

- testicular steroidogenesis: A possible estrogenic effect of di(n-butyl) phthalate. *Reproduction* 139, 427-437, (2010).
7. Ohsako S, Fukuzawa N, Ishimura R, Kawakami T, Wu Q, Nagano R, Zaha H, Sone H, Yonemoto J, and Tohyama C. Comparative contribution of the aryl hydrocarbon receptor gene to perinatal stage development and dioxin-induced toxicity between the urogenital complex and testis in the mouse. *Biol Reprod* 82, 636-643, (2010).
 8. Ishihara K, Ohsako S, Tasaka K, Harayama H, Miyake M, Warita K, Tanida T, Mitsuhashi T, Nanmori T, Tabuchi Y, Yokoyama T, Kitagawa H, and Hoshi N. When does the sex ratio of offspring of the paternal 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) exposure decrease: In the spermatozoa stage or at fertilization? *Reprod Toxicol* 29, 68-73, (2010).
 9. Alam MS, Ohsako S, Tay TW, Tsunegawa N, Kanai Y, and Kurohmaru M. Di(n-butyl) phthalate induces vimentin filaments disruption in rat Sertoli cells: A possible relation with spermatogenic cell apoptosis. *Anatomia Histologia Embryologia* 39, 189-193, (2010).
 10. Kawakami T, Ito T, Ohsako S, Shiizaki Y, Murakami Y, Hirowatari K, Sato M, and Tohyama C. Possible Involvement of arylhydrocarbon receptor variants in TCDD-induced thymic atrophy and XRE-dependent transcriptional activity in Wistar Hannover GALAS rats. *J Toxicol Sci* 34, 209-220, (2009).
 11. Ishimura R, Kawakami T, Ohsako S, and Tohyama C. Dioxin-induced toxicity on vascular remodeling of the placenta. *Biochemical Pharmacol* 77, 660-669, (2009).
 12. 大迫誠一郎、エピジェネティクスと環境医学、分子予防環境医学、本の泉社、568-575、(2010).
 13. 大迫誠一郎、プログラムされる“病”の新たな仮説-環境化学物質による代謝系遺伝子の次世代エピゲノム変化、科学、岩波書店、984-989、(2009).
- ## 2. 学会発表
1. 大迫誠一郎, 山根順子, 今西哲, 遠山千春. ヒト ES 細胞を用いた神経系誘導培養系における Ahr アゴニストの影響. 第 11 回日本再生医療学会総会, 横浜 (2012) 6 月 12 日
 2. Seiichiroh Ohsako, Junko Yamane, Satoshi Imanishi, Chiharu Tohyama Effects of Ahr agonist on neuronal cell differentiation from human ES cells. The 10th Annual Meeting of International Society for Stem Cell Research, June 2012, Yokohama, Japan
 3. 大迫誠一郎, 永野麗子, 何小明, 今西哲, 赤沼宏美, 山根順子, 藤渕航, 曾根秀子. 胚性幹細胞試験を用いたメチル水銀の神経発生毒性の影響評価. 第 82 回日本衛生学会、京都 (2012) 3 月
 4. 大迫誠一郎. 環境汚染化学物質の周産期曝露による表現型変化—エピジェネティクスと環境毒性学—. 日本獣医学会、さいたま (2012) 3 月 28 日大宮ソニックシティ
 5. 大迫誠一郎. 環境汚染物質の胎生期曝露による生後の化学発癌感受性亢進とエピゲノム変化. シンポジウム「環境に起因する疾患エピゲノム変化」日本人類遺伝学会、千葉 (2011) 11 月 10 日幕張メッセ
 6. 大迫誠一郎, 永野麗子, 何小明, 今西哲, 藤渕航, 赤沼宏美, 曾根秀子. マウスおよびヒト ES 細胞を用いた神経分化培養系におけるメチル水銀の毒性影響評価. 第 13 回日本外

- 因性内分泌攪乱化学物質学会、東京（2010）
12月16日東京大学山上会館
7. 赤沼宏美, 永野麗子, 座波ひろ子, 大迫誠一郎, 曾根秀子. 胎生プログラミング異常を検出するためのマルチプロファイリング技術の確立. 第13回日本外因性内分泌攪乱化学物質学会、東京（2010）12月16日東京大学山上会館
 8. 曾根秀子, 永野麗子, 赤沼宏美, 座波ひろ子, 大迫誠一郎. ヒトES細胞の神経分化系を用いたサリドマイドの影響についての研究. 第13回日本外因性内分泌攪乱化学物質学会、東京（2010）12月16日
 9. 永野麗子, 何小明, 横山雅美, 赤沼宏美, 座波ひろ子, 末盛博文, 大迫誠一郎, 曾根秀子. マウスおよびヒトES細胞の神経分化系を用いたサリドマイドの影響についての研究. 第36回日本トキシコロジー学会、沖縄（2010）6月17日沖縄コンベンションセンター
 10. 何小明, 永野麗子, 赤沼宏美, 座波ひろ子, 遠山千春, 曾根秀子, 大迫誠一郎. マウスES細胞を用いた神経系分化系におけるメチル水銀の毒性影響評価. 第36回日本トキシコロジー学会、沖縄（2010）6月17日沖縄コンベンションセンター
 11. Hideko Sone, Hiromi Akanuma, Reiko Nagano, Seichiroh Ohsako. Multi-profiling analysis in neuronal cells derived from embryonic stem cells to identify fetal programming. The 12th International Congress of Toxicology. 2010
 12. 赤沼宏美, 永野麗子, 座波ひろ子, 大迫誠一郎, 曾根秀子. 胎生プログラミング異常を検出するためのマルチプロファイリング技術の確立. 第13回日本外因性内分泌攪乱化学物質学会、東京（2010）12月16日東京大学山上会館
 13. 曾根秀子, 永野麗子, 赤沼宏美, 座波ひろ子, 大迫誠一郎. ヒトES細胞の神経分化系を用いたサリドマイドの影響についての研究. 第13回日本外因性内分泌攪乱化学物質学会、東京（2010）12月16日
 14. 大迫誠一郎, 永野麗子, 何小明, 今西哲, 藤渕航, 赤沼宏美, 曾根秀子. マウスおよびヒトES細胞を用いた神経分化培養系におけるメチル水銀の毒性影響評価. 第13回日本外因性内分泌攪乱化学物質学会、東京（2010）12月16日東京大学山上会館
- #### H. 知的財産権の出願・登録状況
1. 特許取得
 1. 曾根秀子, 大迫誠一郎, 永野麗子, 今西哲, 赤沼宏美, 宮崎航. 「胎生プログラミングに対する影響を評価するための方法」. 特願 2009-81497（識別番号100078662）（2009）.
 2. 大迫誠一郎, 栗田尚佳. 「CpGメチル化頻度の変動をゲノムワイドに比較解析するためのDNA試料作成方法」. 米国仮出願（出願番号：61/309971）（2010）.
 2. 実用新案登録
なし
 3. その他
なし

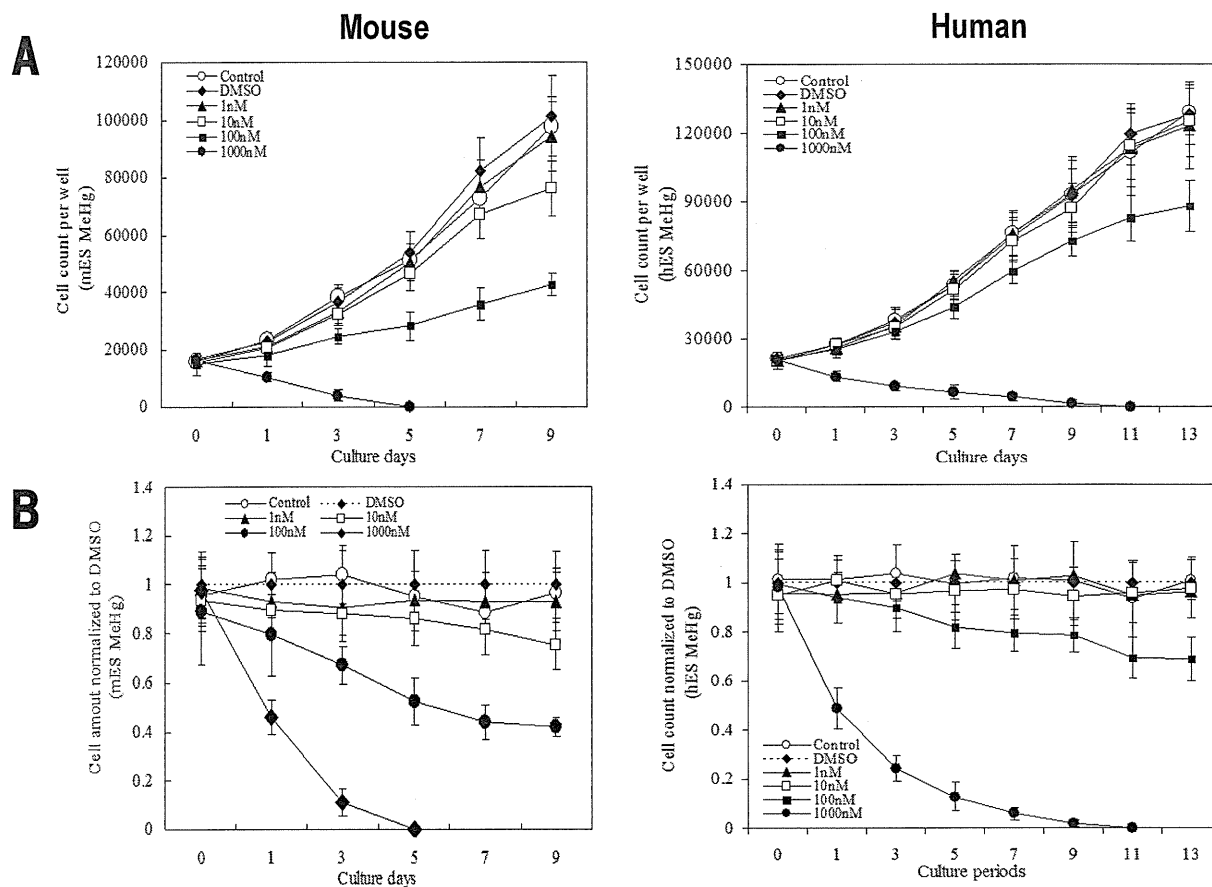


図1. マウスES細胞およびヒトES細胞の神経系細胞培養系に及ぼすMeHgの影響。MTTアッセイによって用量依存的な細胞活性の測定を行った。MeHg濃度：1, 10, 100, 1000 nM。A, 培養時間によるウェル内細胞数の推移。B, 培養時間による細胞活性の推移を細胞数当たり換算して計算した。各時間後とのDMSO対照群の活性を1とした。

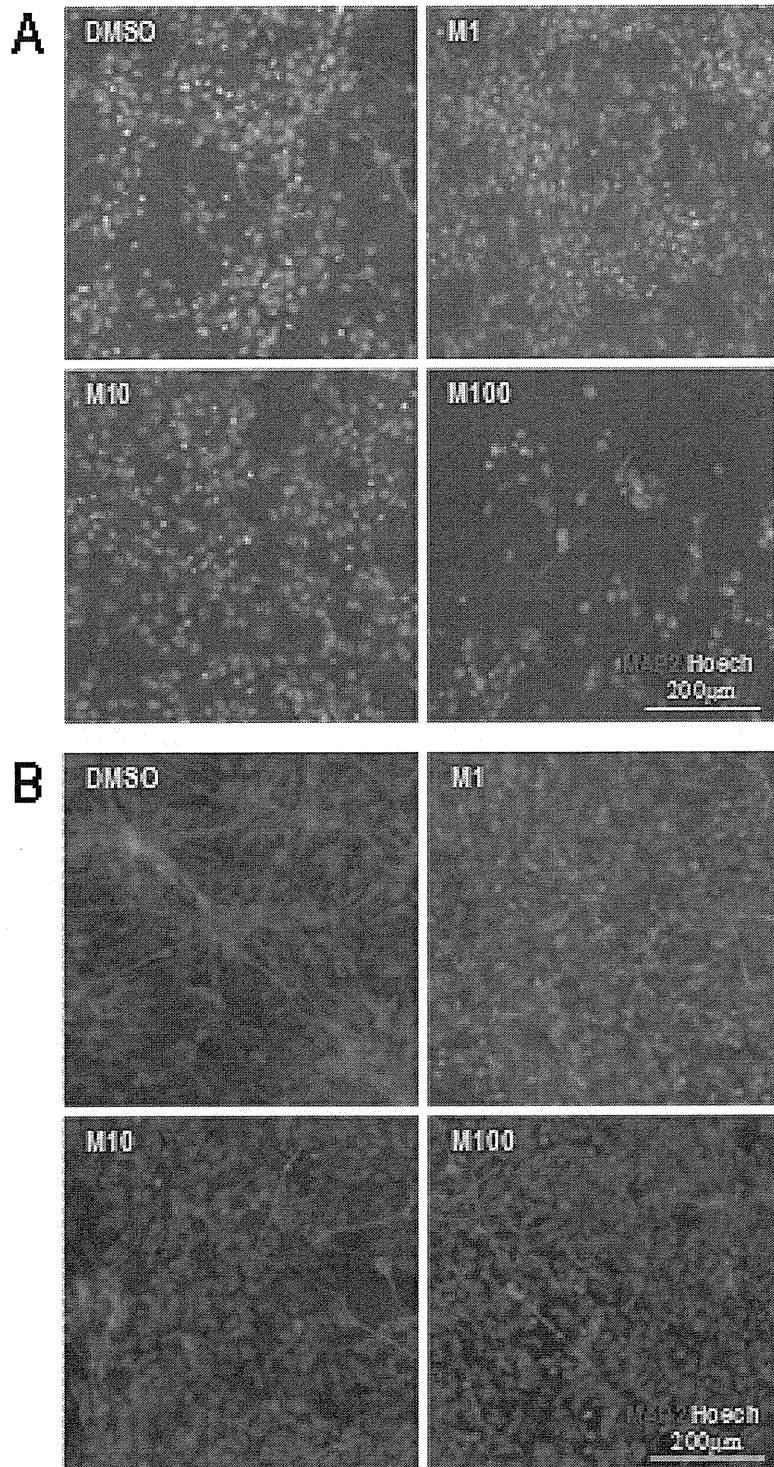


図2. マウスES細胞(A)およびヒトES細胞(B)の神経系細胞培養系に及ぼすMeHgの影響。DMSO, 対照; M1, 1 nM MeHg; M10, 10 nM MeHg; M100, 100 nM MeHg。

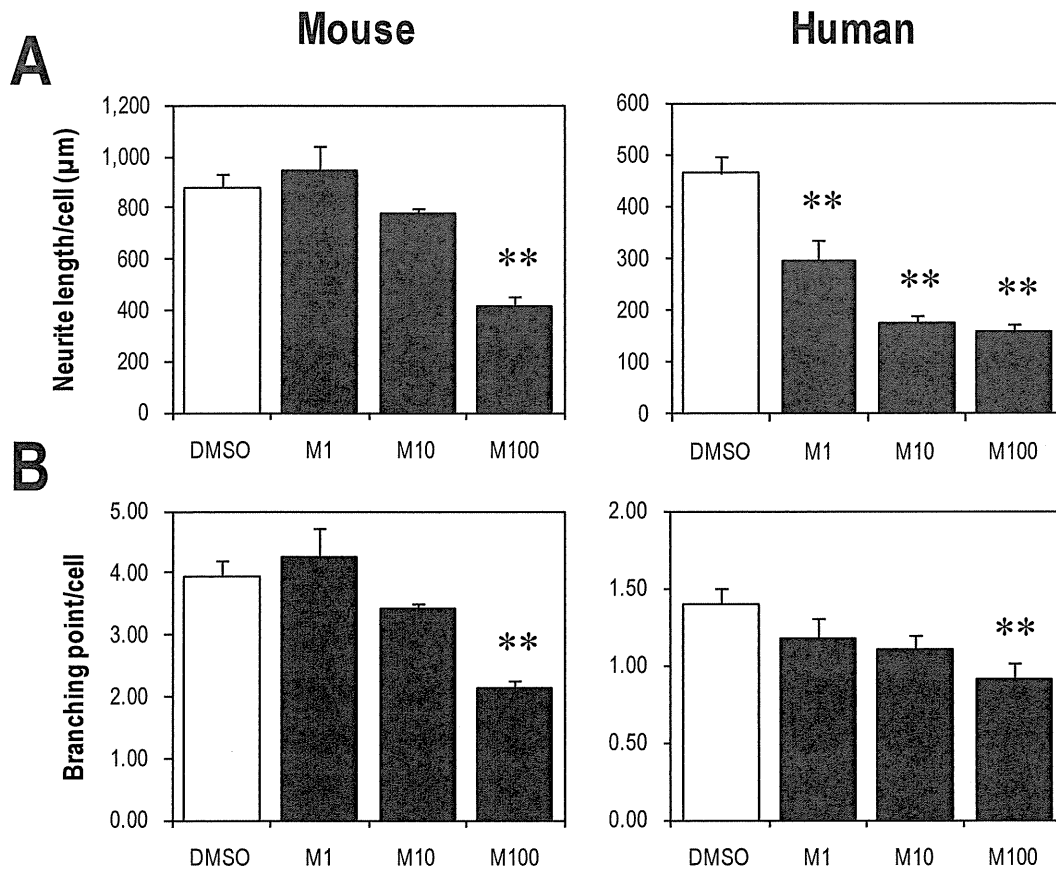


図3. マウスES細胞およびヒトES細胞の神経系細胞培養系に及ぼすMeHgの影響。全自動細胞画像解析装置でMAP2陽性細胞の神経突起長(Neurite length, A)と分岐点数(Branching points, B)の測定を行った。DMSO, 対照; M1, 1 nM MeHg; M10, 10 nM MeHg; M100, 100 nM MeHg。DMSO対照群との比較(*, $p < 0.05$; **, $p < 0.01$)。

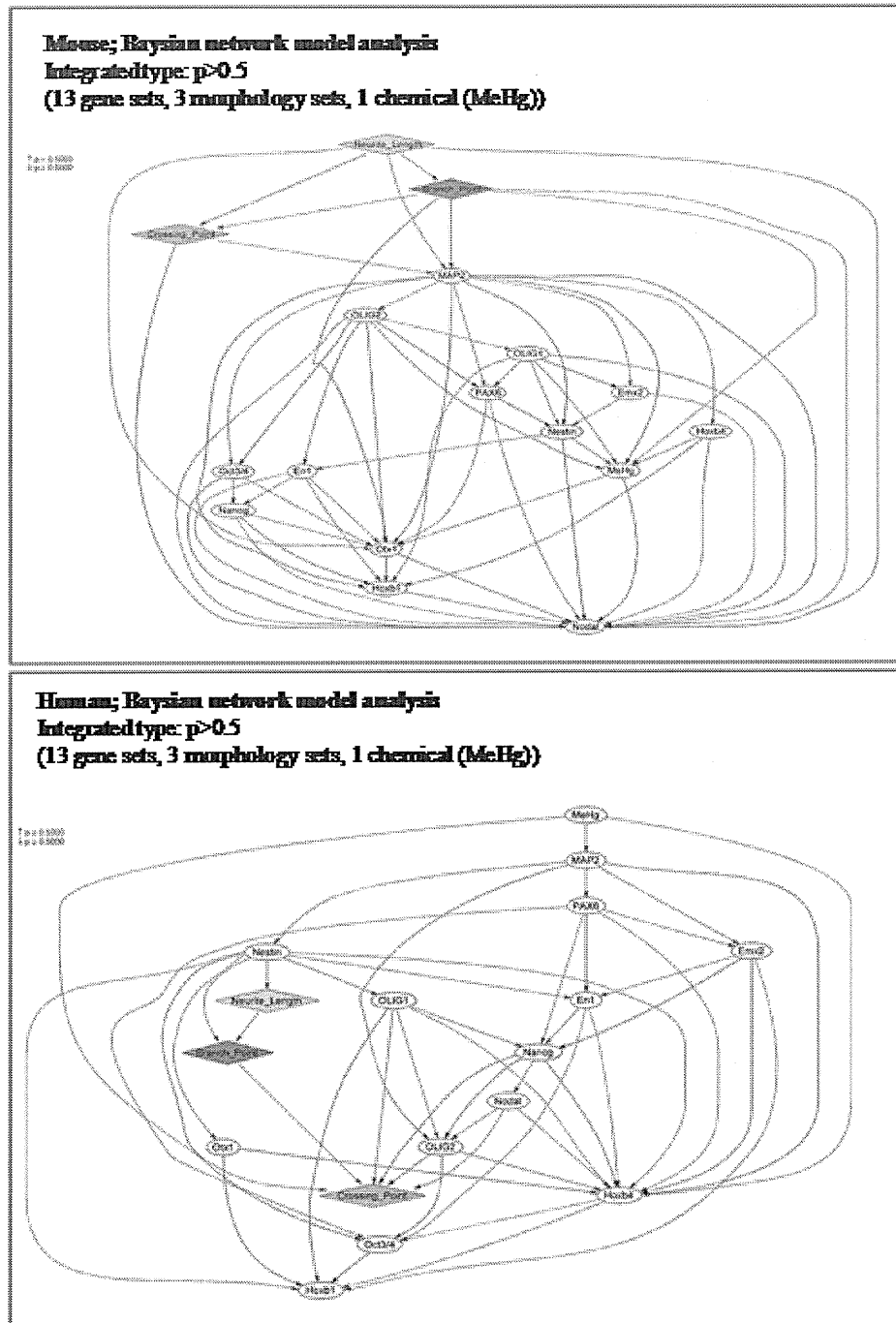


図4. マウスES細胞およびヒトES細胞の神経系細胞培養系に及ぼすMeHgの影響の確率推論アルゴリズム(ベイジアンネットワークモデル)に基づく統合化ネットワーク解析結果。マウス細胞培養系のDay 23ならびにヒトES細胞培養系のDay50における13種の遺伝子発現データと3種の形態情報(Neurite length, Branching points, Crossing points)ならびにMeHgとの関係をコンピュータ解析し自動的に図式化させた。ベイズモデルにおけるベータ値が正の関係を赤い矢印、負の関係を青い矢印で示した。相互作用確率で $p > 0.5$ のものを抽出した。

ヒト ES 細胞の培養ステージの違いにダイオキシン（TCDD）曝露の晩発的影響

研究分担者

大迫誠一郎

東京大学 疾患生命工学センター 准教授

研究要旨

アリールハイドロカーボン受容体（Ahr）は多環芳香族炭化水素やダイオキシンなどの体外異物に反応する核内受容体であり、個体発生過程において活性化が起きると催奇形性等の様々な生体影響を引き起こす。ここではヒト ES 細胞を用いた神経系細胞発生過程においてダイオキシンによる Ahr の活性化の影響を調べた。

ヒト ES 細胞（KhES-1）を MEF を用いて増殖させ EB 形成させた。ES 細胞のプレート播種直後（Day0）から 24 時間、2,3,7,8-tetrachlorodibenzo-p-dioxin（TCDD、0.1 nM-10 nM）に曝露した。また、Day8 で神経誘導培地に変更し Day9 で TCDD に曝露した群も設けた。Day10 でオルニチンラミニンコートプレートに播種、Day12 から神経増殖培地に変え、Day28 で再播種し、Day40 まで培養して神経系細胞集団を分化させた。対照群の細胞では Day35 での TCDD 曝露も行った。

Ahr 活性化のバイオマーカーである CYP1A1 の誘導は Day9 曝露のみで認められ、Day0 曝露群と Day35 曝露群での反応性は認められなかった。Day40 まで培養した結果、どの曝露群においても神経系の十分な分化が認められた。Oct3/4 や神経系分化マーカーである Map2 の発現にも TCDD 曝露 24 時間目ならびに Day40 での TCDD による変動は認められなかった。Day9 曝露群を Day25 まで神経誘導培養した結果、ニューラルロゼッタの形成率が対照群や他の処理群より上昇しており、NESTIN、MTAP2 のレベルが高く、SOX17、FOXA2 の低下が著しいことがわかった。

今回の結果は、EB 形成後期の神経系誘導条件では TCDD に対する反応性が他のステージより高く、またその結果として内胚葉系の細胞の分化を阻害し、外胚葉系の細胞の分化率を上昇させることを示唆している。

A. 研究目的

環境汚染化学物質のなかでもダイオキシン類は、急性毒性として皮膚の塩素ザソウや肝機能異常による消耗性症候群とそれに引き続く死亡を伴う、極めて毒性の強い化合物である。特に胎児期曝露の場合では数ナノグラムの単回摂取でも生まれてくる個体の不可逆的变化、すなわち生殖次世代影響として種々の影響を示すことが知られている。これら生殖発生学的影響（催奇形性・胎仔死亡・生殖機能異常・行動異常等）は、高感受性と言う点からリスク

管理上注視すべき指標である。しかし各指標とも発生学的な病態発生の分子機構は未知な部分が多い。上記の現象はダイオキシン受容体であるアリールハイドロカーボン受容体（Ahr）のノックアウトマウスで見られないことから、Ahr がプライマリーの標的分子であることは間違いない。Ahr は多環芳香族炭化水素やダイオキシンなどの体外異物に反応する核内受容体であり、生体防御に必須で有り、正常な個体発生過程において不可欠な遺伝子である。ここでは、我々の保持するヒト ES 細胞を用いた神

経系細胞発生過程においてダイオキシンによる Ahr の活性化の影響を調べた。培養系の種々の分化ステージで曝露し、十分に分化した細胞との比較を通じて個体発生過程のダイオキシンによる影響を、初めてヒトを対象として解析する試みを行った。

（倫理面への配慮）

東京大学ライフサイエンス委員会倫理審査専門委員会にて2009年12月機関承認を得、それにより、2010年1月文部科学省より使用許可を得た。

B. 研究方法

ヒトES細胞は京大再生医科学研究所幹細胞センターの末盛博文博士から提供されたKhES-1株 (XX genotype) を使用した。ES細胞の維持分化にはマウス胚性線維芽細胞 (MEF) をフィーダー細胞として継代を行った。胚様体 (EB) の形成にはMEF除去後にSUMILON PrimeSurfase 96U (住友ベークライト社製) を用いた (図1)。簡単には、MEFを酵素的に浮遊させ、ES細胞集塊のみをディッシュ上に残した上で、Rock阻害剤Y-27632を添加したEB培地に懸濁し、 9.0×10^3 個/well のhES細胞をSUMILON PrimeSurfase 96U に播種した (Day0) (図1)。この播種時点から24時間、2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD、0.1nM-10nM) を添加したEB培地でEB形成を行った。Day8で神経誘導培地 (NIM) に変更しDay9でTCDDに曝露した群も設けた。形成されたEBをDay10でオルニチンラミニン (O/L) コートプレートに播種、Day12から神経増殖培地 (NPM) に変え、Day28で再播種、Day40まで培養して神経系細胞集団を分化させた (図1)。対照群の細胞ではDay35でのTCDD曝露も行った (図1)。

各ステージ曝露後24時間目のRNAを回収し、*Pou5f1*, *Mtap2*, *Cyp1a1*, *Cyp1b1*, *Ahr*, *Actb* の mRNA を定量 RT-PCR で測定した。また培養28日目 (Day28) および最終段階 (Day40) のRNAを回収し、*Mtap2*, *Gfap*, *Nes*, *Th*, *Sox17*, *Foxa2*, *Pdgfr*, *Kdr*,

Ahr, *Actb* の mRNA を定量 RT-PCR で測定した。またMAP2抗体 (デンドライトマーカー) およびTH抗体 (ドパミン神経マーカー) による免疫染色とHoechst 33342による核染色を施し神経系細胞の分化状態を観察した。

C. 研究結果

Ahr 活性化のバイオマーカーであるCYP1A1の誘導はDay9曝露のみで認められ、Day0曝露群とDay35曝露群での反応性は認められなかった (図2)。Day40まで培養した結果、どの曝露群においても神経系の十分な分化が認められた。Day40までの培養で*Pou5f1* や神経系分化マーカーである*Mtap2* の発現にはDay0曝露群でもDay35曝露群でもTCDDによる変動は認められなかった (図2)。しかし、Day9曝露群をDay25まで神経誘導培養した段階でのニューラルロゼッタ (胚児の神経管と同義の構造体) の形成率が対照群や他の処理群より上昇していることが観察された (図3)。そこでDay28における培養系からRNAを採取し、神経細胞マーカーならびに胚葉系マーカーの発現レベルを見たところ、*Mtap2*, *Gfap*, *Th* のレベルが1nM以上のTCDD曝露群で高く、*Sox17*, *Foxa2*, *Pdgfr*, *Kdr* のレベルが1nM以上のTCDD曝露群で著しく低下していることがわかった (図4)。

D. 考察

今回の結果は、EB形成後期の神経系誘導条件ではTCDDに対する反応性が他のステージより高いことを示している。またその結果として内胚葉系ならびに中胚葉系の細胞の分化が阻害され、外胚葉系の細胞の分化率を上昇させることを示唆している。発生過程のヒト胚児では神経を形成する外胚葉系の細胞群より、むしろ中内胚葉系の細胞の感受性が高く、分化あるいは増殖の阻害が起きるものと考えられた。マウスやラットにおけるダイオキシン胎児期曝露試験では各種の脳発達影響が報告されている。ヒトにおける影響がいかなるものか定かでない

いが、仮に影響があると想定した場合、それは神経系細胞意外の他の細胞を構成する胚葉系の分化増殖に変化が生じた結果間接的に起きるものであると考えられる。

E. 結論

ヒト ES 細胞からの神経系細胞分化培養系を用いてダイオキシン (TCDD) の分化に及ぼす影響を初めて検討した。ヒトにおいて TCDD は神経系の細胞の分化増殖を阻害せず、むしろ相対的に分化率を増加させた。

F. 健康危険情報

なし。

G. 研究発表

1. 論文発表

1. He X, Imanishi S, Sone H, Nagano R, Qin XY, Yoshinaga J, Akanuma H, Yamane J, Fujibuchi W, and Ohsako S. Effects of Methylmercury Exposure on Neuronal Differentiation on Mouse and Human Embryonic Stem Cells. *Toxicol Lett* (2012) (in press)
2. Yoshioka W, Aida-Yasuoka K, Fujisawa N, Kawaguchi T, Ohsako S, Hara S, Uematsu S, Akira S, and Tohyama C. Critical role of mPGES-1 in the pathogenesis of hydronephrosis caused by lactational exposure of mice to dioxin. *Toxicol Sci* (2012) (in press).
3. Nagano R, Akanuma H, Qin XY, Imanishi S, Toyoshiba H, Yoshinaga J, Ohsako S, and Sone H. Multi-Parametric Profiling Network Based on Gene Expression and Phenotype Data: A Novel Approach to Developmental Neurotoxicity Testing. *Int J Mol Sci* 13(1), 187-207, (2012).
4. Ohsako S. Perinatal exposure to environmental chemicals induces epigenomic changes in offspring. *Genes Environ* 33, 43-49, (2011).
5. Sone H, Okura M, Zaha H, Fujibuchi W, Taniguchi T, Akanuma H, Nagano R, Ohsako S, and Yonemoto J. pCEC: a toxicogenomics database with a toxicoinformatics system for risk evaluation and toxicity prediction of environmental chemicals. *J Toxicol Sci* 35, 115-123, (2010).
6. Alam MS, Ohsako S, Matsuwaki T, Zhu XB, Tsunekawa N, Kanai Y, Sone H, Tohyama C, and Kurohmaru M. Induction of spermatogenic cell apoptosis in prepubertal rat testes irrespective of testicular steroidogenesis: A possible estrogenic effect of di(n-butyl) phthalate. *Reproduction* 139, 427-437, (2010).
7. Ohsako S, Fukuzawa N, Ishimura R, Kawakami T, Wu Q, Nagano R, Zaha H, Sone H, Yonemoto J, and Tohyama C. Comparative contribution of the aryl hydrocarbon receptor gene to perinatal stage development and dioxin-induced toxicity between the urogenital complex and testis in the mouse. *Biol Reprod* 82, 636-643, (2010).
8. Ishihara K, Ohsako S, Tasaka K, Harayama H, Miyake M, Warita K, Tanida T, Mitsuhashi T, Nanmori T, Tabuchi Y, Yokoyama T, Kitagawa H, and Hoshi N. When does the sex ratio of offspring of the paternal 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) exposure decrease: In the spermatozoa stage or at fertilization? *Reprod Toxicol* 29, 68-73, (2010).
9. Alam MS, Ohsako S, Tay TW, Tsunegawa N, Kanai Y, and Kurohmaru M. Di(n-butyl) phthalate induces vimentin filaments disruption in rat Sertoli cells: A possible relation with spermatogenic cell apoptosis. *Anatomia Histologia Embryologia* 39, 189-193, (2010).
10. Kawakami T, Ito T, Ohsako S, Shiizaki Y, Murakami Y, Hirowatari K, Sato M, and Tohyama C. Possible Involvement of arylhydrocarbon receptor variants in TCDD-induced thymic atrophy

- and XRE-dependent transcriptional activity in Wistar Hannover GALAS rats. *J Toxicol Sci* 34, 209-220, (2009).
11. Ishimura R, Kawakami T, Ohsako S, and Tohyama C. Dioxin-induced toxicity on vascular remodeling of the placenta. *Biochemical Pharmacol* 77, 660-669, (2009).
 12. 大迫誠一郎、エピジェネティクスと環境医学、分子予防環境医学、本の泉社、568-575、(2010).
 13. 大迫誠一郎、プログラムされる“病”の新たな仮説-環境化学物質による代謝系遺伝子の次世代エピゲノム変化、科学、岩波書店、984-989、(2009).
- ## 2. 学会発表
1. 大迫誠一郎、山根順子、今西哲、遠山千春. ヒト ES 細胞を用いた神経系誘導培養系における Ahr アゴニストの影響. 第 11 回日本再生医療学会総会、横浜（2012）6 月 12 日
 2. Seichiroh Ohsako, Junko Yamane, Satoshi Imanishi, Chiharu Tohyama Effects of Ahr agonist on neuronal cell differentiation from human ES cells. The 10th Annual Meeting of International Society for Stem Cell Research, June 2012, Yokohama, Japan
 3. 大迫誠一郎、永野麗子、何小明、今西哲、赤沼宏美、山根順子、藤渕航、曾根秀子. 胚性幹細胞試験を用いたメチル水銀の神経発生毒性の影響評価. 第 82 回日本衛生学会、京都（2012）3 月
 4. 大迫誠一郎. 環境汚染化学物質の周産期曝露による表現型変化—エピジェネティクスと環境毒性学—. 日本獣医学会、さいたま（2012）3 月 28 日大宮ソニックシティ
 5. 大迫誠一郎. 環境汚染物質の胎生期曝露による生後の化学発癌感受性亢進とエピゲノム変化. シンポジウム「環境に起因する疾患エピゲノム変化」日本人類遺伝学会、千葉（2011）11 月 10 日幕張メッセ
 6. 大迫誠一郎、永野麗子、何小明、今西哲、藤渕航、赤沼宏美、曾根秀子. マウスおよびヒト ES 細胞を用いた神経分化培養系におけるメチル水銀の毒性影響評価. 第 13 回日本外因性内分泌攪乱化学物質学会、東京（2010）12 月 16 日東京大学山上会館
 7. 赤沼宏美、永野麗子、座波ひろ子、大迫誠一郎、曾根秀子. 胎生プログラミング異常を検出するためのマルチプロファイリング技術の確立. 第 13 回日本外因性内分泌攪乱化学物質学会、東京（2010）12 月 16 日東京大学山上会館
 8. 曾根秀子、永野麗子、赤沼宏美、座波ひろ子、大迫誠一郎. ヒト ES 細胞の神経分化系を用いたサリドマイドの影響についての研究. 第 13 回日本外因性内分泌攪乱化学物質学会、東京（2010）12 月 16 日
 9. 永野麗子、何小明、横山雅美、赤沼宏美、座波ひろ子、末盛博文、大迫誠一郎、曾根秀子. マウスおよびヒト ES 細胞の神経分化系を用いたサリドマイドの影響についての研究. 第 36 回日本トキシコロジー学会、沖縄（2010）6 月 17 日沖縄コンベンションセンター
 10. 何小明、永野麗子、赤沼宏美、座波ひろ子、遠山千春、曾根秀子、大迫誠一郎. マウス ES 細胞を用いた神経系分化系におけるメチル水銀の毒性影響評価. 第 36 回日本トキシコロジー学会、沖縄（2010）6 月 17 日沖縄コンベンションセンター
 11. Hideko Sone, Hiromi Akanuma, Reiko Nagano, Seichiroh Ohsako. Multi-profiling analysis in neuronal cells derived from embryonic stem cells to identify fetal programming. The 12th International Congress of Toxicology. 2010
 12. 赤沼宏美、永野麗子、座波ひろ子、大迫誠一郎、曾根秀子. 胎生プログラミング異常を検出するためのマルチプロファイリング技術の確立. 第 13 回日本外因性内分泌攪乱化学物質学会、東

京（2010）12月16日東京大学山上会館

13. 曾根秀子, 永野麗子, 赤沼宏美, 座波ひろ子, 大迫誠一郎. ヒト ES 細胞の神経分化系を用いたサリドマイドの影響についての研究. 第13回日本外因性内分泌攪乱化学物質学会、東京（2010）12月16日
14. 大迫誠一郎, 永野麗子, 何小明, 今西哲、藤渕航, 赤沼宏美, 曾根秀子. マウスおよびヒト ES 細胞を用いた神経分化培養系におけるメチル水銀の毒性影響評価. 第13回日本外因性内分泌攪乱化学物質学会、東京（2010）12月16日東京大学山上会館

H. 知的財産権の出願・登録状況

1. 特許取得

1. 曾根秀子、大迫誠一郎、永野麗子、今西哲、赤沼宏美、宮崎航. 「胎生プログラミングに対する影響を評価するための方法」. 特願 2009-81497（識別番号 100078662）（2009）.
2. 大迫誠一郎、栗田尚佳. 「CpG メチル化頻度の変動をゲノムワイドに比較解析するための DNA 試料作成方法」. 米国仮出願（出願番号：61/309971）（2010）.

2. 実用新案登録

なし

3. その他

なし

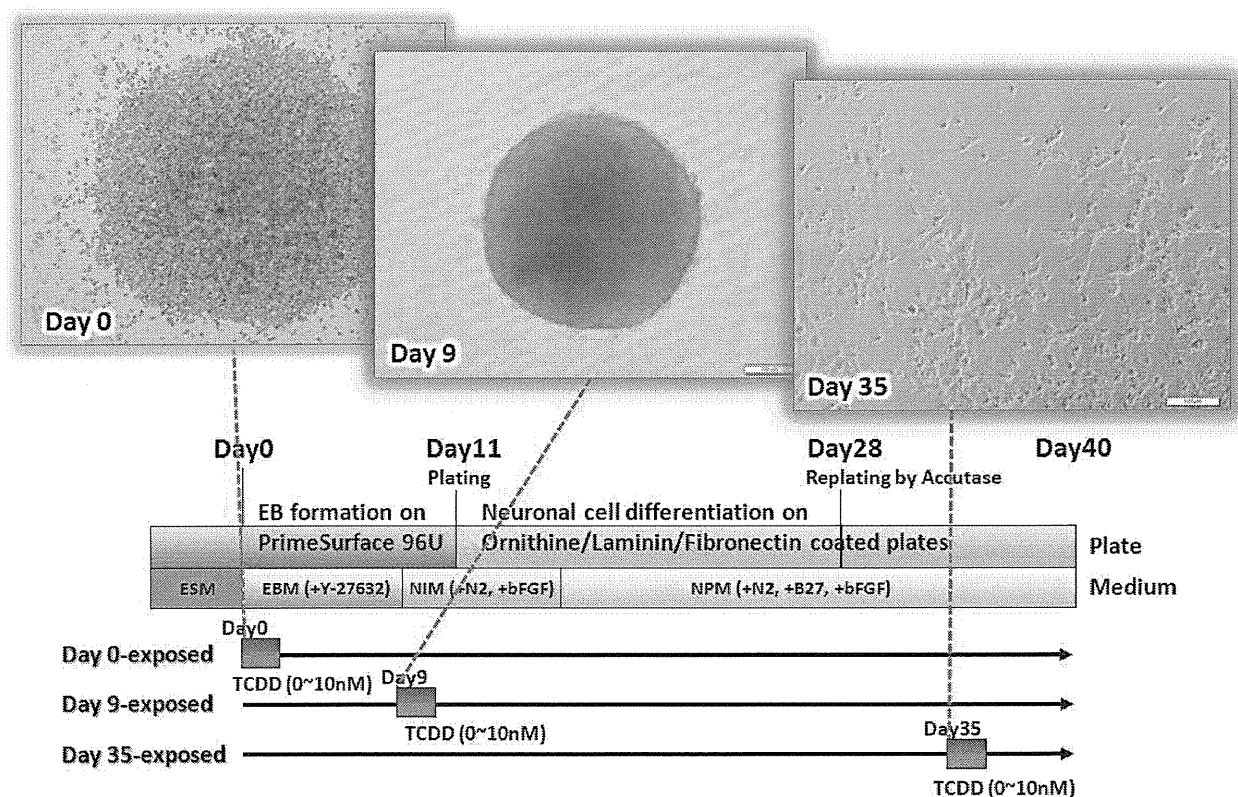


図 1. 培養スケジュールと TCDD 曝露時期。

KhES-1 を MEF との共培養で増幅した（培養液 ES 培地（ESM））。十分なクランプ形成を確認し、解離液で MEF を分離してシングルセルとした。これを EB 形成デバイスである SUMILON PrimeSurface 96U に播種し Day0 とした。この段階で TCDD を 24 時間曝露した群、Day9 から 24 時間曝露した群を設けた。培養液には Rock 阻害剤 Y-27632 を添加した EB 培地を Day8 まで使用した。Day9 からは NIM に変更し神経幹細胞誘導を行った。Day15 から NPM に換え、Day28 で再播種し、Day40 まで神経細胞誘導を行った。Day35 で TCDD を曝露する群も設けた。

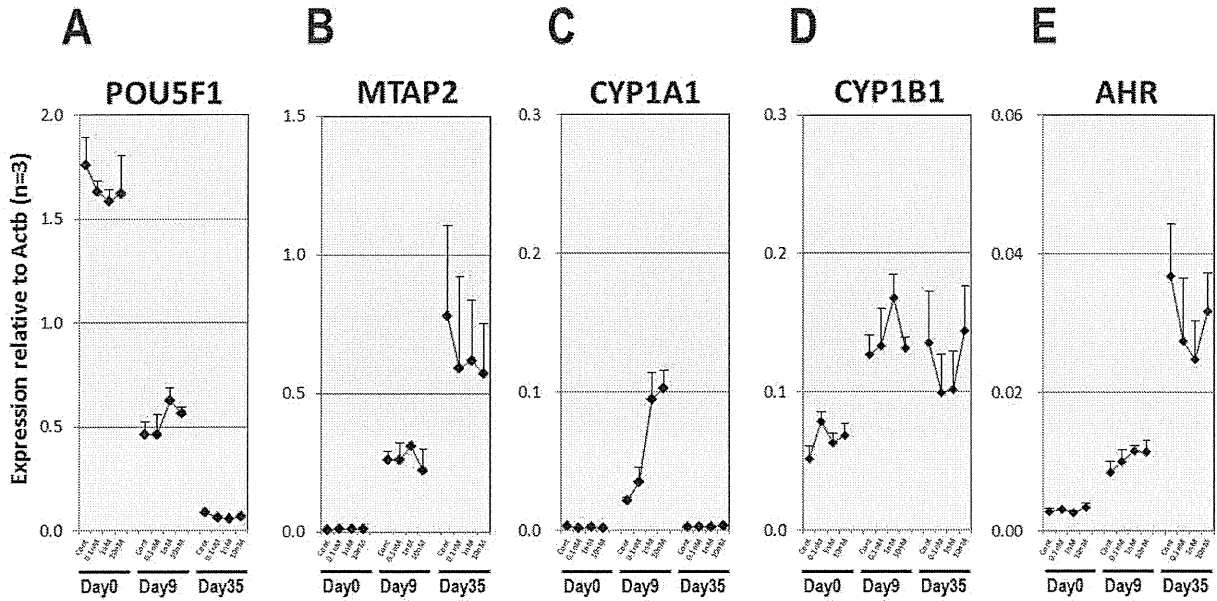


図 2. 各曝露ステージにおける分化マーカーとダイオキシンバッチリ遺伝子の変動。

定量 RT-PCR 法でベータアクチンを内部標準として定量した。A) *Pouf51*. B) *Mtap2*. C) *Cyp1a1*. D) *Cyp1b1*. E) *Ahr*. 幹細胞マーカーである *Pouf51* および神経細胞 dendrocyte マーカーである *Mtap2* の発現が分化ステージを各々減少と増加しており分化培養系は成立していることがうかがわれる。ダイオキシンバイオマーカーである *Cyp1a1* は Day9 でのみ反応することがわかり、ES 細胞や十分に分化した神経細胞では誘導がかからないことがわかった。

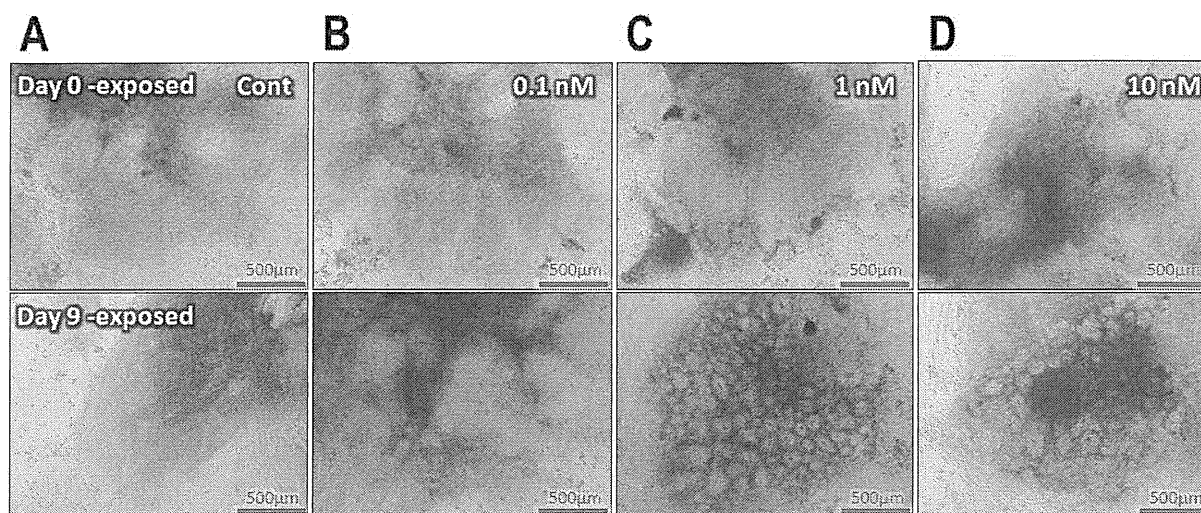


図 3. Day0 および Day9 での 24 時間 TCDD 曝露後の Day25 におけるプレート上での分化状態。

定量 RT-PCR 法でベータアクチンを内部標準として定量した。A) DMSO. B) 0.1nM TCDD. C) 1nM TCDD. D) 10nM TCDD. 上段 (Day0 曝露)。下段 (Day9 曝露)。肉眼的に観察されるニューラルロゼッタが Day9 曝露群 (1nM および 10nM) で多くなることがわかった。

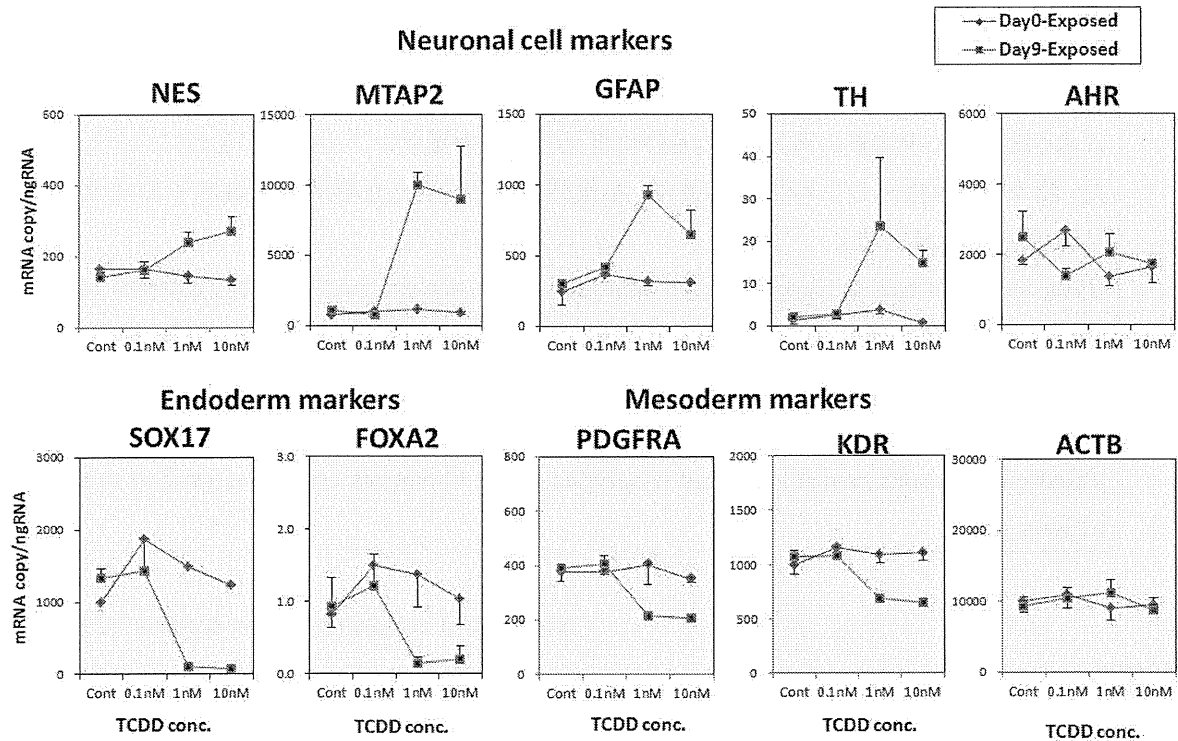


図 4. Day0 および Day9 での 24 時間 TCDD 曝露後の Day28 における分化マーカー遺伝子レベル。
定量 RT-PCR 法でコピー数を測定した。神経細胞マーカー (*Nes*, *Mtap2*, *Gfap*, *Th*)、内胚葉マーカー (*Sox17*, *Foxa2*)、中胚葉マーカー (*Pdgfr*, *Kdr*) および (*Ahr*, *Actb*) のレベルを測定した。Day9 曝露 (1nM および 10nM TCDD) で神経細胞マーカーの上昇と内胚葉および中胚葉マーカーの減少が著しいことがわかった。

トキシコゲノミクスデータ解析：類型化・ランキング・次元削減
Analysis of toxicogenomic data: Classification, ranking and dimensionality reduction

Jean-Francois Pessiot, Wataru Fujibuchi

研究要旨

いくつかの機械学習ベースのアプローチを使用して、本事業で取得したトキシコゲノミクスデータセットを4つの異なる側面から分析した。具体的には：化学物質の類型、用量による変化、化学物質の可分性、ならびに低用量と高用量サンプルの違いを基に分析した。我々の予測ならびに可視化結果は、本実験で使用された化学物質と遺伝子の重要な特性を明らかにした。

Using several machine learning based approaches, we analyze four different aspects of the toxicogenomic dataset: the classification of chemical types, the differences between dose samples and control samples, the chemicals separability, and the differences between low dose samples and high dose samples. Our prediction results and visualization results reveal important properties of the chemicals and genes used in this experiment.

A. PURPOSE

In this work, we used a machine learning approach to analyze a toxicogenomic dataset. First, we studied the problem of predicting a chemical's class given its gene expression patterns. This is a very important problem, because an accurate prediction system would be very helpful to study the properties of new chemicals. Another example of application is a scenario where a sample has been contaminated by an unknown chemical. In order to provide an appropriate treatment, it would be useful to predict the class.

Then, we studied the difference between dose samples and control samples. If a chemical is active, then the gene expression patterns between dose and control should be different. This is an important way to check the chemicals properties (e.g., chemicals which show no gene expression changes between dose samples

and control samples might be considered harmless).

Then, we studied the separability of the dataset at the chemical level. We wanted to know if, for a given drug, the gene expression vectors for different time points are similar. This is a way to analyze the influence of time on a given chemical.

Last, we studied the difference between low dose samples and high dose samples. If a chemical is only active above a specific dose threshold, then the gene expression patterns of low dose samples should differ from the patterns of high dose samples. Accurate identification of such thresholds may improve the understanding of these chemicals, as well as setting specific safety guidelines regarding exposure to chemicals.

B. METHODS

Pre-processings.

All gene expression values were normalized by their corresponding b-ACTIN values. All feature vectors were normalized such that their Euclidean norms equal one.

Support Vector Machine.

SVM is considered the state-of-the-art algorithm for solving binary classification problems, and is available in many free and public libraries [33-b]. The goal of an SVM is to learn a separating hyperplane between two classes, in such a way that the margin between the classes is large. SVM can be trained using kernels functions, which can thought of as similarity measures between the samples [39-b]. Let $K(x, y)$ be the kernel function between any two gene expression vectors x and y . Given a set of n gene expression vectors x_1, \dots, x_n that belong to the positive class (labeled +1) or the negative class (labeled -1), the decision function learned by an SVM for a new test vector x is as follows:

$$f(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$$

The goal of the SVM is to compute the optimal weights α_i , such that genes in the class +1 have large positive scores and genes in the class -1 have large negative scores. The regularization parameter C has to be set by the user and is used to control the overfitting problem on the training set. The SVM function $f(x)$ can then be used to generate predictions for new vectors for which class labels are unknown.

In our experiments, we used the popular Gaussian kernel between two vectors u and v :

$$K(u, v) = \exp\left(-\frac{\|u-v\|^2}{2\sigma^2}\right)$$

The second parameter “sigma” is the bandwidth parameter of the Gaussian kernel.

Pairwise Ranking Component Analysis.

Distance learning. Distance metrics are essential components in many machine learning algorithms [42,34]. In the classification setting, good distance metrics should improve the accuracy of such algorithms as the Nearest Neighbour [19,41]. In the unsupervised setting, such algorithms as K-Means or multi-dimensional scaling critically rely on the choice of the distance function. Distance metrics can also be used for visualization and interpretation purposes, and therefore are also important for understanding data. However, choosing the appropriate distance function manually is a difficult problem that requires prior knowledge. This underscores the need to automatically learn good distance metrics from available data.

A common approach to distance metric learning takes place in the classification setting. The learned distance metric should help improve classification accuracy.

Many distance metric learning methods have been proposed for classification and most of them learn a Mahalanobis distance metric to optimize the classification error directly or indirectly [21,23,36]. Among them, Linear Discriminant Analysis (LDA) learns a Mahalanobis distance metric such that the classes are compactly grouped around their centers, while being far from each other [14]. The intuition is that the classification task might become easier if the classes are well separated. More recently, Neighbourhood Components Analysis (NCA) directly optimizes the classification error of the nearest

neighbour classifier [19]. Other related studies optimize the margin [41] or the class compactness [18], learn the local structure of the data [36], use information theory [10, 24], or work in the online learning setting [33,22]. All these methods share a common objective: they learn a Mahalanobis distance metric so that the classification task would become easier [29].

However, ranking the data vectors is sometimes more important than predicting the labels themselves [9,27]. In the information retrieval (IR) setting, for example, a user submits a query to an IR system. The IR system should then return a list of gene expression vectors ranked with respect to their relevance to the query [31]. The most relevant gene expression vectors should appear at the top of the list and the less relevant ones should appear at the bottom of the list. Online search engines, such as Google, are typical examples of this setting. The quality of such systems relies on their ability to correctly rank the web pages rather than their ability to classify them as relevant or not. This problem has attracted considerable attention in the last few years (see e.g. [17,20,5,1,12]).

We are interested in learning a distance metric for the ranking setting [6]. Our primary goal is to improve the ranking performance in IR systems that return ranked lists of gene expression vectors to users. Previous works have proposed solutions to this problem [32,11]. However, those works mainly focus on the prediction accuracy and the scalability. Although those criteria are important, there are other desirable properties for a distance metric learning method: interpretability [8,25,26,43] and simplicity. In this work, we focus on these two properties. Interpretability is crucial to discover the latent structure of a dataset [13,37,40]. The latent structure can then be used to visualize, understand, and summarize the data [3,4,30,35,44].

This is especially useful for large-scale datasets, for which the analysis of each item would be prohibitively time-consuming. Simplicity is also an important property of our methodology, because databases are becoming increasingly common tools to store and query data. We want our distance metric learning method to be easy to understand and to implement by non-specialists.

Description of PARCA. We consider an IR setting [27,28,31,7]. A user submits a query (a gene expression vector) to the IR system. The goal of the IR system is to present to the user a list of the database's gene expression vectors, sorted with respect to their similarity to the query: the most similar gene expression vectors appear at the top of the list and the less similar ones appear at the bottom.

One simple way to do this is to compare the query with the gene expression vectors using a distance function (e.g., Euclidean distance). Given the vector representation of the query and the gene expression vectors [15], we simply compute the distances between the query vector and the gene expression vectors. If gene expression vector A is closer to the query than gene expression vector B, then A is considered more relevant and should be ranked higher than B. The result is an ordered list of ranked gene expression vectors, from the most relevant gene expression vectors to the least relevant ones. This ordered list of gene expression vectors can then be presented to the user.

In order to design efficient IR systems, we need to measure the quality of such ordered lists. In other words, we need to define a performance measure for the evaluation of the ranking accuracy. Many criteria and performance measures have been studied in the

IR literature (see e.g., [2]). In this work, we focus on the pairwise ranking error.

Pairwise ranking error. First, let us consider a set of examples $S = \{(x_i, y_i)\}_{i=1..N}$, where each example is composed of a vector $x_i \in \mathbb{R}^D$ and a class label $y_i \in \{1, \dots, C\}$. In practice, S may be a gene expression vector database where x_i represents the gene expression levels of sample i , and y_i represents the gene expression vector's label (e.g. cancer or healthy).

To measure the quality of the IR system, we consider the pairwise ranking error [20,39]. First, let us consider the gene expression vector $(x_i, y_i) \in S$ as the query. The user is only interested in other gene expression vectors similar to the query, i.e., gene expression vectors with the same class label y_i . Now let us consider two gene expression vectors $(x_j, y_j) \in S$ and $(x_k, y_k) \in S$, such that x_j belongs to the same class as x_i (i.e., $y_j = y_i$), and x_k belongs to any other class (i.e., $y_k \neq y_i$). As x_i and x_j belong to same class, we expect x_i to be closer to x_j than to x_k . If we consider the Euclidean distance, this means that $\|x_i - x_j\|^2$ should be smaller than $\|x_i - x_k\|^2$. When this is not the case, there is a ranking error. We can now consider the following quantity:

$$\mathbb{I}[\|x_i - x_j\|^2 > \|x_i - x_k\|^2] = \begin{cases} 1 & \text{if } x_i \text{ is closer to } x_k \text{ than } x_j \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbb{I}[p] = 1$ if the predicate p is true, 0 otherwise. In other words, this quantity equals 1 when there is a ranking error, 0 otherwise. By doing this, we are essentially casting the ranking problem in the classification setting (see e.g. [20,38]). Considering (x_i, y_i) as the query, we can now compute the ranking error E_i :

$$E_i = \frac{1}{\alpha_i} \sum_{\substack{j \in C_i \\ j \neq i}} \sum_{k \notin C_i} \mathbb{I}[\|x_i - x_j\|^2 > \|x_i - x_k\|^2]$$

where $C_i = \{j = 1 \dots N \mid y_j = y_i\}$ is the set of indices corresponding to class y_i and α_i is the number of pairs (j, k) such that x_i and x_j belong to the same class while x_i and x_k belong to different classes: $\alpha_i = \{ (y_j, y_k) \mid j \in C_i, j \neq i, k \notin C_i \}$.

With such a normalization, we have $E \in [0, 1]$, with $E = 0$ if all the pairs are correctly ordered and $E_i = 1$ if they are all incorrectly ordered. The total ranking error E is the sum of the normalized ranking errors over the whole dataset:

$$E = \sum_{(x_i, y_i) \in S} \frac{1}{\alpha_i} \sum_{\substack{j \in C_i \\ j \neq i}} \sum_{k \notin C_i} \mathbb{I}[\|x_i - x_j\|^2 > \|x_i - x_k\|^2]$$

Our goal is to learn a distance metric that minimizes the pairwise ranking error.

The ranking error E uses the standard Euclidean distance function for ranking the data vectors: $d(x_i, x_j)^2 = (x_i - x_j)^T (x_i - x_j) = \|x_i - x_j\|^2$. However, the initial feature space may not be optimal for the ranking task, thus leading to a large ranking error. Therefore, our goal is to learn an appropriate distance function from the available data. To this end, we consider the Mahalanobis distance:

$$d(x_i, x_j)^2 = (x_i - x_j)^T A^T A (x_i - x_j) = \|Ax_i - Ax_j\|^2$$

parameterized by the matrix $A \in \mathbb{R}^{K \times D}$, where D is the number of dimensions of the initial feature space and $K \leq D$ is the number of dimensions of the new feature space. It is clear that the Mahalanobis distance is equivalent to a linear projection from \mathbb{R}^D to \mathbb{R}^K using

A, followed by the standard Euclidean distance.

Replacing the Euclidean distance with the Mahalanobis distance in the definition of E, the ranking error now writes:

$$E(\mathbf{A}) = \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in S} \frac{1}{\alpha_i} \sum_{\substack{j \in C_i \\ j \neq i}} \sum_{\substack{k \in C_i \\ k \neq i}} \mathbb{I}[\|\mathbf{Ax}_i - \mathbf{Ax}_j\|^2 > \|\mathbf{Ax}_i - \mathbf{Ax}_k\|^2]$$

A good projection matrix should have a low ranking error. Therefore, the projection matrix that minimizes the ranking error is the solution of:

$$\min_{\mathbf{A} \in \mathbb{R}^{K \times D}} E(\mathbf{A})$$

The definition of E(A) contains boolean predicates that are not differentiable. Therefore, the cost function E(A) is difficult to minimize. As [16] suggested, we use $\mathbb{I}[x < 0] < \exp(-x)$ and consider the following upper bound:

$$F(\mathbf{A}) = \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in S} \frac{1}{\alpha_i} \sum_{\substack{j \in C_i \\ j \neq i}} \sum_{\substack{k \in C_i \\ k \neq i}} \exp\left(\|\mathbf{Ax}_i - \mathbf{Ax}_j\|^2 - \|\mathbf{Ax}_i - \mathbf{Ax}_k\|^2\right)$$

It is clear that we have $E(\mathbf{A}) < F(\mathbf{A})$. The new cost function is now differentiable and can be minimized with standard gradient descent strategies. However, the computation of F(A) is expensive as it requires consideration of triplets of data vectors $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$. A naive implementation would have a computational complexity of $O(N^3K + NDK)$, where N is the total number of examples, D is the number of dimensions, and K is the distance metric's rank. As we want to use PARCA with large-scale datasets, a more efficient computation is required.

Denoting $x_{ij} = \mathbf{x}_i - \mathbf{x}_j$, we write F(A) as:

$$\begin{aligned} F(\mathbf{A}) &= \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in S} \frac{1}{\alpha_i} \sum_{\substack{j \in C_i \\ j \neq i}} \sum_{\substack{k \in C_i \\ k \neq i}} \exp\left(\|\mathbf{Ax}_{ij}\|^2 - \|\mathbf{Ax}_{ik}\|^2\right) \\ &= \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in S} \frac{1}{\alpha_i} \sum_{\substack{j \in C_i \\ j \neq i}} \exp\left(\|\mathbf{Ax}_{ij}\|^2\right) \sum_{\substack{k \in C_i \\ k \neq i}} \exp\left(-\|\mathbf{Ax}_{ik}\|^2\right) \end{aligned}$$

With this form, the computational complexity of F(A) is now $O(N^2K + N^2DK)$.

Derivatives. Let us differentiate the cost function F with respect to the projection matrix A:

$$\begin{aligned} \frac{\partial F}{\partial \mathbf{A}} &= -2\mathbf{A} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in S} \frac{1}{\alpha_i} \left(\sum_{\substack{j \in C_i \\ j \neq i}} \exp\left(\|\mathbf{Ax}_{ij}\|^2\right) \times \sum_{\substack{k \in C_i \\ k \neq i}} \mathbf{x}_{ik} \mathbf{x}_{ik}^T \exp\left(-\|\mathbf{Ax}_{ik}\|^2\right) \right. \\ &\quad \left. + \sum_{\substack{j \in C_i \\ j \neq i}} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \exp\left(\|\mathbf{Ax}_{ij}\|^2\right) \times \sum_{\substack{k \in C_i \\ k \neq i}} \exp\left(-\|\mathbf{Ax}_{ik}\|^2\right) \right) \end{aligned}$$

The computational complexity of the derivative $\partial F / \partial \mathbf{A}$ is $O(N^2DK)$. We can now apply standard gradient descent strategies to minimize F. Such strategies are implemented in most scientific computing libraries, e.g., scipy for Python or GSL for C/C++. However, as the function is not convex, the optimization only results in local minima. In practice, we run several gradient descents with random initializations and keep the best result.

C. RESULTS

Chemical class identification.

We considered the twenty-two chemicals belonging to the following three classes: neural, carcinogenic (genetic) and carcinogenic (non-genetic). In this section, our goal is to identify which class a given chemical belongs to. Such a prediction problem is useful when unknown chemicals are detected, as appropriate treatments require some information about the chemical type.