Characterization of Gene Expression Patterns among Artificially Developed Cancer Stem Cells Using Spherical Self-Organizing Map

Akimasa Seno[1,*], Tomonari Kasai[1,*], Masashi Ikeda[1], Arun Vaidyanath[1], Junko Masuda[1], Akifumi Mizutani[1], Hiroshi Murakami[1], Tetsuya Ishikawa[2,3] and Masaharu Seno[1]

[1]Laboratory of Nano-Biotechnology, Department of Medical Bioengineering Science, Graduate School of Natural Science and Technology, Okayama University, Kita-ku, Okayama, Japan. [2]Cell Biology, Core Facilities for Research and Innovative Medicine, National Cancer Center Research Institute, Chuo-ku, Tokyo, Japan. [3]Central Animal Division, Fundamental Innovative Oncology Core Center, National Cancer Center Research Institute, Chuo-ku, Tokyo, Japan. *These two authors equally contributed to the present study.

ABSTRACT: We performed gene expression microarray analysis coupled with spherical self-organizing map (sSOM) for artificially developed cancer stem cells (CSCs). The CSCs were developed from human induced pluripotent stem cells (hiPSCs) with the conditioned media of cancer cell lines, whereas the CSCs were induced from primary cell culture of human cancer tissues with defined factors (*OCT3/4*, *SOX2*, and *KLF4*). These cells commonly expressed human embryonic stem cell (hESC)/hiPSC-specific genes (*POU5F1, SOX2, NANOG, LIN28*, and *SALL4*) at a level equivalent to those of control hiPSC 201B7. The sSOM with unsupervised method demonstrated that the CSCs could be divided into three groups based on their culture conditions and original cancer tissues. Furthermore, with supervised method, sSOM nominated *TMED9, RNASE1, NGFR, ST3GAL1, TNS4, BTG2, SLC16A3, CD177, CES1, GDF15, STMN2, FAM20A, NPPB, CD99, MYL7, PRSS23, AHNAK*, and *LOC152573* genes commonly upregulating among the CSCs compared to hiPSC, suggesting the gene signature of the CSCs.

KEYWORDS: cancer stem cell, spherical self-organizing map, hiPSC

CITATION: Seno et al. Characterization of Gene Expression Patterns among Artificially Developed Cancer Stem Cells Using Spherical Self-Organizing Map. *Cancer Informatics* 2016:15 163–178 doi: 10.4137/CIN.S39839.

TYPE: Original Research

RECEIVED: April 04, 2016. **RESUBMITTED:** May 15, 2016. **ACCEPTED FOR PUBLICATION:** May 30, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 2,437 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by JSPS Grants-in-aid for Scientific Research (A) no. 25242045 and (C) no. 16K07135 and for Challenging Exploratory Research no. 26640079 from the Ministry of Education, Culture, Sports, Science and Technology. Life-Science Intellectual Property Platform Fund also supported this work. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: TI discloses patents pending (WO2011148983 A1 and WO2013081188 A1) relevant to the work presented here. MS discloses collaboration with Mitsubishi Tanabe Pharma, Toppan Printing Co Ltd, Carna Biosciences Inc, Wako Pure Chemical Industries, and Ensuiko Sugar Refining Co Ltd, outside the work presented here. In addition, MS discloses patents pending (PCT/JP2014/057572 and PCT/JP2016/060309) relevant to the work presented here. Other authors disclose no potential conflicts of interest.

CORRESPONDENCE: mseno@okayama-u.ac.jp; teishika@ncc.go.jp

Paper subject to independent expert blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Cancer stem cells (CSCs) are thought to possess stemness, the capacity of self-renewal and multipotent differentiation. Such CSCs have been found in patients with acute myeloid leukemia[1] and other cancers.[2–8] As these cells might cause relapse, metastasis, and drug resistance of cancer, cancer therapy targeting CSCs would be an attractive strategy to cure cancer patients. Although it is important to identify the characteristic markers of CSCs, they would constitute only a small population in cancer tissues to analyze. Recently, induced pluripotent stem cells (iPSCs) have been generated from somatic cells by reprograming to have the ability of self-renewal and pluripotency.[9] With this technique, the development of artificial CSCs has been reported. We converted mouse iPSCs to have CSC properties, and another group reprogrammed human cancer cell lines to have CSC properties through the process of iPSC preparation. Both approaches were successful to demonstrate CSC properties to form spheres *in vitro* and malignant tumors *in vivo*.[10]

In this study, we newly developed the CSCs that were derived from human iPSCs (hiPSCs) with the conditioned media of cancer cell lines or that were induced from primary cell culture of human cancer tissues with defined factors (*OCT*3/4, *SOX*2, and *KLF*4). The CSCs were analyzed using gene expression microarray coupled with the clustering procedure of spherical self-organizing map (sSOM).

Materials and Methods

Induction of CSCs from primary cell culture of human cancer tissues with defined factors. The anonymous remnant human cancer tissue samples were provided via the Health Science Research Resources Bank. Written informed consent from donors was obtained for the use of these samples in research. The study was done under the approval of the Institutional Review Boards of the National Cancer Center of Japan and the Japan Health Sciences Foundation/the Health Science Research Resources Bank. The Health Science Research Resources Bank has been currently transferred to Japanese

Collection of Research Bioresources, National Institutes of Biomedical Innovation, Health and Nutrition (http://bioresource.nibiohn.go.jp/human/index.html). The cancer tissues were derived from pathologically defined cancerous parts of the colon (from a Japanese male, 55 years old) and the stomach (from a Japanese male, 67 years old) as surgical waste after an operation. The cancer tissue-derived cell suspensions were prepared as previously described.[11]

The cancer tissue-derived cells were seeded on collagen-coated dishes with Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum. One day later, the cells at approximately 5%–10% confluency were incubated with the pantropic retrovirus vector solution (*OCT3/4, KLF4*, and *SOX2*) at 37 °C for one day. The pantropic retrovirus vector solution was prepared as previously described.[11] The study was approved by the Institutional Recombinant DNA Advisory Committee of the National Cancer Center. Mitomycin C-treated mouse embryonic fibroblasts (MEFs) were seeded following the infection. The culture was replaced with confluency. The confluent culture was further refreshed with mTeSR1 medium (STEMCELL Technologies) every day from day 22 (for colon cancer tissue-derived cells) and day 15 (for gastric cancer tissue-derived cells). Clones iPS-CC1-4, iPS-CC1-10, iPS-CC1-11, iPS-CC1-17, iPS-CC1-18, and iPS-CC1-25 were isolated from primary cell culture of human colon cancer tissues. These clones were designated as iPS-CC1. Clones iPS-GC1-1, iPS-GC1-2, iPS-GC1-3, iPS-GC1-5, iPS-GC1-7, and iPS-GC1-8 were isolated from primary cell culture of human gastric cancer tissues. These clones were designated as iPS-GC1.

The isolated clones were subcultured in each well of gelatin-coated 24-well plates. After an expansion culture, each clone was further cultured in each well of gelatin-coated six-well plates and finally cultured in a gelatin-coated 100-mm dish. The expanded clones were treated with a dissociation solution (0.25% trypsin–EDTA; Gibco, and 1% collagenase; Invitrogen) or 0.25% trypsin–EDTA and passaged in mTeSR1 supplemented with 10–20 μM Y-27632 (Calbiochem and Wako) to avoid cell death as previously described.[12] The clones were cultured with the MEFs (5×10^4 cells/cm^2) mainly in TeSR1 medium and occasionally in primate ESC medium (ReproCell) in gelatin-coated dishes. Using the AllPrep DNA/RNA Mini Kit (Qiagen), total RNA was prepared from each clone that was cultured with the MEFs (5×10^4 cells/cm^2) in mTeSR1 medium in gelatin-coated 100-mm dishes before long-term serial passage.

Induction of CSCs from hiPSCs. The cancer cell lines listed in Table 1 were cultured in adherent 100-mm-diameter culture dishes (Techno Plastic Products AG) in DMEM medium or RPMI1640 medium containing 10% FBS supplemented with 1% penicillin/streptomycin at 37 °C under the atmosphere of 5% CO_2. The conditioned medium from each of the cell lines was collected and mixed with Repro FF2, Repro stem (ReproCELL Inc.), or bFGF-free human iPS stem cell medium consisting of DMEM-F12 medium supplemented with nonessential amino acid, 2.5 mM L-glutamine, KnockOut Serum Replacement (Thermo Fisher Scientific), and 0.1 mM 2-mercaptoethanol at a ratio of 1:1 to prepare a differentiation induction medium. hiPSCs (201B; RIKEN BioResource Center)[13] kept undifferentiated were cultured

Table 1. CSCs developed from hiPSCs and human cancer cell lines, of which conditioned medium prepared for the treatment.

CSC NAME	CANCER CELL LINE	ORIGIN	MEDIUM
OCC-hiPS-6	ZR-75-1	Breast	DMEM
OCC-hiPS-10	HT-29	Colon	DMEM
OCC-hiPS-12	SKOV3	Ovary	DMEM
OCC-hiPS-16	ECC4	Gastrointestinal	DMEM
OCC-hiPS-17	CW-2	Colon	DMEM
OCC-hiPS-19	MY	Lymphocyte	DMEM
OCC-hiPS-20	MOLT4	T-cell leukemia	DMEM
OCC-hiPS-25	Li-7	Hepatocellular	RPMI 1640
OCC-hiPS-27	Lu99B	Lung	RPMI 1640

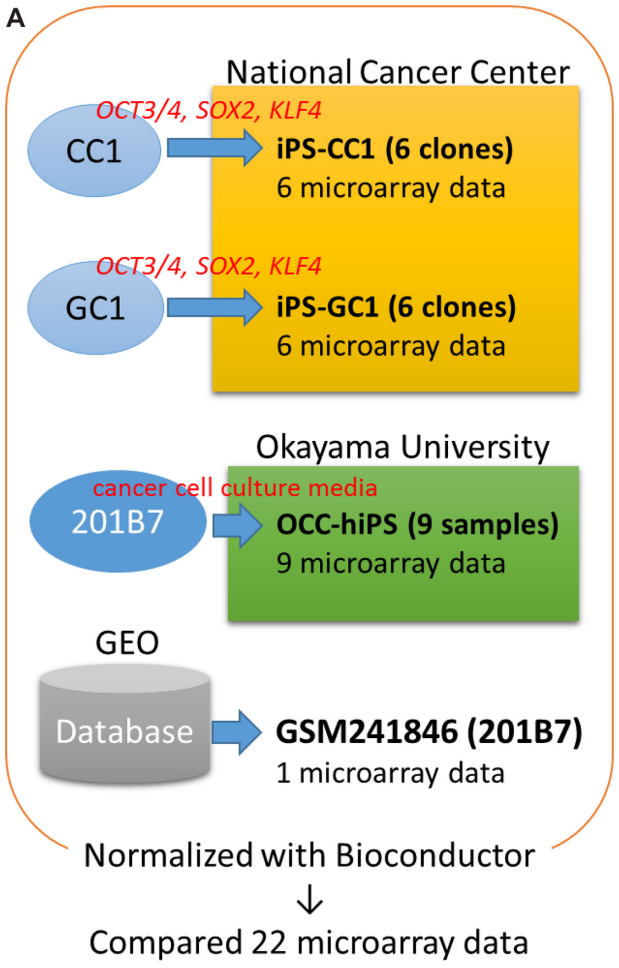

Figure 1. (*Continued*)

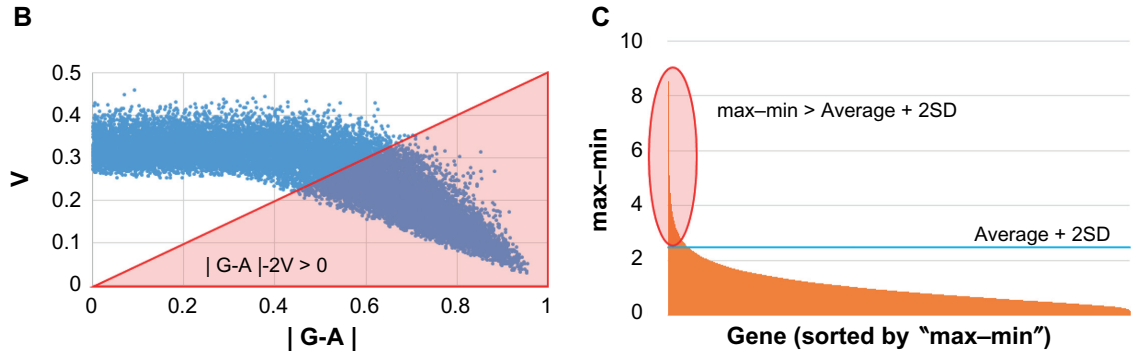

Figure 1. Flowchart of the experimental procedure.

Notes: (**A**) Twenty-two samples were analyzed with microarray experiments, and the data were compared with hiPSC 201B7 data from GEO (GSM241846) after normalization. (**B** and **C**) To detect differentially expressed genes/probes, two parameters were used for gene selection; one is (**B**) |G-A|-2V > 0 and another is (**C**) max–min > average + 2SD. (**B**) G, A, and V are denoted as follows: the average of gene expression level among the CSCs, the gene expression level of hiPSC 201B7, and the SD of the gene expression level among the CSCs, respectively. These values were calculated with *I*, which was described in the "Materials and methods" section. (**C**) *Average + 2SD* was calculated with the *max–min* value. These values were calculated with Bioconductor normalized intensity for each gene. Normalized intensity *i'* was shown in base-2 logarithm on Y-axis. For Figure 2, a gene set was made by only one parameter (**B**). To list up genes that have much difference, parameter (**C**) in addition to (**B**) was used for each gene set of Figures 3–5. Using each gene set, sSOM analysis was performed with *I*.

in induction medium to allow differentiation. During the induction of differentiation, half of the medium was exchanged every day, and the cells were passaged once or twice every two weeks. The period of induction of differentiation was at least 28 days, and at most two months. The cells were cultured at 37 °C under the atmosphere of 2% CO_2.

Gene expression analysis. For iPS-CC1 and iPS-GC1, the microarray study was carried out using a Whole Human Genome Oligo Microarray 4x44 K (Agilent Technologies). The analysis was performed according to the Agilent technical protocols. RNA was quantified using a NanoDrop ND-1000 spectrophotometer, and quality was monitored using the Agilent 2100 Bioanalyzer (Agilent Technologies). Cyanine-3 (Cy3)-labeled cRNA was prepared from 0.5 µg RNA using the One-Color Low RNA Input Linear Amplification PLUS Kit (Agilent Technologies) according to the manufacturer's instructions, followed by RNeasy column purification (Qiagen).

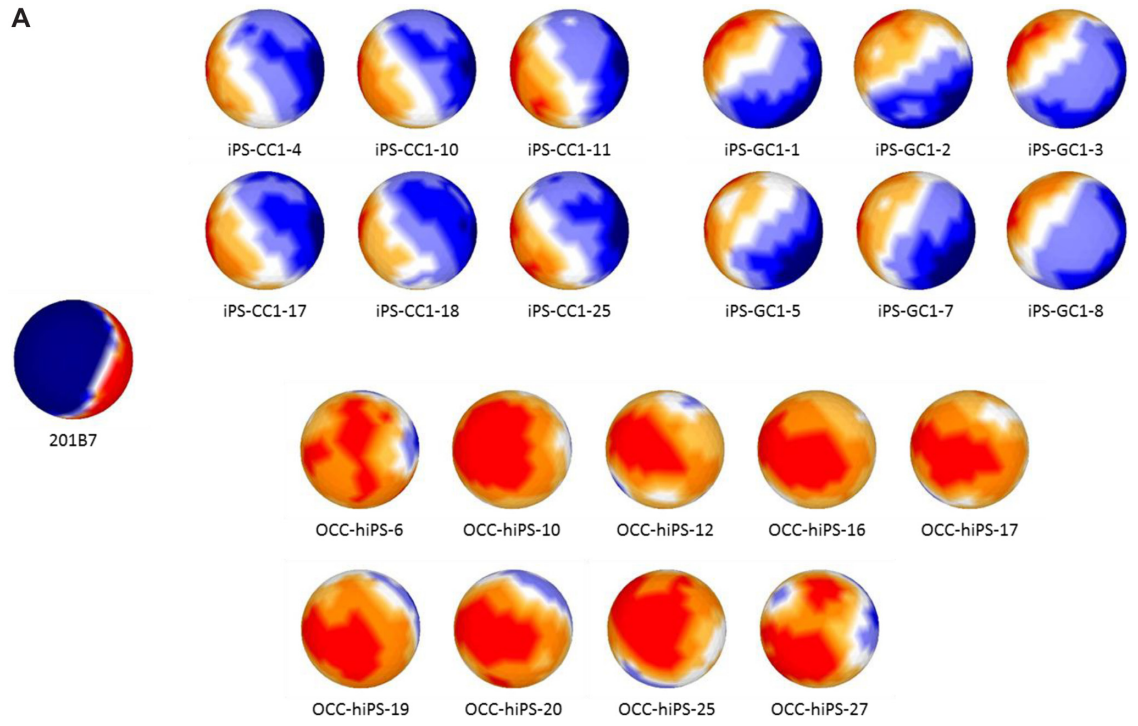

Figure 2. (*Continued*)

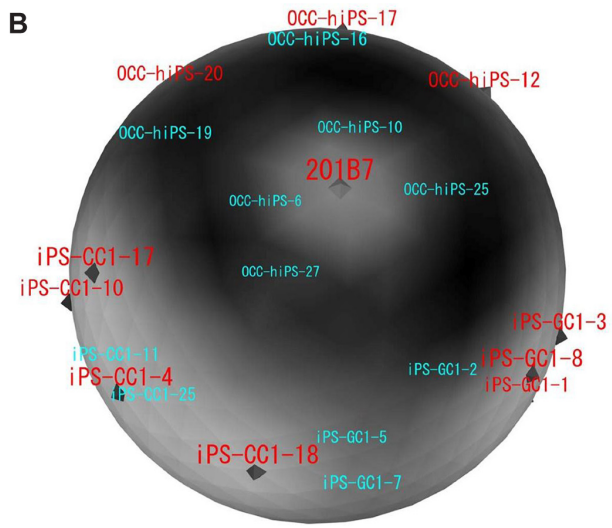

Figure 2. Mapping and clustering of normal hiPS and all the CSCs with sSOM. Microarray data of hiPSC 201B7 were obtained from NCBI GEO (GSM241846), and those of the CSCs were obtained as our original data.

Notes: (**A**) Gene expression patterns were analyzed by sSOM with the microarray data of GSM241846 and the CSCs. The data were used 2678 probes, which were extracted by |A-G|-2V > 0. (**B**) Each of the CSCs and hiPSC 201B7 was mapped on a sphere by sSOM analysis. The CSCs were clustered into three groups with sSOM. Each of analyzed CSCs was depicted on a sphere. The CSCs named in red color were mapped on the front side of the sphere. The CSCs named in light blue color were mapped on the back side of the sphere.

Dye incorporation and cRNA yield were checked using the NanoDrop ND-1000 Spectrophotometer. A total of 1.5 μg of Cy3-labeled cRNA (specific activity >10.0 pmol Cy3/μg cRNA) was fragmented at 60 °C for 30 minutes in a reaction volume of 250 mL containing 1 × Agilent fragmentation buffer and 2 × Agilent blocking agent following the manufacturer's instructions. On completion of the fragmentation reaction, 250 mL of 2 × Agilent hybridization buffer was added to the fragmentation mixture and hybridized to Agilent Whole Human Genome Oligo Microarrays (G4112 A) for 17 hours at 65 °C in a rotating Agilent hybridization oven. After hybridization, microarrays were washed for one minute at room temperature with GE Wash Buffer 1 (Agilent Technologies) and one minute with 37 °C GE Wash Buffer 2 (Agilent Technologies) and then dried immediately by brief centrifugation. Slides were scanned immediately after washing on the Agilent DNA Microarray Scanner (G2505B) using one color scan setting for 1x44k array slides (scan area 61 mm × 21.6 mm, scan resolution 10 μm, dye channel is set to Green, and Green PMT is set to 100%). The scanned images were analyzed with Feature Extraction Software 9.1 (Agilent Technologies) using default parameters (protocol GE1-v5_95_Feb07 and Grid: 014850_D_20070820) to obtain background subtracted and spatially detrended Processed Signal intensities. Features flagged in feature extraction as feature nonuniform outliers were excluded. Data (GSM241846) from the Gene Expression Omnibus was used as typical hiPSCs (201B7).[13]

For Okayama CSC collection (OCC)-hiPS cells, a SurePrint G3 Human GE 8x60 K v2 Microarray (Agilent Technologies) was used for the microarray study. RNA was quantified using a NanoDrop, and quality was monitored with the Agilent 2100 Bioanalyzer (Agilent Technologies). Cy3-labeled cRNA was prepared from 10 to 200 ng RNA using Low Input Quick Amp Labeling Kit, one-color (Agilent Technologies) according to the manufacturer's instructions, followed by RNeasy column purification (Qiagen). A total of 600 ng of Cy3-labeled cRNA was fragmented at 60 °C for 30 minutes and hybridized for 17 hours at 65 °C with Gene Expression Hybridization Kit (Agilent Technologies). After hybridization, microarrays were washed with Gene Expression Wash Buffers Pack (Agilent Technologies) and scanned on the Agilent DNA Microarray Scanner (G2565CA). The scanned images were analyzed with Feature Extraction Software 10.10.1.1 (Agilent Technologies) using parameters (protocol GE1_1010_Sep10 and Grid: 039494_D_F_20120628) to obtain background subtracted and spatially detrended Processed Signal intensities. Features flagged in feature extraction as feature nonuniform outliers were excluded.

Numeric intensity data were normalized with Bioconductor[14,15] package *agilp* (ver.3.2.0, https://bioconductor.org/packages/release/bioc/html/agilp.html)[16] as directed by maintainer's manual. Briefly, the raw intensity data were mapped to the same ID with *IDswop*. Mapped data were trimmed with *Equaliser* so as to include only the set of genes that are common to all data. Then, a baseline was generated by *Baseline*, and a set of gene expression data files were normalized by *AALoess*. Through these procedures, 18,561 genes were assessed for the expression from all the set of data. After this normalized procedure, housekeeping genes (*ACTB, ATP5F1, GAPDH, GAPDHS, GUSB, GUSBL1, GUSBL2, HPRT1, PGK1, PPIA, PPIAL4, RPLP0, RPLP1, RPLP2, RPS18, TBP, TBPL1, TFRC*, and *YWHAZ*) and hESC/hiPSC-enriched genes (*POU5F1, SOX2, NANOG, LIN28, SALL4, TDGF1, DNMT3B, ZFP42, TERT, GDF3, CYP26A1, DPPA4, PODXL*, and *ZIC3*) of the CSCs expressed at a level equivalent to those of hiPSCs (201B7) (Supplementary Fig. 1).

Data filtering and sSOM analysis. A data filtering (with parameter (B) shown in Fig. 1) was performed to extract genes of which expression showed significant difference between the CSCs prepared in this study and normal iPSC 201B7. The feature scaled intensity (I) was defined as following:

$$I = \frac{i' - \min}{\max - \min},$$

where i': normalized intensity of each probe, min: the minimum normalized intensity of a probe among all analyzed

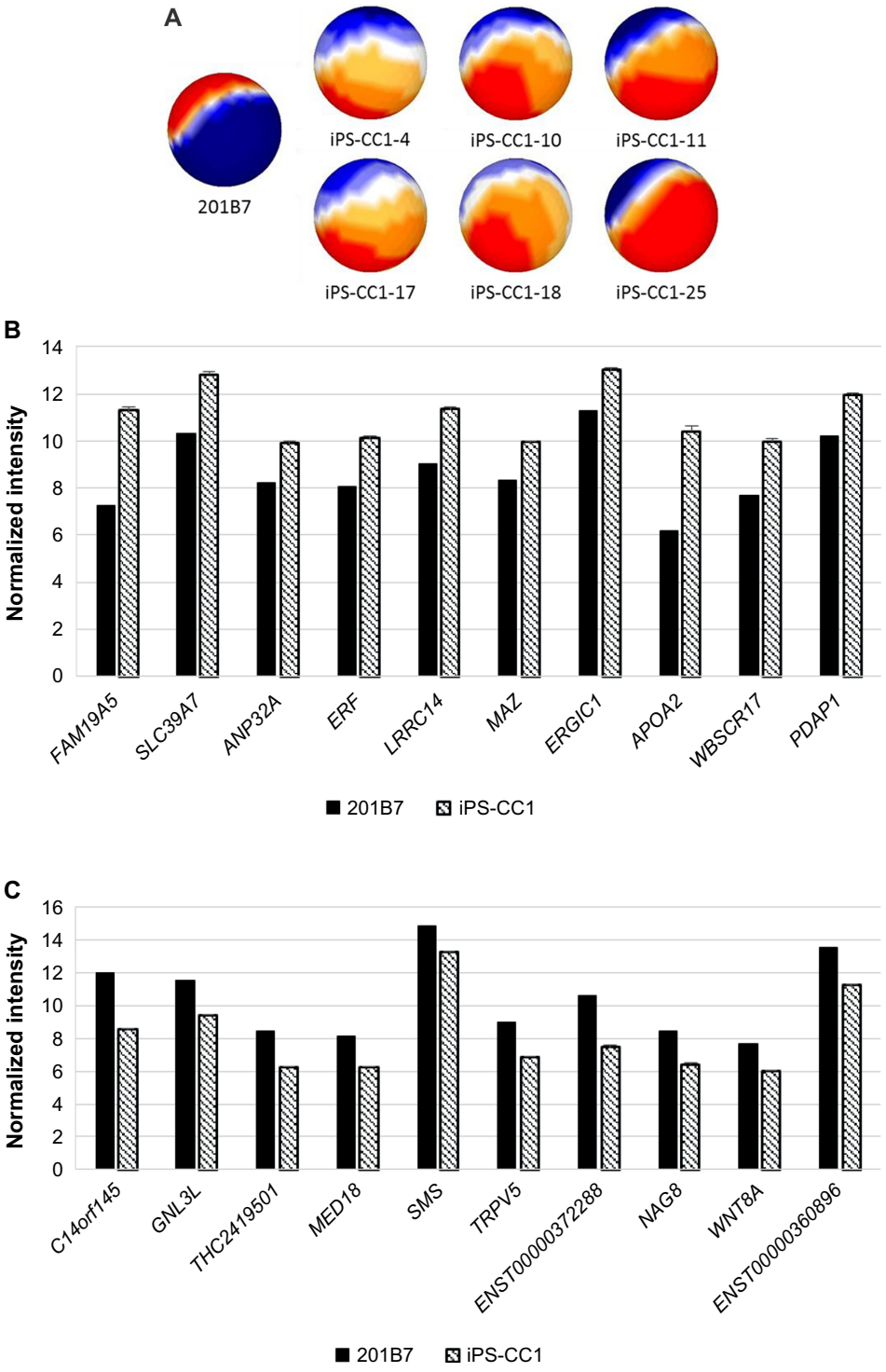

Figure 3. Mapping and comparison of normal hiPSC and iPS-CC1 cells with sSOM.

Notes: (**A**) Gene expression patterns analyzed by sSOM with the microarray data of 201B7 (GSM241846) and iPS-CC1. The data were used 598 genes, which were extracted by the two parameters (see Fig. 1). Each of iPS-CC1 was mapped as a sphere by sSOM analysis. The normalized intensities of 323 upregulating genes (**B**) or 275 downregulating genes (**C**) in iPS-CC1, which were compared to GSM241846, were analyzed by sSOM. Ten genes close to the IP were aligned by the order of NSD as listed in Tables 2 and 3. Graphs were depicted as mean + SD. Normalized intensity *i'* was shown in base-2 logarithm on Y-axis. Y-linked genes were eliminated from the list because sex differences were confounding factor.

samples, and max: the maximum normalized intensity of a probe among all analyzed samples. Probes were extracted with the *I* value for each probe that was evaluated with the scores defined by a filtering formula '|G-A|-2V', where *'G'*, *'A'*, and *'V'* denote the average expression level of a gene among the CSCs, the expression level of a gene of normal iPSC, and the standard deviation (SD) of a gene expression level among the CSCs, respectively. As an additional filtering

Table 2. Top 10 upregulating genes of hiPS-CC1 compared with hiPSC 201B7 except for Y-related genes.

NSD	GENE	DESCRIPTION	ACCESSION NO.
1.07	*FAM19A5*	Family with sequence similarity 19 (chemokine (C-C motif)-like), member A5 (FAM19A5), mRNA	NM_015381
1.76	*SLC39A7*	Homo sapiens solute carrier family 39 (zinc transporter), member 7 (SLC39A7), mRNA	NM_006979
1.82	*ANP32A*	Acidic (leucine-rich) nuclear phosphoprotein 32 family, member A (ANP32A), mRNA	NM_006305
1.86	*ERF*	Ets2 repressor factor (ERF), mRNA	NM_006494
2.30	*LRRC14*	Leucine rich repeat containing 14 (LRRC14), mRNA	NM_014665
3.14	*MAZ*	MYC-associated zinc finger protein (purine-binding transcription factor) (MAZ), transcript variant 1, mRNA	NM_002383
3.15	*ERGIC1*	Endoplasmic reticulum-golgi intermediate compartment (ERGIC) 1 (ERGIC1), transcript variant 1, mRNA	NM_001031711
3.15	*APOA2*	Apolipoprotein A-II (APOA2), mRNA	NM_001643
3.16	*WBSCR17*	Williams-Beuren syndrome chromosome region 17 (WBSCR17), mRNA	NM_022479
3.17	*PDAP1*	PDGFA associated protein 1 (PDAP1), mRNA	NM_014891

Abbreviation: NSD, nonsignificant distance.

Table 3. Top 10 downregulating genes of hiPS-CC1 compared with hiPSC 201B7.

NSD	GENE	DESCRIPTION	ACCESSION NO.
0.37	*C14orf145*	Chromosome 14 open reading frame 145 (C14orf145), mRNA	NM_152446
0.37	*GNL3L*	Guanine nucleotide binding protein-like 3 (nucleolar)-like (GNL3 L), mRNA	NM_019067
0.84	*THC2419501*	RL31_HUMAN (P62899) 60S ribosomal protein L31, partial (97%)	THC2419501
1.41	*MED18*	Mediator of RNA polymerase II transcription, subunit 18 homolog (S. cerevisiae) (MED18), mRNA	NM_017638
1.48	*SMS*	Spermine synthase (SMS), mRNA	NM_004595
1.50	*TRPV5*	Transient receptor potential cation channel, subfamily V, member 5 (TRPV5), mRNA	NM_019841
1.52	*ENST00000372288*	PREDICTED: similar to nuclear DNA-binding protein (LOC642521), mRNA	XM_926017
1.54	*NAG8*	Nasopharyngeal carcinoma associated gene protein-8 (NAG8), mRNA	NM_014411
1.58	*WNT8A*	Wingless-type MMTV integration site family, member 8A (WNT8A), mRNA	NM_058244
1.76	*ENST00000360896*	Full-length cDNA clone CS0DL004YD09 of B cells (Ramos cell line) Cot 25-normalized of (human)	CR614522

Abbreviation: NSD, nonsignificant distance.

(with parameter (c) shown in Fig. 1), to find a significant difference between the CSCs and hiPSC, the max–min difference (max-min) of normalized intensity (i') for each probe was calculated and then a probe was chosen if 'max-min' of each probe was larger than 'average+2SD' of 'max-min' of all probes (max-min > average+2SD) (Fig. 1C). The resulting data set (using parameter (B) or (B) plus (C)) was used for mapping probes by the sSOM software *Blossom* (SOM Japan; http://www.somj.com/). In clustering of probes, IP was included as an *ideal probe* of virtual probe with all $I = 1$ or 0 of the CSCs while $I = 0$ or 1 in normal hiPSC, respectively. Nonsignificant distance (NSD) was calculated as the distance between each probe and IP under the default sSOM parameters. To integrate the resolution, the top 50 probes mapping at the positions closest to IP were selected and the selected probes were subjected to sSOM analysis again to select the top 10 probes.

Results

Visualization of expression patterns by sSOM clustering. DNA microarray analysis was performed to characterize the CSCs that were induced from the cancer tissue-derived cells with defined factors and that were converted from hiPSC 201B7 with the conditioned media of cancer cell lines. As a common control, hiPSC 201B7 (GSM241846) was employed, which had been scanned by an Agilent DNA microarray scanner G2505B.[13] Although the microarray scanning of the CSCs was independently performed, the data could be normalized with Bioconductor package called 'agilp', which was specialized in normalizing Agilent microarray data (Fig. 1A).

For sSOM analysis, normalized intensities were used, which were feature scaled (0-1) as I defining in Material and Methods. As a result of data filtering with '$|G\text{-}A|\text{-}2V > 0$', which was modified from our previous reports,[17,18] 2678 probes were extracted with potentially significant differences

Table 4. Top 10 upregulating genes of hiPS-GC1 compared with hiPSC 201B7 except for Y-related genes.

NSD	GENE	DESCRIPTION	ACCESSION NO.
0.62	*H2AFY2*	H2A histone family, member Y2 (H2AFY2), mRNA	NM_018649
0.80	*MT2A*	Metallothionein 2A (MT2A), mRNA	NM_005953
1.37	*APH1A*	Anterior pharynx defective 1 homolog A (C. elegans) (APH1A), mRNA	NM_016022
1.76	*ID2*	Inhibitor of DNA binding 2, dominant negative helix-loop-helix protein (ID2), mRNA	NM_002166
1.76	*ID2*	Inhibitor of DNA binding 2, dominant negative helix-loop-helix protein (ID2), mRNA	NM_002166
1.78	*PBX2*	Pre-B-cell leukemia transcription factor 2 (PBX2), mRNA	NM_002586
1.83	*PSPH*	Phosphoserine phosphatase (PSPH), mRNA	NM_004577
1.84	*FAM19A5*	Family with sequence similarity 19 (chemokine (C-C motif)-like), member A5 (FAM19A5), mRNA	NM_015381
1.87	*ANP32D*	Acidic (leucine-rich) nuclear phosphoprotein 32 family, member D (ANP32D), mRNA	NM_012404
1.87	*ANP32A*	Acidic (leucine-rich) nuclear phosphoprotein 32 family, member A (ANP32A), mRNA	NM_006305

Abbreviation: NSD, nonsignificant distance.

Table 5. Top 10 downregulating genes of hiPS-GC1 compared with hiPSC 201B7.

NSD	GENE	DESCRIPTION	ACCESSION NO.
0.40	*GNPTAB*	N-acetylglucosamine-1-phosphate transferase, alpha and beta subunits (GNPTAB), mRNA	NM_024312
0.67	*CDC2L5*	Cell division cycle 2-like 5 (cholinesterase-related cell division controller) (CDC2L5), transcript variant 1, mRNA	NM_003718
0.72	*C14orf145*	Chromosome 14 open reading frame 145 (C14orf145), mRNA	NM_152446
0.82	*ENST00000372288*	PREDICTED: similar to nuclear DNA-binding protein (LOC642521), mRNA	XM_926017
0.93	*SOHLH2*	Spermatogenesis and oogenesis specific basic helix-loop-helix 2 (SOHLH2), mRNA	NM_017826
0.96	*SYT1*	Synaptotagmin I (SYT1), mRNA	NM_005639
0.96	*MPPED2*	Metallophosphoesterase domain containing 2 (MPPED2), mRNA	NM_001584
1.41	*LHX4*	LIM homeobox 4 (LHX4), mRNA	NM_033343
1.41	*GNL3L*	Guanine nucleotide binding protein-like 3 (nucleolar)-like (GNL3 L), mRNA	NM_019067
1.43	*RUNX1T1*	Runt-related transcription factor 1; translocated to, 1 (cyclin D-related) (RUNX1T1), transcript variant 1, mRNA	NM_004349

Abbreviation: NSD, nonsignificant distance.

(Fig. 1B). The resulting probes were then analyzed by sSOM software with unsupervised method. The results of sSOM were mapped as the gene expression patterns visualizing on the spherical surfaces (Fig. 2A and Supplementary Fig. 2). It is noteworthy that each pattern of the CSCs appeared similar one another in each of three clustered CSC group but different from that of iPSC 201B7. Otherwise, the grouping of the CSCs was indicated by spotting each of the CSCs on a sphere, which were characterized using the identical gene set of Figure 2A. As shown in Figure 2B, the grouping of the CSCs was indicated by spotting each of the CSCs on a sphere, which were characterized using the identical gene set of Figure 2A. The CSCs were also confirmed to be clustered into the three groups different from hiPSC 201B7 by sSOM. Thus, the gene expression profiles were considered to be visualized by the sSOM mapping (Fig. 2A) and clustering (Fig. 2B) even when judged at a glance. The differences of three CSC groups were easily distinguished from one another and different from normal hiPSC as the mapping patterns.

To identify genes, which were commonly expressed in high or low level among all the CSCs in contrast to hiPSC, an ideal probe *IP* was inserted into the data set and analyzed with the 2,678 probes. 'IP' is defined as an ideal gene of which expression is limited only to either all the CSCs or hiPSC.[19,20] Theoretically, a gene of which expression is similar among those of all the CSCs should be located around IP by sSOM

mapping. Another factor was necessary to extract probes that show much difference because IP did not reflect the difference of *i'* between normal hiPSC and CSCs. Since the CSCs were clustered into three groups, each CSC group could be compared with normal hiPSC to investigate their significant difference, respectively. In the case, a difference between the maximum and the minimum value (max-min) was calculated for each of 18,561 probes. Probes were extracted when the 'max-min' was larger than the 'average+2SD' of all probes (Fig. 1C). For sSOM analysis, datasets of 598, 439 and 402 probes were utilized for comparisons between normal hiPSC and iPS-CC1, iPS-GC1, or OCC-hiPS, respectively.

***FAM19A5* is significantly upregulated in iPS-CC1.** One method to prepare the CSC was to infect defined factors (*OCT3/4, SOX2*, and *KLF4*) to primary culture cells derived from cancer tissues. With this idea, iPS-CC1 was induced from cells derived from colon cancer tissues and analyzed with sSOM. The sSOM sphere of gene expression is shown in Figure 3A. Their gene expression patterns on sSOM sphere were remarkably different from that of hiPS 201B7 cells, suggesting that the sSOM analysis was effective to select genes characteristic to iPS-CC1. Among significantly upregulating genes, IP showed *RPS4Y2/1* the most characteristic in iPS-CC1 (Supplementary Fig. 3 and Supplementary Table 1). However, this was probably a difference between the sex because 201B7 has 46 XX chromosome, whereas iPS-CC1 was induced from a male cancer patient (46XY). Except for Y chromosome-related genes, *FAM19A5* was considered as the most characteristic gene of iPS-CC1 (Fig. 3B and Table 2). In contrast, *C14orf145* and *GNL3L* were the most downregulating genes of iPS-CC1, although their difference of intensity was not so much large as those of upregulating genes (Fig. 3C and Table 3).

***MT2A* is significantly upregulated in iPS-GC1.** Six clones of iPS-GC1 series were induced from primary culture cells derived from gastric cancer tissues with defined factors. iPS-GC1 series were analyzed with sSOM. The sSOM sphere of gene expression was shown in Figure 4A. Their gene expression patterns on sSOM sphere of iPS-GC1 series were evidently distinct from that of hiPSC 201B7. A clustering analysis with IP extracted genes significantly upregulating among iPS-GC1. *PRS4Y1/2* was ranked as the top with the shortest NSD in iPS-GC1 because of the inconsistency of sex between iPS-GC1 from male and hiPS 201B7 cells from female (Supplementary Fig. 4 and Supplementary Table 2). Except for Y chromosome-related genes, *MT2A* was a gene closest to IP as significantly upregulating gene (Fig. 4B and Table 4). In contrast, *GNPTAB* was ranked as the most significantly downregulating gene in iPS-GC1 (Fig. 4C and Table 5).

Histones, TMED9, and CASKIN1 are indicated as the characteristic gene in the series of OCC-hiPS. The CSCs, which were converted from hiPSCs with the condition media of cancer cell lines, are now being registered in the OCC.[10] They were mapped by sSOM (Fig. 5A). Each sample of OCC-hiPS series was prepared from hiPSC 201B7 with each conditioned medium of each different cancer cell lines. As a result, each different gene expression pattern of the CSCs indicating different phenotypes could be induced under each different conditioned medium, as demonstrated in mouse iPSCs.[10,21] From a NSD by sSOM analyses, histone genes were nominated as the most upregulating genes in the CSCs of OCC-hiPS (Fig. 5B and Table 6). Histones might regulate gene expression causing various cell dysfunction, although further investigation would be required to clarify the mechanism underlying the upregulation. Transmembrane emp24 protein transport domain containing 9 (*TMED9*) and *CASKIN1* might be other candidates of the characteristic genes upregulating in the CSCs of OCC-hiPS. On the other hand, *AL832540* was ranked as the most downregulating gene in the CSCs of OCC-hiPS with the shortest NSD (Fig. 5C and Table 7).

***TMED9* is upregulated in all CSCs.** The aim of the sSOM analysis with microarray data was to identify a gene whose expression was commonly up/downregulated in all the CSCs. For this purpose, significantly up/downregulating genes, which were nominated in Figures 3–5, were summarized in the Venn diagrams (Fig. 6A and B). Using the diagrams, 18-upregulating genes and 15-downregulating genes were extracted from three of the CSC groups in common. Of these commonly upregulating genes, *TMED9* was the most characteristic gene in all the CSCs (Fig. 6C and D and Tables 8 and 9). *NPPB* seems to also have significance, although its NSD was larger than that of *TMED9*. Downregulating genes seem to be much less significant than upregulating genes (Figs. 3–5).

Discussion

The CSCs were converted from normal hiPSCs with the conditioned media of cancer cell lines; otherwise, the CSCs were induced from primary cell culture of human cancer tissues with defined factors. Gene expression microarray experiments confirmed that the CSCs and typical hiPSC commonly expressed many hESC/hiPSC-specific or -enriched genes. The sSOM demonstrated that the CSCs could be clustered into three groups due to their origins with unsupervised method. Nevertheless, the supervised method of sSOM identified *TMED9, RNASE1, NGFR, ST3GAL1, TNS4, BTG2, SLC16A3, CD177, CES1, GDF15, STMN2, FAM20A, NPPB, CD99, MYL7, PRSS23, AHNAK*, and *LOC152573* genes commonly upregulating among all the CSCs compared with normal hiPSC.

DNA microarray analyses allow us to perform large-scale and high-throughput screening of differentially expressed genes among many samples. To reveal the patterns of gene expression, a method to analyze and evaluate the large data generated by series of microarray experiments plays an important role. Hierarchical clustering and SOM clustering have

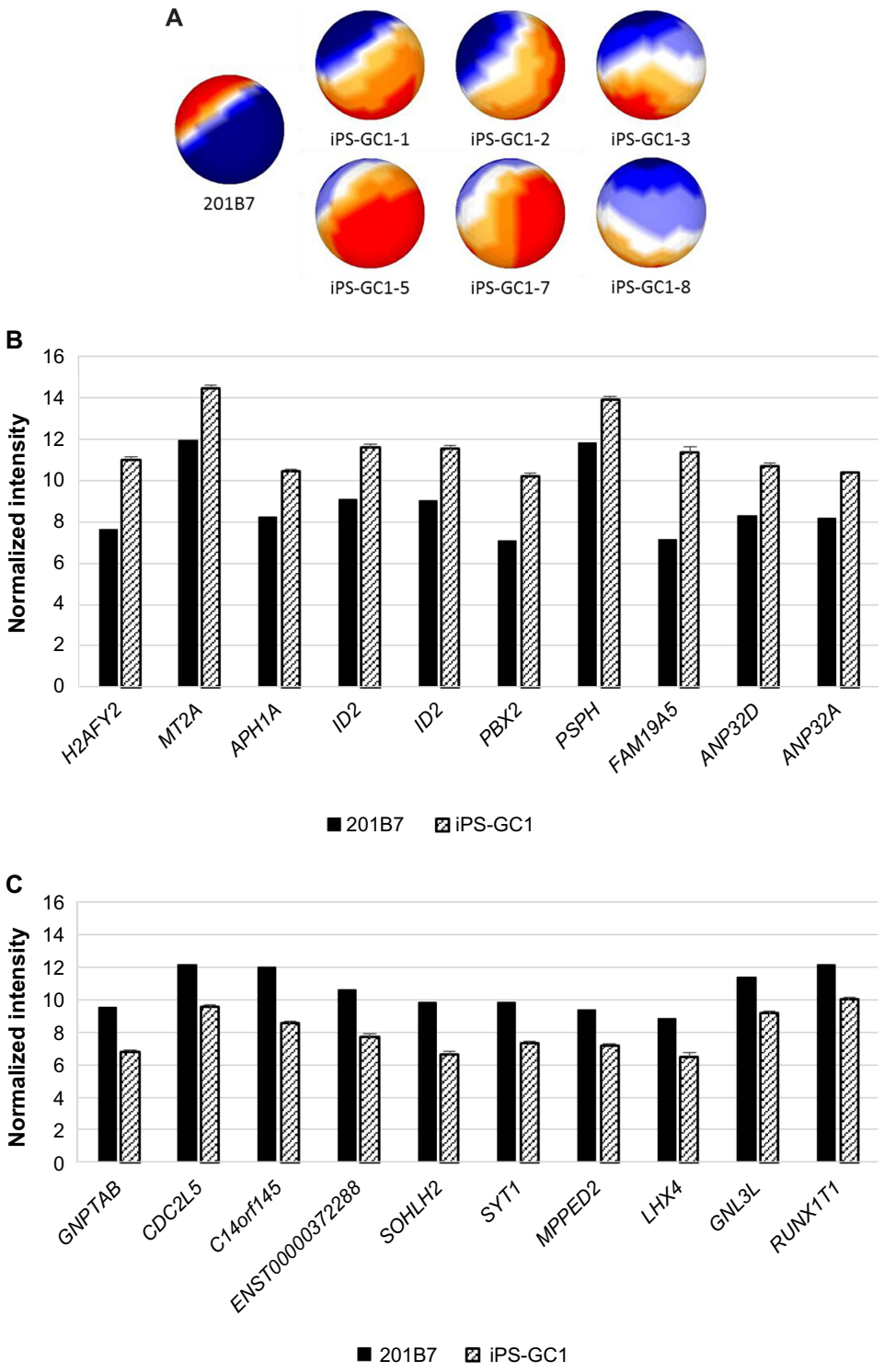

Figure 4. Mapping and comparison of normal hiPSC and iPS-GC1 with sSOM.
Notes: (**A**) Gene expression patterns analyzed by sSOM with the microarray data of 201B7 (GSM251846) and iPS-GC1. The data were used 439 genes, which were extracted by the two parameters (see Fig. 1). Each of iPS-GC1 was mapped as a sphere by sSOM analysis. The normalized intensities of 328 upregulating genes (**B**) or 111 downregulating genes (**C**) of iPS-GC1, which were compared to GSM241846, were analyzed by sSOM, and 10 genes close to the IP were aligned by the order of NSD as listed in Tables 4 and 5. Graphs are depicted as mean + SD. Normalized intensity *i'* was shown in base-2 logarithm on Y-axis. Y-linked genes were eliminated from the list because sex differences were confounding factor.

widely been used to extract useful information from expression profiles. Compared with hierarchical clustering, SOM has a number of features well suited to cluster genes by their expression patterns. It also has good computational properties and is easy to run and fast.[22,23] A conventional plane SOM (2D SOM) has not yet common in gene clustering procedure.

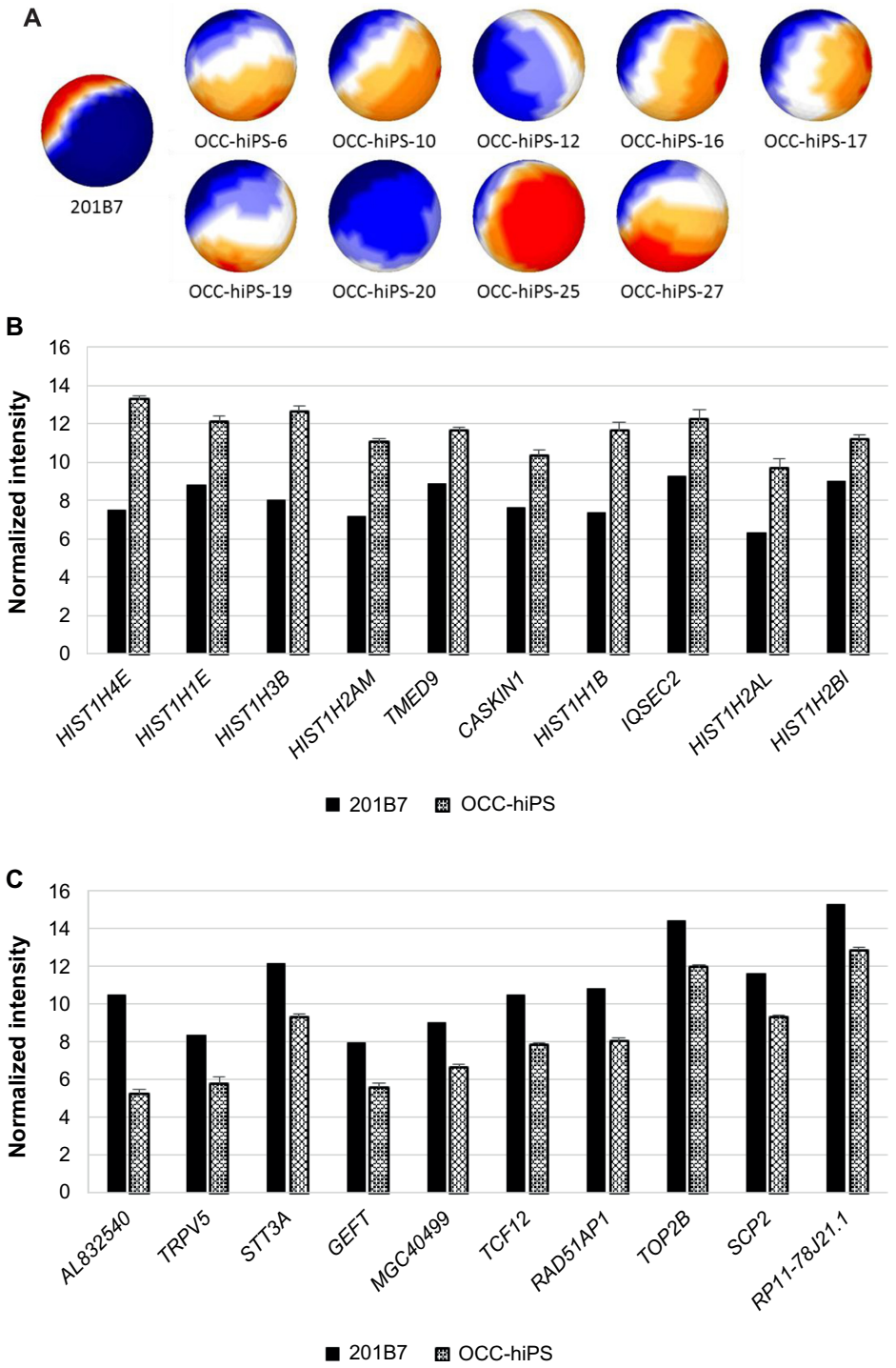

Figure 5. Mapping and comparison of normal hiPSC and OCC-hiPS with sSOM.
Notes: (**A**) Gene expression patterns analyzed by sSOM with the microarray data of 201B7 (GSM241846) and OCC-hiPS. The data were used 402 genes, which was extracted by the two parameters (see Fig. 1). Each of OCC-hiPS was mapped as a sphere by sSOM analysis. The normalized intensities of 255 upregulating genes (**B**) or 147 downregulating genes (**C**) of OCC-hiPS, which were compared to GSM241826, were analyzed by sSOM. Ten genes close to the IP were aligned by the order of NSD as listed in Tables 6 and 7. Graphs were depicted as mean + SD. Normalized intensity *i'* was shown in base-2 logarithm on Y-axis.

The reason might be that the grid units at the boundary of the 2D SOM result have fewer neighbors than the units inside the map, which often cause the *border effect* – the weight vectors of these units *collapse to the center of the input space*.[24] To solve this, sSOM is suitable for data with underlying directional structures. sSOM has been shown effective to remove the border effect and is useful to convey the information of distance and direction with running speed comparable to the conventional 2D SOM.[25,26] Since we have successfully applied sSOM for the analytical procedure of microarray data,[17,19,20,27]

Table 6. Top 10 upregulating genes of OCC-hiPS compared with hiPSC 201B7.

NSD	GENE	DESCRIPTION	ACCESSION NO.
0.06	*HIST1H4E*	Histone 1, H4e, mRNA	NM_003545
0.33	*HIST1H1E*	Histone 1, H1e, mRNA	NM_005321
0.64	*HIST1H3B*	Histone 1, H3b, mRNA	NM_003537
0.64	*HIST1H2AM*	Histone 1, H2am, mRNA	NM_003514
1.01	*TMED9*	Transmembrane emp24 protein transport domain containing 9, mRNA	NM_017510
1.05	*CASKIN1*	CASK interacting protein 1, mRNA	NM_020764
1.31	*HIST1H1B*	Histone 1, H1b, mRNA	NM_005322
1.34	*IQSEC2*	IQ motif and Sec7 domain 2, mRNA	NM_015075
1.41	*HIST1H2AL*	Histone 1, H2al, mRNA	NM_003511
1.46	*HIST1H2BI*	Histone 1, H2bi, mRNA	NM_003525

Abbreviation: NSD, nonsignificant distance.

Table 7. Top 10 downregulating genes of OCC-hiPS compared with hiPSC 201B7.

NSD	GENE	DESCRIPTION	ACCESSION NO.
0.36	*AL832540*	mRNA; cDNA DKFZp547A0117 (from clone DKFZp547A0117)	AL832540
1.18	*TRPV5*	Transient receptor potential cation channel, subfamily V, member 5, mRNA	NM_019841
1.29	*STT3A*	STT3, subunit of the oligosaccharyltransferase complex, homolog A (S. cerevisiae), mRNA	NM_152713
1.34	*GEFT*	RAC/CDC42 exchange factor, transcript variant 2, mRNA	NM_133483
2.58	*MGC40499*	PRotein Associated with Tlr4, mRNA	NM_152755
2.59	*TCF12*	Transcription factor 12 (HTF4, helix-loop-helix transcription factors 4), transcript variant 4, mRNA	NM_207038
2.61	*RAD51AP1*	RAD51 associated protein 1, mRNA	NM_006479
2.61	*TOP2B*	Topoisomerase (DNA) II beta 180kDa, mRNA	NM_001068
2.61	*SCP2*	Sterol carrier protein 2, transcript variant 1, mRNA	NM_002979
2.63	*RP11-78J21.1*	Heterogeneous nuclear ribonucleoprotein A1-like (LOC144983), transcript variant 1, mRNA	NM_001011724

Abbreviation: NSD, nonsignificant distance.

we employed sSOM for the data analysis in this study. By comparing the sSOM patterns of CSCs with that of normal hiPSC, we successfully demonstrated a simple and easy way to screen for the candidate genes commonly and specifically expressed among all the CSCs. One reason for this success would be due to the feature scaling to 0 to 1 (*I*). Without this scaling, the result of sSOM sphere is shown in Supplementary Figure 5. It would be difficult to distinguish from each other in this figure. There are no enough data for the copy number changes or chromosomal abnormalities in original cells. It is needed to study the relationship between our CSCs and these problems.

Normalization of data is another important issue in data mining. Standard protocols for the normalization to make various dataset comparable should be available but not always available even it is often necessary to search relations between data. If the normalizing process was skipped or inadequate, analyses would result in nothing or false. Since all microarray data were obtained by Agilent microarray system in this study, Bioconductor package *agilp*, which was developed by Thomas et al to normalize the microarray data, was employed.[28] They clearly showed a relationship between T-cell population and T-cell signature score obtained from various microarray data provided by independent groups.[28]

Through the data mining procedure described earlier, *FAM19A5* was significantly upregulated in iPS-CC1, *MT2A* in GC1-iPS, and *TMED9* in all the CSCs including OCC-hiPS. *FAM19A5* has been reported as cholangiocarcinoma marker by protein analysis.[29] *MT2A* was expressed in a subgroup of patients with acute myelomonocytic leukemia[30] and also recently identified as a gastric cancer-related gene.[31] Although these independent studies support our results, little information related to cancer is known about *TMED9*. *TMED9* is also called as $p24\alpha_2$ or p25 and a transporter protein expressed on endoplasmic reticulum membrane. It had found to maintain endoplasmic reticulum exit sites and

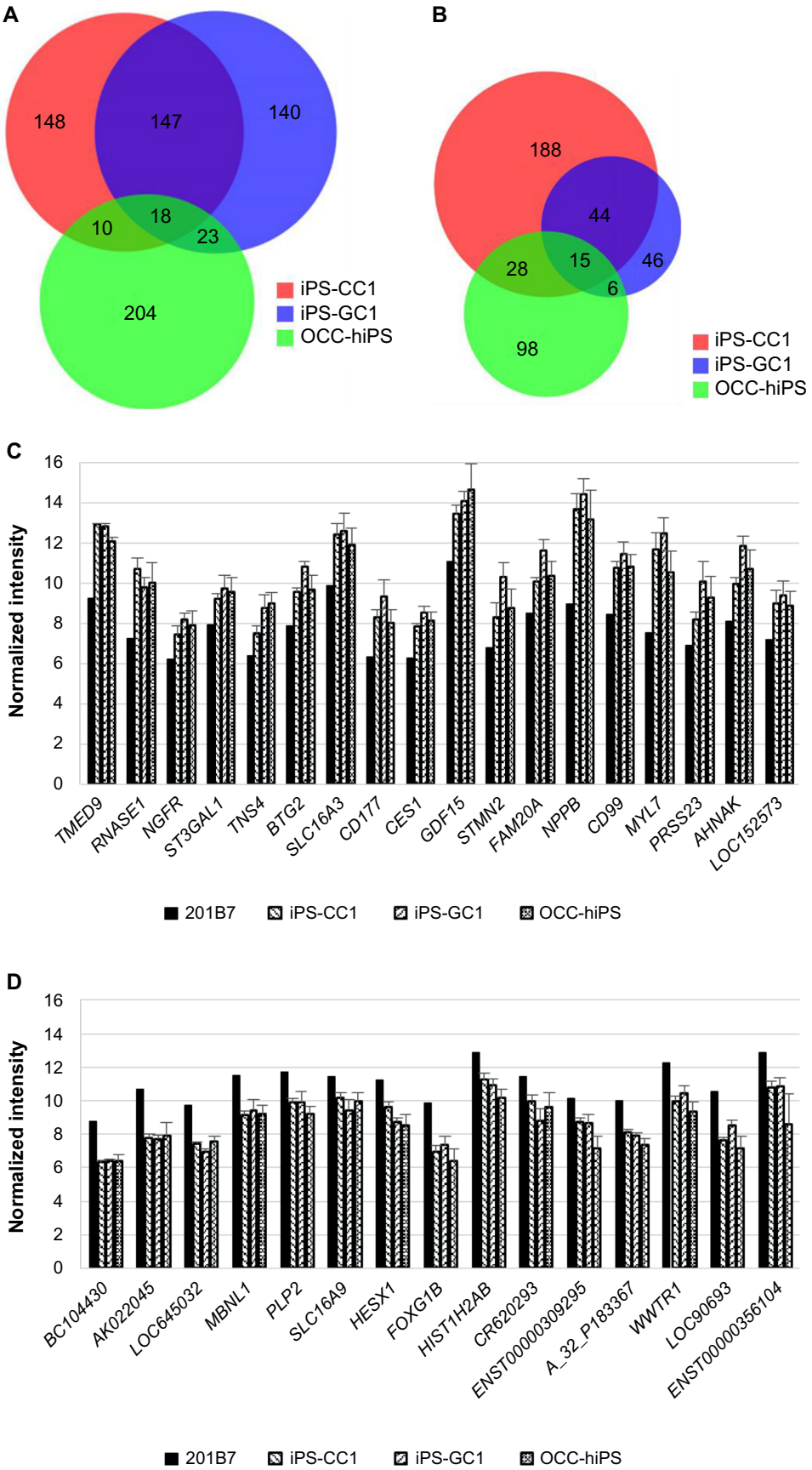

Figure 6. Venn diagrams of upregulating genes or downregulating genes extracted by each sSOM analysis of iPS-CC1, iPS-GC1, and OCC-hiPS.
Notes: Venn diagrams of upregulating genes (**A**) or downregulating genes (**B**) describing in the legends of Figures 2–4. The numbers of genes were in each area. Each of the upregulating (**C**) or downregulating (**D**) genes commonly among all the CSCs was shown left to right in the order of NSD from the IP. Graphs were depicted as mean + SD. Normalized intensity *i'* was shown in base-2 logarithm on Y-axis.

Table 8. Common upregulating genes of all the CSCs compared with hiPSC 201B7.

NSD	GENE	DESCRIPTION	ACCESSION NO.
0.49	*TMED9*	Transmembrane emp24 protein transport domain containing 9, mRNA	NM_017510
3.30	*RNASE1*	Ribonuclease, RNase A family, 1 (pancreatic), transcript variant 3, mRNA	NM_198232
5.34	*NGFR*	Nerve growth factor receptor (TNFR superfamily, member 16), mRNA	NM_002507
5.43	*ST3GAL1*	ST3 beta-galactoside alpha-2,3-sialyltransferase 1, transcript variant 1, mRNA	NM_003033
5.44	*TNS4*	Tensin 4, mRNA	NM_032865
5.46	*BTG2*	BTG family, member 2, mRNA	NM_006763
5.52	*SLC16A3*	Solute carrier family 16 (monocarboxylic acid transporters), member 3, transcript variant 2, mRNA	NM_004207
5.56	*CD177*	mRNA for NB1 glycoprotein (NB1 gene), negative phenotype #1	AJ310433
5.61	*CES1*	Carboxylesterase 1 (monocyte/macrophage serine esterase 1) transcript variant 3, mRNA	NM_001266
5.65	*GDF15*	Growth differentiation factor 15, mRNA	NM_004864
5.68	*STMN2*	Stathmin-like 2, mRNA	NM_007029
5.77	*FAM20A*	Family with sequence similarity 20, member A, mRNA	NM_017565
5.82	*NPPB*	Natriuretic peptide precursor B, mRNA	NM_002521
5.84	*CD99*	CD99 molecule, mRNA	NM_002414
5.85	*MYL7*	Myosin, light polypeptide 7, regulatory, mRNA	NM_021223
5.87	*PRSS23*	Protease, serine, 23, mRNA	NM_007173
5.88	*AHNAK*	AHNAK nucleoprotein (desmoyokin), transcript variant 1, mRNA	NM_001620
5.99	*LOC152573*	Clone IMAGE:4477067, mRNA, partial cds.	BC012029

Abbreviation: NSD, nonsignificant distance.

Table 9. Common downregulating genes of all the CSCs compared with hiPSC 201B7.

NSD	GENE	DESCRIPTION	ACCESSION NO.
0.47	*BC104430*	cDNA clone IMAGE: 40021976	BC104430
1.25	*AK022045*	cDNA FLJ11983 fis, clone HEMBB1001337	AK022045
1.63	*LOC645032*	PREDICTED: hypothetical protein LOC645032, mRNA	XM_928089
3.23	*MBNL1*	Muscleblind-like (Drosophila), transcript variant 1, mRNA	NM_021038
4.58	*PLP2*	Proteolipid protein 2 (colonic epithelium-enriched), mRNA	NM_002668
4.58	*SLC16A9*	Solute carrier family 16 (monocarboxylic acid transporters), member 9, mRNA	NM_194298
4.59	*HESX1*	Homeobox, ES cell expressed 1, mRNA	NM_003865
4.68	*FOXG1B*	Forkhead box G1B, mRNA	NM_005249
4.72	*HIST1H2AB*	Histone 1, H2ab, mRNA	NM_003513
4.74	*CR620293*	Full-length cDNA clone CS0DF028YD24 of Fetal brain of (human)	CR620293
4.85	*ENST00000309295*	mRNA; cDNA DKFZp762C186 (from clone DKFZp762C186)	AL834433
4.86	*A_32_P183367*	Unknown	–
4.88	*WWTR1*	WW domain containing transcription regulator 1, mRNA	NM_015472
4.88	*LOC90693*	LOC90693 protein, mRNA	NM_138771
4.94	*ENST00000356104*	Unknown	–

Abbreviation: NSD, nonsignificant distance.

vesicular-tubular clusters[32] and has special domain to form membrane fidelity.[33]

Although NSD of *NPPB* is much larger than that of *TMED9*, *NPPB* reasonably appears to be upregulated with much difference between CSCs and hiPSC 201B7 in the average value (Table 8 and Fig. 5A and C). Actually, NPPB is reported as a biomarker for a cancer-associated fibroblast in ovarian cancer,[34] and these cancer-associated fibroblasts are thought as *feeder cells* of tumor including CSCs.[35,36] The feeder cells might be related with the progeny of CSCs

and supporting the self-renewal of CSCs.[37] Similarly with NPPB, *MYL7*, which encodes Atrial Light Chain-2, is also listed as one of the 18 genes (Table 8). This gene is considered to be related to cell stemness rather than tumorigenesis.[38,39] RNASE1 is a member of ribonuclease family and cleaves phosphodiester double-stranded RNA bonds. This protects its host against viruses. Moreover, it is also known that glycosylated Asn[88] on this molecule is correlated with the pancreatic cancer.[40] NGFR is a one of the growth factor receptors that binds to neurotrophins. This receptor was referred as breast cancer marker before.[41] ST3GAL1 is a type II membrane protein that catalyzes sialylation. Chong et al reported that it has a critical role to sustain glioblastoma growth.[42] TNS4 is an adhesion protein mediating integrin. This protein is upregulated by ERK1/2 and enhance cancer cell migration.[43] *SLC16A3* encodes MCT4 that mediates lactate transportation[44] and whose expression is upregulated in clear cell renal cell carcinoma.[45] CD177 is known as a neutrophil-specific antigen and could be a gastric cancer marker.[46] GDF15 is a member of TGF-β superfamily and contributes to host defense from injury or disease.[47] The expression of GDF15 is upregulated by various stimuli and, because of this, it could be a biomarker of many cancers.[48] STMN2 was first identified as a neuron-specific, developmentally regulated protein.[49] This gene is upregulated in hepatoma cells and might play a critical role in β-catenin/TCF-mediated carcinogenesis.[50] CD99 is expressed on leukocytes and helps T-cell adhesion.[51] Among patients with diffuse large B-cell lymphoma, two-year event-free survival gets worse when CD99 is positive in germinal center B-cells.[52] PRSS23 is a serine protease and coexpressed with estrogen receptor α, which is a biomarker for human breast cancer.[53] Chan et al found that PRSS23 might be critical for estrogen-induced cell proliferation of estrogen receptor α-positive breast cancer cells.[54] CES1 is a serine esterase and involved in the activation of prodrugs like angiotensin-converting enzyme inhibitors.[55] This gene is highly expressed in human colorectal cancers.[56] These molecules are thought to be positively correlated with cancer. On the other hand, AHNAK is expressed in various cell types and involved in many cellular processes such as calcium regulation and actin organization.[57–59] *Ahnak*$^{-/-}$ mouse showed progressed hyperplasia of mammary glands, and their expression was low in human breast cancer tissues than that in controls.[60] BTG2 regulates cell cycle in a p53-dependent way.[61] Some studies have shown that BTG2 expression is downregulated in cancer tissues.[62–64] These molecules could be tumor suppression marker. There is little information about FAM20A that has a functional locus in hematopoiesis[65] and LOC152573.

Collectively, iPSC technology and gene ontology were embodied that the sSOM analysis could depict the gene signature of the CSCs and list up the marker genes, although further biological study would be needed for the relationships between the nominated genes and the CSCs.

Conclusion

We newly developed artificial CSCs commonly expressing hESC/hiPSC-enriched genes at a level equivalent to those of typical hiPSCs (201B7). The unsupervised method of the sSOM analysis demonstrated that the CSCs could be divided into distinct groups due to their culture conditions and original cancer tissues. Furthermore, the supervised method of the SOM analysis suggested the gene signature and the marker genes of the CSCs.

Acknowledgment

The authors thank Dr. Heizo Tokutaka for developing sSOM software Blossom, Dr. Toshio Kitamura for Plat-GP packaging cells, Japanese Collection of Research Bioresources for providing hiPSC 201B7, and National Institutes of Biomedical Innovation, Health and Nutrition for human cancer tissues. The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus[66] and are accessible through GEO Series accession number GSE83883 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83883).

Author Contributions

Conceived and designed the experiments: AS, TI, MI, MS. Analyzed the data: AS, TI. Wrote the first draft of the article: AS. Contributed to the writing of the article: AS, TI, AV, TK, MS. Agreed with the article results and conclusions: AS, TI, TK, AV, MS. Jointly developed the structure and arguments for the article: AS, TI, TK, MI, AV, JM, AM, HM, MS. Made critical revisions and approved the final version: AS, TI, MS. All authors reviewed and approved the final article.

Supplementary Material

Supplementary Figure 1. Normalized intensities of (**A**) hESC/hiPSC-enriched genes and (**B**) housekeeping genes. Each rectangular shows min to "min + (average+2SD)" which we used as threshold in Figure 1C. Normalized intensity *i*' was shown in base-2 logarithm on Y-axis.

Supplementary Figure 2. Backside of spheres shown in Figure 2A.

Supplementary Figure 3. Comparison of the normalized intensities of top 10 genes including Y-linked genes nearest to the IP in iPS-CC1. These genes were analyzed by sSOM with IP and aligned by the order of NSD. Graphs were depicted as mean + SD. Normalized intensity *i*' was shown in base-2 logarithm on Y-axis.

Supplementary Figure 4. Comparison of the normalized intensities of top 10 genes including Y-linked genes nearest to the IP in iPS-GC1. These genes were analyzed by sSOM with IP and aligned by the order of NSD. Graphs were depicted as mean + SD. Normalized intensity *i*' was shown in base-2 logarithm on Y-axis.

Supplementary Figure 5. sSOM results without feature scaling. These spheres were sSOM-calculated with normalized intensity (*i*').

Supplementary Table 1. Top 10 upregulating genes including Y-linked genes in hiPS-CC1 compared with hiPSC 201B7. Graphs of these genes were shown in Supplementary Figure 3.

Supplementary Table 2. Top 10 upregulating genes including Y-linked genes in hiPS-GC1 compared with hiPSC 201B7. Graphs of these genes were shown in Supplementary Figure 4.

Note: *NSD: non-significant distance.

REFERENCES

1. Bonnet D, Dick JE. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med*. 1997;3(7):730–7.
2. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A*. 2003;100(7):3983–8.
3. Singh SK, Clarke ID, Terasaki M, et al. Identification of a cancer stem cell in human brain tumors. *Cancer Res*. 2003;63(18):5821–8.
4. Richardson GD, Robson CN, Lang SH, Neal DE, Maitland NJ, Collins AT. CD133, a novel marker for human prostatic epithelial stem cells. *J Cell Sci*. 2004;117(pt 16):3539–45.
5. Xin L, Lawson DA, Witte ON. The Sca-1 cell surface marker enriches for a prostate-regenerating cell subpopulation that can initiate prostate tumorigenesis. *Proc Natl Acad Sci U S A*. 2005;102(19):6942–7.
6. Ricci-Vitiani L, Lombardi DG, Pilozzi E, et al. Identification and expansion of human colon-cancer-initiating cells. *Nature*. 2007;445(7123):111–5.
7. Prince ME, Sivanandan R, Kaczorowski A, et al. Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma. *Proc Natl Acad Sci U S A*. 2007;104(3):973–8.
8. Li C, Heidt DG, Dalerba P, et al. Identification of pancreatic cancer stem cells. *Cancer Res*. 2007;67(3):1030–7.
9. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*. 2006;126(4):663–76.
10. Chen L, Kasai T, Li Y, et al. A model of cancer stem cells derived from mouse induced pluripotent stem cells. *PLoS One*. 2012;7(4):e33544.
11. Ishikawa T, Kobayashi M, Yanagi S, et al. Human induced hepatic lineage-oriented stem cells: autonomous specification of human iPS cells toward hepatocyte-like cells without any exogenous differentiation factors. *PLoS One*. 2015;10(4):e0123193.
12. Masaki H, Ishikawa T, Takahashi S, et al. Heterogeneity of pluripotent marker gene expression in colonies generated in human iPS cell induction culture. *Stem Cell Res*. 2007;1(2):105–15.
13. Takahashi K, Tanabe K, Ohnuki M, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*. 2007;131(5):861–72.
14. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80.
15. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015;12(2):115–21.
16. Chain B. agilp: Agilent expression array processing package, R package version 3.2.0. 2012.
17. Sugii Y, Kasai T, Ikeda M, et al. A unique procedure to identify cell surface markers through a spherical self-organizing map applied to DNA microarray analysis. *Biomark Cancer*. 2016;8:17–23.
18. Ikeda M, Kumon K, Omoto K, et al. Spherical self-organizing map detects MYBL 1 as candidate gene for triple-negative breast cancer. *Neurosci Biomed Eng*. 2015;3(2):94–101.
19. Tuoya, Sugii Y, Satoh H, et al. Spherical self-organizing map as a helpful tool to identify category-specific cell surface markers. *Biochem Biophys Res Commun*. 2008;376(2):414–8.
20. Yuh S, Takayuki K, Takayuki O, Masashi I, Heizo T, Masaharu S. Clustering genes, tissues, cells and bioactive chemicals by sphere SOM. In: Igadwa MJ, ed. *Self Organizing Maps – Applications and Novel Algorithm Design*. Rijeka: InTech; 2011:371–86.
21. Yan T, Mizutani A, Chen L, et al. Characterization of cancer stem-like cells derived from mouse induced pluripotent stem cells transformed by tumor-derived extracellular vesicles. *J Cancer*. 2014;5(7):572–84.
22. Wang J, Delabie J, Aasheim H, Smeland E, Myklebost O. Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC Bioinformatics*. 2002;3:36.
23. Kohonen T. Essentials of the self-organizing map. *Neural Netw*. 2013;37:52–65.
24. Sarle WS. *Neural Network FAQ*. SAS Institute; 2002. Available at: ftp://ftp.sas.com/pub/neural/FAQ.html.
25. Sangole A, Knopf GK. Visualization of randomly ordered numeric data sets using spherical self-organizing feature maps. *Comput Graph*. 2003;27(6):963–76.
26. Wu Y, Takatsuka M. Spherical self-organizing map using efficient indexed geodesic data structure. *Neural Netw*. 2006;19(6):900–10.
27. Samah AS, Yuh S, Tuoya, et al. Identification of TM9SF2 as a candidate of the cell surface marker common to breast carcinoma cells. *Chin J Clin Oncol*. 2009;6:1.
28. Thomas N, Heather J, Pollara G, et al. The immune system as a biomonitor: explorations in innate and adaptive immunity. *Interface Focus*. 2013;3(2):20120099.
29. Janvilisri T, Leelawat K, Roytrakul S, Paemanee A, Tohtong R. Novel serum biomarkers to differentiate cholangiocarcinoma from benign biliary tract diseases using a proteomic approach. *Dis Markers*. 2015;2015:105358.
30. Le Beau MM, Diaz MO, Karin M, Rowley JD. Metallothionein gene cluster is split by chromosome 16 rearrangements in myelomonocytic leukaemia. *Nature*. 1985;313(6004):709–11.
31. Kim JM, Sohn HY, Yoon SY, et al. Identification of gastric cancer-related genes using a cDNA microarray containing novel expressed sequence tags expressed in gastric cancer cells. *Clin Cancer Res*. 2005;11(2 pt 1):473–82.
32. Lavoie C, Paiement J, Dominguez M, et al. Roles for alpha(2)p24 and COPI in endoplasmic reticulum cargo exit site formation. *J Cell Biol*. 1999;146(2):285–99.
33. Emery G, Parton RG, Rojo M, Gruenberg J. The trans-membrane protein p25 forms highly specialized domains that regulate membrane composition and dynamics. *J Cell Sci*. 2003;116(pt 23):4821–32.
34. Lawrenson K, Grun B, Lee N, et al. NPPB is a novel candidate biomarker expressed by cancer-associated fibroblasts in epithelial ovarian cancer. *Int J Cancer*. 2015;136(6):1390–401.
35. Karagiannis GS, Poutahidis T, Erdman SE, Kirsch R, Riddell RH, Diamandis EP. Cancer-associated fibroblasts drive the progression of metastasis through both paracrine and mechanical pressure on cancer tissue. *Mol Cancer Res*. 2012;10(11):1403–18.
36. Chen WJ, Ho CC, Chang YL, et al. Cancer-associated fibroblasts regulate the plasticity of lung cancer stemness via paracrine signalling. *Nat Commun*. 2014;5:3472.
37. Matsuda S, Yan T, Mizutani A, et al. Cancer stem cells maintain a hierarchy of differentiation by creating their niche. *Int J Cancer*. 2014;135(1):27–36.
38. Hernández D, Millard R, Sivakumaran P, et al. Electrical stimulation promotes cardiac differentiation of human induced pluripotent stem cells. *Stem Cells Int*. 2016;2016:1718041.
39. Maves L, Tyler A, Moens CB, Tapscott SJ. Pbx acts with Hand2 in early myocardial differentiation. *Dev Biol*. 2009;333(2):409–18.
40. Nakata D. Increased N-glycosylation of Asn^{88} in serum pancreatic ribonuclease 1 is a novel diagnostic marker for pancreatic cancer. *Sci Rep*. 2014;4:6715.
41. Tsang JY, Wong KH, Lai MW, et al. Nerve growth factor receptor (NGFR): a potential marker for specific molecular subtypes of breast cancer. *J Clin Pathol*. 2013;66(4):291–6.
42. Chong YK, Sandanaraj E, Koh LW, et al. ST3GAL1-associated transcriptomic program in glioblastoma tumor growth, invasion, and prognosis. *J Natl Cancer Inst*. 2016;108(2).
43. Chan LK, Chiu YT, Sze KM, Ng IO. Tensin4 is up-regulated by EGF-induced ERK1/2 activity and promotes cell proliferation and migration in hepatocellular carcinoma. *Oncotarget*. 2015;6(25):20964–76.
44. Price NT, Jackson VN, Halestrap AP. Cloning and sequencing of four new mammalian monocarboxylate transporter (MCT) homologues confirms the existence of a transporter family with an ancient past. *Biochem J*. 1998;329(pt 2):321–8.
45. Fisel P, Kruck S, Winter S, et al. DNA methylation of the SLC16 A3 promoter regulates expression of the human lactate transporter MCT4 in renal cancer with consequences for clinical outcome. *Clin Cancer Res*. 2013;19(18):5170–81.
46. Toyoda T, Tsukamoto T, Yamamoto M, et al. Gene expression analysis of a Helicobacter pylori-infected and high-salt diet-treated mouse gastric tumor model: identification of CD177 as a novel prognostic factor in patients with gastric cancer. *BMC Gastroenterol*. 2013;13:122.
47. Hsiao EC, Koniaris LG, Zimmers-Koniaris T, Sebald SM, Huynh TV, Lee SJ. Characterization of growth-differentiation factor 15, a transforming growth factor beta superfamily member induced following liver injury. *Mol Cell Biol*. 2000;20(10):3742–51.
48. Vaňhara P, Hampl A, Kozubík A, Souček K. Growth/differentiation factor-15: prostate cancer suppressor or promoter? *Prostate Cancer Prostatic Dis*. 2012;15(4):320–8.
49. Antonsson B, Lütjens R, Di Paolo G, et al. Purification, characterization, and in vitro phosphorylation of the neuron-specific membrane-associated protein SCG10. *Protein Expr Purif*. 1997;9(3):363–71.
50. Lee HS, Lee DC, Park MH, et al. STMN2 is a novel target of beta-catenin/TCF-mediated transcription in human hepatoma cells. *Biochem Biophys Res Commun*. 2006;345(3):1059–67.

51. Gelin C, Aubrit F, Phalipon A, et al. The E2 antigen, a 32 kd glycoprotein involved in T-cell adhesion processes, is the MIC2 gene product. *EMBO J*. 1989;8(11):3253–9.
52. Hong J, Park S, Park J, et al. CD99 expression and newly diagnosed diffuse large B-cell lymphoma treated with rituximab-CHOP immunochemotherapy. *Ann Hematol*. 2012;91(12):1897–906.
53. Sommer S, Fuqua SA. Estrogen receptor and breast cancer. *Semin Cancer Biol*. 2001;11(5):339–52.
54. Chan HS, Chang SJ, Wang TY, et al. Serine protease PRSS23 is upregulated by estrogen receptor α and associated with proliferation of breast cancer cells. *PLoS One*. 2012;7(1):e30397.
55. Thomsen R, Rasmussen HB, Linnet K, Consortium I. In vitro drug metabolism by human carboxylesterase 1: focus on angiotensin-converting enzyme inhibitors. *Drug Metab Dispos*. 2014;42(1):126–33.
56. Amissah F, Duverna R, Aguilar BJ, Poku RA, Lamango NS. Polyisoprenylated methylated protein methyl esterase is both sensitive to curcumin and overexpressed in colorectal cancer: implications for chemoprevention and treatment. *Biomed Res Int*. 2013;2013:416534.
57. Haase H, Podzuweit T, Lutsch G, et al. Signaling from beta-adrenoceptor to L-type calcium channel: identification of a novel cardiac protein kinase A target possessing similarities to AHNAK. *FASEB J*. 1999;13(15):2161–72.
58. Gentil BJ, Delphin C, Mbele GO, et al. The giant protein AHNAK is a specific target for the calcium- and zinc-binding S100B protein: potential implications for Ca2+ homeostasis regulation by S100B. *J Biol Chem*. 2001;276(26):23253–61.
59. Hohaus A, Person V, Behlke J, Schaper J, Morano I, Haase H. The carboxyl-terminal region of ahnak provides a link between cardiac L-type Ca2+ channels and the actin-based cytoskeleton. *FASEB J*. 2002;16(10):1205–16.
60. Lee IH, Sohn M, Lim HJ, et al. Ahnak functions as a tumor suppressor via modulation of TGFβ/Smad signaling pathway. *Oncogene*. 2014;33(38):4675–84.
61. Rouault JP, Falette N, Guéhenneux F, et al. Identification of BTG2, an antiproliferative p53-dependent component of the DNA damage cellular response pathway. *Nat Genet*. 1996;14(4):482–6.
62. Liu M, Wu H, Liu T, et al. Regulation of the cell cycle gene, BTG2, by miR-21 in human laryngeal carcinoma. *Cell Res*. 2009;19(7):828–37.
63. Struckmann K, Schraml P, Simon R, et al. Impaired expression of the cell cycle regulator BTG2 is common in clear cell renal cell carcinoma. *Cancer Res*. 2004;64(5):1632–8.
64. Takahashi F, Chiba N, Tajima K, et al. Breast tumor progression induced by loss of BTG2 expression is inhibited by targeted therapy with the ErbB/HER inhibitor lapatinib. *Oncogene*. 2011;30(27):3084–95.
65. Nalbant D, Youn H, Nalbant SI, et al. FAM20: an evolutionarily conserved family of secreted proteins expressed in hematopoietic cells. *BMC Genomics*. 2005;6:11.
66. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002; 30(1):207–10.