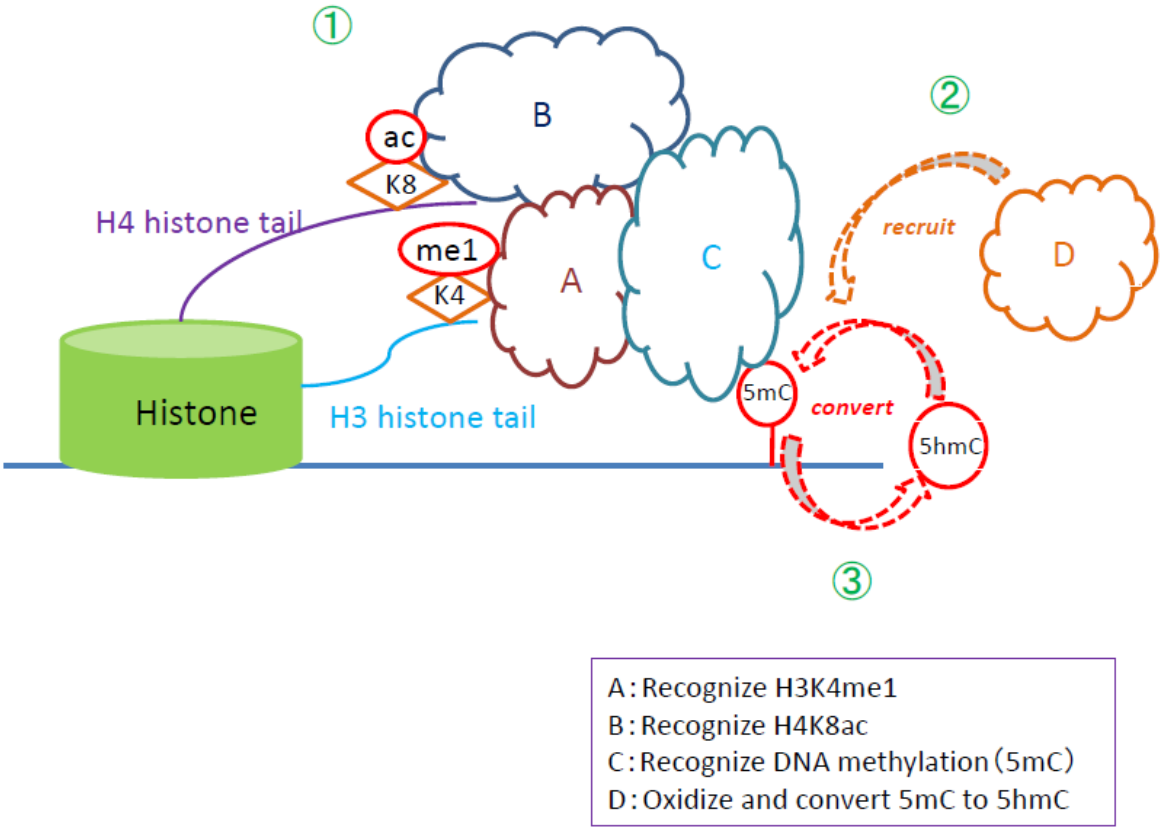


図 29：hESC における、DNA ヒドロキシメチル化機構（予想図）。H3K4me1 を認識するドメイン（A）、H4K8ac を認識するドメイン（B）、及び DNA メチル化を認識するドメイン（C）を併せ持つある複合体が、H3K4me1 と H4K8ac を目印にして、その DNA 上の部位へ呼び寄せられる。さらにドメイン C により 5mC を捉え、そこに DNA 脱メチル化関連酵素（D）を呼び寄せることにより、5mC を 5hmC に変換する。

ヒトには Tip60（Alias：KAT5）というタンパク質が存在し、これは H3K4me1 を認識して H4K8 をアセチル化することができる。しかし PFAM データベース（http://pfam.xfam.org/）には、この Tip60 に DNA メチル化を認識する MBD ドメインは登録されていない。しかしこの Tip60 が、MBD ドメインを持つある複合体のサブユニットとなることにより、今回予測した機構を達成できる可能性はある。そこで hESC における Tip60 の発現量を調べた結果、全遺伝子の平均発現量と比較して顕著な発現量増加が見られた（図 30）。この結果は Tip60 の、この機構に対する候補タンパクとしての可能性を支持するものである。

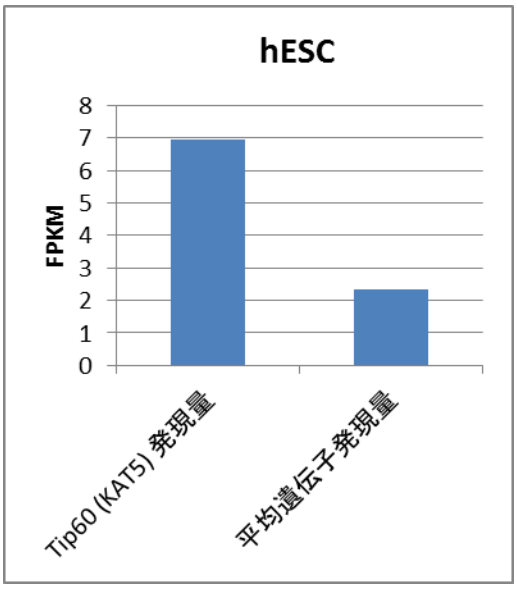


図 30：hESC における遺伝子発現量比較。Tip60 の発現量は、全遺伝子平均と較べて顕著な増加傾向を示している。

さらに DNA 脱メチル化タンパク（TET）の分布強度を、5hmC、H3K4me1、及び H4K8ac の分布強度と比較して、図 12 及び図 13 と同様の方法で相関係数を調べた結果、中程度の相関が見られた（図 31）。いずれも顕著に高い値ではないが、これは使用したヒト ES 細胞株の違いが一因であると考えられる（TET は HUES8 細胞株のデータを使用し、5hmC、H3K4me1、及び H4K8ac は H1 細胞株のデータを使用した）。5hmC が TET タンパクの作用により生成することからこれらは同じ領域に分布し、その相関は高くなることが予想される。この TET と 5hmC の相関（PCC：0.42）と相対的に同程度の相関が、TET と H3K4me1（PCC：0.44）及び TET と H4K8ac（PCC：0.30）間においても観察されたことは、TET が図 29 の機構図における DNA 脱メチル化関連酵素（図中の D）である可能性を示唆していると考えられる。

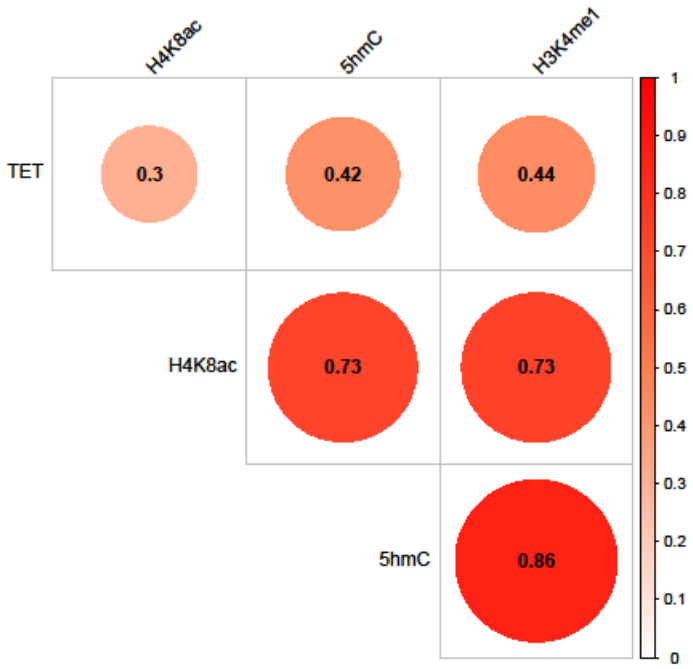


図 31 : TET タンパクの遺伝子毎の分布強度と、他のファクター（5hmC、H3K4me1、及び H4K8ac）強度間のピアソン相関係数。

今後に向けての課題として、今回の研究では共起ファクター間の生成順序について言及したが、より詳細なファクター相互作用機序・因果関係について解明するためには公共データに加えて、三次元構造に関するデータ（Hi-C）や 5hmC のさらなる酸化中間体である 5fC・5caC データ（図 1）などを揃えた、統合的な解析が不可欠である。さらにそれぞれのファクターに関する ESC 分化系データを加えて解析することにより、この共起が持つ生物学的特徴や時系列に沿った各ファクターの変動傾向などについて、より一層深い考察が得られるはずである。また今回は gene body 及び promoter 領域に焦点を当てて研究を行ったが、H3K4me1 はエンハンサー領域のマーカーとしても認識されており[51]、この領域における共起についても解析が求められる。

今回発見した 5hmC、H3K4me1、及び H4K8ac からなる共起は新規のエピジェネティックコード候補となり、多能性幹細胞の性質に関わっている可能性を秘めている。さらには従来の細胞間の遺伝子発現変化を基本とした解析法を補完することができる、エピジェネティック解析の新たな有用性をも示すものである。

謝 辞

本研究を行うにあたり、皆様方より終始適切な助言を賜り、また丁寧に御指導頂きました。ここに感謝の意を表したいと思います。東京大学新領域創成科学研究科の中井謙太教授には、研究テーマを始めとする多くの面で御指導頂きました。また、東京大学医科学研究所の朴聖俊特任講師にはコンピューターに関する基本的知識から始まり、研究の進め方、論文の書き方に至るまで幅広く相談に乗って頂き、終始適切な助言を頂きました。最後に、研究生活において大変お世話になりました中井研究室の皆様方に心から御礼申しあげます。本当にありがとうございました。

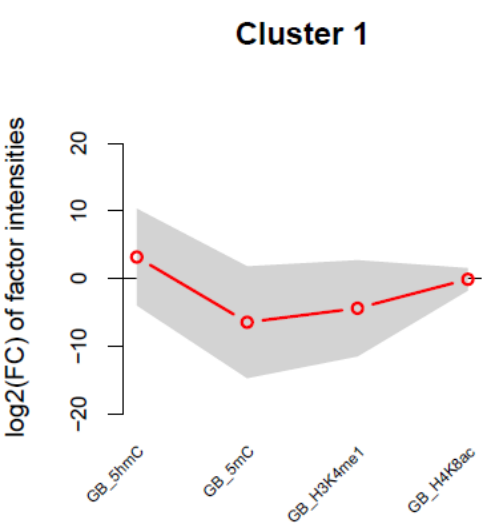
参考文献

1. Kim, K., et al., *Epigenetic memory in induced pluripotent stem cells.* Nature, 2010. **467**(7313): p. 285-90.
2. Pastor, W.A., et al., *Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells.* Nature, 2011. **473**(7347): p. 394-397.
3. Szulwach, K.E., et al., *Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells.* PLoS Genet, 2011. **7**(6): p. e1002154.
4. Bernstein, B.E., et al., *A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells.* Cell, 2006. **125**(2): p. 315-326.
5. Vastenhouw, N.L. and A.F. Schier, *Bivalent histone modifications in early embryogenesis.* Curr Opin Cell Biol, 2012. **24**(3): p. 374-86.
6. Wu, H., et al., *Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells.* Nature, 2011. **473**(7347): p. 389-93.
7. Neri, F., et al., *Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells.* Genome Biol, 2013. **14**(8): p. R91.
8. Yang, Y.A., et al., *FOXA1 potentiates lineage-specific enhancer activation through modulating TET1 expression and function.* Nucleic Acids Res, 2016. **44**(17): p. 8153-64.
9. Mahe, E.A., et al., *Cytosine modifications modulate the chromatin architecture of transcriptional enhancers.* Genome Res, 2017. **27**(6): p. 947-958.
10. Maleszka, R., *Epigenetic code and insect behavioural plasticity.* Curr Opin Insect Sci, 2016. **15**: p. 45-52.
11. Maleszka, R., P.H. Mason, and A.B. Barron, *Epigenomics and the concept of degeneracy in biological systems.* Brief Funct Genomics, 2014. **13**(3): p. 191-202.
12. Turner, B.M., *Defining an epigenetic code.* Nat Cell Biol, 2007. **9**(1): p. 2-6.
13. Laird, P.W., *Principles and challenges of genomewide DNA methylation analysis.* Nat Rev Genet, 2010. **11**(3): p. 191-203.
14. Li, E., *Chromatin modification and epigenetic reprogramming in mammalian development.* Nat Rev Genet, 2002. **3**(9): p. 662-73.
15. Tucker, K.L., *Methylated cytosine and the brain: a new base for neuroscience.* Neuron, 2001. **30**(3): p. 649-52.
16. Okano, M., S. Xie, and E. Li, *Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases.* Nat Genet, 1998. **19**(3): p. 219-20.
17. Ball, M.P., et al., *Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells.* Nat Biotechnol, 2009. **27**(4): p. 361-8.
18. Hackett, J.A., et al., *Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine.* Science, 2013. **339**(6118): p. 448-52.

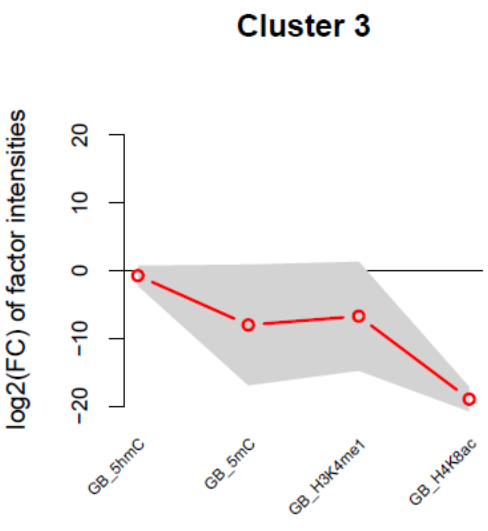
19. Kriaucionis, S. and N. Heintz, *The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain.* Science, 2009. **324**(5929): p. 929-30.
20. Munzel, M., et al., *Quantification of the sixth DNA base hydroxymethylcytosine in the brain.* Angew Chem Int Ed Engl, 2010. **49**(31): p. 5375-7.
21. Szwagierczak, A., et al., *Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA.* Nucleic Acids Res, 2010. **38**(19): p. e181.
22. Pfeifer, G.P., S. Kadam, and S.G. Jin, *5-hydroxymethylcytosine and its potential roles in development and cancer.* Epigenetics Chromatin, 2013. **6**(1): p. 10.
23. He, Y.F., et al., *Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA.* Science, 2011. **333**(6047): p. 1303-7.
24. Tahiliani, M., et al., *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, 2009. **324**(5929): p. 930-5.
25. Penn, N.W., et al., *The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid.* Biochem J, 1972. **126**(4): p. 781-90.
26. Zhao, M.T., et al., *Methylated DNA immunoprecipitation and high-throughput sequencing (MeDIP-seq) using low amounts of genomic DNA.* Cell Reprogram, 2014. **16**(3): p. 175-84.
27. Nestor, C.E. and R.R. Meehan, *Hydroxymethylated DNA immunoprecipitation (hmeDIP).* Methods Mol Biol, 2014. **1094**: p. 259-67.
28. Peng, J., B. Xia, and C. Yi, *Single-base resolution analysis of DNA epigenome via high-throughput sequencing.* Sci China Life Sci, 2016. **59**(3): p. 219-26.
29. Yu, M., et al., *Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome.* Cell, 2012. **149**(6): p. 1368-1380.
30. Lee, D.D. and H.S. Seung, *Learning the parts of objects by non-negative matrix factorization.* Nature, 1999. **401**(6755): p. 788-91.
31. Yang, Z. and G. Michailidis, *A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data.* Bioinformatics, 2016. **32**.
32. Li, Y. and A. Ngom, *The non-negative matrix factorization toolbox for biological data mining.* Source code for biology and medicine, 2013. **8**.
33. Brunet, J.P., et al., *Metagenes and molecular pattern discovery using matrix factorization.* Proc Natl Acad Sci U S A, 2004. **101**(12): p. 4164-9.
34. Hutchins, L.N., et al., *Position-dependent motif characterization using non-negative matrix factorization.* Bioinformatics, 2008. **24**(23): p. 2684-90.
35. Frigyesi, A. and M. Hoglund, *Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes.* Cancer Inform, 2008. **6**: p. 275-92.
36. Wang, T., et al., *Subtelomeric hotspots of aberrant 5-hydroxymethylcytosine-mediated epigenetic modifications during reprogramming to pluripotency.* Nat Cell Biol, 2013. **15**(6): p. 700-11.

37. Hanzelmann, S., et al., *Replicative senescence is associated with nuclear reorganization and with DNA methylation at specific transcription factor binding sites.* Clin Epigenetics, 2015. **7**: p. 19.
38. Sammons, M.A., et al., *TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity.* Genome Res, 2015. **25**(2): p. 179-88.
39. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.
40. Ferrari, K.J., et al., *Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity.* Mol Cell, 2014. **53**(1): p. 49-62.
41. Ma, H., et al., *Abnormalities in human pluripotent cells due to reprogramming mechanisms.* Nature, 2014. **511**(7508): p. 177-83.
42. Choi, J., et al., *A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs.* Nat Biotechnol, 2015. **33**(11): p. 1173-81.
43. Schmieder, R., et al., *TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets.* BMC Bioinformatics, 2010. **11**: p. 341.
44. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
45. Kim, D., et al., *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.
46. Stroud, H., et al., *5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells.* Genome Biology, 2011. **12**(6): p. R54.
47. Cho, H., et al., *A human RNA polymerase II complex containing factors that modify chromatin structure.* Mol Cell Biol, 1998. **18**(9): p. 5355-63.
48. Drukker, M., et al., *Human embryonic stem cells and their differentiated derivatives are less susceptible to immune rejection than adult cells.* Stem Cells, 2006. **24**(2): p. 221-9.
49. Pearl, J.I., et al., *Pluripotent stem cells: immune to the immune system?* Sci Transl Med, 2012. **4**(164): p. 164ps25.
50. Okita, K., N. Nagata, and S. Yamanaka, *Immunogenicity of induced pluripotent stem cells.* Circ Res, 2011. **109**(7): p. 720-1.
51. Local, A., et al., *Identification of H3K4me1-associated proteins at mammalian enhancers.* Nat Genet, 2018. **50**(1): p. 73-82.

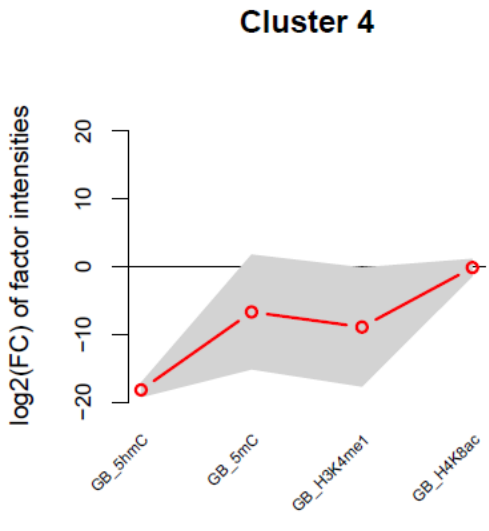
Supplemental figures



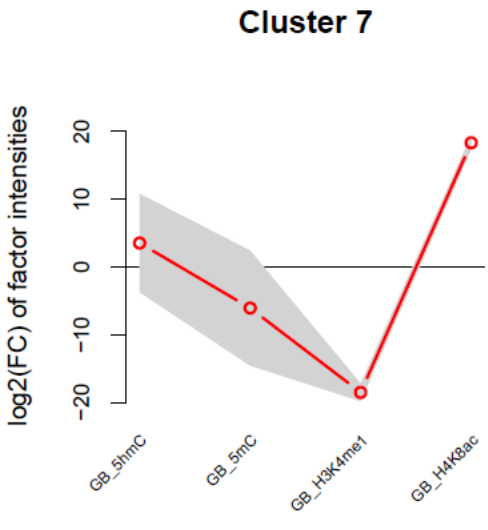
(2186 genes)



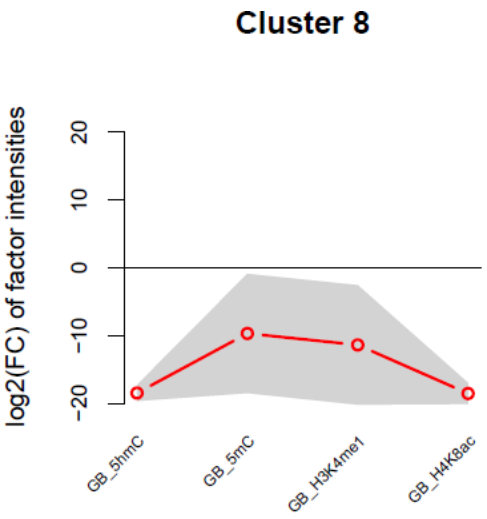
(449 genes)



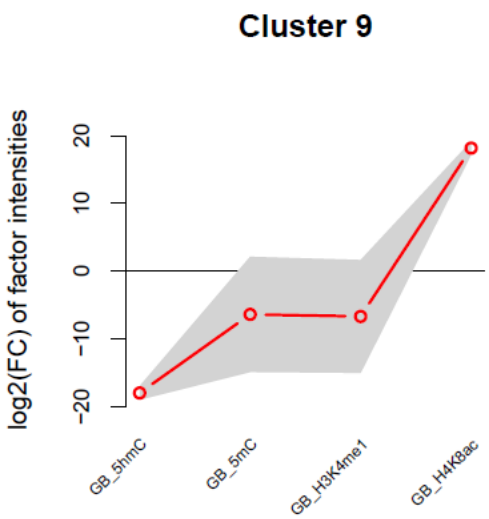
(340 genes)



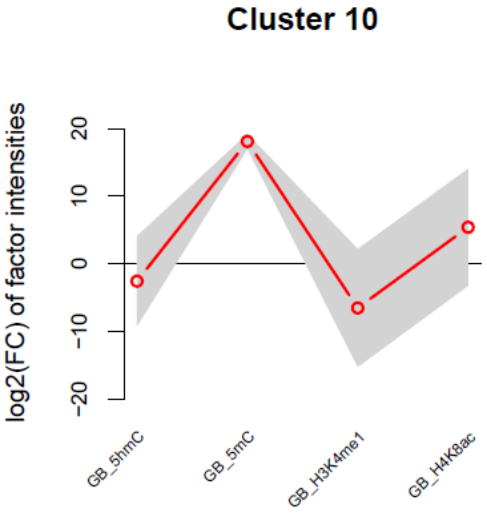
(198 genes)



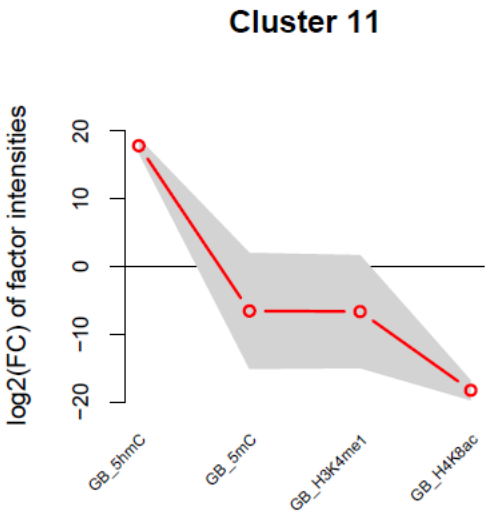
(167 genes)



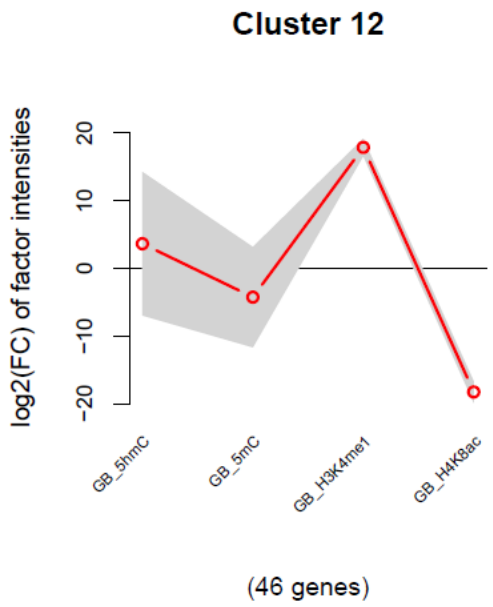
(132 genes)



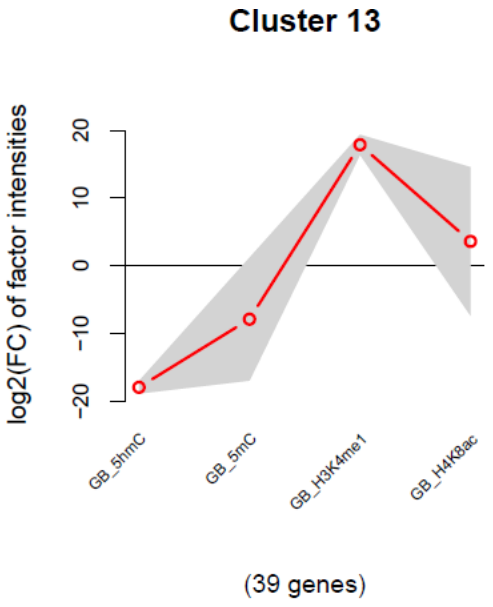
(106 genes)



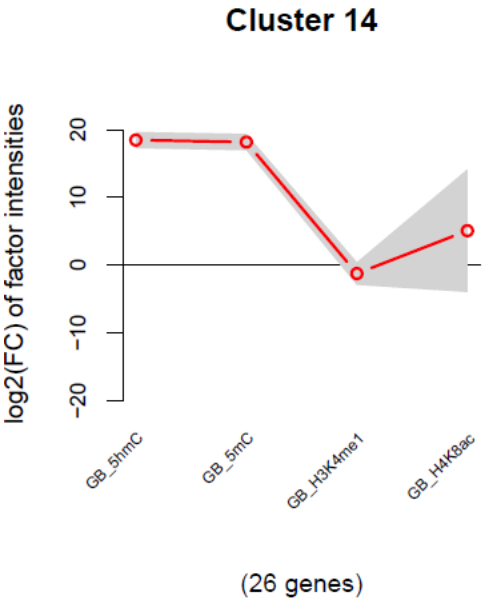
(89 genes)



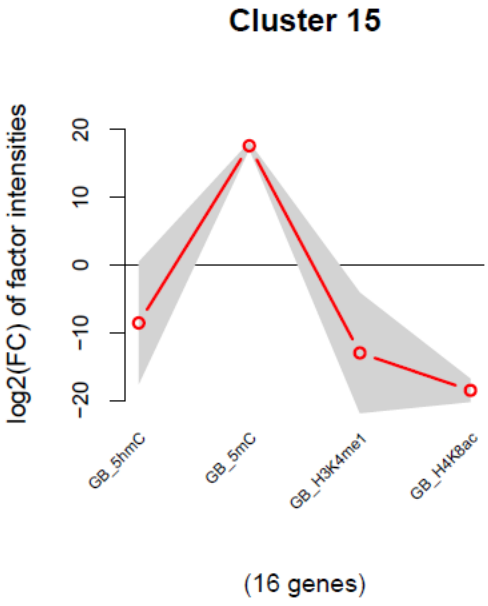
(46 genes)



(39 genes)



(26 genes)



(16 genes)

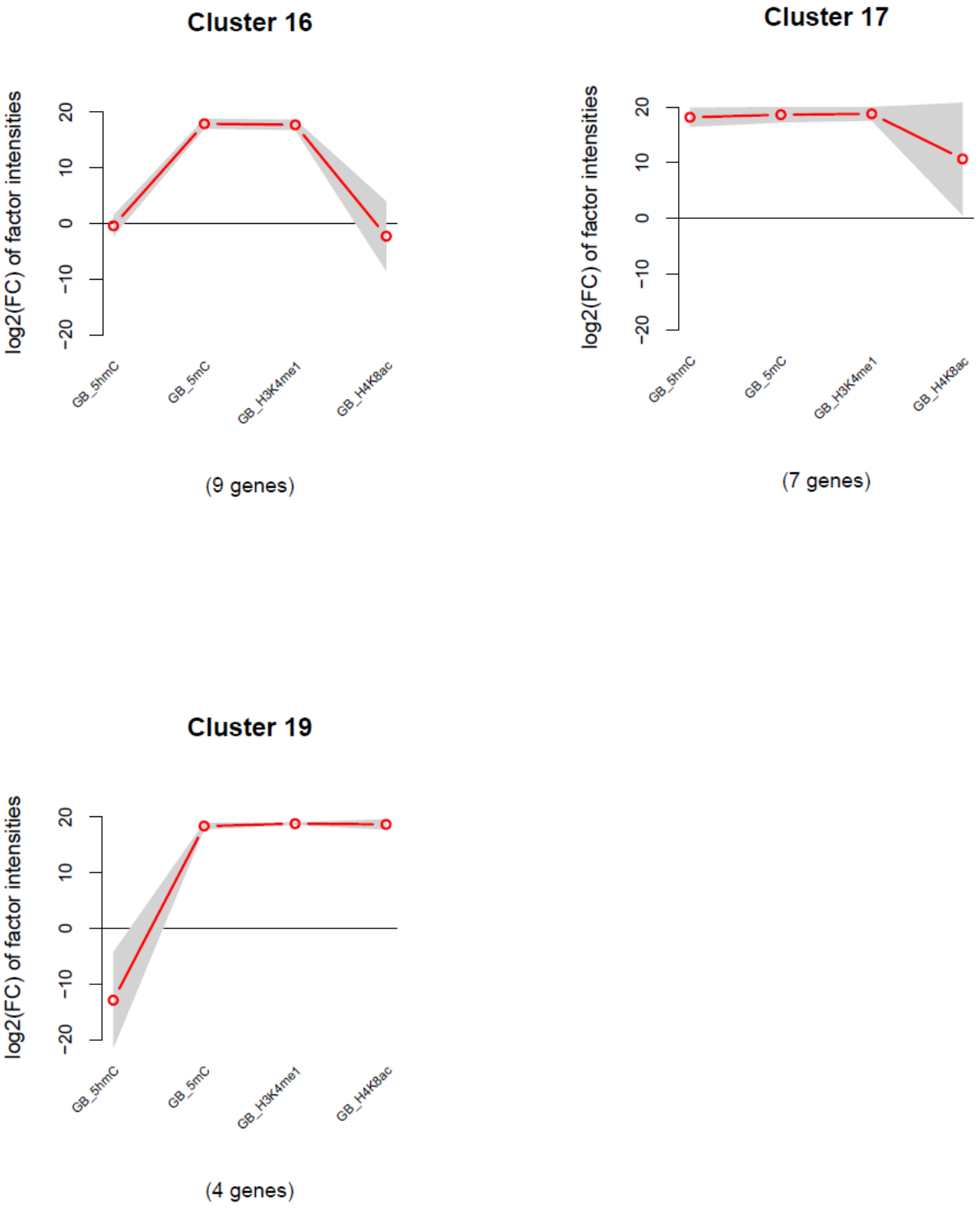


図 S1：Gene body 領域を対象とした階層的クラスタリングによる、19 個のクラスター遺伝子群が持つ 5hmC、5mC、H3K4me1、及び H4K8ac 変動パターン（本文中の cluster 2、5、6、及び 18 を除く）。GB は gene body を意味する。縦軸は体細胞・hESC 間の各ファクター強度変化比を log で表している。

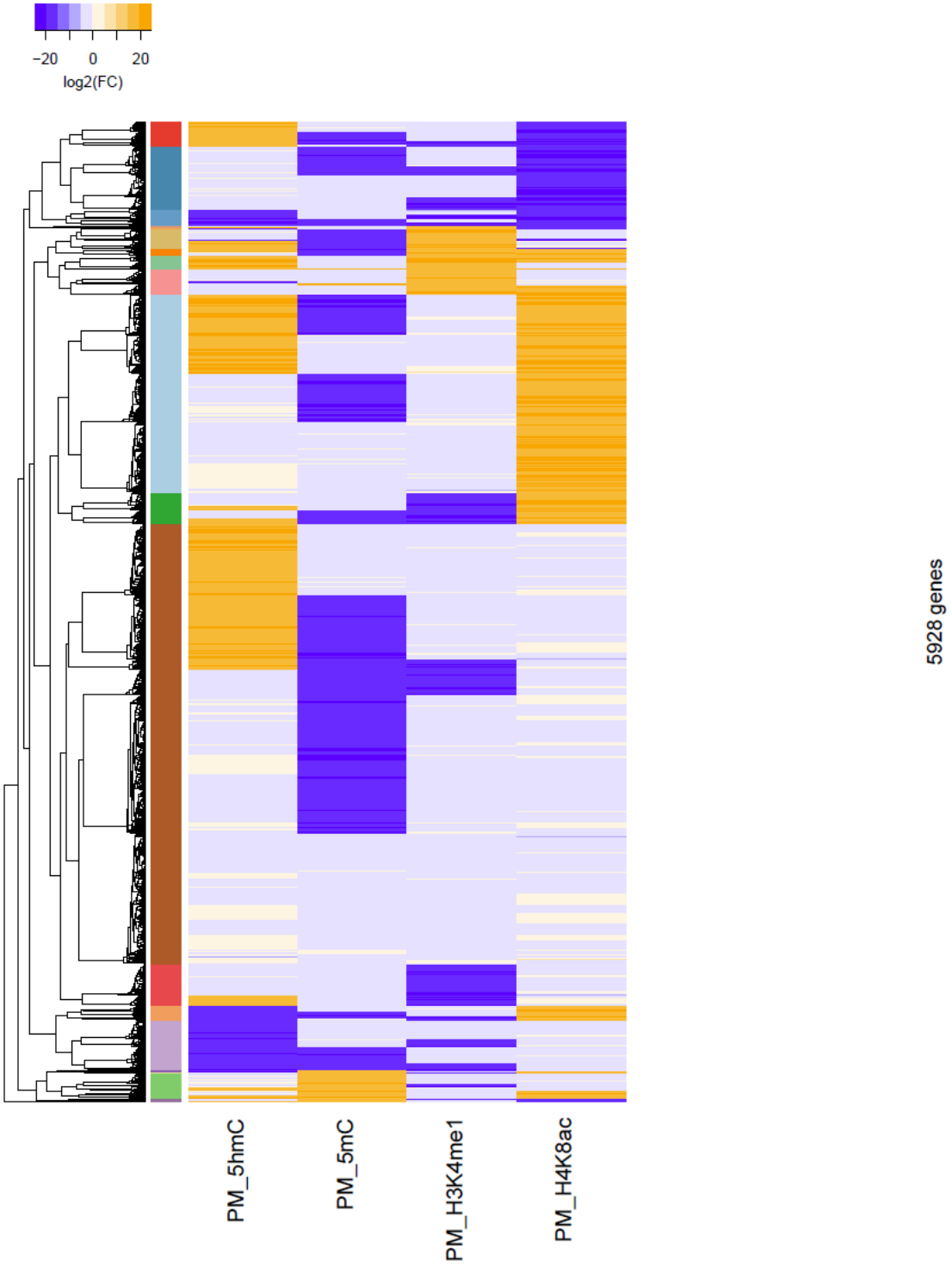
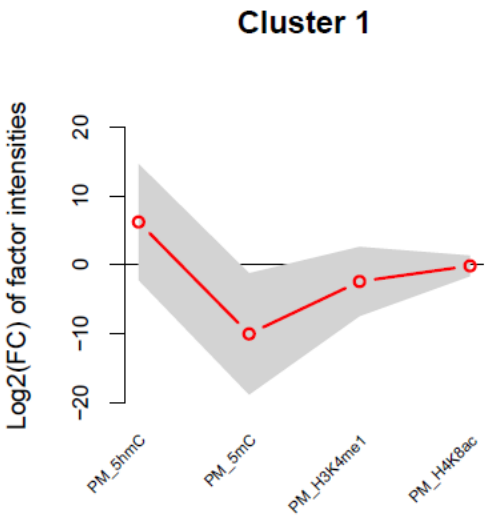
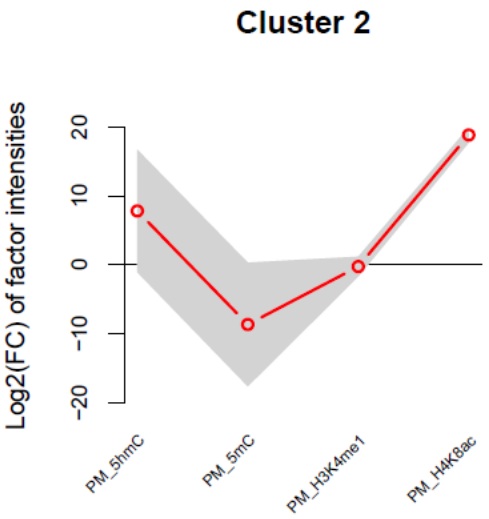


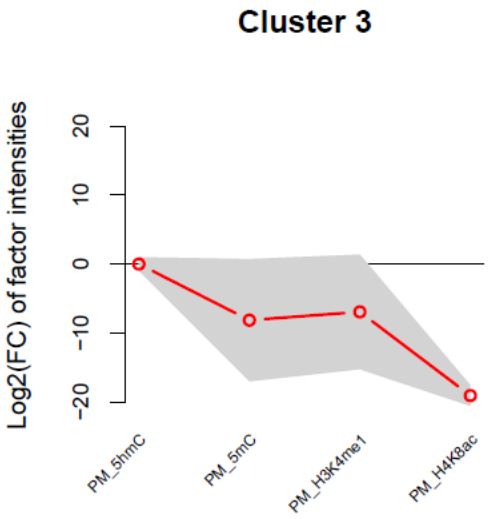
図 S2：各エピジェネティックファクターの promoter 領域における、体細胞・ESC 間変動パターンに基づいた全遺伝子の階層的クラスタリング。PM は promoter を意味する。いずれのファクターにも変化が見られなかった遺伝子群を除外した結果、5928 個の遺伝子群に 20 種類の特徴的なエピジェネティックファクター変動パターンが見られた。



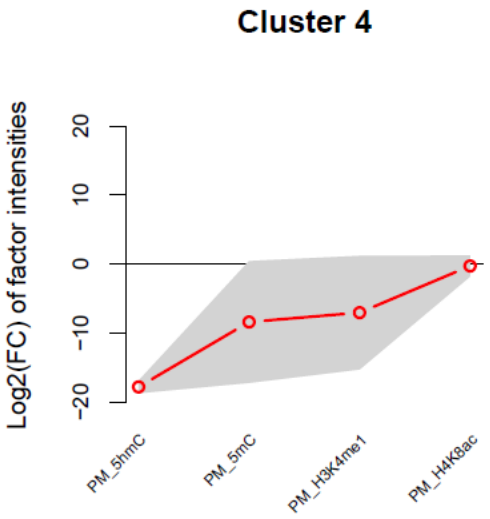
(2662 genes)



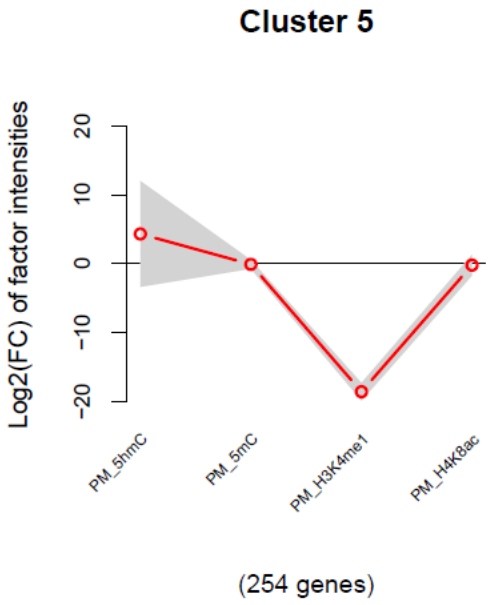
(1202 genes)



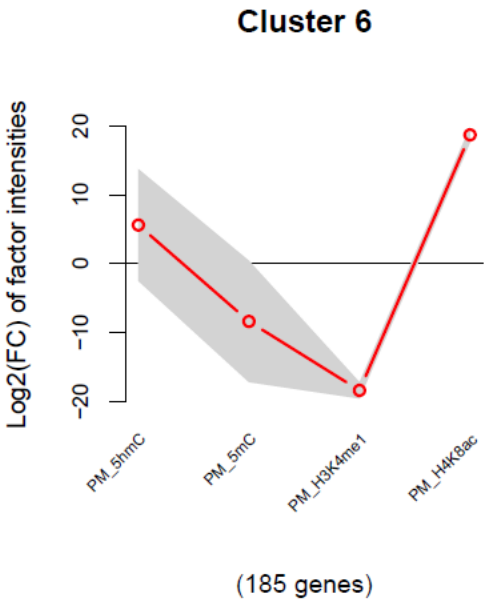
(385 genes)



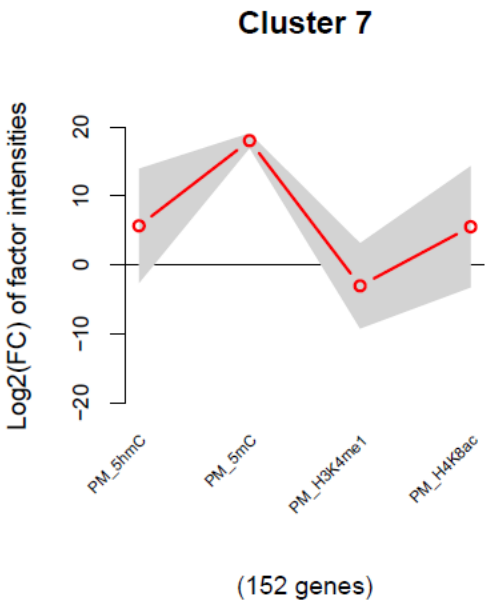
(296 genes)



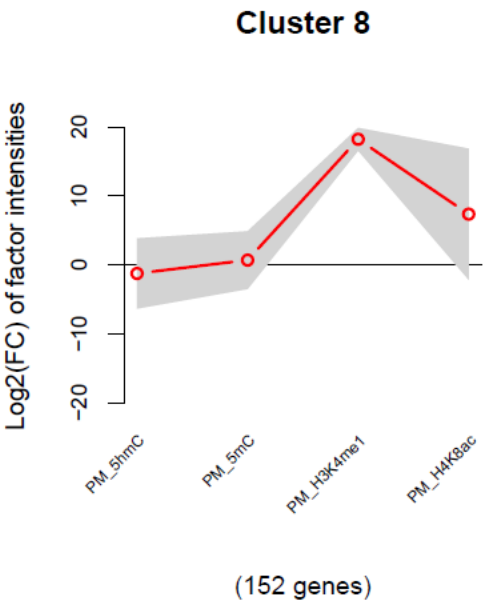
(254 genes)



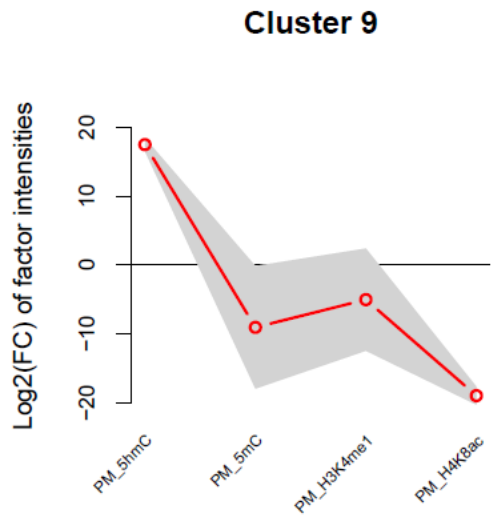
(185 genes)



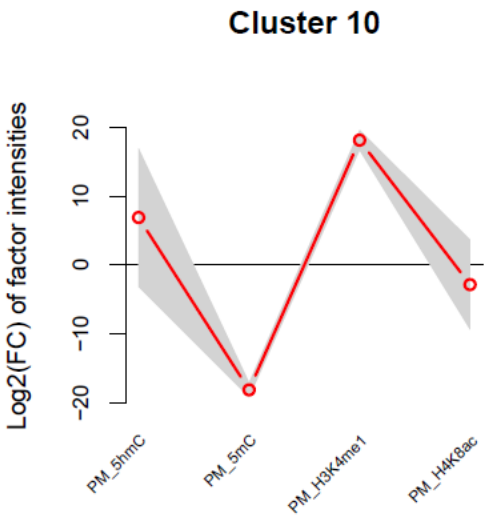
(152 genes)



(152 genes)

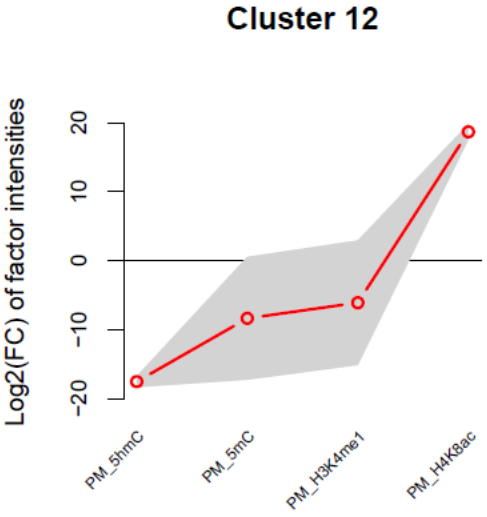


(142 genes)

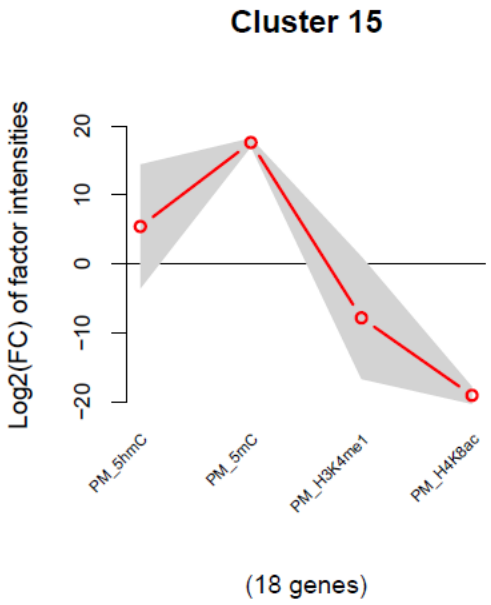


(121 genes)

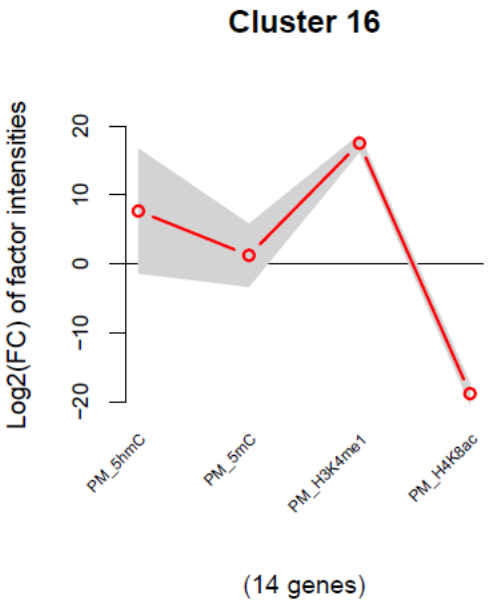
(96 genes)



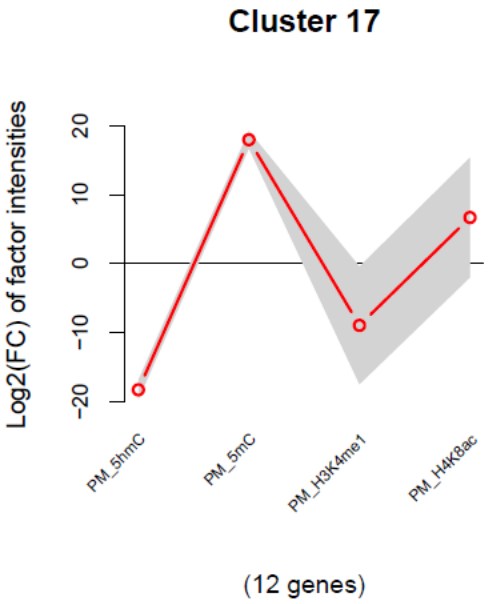
(92 genes)



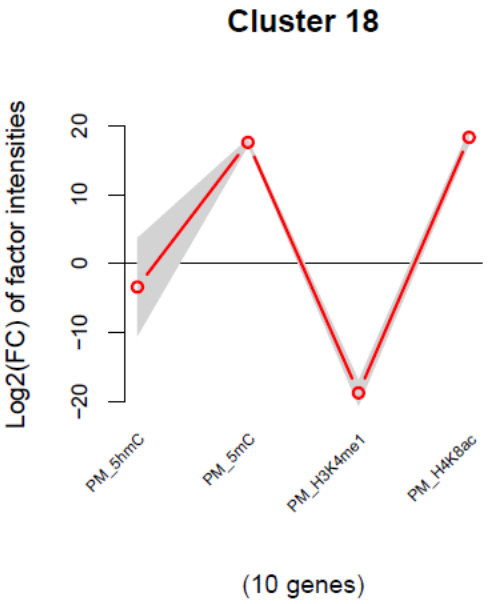
(18 genes)



(14 genes)



(12 genes)



(10 genes)

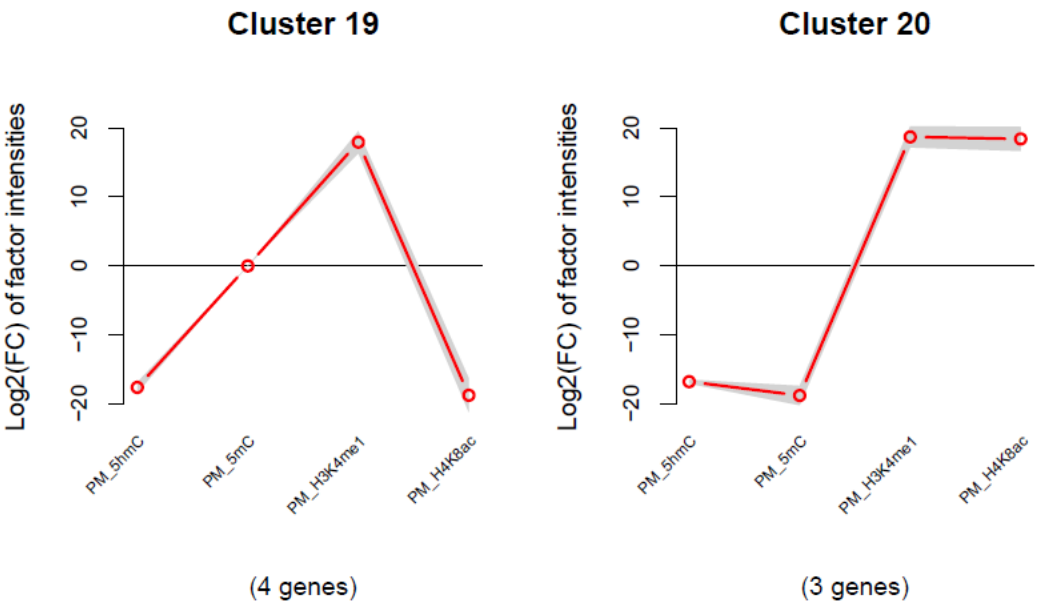
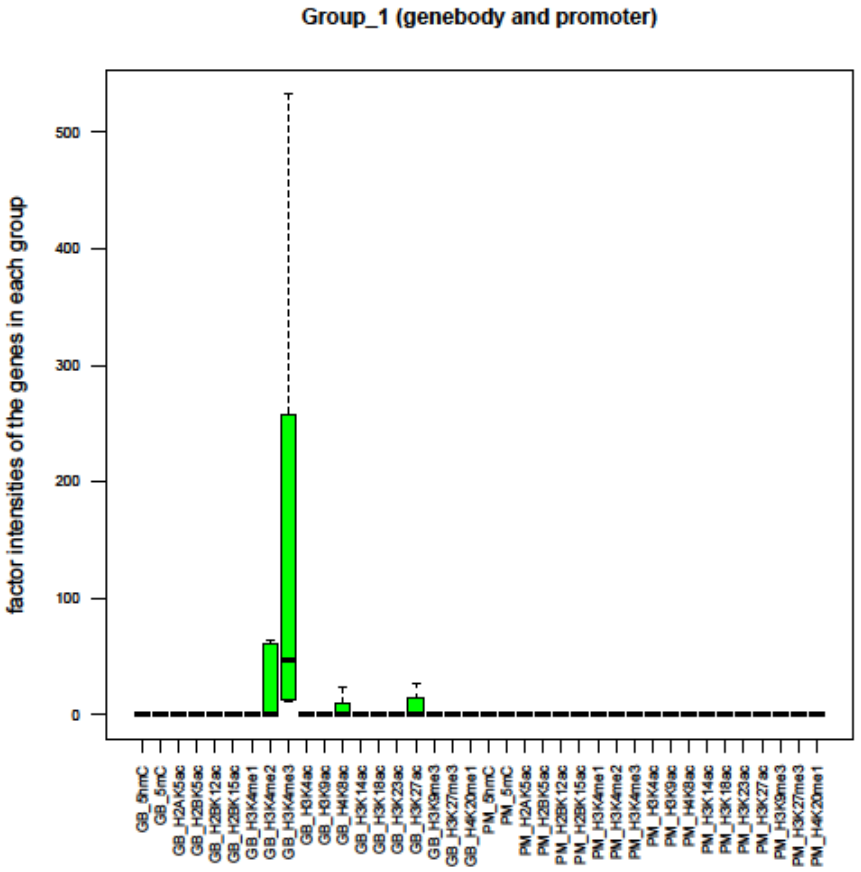
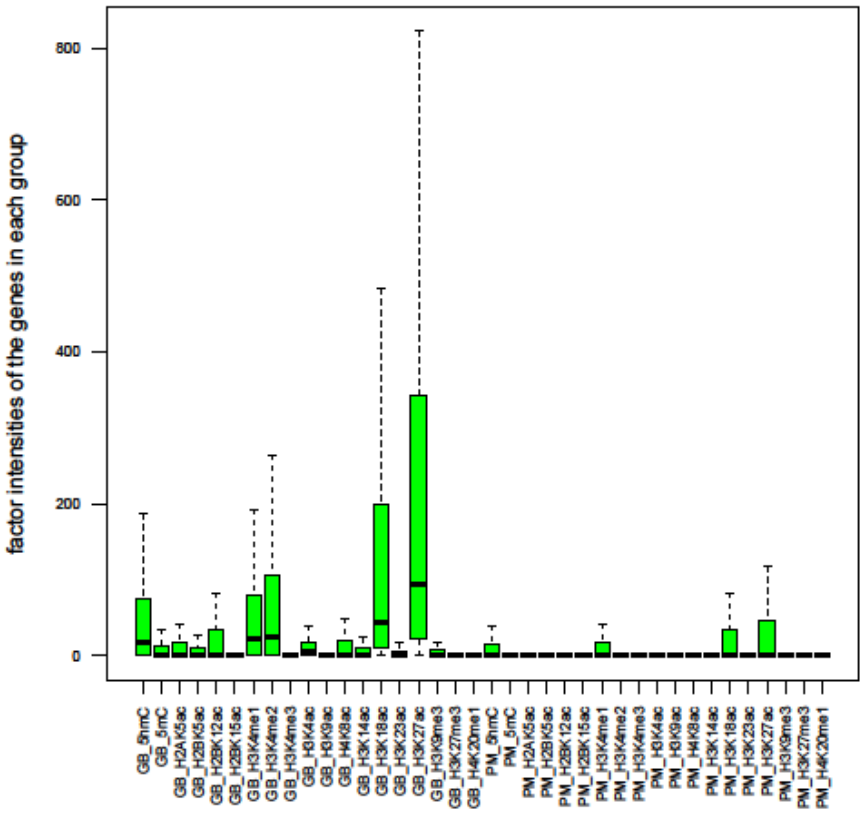
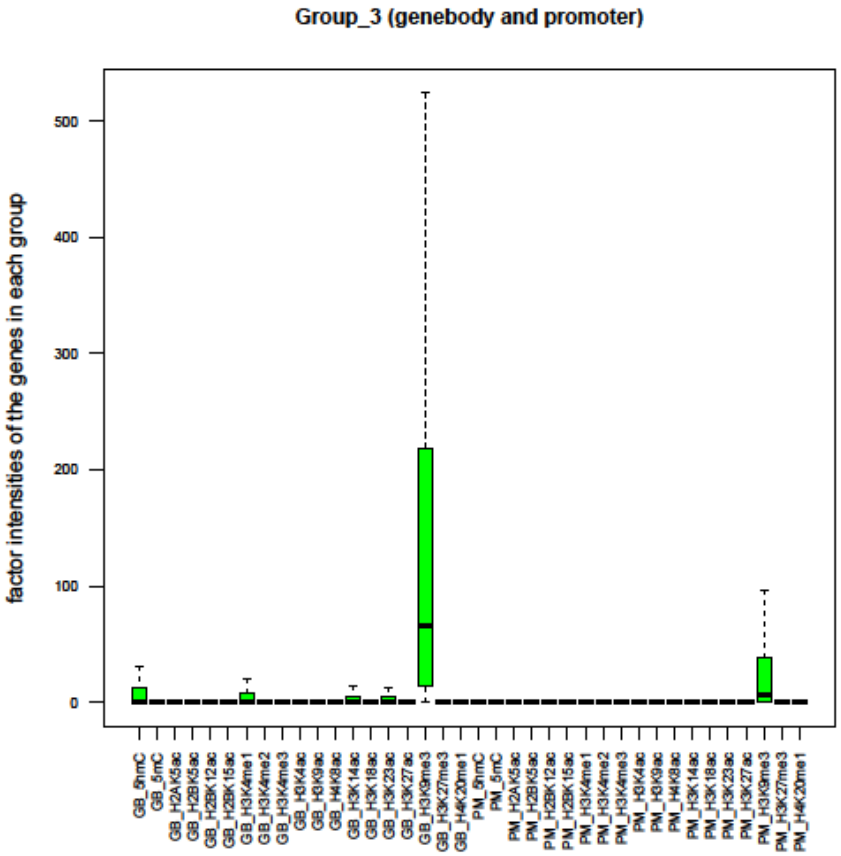
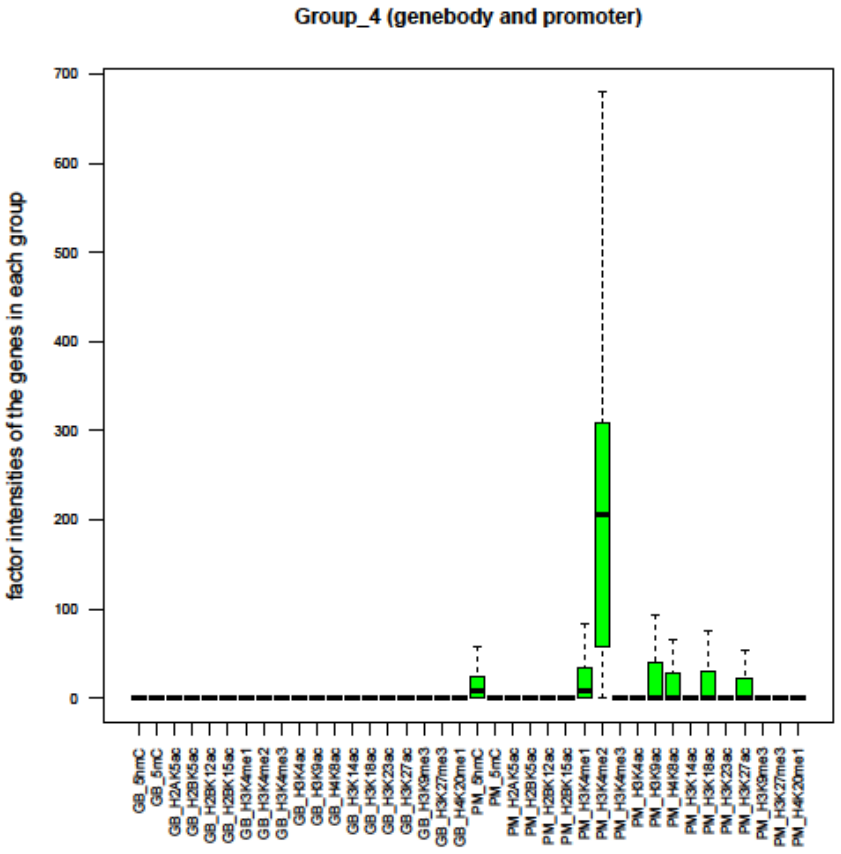


図 S3：Promoter 領域を対象とした階層的クラスタリングによる、20 個のクラスター遺伝子群が持つ 5hmC、5mC、H3K4me1、及び H4K8ac 変動パターン（本文中の cluster13 及び 14 を除く）。PM は promoter を意味する。縦軸は体細胞・hESC 間の各ファクター強度変化比を log で表している。


Group_1 (genebody and promoter)
factor intensities of the genes in each group
0
100
200
300
400
500
GB_5hmC
GB_5mC
GB_H2AK5ac
GB_H2BK5ac
GB_H2BK12ac
GB_H2BK15ac
GB_H3K4me1
GB_H3K4me2
GB_H3K4me3
GB_H3K4ac
GB_H3K9ac
GB_H4K8ac
GB_H3K14ac
GB_H3K18ac
GB_H3K23ac
GB_H3K27ac
GB_H3K9me3
GB_H3K27me3
GB_H4K20me1
PM_5hmC
PM_5mC
PM_H2AK5ac
PM_H2BK5ac
PM_H2BK12ac
PM_H2BK15ac
PM_H3K4me1
PM_H3K4me2
PM_H3K4me3
PM_H3K4ac
PM_H3K9ac
PM_H4K8ac
PM_H3K14ac
PM_H3K18ac
PM_H3K23ac
PM_H3K27ac
PM_H3K9me3
PM_H3K27me3
PM_H4K20me1


Group_2 (genebody and promoter)
factor intensities of the genes in each group
0
200
400
600
800
GB_5hmC
GB_5mC
GB_H2AK5ac
GB_H2BK5ac
GB_H2BK12ac
GB_H2BK15ac
GB_H3K4me1
GB_H3K4me2
GB_H3K4me3
GB_H3K4ac
GB_H3K9ac
GB_H4K8ac
GB_H3K14ac
GB_H3K18ac
GB_H3K23ac
GB_H3K27ac
GB_H3K9me3
GB_H3K27me3
GB_H4K20me1
PM_5hmC
PM_5mC
PM_H2AK5ac
PM_H2BK5ac
PM_H2BK12ac
PM_H2BK15ac
PM_H3K4me1
PM_H3K4me2
PM_H3K4me3
PM_H3K4ac
PM_H3K9ac
PM_H4K8ac
PM_H3K14ac
PM_H3K18ac
PM_H3K23ac
PM_H3K27ac
PM_H3K9me3
PM_H3K27me3
PM_H4K20me1


Group_3 (genebody and promoter)
factor intensities of the genes in each group
0
100
200
300
400
500
GB_5hmC
GB_5mC
GB_H2AK5ac
GB_H2BK5ac
GB_H2BK12ac
GB_H2BK15ac
GB_H3K4me1
GB_H3K4me2
GB_H3K4me3
GB_H3K4ac
GB_H3K9ac
GB_H4K8ac
GB_H3K14ac
GB_H3K18ac
GB_H3K23ac
GB_H3K27ac
GB_H3K9me3
GB_H3K27me3
GB_H4K20me1
PM_5hmC
PM_5mC
PM_H2AK5ac
PM_H2BK5ac
PM_H2BK12ac
PM_H2BK15ac
PM_H3K4me1
PM_H3K4me2
PM_H3K4me3
PM_H3K4ac
PM_H3K9ac
PM_H4K8ac
PM_H3K14ac
PM_H3K18ac
PM_H3K23ac
PM_H3K27ac
PM_H3K9me3
PM_H3K27me3
PM_H4K20me1


Group_4 (genebody and promoter)
factor intensities of the genes in each group
0
100
200
300
400
500
600
700
GB_5hmC
GB_5mC
GB_H2AK5ac
GB_H2BK5ac
GB_H2BK12ac
GB_H2BK15ac
GB_H3K4me1
GB_H3K4me2
GB_H3K4me3
GB_H3K4ac
GB_H3K9ac
GB_H4K8ac
GB_H3K14ac
GB_H3K18ac
GB_H3K23ac
GB_H3K27ac
GB_H3K9me3
GB_H3K27me3
GB_H4K20me1
PM_5hmC
PM_5mC
PM_H2AK5ac
PM_H2BK5ac
PM_H2BK12ac
PM_H2BK15ac
PM_H3K4me1
PM_H3K4me2
PM_H3K4me3
PM_H3K4ac
PM_H3K9ac
PM_H4K8ac
PM_H3K14ac
PM_H3K18ac
PM_H3K23ac
PM_H3K27ac
PM_H3K9me3
PM_H3K27me3
PM_H4K20me1

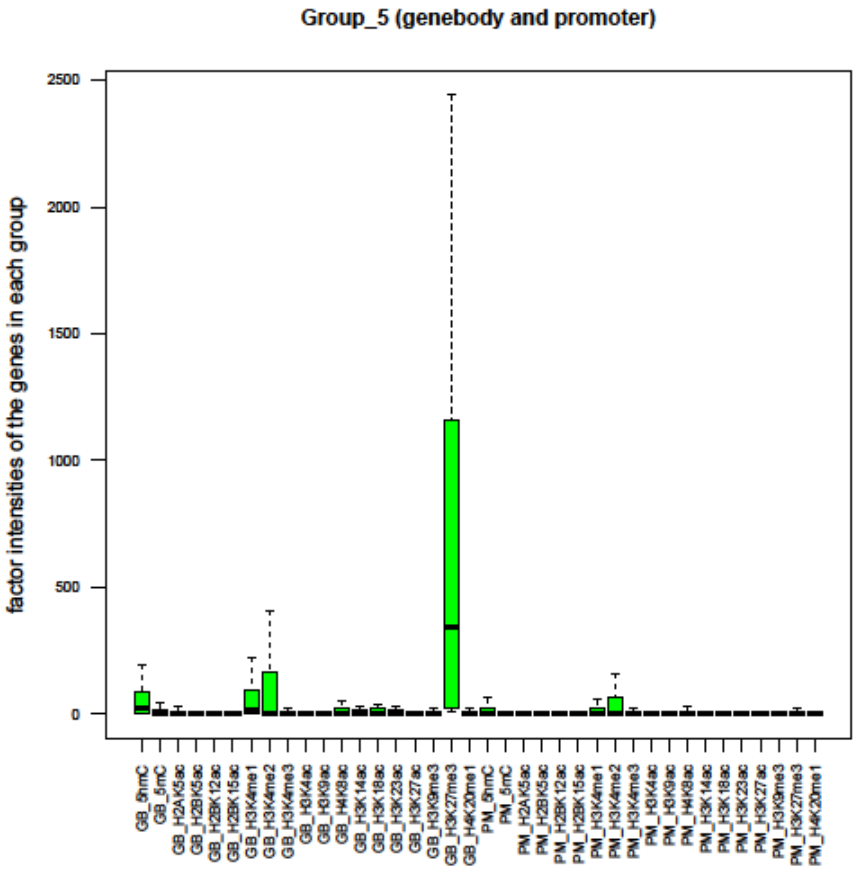
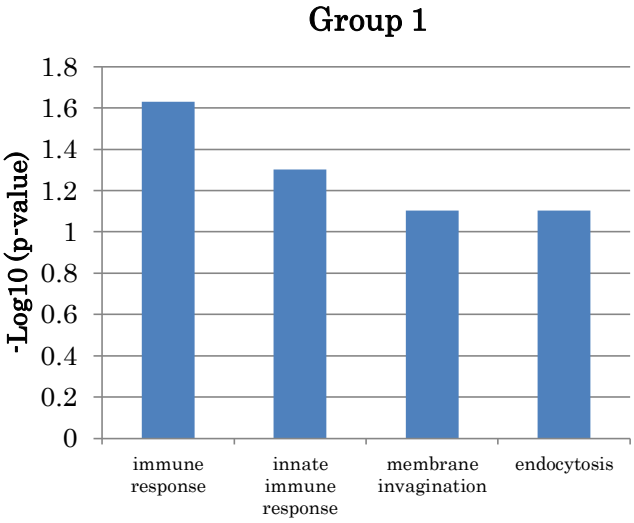
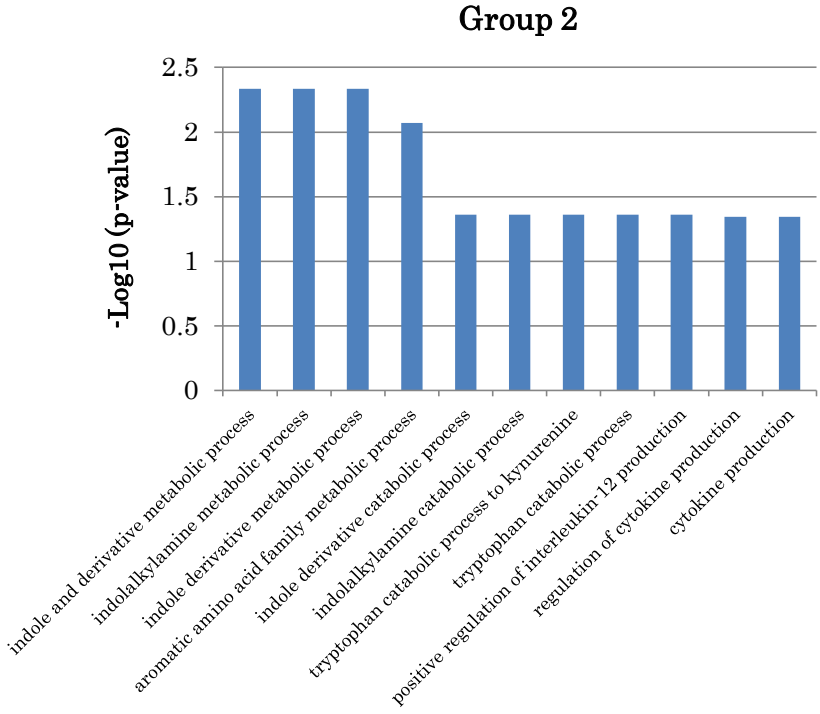
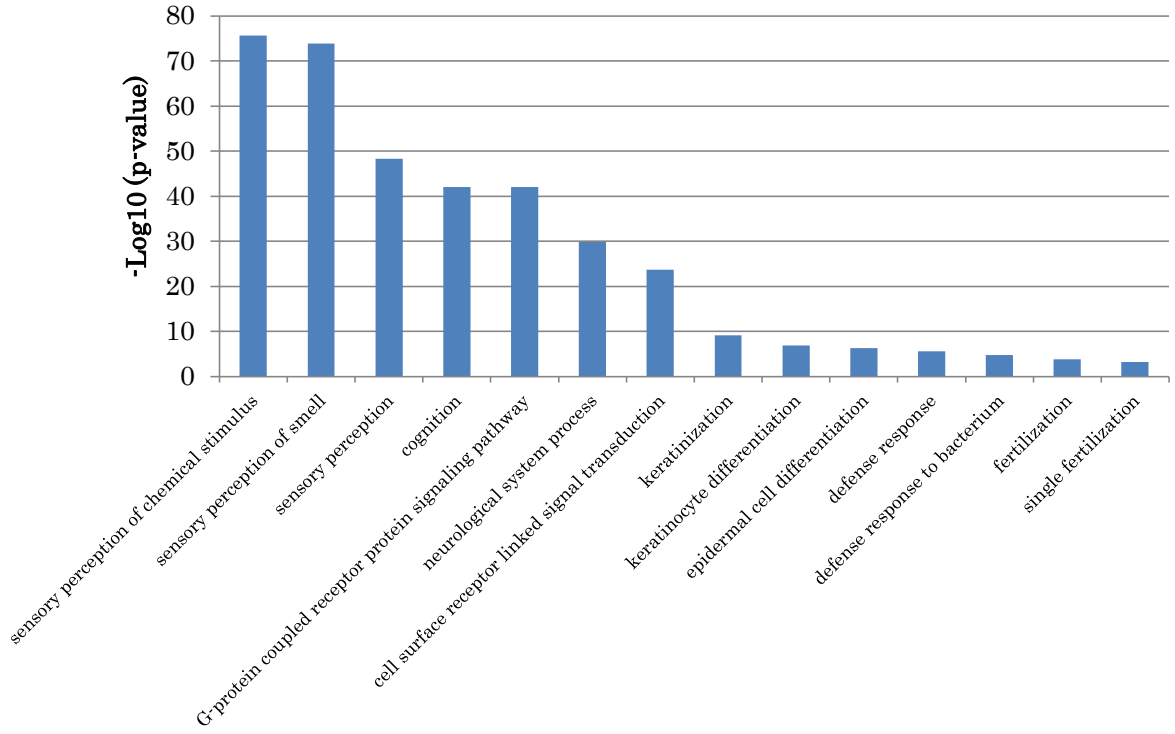


図 S4：NMF 解析により得られた各グループに含まれる遺伝子群の、エピジェネティックファクター強度分布（本文中の group 6、7、及び 8 を除く）。ボックスプロットの縦軸は各エピジェネティックファクターの強度を表しており、横軸の GB は gene body、PM は promoter を意味する。

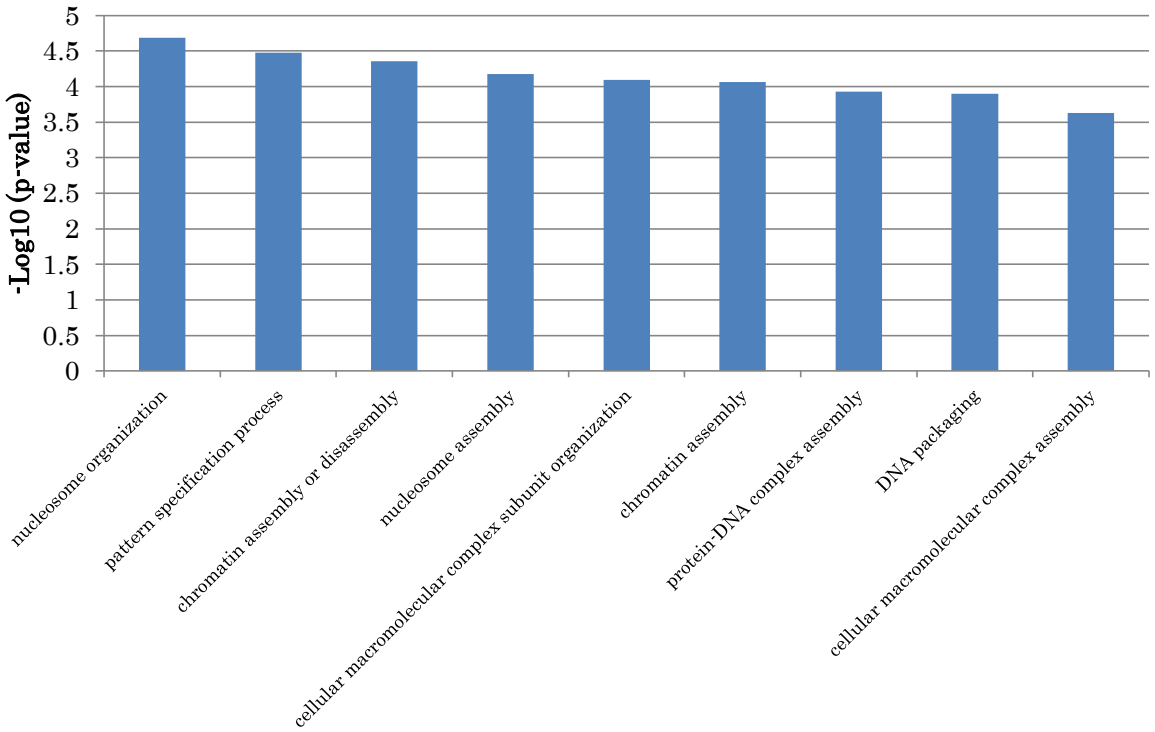

Group 1
-Log10 (p-value)
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
immune response
innate immune response
membrane invagination
endocytosis


Group 2
-Log10 (p-value)
2.5
2
1.5
1
0.5
0
indole and derivative metabolic process
indolalkylamine metabolic process
indole derivative metabolic process
aromatic amino acid family metabolic process
indole derivative catabolic process
indolalkylamine catabolic process
tryptophan catabolic process to kynurenine
tryptophan catabolic process
positive regulation of interleukin-12 production
regulation of cytokine production
cytokine production

Group 3



Group 4



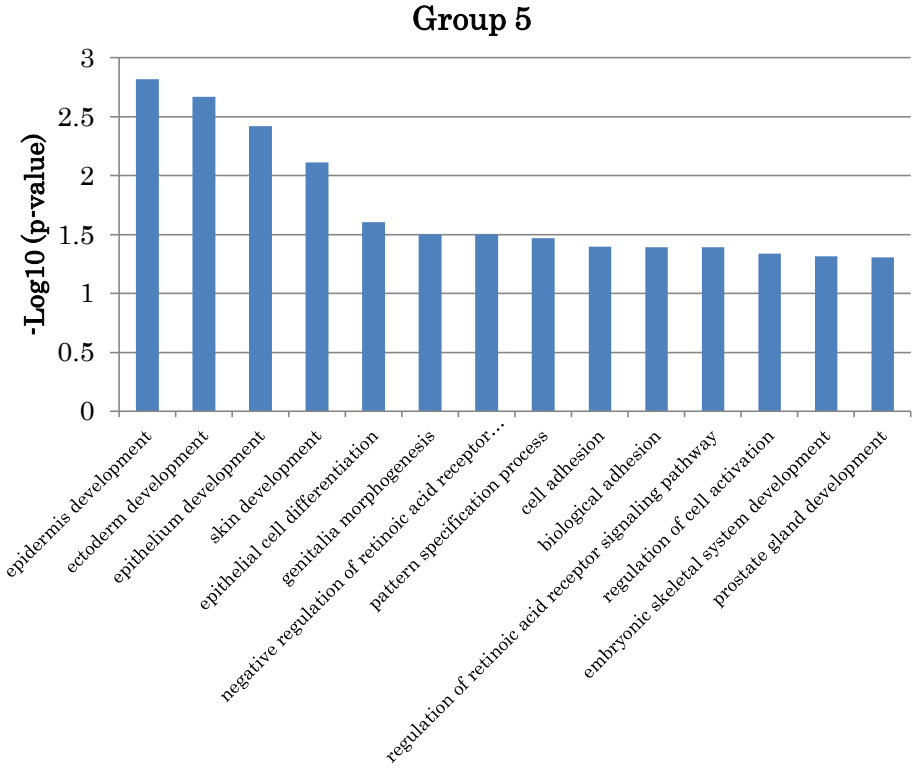


図 S5：NMF 解析により得られた各グループに含まれる遺伝子群の、gene ontology 解析結果（本文中の group 6、7、及び 8 を除く）。Y 軸は解析により得られた p-value 値を対数で表したものである。

Supplemental table

表 S1 : NMF 解析による group 8 遺伝子群の中の、“immune response” 関連遺伝子。

Gene symbol	Gene name
CLEC10A	C-type lectin domain family 10, member A
CLEC4C	C-type lectin domain family 4, member C
CD14	CD14 molecule
CD27	CD27 molecule
CD7	CD7 molecule
CD79B	CD79b molecule, immunoglobulin-associated beta
EBI3	Epstein-Barr virus induced 3
FCER1G	Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide
GPR183	G protein-coupled receptor 183
LIME1	Lck interacting transmembrane adaptor 1
ARHGDIB	Rho GDP dissociation inhibitor (GDI) beta
WAS	Wiskott-Aldrich syndrome (eczema-thrombocytopenia)
AQP9	aquaporin 9
CARD9	caspase recruitment domain family, member 9
CCL17	chemokine (C-C motif) ligand 17
CCR4	chemokine (C-C motif) receptor 4
CSF2	colony stimulating factor 2 (granulocyte-macrophage)
C4BPB	complement component 4 binding protein, beta
CCR6, CCNL2	cyclin L2; chemokine (C-C motif) receptor 6
CST7	cystatin F (leukocystatin)
ERAP2	endoplasmic reticulum aminopeptidase 2
FCN2	ficolin (collagen/fibrinogen domain containing lectin) 2 (hucolin)
FCN1	ficolin (collagen/fibrinogen domain containing) 1
FCN3	ficolin (collagen/fibrinogen domain containing) 3 (Hakata antigen)
GBP4	guanylate binding protein 4
ITGAD	integrin, alpha D
IL1R2	interleukin 1 receptor, type II
IL10	interleukin 10
IL31	interleukin 31
LILRB2	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 2
LST1	leukocyte specific transcript 1
LAIR1	leukocyte-associated immunoglobulin-like receptor 1
LAT	linker for activation of T cells
LAT2	linker for activation of T cells family, member 2
PGLYRP1	peptidoglycan recognition protein 1
PDCD1	programmed cell death 1
RAG2	recombination activating gene 2
PRKDC	similar to protein kinase, DNA-activated, catalytic polypeptide; protein kinase, DNA-activated, catalytic polypeptide
S1PR4	sphingosine-1-phosphate receptor 4
TOLLIP	toll interacting protein
TLR7	toll-like receptor 7
TLR8	toll-like receptor 8
TLR9	toll-like receptor 9
TNF	tumor necrosis factor (TNF superfamily, member 2)
TNFSF10	tumor necrosis factor (ligand) superfamily, member 10
TNFSF13B	tumor necrosis factor (ligand) superfamily, member 13b
TNFRSF4	tumor necrosis factor receptor superfamily, member 4
ZAP70	zeta-chain (TCR) associated protein kinase 70kDa