Reproductive Toxicology 29 (2010) 68–73

Contents lists available at ScienceDirect

Reproductive Toxicology

journal homepage: www.elsevier.com/locate/reprotox

When does the sex ratio of offspring of the paternal 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) exposure decrease: In the spermatozoa stage or at fertilization?

Kana Ishihara[a], Seiichiroh Ohsako[b], Ken Tasaka[a], Hiroshi Harayama[a], Masashi Miyake[a], Katsuhiko Warita[c], Takashi Tanida[a], Tomoko Mitsuhashi[a], Takashi Nanmori[a], Yoshiaki Tabuchi[d], Toshifumi Yokoyama[a], Hiroshi Kitagawa[a], Nobuhiko Hoshi[a,*]

[a] *Department of Animal Science, Graduate School of Agricultural Science, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan*
[b] *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, Japan*
[c] *Department of Anatomy and Neurobiology, Faculty of Medicine, Kagawa University, 1750-1 Ikenobe, Miki-cho, Kita-gun, Kagawa 761-0793, Japan*
[d] *Division of Molecular Genetics Research, Life Science Research Center, University of Toyama, 2630 Sugitani, Toyama 930-0194, Japan*

ARTICLE INFO

Article history:
Received 17 August 2009
Accepted 26 September 2009
Available online 4 October 2009

Keywords:
TCDD
Paternal exposure
Sex ratio
Spermatozoa
2-cell embryo

ABSTRACT

Recent animal experiments confirmed that paternal 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) exposure decreases the sex ratio of offspring at birth without altering litter size. However, the timing of this decrease remained unclear. Male mice were administered TCDD at 7–12 weeks of age and mated with non-treated females. The Y-bearing/X-bearing sperm ratio was examined by real-time PCR and FISH methods, and the sex ratio of the 2-cell embryos collected from non-treated females that had been mated with TCDD-exposed males were investigated by nested PCR. The Y-bearing/X-bearing sperm ratio was not significantly decreased in the TCDD group. However, the sex ratio of the 2-cell embryos of the TCDD group was significantly lower than that of the control group. These results may have resulted from a decrease in fertility of Y-bearing sperm. Thus, the results of this study suggested that the sex ratio of the offspring was decreased at fertilization and not during the spermatozoa stage.

1. Introduction

In recent years the public has become more aware that exposure of males to certain environmental or occupational agents affects their offspring. Occupational exposure in various industries has led to increased incidences of miscarriage [1] and various birth defects [2]. It has also been shown that paternal cranial irradiation leads to epigenetic alterations in offspring [3]. Prominent among these reports are those by Mocarelli et al. [4,5] showing that paternal 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) exposure in Seveso, Italy, led to a decrease in male offspring.

In 1976, TCDD was released in an explosion at a chemical plant near Seveso, Italy, resulting in the highest concentrations of TCDD ever recorded in humans. Subsequent data linked a decrease in male births in Seveso to the increased TCDD concentration in fathers, and the altered sex ratio (the proportion of male offspring) was especially pronounced in the children of fathers exposed to TCDD before age 19 [4,5]. In addition to the Seveso incident, Ryan et al. [6], in a study conducted in Ufa, Russia, suggested that human exposure to high levels of dioxin is associated with the birth of more girls only in cases of paternal exposure. In our previous study, we exposed young male mice to two concentrations of TCDD (TCDD2/0.4 group and TCDD2000/400 group; an initial loading dose of 2 or 2000 ng TCDD/kg, followed by a weekly maintenance dose of 0.4 or 400 ng TCDD/kg) to re-create the Seveso incident and evaluated the sex ratio of their offspring at birth [7]. The reason why we used young male mice (7 weeks at the start of administration) in that study was that the sex ratio of offspring of males exposed to TCDD during adolescence showed a greater decrease than that of males who were older than that when the incident occurred [4,5]. The results from the previous study revealed that paternal TCDD administration produced a dose-dependent reduction in the sex ratio of offspring (F1) and a significantly lower proportion of male offspring in the high-dose (TCDD2000/400) group. In addition, the induction intensity of CYP1A1 in the liver varied among individuals in the TCDD group, and the dimensions of the CYP1A1 immunoreactive area were correlated with the sex ratio of the offspring. This means that the high sensitivity subgroup of male parents to TCDD was strongly related to the decrease in male offspring. We also reported that TCDD exposure does not influence

* Corresponding author. Tel.: +81 78 803 5811; fax: +81 78 803 5811.
E-mail address: nobhoshi@kobe-u.ac.jp (N. Hoshi).

0890-6238/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.reprotox.2009.09.009

litter size; the number of male offspring decreases by TCDD exposure, while the number of female offspring increases. From these data, we presumed that paternal TCDD exposure decreased the sex ratio of offspring and altered it before implantation occurred. However, the mechanisms underlying the reduction in male offspring and the timing of this change have remained unclear.

The purpose of this study was to investigate when the sex ratio of the offspring decreases. We examined the Y-bearing/X-bearing sperm ratio as well as the sex ratio of 2-cell embryos by exposure to the same dose (an initial loading dose of 2000 ng TCDD/kg, followed by a weekly maintenance dose of 400 ng TCDD/kg) as used in our previous study [7], because this dose group showed a significant decrease in the sex ratio of offspring at birth. The effects of TCDD exposure to sperm concentration and motility were also examined.

2. Materials and methods

2.1. Chemicals

TCDD was obtained from Wako Pure Chemical Industries, Ltd. (Osaka, Japan). Sesame oil, used for dissolving TCDD and for vehicle treatment as a control, was purchased from Kanto Chemical Co., Inc. (Tokyo, Japan).

2.2. Animals and treatments

Male and female ICR mice were purchased from Japan SLC, Inc. (Hamamatsu, Japan) and bred at Kobe University (Kobe, Japan). They were maintained under controlled conditions of temperature (23 ± 2 °C) and humidity (50 ± 10%) on a 12 h light, 12 h dark cycle. The animals were given an MR-A1 laboratory diet (Nosan Corporation, Yokohama, Japan) and filtered water *ad libitum* throughout the experiments. This study was approved by the Institutional Animal Care and Use Committee (Permission number: 19-5-46) and carried out according to the Kobe University Animal Experimentation Regulations.

We used 7-week-old male mice at the start of administration to re-create the Seveso incident. These male mice were divided into a control group ($n = 59$) and a TCDD group ($n = 49$), and were administered TCDD orally by gastric sonde with an initial loading dose of 2000 ng TCDD/kg body weight or an equivalent volume of sesame oil (vehicle) as a control, followed by a weekly maintenance dose of 400 ng TCDD/kg body weight to maintain the body TCDD burden as constantly as possible or an equivalent volume of sesame oil until the mice reached 12 weeks of age. We also used 6-week-old female mice ($n = 62$) for mating with males.

2.3. Epididymal sperm concentration and motility analysis

Mouse spermatozoa were collected from TCDD-treated mice and control mice for examination of sperm concentration and sperm motility. The mice were sacrificed by cervical dislocation under anesthesia using ether, and the caudal epididymides were then dissected out. The epididymis was cut at one or two points with a scissors, and white sperm pellets were released into a TYH-HEPES medium droplet that was covered with mineral oil (Nacalai Tesque Inc., Kyoto, Japan) on a 37.5 °C heater plate. The TYH-HEPES medium was composed of 119.37 mM NaCl, 4.78 mM KCl, 1.71 mM $CaCl_2$, 1.19 mM KH_2PO_4, 1.19 mM $MgSO_4$, 1.00 mM sodium pyruvate, 5.56 mM glucose, 25.07 mM HEPES, 0.05 g/l streptomycin sulfate, 100 U/ml potassium penicillin G, 0.1% polyvinylalcohol (Sigma–Aldrich Co., St. Louis, MO, USA) and 5 mg/l phenol red. The recovered spermatozoa were diffused for 5 min in the droplet. The sperm concentration was measured using a hemocytometer chamber. For sperm motility analysis, a 2 μl drop of sperm suspension was put on a 1-mm-deep stage of a glass plate to assess sperm motility (Fujihira Industry Co., Ltd., Tokyo, Japan), covered with the coverslip and then placed on the heated plate (37.5 °C) under a bright-field microscope (EX41; Olympus Co., Tokyo, Japan). Sperm motility was recorded with a CCD camera (CS230B; Olympus Co.) and a DVD recorder (DVR-7000; Pioneer Co., Tokyo, Japan). The motility patterns of more than 100 spermatozoa (except large sperm clumps) were randomly recorded with the CCD camera and DVD recorder in each sample. The recorded movies were played in the slowest mode. At least 10 microscopic fields were observed for each sample, and the percentage of motile sperm was determined.

2.4. Sperm DNA isolation

Mouse spermatozoa for DNA isolation were collected from TCDD-treated and control mice by pricking the caudal epididymides with needles, and this manipulation was done in a TYH-HEPES medium droplet that was covered with mineral oil under a stereomicroscope so as not to contaminate the other male tissues. The sperm pellets were resuspended with a 1–1.5 volume of lysis buffer containing 0.5 M EDTA, 2-mercaptoethanol (Sigma–Aldrich Co.) and 10 mg/ml proteinase K (Takara Bio Inc., Shiga, Japan), and then incubated in a shaking water bath at 55 °C overnight. The 2-mercaptoethanol was used to obtain high-quality spermatozoa DNA yields for real-time PCR, because the spermatozoa DNA is tightly packed into protamines. Sperm DNA extraction was performed using the Wizard genomic DNA purification kit (Promega, Madison, WI, USA) following the manufacturer's instructions.

2.5. Real-time PCR and calculation of Y-bearing/X-bearing sperm ratio

To gain an accurate prediction of the X and Y chromosome content in the sperm DNA samples, quantitative real-time polymerase chain reaction (PCR) analysis was performed using a LightCycler (Roche Co., Basel, Switzerland). We selected the *Sry* primer pair (forward: 5′-ATGGAGGGCCATGTCAAGCG-3′ and reverse: 5′-GGGTATTTCTCTCTGTGTAGGATCTTCAA-3′) as Y-bearing sperm-specific primers and the *AR* primer pair (forward: 5′-ATGGAGGTGCAGTTAGGGCT-3′ and reverse: 5′-TCCTCAGTGTGCTGCTGCC-3′) as X-bearing sperm-specific primers. To calculate Sry and AR gene copy numbers in the isolated genomic DNA, the basic protocol for real-time PCR was briefly modified [8]. The Sry and AR gene PCR products were used for calibration by calculating the molecular weight and making stock dilutions from 2×10^8 to 2×10^3 copies/μl. Aliquots (2 μl) of DNAs or standard DNA solution were amplified with a master mixture (SYBR Premix Ex Taq (Perfect Real Time), TaKaRa Bio Inc.) containing the Sry- or AR-specific primers described above in a final volume of 20 μl. Fluorescent products were detected at the end of the extension period. The specificity of the amplified PCR products was confirmed by melting curve analysis. Unknown concentrations of samples were extrapolated by a comparison with standards amplified under the same conditions using Lightcycler software. The Y-bearing/X-bearing sperm ratio [DNA concentration of Sry/DNA concentration of AR] of each animal was calculated.

2.6. Preparation of epididymal sperm smears for FISH

Mouse spermatozoa for FISH analyses were collected from TCDD-treated and control mice. Mice were killed by cervical dislocation, and then the caudal epididymides were dissected out. The epididymis was cut at 1 or 2 points with a scissors, and the white sperm pellet was placed into 0.1 ml of 2.2% sodium citrate at 32 °C for 5 min to allow the sperm to swim out into the solution. The sperm suspension was centrifuged at 1500 rpm for 5 min. The sperm pellet was resuspended into 0.1 ml of 0.075 M KCl for 45 min at 37 °C. Sperm suspension (5 μl) was pipetted onto a dry glass slide that had been precleaned by soaking in 100% ethanol for at least 24 h. The smears were allowed to air-dry overnight and then stored at −20 °C until used.

2.7. FISH in spermatozoa and calculation of Y-bearing sperm ratio

Smears of mice were each fixed in 3:1 methanol: acetic acid and air dried before pretreatment commenced. To prepare the smears for FISH, spermatozoa were decondensed by incubating the slides for 30 min in 0.2 ml of 10 mM dithiothreitol (Sigma–Aldrich Co.) on ice. Slides were briefly rinsed in D.W. and allowed to dry completely at room temperature before they were used for hybridization. The probes specific for mouse Y chromosome (Cy3-labeled) were obtained from Cambio (Cambridge, UK). The probe mixture, which included labeled probes and hybridization buffer, was denatured at 72 °C for 10 min. The sperm smears were denatured at 78 °C for 6 min in 70% formamide/2xSSC and then dehydrated in an alcohol series (70%, 85% and 100%, 2 min each). The probe mixture was applied to the sperm smears and incubated at 37 °C overnight. After hybridization, the slides were washed six times: three times in washing solution (50% formamide/2xSSC), twice in 2xSSC, and once in PN buffer for 5 min each at 37 °C. The nuclei were counterstained with 4, 6-diamidino-2-phenylindole (DAPI) (Cambio Ltd.). Coverslips were applied and sealed with clear nail polish. Slides were then viewed.

The slides were examined with a Leica FW4000 fluorescence microscope (Leica Co., Wetzlar, Germany) (magnification: 400×) equipped with single and double band-pass filters to detect DAPI and Cy3. The investigation was performed blindly, and at least 1000 cells per slide were scored as much as possible by assessing randomly selected visual fields. Sperm nuclei were scored only when they were morphologically preserved, not clumped or overlapping, when they showed well-defined outlines and when the sperm heads were not decondensed to more than twice the size of normal non-decondensed spermatozoa. In every sample, the proportion of sperm presenting with a clear signal was ≥95%.

The Y-bearing sperm ratio [Y-chromosome-bearing spermatozoa/DAPI-positive spermatozoa × 100] of each animal was calculated.

2.8. Embryo collection for nested PCR

Six-week-old female mice were superovulated by intraperitoneal injections of 5 IU eCG (Teikoku Hormone Medical Co., Ltd., Tokyo, Japan) followed 48 h later by intraperitoneal injections of 5 IU of hCG (Teikoku Hormone Medical Co., Ltd.). One week after the last exposure of TCDD to male mice, each superovulated female mouse that had not been exposed to TCDD was paired with a male, with one pair per cage. After confirmation of vaginal plug in each female, embryos were collected 44 h after hCG injection from the oviducts of the females at the 2-cell stage in 37 °C FHM medium, which was composed of 95 mM NaCl, 2.5 mM KCl, 0.35 mM KH_2PO_4, 0.2 mM $MgSO_4 \cdot 7H_2O$, 4 mM $NaHCO_3$, 0.2 mM glucose, 0.2 mM glutamine, 0.2 mM Na-pyruvate, 10 mM Na-lactate, 1.71 mM $CaCl_2 \cdot 2H_2O$, 0.01 mM EDTA·4Na, 20 mM HEPES, 80 mg/l kanamycin and 1 g/l BSA.

The embryos were transferred to 80 μl drops of KSOM medium, which was composed of 94.97 mM NaCl, 2.55 mM KCl, 0.37 mM KH_2PO_4, 0.23 mM $MgSO_4 \cdot 7H_2O$, 25 mM $NaHCO_3$, 0.22 mM glucose, 1.0 mM glutamine, 0.2 mM Na-pyruvate, 10 mM Na-lactate, 1.7 mM $CaCl_2 \cdot H_2O$, 80 mg/l kanamycin, 3 g/l BSA, 0.1 mM EDTA–2Na under mineral oil and separated by gentle pipetting of the groups of 2-cell embryos using a micropipette with an internal diameter of 70–100 μm. Each 2-cell embryo was put into a PCR micro-test-tube containing 15 μl of 0.15 mg/ml proteinase K with the help of a dissection microscope. After the embryos were collected, the PCR tubes were heated for 60 min at 50 °C and then for 10 min at 100 °C to inactivate the proteinase K. The cells in the PCR micro-test-tubes were stored at −80 °C until PCR analysis was performed.

2.9. Oligonucleotide primers for nested PCR

The *Sry* gene was chosen as the Y chromosome-specific gene and the *IL-3* gene was chosen to detect both XX and XY embryos. For every locus, two sets of primers were used, an inner and an outer set. The primers were obtained from Hokkaido System Science Co., Ltd. (Sapporo, Japan). The oligonucleotide sequences from the outer primers used in the initial PCR were as follows:

Sry-outer forward 5′-TCTTAAACTCTGAAGAAGAGAC-3′
Sry-outer reverse 5′-GTCTTGCCTGTATGTGATGG-3′
IL-3-outer forward 5′-GGGACTCCAAGCTTCAATCA-3′
IL-3-outer reverse 5′-TGGAGGAGGAAGAAAAGCAA-3′
The inner primer sequences used in the second-round PCR were as follows:
Sry-inner forward 5′-TTCCAGGAGGCACAGAGATT-3′
Sry-inner reverse 5′-GTCCCACTGCAGAAGGTTGT-3′
IL-3-inner forward 5′-GGGAAGCTCCCAGTGAGTAA-3′
IL-3-inner reverse 5′-GGTTCACCCACAGCTGTCTT-3′

The sequences of Sry-outer primers were the same as those used in Kunieda et al. [9].

2.10. Nested PCR and the sex ratio of the 2-cell embryos

A multiplex PCR reaction using nested primers was performed in two rounds for simultaneous amplification of the *Sry* and *IL-3* sequences. Samples were thawed on ice, and a 15 μl first-round PCR mix containing the outer primers was added. After the first amplification, 1 μl portions of the products of the amplification reactions were dispensed respectively into each tube, and each was subjected to second-step amplification with 9 μl of a reaction mixture containing the corresponding inner primers. Both steps of the PCR were carried out using a reaction mixture consisting of a 5× PCR buffer, 2.5 mM dNTPs, 50 μM each of the oligonucleotide primers and 1.25 U/μl of Taq DNA polymerase (Promega). PCR was performed for 30 cycles, each consisting of 1 min denaturation at 94 °C followed by annealing and extension for 1 and 1.5 min each at 55 and 72 °C, respectively. The PCR products were subjected to electrophoresis in a 2.5% agarose gel containing 0.005% ethidium bromide. The male embryos showed both Sry and IL-3 signals, but the female embryos showed only the IL-3 signal.

The sex ratio of the 2-cell embryos [number of male embryos/number of (male + female) embryos × 100] of each animal was calculated.

2.11. Statistical analysis

SAS Ver 5.0 (SAS Institute, Cary, NC, USA) was used to analyze the present data. In comparing every parameter between two groups (control and TCDD groups), the equality of variances was evaluated by using the *F* test to select statistical tests. If variances were homoscedastic, we used Student's *t*-test. In contrast, if the variances of parameters were not homoscedastic, the Welch test was applied. *P*-values < 0.05 were considered statistically significant.

3. Results

3.1. Epididymal sperm concentration and motility

Epididymal sperm concentrations of five males per group and sperm motility of seven males in the control group and five males in the TCDD group were examined. Figs. 1 and 2 show the effects of TCDD exposure on sperm concentration and motility. Compared with the control group, the TCDD group showed a lower level of caudal epididymal sperm concentration [Control: $4.28 \pm 0.79 \times 10^7$/ml; TCDD: $2.29 \pm 0.70 \times 10^7$/ml; $P = 0.10$], because one of the mice in the TCDD group showed a very low sperm concentration. The mean value of sperm motility in the TCDD group was lower than that of the control group, but the difference was not significant [Control: 45.39 ± 7.19%; TCDD: 42.48 ± 5.71%; $P = 0.77$].

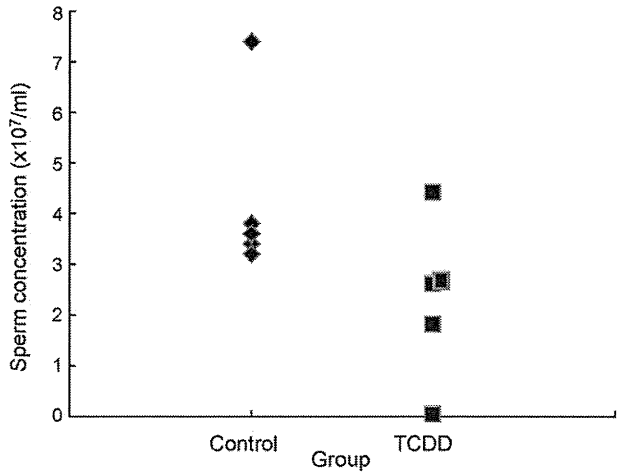

Fig. 1. Concentration of epididymal spermatozoa. The sperm concentration of the TCDD group was lower than that of the control group (Control: $4.28 \pm 0.79 \times 10^7$/ml; TCDD: $2.29 \pm 0.70 \times 10^7$/ml; $P = 0.10$). The sample size was five per group ($n = 5$).

3.2. Y-bearing/X-bearing sperm ratio

The results of quantitative PCR are shown in Fig. 3A–C. The control group consisted of 20 male mice and the TCDD group consisted of 14 male mice. Although the Y-bearing/X-bearing sperm ratio of the TCDD group was lower than that of the control group [Control: 2.68 ± 0.15, TCDD: 2.36 ± 0.04; $P = 0.060$], the difference was not significant. In addition, the TCDD group tended to have lower levels of Sry DNA concentrations [Control: 28.12 ± 1.20 ng/μl, TCDD: 25.80 ± 0.61 ng/μl ($P = 0.096$)], which is a sensitive and specific marker of the Y chromosome. This difference also did not attain statistical significance. The DNA concentrations of AR showed no differences between the control and TCDD groups [Control: 10.87 ± 0.57 ng/μl, TCDD: 10.95 ± 0.29 ng/μl, $P = 0.895$].

3.3. Y-bearing sperm ratio examined by FISH method

A total of 10,083 spermatozoa from five males per group were scored with the FISH technique (Fig. 4). As shown in Table 1, no significant differences were found between the groups. The sperm

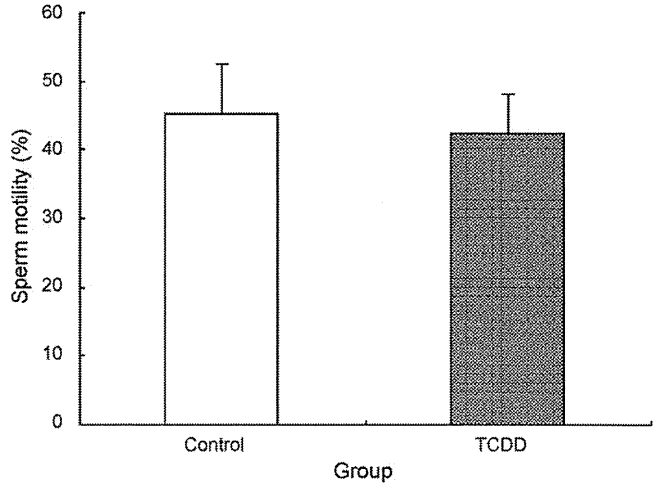

Fig. 2. Motility of epididymal spermatozoa. There were no apparent changes in the motility of the cauda epididymal spermatozoa in the TCDD group compared with the control group ($P = 0.77$). However, the mean value in the TCDD group was lower than that in the control group (Control: 45.39 ± 7.19%; TCDD: 42.48 ± 5.71%). The sample size of the control group was seven, and that of the TCDD group was five. Each histogram and bar indicate mean ± SE.

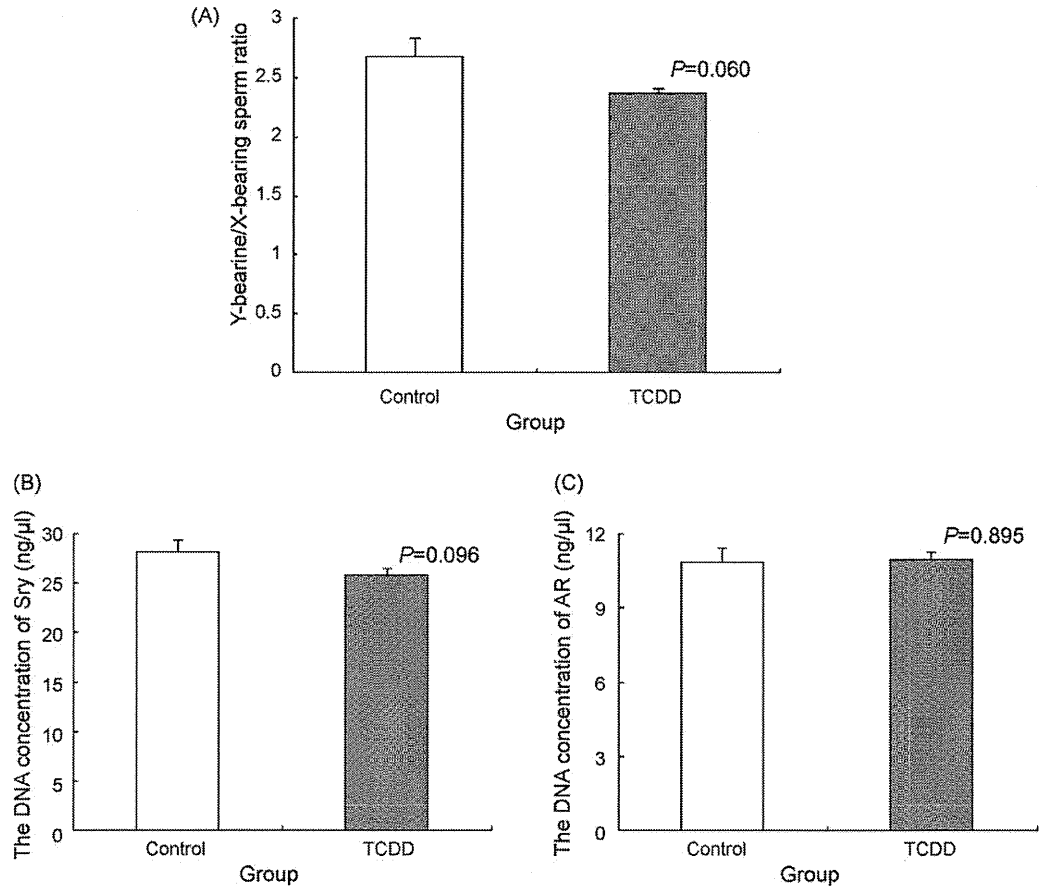

Fig. 3. (A) Y-bearing/X-bearing sperm ratio. The Y-bearing/X-bearing sperm ratio in the TCDD group did not show a significant decrease compared with the control group. However, the Y-bearing/X-bearing sperm ratio of the TCDD group was lower than that of the control group. The sample size of the control group was 20, and that of the TCDD group was 14. (B) The concentrations of the sperm DNA of Sry. The Sry primer was used to detect Y-bearing sperm. The concentrations of sperm DNA of Sry in the TCDD group were lower levels than those of the control group. (C) The concentrations of the sperm DNA of AR. The AR primer was used to detect X-bearing sperm. The concentrations of AR showed no apparent changes between the control group and the TCDD group. Each histogram and bar indicate mean ± SE.

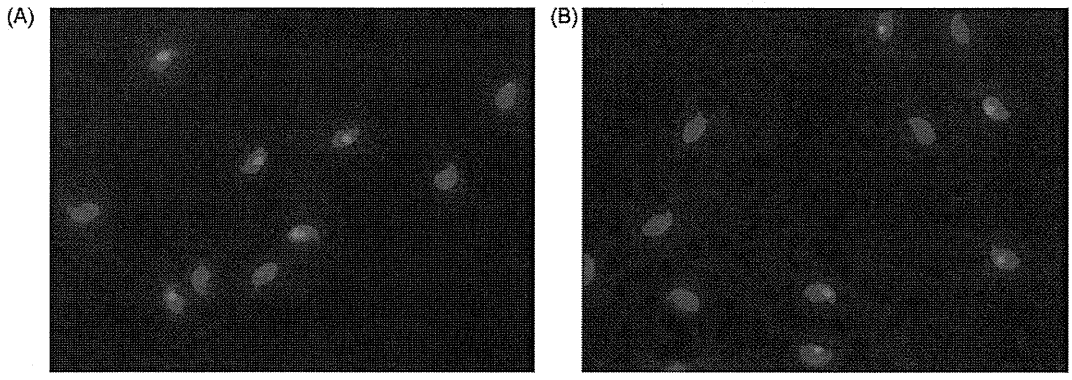

Fig. 4. Microscope images of FISH in epididymal spermatozoa. Samples were hybridized with chromosome-specific painting DNA probes for Y chromosomes (Cy3/red). The sperm nuclei were stained with DAPI and appear as blue. (A) Sperm samples of the control group yielded nearly equal ratios of X and Y chromosome-bearing cells. (B) Microscope images of FISH in epididymal spermatozoa of the TCDD group. There was no significant change in the Y-bearing sperm ratio of the TCDD group compared with the control group. The sperm nuclei were stained with DAPI.

Table 1
Y-bearing sperm ratio examined by FISH method.

Group	Control						TCDD					
Animal no.	1	2	3	4	5	Total	1	2	3	4	5	Total
Cells scored	1055	1071	1080	1086	1068	5360	1074	1079	1094	308	1168	4723
Y sperm	532	539	541	553	536	2701	539	540	549	149	586	2363
X sperm	523	532	539	533	532	2659	535	539	545	159	582	2360
Y-bearing sperm ratio (%)	50.43	50.33	50.09	50.92	50.19	50.39	50.19	50.05	50.18	48.38	50.17	50.03

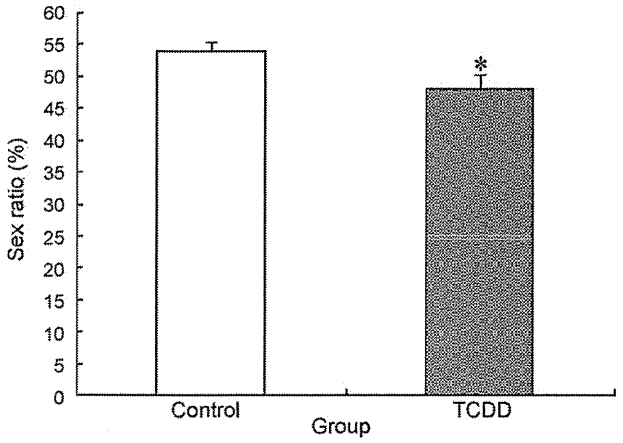

Fig. 5. Sex ratio of 2-cell embryos. A total of 614 embryos of the control group and 646 embryos of the TCDD group were examined. The proportion of XY 2-cell embryos from the TCDD-exposed sires was significantly lower than that of the control group. Each histogram and bar indicate mean ± SE. *: $P < 0.05$.

count scored with the FISH technique from one of the five mice in the TCDD group was very low, because the total number of sperm collected from the cauda epididymides of that mouse was much lower than that of any of the others.

3.4. Sex ratio of 2-cell embryos

The sex ratio of the 32 litters from the control group (614 embryos) and 30 litters from the TCDD group (646 embryos) were sexed by nested PCR using XY embryo-specific primers (Fig. 5). We found a statistically significant decrease in the sex ratio of embryos in the TCDD group compared with the control group [Control: 53.95 ± 1.54%, TCDD: 47.92 ± 2.20%, $P < 0.05$].

4. Discussion

In theory, the sex allocation ratio of many species of mammals is almost 1:1. However, in reality this is not necessarily true. There have been many reports about changes in the sex ratio of human offspring. Severe periconceptional life events have been found to reduce the sex ratio in offspring [10]. The proportion of male offspring decreases with increasing parental age, and is higher in white people than in black people [11]. The sex ratio varies with the coital rate and with the time taken to achieve conception [12]. The proportion of male offspring to female offspring decreased after the Kobe earthquake [13]. Among these reports about alterations in the sex ratio of offspring, reports that TCDD exposure altered the ratio have attracted attention. A decreased male/female sex ratio among children born to males exposed to TCDD at a relatively young age compared with unexposed males has been reported in Seveso, Italy [4,5] and Ufa, Russia [6]. Still other reports found no significant association between paternal serum TCDD levels and the sex ratio of offspring in the USA [14] and Japan [15]. In animal experiments, Ikeda et al. [16] showed that TCDD exposure to male rats *in utero* significantly decreased the number of male offspring. In contrast, Rowlands et al. [17] showed that the number of rat male offspring was not decreased by *in utero* TCDD exposure. Thus, based on these previous articles, it was unclear whether or not TCDD exposure altered the sex ratio of offspring, and no experiments re-creating the Seveso incident had been conducted. Therefore, in our previous study we exposed two different doses of TCDD only to sexually mature young male mice (F0) and examined TCDD's effect on the sex ratio of their offspring (F1) at birth [7]. The results showed that the male/female sex ratio of the offspring dose-dependently decreased in the TCDD groups, and that of the high-dose group significantly decreased. Despite changes in the sex ratio, no alteration was found in the litter size of the TCDD group compared to the control group [7]. The data suggest that direct paternal TCDD exposure, not *in utero* exposure, decreases the sex ratio of offspring, and that the sex ratio of offspring already changes before implantation signs are seen. However, the timing of the alteration of the sex ratio of offspring in TCDD-exposed male parents remained unclear, so we investigated TCDD exposure's effect on the Y-bearing/X-bearing sperm ratio and the sex ratio of 2-cell embryos.

In the present study, the Y-bearing/X-bearing sperm ratio was examined by using a modified version of Parati's method [18] of quantitative PCR. This ratio did not show a significant decrease between the control group and the TCDD group. However, the Y-bearing/X-bearing sperm ratio and DNA concentrations of Sry of the TCDD group tended to decrease, with $P = 0.06$. On the other hand, the DNA concentration of AR was not affected by TCDD exposure. The Y-bearing sperm ratio was also checked by the FISH method, which revealed no marked differences between the groups. This result supports the findings of real-time PCR.

In this study, sperm motility was not significantly decreased by TCDD administration, but the mean value of that of the TCDD group was lower than that of the control group. In addition, the sperm concentration of the TCDD group was lower than those of the control group, with $P = 0.10$. Mocarelli et al. [19] reported that exposure to TCDD in infancy significantly reduced sperm concentration and motility, whereas an opposite effect was seen with exposure during puberty in Seveso, Italy. However, in the present study, the sperm concentration and motility were not significantly decreased. Our results did not coincide with those of a report on humans [19]. The causes of difference between ours and humans are that the parameters from sperm samples collected 22 years after the accident were measured in the human study, whereas we exposed TCDD to male mice from 7 to 12 weeks of age after puberty and examined their sperm almost immediately (1 week) after exposure. In addition, our TCDD administration dose was much lower than that in the Seveso incident. Thus, the differences in when the sperm samples were collected and in the exposed dose between Mocarelli's study [19] and the present one might account for the lack of a significant decreases in sperm concentration and motility in the present study.

The sex ratio of the 2-cell embryos in the TCDD group showed a significant decrease compared with the control group. This indicates that the sex ratio of the TCDD group had already decreased at the 2-cell embryo stage.

Thus, TCDD exposure significantly reduced the sex ratio of the embryos and the sex ratio at birth without altering litter size. However, TCDD exposure did not affect the Y-bearing/X-bearing sperm ratio. These results suggest that the sex ratio of offspring was decreased at fertilization, and thus that the sex ratio of neonates was also decreased. In the case of paternal TCDD exposure, the number of eggs per litter was not affected, therefore we think this is because the fertility of Y-bearing sperm might have been affected by TCDD exposure, and thus more X-bearing spermatozoa than Y-bearing spermatozoa were fertilized. As a result, both the sex ratio of embryos and that at birth were decreased in the TCDD group with no change in litter size. This distortion of the sex ratio might be produced by epigenetic alterations induced by TCDD exposure [20,21]. However, the mechanisms of decrease in the sex ratio, especially in the fertility of Y-bearing spermatozoa, remain to be investigated.

In conclusion, this study suggests that TCDD exposure to only male mice significantly decreases the sex ratio of 2-cell embryos without altering the Y-bearing/X-bearing sperm ratio. Our previous study revealed that the sex ratio at birth was significantly decreased by TCDD exposure despite no difference in litter size [7]. That finding, together with the current results, leads us to conclude that the sex ratio of offspring was decreased at fertilization in the TCDD exposure cases. Further investigation of the

differences in fertility between Y-bearing and X-bearing spermatozoa is needed.

Conflicts of interest

The authors have no conflicts of interest that would have inappropriately influenced the work presented in this manuscript.

Acknowledgments

The authors would like to express their heartfelt thanks to Prof. Chiharu Tohyama (The University of Tokyo) for his excellent advice during this research.

This work was supported in part by Grants-in-Aid for Scientific Research (B) (15390510, 18380089) and for Scientific Research on Priority Areas (1) (14042260) from the Ministry of Education, Culture, Sports, Science and Technology of Japan to N.H.

References

[1] Savitz DA. Paternal exposures and pregnancy outcome: miscarriage, stillbirth, low birth weight, preterm delivery. In: Olshan AF, Mattison DR, editors. Male-mediated developmental toxicity. New York: Plenum Press; 1994. p. 177–96.
[2] Olshan AF, Schnitzer PG. Paternal occupation and birth defects. In: Olshan AF, Mattison DR, editors. Male-mediated developmental toxicity. New York: Plenum Press; 1994. p. 153S–67S.
[3] Tamminga J, Koturbash I, Baker M, Kutanzi K, Kathiria P, Pogribny IP, et al. Paternal cranial irradiation induces distant bystander DNA damage in the germline and leads to epigenetic alterations in the offspring. Cell Cycle 2008;7:1238–45.
[4] Mocarelli P, Brambilla P, Gerthoux PM, Patterson Jr DG, Needham LL. Change in sex ratio with exposure to dioxin. Lancet 1996;348:409.
[5] Mocarelli P, Gerthoux PM, Ferrari E, Patterson Jr DG, Kieszak SM, Brambilla P, et al. Paternal concentrations of dioxin and sex ratio of offspring. Lancet 2000;355:1858–63.
[6] Ryan JJ, Amirova Z, Carrier G. Sex ratios of children of Russian pesticide producers exposed to dioxin. Environmental Health Perspectives 2002;110:699–701.
[7] Ishihara K, Warita K, Tanida T, Sugawara T, Kitagawa H, Hoshi N. Dose paternal exposure to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) affect the sex ratio of offspring? The Journal of Veterinary Medical Science 2007;69:347–52.
[8] 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. Toxicology Letters 2005;160:22–33.
[9] Kunieda T, Xian M, Kobayashi E, Imamichi T, Moriwaki K, Toyoda Y. Sexing of mouse perimplantation embryos by detection of Y chromosome-specific sequences using polymerase chain reaction. Biology of Reproduction 1992;46:692–7.
[10] Hansen D, Møller H, Olsen J. Sever periconceptional life events and the sex ratio in offspring: follow up study based on five national registers. British Medical Journal 1999;319:548–9.
[11] Mathews TJ, Hamilton BE. Trend analysis of the sex ratio at birth in the United States. National Vital Statistics Reports 2005;53:1–17.
[12] James WH. The variations of human sex ratio at birth with time of conception within the cycle, coital rate around the time of conception, duration of time taken to achieve conception, and duration of gestation: a synthesis. Journal of Theoretical Biology 2008;255:199–204.
[13] Fukuda M, Fukuda K, Shimizu T, Møller H. Decline in sex ratio at birth after Kobe earthquake. Human Reproduction 1998;13:2321–2.
[14] Schnorr TM, Lawson CC, Whelan EA, Dankovic DA, Deddens JA, Piacitelli LA, et al. Spontaneous abortion, sex ratio, and paternal occupational exposure to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin. Environmental Health Perspectives 2001;109:1127–32.
[15] Yoshimura T, Kaneko S, Hayabuchi H. Sex ratio in offspring of those affected by dioxin and dioxin-like compounds: the Yusho, Seveso and Yucheng incidents. Occupational and Environmental Medicine 2001;58:540–1.
[16] Ikeda M, Tamura M, Yamashita J, Suzuki C, Tomita T. Repeated in utero and lactational 2,3,7,8-tetrachlorodibenzo-*p*-dioxin exposure affects male gonads in offspring, leading to sex ratio changes in F2 progeny. Toxicology and Applied Pharmacology 2005;206:351–5.
[17] Rowlands JC, Budinsky RA, Aylward LL, Faqi AS, Carney EW. Sex ratio of the offspring of Sprague-Dawley rats exposed to 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) in utero and lactationally in a three-generation study. Toxicology and Applied Pharmacology 2006;216:29–33.
[18] Parati K, Bongioni G, Aleandri R, Galli A. Sex ratio determination in bovine semen: a new approach by quantitative real time PCR. Theriogenology 2006;66:2202–9.
[19] Mocarelli P, Gerthoux PM, Patterson Jr DG, Milani S, Limonta G, Bertona M, et al. Dioxin exposure, from infancy through puberty, produces endocrine disruption and affects human semen quality. Environmental Health Perspectives 2008;116:70–7.
[20] Crews D, McLachlan JA. Epigenetics, evolution, endocrine disruption, health, and disease. Endocrinology 2006;147:S4–10.
[21] Crews D, Gore AC, Hsu TS, Dangleben NL, Spinetta M, Schallert T, et al. Transgenerational epigenetic imprints on mate preference. Proceedings of the National Academy of Sciences of the United States of America 2007;104:5942–6.

DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2011, Article ID bar046, doi:10.1093/database/bar046

Original article

CELLPEDIA: a repository for human cell information for cell studies and differentiation analyses

Akiko Hatano[1,2], Hirokazu Chiba[1], Harry Amri Moesa[1,3], Takeaki Taniguchi[4], Satoshi Nagaie[1,2], Koji Yamanegi[5], Takako Takai-Igarashi[7], Hiroshi Tanaka[7] and Wataru Fujibuchi[1,6,7,*]

[1]Computational Biology Research Center, Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan, [2]Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan, [3]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan, [4]Advanced Business Division, Mitsubishi Research Institute, Inc., 2-10-3 Nagatacho, Chiyoda-ku, Tokyo 100-8141, Japan, [5]Department of Clinical Pathology, Hyogo College of Medicine, 1-1 Mukogawa-cho, Nishinomiya, Hyogo 663-8501, Japan, [6]Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan and [7]Department of Computational Biology, Graduate School of Biomedical Science, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan

***Corresponding author:** Tel: +81 3 3599 8619; Fax: +81 3 3599 8081; Email: w.fujibuchi@aist.go.jp

Submitted 13 February 2011; Revised 28 July 2011; Accepted 21 September 2011

CELLPEDIA is a repository database for current knowledge about human cells. It contains various types of information, such as cell morphologies, gene expression and literature references. The major role of CELLPEDIA is to provide a digital dictionary of human cells for the biomedical field, including support for the characterization of artificially generated cells in regenerative medicine. CELLPEDIA features (i) its own cell classification scheme, in which whole human cells are classified by their physical locations in addition to conventional taxonomy; and (ii) cell differentiation pathways compiled from biomedical textbooks and journal papers. Currently, human differentiated cells and stem cells are classified into 2260 and 66 cell taxonomy keys, respectively, from which 934 parent–child relationships reported in cell differentiation or transdifferentiation pathways are retrievable. As far as we know, this is the first attempt to develop a digital cell bank to function as a public resource for the accumulation of current knowledge about human cells. The CELLPEDIA homepage is freely accessible except for the data submission pages that require authentication (please send a password request to cell-info@cbrc.jp).

Database URL: http://cellpedia.cbrc.jp/

Introduction

The human body consists of more than 10 trillion cells that have highly diverse structures and functions and play specific physiological roles (1). Many histology and molecular biology textbooks state that there are ~200 types of cells in the adult human body (2). However, this classical observation is mainly based on the morphological or histological perspective, and no cell classification system based on gene expression levels has yet been systematically established. This sometimes causes confusion in modern cell biology. Indeed, there are cells that have the same name but show different characteristics. One example is 'fibroblasts', which are found in multiple tissues but are known to have different characteristics, such as cell differentiation capacities, depending on their lineages and surrounding conditions (3). Similarly, the existence of different gene expression levels among cells found in slightly different locations, such as the proximal and distal parts of the colon, has been reported (4). In stem cell research, it is recently reported that only a particular group of sinusoidal endothelial cells in the bone marrow with distinct gene expression patterns regulate the properties of hematopoietic stem cells (5). Recently, several efficient strategies for such studies, including microarray

and other high-throughput technologies at single-cell resolution, have been developed to detect minute differences between cells (6, 7). Gene expression analysis with these technologies may reveal the different characteristics of individual cells in the same or different tissues in the near future.

Extensive studies of cell differentiation using embryonic stem (ES) or induced pluripotent stem (iPS) cells have made remarkable progress (8, 9) and have generated cells that possess characteristics different from those of naturally occurring cells in the human body (10). Some reports have shown that several types of cells can be directly generated by reprogramming from other differentiated cells (11, 12). Therefore, a more detailed and well-structured classification system is urgently required to provide reference cells with which to characterize artificially produced cells or distinguish functionally similar but differently localized cells. In addition, there is a compelling need to accumulate and provide cell differentiation or transdifferentiation information for both naturally existing and artificially produced cells to further enhance studies of design methods or routes to produce particular types of cells.

Inspired by the importance of cell studies early on, we developed a gene expression profile data search system called 'CellMontage' in 2004, where we also provided a draft version of a cell catalogue (13). In the previous version, we used classical cell taxonomy and stored cell illustrations for ~200 cell types. In CELLPEDIA, a new and detailed cell classification system is proposed by combining conventional taxonomy with physical mappings, allowing the unique identification of any naturally existing cells. With the improved taxonomy as its backbone, CELLPEDIA provides a repository for cell image, gene expression and cell differentiation data, serving as a digital dictionary of human cells for the biomedical field. This database is expected to become an increasingly important tool for cell-related research, including regenerative medicine and cell therapies, in the near future.

General features of CELLPEDIA

Database structure

The latest version of CELLPEDIA consists of two major cell categories: 'differentiated cell' and 'stem cell'. The 'differentiated cell' category contains information of fully differentiated or mature cells that represent specific functions or characteristics, whereas the 'stem cell' category contains information of undifferentiated or immature cells, such as progenitor or multipotent stem cells. Each cell entry in either category contains three primary (i.e. raw) data: (i) digital images of cells and tissues; (ii) gene expression data; and (iii) journal articles on cell functions and differentiations. The cell images and the gene expression data are basically collected via submission pages from either laboratories or public data. The journal articles are manually compiled from PubMed (14) by the authors. Currently, 365 images, 878 gene expression profiles and 336 articles for 'differentiated cell' are registered. For 'stem cell', 36 images, 92 gene expression profiles and 83 articles are registered. These primary data are linked to each other by our own cell taxonomy keys (explained in the 'Enhanced cell classification system' section) in the relational database structure. All the three types of primary data cited above are retrievable from an integrated page, together with additional secondary information that includes (iv) image parameters regarding cell shape; (v) the Open Biomedical Ontology (OBO) Cell Type Ontology; and (vi) cell differentiation neighborhood links (Figure 1).

To access any data easily, a keyword search box on the top page is available for our cell taxonomy, all the primary and OBO Cell Type Ontology information cited above. Figure 1a shows an example of a search result that is queried by the keywords 'hepatic, lobule, liver' in the search target 'human cell taxonomy'. A user can select a favorite entry from the list of retrieved cell types and an integrated page for the selected cell is shown by clicking the Image ID. The locations of cells or tissues are shown by voxel models of male and female human bodies provided by National Institute of Information and Communications and Technology (NICT) (15). We also provide precalculated gene modules or biclusters extracted from the collected gene expression data described above using our software program called System for Assembling Modules by Ultra Rapid Algorithm on Itemsets (SAMURAI) (16). A gene module consists of a subset of genes and a subset of experiments, and SAMURAI exhaustively extracts gene modules that share common gene expression patterns in both query and gene expression databases (17).

The CELLPEDIA database system is developed and implemented in Perl CGI script on the MySQL relational database. The current Web server runs with Apache daemon on a Linux system. The database schema is designed so that as many as 15 tables in the database can be interconnected by the backbone of the cell taxonomy keys (differentiated cell IDs or stem cell IDs) and other branch group keys, such as Image IDs, Expression IDs, etc. A simplified database schema is shown in Figure 2.

Enhanced cell classification system

In typical textbooks, cells with equivalent or similar functions are given the same name even if they exist in different physical locations. For instance, 'goblet cells' are found in the nose, lung and small and large intestines. Similarly, 'fibroblasts' exist in many connective tissues, including loose, dense and reticular tissues that have different properties. These cells may have different gene expression

patterns that are attributable to various surrounding conditions (e.g. pH, ECM components, mechanical tension, etc.) (18–21). Particularly, ECM-associated molecules and their mechanical properties often influence cell shape and differentiation. These properties are particularly important in the field of regenerative medicine (22–25). Figure 3 shows two types of cells and examples of their different physical locations. Table 1 summarizes the different physical locations of six well-known cell types: fibroblast cells, smooth muscle cells, epithelial cells, endothelial cells, nerve cells and goblet cells. For example, fibroblast cells are found in as many as 297 physical locations, whereas goblet cells are found in 30 locations.

CELLPEDIA defines its own enhanced cell taxonomy for physically distinct cells, establishing a hierarchical structure of up to eight levels from organs (or tissues) to cells. It presents a combination of both conventional cell taxonomy and their physical locations (Figure 1a). In total, 2260 types of differentiated human cells are listed and assigned unique cell taxonomy keys. Similarly, stem cells and progenitor cells are classified into 66 taxonomy keys. Some of the differentiated cells and stem cells are interconnected based on the cell differentiation information by taxonomy keys.

In order to verify our own hierarchical taxonomy, we use OBO UBERON (26, 27), which is a multi-species anatomy ontology, to check the agreement of classification names and structures. Among the non-redundant 551 hierarchical node terms (excluding the cell name itself) in the eight levels of CELLPEDIA taxonomy, 505 (91.7%) exactly correspond to UBERON anatomical terms. The remaining 46 terms are not found in UBERON exactly; nevertheless, most of them belong to subclass of the end terms of UBERON, i.e. they are mapped to more detailed portions of the human anatomy than those of UBERON. The 46 terms are listed in Supplementary Table S1.

We also compare the orders of hierarchical node terms for all of the 2260 differentiated cells in CELLPEDIA with those in UBERON. We find that as few as 1259 (55.7%) hierarchical paths match the OBO 'part_of' or 'is_a' relationships in the UBERON anatomical terms. The main reason for this low consistency is the vascular system; CELLPEDIA distinguishes each blood vessel according to physical location, whereas UBERON recognizes only one independent blood system. If we exclude the vascular system, as many as 1951 (86.3%) paths would match the OBO 'part_of' or 'is_a' relationships in the UBERON anatomical terms. If we allow one mismatch term in the path, almost all of the 2227 (98.5%) paths would match. From these observations, we conclude that CELLPEDIA taxonomy is highly compatible with the UBERON anatomical terms and their relationships are thus reliable and useful.

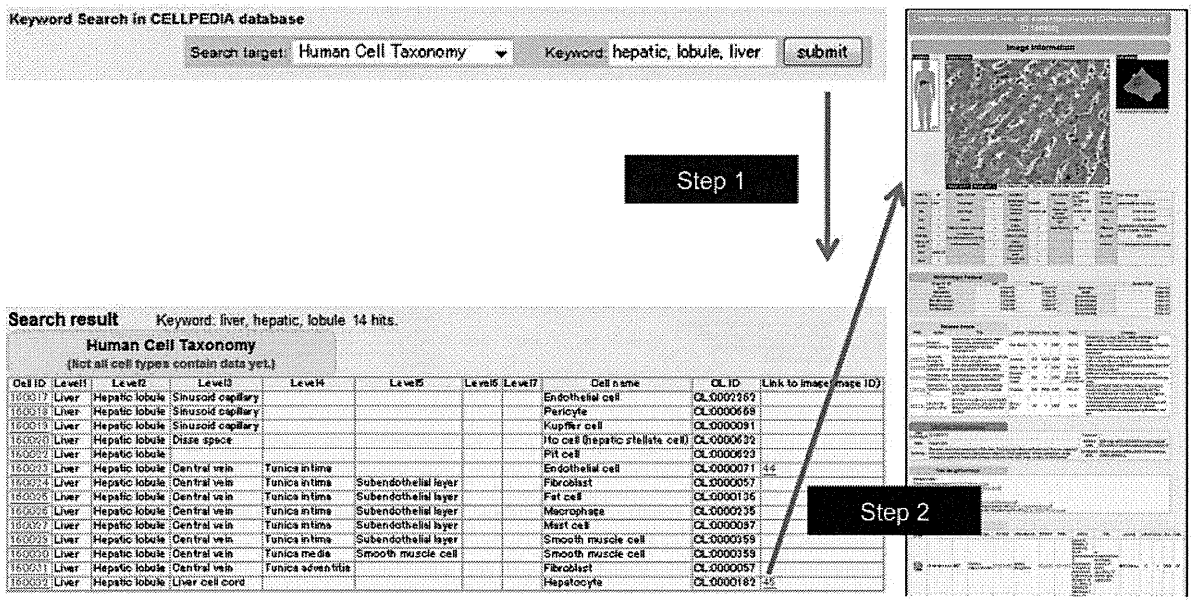

Figure 1. Example of keyword search and integrated page of hepatocyte (differentiated cell ID: 160032). (**a**) Retrieval of cell taxonomy entries by keyword search (keywords: hepatic, lobule, liver) is shown (Step 1). Clicking on an image ID in the cell taxonomy page will lead to the integrated page (Step 2). The integrated page is composed of three types of primary data: (**b**) digital images of cells (and tissues); (**c**) gene expression data; and (**d**) journal articles. Additional secondary information is also provided, including (b) image parameters regarding cell shape (table of morphologic features); (**e**) the OBO Cell Type Ontology; and (**f**) cell differentiation neighborhood links.

Database, Vol. 2011, Article ID bar046, doi:10.1093/database/bar046

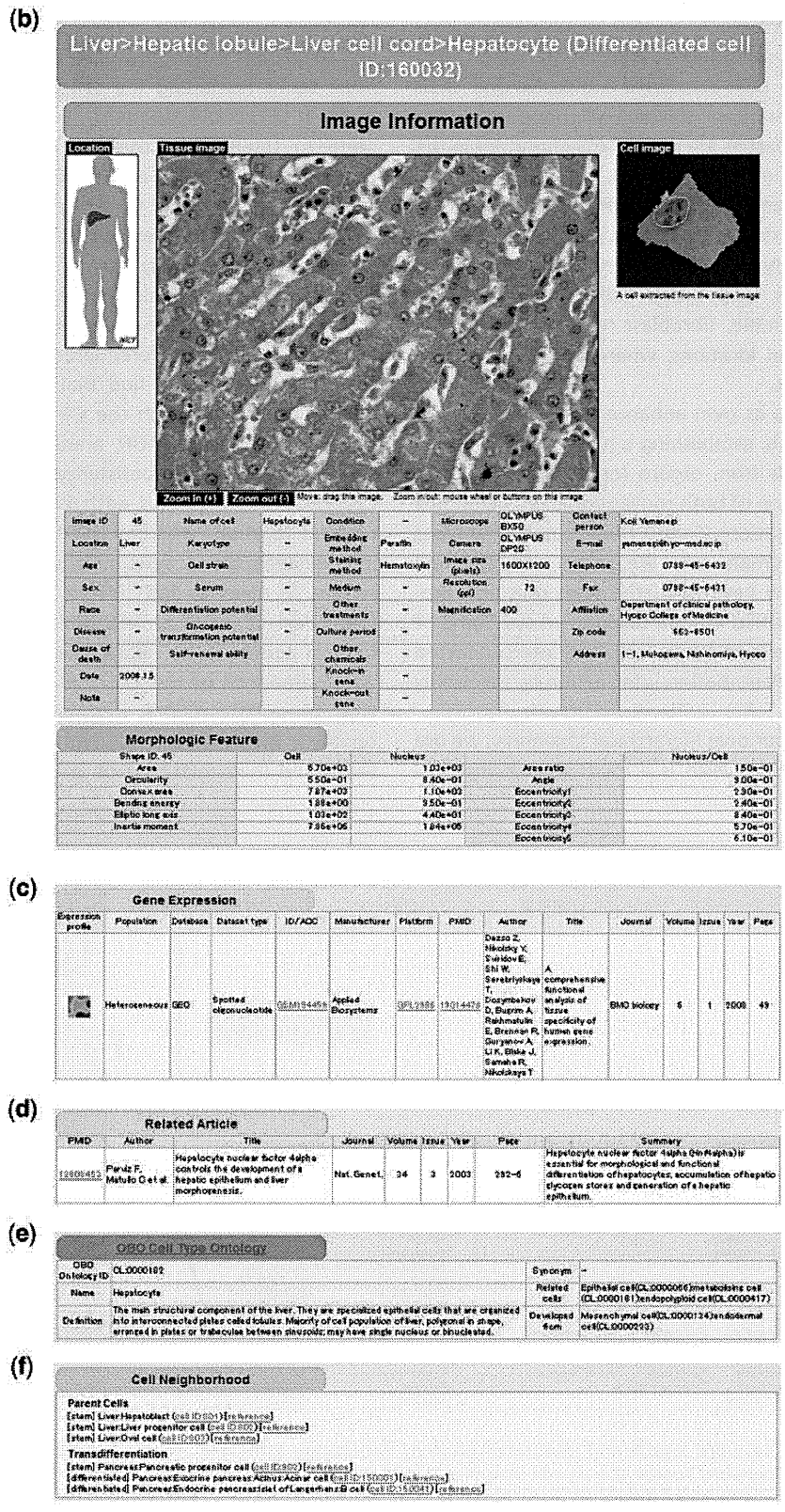

Figure 1 Continued

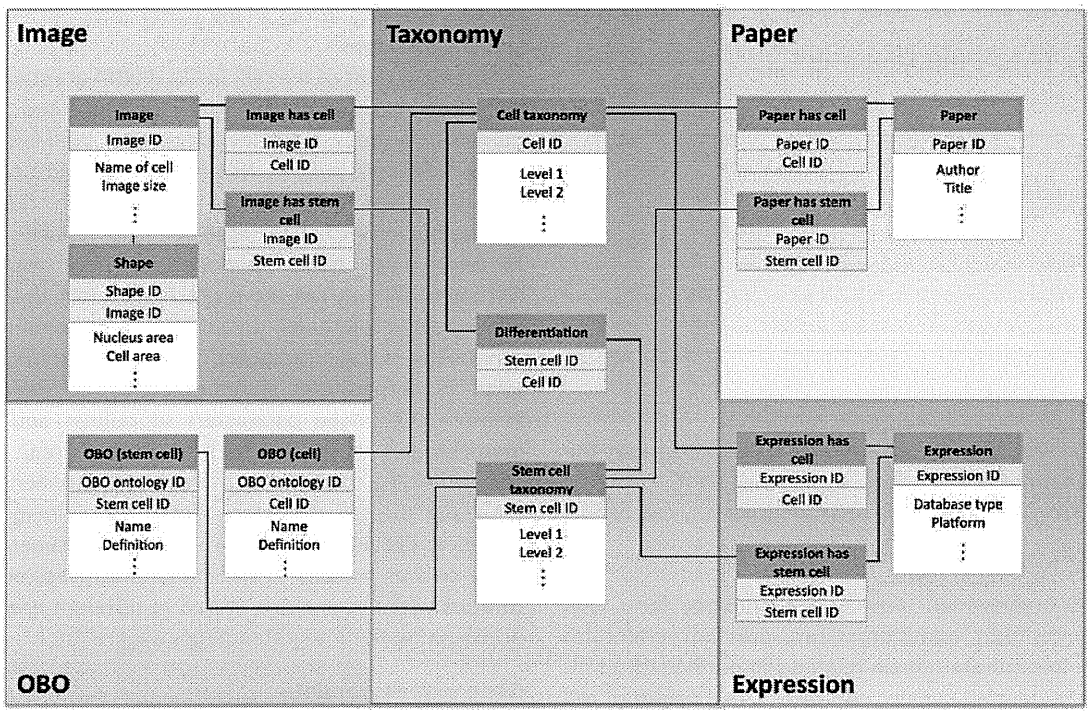

Figure 2. Simplified view of database schema used in CELLPEDIA. The database schema is designed so that as many as 15 tables in the database can be interconnected by the backbone of the cell taxonomy keys (differentiated cell ID or stem cell ID) and other branch group keys.

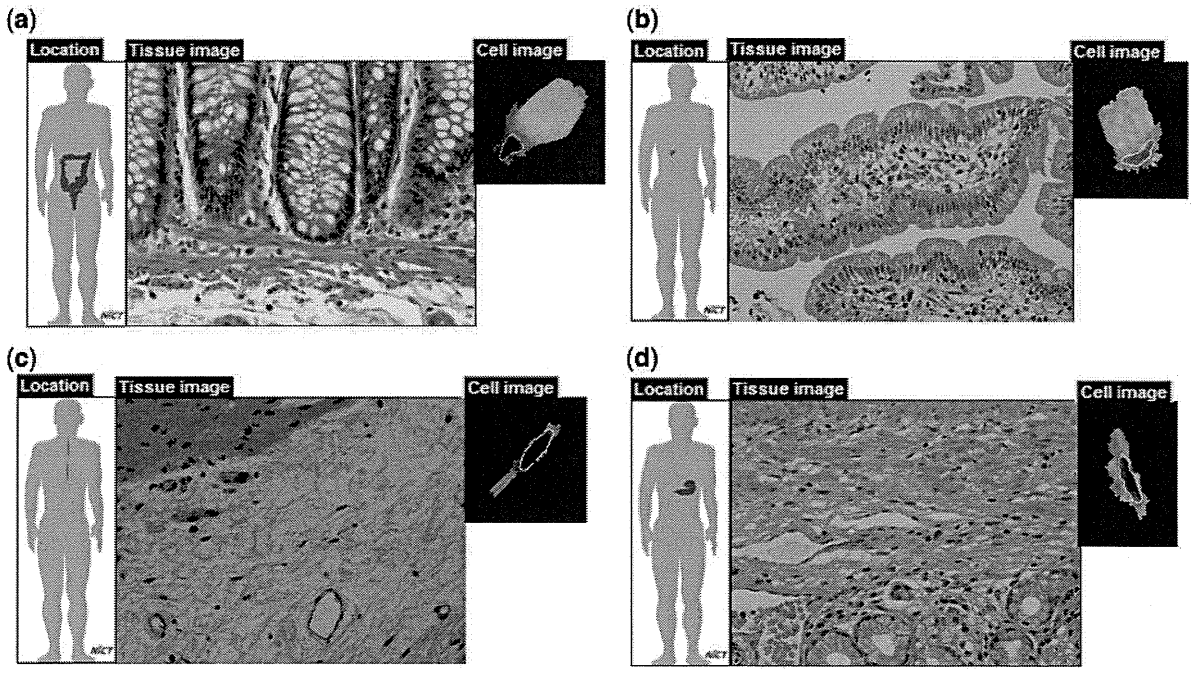

Figure 3. Examples of cells with the same names. Goblet cells of (**a**) large intestine and (**b**) small intestine; and fibroblasts of (**c**) esophageal submucosa and (**d**) pylorus.

Table 1. Cells with the same names in different physical locations

Cell name	Number of physical locations
Fibroblast	297
Smooth muscle	269
Epithelial	122
Endothelial	117
Nerve	44
Goblet	30

Note that UBERON contains multiple paths to reach the same anatomical location from the root, whereas CELLPEDIA contains a single path to locate any cell types.

OBO cell type ontology

The OBO organization provides the ontology for ~1500 types of cells from various organisms, including vertebrates (28). We incorporate human-related cell ontological terms into our integrated pages to provide formal names of cell types under the defined vocabulary. Currently, 2240 (99.5%) and 48 (72.7%) taxonomy keys for differentiated cells and stem cells, respectively, are linked to 269 ontological terms of OBO cell types and displayed on the integrated pages (Figure 1e).

Cell images and morphological parameters

We have collected cell images corresponding to our human cell taxonomy keys to provide visual information about the cells. The cell images are primarily collected by submission from various laboratories and manually checked by the curators for their quality and tissue origins. Currently, we have 365 images for differentiated cells and 36 images for stem cells registered in the database. Nevertheless, when we count the number of cell types containing images in a conventional way, i.e. counting the cell types excluding the redundant names, we find that 57 of a total of 170 differentiated cell types are already registered.

As the resolutions of the collected cell images vary, we provide zoom-in and zoom-out functions for the tissue images to enable the user to choose the proper scale on the integrated pages. A magnified image of the original image is also shown on the right side. Image attributes provided by the submitters, such as donor information, cell culture conditions, microscope platforms, and contact persons, are also collected. Some additional attributes for stem cells are also provided, such as types of feeder cells, differentiation ability, self-renewal ability and so on (Figure 1b). It is noteworthy that multiple images from various submitters for the same cell taxonomy key are stored to record variability in cell shape under different conditions and environments.

Currently, most of the cells in CELLPEDIA are stained with hematoxylin and eosin (HE), which is mainly used to distinguish the nucleus from the cytoplasm. Well-known cell image analyzers, such as NIH Image (29) and CellProfiler (30), recognize cell outlines very well in fluorescent images but often fail to recognize cells stained with HE. Therefore, we have developed a new tool called 'Cytometrica' that can more precisely detect the outlines of HE-stained cells based on a dynamic programming algorithm (Moesa,H.A., Taniguchi,T. and Fujibuchi,W., manuscript in preparation). Cytometrica is also designed to measure the image parameters of cellular and nuclear morphologies, such as the surrounding area, bending energy, inertia moment, and area ratio between the nucleus and the cell. Therefore, when a raw image is submitted to CELLPEDIA, the morphological parameters of as many cells as possible in the image are measured by the curators with Cytometrica, stored as different entries, and displayed on the integrated pages (Figure 1b).

Gene expression data

We have collected 878 sets of gene expression data for differentiated cells and 92 sets for stem cells from such public databases as the Gene Expression Omnibus (31) and ArrayExpress (32). We also provide a data submission page for gene expression data for future use. Due to limitations in the current techniques for single-cell analysis and available data, we have so far accumulated gene expression data obtained from tissues or multiple cells. To avoid confusion, we have added a 'Population' category in the gene expression data table to provide information of the data resource. 'Heterogeneous' means that the gene expression data are measured from a heterogeneous cell population. 'Multiple' means that the data are measured from unique but multiple cells. 'Single' means that the data are measured from a single cell.

To visualize our gene expression data, we generate self-organization maps using all the collected data (Figure 1c), which are cross-referenced with Entrez gene IDs. We also generate a list of genes that are up- and down-regulated in each cell type and make them available for downloading. In addition, tissues or cells with similar gene expression patterns are connected with the correlation coefficients of collected gene expression data, and in order to survey the relationships of gene expression data, the results are displayed as a minimum spanning tree in each of 'differentiated cell' and 'stem cell' top pages.

Journal articles

Reference journal articles for each cell taxonomy key are manually curated by the authors to glean existing knowledge about mammalian cells. We have so far preferentially

collected articles for cells with image data only as the database is still in its infancy. We have summarized 336 papers for differentiated cells and 86 papers for stem cells, respectively, from the perspectives of cell function and differentiation. We also provide hyperlinks to corresponding entries in PubMed on the integrated pages. In CELLPEDIA, the contents of the journal articles are represented by a concise sentence incorporating the most important information, so that the user can instantly access biological functions or differentiation factors of cells (Figure 1d).

Applications of CELLPEDIA

Gene expression similarity

Although not yet available at the single-cell level, the gene expression data compiled in CELLPEDIA obviously provide important information about the gene expression profiles of cells. We have developed and maintained a matching system for gene expression data called 'CellMontage' in our laboratory (33), and have added the CELLPEDIA dataset to the database selection menu on the CellMontage homepage (http://cellmontage.cbrc.jp/) so that the user can check whether the expression profiles of artificially produced cells are similar to the profiles of naturally existing cells. As we consider CellMontage to be a part of CELLPEDIA project, the 'Profile Matching' and 'Profile Retrieval' pages of CellMontage are now available in the 'Cell Analysis Tools' in CELLPEDIA. We also plan to release a new version of CellMontage, which will allow more accurate ranking with a machine-learning-based approach (to be published elsewhere). This approach will be quite useful in predicting the characteristics of iPS cells, such as multipotency, tumorigenicity and cell stability.

Cell neighborhood

It would be useful if we could retrieve accumulated information on cell differentiation or transdifferentiation in a deductive way, to explore new pathways for the production of cells. CELLPEDIA currently provides 934 binary (parent–child) relationships involved in cell differentiation, collected from textbooks and journals, as the 'cell neighborhood' (Figures 1f and 4).

Figure 4a shows an example of deducing cell differentiation or transdifferentiation pathways from the integrated page of hepatocyte (differentiated cell ID: 160032) using cell neighborhood links. Three linked cells are found for both parental and transdifferentiation relationships from the hepatocyte page. It is noteworthy that distant relationships between cells can be deduced by following consecutive neighborhood links. For example, the differentiation route of liver progenitor cells to pancreatic beta cells, which is already reported (34), is determined by successively following the links from 'liver progenitor cell (stem cell ID:802, Figure 4a top) to hepatocyte (differentiated cell ID:160032, Figure 4a bottom)' and from 'hepatocyte to pancreatic beta cell (differentiated cell ID:150041, Figure 4a right)'.

As an option, the cell differentiation pathway map can be visualized from the collected binary relationships using such public software as yEd Graph Editor (yWorks software, Tubingen, Germany) or graphviz (AT&T Inc, Dallas, TX, USA). Figure 4b shows an example of such a map constructed using all of the collected 934 cell differentiation and transdifferentiation data. As a result, three large clusters are recognized. The largest one (Figure 4b top left and in the red box) is a mesenchymal stem cell cluster and its differentiated cells. The rest of the clusters are hematopoietic system (Figure 4b top middle) and endoderm (Figure 4b top right) clusters. The endoderm cluster includes most of the registered endodermal tissues, such as liver, pancreas and large and small intestines. The region that contains transdifferentiation pathways (indicated by red arrows) is shown in a magnified image (Figure 4b blue box). In the future, we expect that with further accumulated binary relationship data under a new deduction system, novel differentiation pathways will be explored and experimentally verified.

Future work

The accumulation of more data of cell locations and surrounding conditions will be increasingly important for cell differentiation and regenerative medicine. We will continue to collect more data, especially about stem cells and cell neighborhoods, as well as improve the database to make it more accessible to such information in an integrative way, in order to contribute to progress in regenerative medicine. Yet unknown but more efficient transdifferentiation pathways may be found by expanding cell neighborhood links.

Furthermore, through discussions with stem cell researchers, we plan to extend our database structure to incorporate recent iPS or ES cell data generated with modern technologies, such as next-generation sequencing and mass spectrometry. RNA-Seq, chromatin remodeling, epigenetics, lectin microarray and proteomics data are possible candidates to be incorporated in future development. We will also prepare another cell category 'cell line' and collect related data because requests are often received from biomedical researchers and pharmaceutical industries. Although this is a long-term goal, we also plan to collect pathological images and information about diseases, such as various types of cancers, in the future version of CELLPEDIA, to help in the analysis of disease mechanisms.

Regarding the technical aspect of the database system, we plan to provide a Web page for data downloading to obtain whole data for each cell type. Similarly, we plan

Database, Vol. 2011, Article ID bar046, doi:10.1093/database/bar046

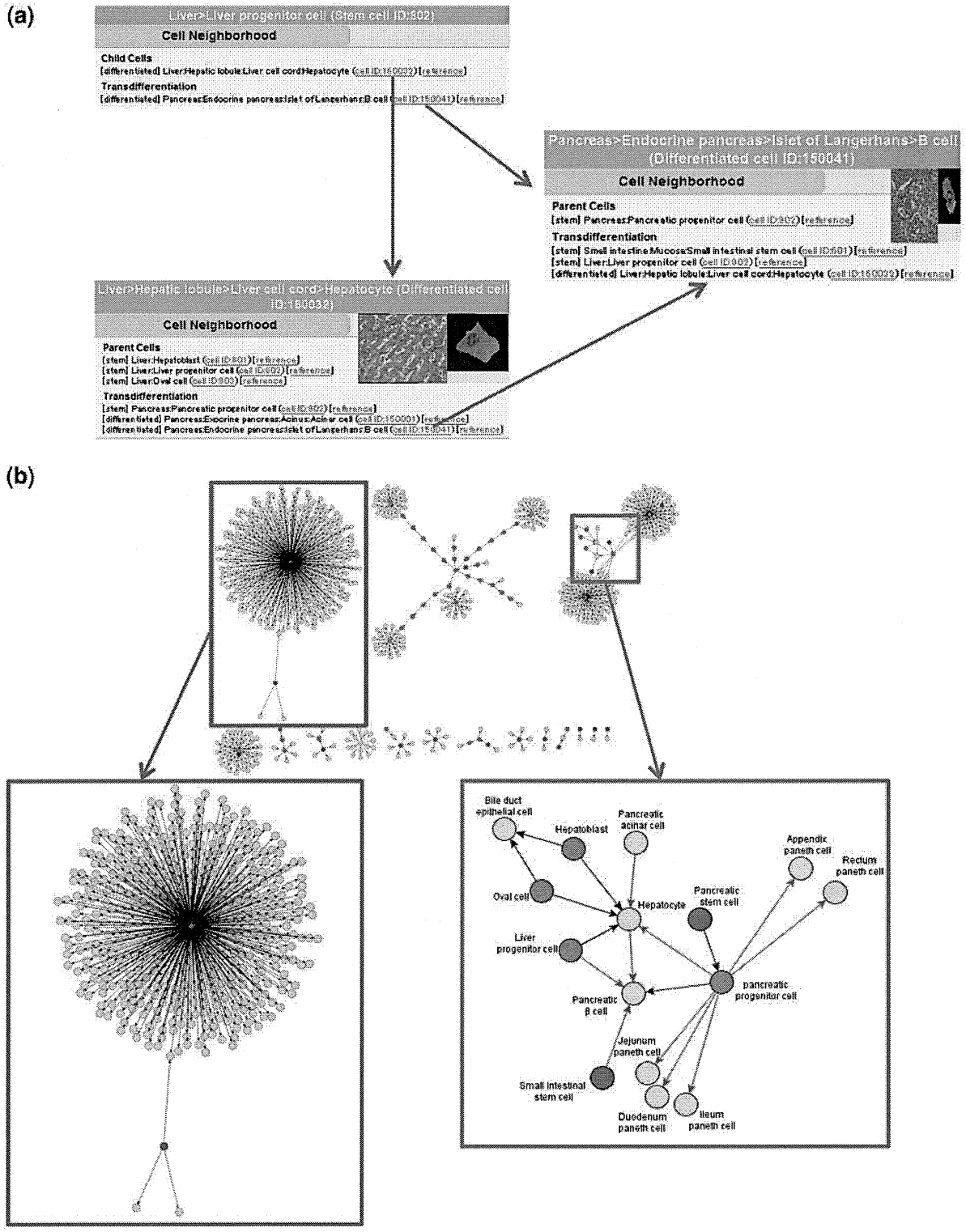

Figure 4. Cell neighborhood links. (**a**) Cell neighborhood links with the hepatocyte: cell differentiation or transdifferentiation pathways can be deductively traced. (**b**) Cell differentiation map constructed with 934 collected binary relationships. Circles filled with different colors indicate levels of differentiated cells (blue: stem cells, wine red: progenitor cells and light green: differentiated cells). Black and red arrows indicate differentiation and transdifferentiation pathways, respectively. The red box shows the mesenchymal stem cell cluster in a magnified view. The blue box shows a part of the endoderm cluster in a magnified view.

to develop an application programming interface, or API, to provide CELLPEDIA entries more efficiently at the request of some pharmaceutical companies.

Availability

CELLPEDIA is freely available for academic and personal use. Commercial users are required to obtain a license when any data of CELLPEDIA are used to generate secondary products.

Supplementary Data

Supplementary data are available at *Database* Online.

Ackowledgements

The authors wish to thank Dr Makoto Asashima and Dr Hiroshi Kuriyama of Advanced Industrial Science and Technology for useful discussions on iPS and ES cell data. The authors also thank Dr Kaoru Mogushi of Tokyo Medical and Dental University for critical reading the manuscript.

Funding

AIST Research Information Database (RIO-DB) team in part. Funding for open access charge: 2011 AIST Management Expenses Grants.

Conflict of interest. None declared.

References

1. Alberts,B., Johnson,A., Lewis,J. *et al.* (2008) *Molecular Biology of the Cell*, 5th edn., Garland Science, Taylor & Francis Group, LLC, New York, NY.
2. Gartner,L.P. and Hiatt,J.L. (2001) *Color Textbook of Histology*, 2nd edn., W.B. Saunders Co, Philadelphia, PA.
3. Lorenz,K., Sicker,M., Schmelzer,E. *et al.* (2008) Multilineage differentiation potential of human dermal skin-derived fibroblasts. *Exp. Dermatol.*, **17**, 925–932.
4. LaPointe,L.C., Dunne,R., Brown,G.S. *et al.* (2008) Map of differential transcript expression in the normal human large intestine. *Physiol. Genomics*, **33**, 50–64.
5. Kobayashi,H., Butler,J.M. and O'Donnell,R. (2010) Angiocrine factors from Akt-activated endothelial cells balance self-renewal and differentiation of haematopoietic stem cells. *Nat. Cell Biol.*, **11**, 1046–1056.
6. Tang,F., Barbacioru,C., Nordman,E. *et al.* (2010) RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nat. Protoc., **5**, 516–535.
7. Kurimoto,K. and Saitou,M. (2010) Single cell cDNA microarray profiling of complex biological processes of differentiation. Curr. Opin. Genet. Dev., **5**, 470–477.
8. Takahashi,K., Tanabe,K., Ohnuki,M. *et al.* (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, **131**, 861–872.
9. Navarro-Alvarez,N., Soto-Gutierrez,A. and Kobayashi,N. (2010) Hepatic stem cells and liver development. *Methods Mol. Biol.*, **640**, 181–236.
10. Tsuji,O., Miura,K., Okada,Y. *et al.* Therapeutic potential of appropriately evaluated safe-induced pluripotent stem cells for spinal cord injury. *Proc. Natl Acad. Sci. USA*, **107**, 12704–12709.
11. Ieda,M., Fu,J.D., Delgado-Olguin,P. *et al.* (2010) Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, **142**, 375–386.
12. Zhou,Q., Brown,J., Kanarek,A. *et al.* (2008) In vivo reprogramming of adult pancreatic exocrine cells to beta-cells. *Nature*, **455**, 627–632.
13. Fujibuchi,W., Kiseleva,L., Taniguchi,T. *et al.* (2005) Development of cell knowledge base and prediction of cell types and characteristics by gene expression profiles. IPSJ SIG Technical Reports 2005-BIO-2, 33–37 (in Japanese).
14. Public PubMed database from the U.S. National Library of Medicine. National Institutes of Health. http://www.ncbi.nlm.nih.gov/pubmed/
15. Nagaoka,T., Watanabe,S., Sakurai,K. *et al.* (2004) Development of realistic high-resolution whole-body voxel models of Japanese adult males and females of average height and weight, and application of models to radio-frequency electromagnetic-field dosimetry. *Phys. Med. Biol.*, **49**, 1–15.
16. Fujibuchi,W., Kim,H., Okada,Y. *et al.* (2009) High-performance gene expression module analysis tool and its application to chemical toxicity data. Methods Mol. Biol., **577**, 55–65.
17. Okada,Y. and Fujibuchi,W. (2007) Mining a large-scale microarray database for similar gene expression modules to find distant relationships between Down syndrome and Huntington's disease. In: *The 7th International Conference for the Critical Assessment of Microarray Data Analysis* (PDF available from http://camda.bioinfo.cipf.es/camda07/).
18. Moré,J., Fioramonti,J., Bénazet,F. *et al.* (1987) Histochemical characterization of glycoproteins present in jejunal and colonic goblet cells of pigs on different diets. A biopsy study using chemical methods and peroxidase-labelled lectins. *Histochemistry*, **87**, 189–194.
19. Stenman,S. and Vaheri,A. (1978) Distribution of a major connective tissue protein, fibronectin, in normal human tissues. J. Exp. Med., **147**, 1054–1064.
20. Chavrier,C., Couble,M.L., Magloire,H. *et al.* (1984) Connective tissue organization of healthy human gingiva. Ultrastructural localization of collagen types I-III-IV. J. Periodont. Res., **19**, 221–229.
21. Camelliti,P., Borg,T.K. and Kohl,P. (2005) Structural and functional characterization of cardiac fibroblasts. Cardiovasc. Res., **65**, 40–51.
22. Guilak,F., Cohen,D.M., Estes,B.T. *et al.* (2009) Control of stem cell fate by physical interactions with the extracellular matrix. *Cell Stem Cell*, **5**, 17–26.
23. Connelly,J.T., Gautrot,J.E., Trappmann,B. *et al.* (2010) Actin and serum response factor transduce physical cues from the microenvironment to regulate epidermal stem cell fate decisions. *Nat. Cell. Biol.*, **12**, 711–718.
24. Bonfanti,P., Claudinot,S., Amici,A.W. *et al.* (2010) Microenvironmental reprogramming of thymic epithelial cells to skin multipotent stem cells. *Nature*, **466**, 978–982.

25. Gilbert,P.M., Havenstrite,K.L., Magnusson,K.E. *et al.* (2010) Substrate elasticity regulates skeletal muscle stem cell self-renewal in culture. *Science*, **329**, 1078–1081.
26. Washington,N.L., Haendel,M., Mungall,C.J. *et al.* (2009) Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.*, **7**, e1000247.
27. Meehan,T.F., Masci,A.M., Abdulla,A. *et al.* (2011) Logical development of the cell ontology. *BMC Bioinformatics*, **12**, 6.
28. Smith,B., Ashburner,M., Rosse,C. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **11**, 1251–1255.
29. The public domain NIH Image program from the U.S. National Institutes of Health. http://rsbweb.nih.gov/nih-image/
30. Carpenter,A.E., Jones,T.R., Lamprecht,M.R. *et al.* (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, **7**, R100.
31. Barrett,T., Troup,D.B., Wilhite,S.E. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res., **37**, D885–D890.
32. Parkinson,H., Kapushesky,M., Kolesnikov,N. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res., **37**, D868–D872.
33. Fujibuchi,W., Kiseleva,L., Taniguchi,T. *et al.* (2007) CellMontage: similar expression profile search server. *Bioinformatics*, **23**, 3103–3104.
34. Jin,C.X., Li,W.L., Xu,F. *et al.* (2008) Conversion of immortal liver progenitor cells into pancreatic endocrine progenitor cells by persistent expression of Pdx-1. *J. Cell Biochem.*, **104**, 224–236.

Prediction of Chemical Toxicity by Network-based SVM on ES-cell Validation System

Wataru Fujibuchi[1] w.fujibuchi@aist.go.jp
Sachiyo Aburatani[1] s.aburatani@aist.go.jp
Junko Yamane[2] yamane-j@m.u-tokyo.ac.jp
Satoshi Imanishi[2] imanishi@m.u-tokyo.ac.jp
Hiromi Akanuma[3] akanuma.hiromi@nies.go.jp
Hideko Sone[3] hsone@nies.go.jp
Seiichiroh Ohsako[2] ohsako@m.u-tokyo.ac.jp

[1] Computational Biology Research Center, Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] Center for Disease Biology and Integrative Medicine, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8654, Japan
[3] Research Center for Environmental Risk, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba 305-8506, Japan

Keywords: Bayesian network, kernel SVM, chemical toxicity, gene expression, stem cell

1 Introduction

Stem cell-based chemical informatics becomes increasingly important in biomedical fields, especially in drug discovery research. Here we describe our advanced bioinformatics technology to predict toxicities of chemicals including methylmercury (causes "Minamata" disease) and thalidomide (causes baby deformities) using the ES cell-based chemical validation system. In our analysis, we find that the Support Vector Machine (SVM)[1] prediction accuracy is increased by inferred gene networks when the weights of network edges are used as SVM features.

2 Method and Results

With RT-PCR gene expression data from ES cells exposed to 15 chemicals that are categorized into "neurotoxic", "tumor-genesis", and "others", we compared two SVM prediction methods: 1) gene expression data only 2) gene expression data + gene network edge weights. For RT-PCR analysis, we chose 9 (+1 internal standard) genes known as important to neuron development and performed RT-PCR experiments with 5 doses and 4 time points, yielding a total of 9x5x4=180 dimensions of data as SVM features. For network analysis, we start with the existing Bayesian network (BN) inference algorithm called TAO-Gen[2,3] and have improved it to the parallel (replica-exchange) version called RX-TAOgen, in order to obtain stable networks. A total of 9x9=81 gene-to-gene weights (β in Fig.1) on network edges are obtained and used as additional features for SVM. The whole scheme is shown in Fig. 1.

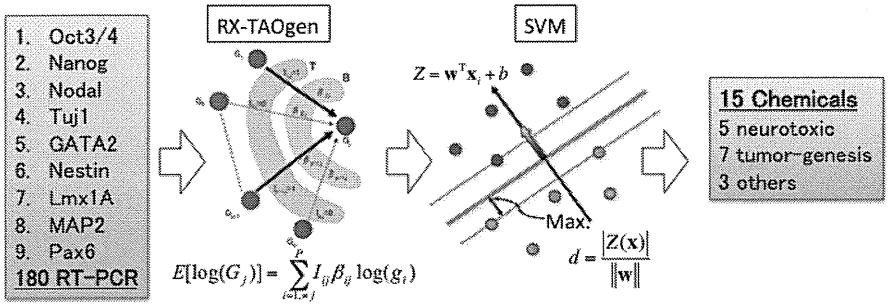

Fig.1 The whole scheme of toxic prediction using Bayesian networks as a part of input vectors of SVM.

2.1 Prediction Accuracy

For each of 15 chemicals, two RT-PCR experiments are performed, thus a total of 30 data are obtained. To evaluate the prediction accuracy, we performed a leave-two-out-prediction (LTOP) test, where two repeat data for each of 15 chemicals are used as test data at one cycle of prediction. In the SVM analysis, we tested three well-known kernels (linear, polynomial, and RBF) and our maximum entropy kernel[4] that is distance-based and the final kernel matrix is obtained by maximizing the kernel entropy. The accuracy is calculated by the ratio of the number of trues (true positives + true negatives) in all of the test data. The results are shown in Table 1.

Table 1: The number of data and accuracies for toxic chemical predictions.
(SVM: support vector machine, BN: Bayesian network)

Toxic category	Data (total of 30)	SVM	BN+SVM
Neurotoxic	10	90.0%	**93.3%**
Tumor-genesis	14	96.7%	**100.0%**
Others	6	–	–

2.2 Network Signatures

According to the results, the use of BN weights as additional features to SVM increases accuracies in both categories of toxic chemicals. The inferred BN structures are quite stable due to the high-tuned Gibbs-sampling method in RX-TAOgen; nevertheless, the obtained structures vary among chemicals. The network examples for 5 neurotoxic and 7 tumor-genesis chemicals are shown in Fig.2.

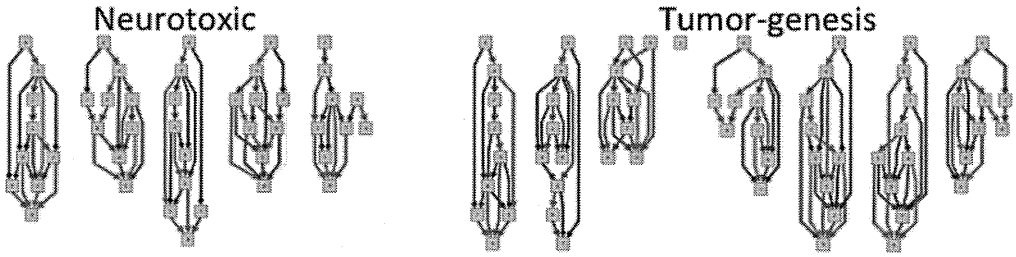

Fig.2 Examples of BN structures inferred from RT-PCR data for two chemical categories. Nine genes are indicated by rectangles and large weights are shown as edges.

3 Discussion

As the BN structures vary in Fig. 2 and it is difficult to recognize any distinct signatures in the same category by human eye, we performed a statistical comparison of the network weights of a pair of inferred networks. Interestingly, we observe that there are more correlations in weights among the same category than the different categories. For example, the average Pearson correlation coefficient of edge weights within 5 neurotoxic chemicals is 0.90, while that of 5 neurotoxic and 7 tumor-genesis chemicals is 0.86. Together, we conclude that the BN structures are weakly conserved in the toxic category, which may give additional information to SVM, resulted in the increase of accuracy.

References

[1] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, 1995.
[2] Yamanaka, T., Toyoshiba, H., Sone, H., Parham, F.M., and Portier C.J., The TAO-Gen algorithm for identifying gene interaction networks with application to SOS repair in E. coli., *Environ. Health. Perspect.*, 112:1614-21, 2004.
[3] Toyoshiba, H., *et al.*, Inference for Bayesian network via Gibbs sampling, Pre-print.
[4] Fujibuchi, W. and Kato, T., Classification of heterogeneous microarray data by maximum entropy kernel., BMC Bioinformatics, 8:267, 2007.

Regular Paper

Learning similarity functions for multi-platform gene expression data

Abstract

The existence of several technologies for measuring gene expression and the growing number of available large-scale gene expression microarrays motivate the need for cross-platform analysis tools. Cross-platform analysis of microarray data is an important problem, which heavily relies on the choice of a similarity function. For a classification task, a good similarity function should improve the prediction performance. It should also be easy to compute, and provide new biological insights of the data. However in practice, choosing a good similarity function for multi-platform microarray data is a difficult problem. In this work, our goal is to improve the performance of microarray search engines such as CellMontage. Therefore, we focus the ranking task rather than the classification task. Our ranking-based approach compares favourably to several similarity functions, including the Pearson and Spearman Correlation coefficients, the Euclidean distance, Linear Discriminant Analysis, and Neighbourhood Component Analysis. Experiments show that our method can be used to differentiate different types of cells with high accuracy, including induced pluripotent stem cells, embryonic stem cells, and cancer cells.