Table 1 Antitumor activities (IC_{50}) of phospha sugar derivatives against leukemia cell lines of K562 and U937 for 48 h at 37°C)

Phospha sugars	K562 IC_{50} (μM)	U937 IC_{50} (μM)
1	>200	>900
2	48	83
3 (DBMPP)	23	24
4 (TBMPP)	3.2	2.3
5	34	ND[a]
6	>100	ND[a]
7	63	ND[a]
8	>100	ND[a]
(Imatinib mesylate)	0.48	>500

[a]ND = not yet determined.

against leukemia cell lines. DBMPP and TBMPP were more active against U937 cell lines than imatinib mesylate (Glivec, Gleevec), however, they did not give any damages against healthy or normal leukocytes (Table 1). Therefore, here we have focused on DBMPP and TBMPP for further investigation of phospha sugar derivatives to develop novel antitumor agents.

DBMPP and TBMPP have wide spectral antitumor activities against various kinds of leukema cell lines (Table 2). The cell cycle analysis revealed that DBMPP induced apoptosis to stop the progress of the cell cycle of K562 and U937 cell lines likewise the manner of imatinib mesylate at the Sub G1 stage (Figure 1: against K562 cell lines). Figure 1 shows that the apoptosis of 78% for the cell cycles was induced by DBMPP (20 μM) against K562 cell lines. TBMPP (20 μM) also induced the apoptosis against various leukemia cell lines about 80% (Figure 2). Branched deoxybromophospha sugars DBMPP and TBMPP have good to excellent antitumor activities against wide spectral leukemia cell lines.

Western Blot Analysis[13]

Western blot analysis for phospha sugars was performed against leukema cell lines. DBMPP enhanced the expression of tumor accelerator factors of FoxM1, KIS, Skp2,

Table 2 Antitumor activities (IC_{50}) of phospha sugar derivatives DBMPP and TBMPP against various kinds of leukemia cell lines (for 48 h at 37°C)

Cell lines	TBMPP (**4**) IC_{50} (μM)	DBMPP (**3**) IC_{50} (μM)
HL60	4.8 ± 1.0	18 ± 1.5
NB4	3.2 ± 0.9	15 ± 1.4
YRK2	5.3 ± 1.3	28 ± 2.6
NOMO-1	5.5 ± 0.8	18 ± 2.1
CEM	6.9 ± 0.3	29 ± 2.4
MOLT4	6.7 ± 1.2	26 ± 1.8
SUP-B15	7.1 ± 1.0	24 ± 2.8
MEG-01	8.6 ± 1.4	27 ± 1.9
SHG3	5.4 ± 0.6	26 ± 2.1
(Healthy or normal leukocyte)[a]	>>200	>>200

[a]Blast 0% (no leukemia cells).

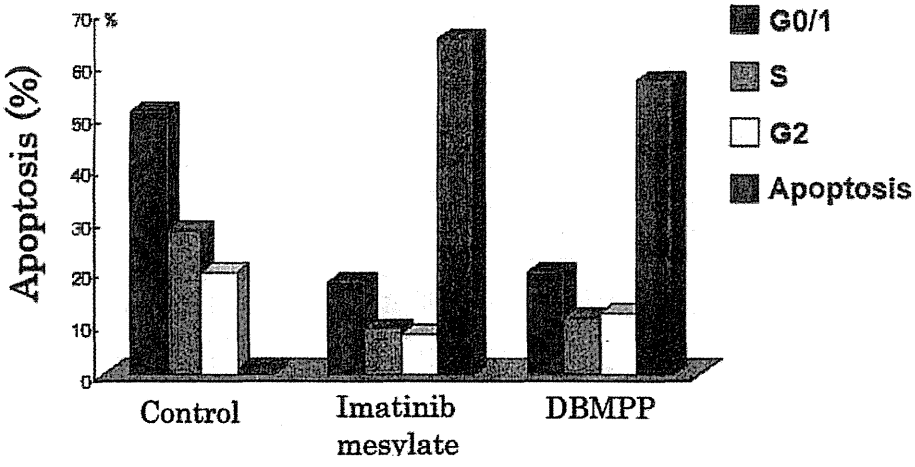

Figure 1 Results of the flow cytometry observed by DBMPP and Imatinib mesylate (Glivec, Gleevec) against K562 cell lines for 24 h at 37°C showing apoptosis (%) induced.

Cyclin D1, Survivin, Aurora-B, Actin against U937 cell lines. On the other hand, DBMPP suppressed the expression of tumor suppressor factors of $p27^{Kip1}$ and $p21^{Cip1}$ against U937 cell lines, and then the phospha sugar induced apoptosis and stops the cell cycle progress (Figure 3). Similarly, DBMPP enhanced the expression of tumor accelerator factors of Aurora-A, Aurora-B, Survivin, FoxM1, Skp2, hKIS, KPC1, and Pirh1 against K562 cell lines, and then affected on the cell cycle progression (Figure 4).[13]

The Western blot analysis for TBMPP against leukemia cell lines showed the enhanced expression of IER5, and then suppressed the expression of Cdc25B (Figure 5).[14] Cdc25B is known to be a common factor of cell cycle progression for many kinds of different tumor cell lines, therefore, phospha sugars TBMPP and DBMPP may be expected

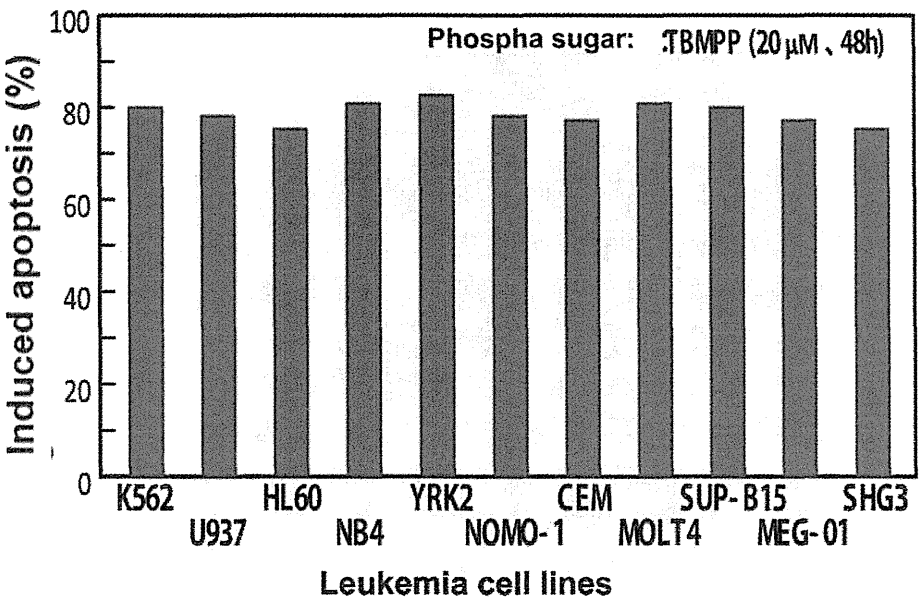

Figure 2 Apoptosis (%) induced by TBMPP (20 μM) for 48 h at 37°C.

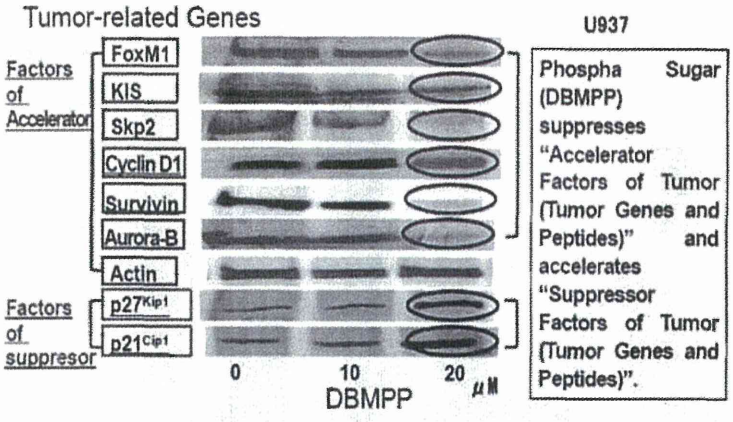

Figure 3 Western blot analysis by DBMPP against U937 cell lines.

to be quite efficient antitumor agents against many kinds of leukemia cell lines and to cure effectively tumor patients suffered from different types of leukemia cell lines.

EXPERIMENTAL

Instruments

TLC (Silica gel: Wako Chromato sheet and/or Merk Kieselgel 60; Eluent : $CHCl_3$: MeOH = 20 : 1, in R*f* value); melting point apparatus (Gallenkamp, in °C) and thermal analysis instrument (Shimazu DTG-60A50AH, TGA and DSC, in °C); HPLC (GL Science: GL-7410 HPLC Pump and GL-7450 UV Detector); MS (MALDI-TOF-MS: GL Science, Voyager-DE Porimerix; Matrix: α-Cyano-4-hydroxycinnamic acid, in *m/z*); IR

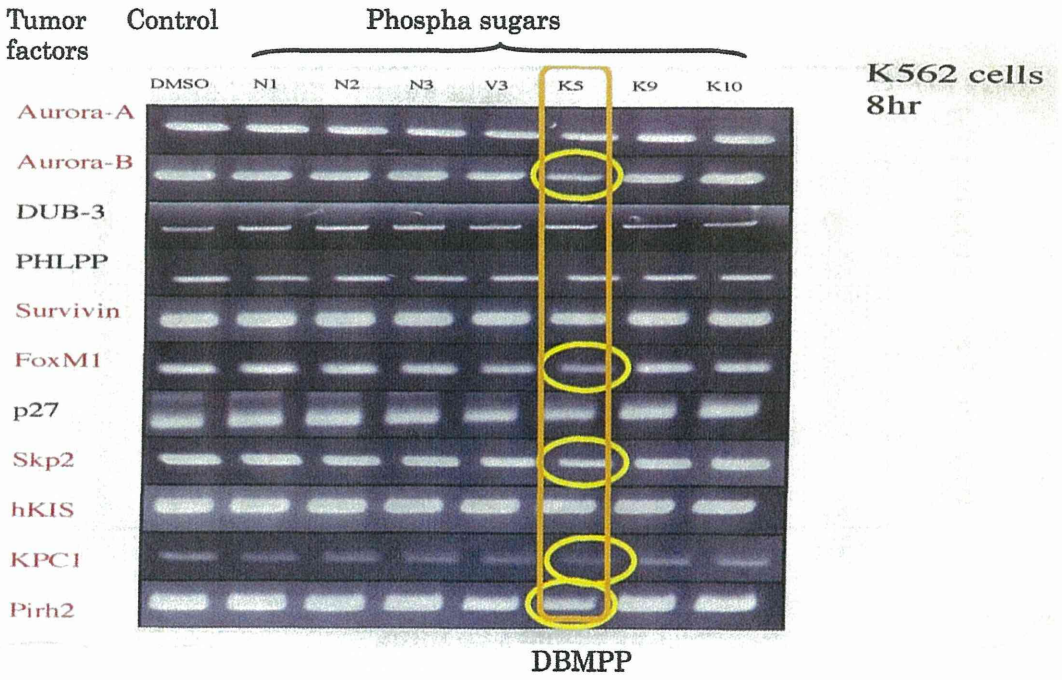

Figure 4 Western blot analysis by DBMPP against K562 cell lines. (Color figure available online).

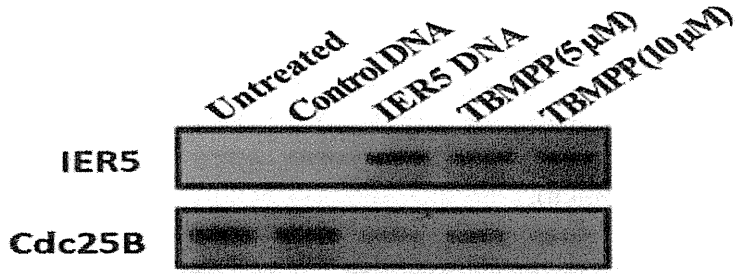

Figure 5 Western blot analysis; Acceleration of IER5 expression and suppression of Cdc25B expression by TBMPP.

(JASCO FT/IR 410 (KBr), in cm^{-1}); ^{1}H-NMR (JEOL JNM-AL300 (300 MHz); Solvent: $CDCl_3$, in δ (ppm) from TMS) were used to record and collect the data for analyzing the products.

Materials

3-Methyl-1-phenyl-2-phospholene 1-oxide (**1**),[15] was prepared and used as the starting material.

Synthesis of 4-bromo-3-methyl-1-phenyl-2-phospholene 1-oxide (2)

3-Methyl-1-phenyl-2-phospholene 1-oxide (**1**: 0.96 g, 5.0 mmol; 1.0 eq.) was dissolved in CCl_4 (5 mL), and to the solution NBS (1.33 g, 1.5 eq.) was added and stirred to make clear solution at 70 °C. To the solution was then added AIBN (0.12 g, 0.75 mmol; 0.15 eq.) and the solution of the reaction mixture was refluxed for 1 day. The reaction mixture was treated with saturated aqueous solution of sodium hydrogensulfite, and then the solution was washed with saturated sodium hydrogencarbonate solution, brine, and water followed by drying over anhydrous Na_2SO_4, which was followed by removal of the solvent afforded crude product. Silica gel column chromatograpy of the product with the mixed eluent of chloroform and methanol (20 : 1) afforded product **2** (0.88 g, 3.3 mmol) in a yield of 65%; Molecular equation: $C_{11}H_{12}BrOP$; M.W.: 271.01; TLC: $R_f = 0.42$ ($CHCl_3$:MeOH = 20 : 1); MS (*m/z*), 271.5(MH^+, 100), 273.5(MH^+, isptope, 90); ^{1}H-NMR ($CDCl_3$, 300 MHz), $\delta = 2.23$ (s, 3H, CH_3), 2.70–2.89 (m, 1H, C5), 2.96–3.15 (m, 1H, C5), 5.14–5.20 (m, 1H, C4), 6.17 (t, $J = 20.1$ Hz, 1H, C2), 7.48–7.65(m, 3H, *o,p*-Ph), 7.80–7.87(m, 2H, *m*-Ph).

Synthesis of 2,3-dibromo-3-methyl-1-phenylphospholane 1-oxide (DBMPP (3))

3-Methyl-1-phenyl-2-phospholene 1-oxide (**1**: 0.21 g, 1.1 mmol; 1.0 eq.) was dissolved in $CHCl_3$ (5 mL), and to the solution $Cu(II)Br_2$ (0.12 g, 0.53 mmol; 0.49 eq.) was added $CHCl_3$ (10 mL) solution of bromine (0.5 mL; 2.7 g, 17 mmol; 16 eq.) dropwisely and stirred for 1 h at room temperature. The reaction mixture was treated with saturated sodium hydrogensulfite solution and was extracted with $CHCl_3$. The $CHCl_3$ extract was washed with saturated sodium hydrogencarbonate, brine, and water, and then dried over anhydrous Na_2SO_4. Evaporation of the solvent from the solution of the product followed by silica gel column chromatography with the mixed eluent of chloroform and methanol (20

: 1) afforded 2,3-dibromo-3-methyl-1-phenylphospholane 1-oxide (**3**; 0.35 g, 0.99 mmol) in 90% yield as a colorless crystal; Molecular equation: $C_{11}H_{13}Br_2OP$; M.W.: 352; TLC: $R_f = 0.23$ ($CHCl_3$:MeOH = 20 : 1); MS (*m/z*), 351.5 (MH^+, 60), 353.5 (MH^+, isptope peak, 100), 355.5 (MH^+, isptope peak, 40); ^{1}H-NMR ($CDCl_3$, 300 MHz), $\delta = 2.17$ (s, 3H, C3-CH_3), 2.44–2.58 (m, 2H, C5), 2.60–2.83 (m, 2H, C4), 4.67 (dd, $J = 1.5$ Hz, $J = 7.2$ Hz, 1H, C2), 7.32–7.56 (m, 5H, Ph).

Synthesis of 2,3,4-tribromo-3-methyl-1-phenylphospholane 1-oxide (TBMPP (4))

Similarly, the reaction of 4-bromo-3-methyl-1-phenyl-2-phospholene 1-oxide (**2**: 0.16 g, 6.0 mmol; 1.0 eq.) with bromine (1.2 mL; 6.6 g, 41 mmol; 6.8 eq.) in CCl_4 under reflux temperature for 24 h and work-up procedure afforded TBMPP (1.7 g, 4.0 mmol) in 69% yield; Molecular equation: $C_{11}H_{12}Br_3OP$; M.W.: 431; TLC: $R_f = 0.62$ ($CHCl_3$: MeOH = 20 : 1); Mp: 142–144 °C; MS (*m/z*): 429.0 (MH^+, 45), 431.1 (MH^+, isptope, 100), 433.1 (MH^+, isptope, 95), 435.1 (MH^+, isptope, 30); ^{1}H-NMR ($CDCl_3$, 300 MHz), $\delta = 2.14$ (s, 3H, C3-CH_3), 2.73–2.90 (m, 1H, C5), 3.04–3.14 (m, 1H C5), 4.26 (d, $J =$ 1.8 Hz,1H, 4C), 4.68–4.86 (m, 1H, C2), 7.52–7.79 (m, 5H, Ph).

Synthesis of 2-bromo-3-hydroxy-3-methyl-1-phenylphospholane 1-oxide (5)

The reaction of 3-methyl-1-phenyl-2-phospholene 1-oxide (**1**: 0.24 g, 1.3 mmol; 1.0 eq.) with bromine (0.5 mL; 2.7 g, 17 mmol; 13 eq.) in the mixed solvent of $CHCl_3$ (2 mL) and water (8 mL (340 eq.) of diluted sodium hydroxide aqueous solution) under stirring for 3 d at room temperature was progressed. The reaction mixture was treated with saturated sodium hydrogensulfite solution and the product was extracted with $CHCl_3$. The extract was washed with saturated sodium hydrogencarbonate solution, brine, and water, and then the solution of the product was dried over anhydrous Na_2SO_4. Evaporation of the solvent followed by silica gel column chromatography with the mixed eluent of chloroform and methanol (20 : 1) afforded product **5** (0.20 g, 0.69 mmol) in 55% yield; Molecular equation $C_{11}H_{14}BrO_2P$; M.W.: 289; TLC: $R_f = 0.45$ ($CHCl_3$: MeOH = 20 : 1); MS (*m/z*): 289.5 (MH^+), 291.5 (MH^+, isptope); ^{1}H-NMR ($CDCl_3$, 300 MHz): $\delta = 1.69$ (s, 3H, C3-CH_3), 2.01–2.51 (m, 2H, C4, C5), 4.32 (m, 1H, C2), 7.48–7.71 (m, 5H, Ph).

Synthesis of 2-bromo-3-(2-hydroxyethoxy)-3-methyl-1-phenylphospholane 1-oxide (6)

The $CHCl_3$ (2 mL) solution of the mixture of 3-methyl-1-phenyl-2-phospholene 1-oxide (**1**: 0.050 g, 0.26 mmol; 1.0 eq.), ethylene glycol (3 mL, 3.3 g, 53 mmol; 200 eq.), and bromine (0.3 mL; 1.6 g, 10 mmol; 39 eq.) was stirred for 2 d at room temperature. The work-up procedure of the reaction mixture by treating with saturated sodium hydrogensulfite and washing with sodium hydrogencarbonate, brine, and water followed by drying over anhydrous Na_2SO_4 and removal of the solvent, and chromatography on silica gel (eluent: $CHCl_3$: MeOH = 20; 1) afforded product **6** (0.077 g, 0.23 mmol) in a yield of 88%; Molecular equation: $C_{13}H_{18}BrO_3P$; M.W.: 333; TLC: $R_f = 0.15$ ($CHCl_3$: MeOH = 20 : 1);

MS (*m/z*): 333.1 (MH^+, 100), 335.1 (MH^+, isptope, 85); ^{1}H-NMR ($CDCl_3$, 300 MHz), δ = 1.57 (s, 3H, C3-CH_3), 2.10–2.31 (m, 2H, C5), 2.40–2.44 (m, 2H, C4), 3.63–3.65 (m, 2H, -CH_2OH), 3.82–3.85 (m, 2H, -OCH_2-), 4.19 (s, 1H, -OH), 7.50–7.82 (m, 5H, Ph)

Synthesis of 2-bromo-3-(2-(2-hydroxyethoxy)-ethoxy)-3-methyl-1-phenylphospholane 1-oxide (7)

Similarly, the reaction of 3-methyl-1-phenyl-2-phospholene 1-oxide (**1**: 0.050 g, 0.26 mmol; 1.0 eq.) with bromine (0.5 mL; 1.6 g, 10 mmol; 38 eq.) in diethylene glycole (3 mL; 3.3 g, 31 mmol; 120 eq.) for 2 d at room temperature under stirring and the work-up procedure followed by silica gel column chromatography (eluent: $CHCl_3$: MeOH = 20 : 1) afforded product **7** (0.055 g, 0.15 mmol) in 58% yield; Molecular equation $C_{15}H_{22}BrO_4P$; M.W.: 377; TLC: R_f = 0.15 ($CHCl_3$: MeOH = 20 : 1); MS (*m/z*): 377.3 (MH^+, 100), 379.3 (MH^+, isptope, 90); ^{1}H-NMR ($CDCl_3$, 300 MHz), δ = 1.53 (s, 3H, C3-CH_3), 2.07–2.13 (m, 2H, C5), 2.34–2.53 (m, 2H, C4), 3.84 (s, 1H, OH), 3.59–3.84 (m, 8H, -CH_2CH_2-), 4.15–4.17 (m, 1H, C2), 7.46–7.77 (m, 5H, Ph)

Synthesis of 2-Imidazolyl-3-methyl-1-phenyl-2-phospholane 1-oxide (8)

To an acetonitril (3 mL) solution of 4-bromo-3-methyl-1-phenyl-2-phospholene 1-oxide (**2**: 0.050 g, 0.18 mmol; 1.0 eq.) was added imidazole (0.062 g, 0.91 mmol; 5.1 eq.) and the mixture was stirred for 24 h at 60°C. The reaction mixture was worked-up by addition of $CHCl_3$ followed by washing with saturated sodium hydrogencarbonate solution, brine, and water, and then dried over anhydrous Na_2SO_4 followed by removal of the solvent to afford crude product **8**, whose chromatography on silica gel with an eluent of mixed solvent ($CHCl_3$: MeOH = 20 : 1) afforded product **8** (0.042 g, 0.16 mmol) in a yield of 89%; Molecular equation: $C_{14}H_{15}N_2OP$; M.W.: 258; TLC: R_f = 0.2 ($CHCl_3$: MeOH = 20 : 1); MS (*m/z*): 259.5 (MH^+, 100); ^{1}H-NMR ($CDCl_3$, 300 MHz), δ = 1.88 (s, 3H, C3-CH_3), 2.36–2.49 (m, 1H, C5), (2.85–2.95 (m, 1H, C5), 5.16(s, 1H, C2), 6.31 (d, J = 21.6 Hz, im-C4-H), 7.13 (d, J = 20.9 Hz, im-C5-H), 7.52–7.72 (m, 5H, Ph), 7.63 (s, 1H, im-C2-H)

Evaluations and Analyses

MTT In Vitro Evaluation for Phospha Sugars Against Leukemic Cells[13]

Antitumor activities of phospha sugars **2–8** prepared were evaluated by MTT method against K562 (human chronic myelogenous leukemia) and U937 (human acute myelogenous leukemia) cell lines as well as HL60, NB4, YRK2, NOMO-1, CEM, MOLT4, SUP-B15, MEG-01, and SHG3 cell lines. Cells were seeded in 96-well flat-bottomed microplates at a density of 5 × 10^4 per well and incubated with various concentrations of phospha sugar derivatives for antitumor activity assay or without any phospha sugars for control experiments, for 48 h at 37°C, and then 10 μL solution of 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) (Sigma) was added to each well at the final concentration of 1.0 μg/mL/well. After incubation for 24 h or 48 h at 37°C, absorbance was measured at a wavelength of 560 nm by using a microplate reader for the *in vitro* evaluation.

Cell Cycle Analyses[13]

Propidium iodide (PI) (Sigma) staining was used to analyze cellular DNA content. Cells treated with Imatinib mesylate (Glivec, Gleevec) (1 μM) or phospha sugars DBMPP or TBMPP were cultured at 37°C in 2 mL of complete medium containing 1×10^6 cells. After incubation for 24 h or 48 h, the cells were washed twice with cold PBS, fixed with 70% ethanol overnight before treatment with 100 μg/mL RNase A, and then stained with 50 μg/mL of PI. For apoptosis analysis, the relative DNA content per cell was measured by flow cytometry using an Epics Elite flow cytometer (Coulter Immunotech). The percentage of cells in the apoptotic sub-G1 phase, as well as G1, S, and G2/M phases, was calculated using the ModFit program (Becton Dickinson).

Western Blot Analyses[13,14]

For Phospha Sugar Treated Leukemia Cells. Leukemia cells treated with phospha sugars (DBMPP or TBMPP) were harvested, and then washed with cold PBS, and resuspended in lysis buffer containing 0.5% Nonidet P-40, 50 mM Tris-HCl (pH 8.0), 0.1 mM EDTA, 150 mM NaCl, 1 mM sodium orthovanadate, and 1 mM dithiothreitol supplemented with one Complete Mini protease inhibitor tablet (Boehringer Mannheim GmBH) per 20 mL lysis buffer immediately before use. The proteins were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and transferred to polyvinylidene difluoride membranes (Millipore). The membranes were then blocked with 0.5% milk in PBS for 1 h at room temperature. After being washed with Tris-buffered saline Tween (TBS-T), the membranes were incubated for 1 h at room temperature with horseradish peroxidase-conjugated goat anti-mouse IgG or anti-rabbit IgG (Amersham Biosciences Inc.) for 1 h and exposed to X-ray film at room temperature. The signal was detected by chemiluminescence using an ECL detection kit (Amersham Bioscience Inc.). The following commercially available antibodies and dilutions were used for Western blot analysis: rabbit polyclonal anti-FoxM1 antibody (MPP2 K-19, 1:500) (Santa Cruz Biotechnology, Inc.), rabbit polyclonal anti-$p27^{Kip1}$ antibody (1:1000) (Santa Cruz Biotechnology, Inc.), mouse monoclonal anti-$p21^{Cip1}$ antibody (1:1000) (Santa Cruz Biotechnology, Inc.), rabbit polyclonal anti-Cdc25B2 antibody (1:500) (Santa Cruz Biotechnology, Inc.), mouse monoclonal anti-Cyclin D1 antibody (1:500) (Santa Cruz Biotechnology, Inc.), mouse monoclonal anti-Cyclin A antibody (1:500) (Santa Cruz Biotechnology, Inc.), rabbit polyclonal anti-KIS antibody (1:500) (ABGENT, Inc.), rabbit polyclonal anti-Aurora-B antibody (1:500) (ABGENT, Inc.), mouse monoclonal anti-bcl-2 antibody (BD Biosciences Pharmingen), mouse monoclonal anti-caspase-9 antibody (BD Biosciences Pharmingen), mouse monoclonal anticaspase-3 (CPP32) antibody (BD Biosciences Pharmingen), and mouse monoclonal anti-PARP antibody (BD Biosciences Pharmingen). To ensure equal protein volume loading, similar experiments were performed by using a mouse monoclonal antiactin antibody (C-4; ICN Biomedicals, Inc., Aurora, OH) as an internal control.

For Transfected Cells. Cells transfected with scrambled shRNA or with IER5 shRNA-#1 or -#2 were harvested. After 3 d, Western blot analysis was performed using the following antibodies: goat polyclonal anti-IER5 (Abcam), rabbit polyclonal anti-Cdc25B (Santa Cruz), anti-CHK1 (Santa Cruz), anti-WEE1 (Santa Cruz), anti-Aurora-B (Santa Cruz), mouse monoclonal anti-Cyclin B1 (Santa Cruz), and anti-Survivin (Santa Cruz). To ensure equal protein volume loading, similar experiments were performed by using a mouse monoclonal antiactin antibody (C-4; ICN Biomedicals, Inc., Aurora, OH) as an internal control.

REFERENCES

1. (a) M. Shan; G. A. O'Doherty. *Org. Lett.*, **2008,** 10, 3381-3384. (b) M. Sollogoub; P. Sinay. "*Organic Chemisry of Sugars*," D. E. Levy and P. Fugedi (Ed.), (CRC Press, Taylor & Francis, New York, NY, USA, **2006**), Chap. 8, pp. 349-381.
2. M. A. Alam; A. Kumar; Y. D. Vankar. *Eur. J. Org. Chem.*, **2008,** 29, 4972-4980.
3. B. Joseph; P. Rollin. *Phosphorus Sulfur Silicon Relat. Elem.*, **1993,** 74, 467-468.
4. M. Yamashita. "*Top. Heterocyclic Chem., Bioactive Heterocycles II*" Ed. by S. Eguchi, (Springer, Berlin, Germany, **2007**), **8**, pp. 173-222.
5. R. L. Whistler; C. C. Wang. *J. Org. Chem.*, **1968,** 33, 4455-4458.
6. (a) S. Inouye; T. Tsuruoka; T. Ito; T. Niida. *Tetrahedron*, **1968,** 24, 2125-2144. (b) M. Shan; G. A. O'Doherty. *Org. Lett.* **2008,** 10, 3381-3384. (c) M. A. Alam; A. Kumar; Y. D. Vankar. *Eur. J. Org. Chem.* **2008,** 29, 4972-4980. (d) B. G. Davis; M. A, Maughan; T. M. Chapman; R. Villard; S. Courtney. *Org. Lett.* **2002,** 4, 103-106. (e) M. Chmielewski; R. L. Whistler. *J. Org. Chem.* **1975,** 40, 639-643. (f) P. Wang; L. A. Agrofoglio; M. G. Newton; C. K. Chu. *J. Org. Chem.* **1999,** 64, 4173-4178.
7. (a) G. Legler; E. Julich. *Carbohydr. Res.* **1984,** 128, 61-72. (b) J. Braanalt; I. Kvarnstroem; G. Niklasson; S. C. T. Svensson; B. Classon; B. Samuelsson. *J. Org. Chem.* 59, 1783-1788 (**1994**).
8. (a) E. M. Dangerfield; C. H. Plunkett; B. L. Stocker; M. S. M. Timmer. *Molecules*, **2009,** 14, 5298-5307. (b) S. Martinez-Montero; S. Fernandez; Y. S. Sanghvi; J. Chattopadhyaya; M. Ganesan; N. G. Ramesh; V. Gotor; M. Ferrero. *J. Org. Chem.*, **2012,** 77, 4671-4678.
9. (a) B. Joseph; P. Rollin. *J. Carbohydr. Chem.*, **1993,** 12, 719-29. (b) V. Blot; J.-F. Brier; M. Davoust; S. Miniere; V. Reboul; P. Metzner. *Phosphorus, Sulfur Silicon Relat. Elem.*, **2005,** 180, 1171-1182.
10. (a) T. Hanaya; K. Sugiyama; Y. Fujii; A Akamatsu; H. Yamamoto. *Heterocycles*, **1994,** 55, 1301-1309. (b) T. Hanaya; S. Kawase; H. Yamamoto. *Heterocycles*, **2005,** 66, 251-261. (c) M. Yamashita; Y. Nakatsukasa; M. Yoshikane; H. Yoshida; T. Ogata; S. Inokawa. *Carbohydr. Res.* **1977,** 59, 12-14. (d) H. Yamamoto; T. Hanaya; H. Kawamoto; S. Inokawa; M. Yamashita; M. A. Armour; T. T. Nakashima. *J. Org. Chem.* **1985,** 50, 3516-3521.
11. (a) M. Yamashita; Y. Kato; K. Suzuki; T. Oshikawa. *Heterocycl. Commun.* **1998,** 4, 411-414. (b) V. Krishna Reddy; B. Haritha; T. Oshikawa; M. Yamashita. *Tetrahedron Lett.* **2004,** 45, 2851-2854. (c) M. Yamada; K. Asai; J. Yamashita; T. Suyama; T. Niimi; K. Maddali; M. Fujie; S. Nakamura; M. Kimura; Y. Tanaka; M. Toda; M. Yamashita. *Heterocycl. Commun.* **2010,** 16, 173-180.
12. (a) M. Yamaoka; M. Yamashita; M. Yamada; M. Fujie; K. Kiyofuji; N. Ozaki; K. Asai; T. Niimi; T. Suyama; J. Yamashita; A. Sawada; R. Makita; M. Sugiyama; M. Toda; S. Nakamura; K. Ohnishi. *Pure Appl. Chem.*, **2012,** 84, 37-48. (b) K. Tsunekawa; M. Yamashita; M. Fujie; T. Niimi; T. Suyama; K. Asai, S. Ito; J. Yamashita; M. Yamada; N. Ozaki; S. Nakamura. *Phosphorus, Sulfur Silicon Relat. Elem.*, **2011,** 186, 936-944. (c) J. Yamashita; T. Suyama; K. Asai; M. Yamada; T. Niimi; M. Fujie; S. Nakamura; K. Ohnishi; M. Yamashita. *Heterocyclic Commun.*, **2010,** 16, 89-97. (d) M. Yamada; M. Yamashita; T. Suyama; J. Yamashita; K. Asai; T. Niimi; N. Ozaki; M. Fujie; K. Maddali; S. Nakamura; K. Ohnishi. *Bioorg. Med. Chem. Lett.*, **2010,** 20, 5943-5946.
13. S. Nakamura; M. Yamashita; D. Yokota; I. Hirano; T. Ono; M. Fujie; K. Shibata; T. Niimi; T. Suyama; K. Maddali; K. Asai; J. Yamashita; Y. Iguchi; K. Ohnishi. *Invest. New Drugs*, **2010,** 28, 381-391.
14. S. Nakamura; Y. Nagata; L. Tan, T. Takemura; K. Shibata; M. Fujie; S. Fujisawa; Y. Tanaka; M. Toda; R. Makita; K. Tsunekawa; M. Yamada; M. Yamaoka; J. Yamashita; K. Ohnishi;, M. Yamashita. *PLoS One*, **2011,** 6, e28011.
15. M. Yamashita; R. P. Mallikarjuna; Y. Kato; V. Krishna Reddy; K. Suzuki; T. Oshikawa, *Carbohydr. Res.*, **2001,** 336, 257-270.

ELSEVIER

Cancer Genetics and Cytogenetics 197 (2010) 8–15

CANCER GENETICS AND CYTOGENETICS

Analysis of poly(ADP-ribose) polymerase-1 (*PARP1*) gene alteration in human germ cell tumor cell lines

Hideki Ogino[a,b], Robert Nakayama[c], Hiromi Sakamoto[c], Teruhiko Yoshida[c], Takashi Sugimura[a], Mitsuko Masutani[a,b,*]

[a]*Biochemistry Division, National Cancer Center Research Institute, 1-1 Tsukiji 5-chome, Chuo-ku, Tokyo 104-0045, Japan*
[b]*ADP-ribosylation in Oncology Project, National Cancer Center Research Institute, 1-1 Tsukiji 5-chome, Chuo-ku, Tokyo 104-0045, Japan*
[c]*Genetics Division, National Cancer Center Research Institute, 1-1 Tsukiji 5-chome, Chuo-ku, Tokyo 104-0045, Japan*

Received 28 June 2009; received in revised form 17 October 2009; accepted 17 October 2009

Abstract The poly(ADP-ribose) polymerase-1 protein (PARP-1) functions in DNA repair, maintenance of genomic stability, induction of cell death, and transcriptional regulation. We previously analyzed alterations of the *PARP1* gene in 16 specimens of human germ cell tumors, and found a heterozygous sequence alteration that causes the amino acid substitution Met129Thr (M129T) in both tumor and normal tissues in a single patient. In this study, aberration of the *PARP1* gene and protein was further analyzed in human germ cell tumor cell lines. We found a nonheterozygous sequence alteration that causes the amino acid substitution Glu251Lys (E251K) located at a conserved peptide stretch of PARP-1 in cell line NEC8. Sequencing of 95 samples from Japanese healthy volunteers revealed that all the samples were homozygous for the wild-type alleles at M129T and E251K. The M129T allele is thus suggested to be a rare single-nucleotide polymorphism (SNP). We observed a decrease in auto-poly(ADP-ribosyl)ation activity of PARP-1 proteins harboring M129T or E251K amino acid substitution, but the difference was not statistically significant. The levels of PARP-1 and poly(ADP-ribosyl)ation were heterogeneous among germ cell tumor cell lines. The SNPs of the *PARP1* gene, as well as differences in the levels of PARP-1 and poly(ADP-ribosyl)ation of proteins, may influence germ cell tumor development and responses to chemotherapy and radiotherapy.

1. Introduction

Poly(ADP-ribose) polymerase-1 (PARP-1) is activated by DNA damage and catalyzes poly(ADP-ribosyl)ation of various proteins, including PARP-1 itself, using nicotinamide adenine dinucleotide (NAD^+) as a substrate. PARP-1 is involved in DNA repair, maintenance of genomic stability, and cell death induction. We and others have previously reported that *Parp-1* knockout (*Parp-1*$^{-/-}$) mice showed higher susceptibility to carcinogenesis induced by alkylating agents in the colon [1] and lung [2], compared with wild-type (*Parp-1*$^{+/+}$) mice. The incidence of spontaneous tumors developed at an advanced age in the liver was also higher in *Parp-1*$^{-/-}$ mice [3,4].

Involvement of PARP-1 in the development of human cancer has not yet been fully clarified. In human cancers, increased expression of the *PARP1* gene has been reported in Ewing's sarcoma [5,6], in malignant lymphoma [7], and in the familial adenomatous polyposis (FAP) tumors [8]. Decreased expression of the *PARP1* gene has been observed in several gastric and colon cancer cell lines [6], grade II and III endometrial carcinomas [9], and in some breast cancers [10].

The A/A homozygotes of the V762A single-nucleotide polymorphism (SNP) in the *PARP1* gene have been reported to be associated with decreased activity of PARP-1. The A/A homozygotes are shown to be associated with an increased risk for prostate cancer in European-origin subjects [11] and in lung cancer and esophageal cancers in Chinese heavy smokers [12,13]. In the case of lung and esophageal cancers, a twofold increase in risk with the A/A homozygotes was observed in Chinese smokers [12,13]. The combination of the 762A allele of the *PARP1* gene and the 399 G allele of the *XRCC1* gene was associated with increased risk of lung, esophageal, and gastric cardia cancers [12–15].

* Corresponding author. Tel.: +81-3-3542-2511; fax: +81-3-2542-2530.
E-mail address: mmasutan@ncc.go.jp (M. Masutani).

0165-4608/10/$ – see front matter
doi:10.1016/j.cancergencyto.2009.10.012

PARP-1 also participates in the transcriptional regulation of some genes [16–18] and in cellular differentiation [18–20]. *Parp-1*$^{-/-}$ mouse embryonic stem cells show preferential induction of the trophoblast lineage [20], including trophoblast giant cells (TGCs), during teratocarcinoma formation in vivo or during cell culture in vitro [19]. The biochemical properties of TGCs resemble those of syncytiotrophoblastic giant cells (STGCs) of human germ cell tumors [21,22]. It is thus suggested that *PARP1* deficiency may possibly trigger differentiation to STGCs during germ cell tumor formation. The appearance of STGCs in trophoblastic or choriocarcinomatous human germ cell tumors has been reported to be associated with poor prognosis [21]. Teratocarcinoma cells undergo differentiation into epithelial cells in vitro, at least in part, in the presence of the PARP inhibitor 3-aminobenzamide [23].

The aberrations of the *PARP1* gene in 16 human germ cell tumors were previously analyzed, and a heterozygous sequence alteration (ATG to ACG) that causes amino acid substitution, Met129Thr (M129T) [24] was found in one patient in both cancer and normal tissues. In the present study, we further analyzed aberration of the *PARP1* gene and poly(ADP-ribosyl)ation level in human germ cell tumor cell lines and evaluated effects of any amino acid alterations found on PARP-1 function.

2. Materials and methods

2.1. Cell culture

Cell lines JEG-3, NCCIT, PA-1, Tera-1, and Tera-2 were purchased from the American Type Culture Collection (ATCC, Manassas, VA). Cell lines ITOII and NEC8 were purchased from the Japanese Collection of Research Bioresources (http://cellbank.nibio.go.jp/cellbank_e.html). Cell lines NEC14 and NEC15 were purchased from the Riken Bioresource Center (http://www.brc.riken.go.jp/inf/en/index.shtml). Each cell line was cultured under the conditions recommended by the providers. The growing cells at mid-late log phase were used as the materials for preparation of genomic DNA, total RNA, and total protein.

For transfection experiments, an immortalized *Parp-1*$^{-/-}$ mouse embryonic fibroblast (MEF) clone, PH13b, established from spontaneously immortalized *Parp-1*$^{-/-}$ MEFs [25] was used. The cells were cultured at 37°C under 5% CO_2 and 95% humidity.

2.2. Direct sequencing of the human PARP1 gene

Polymerase chain reaction (PCR)-based direct sequencing of all 23 exons of the *PARP1* gene in germ cell tumor cell lines was performed as previously described [24]. Oligonucleotide primer sets for the 23 exons were designed from intron sequences of each exon as previously described [24]. Amplified PCR products were subjected to sequence analysis (ABI PRISM 310 genetic analyzer, Applied Biosystems, Carlsbad, CA; model CEQ8000, Beckman Coulter, Fullerton, CA). Sequence comparison was performed against the sequence of the human *PARP1* gene (NCBI accession numbers NT_004559 and NT_167186) and its cDNA (NCBI accession numbers M18112, M32721, M17081, J03473, BC037545, and BC014206).

2.3. Pyrosequencing

Pyrosequencing for codon 129 and codon 251 of the *PARP1* gene was performed as previously described [26]. Briefly, genomic DNA samples were extracted from blood of 95 healthy Japanese volunteers, and samples were subjected to genotyping by pyrosequencing using the PSQ96 System (Pyrosequencing, Uppsala, Sweden). For sequencing of codons 129 and 251, we used 5′-CGTGCAAGGGGTGTA-3′ and 5′-TGAACACACTTTCTTTAGC-3′ as sequencing primers, respectively. In the case of codon 251, the sequencing result of one sample was not informative.

All subjects provided informed consent, and the study was approved by the Ethical Committee of the National Cancer Center of Japan.

2.4. Construction of PARP1 mutants

PARP1 mutant cDNA harboring either M129T, E251K, or K940R amino acid substitution was prepared using primers harboring respective mutation and Phusion polymerase (Fynnzymes, Espoo, Finland) according to the protocol of the QuikChange site-directed mutagenesis kit (Stratagene, La Jolla, CA). The human *PARP1* cDNA [27] was modified using PCR to harbor restriction enzyme recognition sites of *Sal*I, *Sma*I, *Age*I, and *Xba*I at the 5′ and 3′ terminus, respectively, and modified Kozack's sequences derived from pEGFP-C1 (Clontech Laboratories, Mountain View, CA) in the 5′-UTR (untranslated region) of the *PARP1* cDNA. The *PARP1* cDNA harboring either M129T, E251K, and K940R amino acid substitution was inserted into pcDNA3.1(+)hygro (Invitrogen, Carlsbad, CA) for measurement of PARP-1 enzymatic activity. pEGFP-C1 was used for construction of the GFP protein fused to the N-terminus of PARP-1 protein for analysis of subcellular localization.

2.5. Transfection of the PARP1 mutant constructs

Four micrograms of each construct was transfected into the *Parp-1*$^{-/-}$ MEFs in six-well plates using Lipofectamine 2000 (Invitrogen). At 24 hours after transfection, whole-cell extracts were prepared by suspending cells in a lysis solution containing 50 mmol/L Tris–HCl (pH 6.8), 10% glycerol, 2% sodium dodecyl sulfate (SDS), complete protease inhibitor cocktail tablets (Roche Applied Science, Mannheim, Germany), and 10% β-mercaptoethanol. The samples were frozen in liquid nitrogen and stored at −80°C until use.

2.6. Activity gel analysis

PARP-1 enzymatic activity was measured as auto-poly(ADP-ribosyl)ation activity of PARP-1, as described elsewhere [28]. Briefly, crude extracts were separated by 6% SDS–polyacrylamide gel electrophoresis (SDS-PAGE) containing 100 μg/mL of sonicated salmon sperm DNA as an activated DNA. After renaturation of proteins in the gel, the gel was incubated in a reaction mixture containing 50 mmol/L Tris–HCl (pH 8.0), 1 mmol/L dithiothreitol, 50 μmol/L [^{32}P]adenylate-labeled nicotinamide adenine dinucleotide (1 μCi/mL, NEN–PerkinElmer, Waltham, MA), and 25 mmol/L $MgCl_2$ at 37°C for 1 hour. The gel was fixed with 10% methanol–30% acetic acid (v/v) solution, and washed with 5% trichloroacetic acid–0.2% sodium pyrophosphate. The radioactivities of the dried gels were analyzed using a BAS-2500 bio-imaging analyzer (Fujifilm, Tokyo, Japan).

2.7. Western blot analysis

Whole-cell extracts were prepared by suspending cells in a lysis solution containing 50 mmol/L Tris-HCl (pH 6.8), 10% glycerol, 2% SDS, complete protease inhibitor cocktail tablets (Roche Applied Science), and 10% β-mercaptoethanol, followed by sonication. Equivalent protein amounts of lysate (8 μg) were separated by 4–20% gradient SDS-PAGE. After transfer of proteins to Immobilon-P polyvinylidene difluoride membranes (Millipore, Billerica, MA), the membrane was incubated with anti-poly(ADP-ribose) monoclonal antibody 10H (Alexis Biochemicals–Enzo Life Sciences, Lausanne, Switzerland) [29], anti-PARP-1 monoclonal antibody C2-10 (Oncogene Research Products, Merck Chemicals, Darmstadt, Germany), F1-23 (Alexis Biochemicals) or anti-α-tubulin monoclonal antibody DM1A (MP Biomedicals, Irvine, CA). Immune complexes were visualized using a horseradish peroxidase-linked secondary antibody and an enhanced chemiluminescence reaction ECL kit (Amersham Biosciences, Piscataway, NJ). The PARP-1 protein level was quantified using a LAS-3000 bio-imaging analyzer (Fujifilm).

3. Results

3.1. Sequence alterations in germ cell tumor cell lines

The human germ cell tumor cell lines used in this study are listed in Table 1. All 23 exons and their flanking regions of the *PARP1* gene in nine germ cell tumor cell lines were sequenced. Sequence alterations and SNPs found in the *PARP1* gene are listed in Table 2. We found that the two missense SNPs, Phe54Leu (F54L) and V762A, were described in the NCBI (National Center for Biotechnology Information) database of SNPs. These SNPs of F54L and V762A in NEC14 and NEC15 were observed as nonheterozygous sequence alterations. A nonheterozygous sequence alteration (GAG to AAG) that causes amino acid substitution Glu251Lys (E251K) in NEC8 was also found (Table 2 and Fig. 1). This sequence alteration has not been listed in the NCBI database of SNPs. Further information for NEC8 was not available, and it is not known whether this is a somatic or germ line mutation or sequence alteration caused during establishment of this cell line.

Table 1
Human germ cell tumor cell lines used in this study

Cell line	Type	Reference
NEC8	Embryonal carcinoma: testis	Motoyama et al., 1987 [39]
NEC14	Embryonal carcinoma, choriocarcinoma: testis	Motoyama et al., 1987 [39]
NEC15	Embryonal carcinoma, yolk sac tumor: testis	Motoyama et al., 1987 [39]
ITOII	Embryonal carcinoma: testis	Motoyama et al., 1987 [39]
Tera-1	Embryonal carcinoma: lung	Fogh et al., 1978 [40]
Tera-2	Embryonal carcinoma: lung	Fogh et al., 1978 [40]
NCCIT	Embryonal carcinoma	Teshima et al., 1988 [41]
PA-1	Teratoma: ovary	Zeuthen et al., 1980 [42]
JEG-3	Choriocarcinoma: placenta	Kohler et al., 1971 [43]

To determine whether E251K is a common SNP in the Japanese population, we sequenced 94 samples from Japanese healthy volunteers using a pyrosequencing method. Similarly, sequencing data were obtained from 95 volunteers for M129T, which we previously found in a germinoma and its normal tissue from a patient [24]. In both cases, all the sequenced samples were homozygous: GAG (251E) and ATG (129M). None were heterozygous, which suggests that the two amino acid substitutions, E251K and M129T, are not common SNPs in the Japanese population.

Three synonymous SNPs were also found at Asp81, Ala284, and Lys352 (Table 2). In the noncoding region, a SNP of G to C in 5′-UTR, 17 bases upstream of the translation initiation site, downstream of a putative ETS-1-binding site (base −26 to −22) [30]), was found (rs907187). NEC15 had a nonheterozygous C allele, and ITOII and Tera-1 had heterozygous G/C alleles. We also noted that the nonheterozygous allele at a SNP in intron 2 (rs1805405) was observed at higher frequency (3/9) than a heterozygous allele (1/9) in germ cell tumor cell lines, but the difference was not statistically significant. NEC8, NEC15, and ITOII had the nonheterozygous A allele, whereas NCCIT had heterozygous C/A alleles.

3.2. Effect of amino acid alteration on the activity of PARP-1

The effect of amino acid substitution of methionine to threonine at codon 129 (Met129Thr) and glutamic acid to lysine at codon 251 (Glu251Lys) on PARP-1 enzymatic activity was examined. We transiently expressed the mutated PARP-1 harboring either M129T or E251K substitution in the *Parp-1*$^{-/-}$ MEFs, and measured the enzymatic activity of PARP-1 by the activity gel method. Both *Parp-1* mutants showed a decrease in PARP-1 activity relative to the wild type, although the difference was not

Table 2
Sequence alterations and SNPs found in the *PARP1* gene in human germ cell tumor cell lines

	Exon	Nucleotide[a]	No.	Germ cell tumor cell line[b]	SNP ID[c]	Heterozygosity[c]
				Amino acid substitution		
Phe54Leu	2	TTC	8	NEC8, NEC14, ITOII, Tera-1, Tera-2, NCCIT, PA-1, JEG-3	rs3738708	0.023
		TTC/TTG	0			
		TTG	1	NEC15		
Glu251Lys	6	GAG	8	NEC14, NEC15, ITOII, Tera-1, Tera-2, NCCIT, PA-1, JEG-3	not listed	not listed
		GAG/AAG	0			
		AAG	1	NEC8		
Val762Ala	17	GTG	7	NEC8, ITOII, Tera-1, Tera-2, NCCIT, PA-1, JEG-3	rs1136410	0.351
		GTG/GCG	0			
		GCG	2	NEC14, NEC15		
				SNPs without amino acid substitution		
Asp81Asp	2	GAC	5	NEC8, Tera-2, NCCIT, PA-1, JEG-3	rs1805404	0.372
		GAC/GAT	2	ITOII, Tera-1		
		GAT	2	NEC14, NEC15		
Ala284Ala	7	GCT	5	NEC15, ITOII, Tera-2, PA-1, JEG-3	rs1805414	0.498
		GCT/GCC	2	NCCIT, Tera-1		
		GCC	2	NEC8, NEC14		
Lys352Lys	8	AAA	8	NEC14, NEC15, ITOII, Tera-1, Tera-2, NCCIT, PA-1, JEG-3	rs1805415	0.379
		AAA/AAG	0			
		AAG	1	NEC8		
				SNPs in noncoding region		
5′-UTR (−17 bp)		G	6	NEC8, NEC14, Tera-2, NCCIT, PA-1, JEG-3	rs907187	0.357
		G/C	2	ITOII, Tera-1		
		C	1	NEC15		
Intron 2 (5592 bp)		C	5	NEC14, Tera-1, Tera-2, PA-1, JEG-3	rs1805405	0.362
		C/A	1	NCCIT		
		A	3	NEC8, NEC15, ITOII		

Abbreviations: SNP, single nucleotide polymorphism; UTR, untranslated region.
[a] The altered nucleotide is indicated by underscoring.
[b] See Table 1 for cell lines.
[c] SNP identifiers and average estimated heterozygosity data are from the NCBI database of SNPs, available at http://www.ncbi.nlm.nih.gov/SNP/index.html.

statistically significant ($P = 0.1266$ for 129T, $P = 0.2752$ for 251K) (Fig. 2A).

Cellular localization of E251K and M129T mutants was also analyzed as GFP-fusion protein expressed in the *Parp-1*$^{-/-}$ MEF. The cellular localization of another *PARP1* mutant harboring K940R amino acid substitution was analyzed as well [24]. Wild-type, 129T, 251K, and 940R protein localized exclusively in the nuclei (Fig. 2B). The localization is observed in a punctuated manner in the nuclei, and the localization pattern did not differ between wild-type and the mutants.

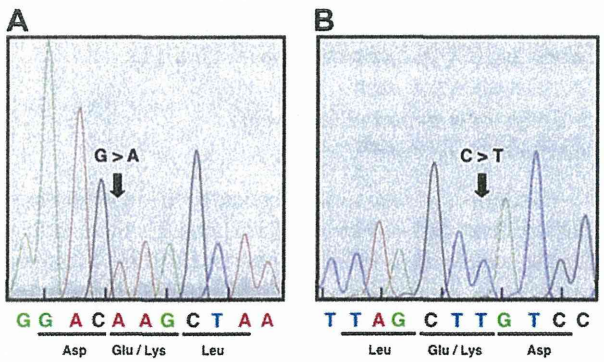

Fig. 1. Electropherograms of sequences surrounding codon 251 of the *PARP1* gene in NEC8 cells. Sequence alterations of both strands were confirmed by sequencing using the sense (A) and anti-sense (B) primers.

3.3. Levels of PARP-1 protein, activity, and poly(ADP-ribosyl)ation

The levels of PARP-1 proteins were measured by Western blot analysis (Fig. 3). PA-1 and NEC8 cells both showed lower levels of PARP-1 proteins. We also examined the enzymatic activity of PARP-1 as PARP-1 auto-poly(ADP-ribosyl)ation activity using whole-cell extracts (Fig. 3, activity gel analysis). PA-1 and NEC8 cells both showed lower levels of auto-poly(ADP-ribosyl)ation activity, compared with the other cell lines analyzed. PA-1, NEC8, and also Tera-2 exhibited lower levels of overall poly(ADP-ribosyl)ation of proteins, compared with the other cell lines.

4. Discussion

Among the germ cell tumor cell lines analyzed in this study, five cell lines were established in Japan, and these

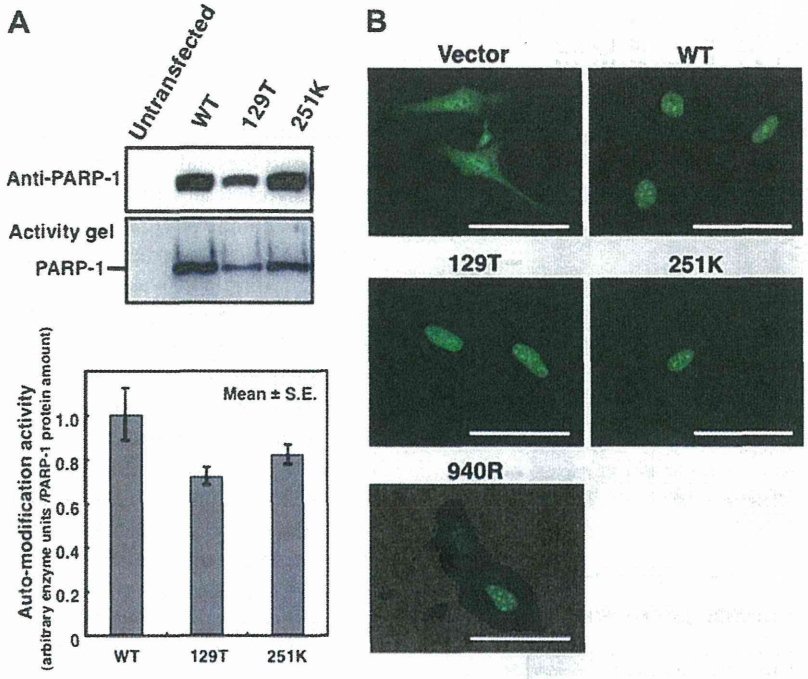

Fig. 2. Effects of amino acid alteration in PARP-1 on enzymatic activity and subcellular localization. (A) Semiquantitative analysis of auto-poly(ADP-ribosyl)ation activity of wild-type and mutant PARP-1 proteins in the whole cell extracts from the *Parp-1*$^{-/-}$ mouse embryonic fibroblasts transfected with either the wild-type (WT) or mutant PARP-1 expression plasmid. The representative result of the activity gel analysis (middle panel) and Western blot analysis of PARP-1 (top panel) is shown. The auto-poly(ADP-ribosyl)ation activity normalized to the expressed PARP-1 level measured by Western blot analysis. A linear relationship between the amount of PARP-1 and auto-poly(ADP-ribosyl)ation activity was confirmed (data not shown). (B) Subcellular localization of GFP-fused PARP-1. At 24 hours after transfection, localization of wild-type PARP-1 and mutant proteins harboring either 129T, 251K, or 940R amino acid substitution was observed exclusively in the nuclei. Scale bars: 10 μm.

may reflect a spectrum of polymorphisms in the Japanese population. We found a nonheterozygous sequence alteration (GAG to AAG) that causes amino acid substitution of E251K of PARP-1 in NEC8. Because the corresponding normal tissue samples were not available for the NEC8 cell line, we could not examine whether E251K is a SNP. It is not known whether the nonheterozygosity represents a homozygosity or an allelic loss, nor could we exclude the possibility that this sequence alteration was introduced during or after establishment of NEC8. We previously reported that a heterozygous sequence alteration that causes amino acid alteration of M129T was observed in a human germ cell tumor specimen [24]. Our analysis in the present study suggests that the sequence alterations of E251K and M129T are not common SNPs in the Japanese population.

E251K is located at a peptide stretch conserved among species in the C-terminus of a DNA binding domain close to the nuclear localization signal. The third zinc-binding motif (codon 295–321) [28,31] is present close to codon 251. The third zinc-binding domain of codon 216–366 is required for dimerization of PARP-1. K249E substitution is reported to decrease PARP-1 enzymatic activity [32]. A decrease in auto-poly(ADP-ribosyl)ation activity but no alteration in nuclear localization of PARP-1 harboring M129T or E251K amino acid substitution was observed; however, we noted that NEC8, which has a nonheterozyous E251K allele, had a lower level of PARP-1 protein in the extract (Fig. 3). The effects on DNA binding or DNA repair regulation, as well as stability, should be further analyzed.

Within nine cell lines, both nonheterozygous minor alleles of V762A and intron 2 (5,592 bp) (SNP ID, rs1805405) showed a tendency of higher frequencies than expected, although it was not statistically significant. These tendencies are similar to the result obtained with our previous study using 16 germ cell tumor specimens [24]. The V762A SNP was found to be associated with the risk of prostate cancers in European-origin subjects, in whom the A/A genotype showed a twofold increase in susceptibility [11]. Recently the *PARP1* V762A polymorphism has been reported to reduce the enzymatic activity of PARP-1 and the ability of interaction with *XRCC1* [15,33]; however, a decrease in PARP-1 auto-poly(ADP-ribosyl)ation activity and overall poly(ADP-ribosyl)ation levels was not observed in NEC14, which harbors a nonheterozygous V762A allele (Table 2 and Fig. 3).

The nonheterozygous allele of SNP rs1805405, located within the polypyrimidine tract close to the 3′ splice acceptor site in the intron 2, was observed at a higher frequency (3/9) than expected, as in the case with human germ cell tumor specimens (3/16) [24]. A stretch of $(C/T)_6$NCAGG(C/T) at a splicing acceptor site is relatively conserved in the introns [34]. The polypyrimidine tract is

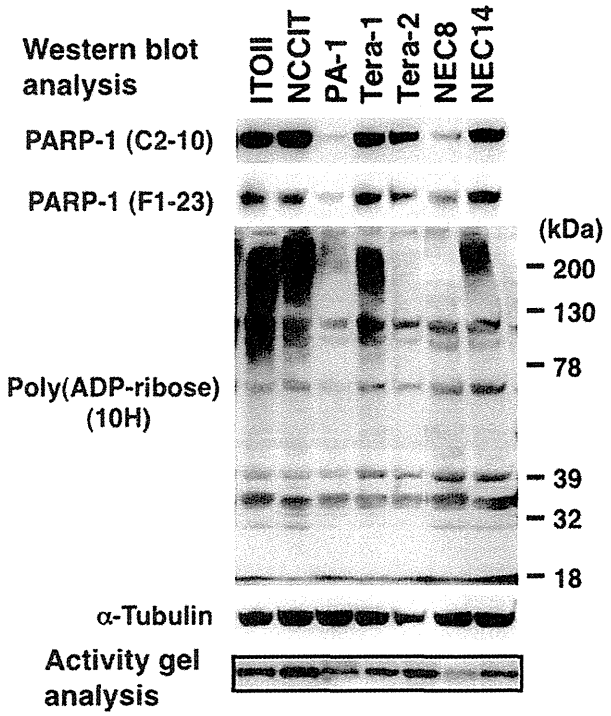

Fig. 3. PARP-1 protein levels and activity, with overall poly(ADP-ribosyl)ation protein levels. Western blot analysis was performed for detection of PARP-1 protein and poly(ADP-ribosyl)ated proteins (upper and middle panels). For detection of PARP-1 enzymatic activity, activity gel analysis was performed (bottom panel) to analyze PARP-1 auto-poly(ADP-ribosyl)ation activity; 30 μg of cell extracts was subjected to 6% SDS-PAGE in the presence of activated DNA in the gel.

highly conserved, and the frequency of adenine at this position is low (~7%) [35]. In the case of all introns of *PARP1* reference sequences (NCBI numbers NT_004559 and NT_167186), the frequency of adenine at this position is ~9%. In the case of hereditary nonpolyposis colorectal cancer proband, a single base-pair T-to-A transversion at position −11 of the *MLH1* gene intron 1 splice acceptor site caused exon 2 skipping [36]. We detected only a full-length transcript of *PARP1* in the cell lines harboring minor alleles of the SNP at intron 2 by Northern blot analysis (data not shown). Whether the SNP at intron 2 of *PARP1* affects the splicing efficiency of exon 3 needs further investigation.

It is noteworthy that the poly(ADP-ribosyl)ation level is lower in NEC8, PA-1, and Tera-2, compared with other germ cell tumor cell lines. The lower poly(ADP-ribosyl)ation level in NEC8 and PA-1 could be explained by the lower PARP-1 level, but that is not the case for Tera-2. Activites of other PARP family proteins or of PARG, a major poly(ADP-ribose) degradation enzyme, may also affect the poly(ADP-ribosyl)ation level in Tera-2. We did not find any SNPs or other base alterations in the 3′UTR of the *PARP1* gene in nine cell lines. Therefore, the mechanism for lower levels of PARP-1 in NEC8 and PA-1 may not be due to altered posttranscriptional regulation of PARP-1.

In this study, we identified sequence alterations, including SNPs, in the *PARP1* gene of human germ cell tumor cell lines. The nonheterozygous minor alleles of SNPs at V762A and intron 2 showed a slightly higher frequency. Differences in the levels of PARP-1 and poly(ADP-ribosyl)ation were observed. Because poly(ADP-ribosyl)ation reaction is involved in several physiological processes in cancer cells, including DNA repair and differentiation, the alteration of PARP-1 activity may affect the development of cancers through multiple processes. It is also suggested that PARP family proteins, including PARP-2, may complement PARP-1 functions [37]. It may therefore be necessary to examine aberrations of PARP-1 and other PARP family proteins in cancers not only at gene expression levels but also at protein or enzymatic activity levels. Clinical trials are ongoing with PARP inhibitors in combination with chemotherapeutic agents [38]. Understanding the mechanisms of functional regulation of PARP-1 in cancer cells is important. Because the activity of PARP-1 and other PARP family members is important in DNA repair and cell death induction, the levels of PARP-1 and poly(ADP-ribosyl)ation activity may also substantially affect the outcome of cancer therapies that target DNA.

Acknowledgments

We thank T. Otsubo, T. Okada, and S. Mimaki for technical help and suggestions. We are grateful for the kind support and suggestions of H. Nakagama. This work was supported in part by a Grant-in-Aid for the Second and Third Term Comprehensive 10-year Strategy for Cancer Control and a Grant-in-Aid for Cancer Research from the Ministry of Health, Labor, and Welfare of Japan.

References

[1] Nozaki T, Fujihara H, Watanabe M, Tsutsumi M, Nakamoto K, Kusuoka O, Kamada N, Suzuki H, Nakagama H, Sugimura T, Masutani M. Parp-1 deficiency implicated in colon and liver tumorigenesis induced by azoxymethane. Cancer Sci 2003;94:497–500.

[2] Tsutsumi M, Masutani M, Nozaki T, Kusuoka O, Tsujiuchi T, Nakagama H, Suzuki H, Konishi Y, Sugimura T. Increased susceptibility of poly(ADP-ribose) polymerase-1 knockout mice to nitrosamine carcinogenicity. Carcinogenesis 2001;22:1–3.

[3] Tong WM, Cortes U, Hande MP, Ohgaki H, Cavalli LR, Lansdorp PM, Haddad BR, Wang ZQ. Synergistic role of Ku80 and poly(ADP-ribose) polymerase in suppressing chromosomal aberrations and liver cancer formation. Cancer Res 2002;62:6990–6.

[4] Masutani M, Nakagama H, Sugimura T. Poly(ADP-ribosyl)ation in relation to cancer and autoimmune disease. Cell Mol Life Sci 2005;62:769–83.

[5] Prasad SC, Thraves PJ, Bhatia KG, Smulson ME, Dritschilo A. Enhanced poly(adenosine diphosphate ribose) polymerase activity and gene expression in Ewing's sarcoma cells. Cancer Res 1990; 50:38–43.

[6] Masutani M, Nozaki T, Sasaki H, Yamada T, Kohno T, Shimizu K, Gotoh M, Shiraishi M, Yokota J, Hirohashi S, Nakagama H, Sugimura T. Poly(ADP-ribose) polymerase-1 gene in human tumor

cell lines: its expression and structural alteration. Proc Jpn Acad Ser B Phys Biol Sci 2004;80B:114–8.

[7] Menegazzi M, Scarpa A, Carcereri de Prati A, Menestrina F, Suzuki H. Correlation of poly(ADP-ribose)polymerase and p53 expression levels in high-grade lymphomas. Mol Carcinog 1999; 25:256–61.

[8] Idogawa M, Yamada T, Honda K, Sato S, Imai K, Hirohashi S. Poly(ADP-ribose) polymerase-1 is a component of the oncogenic T-cell factor-4/β-catenin complex. Gastroenterology 2005;128: 1919–36.

[9] Ghabreau L, Roux JP, Frappart PO, Mathevet P, Patricot LM, Mokni M, Korbi S, Wang ZQ, Tong WM, Frappart L. Poly(ADP-ribose) polymerase-1, a novel partner of progesterone receptors in endometrial cancer and its precursors. Int J Cancer 2004;109:317–21.

[10] Bièche I, de Murcia G, Lidereau R. Poly(ADP-ribose) polymerase gene expression status and genomic instability in human breast cancer. Clin Cancer Res 1996;2:1163–7.

[11] Lockett KL, Hall MC, Xu J, Zheng SL, Berwick M, Chuang SC, Clark PE, Cramer SD, Lohman K, Hu JJ. The *ADPRT* V762A genetic variant contributes to prostate cancer susceptibility and deficient enzyme function. Cancer Res 2004;64:6344–8.

[12] Zhang X, Miao X, Liang G, Hao B, Wang Y, Tan W, Li Y, Guo Y, He F, Wei Q, Lin D. Polymorphisms in DNA base excision repair genes *ADPRT* and *XRCC1* and risk of lung cancer. Cancer Res 2005;65:722–6.

[13] Hao B, Wang H, Zhou K, Li Y, Chen X, Zhou G, Zhu Y, Miao X, Tan W, Wei Q, Lin D, He F. Identification of genetic variants in base excision repair pathway and their associations with risk of esophageal squamous cell carcinoma. Cancer Res 2004;64:4378–84.

[14] Zhang Z, Miao XP, Tan W, Guo YL, Zhang XM, Lin DX. Correlation of genetic polymorphisms in DNA repair genes *ADPRT* and *XRCC1* to risk of gastric cancer [In Chinese]. Ai Zheng 2006;25:7–10.

[15] Miao X, Zhang X, Zhang L, Guo Y, Hao B, Tan W, He F, Lin D. Adenosine diphosphate ribosyl transferase and x-ray repair cross-complementing 1 polymorphisms in gastric cardia cancer. Gastroenterology 2006;131:420–7.

[16] Ogino H, Nozaki T, Gunji A, Maeda M, Suzuki H, Ohta T, Murakami Y, Nakagama H, Sugimura T, Masutani M. Loss of *Parp-1* affects gene expression profile in a genome-wide manner in ES cells and liver cells. BMC Genomics 2007;8:41.

[17] Krishnakumar R, Gamble MJ, Frizzell KM, Berrocal JG, Kininis M, Kraus WL. Reciprocal binding of PARP-1 and histone H1 at promoters specifies transcriptional outcomes. Science 2008;319: 819–21.

[18] Quenet D, Gasser V, Fouillen L, Cammas F, Sanglier-Cianferani S, Losson R, Dantzer F. The histone subcode: poly(ADP-ribose) polymerase-1 (Parp-1) and Parp-2 control cell differentiation by regulating the transcriptional intermediary factor TIF1β and the heterochromatin protein HP1α. FASEB J 2008;22:3853–65.

[19] Nozaki T, Masutani M, Watanabe M, Ochiya T, Hasegawa F, Nakagama H, Suzuki H, Sugimura T. Syncytiotrophoblastic giant cells in teratocarcinoma-like tumors derived from *Parp*-disrupted mouse embryonic stem cells. Proc Natl Acad Sci USA 1999;96: 13345–50.

[20] Hemberger M, Nozaki T, Winterhager E, Yamamoto H, Nakagama H, Kamada N, Suzuki H, Ohta T, Ohki M, Masutani M, Cross JC. Parp1-deficiency induces differentiation of ES cells into trophoblast derivatives. Dev Biol 2003;257:371–81.

[21] von Hochstetter AR, Sigg C, Saremaslani P, Hedinger C. The significance of giant cells in human testicular seminomas: a clinico-pathological study. Virchows Arch A Pathol Anat Histopathol 1985;407:309–22.

[22] Masutani M, Nozaki T, Watanabe M, Ochiya T, Hasegawa F, Nakagama H, Suzuki H, Sugimura T. Involvement of poly(ADP-ribose) polymerase in trophoblastic cell differentiation during tumorigenesis. Mutat Res 2001;477:111–7.

[23] Ohashi Y, Ueda K, Hayaishi O, Ikai K, Niwa O. Induction of murine teratocarcinoma cell differentiation by suppression of poly(ADP-ribose) synthesis. Proc Natl Acad Sci USA 1984;81: 7132–6.

[24] Shiokawa M, Masutani M, Fujihara H, Ueki K, Nishikawa R, Sugimura T, Kubo H, Nakagama H. Genetic alteration of poly(ADP-ribose) polymerase-1 in human germ cell tumors. Jpn J Clin Oncol 2005;35:97–102.

[25] Nozaki T, Fujihara H, Kamada N, Ueda O, Takato T, Nakagama H, Sugimura T, Suzuki H, Masutani M. Hyperploidy of embryonic fibroblasts derived from *Parp-1* knockout mouse. Proc Jpn Acad Ser B Phys Biol Sci 2001;77:121–4.

[26] Nakayama R, Sato Y, Masutani M, Ogino H, Nakatani F, Chuman H, Beppu Y, Morioka H, Yabe H, Hirose H, Sugimura H, Sakamoto H, Ohta T, Toyama Y, Yoshida T, Kawai A. Association of a missense single nucleotide polymorphism, Cys1367Arg of the *WRN* gene, with the risk of bone and soft tissue sarcomas in Japan. Cancer Sci 2008; 99:333–9.

[27] Uchida K, Morita T, Sato T, Ogura T, Yamashita R, Noguchi S, Suzuki H, Nyunoya H, Miwa M, Sugimura T. Nucleotide sequence of a full-length cDNA for human fibroblast poly(ADP-ribose) polymerase. Biochem Biophys Res Commun 1987;148:617–22.

[28] Masutani M, Nozaki T, Hitomi Y, Ikejima M, Nagasaki K, de Prati AC, Kurata S, Natori S, Sugimura T, Esumi H. Cloning and functional expression of poly(ADP-ribose) polymerase cDNA from *Sarcophaga peregrina*. Eur J Biochem 1994;220:607–14.

[29] Kawamitsu H, Hoshino H, Okada H, Miwa M, Momoi H, Sugimura T. Monoclonal antibodies to poly(adenosine diphosphate ribose) recognize different structures. Biochemistry 1984;23: 3771–7.

[30] Soldatenkov VA, Albor A, Patel BK, Dreszer R, Dritschilo A, Notario V. Regulation of the human poly(ADP-ribose) polymerase promoter by the ETS transcription factor. Oncogene 1999;18: 3954–62.

[31] Langelier MF, Servent KM, Rogers EE, Pascal JM. A third zinc-binding domain of human poly(ADP-ribose) polymerase-1 coordinates DNA-dependent enzyme activation [Erratum in: J Biol Chem 2008;283:22884]. J Biol Chem 2008;283:4105–14.

[32] Trucco C, Flatter E, Fribourg S, de Murcia G, Ménissier-de Murcia J. Mutations in the amino-terminal domain of the human poly(ADP-ribose) polymerase that affect its catalytic activity but not its DNA binding capacity. FEBS Lett 1996;399:313–6.

[33] Wang XG, Wang ZQ, Tong WM, Shen Y. *PARP1* Val762Ala polymorphism reduces enzymatic activity. Biochem Biophys Res Commun 2007;354:122–6.

[34] Nakata K, Kanehisa M, DeLisi C. Prediction of splice junctions in mRNA sequences. Nucleic Acids Res 1985;13:5327–40.

[35] Stephens RM, Schneider TD. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J Mol Biol 1992;228:1124–36.

[36] Clarke LA, Veiga I, Isidro G, Jordan P, Ramos JS, Castedo S, Boavida MG. Pathological exon skipping in an HNPCC proband with *MLH1* splice acceptor site mutation. Genes Chromosomes Cancer 2000;29:367–70.

[37] Miwa M, Masutani M. PolyADP-ribosylation and cancer. Cancer Sci 2007;98:1528–35.

[38] Ratnam K, Low JA. Current development of clinical inhibitors of poly(ADP-ribose) polymerase in oncology. Clin Cancer Res 2007; 13:1383–8.

[39] Motoyama T, Watanabe H, Yamamoto T, Sekiguchi M. Human testicular germ cell tumors in vitro and in athymic nude mice. Acta Pathol Jpn 1987;37:431–48.

[40] Fogh J. Cultivation, characterization, and identification of human tumor cells with emphasis on kidney, testis, and bladder tumors. Natl Cancer Inst Monogr 1978;49:5–9.

[41] Teshima S, Shimosato Y, Hirohashi S, Tome Y, Hayashi I, Kanazawa H, Kakizoe T. Four new human germ cell tumor cell lines. Lab Invest 1988;59:328–36.

[42] Zeuthen J, Nørgaard JO, Avner P, Fellous M, Wartiovaara J, Vaheri A, Rosén A, Giovanella BC. Characterization of a human ovarian teratocarcinoma-derived cell line. Int J Cancer 1980;25: 19–32.

[43] Kohler PO, Bridson WE. Isolation of hormone-producing clonal lines of human choriocarcinoma. J Clin Endocrinol Metab 1971;32: 683–7.

Journal of Human Genetics (2010) 55, 436–440

www.nature.com/jhg

ORIGINAL ARTICLE

Detection of inappropriate samples in association studies by an IBS-based method considering linkage disequilibrium between genetic markers

Masataka Andoh[1], Yasunori Sato[2], Hiromi Sakamoto[2], Teruhiko Yoshida[2] and Megu Ohtaki[3]

An association study is a popular study design to identify susceptibility genes for common complex diseases. In such a study, the presence of inappropriate samples, such as those derived from close relatives or showing DNA contamination, causes an inflation of type I error or a decrease of power. Here we propose an identity-by-state (IBS)-based detection method of inappropriate samples taking linkage disequilibrium (LD) into consideration. The test statistics is the mean of the proportion of alleles that are shared identical by state at each single nucleotide polymorphism (SNP) between each sample pair in an association study. A covariance of the number of shared alleles between two SNPs is introduced to consider LD. We show that type I error and power are estimated accurately in computer-simulated data, and that if the number of SNPs analyzed is small, the performance of detection of inappropriate samples is superior to the previous method in simulated LD. An application to real association study data showed that accuracy in estimating the distribution of test statistics improved if LD was considered. Sample pairs considered to be siblings were detected. These results suggested that an LD-considered IBS-based detection method is useful in identifying inappropriate samples in an association study.
Journal of Human Genetics (2010) **55,** 436–440; doi:10.1038/jhg.2010.43; published online 7 May 2010

Keywords: association study; IBS; linkage disequilibrium; normal distribution; quality control; SNP

INTRODUCTION

An association study is a popular study design to identify susceptibility genes for common complex diseases.[1] Under the common disease-common variant (CD-CV) hypothesis,[2] the power of an association study is generally higher than a linkage study for identification of disease susceptibility genes. Most association studies search for genetic markers that are related to a disease by comparing the frequency between the case (disease) and control (non-disease) populations. A disease-susceptibility gene may then be identified in the region of linkage disequilibrium (LD) corresponding to an associated genetic marker. Recently, biallelic single nucleotide polymorphisms (SNPs) are widely used as genetic markers.

A number of biases can be introduced in case–control association studies, making it very important to deal with them appropriately because they cause a significant inflation of type I error or a decrease of power. Quality control (QC), a series of operations to detect and remove biases, includes such possible causes as population stratification, sample contamination and cryptic relatedness.[1,3] Sample contamination can occur when samples of different individual origin are mixed by error in the experimental process of, for example, DNA extraction or SNP typing. Cryptic relatedness is observed when some close relatives are enrolled in a study by chance without the knowledge of investigators, which can cause an inflation of type I error.[3]

For general detection of related samples, a likelihood ratio test based on posterior probability of genotype under certain relationships was proposed.[4] In the case of a family-based study, an identity-by-state (IBS)-based method[5,6] for a detection of errors in a sib-pair relationship was proposed, with the method using the summation of the IBS for a pair of sibs. Conversely, an identity-by-descent (IBD)-based method (PLINK[7]) was proposed. PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) estimates genome-wide IBD-sharing coefficients between unrelated samples from genome-wide data. This metrics is useful for QC by diagnosing pedigree errors, undetected relationships, and sample swap, duplication and contamination events. It calculates $\hat{\pi}$ (the proportion of alleles shared IBD) for each sample pairs, and contamination events are considered as outliers of $\hat{\pi}$. In these previous studies, however, SNPs were assumed to be mutually independent, and LD was not taken into consideration. However, in many association studies, the LD between marker SNPs cannot be neglected.

Here we propose an IBS-based detection method to detect inappropriate samples (for example, contamination, close relatives) in an

[1]Graduate School of Biomedical Sciences, Hiroshima University, Hiroshima, Japan; [2]Genetics Division, National Cancer Center Research Institute, Tokyo, Japan and [3]Department of Environmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University, Hiroshima, Japan
Correspondence: M Andoh, Graduate School of Biomedical Sciences, Hiroshima University, 1-2-3 Kasumi, Minami-Ku, Hiroshima 734-8551, Japan.
E-mail: andohmstk@hiroshima-u.ac.jp
Received 19 October 2009; revised 1 April 2010; accepted 2 April 2010; published online 7 May 2010

association study, which relies on SNP markers with or without LD. We evaluated a type I error and the power of the proposed method and estimated the number of SNPs required to detect inappropriate samples for marker SNPs in either LD or linkage equilibrium (LE). The proposed method was compared with the previous method by simulation. Finally, an application of the proposed method to an example of real data in a genome-wide association study suggested the practical relevance of our discussion.

MATERIALS AND METHODS

Test statistics

For K SNPs, the proposed test statistics is calculated as a mean proportion of alleles that are shared IBS over all SNPs. Let i and j denote the samples in an association study, and let $T_k^{(i,j)} \in \{0, 0.5, 1\}$ be the proportion of shared alleles at SNP k for sample (i, j) as follows:

$$T_k^{(i,j)} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ share 2 alleles at SNP } k \\ 0.5 & \text{if } i \text{ and } j \text{ share 1 alleles at SNP } k \\ 0 & \text{if } i \text{ and } j \text{ share 0 alleles at SNP } k \end{cases}$$

Then we define the test statistics as follows:

$$Y_{i,j} = \frac{1}{K}\sum_{k=1}^{K} T_k^{(i,j)}.$$

It is assumed that almost all samples are not inappropriate (that is, no contamination or relatedness), so that the distribution of $Y_{i,j}$ under the null hypothesis $\mathbf{H}_0$ ((i,j) are independent) can be approximated by the normal distribution with mean $\mu_{i,j}$ and variance $\sigma^2_{i,j}$. No fixed effect for $\mu_{i,j}$, $\sigma^2_{i,j}$ was observed from real data, so that $\mu_{i,j}$, $\sigma^2_{i,j}$ are assumed to be independent of (i,j). Then $\mu_{i,j} \approx \mu$, $\sigma_{i,j}{}^2 \approx \sigma^2$ are given by

$$\mu = E(Y) = \frac{1}{K}\sum_{k=1}^{K} E(T_k|R=1),$$

$$\sigma^2 = V(Y) = \frac{1}{K^2}\left\{\sum_{k=1}^{K} E(T_k^2|R=1) - E(T_k|R=1)^2 + 2\sum_{k_1=1}^{K}\sum_{k_2=k_1+1}^{k_1+w} \mathrm{Cov}\,(T_{k_1}, T_{k_2}|R=1)\right\}.$$

We then take LD into account as covariance $\mathrm{Cov}\,(T_{k_1}, T_{k_2})$ between SNP k_1 and SNP k_2, and denote the range of LD by w, $\{w|0<w<K\}$. In this study, we assume unrelated individuals, parent–child, siblings and contamination as the four types of sample pair relationships, denoted as R=1, 2, 3 and 4, respectively.

Parameter estimation and detection method

To estimate the $E(Y)$ and $V(Y)$, we note that $E(T_k|R=1)$, $E(T_k^2|R=1)$ and $\mathrm{Cov}\,(T_{k_1}, T_{k_2}|R=1)$ are expressed as

$$E(T_k|R=1) = p_k^4 + 2p_k^3 q_k + 4p_k^2 q_k^2 + 2p_k q_k^3 + q_k^4,$$

$$E(T_k^2|R=1) = p_k^4 + p_k^3 q_k + 4p_k^2 q_k^2 + p_k q_k^3 + q_k^4,$$

$$\mathrm{Cov}\,(T_{k_1}, T_{k_2}|R=1) = \sum_{T_{k_1}}\sum_{T_{k_2}} T_{k_1} T_{k_2} p(T_{k_1}, T_{k_2}|R=1) - E(T_{k_1}|R=1)E(T_{k_2}|R=1).$$

respectively, where p_k and q_k are the allele frequencies for SNP k ($p_k+q_k=1$), and the joint probability $p(T_{k_1}, T_{k_2}|R=1)$ is calculated with the estimated haplotype frequency[8] for SNP k_1 and SNP k_2 (see Mathematical details in Supplementary Information).

T_k of close relatives is higher than that of unrelated individuals, $E(T_k|R=3) > E(T_k|R=2) > E(T_k|R=1)$, and the genotypes of the contaminated samples should appear the same for each, so that inappropriate samples are detected as outliers (Y) of the distribution Y under the null hypothesis $\mathbf{H}_0$. In this method, the sample pairs whose Y is more than the threshold $s=\{E(Y|R=1)+E(Y|R=2)\}/2$ are considered as outlier pairs.

Simulation study

We conducted two types of simulation. First, we evaluated the type I error and the power of the proposed method. Second, we compared the proposed method with the previous method in terms of the performance of type I error and power.

Simulation 1. We set the following conditions in the case of LE:

Condition 1: the number of samples (sum of case and control samples) is N=200, 600, 1000 and the number of SNPs is K=100, 200, 400, 800 and 1000.

Condition 2: under a null hypothesis $\mathbf{H}_0$ the allele frequency of each SNP has independent uniform distribution; the genotype frequency follows the Hardy-Weinberg equilibrium.

Condition 3: under an alternative hypothesis $\mathbf{H}_1$ there are three inappropriate sample pairs (parent–child, sibling and contaminated sample pairs, $(Y|R\neq 1)$) in the simulation data.

Condition 4: type I error is estimated as $u=\#\{Y|Y>s,R=1\}/(N(N-1)/2)$, which is the proportion of independent sample pairs (R=1) that are misjudged as inappropriate sample pairs. Power is estimated as $v=\#\{Y|Y>s,R=r\}$,r=2,3,4, which is the number of inappropriate sample pairs that are detected correctly. In this simulation there is one inappropriate sample pair for each type (parent–child, sibling and contaminated sample pairs), so v=0 (not detected) or v=1 (detected). The threshold is $s=\{E(Y|R=1)+E(Y|R=2)\}/2$.

Condition 5: the genotype of parent–child and sibling samples is assigned by using the conditional probability under each relationship.[4] Genotypes of contaminated samples are assigned by the following process: if the genotypes of two samples at an SNP are the same, the genotypes of the two samples are not changed. If the genotypes are different, the genotypes of the two samples are assigned stochastically either as heterozygous or unchanged (the probability of the heterozygous and unchanged status is the same, 0.5 for each) for both samples.

The procedure to conduct the simulation experiment was as follows:

Step 1. Set K and N according to the simulation conditions.
Step 2. Generate SNP genotype data $(Y|R=1)$ under null hypothesis $\mathbf{H}_0$. Under alternative hypothesis $\mathbf{H}_1$ three normal sample pairs are replaced by inappropriate pairs $(Y|R\neq 1)$ in the original data of null hypothesis.
Step 3. Calculate the threshold s for each data ($\mathbf{H}_0$ and $\mathbf{H}_1$), and estimate the u and v, respectively.
Step 4. Repeat first three steps 1000 times and calculate the mean of u and v.

Simulation 2. We set the following conditions in the case of LD:

Condition 1: the number of samples (sum of case and control samples) is N=200, 600, 1000 and the number of SNPs is K=100, 200, 1000, 3000 and 5000.

Condition 2: under null hypothesis $\mathbf{H}_0$ the allele frequency of each SNP and the LD coefficient between SNPs is calculated from reference 100-SNP data selected from the HapMap (http://www.hapmap.org/) JPT data. The reference SNP data is composed of 10 regions on the genome; each region has 10 SNPs that are in the neighborhood of each other. In this simulation, these regions are called strong LD regions (w=10; Supplementary Figure 1).

Conditions 3 and 4: same as those previously described for the LE case.

Condition 5: the genotype of parent–child and siblings samples is assigned by the following process: first, any two specimens are sampled from association study data. Next, in each strong LD region the haplotype data of parent–child or sibling sample are generated by combining the four haplotypes of two samples according to Mendel's law. Genotypes of contaminated samples are assigned by the same process as in the case of LE.

The procedure to conduct the simulation experiment was as follows:

Steps 1, 3 and 4 are the same as those for LE.

Step 2: generate haplotype data based on the allele frequency and the coefficient of reference data by using bivariate normal distribution in each strong LD region.[9] The association study SNP data are generated by the combination of any two haplotypes $(Y|R=1)$ under a null hypothesis $\mathbf{H}_0$. Under an alternative hypothesis $\mathbf{H}_1$ three normal sample pairs are replaced by inappropriate pairs $(Y|R\neq 1)$ in the original data of the null hypothesis.

We set the following parameters for PLINK (version 1.05) to compare the performance of the proposed method and IBD estimation of PLINK, '--genome --genome-full --min 0 --max 1'.

Real data

The proposed method was applied to our real association study data on 1498 samples and 2665 SNPs, which are the second screening data for the JSNP genome scan for gastric cancer.[10] After a routine QC in the typing laboratory removed samples that have many missing values and a high proportion of heterozygotes, the allele frequency showed a nearly uniform distribution and there is a weak LD overall (Figure 1).

RESULTS

Simulation Study

We evaluated the type I error and power (R=2, 3, 4) in the simulation data for SNP markers showing LE or LD (Tables 1 and 2). Type I error and power were calculated accurately by assuming the distribution of Y to be a normal distribution with mean $E(Y)$ and variance $V(Y)$ in both cases. In the case of LE, more than 800 SNPs were required to detect parent–child samples correctly ($\hat{v}=1$) and to avoid exclusion of normal samples from the case–control data ($\hat{u}N(N-1)/2<1$) (Table 1). Conversely, more than 3000 SNPs are required in the case of LD (Table 2). Because the correlation between the SNPs increases the variance $V(Y)$, the type I error is inflated, the power decreases, and the necessary number of SNPs increases. Siblings or contaminated samples were also detected by the smaller number of SNPs (Tables 1 and 2). In comparing u_0(w=0) and u_{10}(w=10) in LD, type I error is calculated more accurately by taking LD into account (Table 2).

To compare the performance of the IBS-based method with the IBD-based method, we applied both methods to the simulation data with LD (K=200, N=200). In the IBD detection method, inappropriate samples were detected by the probability $P(Z)$ of IBD

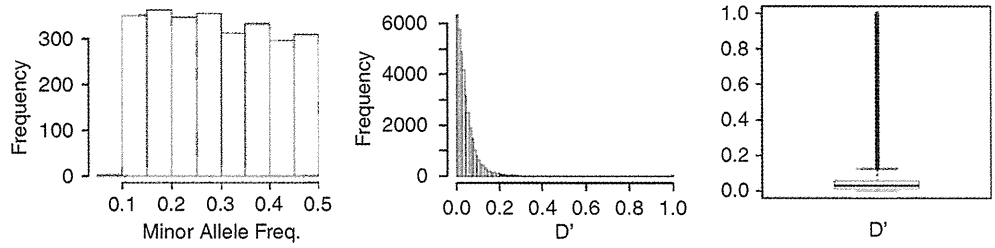

Figure 1 Allele frequency and the distribution of the linkage disequilibrium coefficient of real association study data.

Table 1 Estimated type I error and power in the case of LE

			N=200				N=600				N=1000			
		v		v̂				v̂				v̂		
K	u	R=2	û	R=2	R=3	R=4	û	R=2	R=3	R=4	û	R=2	R=3	R=4
100	0.042	0.988	0.039	0.986	0.996	0.999	0.041	0.985	0.996	0.999	0.041	0.982	0.994	0.999
200	7.42E−3	0.999	6.21E−3	1.0	0.999	1.0	6.57E−3	1.0	1.0	1.0	6.66E−3	0.999	1.0	1.0
400	2.84E−4	1.0	2.67E−4	1.0	1.0	1.0	2.17E−4	1.0	1.0	1.0	2.21E−4	1.0	1.0	1.0
800	5.47E−7	1.0	3.52E−7	1.0	1.0	1.0	3.39E−7	1.0	1.0	1.0	2.76E−7	1.0	1.0	1.0
1000	2.54E−8	1.0	0.0	1.0	1.0	1.0	5.56E−9	1.0	1.0	1.0	6.01E−9	1.0	1.0	1.0

Type I error ($\hat{u}$) is calculated as the average of the number of {YIY>s,R=1}/(N(N−1)/2) in simulation times (1000), and the power ($\hat{v}$) is the average of the cardinal number of {YIY>s,R=r}, r=2, 3, 4 in simulation times. The expected value for type I error and power (u, v) is calculated by the upper/lower probability of the normal distribution with mean $E(Y)$ and variance $V(Y)$ and w=0 for Y.

Table 2 Estimated type I error and power in the case of LD

				N=200				N=600				N=1000			
					v̂				v̂				v̂		
			v_{10}												
K	u_{10}	u_0	R=2	û	R=2	R=3	R=4	û	R=2	R=3	R=4	û	R=2	R=3	R=4
100	0.207	3.90E−2	0.855	0.203	0.849	0.882	0.981	0.204	0.864	0.880	0.977	0.204	0.843	0.867	0.979
200	0.124	6.34E−3	0.933	0.114	0.921	0.949	1.0	0.115	0.933	0.950	0.998	0.116	0.928	0.945	0.996
1000	4.83E−3	1.25E−8	1.0	2.96E−3	1.0	1.0	1.0	3.13E−3	1.0	1.0	1.0	3.20E−3	1.0	1.0	1.0
3000	3.70E−6	0	1.0	8.54E−7	1.0	1.0	1.0	9.63E−7	1.0	1.0	1.0	1.06E−6	1.0	1.0	1.0
5000	3.61E−9	0	1.0	5.03E−8	1.0	1.0	1.0	0.0	1.0	1.0	1.0	2.00E−9	1.0	1.0	1.0

Type I error ($\hat{u}$) is calculated as the average of the number of {YIY>s,R=1}/(N(N−1)/2 in simulation times (1000), and the power ($\hat{v}$) is the average of the cardinal number of {YIY>s,R=r}, r=2,3,4 in simulation times. The expected value for type I error and power (u, v) is calculated by the upper/lower probability of the normal distribution with mean $E(Y)$ and variance $V(Y)$ and w=0 for Y. (u_0, u_{10}) correspond to the expected type I error for w=0, 10, respectively.

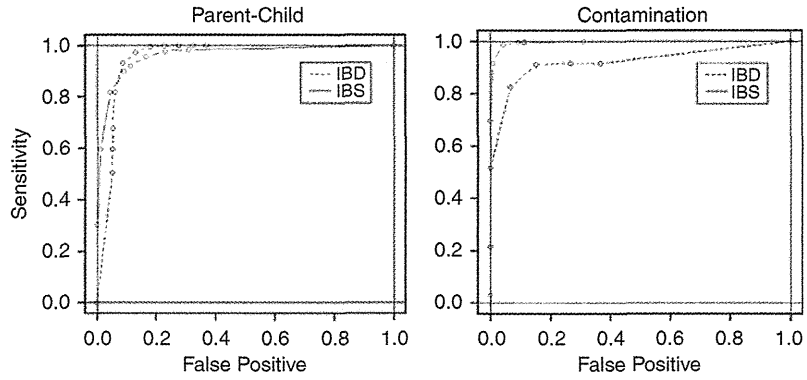

Figure 2 ROC curve for the performance of IBD/IBS-based methods applied to LD simulation data (K=200, N=200). AUC is 0.95 (IBD) and 0.96 (IBS) for parent–child, 0.92 (IBD) and 0.99 (IBS) for contamination.

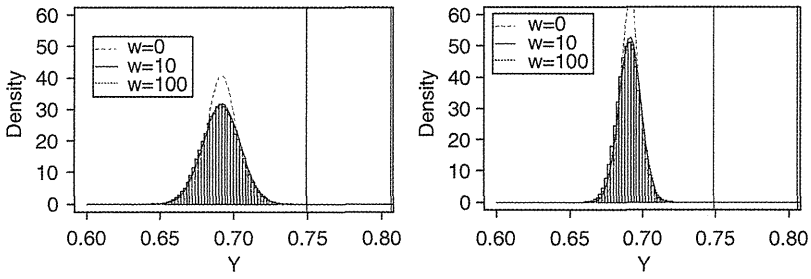

Figure 3 Histogram of real case–control data and theoretical distribution of Y, (K=1000, 2665). The threshold is s=0.75.

state Z=z for the entire genome. The IBD-based method can detect parent–child samples by estimating $P(Z=1)$, and can detect contaminated samples by estimating $\hat{\pi} = P(Z=1)/2+P(Z=2)$. As for parent–child samples, the IBS-based method in this simulation suppressed the false positive to a lower level than the IBD-based method. For contaminated samples, detection by the IBS-based method was more accurate than the IBD-based method (Figure 2). Sample contamination leads to too many heterozygote calls, which leads to fewer IBS 0 calls, which leads to an overestimated IBD and increased $\hat{\pi}$. However, in the contamination process simulated in this study, the genotypes of the contaminated samples were stochastically assigned as heterozygote and the IBD estimates did not increase as much, so the performance of the IBD-based method was low.

Although the number of SNPs is insufficient to detect inappropriate samples accurately according to Table 2, we focus in this simulation on the association study in which the number of SNPs is less than 1000. Moreover, we confirm that there is no difference in the performance between the two methods in the case of 1000 SNPs, and that both methods detect inappropriate samples accurately (data not shown).

Real data analysis

We applied the IBS-based method to the real association study data while changing the number of SNPs (K=200, 600, 1000 and 2665). This real data had a weak LD overall (Figure 1). It was possible to approximate the distribution of Y by a normal distribution, and there was little difference between w=10 and w=100 (Figure 3). In the case of a weak LD, the estimation accuracy of Y could be improved by considering LD. The number of detected sample pairs was estimated accurately by an upper probability of normal distribution (Table 3). The detected two sample pairs were rechecked by clinical-side investigators, and a sibling relationship was in fact strongly suggested.

DISCUSSION

In an association study, a series of QC is essential to maintain research quality. In this study, we focused on the detection of inappropriate samples. Previously, the IBS-based detection methods were proposed in family-based studies.[5,6] However, these methods did not consider LD among genetic markers and thus cannot be applied to association study data with LD. Our new IBS-based detection method can consider LD by using the covariance of Y, and the type I error and the power of the proposed method were able to be evaluated accurately by a simulation study. In a typical association study with only a few inappropriate samples, it is necessary to evaluate type I error correctly to avoid inadvertent exclusion of appropriate samples. In the simulation data, the proposed method detected inappropriate samples correctly and more accurately than did the IBD-based method.

In our simulation study the number of false positive drastically decreases when more than 1000 SNPs are analyzed (Table 2), and the PLINK website also reports that a large number of SNPs (1000 independent SNPs at a minimum) is required to calculate genome-wide IBD given IBS information. Taken together, this means that more than 1000 SNPs are required to detect inappropriate samples accurately. However, in some candidate gene approaches the target genes have been defined already and the number of typing SNPs on these genes is less than 1000 SNPs. In such a case we recommend the proposed method.