

# テンソル分解と主成分分解を用いた教師なし学習による 社会性昆虫のカーストの脳の遺伝子発現プロファイルと DNAメチル化の解析

田口善弘<sup>1,a)</sup>

**概要:** ゲノムが表現型の主たる支配者だという前提は揺らがないにしろ、同じゲノムを保持していると目される同種内の個体でも大きな表現型の差は頻繁に見られる。ゲノムでは性差が決まらない種（魚など）の雌雄差、幼体と成体の大きな差（ホヤ、両生類）などはよく知られた例である。なかでも社会性昆虫におけるカースト（女王 vs ワーカー）間個体差は表現型だけではなく行動差（役割の差）をもたらすという意味で何がキーとなってカーストを引き起こしているのかがよく分かっていない。本稿ではゲノムのメチル化というエピジェネティックな差がカースト間の差を引き起こしているのではという前提で mRNA の発現差とメチル化の発現差をアリとハチという社会性昆虫のカースト間で比較し、ハチについてはメチル化差と発現差がカースト間で同時に起きている遺伝子を同定することに成功したので報告する。

## 1. はじめに

ゲノムが表現型の主たる支配者だという事実は進化を議論する上の根幹であり揺らがない。その一方で、同じゲノムを保持すると目される同種の個体間で顕著な表現型差がある例も多く見られる。例えば、人間の場合、性差はゲノムに刻み込まれているので、男女間で表現型に大きな差があることに疑問を感じることはないだろうが、性差がゲノムではなく生育環境で変わってしまう種は実は多い。例えば、魚類 [2] では、性別は成長過程での環境に左右されることは有名である（養殖うなぎは全てオスになるなどの事例がある）。また、幼体と成体の著しい体型の差異は広く観測される。例えば、ホヤ（幼体はオタマジャクシ様だが、成体は固着生物）や両生類（幼体はエラ呼吸水棲生物であるが、成体は肺呼吸（半）陸上生物）、完全変態昆虫（蝶、甲虫などは幼体はいわゆる芋虫型であるが、成体は羽をもった飛翔生物である）などは、同一個体であるので、ゲノムは共通であるはずであるが、表現型は著しい差異を持っている。このような差異がなぜ生じるかの分子生物学的な基盤の理解は非常に不十分である。

このような表現型の差異にエピジェネティックスの差異が貢献している可能性は古くから指摘されてきた。エピジェ

ネティックスはDNAのメチル化、ヒストン修飾、非コードRNAの発現量、クロマチンの状態、など、DNAの塩基配列に影響を与えること無く、遺伝子の発現パターンを変えてしまう全てのファクターを意味するが、もともと、同じDNAを持っている細胞が異なった分化をするためには、このような因子が影響を与えていることは想像に難しくなく、従って、それが個体レベルの表現型にも影響を与えていることは当然、想定されるからだ。

本稿ではこのような観点から社会性昆虫のカースト [3] に焦点を当てる。社会性昆虫とは、アリ、ハチやシロアリの様に、単に一箇所に集まって生活を営むにとどまらず、分業に基づく資源の共有や、一個体では作成不可能な構造物の創出を行う昆虫である。例えば、ハチでは、個体のサイズに比べて分不相応な巨大な巣を作る種が存在する（ミツバチ、スズメバチなど）。これらの巣は、巣のサイズが個体サイズの数十倍から百倍にまで達するために、共同作業なしに構築することはほぼ不可能である。また、餌の取得についても、アリ、ハチ、シロアリ、などは巣の外部に赴いて資源採集を行う個体群と、巣の内部にとどまり次世代の育成に関わる個体群が厳然と別れている（時系列的に役割を異にする個体群につきつぎと渡り歩くことはありうる）。このような場合、外見の差異（女王と呼ばれる、通常、各コロニーに一体しか存在しない、産卵だけを業務とする個体とワーカーと呼ばれる労働階級の個体では女王の方が数倍体が大きい）もさることながら、行動的な差異も大き

<sup>1</sup> 中央大学理工学部物理学科

<sup>a)</sup> tag@granular.com

本研究は国際会議 APBC2018 で発表済み、プロシーディングスとして出版済みである [1]。

く、このような差異がどんな分子生物学的な基盤に支えられているのかはまだ明らかになっていない。

この様な問題についてエピジェネティクスの方面からアプローチしようという試みが近年見られる [4]。Patalano らは *Polistes canadensis* (英名: red paper wasp, 和名なし、スズメバチの一種) と *Dinoponera quadriceps* (英名: 不明, 和名: ディノハリアリ) の mRNA 発現量と DNA のメチル化をシーケンシング技術で解析し、DNA メチル化と mRNA 発現量の女王とワーカーの差異を解明しようと試みた [5]。また、Ferreira らは *Polistes canadensis* の 4 つのカーストについて mRNA 発現プロファイルを測定し、解析した [6]。本研究では彼らのデータを「主成分分析を用いた教師なし学習による変数選択」及び「テンソル分解を用いた教師なし学習による変数選択」の 2 つの方法 [7] で再解析した結果を報告したい。

## 2. データと解析手法

### 2.1 mRNA 発現プロファイルと DNA メチル化のデータ

Patalano ら [5] の *P. canadensis* と *D. quadriceps* の mRNA 発現プロファイルと DNA メチル化のデータは Gene Expression Omnibus (GEO) [8] から GEO ID:GSE59525 のデータセットをダウンロードした。mRNA 発現プロファイルについては、*P. canadensis* では女王 4 サンプル、ワーカー 6 サンプルの計 10 サンプル、*D. quadriceps* では、女王 7 サンプル、ワーカー 6 サンプルの計 13 サンプルを使用した。いずれも著者らが計算した RPKM 化された発現量のファイルを用いた。DNA メチル化については、*P. canadensis* と *D. quadriceps* 共に、コントロール 1 サンプル、女王 3 サンプル、ワーカー 3 サンプルの計 7 サンプルが提供されていた。

メチル化については遺伝子と関係づけるため

$$\frac{\sum_{\text{メチル化サイト} \in \text{遺伝子}} \text{メチル化値}}{\sum_{\text{メチル化サイト} \in \text{遺伝子}} \text{メチル化値} + \sum_{\text{メチル化サイト} \in \text{遺伝子}} \text{非メチル化値}}$$

を各遺伝子の領域に属するメチル化サイトで和をとった量を計算した (昆虫ではプロモーターではなく、遺伝子領域のメチル化が発現量に関係していると思われるため)。

Ferreira ら [6] の mRNA 発現プロファイルについては、当該論文の supplementary file で提供されているものを使用した (Additional file 9)。このファイルには、女王とワーカー以外に 2 つのカースト (foundress, 新女王、及び、callow, 若年ワーカー) の計 4 つのカーストを含んでおり、サンプル数はそれぞれ、4、6、2、2 であった。

### 2.2 主成分分析を用いた教師なし学習による変数選択

$x_{ij} \in \mathbb{R}^{N \times M}$  を  $i$  番目の遺伝子の  $j$  番目のサンプルの遺伝子発現プロファイル、または、メチル化値とする。 $x_{ij}$  は  $\sum_i x_{ij} = 0, \sum_i x_{ij}^2 = N$  となるように規格化されていると

する。ここで  $i$  番目の遺伝子の第  $k$  主成分得点は

$$XX^T \mathbf{u}_k = \lambda_k \mathbf{u}_k$$

で求まる固有ベクトル  $\mathbf{u}_k$  の第  $i$  成分である。ここで  $X$  は  $x_{ij}$  を成分とする行列、 $\mathbf{u}_k \in \mathbb{R}^N$  は正方対角行列  $XX^T \in \mathbb{R}^{N \times N}$  の第  $k$  固有ベクトル、 $\lambda_k$  は同行列の第  $k$  固有値である。また、 $j$  番目のサンプルの第  $k$  主成分負荷量は

$$\mathbf{v}_k = X^T \mathbf{u}_k$$

で定義されるベクトル  $\mathbf{v}_k \in \mathbb{R}^M$  の第  $j$  成分である。 $\mathbf{v}_k$  は

$$X^T X \mathbf{v}_k = X^T X X^T \mathbf{u}_k = X^T \lambda_k \mathbf{u}_k = \lambda_k \mathbf{v}_k$$

という計算より  $X^T X \in \mathbb{R}^{M \times M}$  の第  $k$  固有ベクトルでもある。

通常、遺伝子発現プロファイルに主成分分析を適用する場合は主成分得点がサンプルに付与されるように適用されるが、今回、それが逆になっていることに注意して頂きたい。

$\mathbf{v}_k$  のうち、注目するサンプル依存性を持っている成分を同定する。この場合、カースト間で差がある  $\mathbf{v}_k$  を選ぶ。選ばれた  $k$  の集合を  $\Omega_k$  としよう。ここで  $\mathbf{u}_k, k \in \Omega_k$  が多重ガウス分布に従っていると仮定して (帰無仮説)、 $i$  番目の遺伝子に棄却確率  $P_i$  を

$$P_i = P_{\chi^2} \left[ > \sum_{k \in \Omega_k} \left( \frac{u_{ki}}{\sigma_k} \right)^2 \right] \quad (1)$$

で付与する。 $\sigma_k$  は  $u_{ki}$  の標準偏差、 $P_{\chi^2}[> x]$  は値が  $x$  より大きい場合の  $\chi$  二乗分布の累積寄与率である。 $P_i$  は Benjamini-Hochberg 法 [9] で多重比較補正する。補正  $P$  値が 0.01 以下の遺伝子を所与のサンプル依存性を有意に持っている遺伝子として選択する。

### 2.3 テンソル分解を用いた教師なし学習による変数選択

$x_{ij} \in \mathbb{R}^{N \times M}$  を  $i$  番目の遺伝子の  $j$  番目のサンプルの遺伝子発現プロファイル、 $x_{i\ell} \in \mathbb{R}^{N \times L}$  を  $i$  番目の遺伝子の  $\ell$  番目のサンプルの DNA メチル化の値、としよう。ここでテンソル  $x_{ij\ell} \in \mathbb{R}^{N \times M \times L}$  を

$$x_{ij\ell} = x_{ij} x_{i\ell}$$

と定義しよう。さらに  $\sum_i x_{ij\ell} = 0, \sum_i x_{ij\ell}^2 = N$  と規格化されているとする。これにテンソル分解 (高次特異値分解、Higher order singular value decomposition, HOSVD [10]) を適用すると

$$x_{ij\ell} = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{k_3=1}^{K_3} G(k_1, k_2, k_3) u_{k_1 i} u_{k_2 j} u_{k_3 \ell}$$

を得る。ここで  $G(k_1, k_2, k_3) \in \mathbb{R}^{N \times M \times L}$  はコアテンソル、

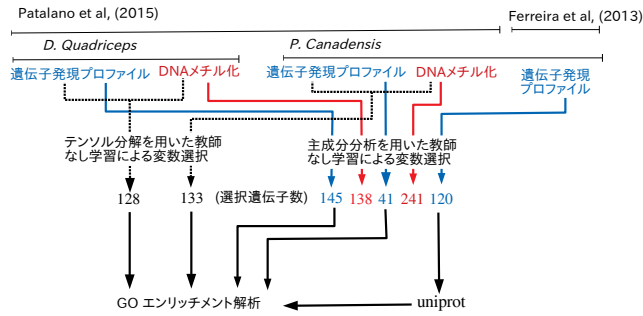


図 1 解析のフローチャート  
Fig. 1 Flow chart

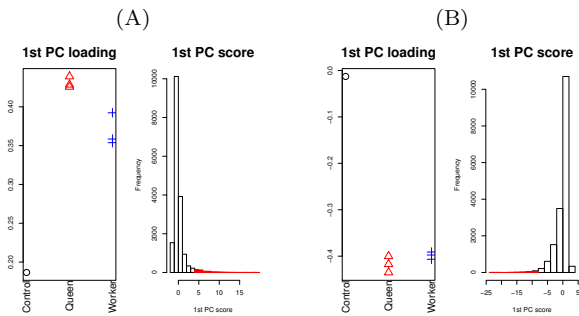


図 2 第一主成分負荷量のカースト依存性 (左) と得点の分布 (右)。  
(A) *P. canadensis* (B) *D. quadriceps*

Fig. 2 The dependence of the first PC loading upon caste(left) and the histogram of the first PC scores(right) (A) *P. canadensis* (B) *D. quadriceps*

$\mathbf{u}_{k_1} \in \mathbb{R}^N$ ,  $\mathbf{u}_{k_2} \in \mathbb{R}^M$ ,  $\mathbf{u}_{k_3} \in \mathbb{R}^L$ , は、特異値ベクトルである。これらを列ベクトルとする行列  $U_{k_1} \in \mathbb{R}^{N \times N}$ ,  $U_{k_2} \in \mathbb{R}^{M \times M}$ ,  $U_{k_3} \in \mathbb{R}^{L \times L}$  は直交行列である。

次に、注目するサンプル依存性をもっている  $\mathbf{u}_{k_2}$ ,  $\mathbf{u}_{k_3}$  を選ぶ。これらの集合をそれぞれ  $\Omega_{k_2}$ ,  $\Omega_{k_3}$  とする。そして  $k_2, k_3$  がこの集合に属する  $G(k_1, k_2, k_3)$  の中で、絶対値が大きい  $G(k_1, k_2, k_3)$  をもつ  $k_1$  を選択する。この集合を  $\Omega_{k_1}$  とする。ここで  $\mathbf{u}_{k_1}, k_1 \in \Omega_{k_1}$  が多重ガウス分布に従っていると仮定して (帰無仮説)、 $i$  番目の遺伝子に棄却確率  $P_i$  を

$$P_i = P_{\chi^2} \left[ > \sum_{k_1 \in \Omega_{k_1}} \left( \frac{u_{k_1 i}}{\sigma_{k_1}} \right)^2 \right] \quad (2)$$

で付与する。 $\sigma_{k_1}$  は  $u_{k_1 i}$  の標準偏差である。補正  $P_i$  値が 0.01 以下の遺伝子を所与のサンプル依存性を有意に持っている遺伝子として選択する。

### 3. 結果

図 1 に今回の解析の概要を示す。詳細は以下のとおりである。

#### 3.1 メチル化

図 2 はメチル化に対する主成分分析の結果である。第一

表 1 主成分分析を用いた教師なし学習による変数選択をメチル化に適用して選ばれた遺伝子 (図 1 赤字) のワーカーと女王の間の差の検定 (P 値)、 $t$ :  $t$  検定、Wilcox: ウィルコクソンの順位和検定、 $KS$ : コルモゴロフシナイ検定。

Table 1  $P$ -values computed by several statistical tests for the distinction of methylation between queen and worker of genes selected by applying PCA based unsupervised FE to methylation (red colored numbers in Fig. 1).  $t$ :  $t$  test, Wilcox: Wilcoxon rank sum test, KS: Kolmogorov-Sinai test.

	$t$	Wilcox	KS
<i>P. canadensis</i>	$8.59 \times 10^{-5}$	$3.10 \times 10^{-3}$	$1.83 \times 10^{-4}$
<i>D. quadriceps</i>	$1.11 \times 10^{-2}$	$5.88 \times 10^{-3}$	$1.75 \times 10^{-3}$

主成分負荷量が *P. canadensis*, *D. quadriceps* 共にコントロールと女王、ワーカーの間で大きく異なっていることが解る。これに対応する第一主成分得点を使って、(1) 式による  $P$  値を多重比較補正して 0.01 以下の遺伝子を選んだ (選択遺伝子数はそれぞれ、241 個と 138 個、図 1 の赤字参照)。また、図 2 (A) では、第一主成分負荷量の値が女王とワーカー > コントロールとなっているのに対して、選択された遺伝子の第一主成分得点がすべて正であることから *P. canadensis* ではメチル化されている遺伝子だけが選ばれていることが解る。同様のことが *D. quadriceps* にも言え、図 2 (B) では、第一主成分負荷量の値が女王とワーカー < コントロールとなっているのに対して、選択された遺伝子の第一主成分得点がすべて負であるのでメチル化されている遺伝子だけが選ばれている。このメチル化はメチル化の増大が主であるという観測は Patalano ら [5] の報告とも一致している。

一方、Patalano ら [5] は女王とワーカーでのメチル化の差異を観測していない。図 2 でも、女王とワーカーの差はわずかだが、差が見えていることは間違いない。そこで、個々の遺伝子のメチル化ではなく、選択された遺伝子全体のメチル化の平均値が女王とワーカーで差があるかどうかを検定で確認した (表 1)。選ばれた遺伝子のメチル化が女王とワーカーで差があることは明らかであり、主成分分析を用いた教師なし学習は、原報 [5] では不可能だった、ワーカーと女王で発現差がある遺伝子群を同定することに成功した。

#### 3.2 mRNA 発現プロファイル

図 3 は mRNA 発現プロファイルに対する主成分分析の結果である。第三、四主成分負荷量が *P. canadensis*, *D. quadriceps* でそれぞれ女王とワーカーの間で大きく異なっていることが解る。これに対応する第三、四主成分得点を使って、(1) 式による  $P$  値を多重比較補正して 0.01 以下の遺伝子を選んだ (選択遺伝子数はそれぞれ、41 個と 145 個、図 1 の青字参照)。

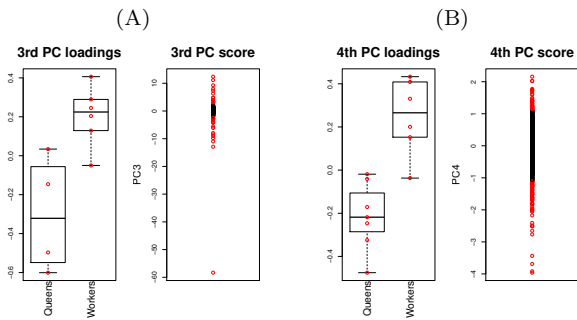


図 3 主成分負荷量のカースト依存性(左)と得点の分布(右)。(A) *P. canadensis*: 第三主成分 (B) *D. quadriceps*: 第四種成分  
**Fig. 3** The dependence of PC loading upon caste(left) and the histogram of PC scores(right) (A) *P. canadensis*:3rd PC (B) *D. quadriceps*:4th PC

表 2 主成分分析を用いた教師なし学習による変数選択を mRNA 発現量に適用して選ばれた遺伝子(図 1 青字)のワーカーと女王の間の差の検定(P値)、*t*:*t*検定、Wilcox: ウィルコクソンの順位和検定、*KS*: コルモゴロフシナイ検定。

**Table 2** *P*-values computed by several statistical tests for the distinction of mRNA expression between queen and worker of genes selected by applying PCA based unsupervised FE to mRNA expression (blue colored numbers in 図 1). *t*: *t* test, Wilcox: Wilcoxon rank sum test, *KS*: Kolmogorov-Sinai test.

	<i>t</i>	Wilcox	<i>KS</i>
<i>P. canadensis</i>	$4.37 \times 10^{-4}$	0.07	$6.45 \times 10^{-2}$
<i>D. quadriceps</i>	$1.73 \times 10^{-12}$	$2.24 \times 10^{-12}$	$5.26 \times 10^{-1}$

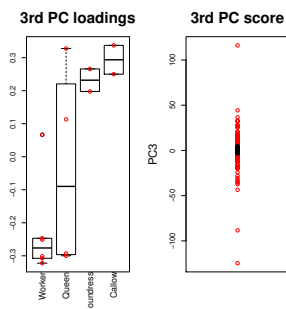


図 4 第三主成分負荷量のカースト依存性(左)と得点の分布(右)  
**Fig. 4** The dependence of the third PC loading upon caste(left) and the histogram of PC scores(right)

選択された遺伝子全体の mRNA 発現量の平均値が女王とワーカーで差があるかどうかを検定で確認した(表 2)。選ばれた遺伝子の mRNA 発現量が女王とワーカーで差があることは明らかである。

### 3.3 mRNA 発現プロファイル: その 2

次に、比較のため、Ferreira ら [6] の *P. canadensis* の mRNA 発現プロファイルにも主成分分析を用いた教師なし学習による変数選択を適用した。図 4 は Ferreira ら [6] の mRNA 発現プロファイルに対する主成分分析の結果であ

る。第三主成分負荷量が 4 種のカースト間で大きく異なっていることが解る。これに対応する第三主成分得点を使って、(1) 式による *P* 値を多重比較補正して 0.01 以下の遺伝子を選んだ(選択遺伝子数は 1 2 0、図 1 の青字参照)。Ferreira ら [6] はペアワイズな二群の比較で遺伝子を選択したため、ペアごとに選択遺伝子が異なってしまったが、我々は共通した遺伝子を選択できた。

選択された遺伝子全体の mRNA 発現量の各カーストの平均値が他のカーストと差があるかどうかを検定で確認した(表 3)。ここで遺伝子は第三種成分得点が正のもの 7 7 遺伝子と負のもの 4 3 遺伝子に分けて解析された。その結果、新女王カーストで発現が上昇している遺伝子以外はすべて、何からの検定で各カーストとその他の間で発現の有意差があることが確認された。

### 3.4 テンソル分解を用いた教師なし学習による変数選択

主成分分析を用いた教師なし学習による変数選択はメチル化、mRNA 発現プロファイル共に、カースト間で差がある遺伝子を見つけることに成功した。特にメチル化についてはカースト間で差がある遺伝子は未発見だったので、差がある遺伝子を発見できたのは非常に有意義だった。

しかし、この研究の本来の目的なカースト形成に対するエピジェネティックスの効果の研究である。ここで選択されたメチル化遺伝子が発現に差をもたらしているかを調べるために、メチル化で選択された遺伝子と mRNA 発現プロファイルで選択された遺伝子に統計的に有意な重なりがあるかを調べたが、そもそもオッズ比が 1 以下になってしまったので、メチル化と mRNA 発現プロファイルで女王とワーカーで同じように差がある遺伝子を選べていないことになる。これでは、メチル化と mRNA 発現プロファイル両方に影響を与えている遺伝子を調べたい、という目的に達していない。

そこで我々は Patalano ら [5] のメチル化と mRNA 発現プロファイルにテンソル分解を用いた教師なし学習による変数選択を適用してみた。

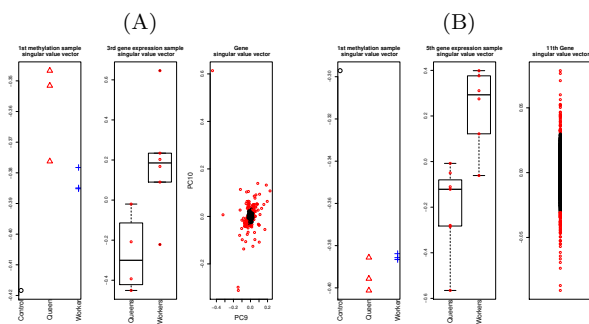
図 5 はテンソル分解の結果である。メチル化の特異値ベクトルについては両種とも第一特異値ベクトル  $u_{1l}$  がカースト依存性を反映していた。mRNA 発現プロファイルについては *P. canadensis* では第三特異値ベクトル  $u_{3j}$  が、*D. quadriceps* では第五特異値ベクトル  $u_{5j}$  が、それぞれワーカーと女王の差を反映していた。これらの特異値ベクトルと絶対値の大きなコアテンソルを共有している遺伝子特異値ベクトルを調べたところ、*P. canadensis* では第 9、10 特異値ベクトル ( $u_{9i}, u_{10i}$ ) が、*D. quadriceps* では第 11 特異値ベクトル ( $u_{11i}$ ) が、それぞれ選択された。これらを用いて、(2) 式を使って、*P* 値を遺伝子に付与し、多重比較補正を行って 0.01 以下の *P* 値を伴った遺伝子を選択したところ、それぞれ 1 3 3 遺伝子、1 2 8 遺伝子が選

**表 3** 主成分分析を用いた教師なし学習による変数選択を mRNA 発現量に適用して選ばれた遺伝子 (図 1 青字) の各カーストとその他の間の差の検定 (P 値)、 $t:t$  検定、Wilcoxon: ウィルコクソンの順位検定、KS: コルモゴロフシナイ検定。

**Table 3** P-values computed by several statistical tests for the distinction of mRNA expression between individual caste and others of genes selected by applying PCA based unsupervised FE to mRNA expression (blue colored numbers in 図 1).  $t:t$  test, Wilcoxon: Wilcoxon rank sum test, KS: Kolmogorov-Sinai test.

第三主成分得点			$t$	Wilcoxon	KS
ワーカー	+	下降	$3.96 \times 10^{-5}$	*	*
対その他	-	上昇	0.22	*	*
女王	+	上昇	0.36	0.10	$8,87 \times 10^{-4}$
対その他	-	下降	0.32	0.69	$4.69 \times 10^{-3}$
新女王	+	上昇	0.05	0.13	0.17
対その他	-	下降	0.11	$3.49 \times 10^{-10}$	$6.27 \times 10^{-14}$
若年ワーカー	+	上昇	0.99	*	*
対その他	-	下降	$1.14 \times 10^{-6}$	$2.7 \times 10^{-7}$	$3.34 \times 10^{-6}$

\* :  $< 2.2 \times 10^{-16}$



**図 5** 特異値ベクトルのカースト依存性 (メチル化  $u_{k_3\ell}$ : 左、mRNA 発現量  $u_{k_2j}$ : 中) と遺伝子依存性 ( $u_{k_1i}$  右)。(A) *P. canadensis* (B) *D. quadriceps*

**Fig. 5** Singular value vectors, dependence upon castes (methylation,  $u_{k_3\ell}$ : left; mRNA expression,  $u_{k_2j}$ : middle), dependence upon genes ( $u_{k_1i}$ : right) (A) *P. canadensis* (B) *D. quadriceps*

扱された (図 1 の黒字参照)。

これらの遺伝子がメチル化と mRNA 発現プロファイルの両方でカースト間差異を有意に持っているかを検討した (表 4)。残念ながら、*D. quadriceps* のメチル化だけは有意差が無かったが、他は有意差があった。このことは *P. canadensis* については mRNA 発現プロファイルとメチル化が同時に女王とワーカーで有意に変化している遺伝子を発見できたことを意味している。

### 3.5 GO エンリッチメント解析

mRNA 発現プロファイルで選ばれた遺伝子の GO エンリッチメント解析を実行した (図 1、表 5)。*D. quadriceps* に対するテンソル分解を用いた教師なし学習による変数選択の結果を除いては、先行研究と同等、あるいはそれ以上の GO エンリッチメント解析の結果が得られている。*D. quadriceps* に対するテンソル分解を用いた教師なし学習に

よる変数選択は表 4 に見るように、mRNA 発現プロファイルとメチル化で同時にワーカーと女王で差がある遺伝子を選択できていないことからみても、まだ、改善の余地があるものと思われる。

## 4. おわりに

本研究では、社会性昆虫の脳の mRNA 発現プロファイルとメチル化の解析に主成分分析を用いた教師なし学習による変数選択を用いた。その結果、先行研究では見出せなかった、カースト間でメチル化が異なっている遺伝子群を見いだせた。また、テンソル分解を用いた教師なし学習による変数選択を mRNA 発現プロファイルとメチル化の統合解析に用いたところ、*P. canadensis* については mRNA 発現プロファイルとメチル化が同時に女王とワーカーで有意に変化している遺伝子を発見できた。さらに選ばれた遺伝子の GO エンリッチメント解析の結果から生物学的に意味がある遺伝子を選択できていることも確認できた。主成分分析およびテンソル分解を用いた教師なし学習による変数選択は有用な解析手法である。

**謝辞** 本研究の原著論文は科研費基盤研究 (C) 17K00417 の助成を受けて行われた。

## 参考文献

- [1] Taguchi, Y.-H.: Tensor decomposition-based and principal-component-analysis-based unsupervised feature extraction applied to the gene expression and methylation profiles in the brains of social insects with multiple castes, *BMC Bioinformatics*, Vol. 19, No. 4, p. 99 (online), DOI: 10.1186/s12859-018-2068-7 (2018).
- [2] 長濱嘉隆孝, 小林亨, 松田勝: 魚類の性決定と生殖腺の性分化/性転換, 蛋白質 酵素 核酸, Vol. 49, No. 2, pp. 116–123 (2004).
- [3] 三浦徹: 社会性昆虫におけるカースト分化の分子・生理的基盤, Vol. 49, pp. 167–174 (1999).
- [4] Weiner, S. A. and Toth, A. L.: Epigenetics in Social In-

表 4 テンソル分解を用いた教師なし学習による変数選択を mRNA 発現プロファイルとメチル化に適用して選ばれた遺伝子 (図 1 黒字) のワーカーと女王の間の差の検定 (P 値)、 $t$ :  $t$  検定、Wilcox: ウィルコクソンの順位和検定、KS: コルモゴロフシナイ検定。

Table 4 P-values computed by several statistical tests for the distinction of methylation and mRNA expression between queen and worker of genes selected by applying TD based unsupervised FE to mRNA expression and methylation (black colored numbers in 図 1).  $t$ :  $t$  test, Wilcox: Wilcoxon rank sum test, KS: Kolmogorov-Sinai test.

		$t$	Wilcox	KS
<i>P. canadensis</i>	mRNA 発現量	$1.71 \times 10^{-3}$	$1.89 \times 10^{-2}$	0.08
	メチル化	$1.74 \times 10^{-4}$	$5.06 \times 10^{-3}$	$1.02 \times 10^{-3}$
<i>D. quadricaps</i>	mRNA 発現量	$2.73 \times 10^{-12}$	$9.05 \times 10^{-12}$	$4.41 \times 10^{-11}$
	メチル化	0.3757	0.7163	0.4413

表 5 mRNA 発現プロファイルで選ばれた遺伝子の GO エンリッチメント解析

Table 5 GO enrichment analysis of the genes selected using gene expression profiles.

<i>P. canadensis</i>	
Patalano ら [5] のデータセット	
本研究の結果 (主成分分析)	
GO:0005319	lipid transporter activity
GO:0006869	lipid transport
本研究の結果 (テンソル分解)	
GO:0005319	lipid transporter activity
GO:0005811	lipid particle
GO:0006869	lipid transport
Patalano ら [5] の結果 エンリッチメント検出無し	
Ferreira ら [6] のデータセット	
本研究の結果 (主成分分析)	
GO:0004129	cytochrome-c oxidase activity
GO:0003735	structural constituent of ribosome
GO:0006412	translation
GO:0005743	mitochondrial inner membrane
GO:0008137	NADH dehydrogenase (ubiquinone) activity
<i>D. quadricaps</i>	
Patalano ら [5] のデータセット	
本研究の結果 (主成分分析)	
GO:0005506	iron ion binding
GO:0009055	electron carrier activity
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0020037	heme binding
GO:0055114	oxidation-reduction process
本研究の結果 (テンソル分解) エンリッチメント検出無し	
Patalano ら [5] の結果	
GO:0003735	structural constituent of ribosome
GO:0005622	intracellular
GO:0005840	ribosome
GO:0005842	cytosolic large ribosomal subunit
GO:0006412	translation

- sects: A New Direction for Understanding the Evolution of Castes, *Genetics Research International*, Vol. 2012, pp. 1–11 (online), DOI: 10.1155/2012/609810 (2012).
- [5] Patalano, S., Vlasova, A., Wyatt, C., Ewels, P., Camara, F., Ferreira, P. G., Asher, C. L., Jurkowski, T. P., Segonds-Pichon, A., Bachman, M., González-Navarrete, I., Minoche, A. E., Krueger, F., Lowy, E., Marcet-Houben, M., Rodriguez-Ales, J. L., Nascimento, F. S., Balasubramanian, S., Gabaldon, T., Tarver, J. E., Andrews, S., Himmelbauer, H., Hughes, W. O. H., Guigó, R., Reik, W. and Sumner, S.: Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies, *Proceedings of the National Academy of Sciences*, Vol. 112, No. 45, pp. 13970–13975 (online), DOI: 10.1073/pnas.1515937112 (2015).
- [6] Ferreira, P. G., Patalano, S., Chauhan, R., Ffrench-Constant, R., Gabaldón, T., Guigó, R. and Sumner, S.: Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes, *Genome Biology*, Vol. 14, No. 2, p. R20 (online), DOI: 10.1186/gb-2013-14-2-r20 (2013).
- [7] Taguchi, Y.: *Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach*, SpringerNature (2020).
- [8] Clough, E. and Barrett, T.: *The Gene Expression Omnibus Database*, pp. 93–110 (online), DOI: 10.1007/978-1-4939-3578-9\_5, Springer New York (2016).
- [9] Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1, pp. 289–300 (online), available from <http://www.jstor.org/stable/2346101> (1995).
- [10] 石黒勝彦, 林 浩平: 関係データ学習 (機械学習プロフェッショナルシリーズ), 講談社 (2016).