non-coding RNAs may be involved. The progress of research on acquired phenotypes will have a significant effect on perinatal medical aspects.

Acknowledgments
The author gratefully acknowledge Dr Kayoko Shimoi of the University of Shizuoka and Dr Tohru Shibuya for providing the opportunity to submit the manuscript to Genes and Environment.

References

1. Barker DJ, Osmond C. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. Lancet 1986; 1: 1077-81.
2. Barker DJ, Shiell AW, Barker ME, Law CM. Growth in utero and blood pressure levels in the next generation. J Hypertens 2000; 18: 843-846.
3. Sinclair KD, Lea RG, Rees WD, Young LE. The developmental origins of health and disease: current theories and epigenetic mechanisms. Soc Reprod Fertil Suppl. 2007; 64: 425-43.
4. Swanson JM, Entringer S, Buss C, Wadhwa PD. Developmental origins of health and disease: environmental exposures. Semin Reprod Med. 2009; 27: 391-402.
5. Rosenfeld CS. Animal models to study environmental epigenetics. Biol Reprod. 2010; 82: 473-88.
6. LeBaron MJ, Rasoulpour RJ, Klapacz J, Ellis-Hutchings RG, Hollnagel HM, Gollapudi BB. Epigenetics and chemical safety assessment. Mutat Res. 2010; 705: 83-95.
7. Bernal AJ, Jirtle RL. Epigenomic disruption: the effects of early developmental exposures. Birth Defects Res A Clin Mol Teratol. 2010; 88: 938-44.
8. Goodman JI, Augustine KA, Cunnningham ML, Dixon D, Dragan YP, Falls JG, Rasoulpour RJ, Sills RC, Storer RD, Wolf DC, Pettit SD. What do we need to know prior to thinking about incorporating an epigenetic evaluation into safety assessments? Toxicol Sci 2010; 116: 375-81.
9. Ho DH, Burggren WW. Epigenetics and transgenerational transfer: a physiological perspective. J Exp Biol. 2010; 213: 3-16.
10. Corpet A, Almouzni G. Making copies of chromatin: the challenge of nucleosomal organization and epigenetic information. Trends Cell Biol. 2009; 19: 29-41.
11. Bostick M, Kim JK, Esteve PO, Clark A, Pradhan S, Jacobsen SE. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science 2007; 317: 1760-4.
12. Bestor T, Laudano A, Mattaliano R, Ingram V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. J Mol Biol. 1988; 203: 971-83.
13. Iida T, Suetake I, Tajima S, Morioka H, Ohta S, Obuse C, Tsurimoto T. PCNA clamp facilitates action of DNA cytosine methyltransferase 1 on hemimethylated DNA. Genes Cells. 2002; 7: 997-1007.
14. Comb M, Goodman HM. CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. Nucleic Acids Res. 1990; 18: 3975-82.
15. Rideout WM, 3rd, Eversole-Cire P, Spruck CH, 3rd, Hustad CM, Coetzee GA, Gonzales FA, Jones PA. Progressive increases in the methylation status and heterochromatinization of the myoD CpG island during oncogenic transformation. Mol Cell Biol. 1994; 14: 6143-52.
16. Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. Science 2001; 293: 1089-93.
17. Shilatifard A. Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. Annu Rev Biochem. 2006; 75: 243-69.
18. Turner BM. Cellular memory and the histone code. Cell 2002; 111: 285-91.
19. Probst AV, Dunleavy E, Almouzni G. Epigenetic inheritance during the cell cycle. Nat Rev Mol Cell Biol. 2009; 10: 192-206.
20. Djupedal I, Ekwall K. Epigenetics: heterochromatin meets RNAi. Cell Res. 2009; 19: 282-95.
21. Francis NJ. Mechanisms of epigenetic inheritance: copying of polycomb repressed chromatin. Cell Cycle 2009; 8: 3513-18.
22. Hales CN, Barker DJ. The thrifty phenotype hypothesis. Br Med Bull. 2001; 60: 5-20.

23. Wadhwa PD, Buss C, Entringer S, Swanson JM. Developmental origins of health and disease: brief history of the approach and current focus on epigenetic mechanisms. Semin Reprod Med. 2009; 27: 358-68.
24. Ohtani-Fujita N, Dryja TP, Rapaport JM, Fujita T, Matsumura S, Ozasa K, Watanabe Y, Hayashi K, Maeda K, Kinoshita S, Matsumura T, Ohnishi Y, Hotta Y, Takahashi R, Kato MV, Ishizaki K, Sasaki MS, Horsthemke B, Minoda K, Sakai T. Hypermethylation in the retinoblastoma gene is associated with unilateral, sporadic retinoblastoma. Cancer Genet Cytogenet 1997; 98: 43-9.
25. Esteller M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene 2002; 21: 5427-40.
26. McCabe MT, Low JA, Daignault S, Imperiale MJ, Wojno KJ, Day ML. Inhibition of DNA methyltransferase activity prevents tumorigenesis in a mouse model of prostate cancer. Cancer Res. 2006; 66: 385-92.
27. Ushijima T. Detection and interpretation of altered methylation patterns in cancer cells. Nat Rev Cancer 2005; 5: 223-31.
28. Ushijima T, Asada K. Aberrant DNA methylation in contrast with mutations. Cancer Sci. 2010; 101: 300-5.
29. McMullen S, Mostyn A. Animal models for the study of the developmental origins of health and disease. Proc Nutr Soc. 2009; 68: 306-20.
30. Liu D, Diorio J, Tannenbaum B, Caldji C, Francis D, Freedman A, Sharma S, Pearson D, Plotsky PM, Meaney MJ. Maternal care, hippocampal glucocorticoid receptors, and hypothalamic-pituitary-adrenal responses to stress. Science 1997; 277: 1659-62.
31. Weaver IC, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, Seckl JR, Dymov S, Szyf M, Meaney MJ. Epigenetic programming by maternal behavior. Nat Neurosci. 2004; 7: 847-54.
32. Yen TT, Gill AM, Frigeri LG, Barsh GS, Wolff GL. Obesity, diabetes, and neoplasia in yellow A(vy)/- mice: ectopic expression of the agouti gene. Faseb J. 1994; 8: 479-88.
33. Wolff GL, Kodell RL, Moore SR, Cooney CA. Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. Faseb J. 1998; 12: 949-57.
34. Waterland RA, Jirtle RL. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. Mol Cell Biol. 2003; 23: 5293-300.
35. Dolinoy DC, Weidman JR, Waterland RA, Jirtle RL. Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. Environ Health Perspect. 2006; 114: 567-72.
36. Dolinoy DC, Huang D, Jirtle RL. Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. Proc Natl Acad Sci U S A 2007; 104: 13056-13061.
37. Bromer JG, Wu J, Zhou Y, Taylor HS. Hypermethylation of homeobox A10 by in utero diethylstilbestrol exposure: an epigenetic mechanism for altered developmental programming. Endocrinology 2009; 150: 3376-82.
38. Onishchenko N, Karpova N, Sabri F, Castren E, Ceccatelli S. Long-lasting depression-like behavior and epigenetic changes of BDNF gene expression induced by perinatal exposure to methylmercury. J Neurochem. 2008; 106: 1378-87.
39. Wu Q, Ohsako S, Ishimura R, Suzuki JS, Tohyama C. Exposure of mouse preimplantation embryos to 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) alters the methylation status of imprinted genes H19 and Igf2. Biol Reprod. 2004; 70: 1790-7.
40. Mably TA, Moore RW, Goy RW, Peterson RE. In utero and lactational exposure of male rats to 2,3,7,8-tetrachlorodibenzo-p-dioxin. 2. Effects on sexual behavior and the regulation of luteinizing hormone secretion in adulthood. Toxicol Appl Pharmacol. 1992; 114: 108-17.
41. Ohsako S, Miyabara Y, Sakaue M, Ishimura R, Kakeyama M, Izumi H, Yonemoto J, Tohyama C. Developmental stage-specific effects of perinatal 2,3,7,8-tetrachlorodibenzo-p-dioxin exposure on reproductive organs of male rat offspring. Toxicol Sci. 2002; 66: 283-92.
42. Ohsako S, Fukuzawa N, Ishimura R, Kawakami T, Wu Q, Nagano R, Zaha H, Sone H, Yonemoto J, 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 2010; 82: 636-43.

43. Whitlock JP, Jr. Induction of cytochrome P4501A1. Annu Rev Pharmacol Toxicol 1999; 39: 103-25.
44. Sims P, Grover PL, Swaisland A, Pal K, Hewer A. Metabolic activation of benzo(a)pyrene proceeds by a diol-epoxide. Nature 1974; 252: 326-8.
45. Brown NM, Manzolillo PA, Zhang JX, Wang J, Lamartiniere CA. Prenatal TCDD and predisposition to mammary cancer in the rat. Carcinogenesis 1998; 19: 1623-9.
46. Muto T, Wakui S, Imano N, Nakaaki K, Takahashi H, Hano H, Furusato M, Masaoka T. Mammary gland differentiation in female rats after prenatal exposure to 3,3',4,4',5-pentachlorobiphenyl. Toxicology 2002; 177: 197-205.
47. Wakui S, Yokoo K, Takahashi H, Muto T, Suzuki Y, Kanai Y, Hano H, Furusato M, Endou H. Prenatal 3,3',4,4',5-pentachlorobiphenyl exposure modulates induction of rat hepatic CYP 1A1, 1B1, and AhR by 7,12-dimethylbenz[a]anthracene. Toxicol Appl Pharmacol. 2006; 210: 200-11.
48. Okino ST, Pookot D, Li LC, Zhao H, Urakami S, Shiina H, Igawa M, Dahiya R. Epigenetic inactivation of the dioxin-responsive cytochrome P4501A1 gene in human prostate cancer. Cancer Res. 2006; 66: 7420-8.
49. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. Nat Rev Genet. 2007; 8: 253-62.
50. Anway MD, Cupp AS, Uzumcu M, Skinner MK. Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science 2005; 308: 1466-9.
51. Skinner MK, Anway MD. Seminiferous cord formation and germ-cell programming: epigenetic transgenerational actions of endocrine disruptors. Ann N Y Acad Sci. 2005; 1061: 18-32.
52. Crews D, Gore AC, Hsu TS, Dangleben NL, Spinetta M, Schallert T, Anway MD, Skinner MK. Transgenerational epigenetic imprints on mate preference. Proc Natl Acad Sci U S A. 2007; 104: 5942-6.
53. Guerrero-Bosagna C, Settles M, Lucker B, Skinner MK. Epigenetic transgenerational actions of vinclozolin on promoter regions of the sperm epigenome. PLoS One 2010; 5: e13100.
54. Inawaka K, Kawabe M, Takahashi S, Doi Y, Tomigahara Y, Tarui H, Abe J, Kawamura S, Shirai T. Maternal exposure to anti-androgenic compounds, vinclozolin, flutamide and procymidone, has no effects on spermatogenesis and DNA methylation in male rats of subsequent generations. Toxicol Appl Pharmacol. 2009; 237: 178-87.
55. Carone BR, Fauquier L, Habib N, Shea JM, Hart CE, Li R, Bock C, Li C, Gu H, Zamore PD, Meissner A, Weng Z, Hofmann HA, Friedman N, Rando OJ. Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. Cell 2010; 143: 1084-96.
56. Rassoulzadegan M, Grandjean V, Gounon P, Vincent S, Gillot I, Cuzin F. RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. Nature 2006; 441: 469-74.
57. Rando OJ, Verstrepen KJ. Timescales of genetic and epigenetic inheritance. Cell 2007; 128: 655-68.
58. Maderspacher F. Lysenko rising. Curr Biol. 2010; 20: R835-7.

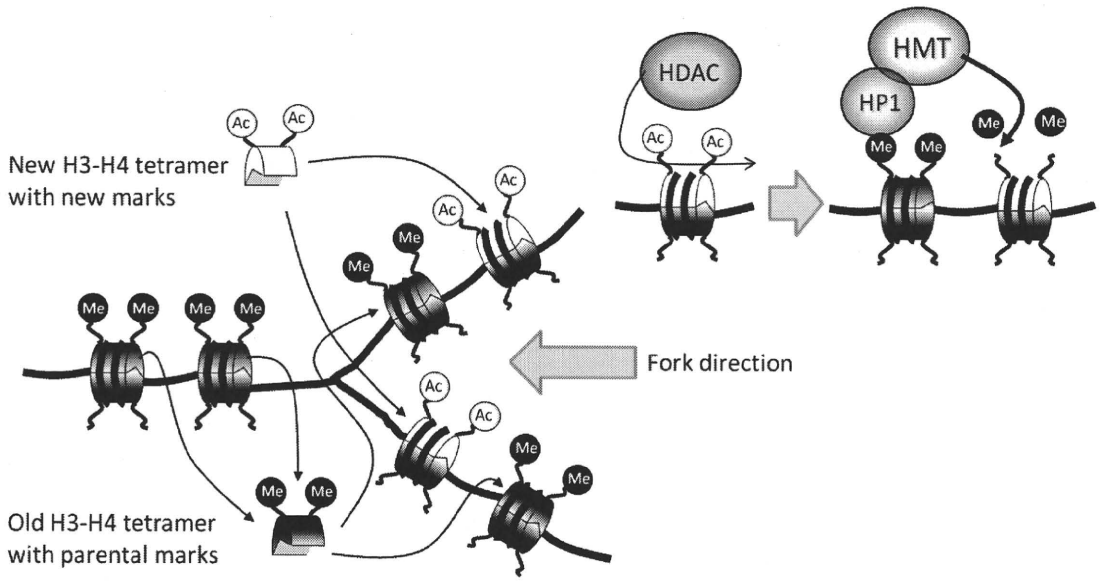

Figure legends

Figure 1. A mechanistic representation of a hypothetical example of histone modification pattern inheritance. On newly synthesized DNA, parental histone H3-H4 tetramers are reused for the formation of the nucleosome of daughter chromatin. At the same time, newly synthesized histone H3-H4 tetramers are incorporated at random. The new histone marks (*e.g.*, acetylated lysine) are erased by HDACs. Methylated marks derived from parental histones neighboring newly incorporated histones are recognized by HP1, which forms a complex with HMT. Erased marks in the newly incorporated histones are quickly methylated by HMT. In this manner, histone modification patterns in a specific region of the genome are precisely copied. (From Probst *et al.*, *Nat Rev Mol Cell Biol.* 2009)

Table 1. Leading reports describing transgenerational effects of chemical exposures on epigenomic alternations in experimental animal models.

	Animal	Chemicals	Dose	Period of exposure	Observed effects on phenotype	Detected changes in DNA methylation
Wu *et al.*, *Biol Reprod* 2004.	ICR mouse	2,3,7,8-tetrachloro dibenzo-*p*-dioxin	10 nM *in vitro*	Fertilized egg with 1-2 cell or in the 8-cell stage	Reduced fetal body weight and reduced H19 mRNA expression	Hypermethyaltion of *H19/Igf2* genomic imprint region of DNA from the fetal body
Anway *et al.*, *Science* 2005. Crews *et al.*, *Proc Natl Acad Sci U S A* 2007.	Sprague-Dawly rats	Vinclozolin Methoxychlor	100 mg/kg/day mother bw ip (Vinclozolin) 200 mg/kg/day mother bw ip (Methoxychlor)	Day 8 to day 15 of gestation	Germ cell apoptosis, decreased sperm count, decreased sexual preference for normal females in male offspring (F1 – F4)	Hypomethylation of LPLase and cytokine-inducible SH2 protein genes in sperm DNA
Doliney *et al.*, *Environ Health Perspect* 2006.	A^{vy} mouse	Genistein	250 mg/kg diet	2 weeks before mating through weaning	Coat-color change (a decease in yellow color and an increase pseudoagouti phenotype)	Hypermethylation of IAP insertion of the Agouti gene promoter region in skin DNA
Doliney *et al.*, *Proc Natl Acad Sci U S A* 2007.	A^{vy} mouse	Bisphenol-A	50 mg/kg diet	2 weeks before mating through weaning	Coat-color change (an increase in yellow color and a decrease in the pseudoagouti phenotype)	Hypomethylation of IAP insertion of the Agouti gene promoter region in skin DNA
Onishchenko *et al.*, *J Neurochem* 2008.	C57BL/6kl mouse	Methylmercury	0.5 mg/kg/day via drinking water	Day 7 of gestation to day 7 of the postnatal period	Depression-like behavior and hippocampal BDNF mRNA suppression	Hypermethyaltion of hippocampal BDNF promoter
Bromer *et al.*, *Endocrinology* 2009.	CD-1 mouse	Diethylstilbestrol	10 µg/kg/day mother bw ip	Day 9 to day 16 of gestation	Abnormalities in the reproductive tract and decreased homeobox A10 expression	Hypermethyaltion of uterus homeobox A10 gene intron

BIOLOGY OF REPRODUCTION **82,** 636–643 (2010)
Published online before print 9 December 2009.
DOI 10.1095/biolreprod.109.080812

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[1]

Seiichiroh Ohsako,[2,7] Noriho Fukuzawa,[3,8] Ryuta Ishimura,[4,8] Takashige Kawakami,[5,8] Qing Wu,[6,8] Reiko Nagano,[9] Hiroko Zaha,[9] Hideko Sone,[9] Junzo Yonemoto,[9] and Chiharu Tohyama[7]

Division of Environmental Health Sciences,[7] Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
Environmental Health Sciences Division[8] and Research Center for Environmental Risk,[9] National Institute for Environmental Studies, Ibaraki, Japan

ABSTRACT

TCDD (2,3,7,8-tetrachlorodebenzo-*p*-dioxin) requires the presence of the aryl hydrocarbon receptor (*Ahr*) gene for its toxic effects, such as reproductive disorders in male offspring of maternally exposed rats and mice. To study the involvement of the *Ahr* gene in producing the toxic phenotype with respect to testicular development, we administered a relatively high dose of TCDD to mice with three different maternally derived *Ahr* genotypic traits, and then compared several *Ahr*-dependent alterations among male reproductive systems on Postnatal Day 14. Reduction in anogenital distance and expression of prostatic epithelial genes in the urogenital complex (UGC) were detected in $Ahr^{+/+}$ and $Ahr^{+/-}$ mice exposed to TCDD, whereas no difference was observed in $Ahr^{-/-}$ mice. In situ hybridization revealed the absence of probasin mRNA expression in the prostate epithelium, despite the obvious development of prostatic lobes in TCDD-exposed mice. In contrast to obvious prostatic dysfunction and induction of cytochrome P450 (CYP) family genes in the UGC by TCDD, no alterations in testicular functions were observed in germ cell/Sertoli cell/interstitial cell marker gene expression or CYP family induction. No histopathological changes were observed among the three genotypes and between control and TCDD-exposed mice. Therefore, mouse external genitalia and prostatic development are much more sensitive to TCDD treatment than testis. Further, the *Ahr* gene, analyzed in this study, does not significantly contribute to testicular function during perinatal and immature stages, and the developing mouse testis appears to be quite resistant to TCDD exposure.

aryl hydrocarbon receptor, developmental biology, dioxin, knockout mouse, prostate, spermatogenesis, testis, toxicology

INTRODUCTION

TCDD (2,3,7,8-tetrachlorodebenzo-*p*-dioxin) is an extremely potent xenobiotic chemical. Maternal exposure to TCDD induces a wide range of physiological alterations and toxicities in the fetus and pups of laboratory animals and perhaps in humans [1]. TCDD induces various toxicological endpoints in male reproductive organs, such as decreased size of sex-accessory glands and reduced sperm counts in testis, epididymis, and ejaculate [2–10]. Although male rat and mouse offspring exposed to TCDD in utero can produce testicular androgen normally, the androgen responsiveness of the ventral prostate is lowered by unknown mechanisms of TCDD [9–12].

Interestingly, the effects of TCDD on testicular development and perinatal stage spermatogenesis are still unclear because the results are contradictory [2–10]. Some studies on testicular development that used conventional experimental animals reported the presence of slight reductions in testicular weight, daily sperm production, and steroidogenesis [2–5]. On the other hand, clear negative data concerning the TCDD effect on testicular development were presented in many other papers [7–10]. Taken together, the impairment in prostate development by in utero TCDD exposure appears to occur in many mammals, including rats and mice, but it is not understood whether or how susceptible the testicular development and perinatal stage spermatogenesis is to the TCDD exposure.

The aryl hydrocarbon receptor (AHR) is a ligand-activated transcription factor that mostly mediates inductions of drug-metabolizing enzymes such as members of the CYP1A family, and it appears to act as a sensor for environmental contaminants such as dioxins and polyaromatic hydrocarbons [13, 14]. Experimental studies with *Ahr* gene knockout mice have already revealed that AHR plays an essential role in the occurrence of multiple TCDD-induced adverse effects, including teratogenic cleft palate and hydronephrosis [15, 16]. Reproductive disorders, such as prostatic growth impairment in male offspring following maternal exposure to TCDD, were also shown to be dependent on the *Ahr* gene [17].

Some evidence also suggested that AHR was involved in developmental signaling in the immune and hepatic systems [18, 19]. In analyses using *Ahr*-null female mice, AHR was

[1]Supported, in part, by the Environmental Technology Development Fund to S.O. and H.S. and the Risk Assessment of Dioxins Fund to C.T. from the Ministry of the Environment, Japan, and by grants from CREST, JST, Japan, to C.T.
[2]Correspondence: Seiichiroh Ohsako, Division of Environmental Health Sciences, Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654. FAX: 81 3 5841 1434;
e-mail: ohsako@m.u-tokyo.ac.jp
[3]Current address: Research Institute of Genome-based Biofactory, AIST, 2-17-2-1 Tsukisamu-higashi, Sapporo, Hokkaido 062-8517, Japan.
[4]Current address: The Jackson Laboratory, 600 Main St., Bar Harbor, ME 04609.
[5]Current address: School of Pharmaceutical Sciences, Tokushima-Bunri University, Tokushima 770-8514, Japan.
[6]Current address: School of Public Health, Fudan University, Shanghai 200032, China.

Received: 13 August 2009.
First decision: 2 September 2009.
Accepted: 17 November 2009.
© 2010 by the Society for the Study of Reproduction, Inc.
eISSN: 1529-7268 http://www.biolreprod.org
ISSN: 0006-3363

Downloaded from www.biolreprod.org.

also required for normal ovarian germ cell dynamics, based on the observation that numbers of nonatretic primordial, primary, and small preantral follicles in ovaries of *Ahr*-null females at early postpartum stages were higher than those in wild-type mice [20–22]. Moreover, AHR cooperates with an orphan nuclear receptor, NR5A1 (also known as Ad4BP/SF-1), to activate P450 aromatase (CYP19) gene transcription in ovarian granulosa cells and modulate endogenous estrogen production in the female reproductive cycle [23]. AHR is also expressed in interstitial cells and male germ cells at specific stages [24]. More recently, AHR has been demonstrated to have a ubiquitin ligase activity, which enhances the degradation of estrogen and androgen receptors, suggesting that AHR may modulate androgen sensitivity in normal development [25]. To date, the possible involvement of AHR in testicular development during immature stages using AHR agonist administration or *Ahr*-null male mice has not been addressed.

To the best of our knowledge, this is the first study that used *Ahr*-null mice to investigate the involvement of the *Ahr* gene in early stages of testicular development and to assess the differences in susceptibility to in utero TCDD exposure between prostate and testicular development.

MATERIALS AND METHODS

Materials

TCDD was purchased from Cambridge Isotope Laboratory (Andover, MA). The purity was higher than 99.5%. Corn oil for dissolving TCDD or the control vehicle was obtained from Sigma-Aldrich (St. Louis, MO). TRIzol reagent, SuperScript III RNase H- Reverse Transcriptase, and oligo(dT)12–18 primer were purchased from Invitrogen (Carlsbad, CA). SYBR Premix Ex Taq (Perfect Real Time) was purchased from TAKARA BIO, Inc. (Otsu, Japan). The plasmid pGEM-TEasy vector was obtained from Promega Corp. (Madison, WI).

Animals and TCDD Administration

Ahr knockout mice were kindly provided by Dr. Yoshiaki Fujii-Kuriyama (Center for Tsukuba Advanced Research Alliance and Institute of Basic Medical Sciences, University of Tsukuba) [15]. They were bred in our own facility at the National Institute for Environmental Studies (NIES). All described procedures were approved by the NIES Institutional Animal Care and Use Committee and were performed in accordance with the Guidelines for Animal Experiments at the NIES. They were maintained in a controlled environment of temperature 24 ± 1°C, humidity 45 ± 5%, and a 12L:12D cycle and were given food and distilled water ad libitum. *Ahr* heterozygous male mice were back-crossed with wild-type female C57BL/6J mice (CLEA Japan, Tokyo, Japan) six times, and heterozygous offspring of both sexes were used in this study. The heterozygous female mice (7- to 10-wk-old) were mated 1:1 with the heterozygous males overnight, and the females that had a vaginal plug on the following morning were designated as being pregnant at Gestational Day 0 (GD0). Dams were housed individually in clear plastic cages with heat-treated wood chips as bedding. On GD13, pregnant mice were given a single dose of TCDD orally (10 μg/kg body weight, close to a lethal dose for a C57BL/6J fetus) or an equivalent volume of vehicle (95% corn oil, 4% n-nonane; 5 ml/kg) as control. Male pups were killed under diethyl ether anesthesia on Postnatal Day 14 (PND14).

Sample Collection

On PND14, immediately before euthanization, the anogenital distance, determined by the length from the base of the genital tubercle to the anterior edge of the anus, was measured with a digital caliper. We also measured the crown-anal length (the distance between the nose and anterior edge of the anus). The testis and epididymis on both sides were excised from the abdomen, and the surrounding adipose tissue was carefully removed. After removing urine from the bladder, the deferent ducts were cut at the base of the bladder. The urogenital complex (UGC), which is a small mass comprising all the lobes of the prostate and seminal vesicle, was then collected by cutting the anterior end of the urethra. All tissue samples were frozen in liquid nitrogen immediately after dissection and kept at −80°C until RNA extraction.

Real-Time RT-PCR

The protocols for real-time RT-PCR quantifications were described previously [26]. Briefly, total RNA was extracted from the UGC (n = 5) and testis (n = 3) using TRIzol reagent. RNA samples were reverse-transcribed with SuperScript III reverse transcriptase and oligo(dT)12–18 primer. Nineteen genes examined in this study: aryl hydrocarbon receptor (*Ahr*), cytochrome P450 1A1 (*Cyp1a1*), cytochrome P450 1A2 (*Cyp1a2*), cytochrome P450 1B1 (*Cyp1b1*), androgen receptor (*Ar*), steroid 5α-reductase type 1 (*Srd5a1*), steroid 5α-reductase type 2 (*Srd5a2*), probasin (*Pbsn*), Mp25 (*Sbp*), PSP94 (*Msmb*), calnexin-t (*Clgn*), Hsp70.2 (*Hspa2*), androgen-binding protein (*Shbg*), cytochrome P450 side chain cleavage (*Cyp11a1*), cytochrome P450 17α/C_{17-20} lyase (*Cyp17a1*), 3β-hydroxysteroid dehydrogenase type I (*Hsd3b1*), 3α-hydroxysteroid dehydrogenase type I (*Akr1c4*), 17β-hydroxysteroid dehydrogenase type III (*Hsd17b3*), and cyclophilin B (*Ppib*). All primer sets are shown in Supplemental Table S1 (available online at www.biolreprod.org). For the real-time RT-PCR, target genes were amplified with SYBR Premix Ex Taq (Perfect Real Time) system by using a LightCycler (Roche, Manheim, Germany). The relative expression level was calculated by normalizing with the average value of each control wild-type ($Ahr^{+/+}$) group. To determine the sequences, the PCR product for each gene was subcloned into pGEM-TEasy vectors or directly sequenced by the dideoxynucleotide chain termination method using the ABI Prism BigDye terminator cycle sequencing kit (PE-Biosystems, Foster City, CA).

In Situ Hybridization

Mouse probasin mRNA was detected in the UGC specimens by in situ hybridization. The UGCs were fixed with neutralized formalin for 48 h, embedded in paraffin, and then cut into 4-μm sections. A template was amplified from the pGEM-TEasy vector inserted with a 377-bp probasin (*Pbsn*) RT-PCR fragment by T7 and SP6 primers to generate sense and antisense transcripts. Digoxigenin-labeled riboprobes were used, and the hybridization was performed using an automated in situ hybridization instrument Gen II (Ventana Medical System, Tucson, AZ). Detection and counterstaining were done with the BlueMap Kit (Ventana Medical System) and Nuclear Fast Red (Sigma-Aldrich).

Immunohistochemistry

Anti-calnexin-t (CLGN), a male germ cell developmental stage-specific protein, was immunostained by the method described previously [27]. Briefly, testes (n = 3; left side) were fixed with Bouin solution and embedded in paraffin. Two different cross-sectional regions (4-μm thickness) from one testis were obtained (six sections from each group). Deparaffinized sections were incubated with anti-mouse calnexin-t antibody, followed by incubation with peroxidase-conjugated goat anti-mouse or rabbit IgG. After washing with PBS, immunoreactivity was detected with diaminobenzidine, followed by hematoxylin counterstaining. Morphometric measurement of the amount of CLGN-positive germ cells was carried out using AxioVision version 4.5 software (Carl Zeiss Co., Ltd., Oberkochen, Germany). The total area of seminiferous tubules within the 1-mm^2 area of the cross-sections from each genotype and treatment group was traced and summed. Then, the CLGN-positive cell numbers were counted and divided by the total area traced for the seminiferous tubules.

Testosterone Assay

Testicular testosterone levels were determined by the enzyme immunoassay (EIA) Kit (Cayman Chemical Co., Ann Arbor, MI). The frozen testis was homogenized in PBS, and the protein concentration was measured by the BCA Protein Assay Kit (Pierce Biotechnology, Inc., Rockford, IL). The homogenate was then extracted with diethyl ether, and the ether phase was air-dried. The dried lipophilic substances were resuspended in the appropriate volume of EIA buffer, and the measurements were done according to the manufacturer's instructions.

Statistical Analysis

For statistical analysis, StatView for Windows version 5.0 (SAS Institute, Cary, NC) was used. All data were expressed relative to the means of the control groups. All results are represented as the mean ± SE. Two-way ANOVA was used for comparison of a given parameter among three control groups ($Ahr^{+/+}$, $Ahr^{+/-}$, $Ahr^{-/-}$), followed by the Fisher PLSD post hoc test. $P < 0.05$ was considered significant.

Downloaded from www.biolreprod.org.

TABLE 1. Reproductive outcomes of male mice exposed to TCDD in utero.[a]

Parameter	Genotype		
	$Ahr^{+/+}$	$Ahr^{+/-}$	$Ahr^{-/-}$
No. of male pups			
Control (n)[b,c]	17 (1.70)	14 (1.40)	6 (0.60)
TCDD (n)[b,d]	6 (0.40)	16 (1.07)	9 (0.60)
Body weight (g)			
Control	6.82 ± 0.41	7.16 ± 0.23	6.04 ± 0.73
TCDD	5.49 ± 0.59	6.59 ± 0.55	7.55 ± 0.27
Crown-anal length (mm)			
Control	54.1 ± 1.2	55.7 ± 0.9	53.0 ± 2.2
TCDD	50.4 ± 2.0	53.1 ± 2.0	56.3 ± 0.9
Anogenital distance (mm)			
Control	3.82 ± 0.11	4.05 ± 0.19	3.66 ± 0.31
TCDD	3.19 ± 0.28*	3.27 ± 0.12**	3.33 ± 0.09
Testicular testosterone level (pg/mg protein)			
Control	140 ± 72 (n = 8)	173 ± 112 (n = 8)	141 ± 86 (n = 5)
TCDD	179 ± 73 (n = 5)	124 ± 33 (n = 8)	197 ± 126 (n = 5)

[a] Data are expressed as means ± SEM, and significant differences were analyzed with ANOVA followed by Fisher PLSD test (versus control of the same genotype, $*P < 0.05$, $**P < 0.01$).
[b] n = The number of male pups per litter.
[c] The number of dams = 10.
[d] The number of dams = 15.

RESULTS

Reproductive Outcome

The number, body weight, and crown-anal length of male pups in each genotype on PND14 are represented in Table 1. There were no statistically significant differences in body weight or crown-anal length among the three genotypes, or between control and TCDD-exposed groups. The number of $Ahr^{+/+}$ male pups exposed to TCDD was only six, which was much lower than the number of control pups, probably due to the fetal death by TCDD exposure in this genotype. Anogenital distance of $Ahr^{+/+}$ and $Ahr^{+/-}$ male pups in the TCDD-exposed groups was significantly reduced compared to control groups ($P = 0.016$ and $P = 0.001$, respectively). In contrast, the anogenital distance of $Ahr^{-/-}$ mice was not significantly different between TCDD-exposed and control mice. ANOVA did not reveal any significant differences in the mean anogenital distance among the three genotypes.

Gene Expressions in UGC

Quantitative RT-PCR analysis of the UGC on PND14 showed that *Ahr* mRNA was detected in $Ahr^{+/+}$ and $Ahr^{+/-}$ mice but not in the $Ahr^{-/-}$ mice. The expression level in the $Ahr^{+/+}$ group was 2-fold higher than that in the $Ahr^{+/-}$ mice, suggesting *Ahr* is transcribed from both alleles in the wild-type mice (Fig. 1). *Cyp1a1*, *Cyp1a2*, and *Cyp1b1* mRNAs, biomarkers of dioxin exposure, were not induced by TCDD

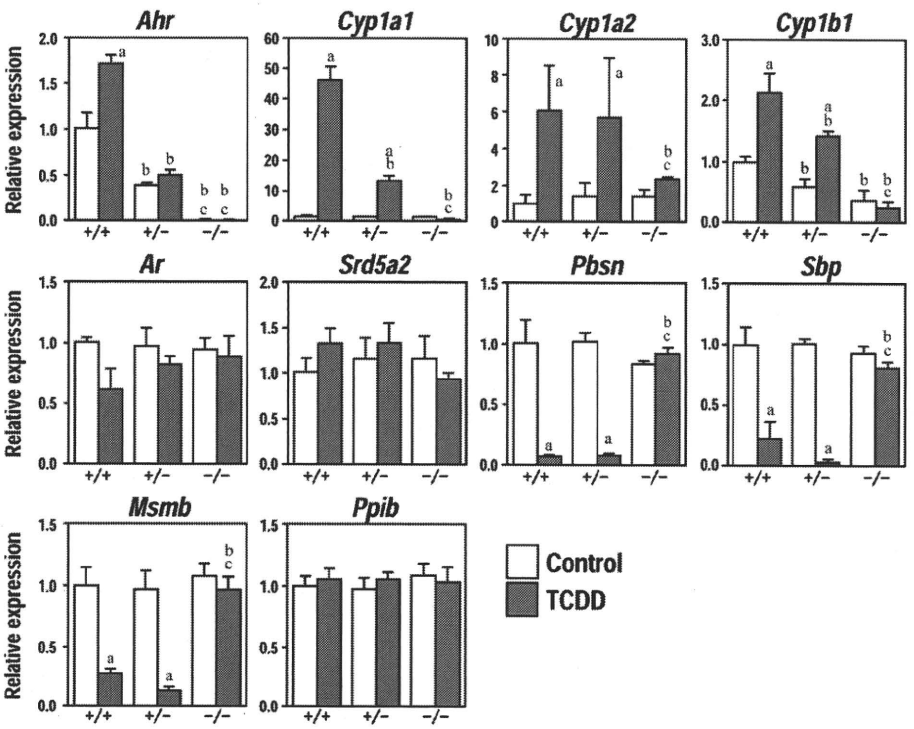

FIG. 1. Quantitative RT-PCR analysis of gene expressions in the urogenital complex of male mouse offspring of three *Ahr* genotypes ($Ahr^{+/+}$, $Ahr^{+/-}$, and $Ahr^{-/-}$) on PND14 with or without TCDD exposure in utero. The values are expressed as mean ± SE for three samples from each group. Note that in utero and lactational exposure to TCDD significantly upregulated *Cyp1a1* and *Cyp1b1* in $Ahr^{+/+}$ and $Ahr^{+/-}$ mice, but not in $Ahr^{-/-}$ mice and that it completely suppressed mRNA expression of the prostatic secretory protein markers *Pbsn*, *Sbp*, and *Msmb* in $Ahr^{+/+}$ and $Ahr^{+/-}$ mice. Significant differences were analyzed with ANOVA followed by the Fisher PLSD test (a, versus control of the same genotype; b, versus the same treatment of wild type; c, versus the same treatment of heterozygous; $P < 0.05$).

Downloaded from www.biolreprod.org.

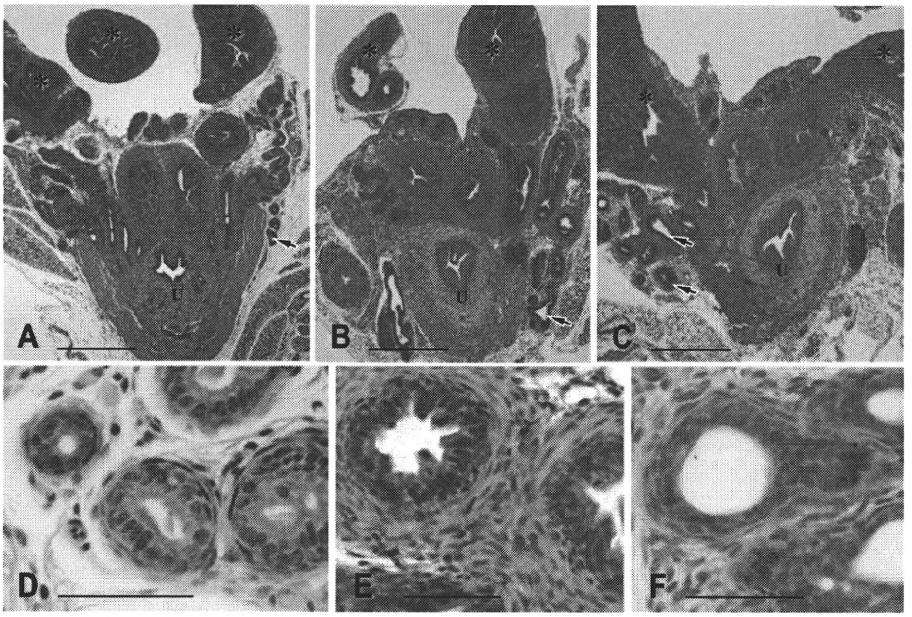

FIG. 2. Histological examinations of the UGC in male mouse pups on PND14. Hematoxylin-eosin staining of $Ahr^{+/+}$ UGCs (**A** and **D**, control; **B**, **C**, **E**, and **F**, TCDD-exposed). Note that control mice had well-developed prostatic lobes on PND14 (**A**). Fewer layers of epithelial cells and increased cell numbers in mesenchymal cells were observed in the dorsolateral prostate lobe of TCDD-exposed animals compared to controls (**E** and **F**). Asterisks, seminal vesicle; arrows, dorsolateral prostate; U, urethra. Bars = 500 μm (**A**, **B**, and **C**), 50 μm (**D**, **E**, and **F**).

in the $Ahr^{-/-}$ mice, but were significantly upregulated in $Ahr^{+/+}$ and $Ahr^{+/-}$ mice. In $Ahr^{+/+}$ and $Ahr^{+/-}$ mice, the *Cyp1b1* expression level in the TCDD-exposed group was higher than that in the control group (Fig. 1). Although a slight decrease of *Ar* mRNA was seen in the $Ahr^{+/+}$ mice (TCDD exposed), we did not detect any significant change in *Ar* and *Srd5a2* mRNA levels among the three TCDD-exposed genotypes. *Pbsn* (dorsolateral), *Sbp* (ventral), and *Msmb* (lateral) were used to investigate functional cytodifferentiation levels of each prostatic epithelia, as reported by others [28]. These three prostate markers were expressed in the control UGCs on PND14. They were barely detectable in TCDD-exposed $Ahr^{+/+}$ and $Ahr^{+/-}$ mice, while TCDD-exposed $Ahr^{-/-}$ mice had quantities of these three marker mRNAs that were very similar to control mice (Fig. 1).

Histopathology of the Urogenital Complex

Prostatic lobes were found to be well developed in the control animals (Fig. 2, A and D). In the TCDD-exposed $Ahr^{+/+}$ animals, the prostatic lobes with existing epithelial layers were clearly observed (Fig. 2, B, C, E, and F). In situ hybridization analysis of the tissue section adjacent to the sections used for histopathological examinations revealed that epithelial cells of dorsolateral prostate lobes had *Pbsn* mRNA signals (Fig. 3A). In accordance with the RT-PCR data, no signals were detected in the epithelia of TCDD-exposed dorsolateral prostates (Fig. 3C).

Gene Expressions in Testis

Consistent with the analysis of UGC (Fig. 1), the expression of *Ahr* mRNA was not detected in the $Ahr^{-/-}$ testis, whereas the *Ahr* expression level was 2-fold higher in the $Ahr^{+/+}$ than in the $Ahr^{+/-}$ mice. No differences were detected in *Ar* mRNA levels (Fig. 4). Among the three CYP1 genes tested, *Cyp1a1* mRNA was not detected in any of the *Ahr* genotypes (data not shown). Although *Cyp1a2* and *Cyp1b1* mRNA was detected in the testis, there were no statistically significant differences among the three genotypes and between the control and TCDD-exposed testes (Fig. 4). The mRNA of *CLGN* and *Hspa2*, male germ cell-specific markers expressed in the pachytene stage of spermatocytes [29, 30], was observed at the same levels among the three genotypes, regardless of TCDD exposure (Fig. 4). No difference was observed in the expression level of *Shbg*, a protein secreted from Sertoli cells [31]. RNA expression levels of four steroidogenic enzyme genes for testosterone synthesis, *Cyp11a1*, *Cyp17a1*, *Hsd3b1*, and *Hsd17b3*, were not affected in the three genotypes under the TCDD dosing regimen used (Fig. 4). Consistently, intratesticular testosterone levels in all genotypes and TCDD-exposed animals were not changed among the three genotypes, regardless of TCDD exposure (Table 1). Additionally the mRNA of *Akr1c4* and *Srd5a1* enzymes for synthesis of 5α-androstane-3α, 17β-diol, the major form of testicular androgen in immature mice [32], was not altered by TCDD exposure and showed no differences among the three genotypes in the testes (Fig. 4).

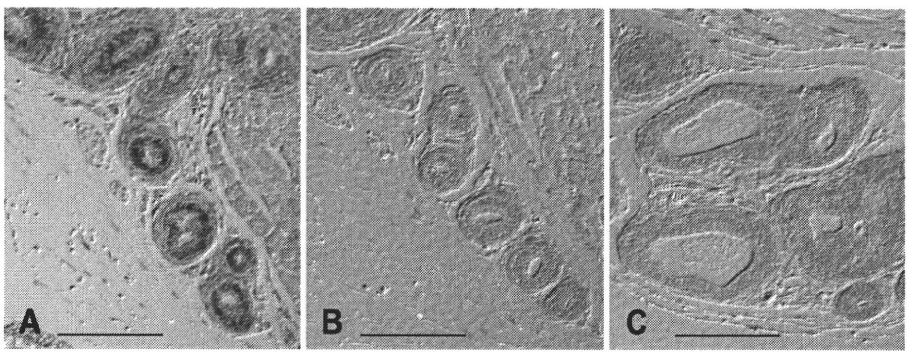

FIG. 3. In situ hybridization analysis of *Pbsn* mRNA expression in the dorsolateral prostate lobe of $Ahr^{+/+}$ mouse UGCs on PND14. **A**) Antisense probe for control mouse. **B**) Sense probe for control mouse. **C**) Antisense probe for TCDD-exposed mouse. Note that epithelial cells of the dorsolateral prostate lobes show *Pbsn* mRNA signals in control mice (**A**), but not in TCDD-exposed mice (**C**). Bar = 200 μm.

Downloaded from www.biolreprod.org.

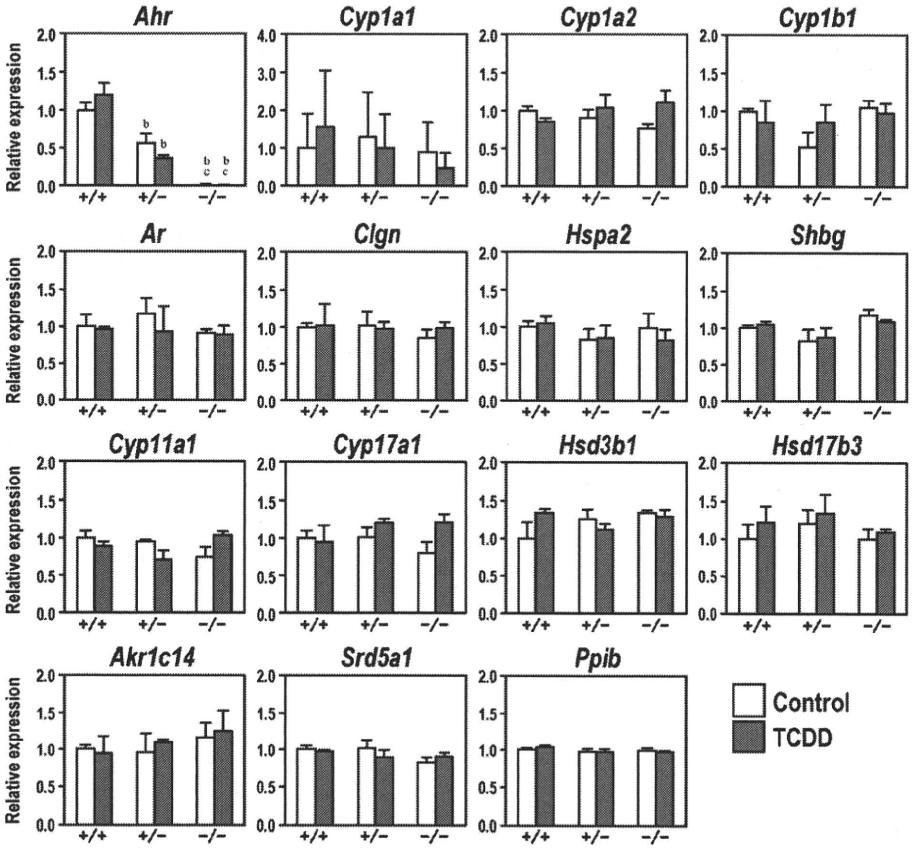

FIG. 4. Quantitative RT-PCR analysis of gene expression in the testes of male pups of three *Ahr* genotypes ($Ahr^{+/+}$, $Ahr^{+/-}$, and $Ahr^{-/-}$) on PND14 with or without TCDD exposure in utero. The values are expressed as the mean ± SE for three samples from each group. Significant differences were analyzed with ANOVA followed by the Fisher PLSD test (b, versus the same treatment of wild-type; c, versus the same treatment of heterozygous; $P < 0.01$).

Histopathology of the Testis

Spermatocytes at the pachytene stage proliferate from the spermatogonium on PND14, and calnexin-t is expressed at this stage [33]. In the present study, germ cells from the three mouse genotypes were immunostained for calnexin-t (Fig. 5, A–C). Positive cell populations were similar in control and TCDD-exposed testes from all genotypes (Fig. 5, D–F).

DISCUSSION

Ahr-*Dependent Reduction in Anogenital Distance and Impairment of Prostatic Development by In Utero and Lactational TCDD Exposure*

In this study, we demonstrated that in utero and lactational TCDD exposure caused reduction of the anogenital distance and impairment of prostatic development in an *Ahr*-dependent manner, because reduction of anogenital distance and disappearance of prostatic epithelial protein mRNAs were observed in $Ahr^{+/+}$ and $Ahr^{+/-}$ but not in the $Ahr^{-/-}$ offspring. The induction of *Cyp1a1* and *Cyp1b1* mRNAs was also observed in $Ahr^{+/+}$ and $Ahr^{+/-}$ mice but not in $Ahr^{-/-}$ mice. Taken together, impairment to the male reproductive system by in utero and lactational TCDD exposure was mainly dependent on the fetal *Ahr* gene.

A study using *Ahr* knockout mice that produced results similar to ours has been reported [17]. In that study, $Ahr^{+/-}$ female and $Ahr^{+/-}$ male mice were mated, and TCDD (5 μg/kg) was injected into the pregnant mice on GD13. The levels of prostatic protein markers were reduced in the TCDD-exposed $Ahr^{+/+}$ mice on PND90, but not in the $Ahr^{-/-}$ mice. Furthermore, those authors reported that TCDD administration on GD13 severely inhibited prostatic bud formation from the urogenital sinus in the fetus [28]. The use of in an vitro organ culture system as well as *Ahr* knockout mice revealed that the inhibition was mediated by AHR expressed in the mesenchy-

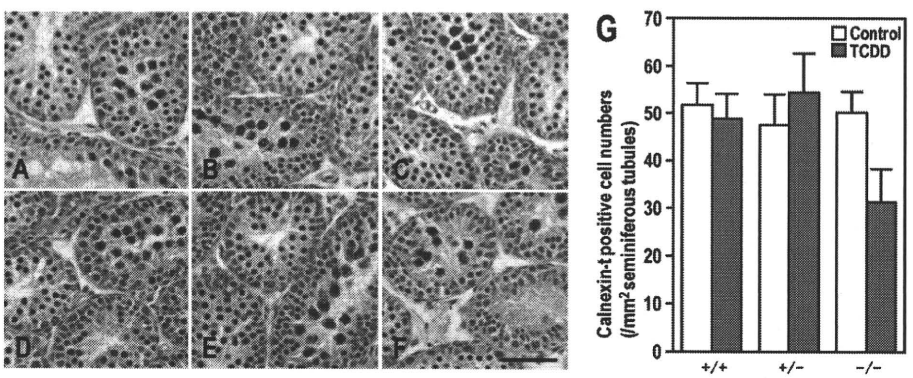

FIG. 5. Immunostaining of CLGN in the testes of male pups on PND14. Testis tissue preparations from each genotype (n = 3) were immunostained. **A**) Control $Ahr^{+/+}$ testis. **B**) Control $Ahr^{+/-}$ testis. **C**) Control $Ahr^{-/-}$ testis. **D**) TCDD-exposed $Ahr^{+/+}$ testis. **E**) TCDD-exposed $Ahr^{+/-}$ testis. **F**) TCDD-exposed $Ahr^{-/-}$ testis. CLGN-positive spermatocytes were observed in all genotypes and TCDD-treatment groups. **G**) Morphometric analysis. There were no statistical differences in the number of positive cells per testis cross section, for all genotypes and treatments. Bar = 100 μm.

Downloaded from www.biolreprod.org.

mal cells of the urogenital sinus [34] and was not caused by interruption of androgen signaling in this tissue [35]. It is still unclear how TCDD impairs the responsiveness of the developing prostate to androgen.

In Utero and Lactational TCDD Exposure Decreased Androgen Responsiveness of the Prostate

In the histological examinations, the prostatic lobes developed with obvious epithelial layers in TCDD-exposed $Ahr^{+/+}$ mice, indicating TCDD-exposed prostates may function as normal exocrine glands (Fig. 2, B and C). However, using RT-PCR and in situ hybridization analyses, we could not detect prostatic epithelial secretory protein mRNAs in the TCDD-exposed UGCs (Figs. 1 and 3). Fewer layers of epithelial cells and increased cell numbers in mesenchymal cells were observed in the TCDD-exposed $Ahr^{+/+}$ animals, suggesting that in utero and lactational exposure to TCDD produced functional abnormalities (Fig. 2, E and F). The prostatic secretory proteins, PBSN, MSMB, and SBP, were reportedly upregulated via the androgen receptor [36–38]. Since no significant differences in intratesticular testosterone levels were found between control and TCDD-exposed animals on PND14, we speculated that abnormal development of prostate glands may be due to decreased androgen sensitivity or that TCDD disrupts mouse prostate epithelial cell differentiation into luminal epithelial cells. This notion is consistent with a previous study [9] in the sense that the ventral prostate of male rat offspring exposed to TCDD in utero and via lactation did not respond to the exogenous androgens testosterone, 5α-dihydrotestosterone (DHT), and 5α-androstane-3α, 17β-diol, in the organ culture system. Administration of the androgen receptor antagonist flutamide and the 5α-reductase inhibitor finasteride to rats in the late pregnancy period did not cause prostate growth in male offspring on PND60 [39], suggesting that DHT is an essential steroid hormone for prostate development. Since the inhibition was found specifically in the prostate but not in the seminal vesicle, the reduction in DHT production in the prostate was initially hypothesized to occur in males exposed to TCDD in utero and via lactation. However, 5α-reductase type 2 enzymatic activity and mRNA expression was elevated, compared to control groups [9, 10]. Thus, it is reasonable to speculate that decreased androgen responsiveness in the TCDD-exposed offspring was caused by other factors. Our previous study using Holtzman rats showed significantly reduced androgen receptor mRNA expression in the ventral prostate following in utero and lactational TCDD exposure, suggesting that the decreased androgen responsiveness might be due to reduced amounts of receptor molecules [10]. However, we could not detect a significant reduction in androgen receptor mRNA here, probably due to a difference in animal species or organs used for RT-PCR analysis.

Function of AHR in the Testis

AHR was reportedly responsible for apoptotic signaling, and the number of primordial cells in the ovarian germ line was not attenuated due to a defect in the apoptosis in *Ahr*-null female mice [20]. More recently, testicular dysfunction was reported in aged *Ahr*-null mice [40], and HSD3B1 expression in Leydig cells was significantly reduced at 24 wk, resulting in serum testosterone decline, lowered sperm number, and reduced size of seminal vesicles. AHR seems to play a role in maintaining normal steroidogenesis in aged animals. However, in that study, there were no significant differences in testicular functions between wild-type and *Ahr*-null mice during younger stages (10 wk old). In our present study, we were unable to find differences in testicular functions among the *Ahr* genotypes, including testosterone production, Sertoli cell differentiation, and spermatogenic cell differentiation. Therefore, it is reasonable to conclude that AHR has very little function in early stages of gonad development. If the AHR functions even at the early stage of development, the functional redundancy with other genes may also exist among *Ahr* and other genes during testicular development.

Resistance of Testicular Development to In Utero and Lactational TCDD Exposure

In our previous study, TCDD administration to pregnant Holtzman rats on GD15 did not alter the testicular weight or serum testosterone levels of male offspring on PND49 [10]. In our present study, we did not find any differences in testicular cell differentiation levels between control *Ahr*-carrying animals and TCDD-exposed animals that survived the high dose of TCDD exposure, including supporting cell marker and spermatogenic cell differentiation markers, which suggested that even in the TCDD-exposed animals testicular differentiation proceeded normally. 5α-androstane-3α, 17β-diol, the major form of testicular androgen in immature mice [32], was previously reported to be slightly reduced in PND21 mice perinatally exposed to TCDD [17]. However, in our present study, in the assayed testes from PND14, no difference was seen in both *Akr1c4* and *Srd5a1* mRNA expression levels among the three genotypes and TCDD treatment. Moreover, in utero and lactational exposure to TCDD did not alter the expression levels of steroidogenic enzyme genes in Leydig cells in $Ahr^{+/+}$ and $Ahr^{+/-}$ animals. At a relatively high dose, testicular CYP11A1 activity was reduced by TCDD exposure [41]. In our previous report, we also found that administration of 100 μg TCDD/kg to adult male mice reduced testicular *Cyp11a1* mRNA and protein levels [42], and in vitro co-planer PCB (3,3′,4,4′,5-pentachlorobiphenyl; PCB126) exposure to neonatal mouse testis downregulated *Cyp11a1* mRNA expression [33]. The reason why we could not detect the reduction in *Cyp11a1* mRNA in the present study is not clear, but it is speculated that intratesticular levels of TCDD in male pups born to dams given TCDD on GD13 was not sufficient to downregulate *Cyp11a1* by PND14.

Although testes and UGCs were collected from the same individual pups, both *Cyp1a2* and *Cyp1b1* mRNAs were not induced in the testes of mice with any of the three genotypes, whereas in the UGCs, *Cyp1a2* and *Cyp1b1* were significantly induced. Moreover, an approximately 30-fold increase in *Cyp1a1* was observed in the $Ahr^{+/+}$ UGCs. Thus, UGCs are much more sensitive to TCDD than testis in terms of *Cyp1a1* mRNA induction. Using a xenobiotic-responsive element connected to the β-galactocidase reporter gene, a transgenic mouse line was generated and then exposed to TCDD in utero and via lactation [43]. X-Gal staining analysis clearly demonstrated that fetal urogenital tracts showed significant induction of the reporter gene, but that the testis did not respond. Based on the above-mentioned results, we speculated that UGCs express modulating factors (modifiers) that enhance AHR-mediated transcription, and that these were lacking or present in small amounts in the testis. Although it cannot be excluded that tissue concentration of TCDD in the testis was lower than that in the UGC, it is more likely that the testis is much more resistant to TCDD exposure than the UGC.

In conclusion, using *Ahr* knockout mice, we confirmed that in utero exposure to TCDD caused *Ahr*-dependent impairment of prostate development and reduced anogenital distances in

Downloaded from www.biolreprod.org.

male offspring, but that the testicular development seemed to be resistant to TCDD exposure. AHR was not associated with testicular development under physiological conditions.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the technical support of Dr. Hiro Nitta (Ventana Medical System, Tucson, AZ) in the in situ hybridization analysis and the technical help by Miss Fumi Kido with RT-PCR analysis. The authors also thank Dr. Yoshiaki Fujii-Kuriyama for providing *Ahr* knockout mice.

REFERENCES

1. Birnbaum LS, Tuomisto J. Non-carcinogenic effects of TCDD in animals. Food Addit Contam 2000; 17:275–288.
2. Mably TA, Moore RW, Peterson RE. In utero and lactational exposure of male rats to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin. 1. Effects on androgenic status. Toxicol Appl Pharmacol 1992; 114:97–107.
3. Mably TA, Bjerke DL, Moore RW, Gendron-Fitzpatrick A, Peterson RE. In utero and lactational exposure of male rats to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin. 3. Effects on spermatogenesis and reproductive capability. Toxicol Appl Pharmacol 1992; 114:118–126.
4. Bjerke DL, Peterson RE. Reproductive toxicity of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin in male rats: different effects of in utero versus lactational exposure. Toxicol Appl Pharmacol 1994; 127:241–249.
5. Faqi AS, Dalsenter PR, Merker HJ, Chahoud I. Effects on developmental landmarks and reproductive capability of 3,3′,4,4′-tetrachlorobiphenyl and 3,3′,4,4′,5-pentachlorobiphenyl in offspring of rats exposed during pregnancy. Hum Exp Toxicol 1998; 17:365–372.
6. Gray LE, Kelce WR, Monosson E, Ostby JS, Birnbaum LS. Exposure to TCDD during development permanently alters reproductive function in male Long Evans rats and hamsters: reduced ejaculated and epididymal sperm numbers and sex accessory gland weights in offspring with normal androgenic status. Toxicol Appl Pharmacol 1995; 131:108–118.
7. Gray LE, Ostby JS, Kelce WR. A dose-response analysis of the reproductive effects of a single gestational dose of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin in male Long Evans hooded rat offspring. Toxicol Appl Pharmacol 1997; 146:11–20.
8. Roman BL, Timms BG, Prins GS, Peterson RE. In utero and lactational exposure of the male rat to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin impairs prostate development. 2. Effects on growth and cytodifferentiation. Toxicol Appl Pharmacol 1998; 150:254–270.
9. Theobald HM, Roman BL, Lin TM, Ohtani S, Chen SW, Peterson RE. 2,3,7,8-Tetrachlorodibenzo-*p*-dioxin inhibits luminal cell differentiation and androgen responsiveness of the ventral prostate without inhibiting prostatic 5alpha-dihydrotestosterone formation or testicular androgen production in rat offspring. Toxicol Sci 2000; 58:324–338.
10. Ohsako S, Miyabara Y, Nishimura N, Kurosawa S, Sakaue M, Ishimura R, Sato M, Takeda K, Aoki Y, Sone H, Tohyama C, Yonemoto J. Maternal exposure to a low dose of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) suppressed the development of reproductive organs of male rats: dose-dependent increase of mRNA levels of 5alpha-reductase type 2 in contrast to decrease of androgen receptor in the pubertal ventral prostate. Toxicol Sci 2001; 60:132–143.
11. Roman BL, Sommer RJ, Shinomiya K, Peterson RE. In utero and lactational exposure of the male rat to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin: impaired prostate growth and development without inhibited androgen production. Toxicol Appl Pharmacol 1995; 134:241–250.
12. Ko K, Theobald HM, Peterson RE. In utero and lactational exposure to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin in the C57BL/6J mouse prostate: lobe-specific effects on branching morphogenesis. Toxicol Sci 2002; 70: 227–237.
13. Swanson HI, Bradfield CA. The AHReceptor: genetics, structure and function. Pharmacogenetics 1993; 3:213–230.
14. Gu YZ, Hogenesch JB, Bradfield CA. The PAS superfamily: sensors of environmental and developmental signals. Annu Rev Pharmacol Toxicol 2000; 40:519–561.
15. Mimura J, Yamashita K, Nakamura K, Morita M, Takagi TN, Nakao K, Ema M, Sogawa K, Yasuda M, Katsuki M, Fujii-Kuriyama Y. Loss of teratogenic response to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) in mice lacking the Ah (dioxin) receptor. Genes Cells 1997; 2:645–654.
16. Peters JM, Narotsky MG, Elizondo G, Fernandez-Salguero PM, Gonzalez FJ, Abbott BD. Amelioration of TCDD-induced teratogenesis in aryl hydrocarbon receptor (AHR)-null mice. Toxicol Sci 1999; 47:86–92.
17. Lin TM, Ko K, Moore RW, Simanainen U, Oberley TD, Peterson RE. Effects of aryl hydrocarbon receptor null mutation and in utero and lactational 2,3,7,8-tetrachlorodibenzo-*p*-dioxin exposure on prostate and seminal vesicle development in C57BL/6 mice. Toxicol Sci 2002; 68:479–487.
18. Fernandez-Salguero P, Pineau T, Hilbert DM, McPhail T, Lee SS, Kimura S, Nebert DW, Rudikoff S, Ward JM, Gonzalez FJ. Immune system impairment and hepatic fibrosis in mice lacking the dioxin-binding AHReceptor. Science 1995; 268:722–726.
19. Schmidt JV, Su GH, Reddy JK, Simon MC, Bradfield CA. Characterization of a murine AHR null allele: involvement of the AHReceptor in hepatic growth and development. Proc Natl Acad Sci U S A 1996; 93: 6731–6736.
20. Robles R, Morita Y, Mann KK, Perez GI, Yang S, Matikainen T, Sherr DH, Tilly JL. The aryl hydrocarbon receptor, a basic helix-loop-helix transcription factor of the PAS gene family, is required for normal ovarian germ cell dynamics in the mouse. Endocrinology 2000; 141:450–453.
21. Benedict JC, Lin TM, Loeffler IK, Peterson RE, Flaws JA. Physiological role of the aryl hydrocarbon receptor in mouse ovary development. Toxicol Sci 2000; 56:382–388.
22. Benedict JC, Miller KP, Lin TM, Greenfeld C, Babus JK, Peterson RE, Flaws JA. Aryl hydrocarbon receptor regulates growth, but not atresia, of mouse preantral and antral follicles. Biol Reprod 2003; 68:1511–1517.
23. Baba T, Mimura J, Nakamura N, Harada N, Yamamoto M, Morohashi K, Fujii-Kuriyama Y. Intrinsic function of the aryl hydrocarbon (dioxin) receptor as a key factor in female reproduction. Mol Cell Biol 2005; 25: 10040–10051.
24. Schultz R, Suominen J, Varre T, Hakovirta H, Parvinen M, Toppari J, Pelto-Huikko M. Expression of aryl hydrocarbon receptor and aryl hydrocarbon receptor nuclear translocator messenger ribonucleic acids and proteins in rat and human testis. Endocrinology 2003; 144:767–776.
25. Ohtake F, Baba A, Takada I, Okada M, Iwasaki K, Miki H, Takahashi S, Kouzmenko A, Nohara K, Chiba T, Fujii-Kuriyama Y, Kato S. Dioxin receptor is a ligand-dependent E3 ubiquitin ligase. Nature 2007; 446:562–566.
26. Shiizaki K, Ohsako S, Koyama T, Nagata R, Yonemoto J, Tohyama C. Lack of CYP1A1 expression is involved in unresponsiveness of the human hepatoma cell line SK-HEP-1 to dioxin. Toxicol Lett 2005; 160:22–33.
27. Ohsako S, Janulis L, Hayashi Y, Bunick D. Characterization of domains in mice of calnexin-t, a putative molecular chaperone required in sperm fertility, with use of glutathione S-transferase-fusion proteins. Biol Reprod 1998; 59:1214–1223.
28. Lin TM, Rasmussen NT, Moore RW, Albrecht RM, Peterson RE. Region-specific inhibition of prostatic epithelial bud formation in the urogenital sinus of C57BL/6 mice exposed in utero to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin. Toxicol Sci 2003; 76:171–181.
29. Ohsako S, Hayashi Y, Bunick D. Molecular cloning and sequencing of calnexin-t. An abundant male germ cell-specific calcium-binding protein of the endoplasmic reticulum. J Biol Chem 1994; 269:14140–14148.
30. Zakeri ZF, Wolgemuth DJ. Developmental-stage-specific expression of the hsp70 gene family during differentiation of the mammalian male germ line. Mol Cell Biol 1987; 7:1791–1796.
31. Wang YM, Sullivan PM, Petrusz P, Yarbrough W, Joseph DR. The androgen-binding protein gene is expressed in CD1 mouse testis. Mol Cell Endocrinol 1989; 63:85–92.
32. Mahendroo M, Wilson JD, Richardson JA, Auchus RJ. Steroid 5alpha-reductase 1 promotes 5alpha-androstane-3alpha,17beta-diol synthesis in immature mouse testes by two pathways. Mol Cell Endocrinol 2004; 222: 113–120.
33. Fukuzawa NH, Ohsako S, Nagano R, Sakaue M, Baba T, Aoki Y, Tohyama C. Effects of 3,3′,4,4′,5-pentachlorobiphenyl, a coplanar polychlorinated biphenyl congener, on cultured neonatal mouse testis. Toxicol In Vitro 2003; 17:259–269.
34. Ko K, Moore RW, Peterson RE. Aryl hydrocarbon receptors in urogenital sinus mesenchyme mediate the inhibition of prostatic epithelial bud formation by 2,3,7,8-tetrachlorodibenzo-*p*-dioxin. Toxicol Appl Pharmacol 2004; 196:149–155.
35. Ko K, Theobald HM, Moore RW, Peterson RE. Evidence that inhibited prostatic epithelial bud formation in 2,3,7,8-tetrachlorodibenzo-*p*-dioxin-exposed C57BL/6J fetal mice is not due to interruption of androgen signaling in the urogenital sinus. Toxicol Sci 2004; 79:360–369.
36. Zhang J, Gao N, Kasper S, Reid K, Nelson C, Matusik RJ. An androgen-dependent upstream enhancer is essential for high levels of probasin gene expression. Endocrinology 2004; 145:134–148.
37. Mills JS, Needham M, Parker MG. Androgen regulated expression of a spermine binding protein gene in mouse ventral prostate. Nucleic Acids Res 1987; 15:7709–7724.
38. Xuan JW, Kwong J, Chan FL, Ricci M, Imasato Y, Sakai H, Fong GH,

Downloaded from www.biolreprod.org.

Panchal C, Chin JL. cDNA, genomic cloning, and gene expression analysis of mouse PSP94 (prostate secretory protein of 94 amino acids). DNA Cell Biol 1999; 18:11–26.
39. Imperato-McGinley J, Sanchez RS, Spencer JR, Yee B, Vaughan ED. Comparison of the effects of the 5alpha-reductase inhibitor finasteride and the antiandrogen flutamide on prostate and genital differentiation: dose-response studies. Endocrinology 1992; 131:1149–1156.
40. Baba T, Shima Y, Owaki A, Mimura J, Oshima M, Fujii-Kuriyama Y, Morohashi KI. Disruption of aryl hydrocarbon receptor (AHR) induces regression of the seminal vesicle in aged male mice. Sex Dev 2008; 2:1–11.
41. Moore RW, Jefcoate CR, Peterson RE. 2,3,7,8-Tetrachlorodibenzo-*p*-dioxin inhibits steroidogenesis in the rat testis by inhibiting the mobilization of cholesterol to cytochrome P450scc. Toxicol Appl Pharmacol 1991; 109:85–97.
42. Fukuzawa NH, Ohsako S, Wu Q, Sakaue M, Fujii-Kuriyama Y, Baba T, Tohyama C. Testicular cytochrome P450scc and LHR as possible targets of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) in the mouse. Mol Cell Endocrinol 2004; 221:87–96.
43. Willey JJ, Stripp BR, Baggs RB, Gasiewicz TA. Aryl hydrocarbon receptor activation in genital tubercle, palate, and other embryonic tissues in 2,3,7,8-tetrachlorodibenzo-*p*-dioxin-responsive lacZ mice. Toxicol Appl Pharmacol 1998; 151:33–44.

Downloaded from www.biolreprod.org.

OPEN ACCESS Freely available online

PLoS one

PeakRegressor Identifies Composite Sequence Motifs Responsible for STAT1 Binding Sites and Their Potential rSNPs

Jean-François Pessiot[1], Hirokazu Chiba[1], Hiroto Hyakkoku[1,2], Takeaki Taniguchi[3], Wataru Fujibuchi[1]*

1 Computational Biology Research Center, Advanced Industrial Science and Technology (AIST), Tokyo, Japan, **2** Waseda University, Tokyo, Japan, **3** Mitsubishi Research Institute, Inc., Tokyo, Japan

Abstract

How to identify true transcription factor binding sites on the basis of sequence motif information (e.g., motif pattern, location, combination, etc.) is an important question in bioinformatics. We present "PeakRegressor," a system that identifies binding motifs by combining DNA-sequence data and ChIP-Seq data. PeakRegressor uses L1-norm log linear regression in order to predict peak values from binding motif candidates. Our approach successfully predicts the peak values of STAT1 and RNA Polymerase II with correlation coefficients as high as 0.65 and 0.66, respectively. Using PeakRegressor, we could identify composite motifs for STAT1, as well as potential regulatory SNPs (rSNPs) involved in the regulation of transcription levels of neighboring genes. In addition, we show that among five regression methods, L1-norm log linear regression achieves the best performance with respect to binding motif identification, biological interpretability and computational efficiency.

Citation: Pessiot J-F, Chiba H, Hyakkoku H, Taniguchi T, Fujibuchi W (2010) PeakRegressor Identifies Composite Sequence Motifs Responsible for STAT1 Binding Sites and Their Potential rSNPs. PLoS ONE 5(8): e11881. doi:10.1371/journal.pone.0011881

Editor: Xiaolin Wu, National Cancer Institute at Frederick, United States of America

Received January 15, 2010; **Accepted** June 7, 2010; **Published** August 27, 2010

Copyright: © 2010 Pessiot et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by general subsidies from the National Institute of Advanced Industrial Science and Technology, Japan. This funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Takeaki Taniguchi is employed by the Mitsubishi Research Institute, Inc. His wages were funded by the company and he participated in performing the computational experiments. This funder also had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Takeaki Taniguchi is employed by the Mitsubishi Research Institute, Inc. There are no patents, products in development, or marketed products related to this research, and his involvement does not alter the adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: w.fujibuchi@aist.go.jp

Introduction

The experimental identification of *cis*-regulatory sites based on transcription factor binding motifs (TFBMs) is a difficult and time-consuming task. In this regard, *in silico* analysis of TFBMs has recently attracted attention as a promising tool for discovering true *cis*-regulatory sites. Previous works attempt to find TFBMs to model the mechanisms underlying the control of gene expression levels [1,2]. They assume that the gene expression levels are determined by the presence of certain motifs in the upstream regions of the genes. Based on this assumption, they find TFBM candidates which show a strong correlation with changes in the gene expression levels. [3] Instead of modeling the expression levels, another solution is to model the binding affinities between a protein and its target genes based on the thermodynamics theory. However, the binding affinities are difficult to measure and related works use transcription factor occupancy to approximate binding affinity [4,5].

In this article, we present PeakRegressor, a new tool for the identification of functional TFBMs from ChIP-Seq data. As far as we know, this is the first attempt at performing peak signal regression based on candidate motif models. Because PeakRegressor is computationally efficient and the models are easy to interpret, it is usable with large-scale datasets. We apply PeakRegressor to two ChIP-Seq datasets and show its ability to recover motifs involved in the binding of STAT1 and RNA Polymerase II.

Results and Discussion

Results with PeakRegressor

Table 1 shows the correlation coefficients between the peak scores and their predicted values by PeakRegressor in the test dataset. We keep the highest correlation coefficient among various β for each iteration of the 30-fold cross-validation, and those 30 correlation coefficients are averaged and shown here. Obviously, the filtering with peak existence probability, i.e., Q-value, over the control experiment enhances the regressions. The filtering with promoter region proximity improves the regressions of RNA Polymerase II but not of STAT1.

In Figure 1, we plot the STAT1 peak scores with two filtering methods such as Q-value $<10^{-3}$ and promoter proximity in the test dataset against their predictions by PeakRegressor. The correlation coefficient is as high as 0.65 between the peak and predicted values for the Q-value filtering, whilst it is as low as 0.41 for promoter proximity filtering. Interestingly, however, the data points that are selected by promoter proximity existed only in a biased region, leading to worse prediction.

In Tables 2 and 3, we show the top ten motifs for STAT1 and RNA Polymerase II identified by PeakRegressor, respectively. The motifs are sorted according to the absolute values of their averaged regression coefficients. A motif with a positive (resp. negative) coefficient is thought to have a strengthening (resp. weakening) effect on the binding. In the case of STAT1,

Table 1. Influence of the peak filtering methods on the correlation coefficients between peak values and their predicted values in the test dataset.

Filtering method	*#Peaks (STAT1/Pol II)*	*STAT1*	*Pol II*
None	36998/24739	0.50	0.44
Promoter proximity	3,907/9,094	0.41	0.53
Q-value $<10^{-3}$	16639/17580	**0.65**	**0.66**

The correlation coefficients are averaged in 30-fold cross-validation.
doi:10.1371/journal.pone.0011881.t001

it is clear that our approach correctly identifies the classical GAS motif TTC[TC]N[GA]GAA as the main binding motif [6]. Meanwhile, the RNA Polymerase II binding motifs also contain known Downstream Promoter Element [AG]G[AT][CT][GAC] and Initiator Site [TC][TC]AN[TA][TC][TC] [7].

STAT1 composite motifs. As the most important feature of PeakRegressor, it can give us a list of putative composite motifs. Basically, it is difficult to evaluate whether a composite motif consists of the same motif or multiple (different) motifs. In order to identify the composite motifs, we proceed as follows. First, we consider the best set of motifs according to PeakRegressor (i.e., the set which corresponds to the best prediction accuracy). Among these, we select 136 motifs which have a normalized coefficient higher than 0.1. We use these motifs to represent each peak sequence as a binary vector, indicating whether a motif is present or not in the peak sequence. Then we cluster the resulting peak vectors using the K-Means algorithm. Thus each cluster contains peak vectors which show similar motif patterns, i.e., sequences containing potential composite motifs.

Here we show an example of a composite motif that are responsible for STAT1 binding signals:

TCACA[TG]G[ACG] + [TC]TT[CA]C[CA][AG][GC][AC]A.

Comparison with other regression methods

PeakRegressor identifies potential TFBMs by solving a regression problem. This regression problem is defined by a set of peak vectors $\{\mathbf{x_i}\}_{i=1...N}$ and their corresponding peak scores $\{y_i\}_{i=1...N}$. The goal is to predict the peak scores from the peak vectors. The fitted regression model is then used to infer the TFBM candidates. We expect the regression method to have three properties. First, it should identify the true binding motifs. Second, it should identify the strengthening and weakening motifs. Third, it should be computationally efficient in order to cope with large ChIP-Seq datasets.

In PeakRegressor, we choose to use the L1-norm log linear regression to solve this problem. This approach favors sparse solutions (i.e., solutions with a small number of motifs) and therefore, we argue that it is more suitable for the TFBM identification problem. However, many other regression methods are available and can be used to solve the regression problem. How do these approaches compare with the L1-norm log linear regression with respect to the desired properties? In the following, we compare our L1-norm log linear regression based approach with other regression methods: linear least squares regression, ridge regression, partial least squares regression, and principal component regression. For each method, we evaluate its performance on the STAT1 and RNA Polymerase II datasets and discuss the results.

Linear least squares regression. In Tables 4 and 5, we show the top ten motifs identified by the linear least squares regression. In the case of STAT1 (Table 4), we can see that the true GAS motif appears within the top ten motifs. However, two problems appear. First, the regression coefficients of the GAS

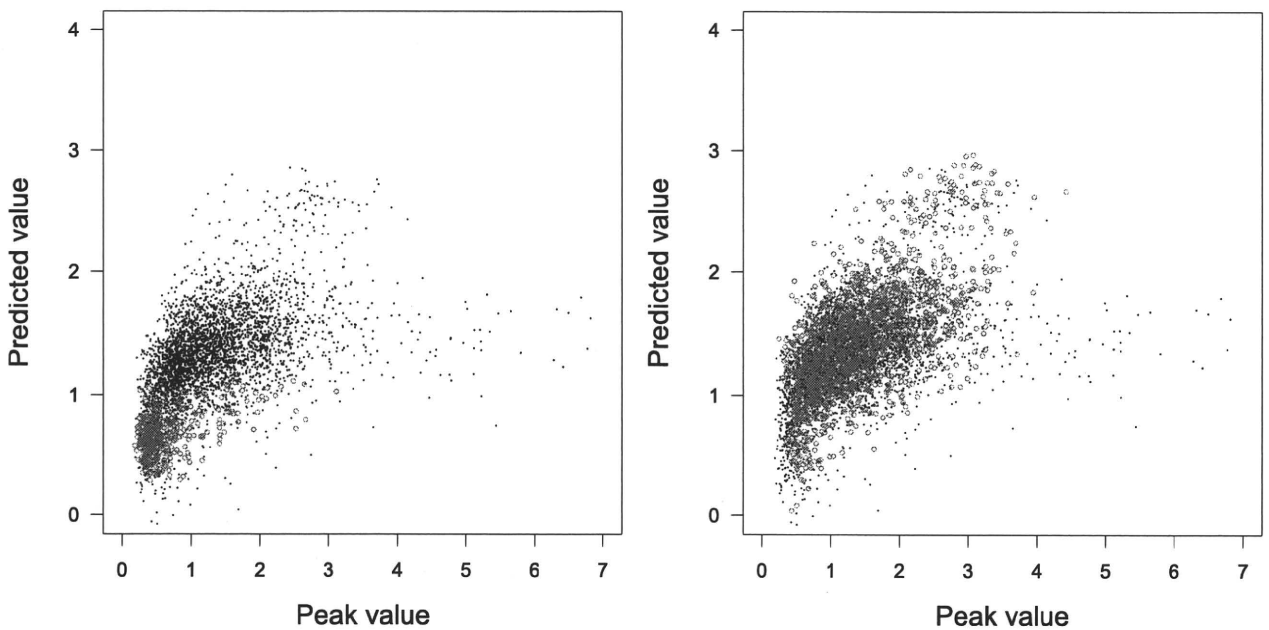

Figure 1. STAT1 regression results with two filtering methods: Q-value (right) and promoter proximity (left). The correlation coefficients on the test data between peak values and their predicted values are 0.65 and 0.41 for Q-value and promoter proximity filterings, respectively.
doi:10.1371/journal.pone.0011881.g001

Table 2. List of putative STAT1 binding motifs identified by PeakRegressor.

STAT1	*Normalized coef.*
CA[TC]GTGACT[TG]C	1.
[TG]G[GTA][GC][AG]T**TT[CA]C[AGC] [GA]GAA**[AC][TG]G[GA][GC]	0.96
TTC[CT][TG][GA]GAAAT[GC][CA] [CA][CAT][AT][TCG][CG][CT]	0.72
[CT][TC]CA[GT]**TTCCAGGAA**[AT]T[CG][CAT]C[CT]	0.65
GGAGGGCG	−0.57
GGACGCCG	−0.56
A[CT]**TTC[TC][TG]GGAA**	0.56
TT[CA]C[TAG][GA]GAA[GA]T	0.55
A[TA]**TTCC[CT][GA]GAA**[AC]T[CG][AC]	0.48
TT[CA][TC][GA]GGAA[AG]	0.47

The classical GAS motifs are shown in boldface.
doi:10.1371/journal.pone.0011881.t002

motif are very low compared to those of the top motifs (between 0.01 and 0.02). This means that according to the linear least squares regression, the true GAS motif has only a minor effect on the binding, which contradicts existing biological knowledge. Second, the most important motifs according to the linear least squares regression are CCCCTCCC and CCCCACCC. However, each of them is associated with opposite coefficients (−1.0 and 0.94 for CCCCTCCC, 0.34 and −0.34 for CCCCACCC). Therefore, each of them is considered to have both a strengthening effect and a weakening effect on the binding, which is a contradictory result.

With the RNA Polymerase II dataset (Table 5), linear least squares regression is able to identify the initiator site and the downstream promoter element. However, the instances of the initiator site have opposite coefficients ([CA]CAGACT with 0.62, T[CT][TA]T[TG][AC][AT] with 0.62, and TT[TAC]TTT[CT] with −0.61). As they are instances of the same motif, we expect them to have the same sign i.e., to have the same effect on the binding. In summary, for both STAT1 and RNA Polymerase II

Table 3. List of putative RNA Polymerase II binding motifs identified by PeakRegressor.

Pol II	*Normalized coef.*
T[AG]**A[GC][TAG]CA**[GCT]A[AC]AA	1.
A[GA]AA[AC][CA]AA[AC]AAA	0.78
C[ACT]**[GT][CG][CT][TA]C**C[AGT]CC[TA]	0.76
C[CT][CG][AT]GGCTGG[AG]G	0.68
TTTCTGC[CT][CT]TT[GT]	0.67
T[TA]T[TC]**[CA]CAGACT**[AT]	0.63
GGAGGGAGGC[AG]G	0.62
AC[AC][CA][AC][AT][AG]AGAAA	0.61
TTTG**T[CT][TA]T[TG][AC][AT]**T	0.54
AAA[AT][GC]AAA[AT]A[GA]A	0.54

The known Downstream Promoter Element and Initiator site motifs are shown in boldface.
doi:10.1371/journal.pone.0011881.t003

Table 4. List of putative STAT1 binding motifs identified by linear least squares regression.

STAT1	*Normalized coef.*
CCCCTCCC	−1.0
CCCCTCCC	0.94
CCCCACCC	0.34
CCCCACCC	−0.34
CA[TC]GTGACT[TG]C	0.02
[TG]G[GTA][GC][AG]T**TT[CA]C[AGC] [GA]GAA**[AC][TG]G[GA][GC]	0.02
[CT][TC]CA[GT]**TTCCAGGAA**[AT]T[CG][CAT]C[CT]	0.01
GGAGGGCG	−0.01
TTC[CT][TG][GA]GAAAT[GC][CA][CA] [CAT][AT][TCG][CG][CT]	0.01
A[CT]**TTC[TC][TG]GGAA**	0.01

The classical GAS motifs are shown in boldface.
doi:10.1371/journal.pone.0011881.t004

datasets, the results of the linear least squares regression are difficult to interpret biologically. This is a typical situation where we would like to reduce the number of motifs used by the regression model. Clearly, this is not possible with the linear least squares regression approach.

Ridge regression. In Tables 6 and 7, we show the top ten motifs identified by the ridge regression. In the case of STAT1 (Table 6), we can see that the ridge regression and the L1-norm log linear regression identify very similar motifs. In both cases, the classical GAS motif is clearly identified as the main binding motif. Both regression methods also identify CA[TC]GTGACT[TG]C as a strengthening motif and GGAGGGCG as a weakening motif. In the case of RNA Polymerase II (Table 7), both methods are able to identify the initiator site (T[CT][TA]T[TG][AC][AT) and the downstream promoter element (A[GC][TAG]CA).

However, they differ greatly with respect to computational complexity. In [8], the authors present an algorithm for computing the L1-norm log linear regression solutions of many regularization parameters for the same computational cost as that of a single least

Table 5. List of putative RNA Polymerase II binding motifs identified by linear least squares regression.

RNA Polymerase II	*Normalized coef.*
T[AG]**A[GC][TAG]CA**[GCT]A[AC]AA	1.0
A[GA]AA[AC][CA]AA[AC]AAA	0.86
C[ACT]**[GT][CG][CT][TA]C**C[AGT]CC[TA]	0.81
C[CT][CG][AT]GGCTGG[AG]G	0.74
TTTCTGC[CT][CT]TT[GT]	0.74
GGAGGGAGGC[AG]G	0.69
AC[AC][CA][AC][AT][AG]AGAAA	0.64
T[TA]T[TC]**[CA]CAGACT**[AT]	0.62
TTTG**T[CT][TA]T[TG][AC][AT]**T	0.62
TT[TAC]TTT[CT]TT[TC]TT	−0.61

The known Downstream Promoter Element and Initiator site motifs are shown in boldface.
doi:10.1371/journal.pone.0011881.t005

Table 6. List of putative STAT1 binding motifs identified by ridge regression.

STAT1	*Normalized coef.*
CA[TC]GTGACT[TG]C	1.
[TG]G[GTA][GC][AG]T**TT[CA]C[AGC] [GA]GAA**[AC][TG]G[GA][GC]	0.89
GGAGGGCG	−0.69
[CT][TC]CA[GT]**TTCCAGGAA**[AT]T[CG][CAT]C[CT]	0.69
A[CT]**TTC[TC][TG]GGAA**	0.68
TTC[CT][TG][GA]GAAAT[GC][CA][CA] [CAT][AT][TCG][CG][CT]	0.65
TT[CA]C[TAG][GA]GAA[GA]T	0.59
TT[CA][TC][GA]GGAA[AG]	0.58
GGACGCCG	−0.57
G[TGC][CGT][AT][TG]**TTCC[TCA][GA][GT]AA**[AG]	0.53

The classical GAS motifs are shown in boldface.
doi:10.1371/journal.pone.0011881.t006

squares fit. As a consequence, using the same STAT1 dataset, a 30-fold cross-validation takes approximately 60 hours with the ridge regression, while it takes only 2.5 hours with the L1-norm log linear regression (i.e., 24 times faster). In summary, although both methods show very similar results with respect to binding motif identification, the ridge regression is slower and more difficult to use with large ChIP-Seq datasets than the L1-norm log linear regression.

Partial least squares regression and principal component regression. In Tables 8 and 9, we show the top ten motifs for STAT1 identified by the partial least squares regression and the principal component regression. We can see that both methods are able to identify the classical GAS motif. In Table 8, the partial least squares regression shows very similar results to the L1-norm log linear regression as both methods identify CA[TC]GTGACT-[TG]C as a strengthening motif and GGAGGGCG as a weakening motif. In Table 9, the principal component regression identifies only the GAS motif and fails to identify any other motifs involved in the binding. In the case of RNA Polymerase II, both partial least

Table 7. List of putative RNA Polymerase II binding motifs identified by ridge regression.

RNA Polymerase II	*Normalized coef.*
T[AG]**A[GC][TAG]CA**[GCT]A[AC]AA	1.0
A[GA]AA[AC][CA]AA[AC]AAA	0.86
C[ACT]**[GT][CG][CT][TA]CC**[AGT]CC[TA]	0.81
C[CT][CG][AT]GGCTGG[AG]G	0.75
TTTCTGC[CT][CT]TT[GT]	0.74
GGAGGGAGGC[AG]G	0.70
AC[AC][CA][AC][AT][AG]AGAAA	0.65
T[TA]T[TC]**[CA]CAGAC**T[AT]	0.63
TTTGT**[CT][TA]T[TG][AC][AT]**T	0.62
TT[TAC]TTT[CT]TT[TC]TT	0.61

The known Downstream Promoter Element and Initiator site motifs are shown in boldface.
doi:10.1371/journal.pone.0011881.t007

Table 8. List of putative STAT1 binding motifs identified by partial least squares regression.

STAT1	*Normalized coef.*
CA[TC]GTGACT[TG]C	1.0
[TG]G[GTA][GC][AG]T**TT[CA]C[AGC] [GA]GAA**[AC][TG]G[GA][GC]	0.80
TTC[CT][TG][GA]GAAAT[GC][CA] [CA][CAT][AT][TCG][CG][CT]	0.58
[CT][TC]CA[GT]**TTCCAGGAA**[AT]T[CG][CAT]C[CT]	0.56
[GA][AG]A[AG][AT][CTG][CA]A[GT][CG]T[GT][CG] [CA]T[TG][CT][CGT]T	0.50
TCACA[TG]G[ACG]	0.42
GGAGGGCG	−0.41
G[TGC][CGT][AT][TG]**TTCC[TCA][GA][GT]AA**[AG]	0.41
TT[CA]C[TAG][GA]GAA[GA]T	0.40
A[TA]**TTCC[CT][GA]GAA**[AC]T[CG][AC]	0.39

The classical GAS motifs are shown in boldface.
doi:10.1371/journal.pone.0011881.t008

squares regression (Table 10) and principal component regression (Table 11) are able to identify the initiator site and the downstream promoter element.

However, the results of the partial least squares regression and the principal component regression are difficult to interpret. In the former (Table 10), different instances of the downstream promoter element have positive or negative coefficients (T[TG]AACA-**CAGTT[TA]** with 1.0, [CT][CG]AG**A[GA]TCC**A[GA][CG] with −0.90, and A[AG][GA][AG]**GGA[GCA]G**A[GA]A with 0.87). As they are instances of the same motif, we expect them to have the same sign, i.e., to have the same effect on the binding. In the latter (Table 11), all the instances of the initiator site and the downstream promoter element have negative coefficients. However, these motifs should strengthen the binding and therefore, we expect their coefficients to be positive.

Table 9. List of putative STAT1 binding motifs identified by principal component regression.

STAT1	*Normalized coef.*
[TAC]**TTCC[CA][GA][GT]AA**[AG][TA]C	1.0
T**TTCC[CT][GA]GAA**AA[CT]TC[AC]TGAA	0.94
TT**TTC[CT][AG]GGAA**[AG][GT]GG[CG][TCA][GA]GG	0.87
T**TTC[TC][TG][GA][GAT]AA**[GA]	0.86
[TC]**TTCC[AC][AG]G[AC]A**	0.85
[GA]GAACC[TC][TG]CAG**TTC[CT][AG]GGAA**	0.82
CC[CTA][CGT]T**TTC[CT]T[GA]GAA**[AG][ACT][CG]	0.82
TTC[CT][TG][GA]GAAAT[GC][CA][CA][CAT]-[AT][TCG][CG][CT]	0.81
T**TTC[CT][AGT]GGAAA**[TG][GA][GA]G[TAC][GA]G	0.80
G[CT]**TT[CA][CT][GAT][GA]GAA**[AG][TG][AGC]-[GA][GCA][TGA]A[CG]	0.78

The classical GAS motifs are shown in boldface.
doi:10.1371/journal.pone.0011881.t009

Table 10. List of putative RNA Polymerase II binding motifs identified by partial least squares regression.

RNA Polymerase II	*Normalized coef.*
T[TG]AACAC**AGTT[TA]**	1.0
C[CT][CG][AT]GGCTGG[AG]G	0.99
G[AG]GG[CG]CCAGAGA	−0.97
[CT][CG]AG**A[GA]TCC**A[GA][CG]	−0.90
CTGG[AC]GCTG[TG][TC][ACG]	−0.89
A[AG][GA][AG]**GGA[GCA]G**A[GA]A	0.87
[CG][AT][CT]T**[GC]C[AT][CG]TCC**[AC]	0.86
GGAGGGAGGC[AG]G	0.86
A[GA]AA[AC][CA]AA[AC]AAA	0.85
[GT]GCCCAGG[CG][TG][GA]G	−0.81

The known Downstream Promoter Element and Initiator site motifs are shown in boldface.

doi:10.1371/journal.pone.0011881.t010

The lack of interpretability of the partial least squares regression and the principal component regression lies in the fact that the regression is performed in a low-dimensional feature space. In the original motif space, the vector representation of the peak sequences has a meaning and each component of a vector measures how similar a motif is to a peak sequence. However, in the low-dimensional feature space computed by the partial least squares regression and the principal component regression, the vector components lose their biological meaning. From the computational complexity perspective, we also mention that both methods are very slow. Using the STAT1 dataset, a 30-fold cross-validation of the partial least squares regression with 10 components takes approximately 240 hours. In summary, the partial least squares regression and the principal component regression are able to identify the classical GAS motif for STAT1 and the initiator site and the downstream promoter element for RNA Polymerase II. However, the results are difficult to interpret biologically and do not allow identification of strengthening or weakening motifs. In addition, they are too slow to be used with large ChIP-Seq datasets.

Table 11. List of putative RNA Polymerase II binding motifs identified by principal component regression.

RNA Polymerase II	*Normalized coef.*
GCT**GG[GT][AC][CT]**[CT]ACA	−1.0
[CG]GCGGCGGCGGC	0.97
GCCCAGGCTG[CG][TA]	−0.96
CA[AC]**AG[TG][GC]C**TG[GA]G	−0.94
CTGG[TC][CT]TCAAA[GC]	−0.90
CTGG[AG]G[TG]GC[AT]G[TG]	−0.89
CTG**GA[GA]T[GT]C**A[GA][TG]	−0.87
[TC]CCA[CA]AG[CAT][AG]CTG	−0.86
[TA]C[AC]T[GA][CG]CCTGT[GT]	−0.84
[CA]TG[AT]CCACAGA[AT]	−0.83

The known Downstream Promoter Element and Initiator site motifs are shown in boldface.

doi:10.1371/journal.pone.0011881.t011

Advantages of L1-norm log linear regression over other methods for TFBM identification. We considered the following regression methods for TFBM identification: L1-norm log linear regression, linear least squares regression, ridge regression, partial least squares regression, and principal component regression. In Table 12, we summarize the correlation coefficients averaged on the test sets. As we can see, all regression methods demonstrate similar performance and are able to identify the classical GAS motif for STAT1 and the initiator site and the downstream promoter element for RNA Polymerase II.

However, they exhibit marked differences with respect to biological interpretability and computational efficiency. The results of the linear least squares regression, the partial least squares regression, and the principal component regression do not allow identification of strengthening or weakening motifs. Therefore, they are difficult to use for binding motif identification. Both L1-norm log linear regression and ridge regression solve this problem by means of regularization. However, the ridge regression is very slow compared to the L1-norm log linear regression. Therefore, the ridge regression is difficult to use with large-scale ChIP-Seq datasets. In summary, the L1-norm log linear regression is the only method that can achieve all the desired goals for our task; it identifies the transcription factor binding motifs, the regression coefficients are easy to interpret biologically, and its implementation with the LASSO algorithm is fast and efficient. This justifies our choice of the L1-norm log linear regression in PeakRegressor.

Parameter setting

The performance of PeakRegressor depends on the choice of parameters that have to be set empirically. In this section, we explain how we choose two important parameters: the length of peak sequences and the number of motif candidates.

Length of peak sequences. In the dataset provided by [9], all the peaks correspond to various DNA sequences. These sequences have different lengths, ranging from 1 bp to several thousand bp. To conduct our analysis, we modify the peak sequences in the following way:

- We shorten long peak sequences for two reasons. First, when using long DNA sequences, the computations of the motif finding algorithm MEME take too much time. Second, finding good quality motifs with MEME is easier with short DNA sequences than with long ones.
- We widen short peak sequences. Due to the noisy nature of ChIP-Seq data, the motifs we are looking for may not be exactly on the provided peak sequence, but in the surrounding DNA neighborhood. Therefore, we decide to choose a uniform length for all the peak sequences. The choice of 200 bp is empirical; we try several values (100 bp, 200 bp, 400 bp, and 800 bp) and consider the one

Table 12. Different regression methods and their correlation coefficients averaged on the test sets.

Regression method	*STAT1 correlation coef.*	*Pol II correlation coef.*
L1-norm log linear regression	**0.65**	**0.66**
Linear least squares regression	0.64	0.64
Ridge regression	0.64	0.64
Partial least squares regression	0.64	0.65
Principal component regression	0.63	0.52

doi:10.1371/journal.pone.0011881.t012

that achieves the best performance, i.e., the highest correlation coefficients (results not shown for other peak lengths).

Number of motif candidates. In the first step of PeakRegressor, we use MEME to find over-represented DNA motifs in the peak sequences. This step results in 800 motif candidates for STAT1 and 880 for RNA Polymerase II. Given the large number of motif candidates, we empirically observe the presence of similar motifs in the set of motif candidates. We may wonder if this redundancy could affect the prediction performance of PeakRegressor. However, we show that this is not the case.

PeakRegressor uses a regression method called L1-norm log linear regression. In contrast with other regression methods, L1-norm log linear regression achieves its best prediction performance by removing redundant or uninformative motifs from the regression model. Therefore, the removal of redundant motifs is automatically performed when using L1-norm log linear regression. Table 2 shows the set of motifs that achieve the best correlation coefficient for STAT1. We can see that some motifs are similar. For example, the motifs A[CT]**TTC[TC][TG]GGAA, TT[CA]C[TAG][GA]GAA** [GA]**T**, A[TA]**TTCC[CT][GA]GAA**[AC]T[CG][AC], and **TT-[CA][TC][GA]GGAA**[AG] are short, similar motifs containing the STAT1 binding motif. In other experiments, we find that the prediction performance worsens when similar motifs are removed (results not shown). Hence, although the motifs appear similar and redundant, they actually contain complementary information for the prediction performance.

Moreover, the motif weights computed by PeakRegressor are all different (resp. 0.56, 0.55, 0.48, and 0.47). Hence, while other approaches, such as motif clustering, would consider all these motifs to be equally important, PeakRegressor is able to detect the relative importance of each motif and compute the corresponding weight. This is explained by the noisy nature of the DNA motifs found by MEME in step 1. For a given binding motif, PeakRegressor needs to use all the noisy PSSM approximations to achieve the best prediction performance. This is an important property of PeakRegressor, especially when the number of noisy motifs is very large.

Candidate motifs and their potential rSNPs

Single or composite motifs found in the PeakRegressor system may reflect actual transcription factor binding sites. If a single nucleotide polymorphism (SNP) occurs within the sites, regulatory control of neighboring gene transcription will be perturbed, thus leading to genetic diseases in some cases [10]. Therefore, true binding sites may have SNPs less frequently than the non-binding sites. As an important verification, we check the number of known SNPs to be found within the STAT1 positions presented by PeakRegressor by using dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/). We find that 0.36% (147 for 40,395 bp) of mapped positions with 10 STAT1 motifs in Table 2 on the peak sequences contains SNPs, while as much as 0.53% (17,852 for 3,344,439 bp) of all positions contains SNPs on the peak sequences. The statistical difference between the above two ratios is highly significant such as $p<3.7^{-7}$ according to the hypergeometric distribution. These sites are possible candidates of rSNPs because the slight change within the motif may affect the change of gene expression level and might cause diseases.

Materials and Methods

PeakRegressor

PeakRegressor is a system to find TFBMs that are statistically important for transcription factor binding signals, by taking ChIP-Seq data as input, and outputs a list of TFBM candidates. In contrast with previous approaches, PeakRegressor uses the peak scores (provided by [9]) as a surrogate for the binding affinities. We argue that the peak scores provide more accurate approximations of the binding affinities than the methods based on transcription factor occupancy [4,5]. Therefore, using the peak scores lead to better identification of functional TFBMs. In addition, PeakRegressor identifies not only primary TFBM candidates but also secondary motifs that may often synergistically strengthen or weaken the binding. The workflow is summarized in Figure 2.

Step 1. First, we define the peak sequences as the 200-bp genomic regions centered around the peaks. Then, we sort the peak sequences according to their ascending scores. We group the peak sequences into clusters such that each cluster contains 200 peaks of consecutive scores. Then, we apply MEME (http://meme.sdsc.edu/) to each peak sequence cluster. For each sequence cluster, MEME is parameterized in ZOOPS mode to find 10 motifs of lengths $8-20$.

This strategy has two advantages. First, it allows us to identify motifs that may be associated with a given binding affinity level. If a cluster contains only low (resp. high) binding affinity peaks, the corresponding sequences may contain weak (resp. strong) binding motifs, i.e., motifs that are specific to low (resp. high) binding affinity. Second, it reduces computational time by parallelizing MEME computations.

Step 2. In order to predict the binding affinity of the peaks, we need to represent each peak as a vector in the motif space. Let seq^i be the DNA sequence of peak i. Let $\text{seq}^i_{j,\ell}$ be the ℓ-length sub-sequence of seq^i, starting from position j. Let S^d be the PSSM of motif d. Let ℓ_i be the length of seq^i and ℓ_d be the length of motif d. We represent peak i as vector $x_i \in R^D$, such that

$$x_{id} = \max_{j=1\ldots\ell_i-\ell_d+1} f(\text{seq}^i_{j,\ell_d}, S^d) - \max(S^d)$$

for $d=1\ldots D$. The quantity $f(\text{seq}^i_{j,\ell_d}, S^d)$ is a sum of log-odd scores, representing how well motif d matches sub-sequence seq^i_{j,ℓ_d}. Hence, the first term of the sum, x_{id}, corresponds to the best match when we slide motif d along sequence seq^i. The term $\max(S^d)$ is the maximum score achievable by any sequence matching with the motif d. Therefore, we always have $x_{id} \leq 0$, with $x_{id}=0$ for the best possible match.

Next, we want all the x_{id} to be positive for interpretability purpose. So we simply shift their values by substracting the lowest component: $x_{id} \leftarrow x_{id} - a$, where a is the minimum value of the original x_{id}. Finally, we normalize each data vector by dividing it with its euclidean norm: $x_i \leftarrow x_i / \|x_i\|^2$.

Step 3. Quantities y_i to be fitted are the log values of the peak enrichment scores, as given by PeakSeq [9]. We can now solve the regression problem defined by (x_i, y_i) pairs for $i=1\ldots N$. Linear regression is a simple and popular approach, but is prone to overfitting. Hence, we choose to regularize the model with L1-norm, i.e., we want to minimize the sum of squared errors and the L1-norm of the regression coefficient vector:

$$\min_{b\in R^D} \beta\|b\| + \sum_{i=1}^{N} (b^T x_i - y_i)^2 \qquad (1)$$

where $\beta>0$ is a user-defined regularization coefficient. The L1-norm log linear regression is able to remove redundant or uninformative features, and to select a small number of features that best explain the fitted quantity [11]. In our case, the features correspond to DNA motifs and hence, the result of this step is a set

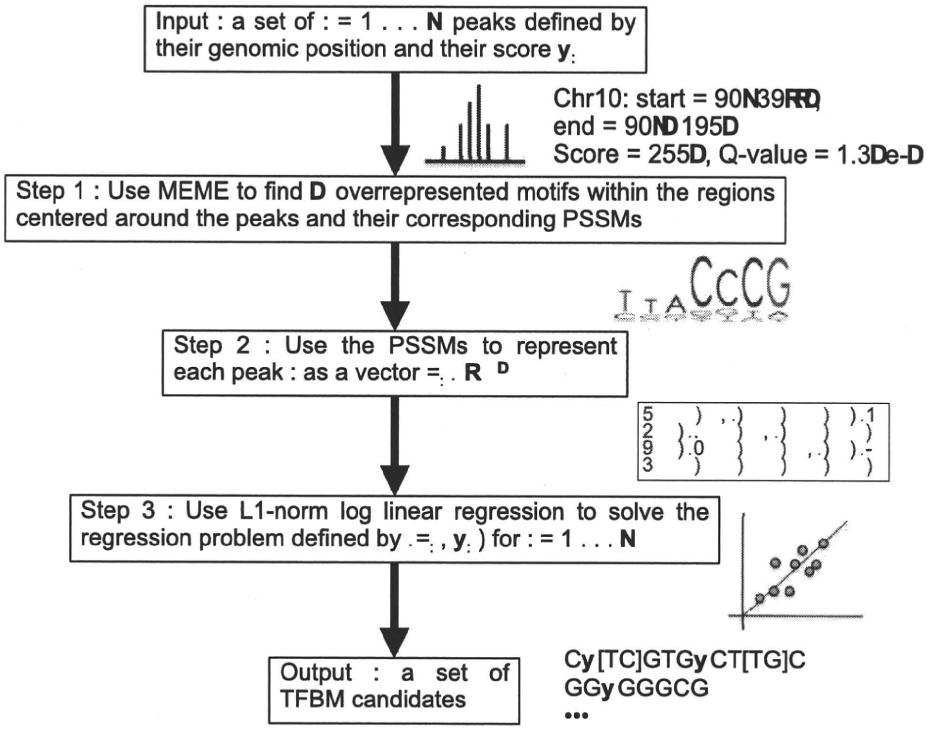

Figure 2. Schematic view of the workflow of PeakRegressor. PeakRegressor takes ChIP-Seq data as input and outputs a list of TFBM candidates and their weights that give the best regression accuracies.
doi:10.1371/journal.pone.0011881.g002

of motifs that best explain the binding signal values from ChIP-Seq dataset. We use Lasso, a popular algorithm for solving L1-norm log linear regression. Lasso is available as part of the LARS package for R (http://www-stat.stanford.edu/~hastie/Papers/LARS/).

Other regression methods

In this section, we present alternatives to the L1-norm log linear regression: linear least squares regression, ridge regression, partial least squares regression, and principal component regression. All these regression methods are used in the following way. Once a regression model is fitted to the peak dataset, we rank the regression coefficients with respect to their absolute values. Using this ranking, the top motifs are the potential TFBMs.

Linear least squares regression. The linear least squares regression is the simplest regression approach. It fits a linear model to the dataset by minimizing the sum of squared errors $\sum_{i=1}^{N}(y_i - b^T x_i)$. Its difference with the L1-norm log linear regression (equation 1) is the absence of a regularization term. Therefore, the linear least squares regression is more prone to overfitting when the regression problem contains more dimensions than samples.

Ridge regression. The ridge regression [12] minimizes $\|b\|^2 + \sum_{i=1}^{N}(y_i - b^T x_i)$, where the regularization term is $\|b\|^2 = \sum_{d=1}^{D} b_d^2$, i.e., the Euclidean norm of b. It is quite similar to the L1-norm log linear regression, and their main difference lies in the regularization term. The ridge regression seeks a solution with a low Euclidean norm. Although the Euclidean norm is a protection against overfitting, it does not favor sparse solutions (i.e., solutions with many motifs) as the L1-norm log linear regression does [11].

Partial least squares regression and principal component regression. The partial least squares regression [13] and the principal component regression are two approaches of the same idea; they perform linear regression using the low-dimensional data matrix Z instead of the initial data matrix X. This approach avoids overfitting problems. Therefore, the partial least squares regression and the principal component regression have been widely used in problems containing several dimensions (i.e., motifs) and few samples (i.e., peaks).

In the principal component regression, the low-dimensional data matrix Z contains the most information about the initial data matrix X (according to the singular value decomposition of X). In the partial least squares regression, the low-dimensional data matrix Z is calculated using both the initial data matrix X and the peak score vector y. In both cases, linear regression is performed using Z instead of the initial data matrix X. Both partial least squares regression and principal component regression are available as part of the PLS package for R (http://mevik.net/work/software/pls.html). Once the regression coefficients have been computed in the low-dimensional space, they are mapped back in the original motif space. Then, these coefficients can be used to identify potential binding motifs.

Input ChIP-Seq datasets

The ChIP-Seq dataset we used is provided by [9] and is publicly available (http://www.camda2009.org/). The dataset provides various information about each peak, including the peak score, the peak center (for STAT1), and the Q-value that reflects the significance of the peak. The Q-values are derived from the P-values. First, they compute the P-values that reflect the significance of peak enrichment in the number of DNA tags, compared to control samples. These P-values are computed using the binomial distribution. Then, to account for multiple hypothesis testing, the Q-values are derived from the P-values. See [9] for more details.