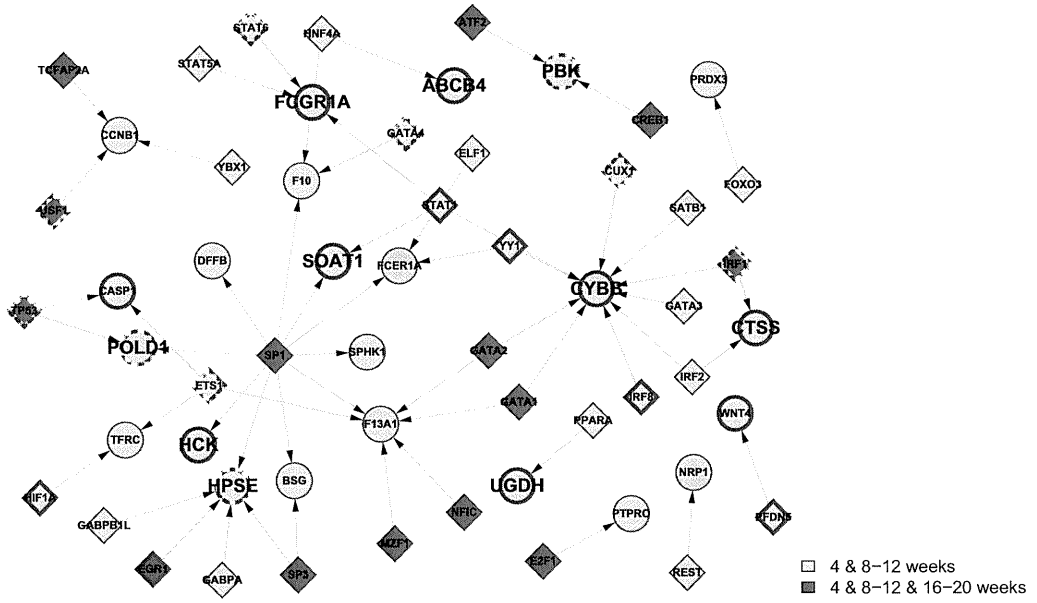


Figure 5 TF-regulatory network in 4w GK. In total, 12 genes were found to be differently expressed between GK and WKY. The nodes are colored by the same way as Figure 1. The node form depends on the expression degree between GK and WKY: bold node means genes that are overexpressed comparing GK to WKY, while dash-line node means genes which expression levels are lower in GK than that in WKY. At the beginning of hyperglycemia, inflammation, bile metabolism dysfunction, decreased proliferation and increased apotosis already show in GK rats liver at 4 weeks of age.

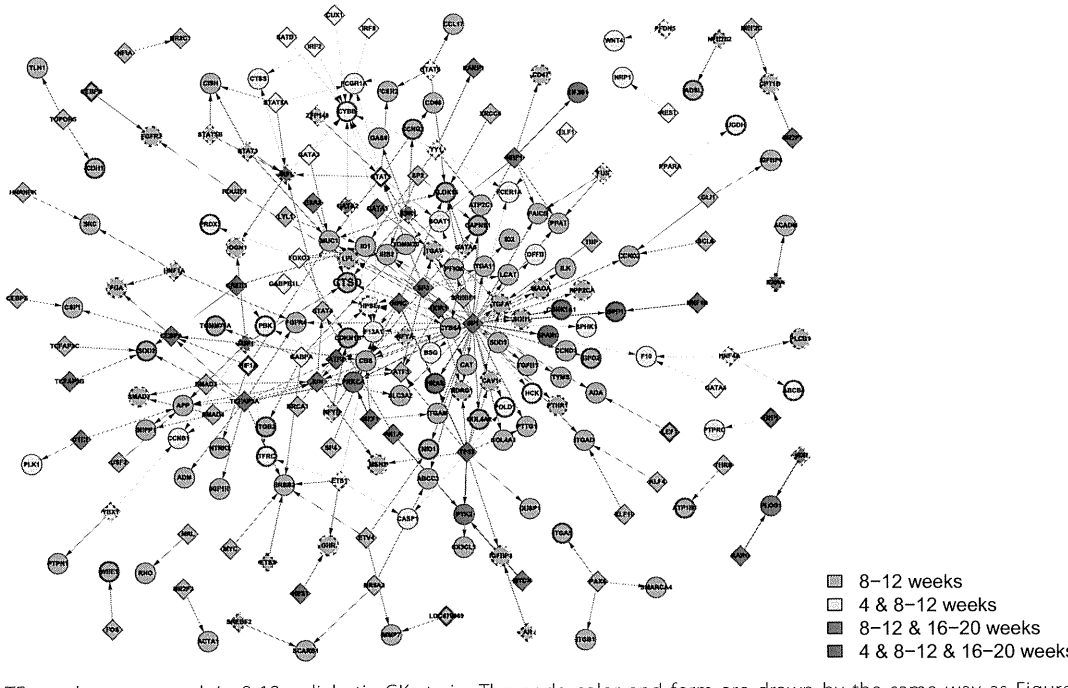


Figure 6 TF regulatory network in 8-12w diabetic GK strain. The node color and form are drawn by the same way as Figures 1 and 5, respectively. Pathways making the chaos directing to the diabetes and networks compensate are both active.

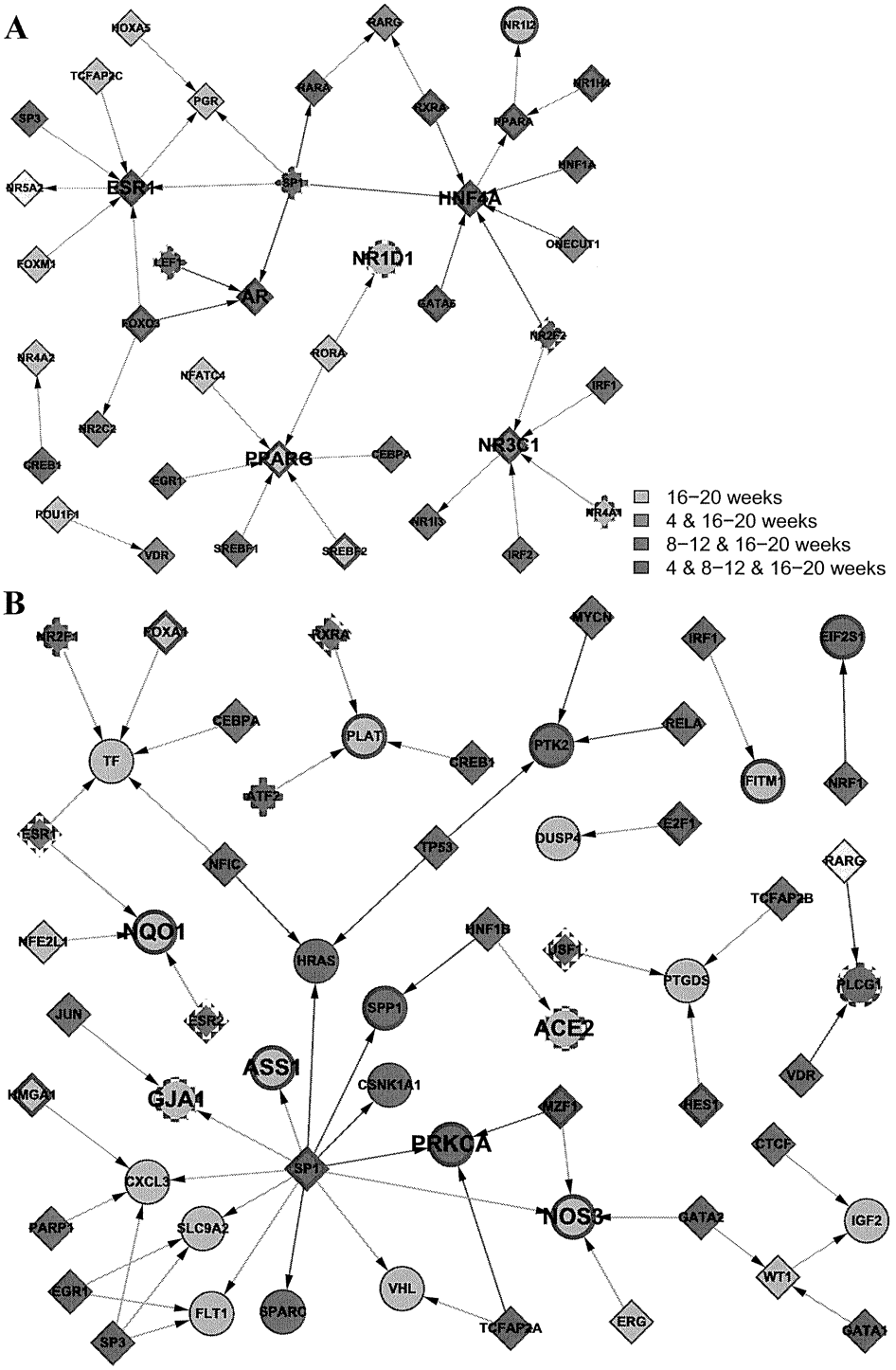


Figure 7 TF-regulatory gene expression networks in 16-20w WKY (A) and GK (B). The node color and form are drawn by the same way as Figures 1 and 5, respectively. Nuclear receptors play important role maintaining the non-diabetic stage in WKY strain. In GK rats, some compensational pathways still exist. However, genes involved in insulin resistance, hypertension and apoptosis are able to cause diabetes progression.

divided into 4 functional groups: immune, metabolism, proliferation and apotosis.

F13A1, CYBB, FCGR1A, HCK, CTSS are involved in inflammation and their expression levels are exclusively increased in GK at 4 weeks of age. Previously we have talked about overexpression of CYBB and FCGR1A inducing inflammation. Although F13A1 is related to thrombosis, it is also been recognized as an inflammation-related gene. Tyrosine-protein kinase (HCK) is an enzyme predominantly expressed in hemopoietic cell types. Overexpression of HCK contributes to inflammation by promoting neutrophil migration and degranulation as well as couple the Fc receptor to the activation of the respiratory burst [32]. Cathepsin S (CTSS) encodes a lysosomal protease that participates in macrophage activation by the degradation of antigens to peptides for presentation [33].

Metabolism group includes higher expression of UGDH, ABCB4, and SOAT1 genes in GK. UDP-glucose 6-dehydrogenase (UGDH) converting UDP-glucose to UDP-glucuronate is significantly increased in DM. The enhanced expression of UGDH is due to excess glucose load. Multidrug resistance protein 3 is a protein that encoded by ABCB4 gene, which transports phospholipids from hepatocytes into bile. Overexpression is associated with progressive familial intrahepatic cholestasis type 3. Sterol O-acyltransferase 1 (SOAT1), also known as acyl-Coenzyme A: cholesterol acyltransferase, forms cholesterol esters from cholesterol located in the endoplasmic reticulum. ABCB4 and SOAT1 are reported coexpressed in gallbladder tissue and participate in bile metabolism [34]. Overexpression of SOAT1 functions to atherosclerosis and accumulates cholesterol in the gallbladder mucosa. Recent studies show that bile metabolism is in close contact with occurrence of T2DM. Disturbed bile metabolism has been reported in animal and human diabetes. Bile acid–binding resin prevents and treats diabetes. Diabetes remission after bariatric surgeries is also suggested to be related to changed bile acid metabolism.

Analyzing genes in proliferation and apoptosis groups reveal decreased replication. The proliferation functional group includes reduced expression of HPSE, PBK and POLD. Heparanase (HPSE) plays an important role in metastasis and angiogenesis. Lymphokine-activated killer T-cell-originated protein kinase (PBK) encodes a mitotic kinase related to mitogen-activated protein kinase kinase (MAPKK) family. DNA polymerase delta catalytic subunit (POLD1) is a DNA polymerase involves in DNA repair synthesis after damage. The apoptosis group including CASP1. Overexpressing CASP1 at 4w GK strain causes increased apoptosis.

The TF regulatory network contributing to initial hyperglycemia at 4w continues to be active in 8-12w diabetic GK strain. In this middle term diabetes, networks making the chaos directing to the diabetes and networks compensate are both active (Figure 6). A good example is the increased expression level of Cathepsin D (CTSD). Animal and human data suggest that CTSD selectively degrades macrophage inflammatory proteins and is possibly used by tumor to escape antitumoral immune response. Higher expression of CTSD may be secondary to the increased inflammation in diabetics. However, CTSD will enhance receptor-mediated insulin degradation in vivo, thus inducing insulin resistance [35]. The insulting and compensating battle slowly progress diabetes to next stage.

At stable hyperglycemia stage, fewer networks are activated compared to middle term stage. However, the insult factors expressed in this stage make diabetes a robust system and unable to return to normals.

• Important networks keep normal or diabetes robustness

Hyperglycemia is consistent in 16-20w GK rat. Thus we believe that genes expressed at this stage in WKY and GK rats are important to keep a steady normal or disease phase.

The first compelling result is the importance of nuclear receptors to maintain the non-diabetic robustness after analyzing TF-regulatory network in the 16-20w WKY (Figure 7A). Nuclear receptors directly bind to DNA, thereby controlling essential biology functions, such as development, homeostasis, and metabolism. HNF4A, NR3C1, ESR1, AR, PPARG, NR1D1 all belong to nuclear receptor family. HNF4A belongs to nuclear receptor subfamily 2. NR3C1, ESR1 and AR are members of subfamily 3, while PPARG and NR1D1 are included in subfamily 1. They work in concert to defense the disturbance outside. Disease states such as diabetes may be induced by the opposite activities of these receptors. Hepatocyte nuclear factor 4 alpha (HNF4A) has been described previously in metabolism section. It directly regulates genes involved in glucose transport and glycolysis. Estrogen receptor alpha (ESR1) and androgen receptor (AR) are activated by the sex hormone estrogen and androgen, respectively. Numerous data suggest that estrogen improves glucose metabolism and plasma lipids in T2DM [36]. AR deficiency plays key roles in the development of insulin and leptin resistance, which explains increased diabetes incidence in elder male [37]. The glucocorticoid receptor, also known as NR3C1 (nuclear receptor subfamily 3, group C, member 1) is expressed in almost every cell controlling the development, metabolism, especially immune response. NR3C1 decreases inflammation. Peroxisome proliferator–activated receptor-γ (PPARG) regulates fatty acid storage and glucose metabolism, thus improve insulin sensitivity without increased insulin secretion.

Many insulin sensitizing drugs are PPARG agonists [38]. N subfamily 1, group D, member 1 (NR1D1) also known as Rev-ErbA activates histone deacetylation, thereby regulating gene expression. Publications indicate that SNPs in these nuclear receptors associate with obesity and/or diabetes. Our data suggest that decreased expression of HNF4A, NR3C1, ESR1, AR, PPARG and NR1D1 overexpression contribute to T2DM.

In GK rats, some compensational pathways still exist, for example a NO synthesis pathway is up-regulated. Three genes nitric oxide synthase 3 (NOS3), argininosuccinate synthetase (ASS1), and NAD(P)H: quinone oxidoreductase (NQ01) related to this pathway are overexpressed. It is well-known that NO decreases blood pressure and promotes vascular actions of insulin. NOS3 catalyzes arginine, oxygen and NADPH to NO and citrulline. ASS1 and NQ01contribute to this metabolism cycle. Many cytokines increase NO regeneration several folds. Increased NO synthesis pathway indicates an inflammation environment in the liver in GK rats. Because reduced cell NO action has been reported in diabetes, the beneficial effects of increased NO production is uncertain. Data analyze reveal increased insulin resistance, hypertension and apoptosis are important to push diabetes to next stage (Figure 7B). Protein kinase C alpha (PRKCA) is mostly expressed in hepatocytes promoting glycogenolysis and gluconeogenesis. Activation of PRKCA mediates serine/threonine phosphorylation of the insulin receptor resulting in decreased active form of insulin receptor, inducing insulin resistance [39]. Angiotensin I converting enzyme 2 (ACE2) is an exopeptidase that catalyses angiotensin peptides and has opposite effects on RAS axis. Thus decreased expression levels of ACE2 accelerate the pathologic process such as hypertension, inflammation, fibrosis and inflammation. Gap junction alpha-1(GJA1) also known as connexin-43, is a component of gap junctions providing a route for cell to cell communication via diffusion materials. Decreased GJA1 expression particularly in hyperglycemia accelerates apoptosis.

• Advantages of network screening over single gene based method

When comparing our results to the original study conducted by Dr. Almon [12], network screening is clearly superior to the single gene-based analysis. One good example is to explain how liver insulin resistance (IR) develops. IR is the major character of T2DM and also present in GK rats after 8 weeks of age. In the original study, authors notice higher expression of P85, thus suspecting interaction of P85 with IRS leading to IR. However, we believe that the developing IR is a dynamic process involving many steps. The first step could be significantly decreased IGF-1R expression after 8 weeks inducing IR in GK. After that, higher expression of CTSD accelerates IR. Compensational pathways also occur, which includes IRS2 overexpression at 8-12w in GK. However as PKC overexpression plus decreased expression of many nuclear factors such as PPARG at 16-20w, IR deteriorates and diabetes becomes unreturnable. Our method is based on the networks and is very different from the gene-based method of identifying the differential expression.

Discussion and Conclusion

T2DM is a complex disease, which is usually not caused by individual gene changes, thereby requiring systems biology methods to understand their mechanisms. In this work, we have performed comprehensive active regulatory network survey by network screening to the published GK *vs.* WKY liver microarray data [12]. Available resources from MSigDB and TRANSFAC are combined together to identify the significant pathways responsive to the status of diabetes or normals. After combining the networks according to features or time points, we built functional or time series TF regulatory network graphs. Analyzing the graphs reveals: 1. More pathways are active during inter-middle stage diabetes; 2. Inflammation, hypoxia, increased apoptosis, decreased proliferation, and altered metabolism are characteristics in GK strain, and displayed as early as 4w. 3. Diabetes progression accompanies insults and compensations. 4. Nuclear receptors work in concert to maintain normal glycemic robustness system.

Network-based analysis based on high throughput data is a challenging issue, which is expected to help us understand complex disease such as diabetes and further elucidate the essential mechanisms of living organisms which would escape conventional single gene-based analysis. In this paper, instead of picking up differently expressed genes from high-throughput data, we use known functional pathways to screen datasets and evaluate significantly activated pathways. Then genes with no annotated linkages to TF are overlooked and the available gene regulatory relationships are integrated to form a comprehensive TF regulatory network, which cannot be achieved by single gene based method. The network shows a whole picture of activated TF regulated functional gene sets under certain conditions and is much easier to bring the biological insights to us.

To our knowledge, two conclusions have not been reported before. The first one comes out from TF regulatory network at 4w GK. It is well-known that the major cause of diabetes in GK rats is insulin secreting beta cell dysfuction. Beta cell mass in GK is only half of that in WKY after birth. To be surprised, we find that at very early age liver already exhibits serious gene expression alternations involving in bile metabolism

dysfunction, inflammation, increased apoptosis and decreased proliferation, which greatly contribute to diabetes development. Another interesting finding is that the 6 nuclear receptors working in concert to maintain robustness of normal blood glucose. Although the relationships of those nuclear receptors with diabetes have been investigated individually before, it is the first time to report how they work together as a fine tune. Restoring their network regulation may have important therapeutic potentials.

This is the first time to use network screening to explain the role of liver in development of diabetes and the underline mechanism. The results provide many important rational information and insights into guiding experiments design. It is worth pointing out that the molecular relationships change dynamically, depending on the conditions in a living cell, which suggests implicitly that all of the relationships in the knowledge-based network do not always exist. Note that some methods are proposed for identifying the active networks from measured data [40]. Our method evaluates the networks from only one set of data measured under one condition to estimate the absolute consistency between network structure and the data, while the other methods generally need the two sets of data to estimate their relative difference by some criteria such as mutual information. We combined various resources together to identify the significant regulatory networks related to the development stages of diabetes. The matching between networks and gene expression profiling was identified by the evaluation of network screening. The active regulatory networks are the potential disease signatures from the comparison of GK and WKY rats. The dynamics of regulatory networks indicate the dysfunctional progression from the network perspective.

In conclusion, network screening is a superior approach to analyze complex disease such as diabetes. The conclusions drawn from this method are more complete and systemic, which gives biologist better guidance for further experiment design.

Actually, we are now extending this approach for screening general biomolecular networks [9,10] with both directed and undirected edges, and in future possibly for studying the problem of networkomics (or netomics) which covers all stable forms of biomolecular networks [41] not only at different biological conditions but also at different spatiotemporal situations.

Abbreviations

T2DM: Type 2 diabetes mellitus; GK: Goto-Kakizaki; WKY: Wistar-Kyoto; IGT: impaired glucose tolerance stage; IFT: impaired fasting glucose stage; GCP: graph consistency probability; TF: transcriptional factor; MSigDB: molecular signatures database; FDR: false discovery rate; DAG: directed acyclic graph; GN: gaussian network; GEO: gene expression omnibus; MODY: maturity-onset non-insulin-dependent diabetes of the young

Acknowledgements

We are grateful to Dr. Jiarui Wu, Dr. Jacob Sten Petersen, Dr. Trine Ryberg Clausen and Mr. Rongkuan Hu for their comments and support. This work was supported by grants from NN-CAS Research Foundation under NO. NNCAS-2009-1 (H.Z.), Major State Basic Research Development Program of China (973 Program) under NO.2011CB504003 (H.Z.), National Natural Science Foundation of China under NO. 81070657 (H.Z.), NO.61072149 and NO.91029301 (L.C. and Z.P.L.), Chief Scientist Program of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences under NO. 2009CSP002 (L.C.), and Development of Analysis Technology for Induced Pluripotent Stem (iPS) Cell, from New Energy and Industrial Technology Development Organization, Japan (S.S. and K.H.). This work was also partically supported by Shanghai Natural Science Foundation under NO. 11ZR1443100 (Z.P.L.) and JSPS FIRST Programm, Japan (L.C.).
This article has been published as part of *BMC Systems Biology* Volume 5 Supplement 1, 2011: Selected articles from the 4th International Conference on Computational Systems Biology (ISB 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1752-0509/5?issue=S1.

Author details

[1]Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China. [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan. [3]INFOCOM Corporation, Tokyo 150-0001, Japan. [4]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei 230027, China. [5]Institute of Systems Biology, Shanghai University, Shanghai 200444, China.

Authors' contributions

HZ, KH and LC conceived the research. HZ, SS and ZPL performed the study. GP and JW gave valuable suggestions and improvements. LC and HZ supervised the project. HZ and ZPL drafted a version of the manuscript. All authors wrote and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 20 June 2011

References

1. Smyth S, Heron A: **Diabetes and obesity: the twin epidemics.** *Nat Med* 2006, **12**:75-80.
2. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P: **A genome-wide association study identifies novel risk loci for type 2 diabetes.** *Nature* 2007, **445**:881-885.
3. Hetherington MM, Cecil JE: **Gene-environment interactions in obesity.** *Forum Nutr* 2010, **63**:195-203.
4. Hayden MR: **Islet amyloid, metabolic syndrome, and the natural progressive history of type 2 diabetes mellitus.** *J Pancreas* 2002, **3**:126-138.
5. Proietto J, Andrikopoulos S, Rosella G, Thorburn A: **Understanding the pathogenesis of type 2 diabetes: can we get off the metabolic merry-go-rounds?** *Aust N Z J Med* 1995, **25**:870-875.
6. Galli J, Fakhrai-Rad H, Kamel A, Marcus C, Norgren S, Luthman H: **Pathophysiological and genetic characterization of the major diabetes locus in GK rats.** *Diabetes* 1999, **48**:2463-2470.
7. Gauguier D, Froguel P, Parent V, Bernard C, Bihoreau MT, Portha B, James MR, Penicaud L, Lathrop M, Ktorza A: **Chromosomal mapping of genetic loci associated with non-insulin dependent diabetes in the GK rat.** *Nat Genet* 1996, **12**:38-43.
8. Permutt MA, Wasson J, Cox N: **Genetic epidemiology of diabetes.** *J Clin Invet* 2005, **115**:1431-1439.
9. Chen L, Wang RS, Zhang X: **Biomolecular Networks: Methods and Applications in Systems Biology.** John Wiley & Sons; 2009.
10. Chen L, Wang RQ, Li G, Aihara K: **Modeling Biomolecular Networks in Cells: Structures and Dynamics.** Springer-Verlag; 2010.

11. Saito S, Aburatani S, Horimoto K: **Network evaluation from the consistency of the graph structure with the measured data.** *BMC Sys Biol* 2008, **2**:84.
12. Almon RR, DuBois DC, Lai W, Xue B, Nie J, Jusko WJ: **Gene expression analysis of hepatic roles in cause and development of diabetes in Goto-Kakizaki rats.** *J Endocrinol* 2009, **200**:331-346.
13. Wingender E: **TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation.** *Brief. Bioinformatics* 2008, 326-332.
14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.** *Proc. Natl. Acad. Sci. USA* 2005, **102**:15545-15550.
15. Pearl J: **Probabilistic Reasoning in Intelligent Systems.** California, Kaufmann Morgan Publishers; 1988.
16. Whittaker J: **Graphical Models in Applied Multivariate Statistics.** New York, John Wiley and Sons; 1990.
17. Coles S: **An Introduction to Statistical Modeling of Extreme Values.** London, Springer-Verlag; 2001.
18. Lee NK, Sowa H, Hinoi E, Ferron M, Ahn JD, Confavreux C, Dacquin R, Mee PJ, McKee MD, Jung DY, Zhang Z, Kim JK, Mauvais-Jarvis F, Ducy P, Karsenty G: **Endocrine regulation of energy metabolism by the skeleton.** *Cell* 2007, **130**:456-469.
19. Beijers HJ, Losekoot M, Odink RJ, Bravenboer B: **Hepatocyte nuclear factor (HNF)1A and HNF4A substitution occurring simultaneously in a family with maturity-onset diabetes of the young.** *Diabet Med* 2009, **26**:1172-1174.
20. Nikkilä EA, Huttunen JK, Ehnholm C: **Postheparin plasma lipoprotein lipase and hepatic lipase in diabetes mellitus. Relationship to plasma triglyceride metabolism.** *Diabetes* 1977, **26**:11-21.
21. Das B, Pawar N, Saini D, Seshadri M: **Genetic association study of selected candidate genes (ApoB, LPL, Leptin) and telomere length in obese and hypertensive individuals.** *BMC Med Genet* 2009, **10**:99.
22. Greenhalgh CJ, Rico-Bautista E, Lorentzon M, Thaus AL, Morgan PO, Willson TA, Zervoudakis P, Metcalf D, Street I, Nicola NA, Nash AD, Fabri LJ, Norstedt G, Ohlsson C, Flores-Morales A, Alexander WS, Hilton DJ: **SOCS2 negatively regulates growth hormone action in vitro and in vivo.** *J Clin Invest 2005* 8211, **115**:397-406.
23. Turnley AM, Faux CH, Rietze RL, Coonan JR, Bartlett PF: **Suppressor of cytokine signaling 2 regulates neuronal differentiation by inhibiting growth hormone signaling.** *Nature Neuroscience* 2002, **5**:1155-1162.
24. Haluzik M, Yakar S, Gavrilova O, Setser J, Boisclair Y, LeRoith D: **Insulin Resistance in the Liver-Specific IGF-1 Gene-Deleted Mouse Is Abrogated by Deletion of the Acid-Labile Subunit of the IGF-Binding Protein-3 Complex Relative Roles of Growth Hormone and IGF-1 in Insulin Resistance.** *Diabetes* 2003, **52**:2483-2489.
25. Rui L, Yuan M, Frantz D, Shoelson S, White MF: **SOCS-1 and SOCS-3 block insulin signaling by ubiquitin-mediated degradation of IRS1 and IRS2.** *J Biol Chem* 2002, **277**:42394-42398.
26. Bolscher BG, de Boer M, de Klein A, Weening RS, Roos D: **Point mutations in the beta-subunit of cytochrome b558 leading to X-linked chronic granulomatous disease.** *Blood* 1991, **77**:2482-2487.
27. Fan F, Jin S, Amundson SA, Tong T, Fan W, Zhao H, Zhu X, Mazzacurati L, Li X, Petrik KL, Fornace AJ Jr, Rajasekaran B, Zhan Q: **ATF3 induction following DNA damage is regulated by distinct signaling pathways and over-expression of ATF3 protein suppresses cells growth.** *Oncogene* 2002, **1**:7488-7496.
28. Kannanayakal TJ, Mendell JR, Kuret J: **Casein Kinase 1 alpha associates with the tau-bearing lesions of inclusion body myositis.** *Neurosci Lett* 2008, **431**:141-145.
29. Bereczky Z, Katona E, Muszbek L: **Fibrin stabilization (factor XIII), fibrin structure and thrombosis.** *Pathophysiol Haemost Thromb* 2005, **33**:430-437.
30. Matsuura E, Kobayashi K, Matsunami Y, Lopez LR: **The immunology of atherothrombosis in the antiphospholipid syndrome: antigen presentation and lipid intracellular accumulation.** *Autoimmun Rev* 2009, **8**:500-505.
31. Auwerx J, Bouillon R, Collen D, Geboers J: **Tissue-type plasminogen activator antigen and plasminogen activator inhibitor in diabetes mellitus.** *Arteriosclerosis* 1988, **8**:68-72.
32. Briggs SD, Bryant SS, Jove R, Sanderson SD, Smithgall TE: **The Ras GTPase-activating protein (GAP) is an SH3 domain-binding protein and substrate for the Src-related tyrosine kinase, Hck.** *J. Biol. Chem* 1995, **24**:14718-14724.
33. Claus V, Jahraus A, Tjelle T, Berg T, Kirschke H, Faulstich H, Griffiths G: **Lysosomal enzyme trafficking between phagosomes, endosomes, and lysosomes in J774 macrophages. Enrichment of cathepsin H in early endosomes.** *J. Biol. Chem* 1998, **273**:9842-9851.
34. Kosters A, Jirsa M, Groen AK: **Genetic background of cholesterol gallstone disease.** *Biochim Biophys Acta* 2003, **1637**:1-19.
35. Nadler ST, Stoehr JP, Schueler KL, Tanimoto G, Yandell BS, Attie AD: **The expression of adipogenic genes is decreased in obesity and diabetes mellitus.** *Proc Natl Acad Sci U S A* 2000, **97**:11371-11376.
36. Geisler JG, Zawalich W, Zawalich K, Lakey JR, Stukenbrok H, Milici AJ, Soeller WC: **Estrogen Can Prevent or Reverse Obesity and Diabetes in Mice Expressing Human Islet Amyloid Polypeptide.** *Diabetes* 2002, **7**:2158-2169.
37. Kalyani RR, Dobs AS: **Androgen deficiency, Diabetes, and the metabolic syndrome in men.** *Curr Opin Endocrinol Diabetes Obes* 2007, **14**:226-234.
38. Lefebvre B, Benomar Y, Guédin A, Langlois A, Hennuyer N, Dumont J, Bouchaert E, Dacquet C, Pénicaud L, Casteilla L, Pattou F, Ktorza A, Staels B, Lefebvre P: **Proteasomal degradation of retinoid X receptor alpha reprograms transcriptional activity of PPARgamma in obese mice and humans.** *J Clin Invest* 2010, **120**:1454-1468.
39. Chin JE, Liu F, Roth RA: **Activation of protein kinase C alpha inhibits insulin-stimulated tyrosine phosphorylation of insulin receptor substrate-1.** *Mol Endocrinol* 1994, **8**:51-58.
40. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis.** *Mol Sys Biol* 2007, **3**:140.
41. Lei HB, Zhang JF, Chen L: **Multi-equilibrium property of metabolic networks: SSI module.** *BMC Sys Biol* 2010, **5(Suppl 1)**:S15.

doi:10.1186/1752-0509-5-S1-S16
Cite this article as: Zhou *et al.*: Network screening of Goto-Kakizaki rat liver microarray data during diabetic progression. *BMC Systems Biology* 2011 5(Suppl 1):S16.

Saito et al. BMC Systems Biology 2011, 5(Suppl 1):S17
http://www.biomedcentral.com/1752-0509/5/S1/S17

BMC Systems Biology

REPORT Open Access

Possible linkages between the inner and outer cellular states of human induced pluripotent stem cells

Shigeru Saito[1,2†], Yasuko Onuma[3†], Yuzuru Ito[3†], Hiroaki Tateno[4†], Masashi Toyoda[5†], Akutsu Hidenori[5], Koichiro Nishino[5], Emi Chikazawa[5], Yoshihiro Fukawatase[5], Yoshitaka Miyagawa[6], Hajime Okita[6], Nobutaka Kiyokawa[6], Yohichi Shimma[4], Akihiro Umezawa[5], Jun Hirabayashi[4], Katsuhisa Horimoto[1,7*], Makoto Asashima[3,8*]

From The 4th International Conference on Computational Systems Biology (ISB 2010)
Suzhou, P. R. China. 9-11 September 2010

Abstract

Background: Human iPS cells (hiPSCs) have attracted considerable attention for applications to drug screening and analyses of disease mechanisms, and even as next generation materials for regenerative medicine. Genetic reprogramming of human somatic cells to a pluripotent state was first achieved by the ectopic expression of four factors (Sox2, Oct4, Klf4 and c-Myc), using a retrovirus. Subsequently, this method was applied to various human cells, using different combinations of defined factors. However, the transcription factor-induced acquisition of replication competence and pluripotency raises the question as to how exogenous factors induce changes in the inner and outer cellular states.

Results: We analyzed both the RNA profile, to reveal changes in gene expression, and the glycan profile, to identify changes in glycan structures, between 51 cell samples of four parental somatic cell (SC) lines from amniotic mesodermal, placental artery endothelial, and uterine endometrium sources, fetal lung fibroblast (MRC-5) cells, and nine hiPSC lines that were originally established. The analysis of this information by standard statistical techniques combined with a network approach, named network screening, detected significant expression differences between the iPSCs and the SCs. Subsequent network analysis of the gene expression and glycan signatures revealed that the glycan transfer network is associated with known epitopes for differentiation, e.g., the SSEA epitope family in the glycan biosynthesis pathway, based on the characteristic changes in the cellular surface states of the hiPSCs.

Conclusions: The present study is the first to reveal the relationships between gene expression patterns and cell surface changes in hiPSCs, and reinforces the importance of the cell surface to identify established iPSCs from SCs. In addition, given the variability of iPSCs, which is related to the characteristics of the parental SCs, a glycosyltransferase expression assay might be established to define hiPSCs more precisely and thus facilitate their standardization, which are important steps towards the eventual therapeutic applications of hiPSCs.

* Correspondence: k.horimoto@aist.go.jp; m-asashima@aist.go.jp
† Contributed equally
[1]Computational Biology Research Center, National Institute of Advanced Industrial Science Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
[3]Research Center for Stem Cell Engineering, National Institute of Advanced Industrial Science Technology (AIST), Tsukuba Central 4, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8562, Japan
Full list of author information is available at the end of the article

Background

Reprogramming of human and mouse fibroblasts to induced pluripotent stem cells (iPSCs) has been achieved by the expression of only four transcription factors, Oct4, Sox2, Klf4, and c-Myc, referred to as the "four factors" [1]. iPSCs hold great promise for human disease analyses and therapies, because they are highly similar to embryonic stem cells (ESCs) in their ability to self-renew and generate all three germ layers. A key question raised by transcription factor-induced reprogramming to self-renewal and pluripotency is how the four factors act to accomplish these changes in the inner and outer cell states.

The morphological changes accompanying the reprogramming of somatic cells to iPSCs can be visually identified by alkaline phosphatase staining. The changes in the outer cellular states are further monitored by characteristic molecular markers. In fact, the monoclonal antibodies currently used to define ESCs and iPSCs, including the globo-series glycosphingolipid epitopes SSEA-3 and SSEA-4, and the keratanase-sensitive glycoprotein associated epitopes Tra 1–60 and Tra 1–81, recognize glycan antigens [2-4]. Recently, global analyses of glycan signatures for pluripotency on the cell surface were reported, by direct observations of glycan structures by MALDI-TOF mass spectrometric and NMR spectrometric profiling in ESCs [5] and indirect observations of lectins by a lectin microarray in stem cells [6]. Furthermore, the extracellular matrix is also important for controlling cellular states through cell-cell interactions [7].

The inner cellular states also change during the remodeling of the somatic cell transcription and chromatin programs to the ES-like state, including the reactivation of the somatically silenced X chromosome, the demethylation of the Oct4 and Nanog promoter regions, and the genome-wide resetting of histone H3 lysine 4 and 27 trimethylation [8]. It is particularly important to determine whether the gene expression differences observed between hiPSCs and the corresponding parental cells actually reflect the differences between these pluripotent cell types, especially between hiPSCs and ESCs [9-12]. Gene expression signatures were reported for reprogrammed cell lines derived in different labs by various methods [13-15]. In addition, genome-wide mapping of transcription factor targets by ChIP, combined with microarrays or sequencing methods, can provide a foundation for understanding transcriptional networks [16-20]. Expanding the number of transcription factors analyzed by ChIP-based methods is especially informative in dissecting system level biological processes. In ESCs, some groups have used new methods for global target mapping to predict the target genes regulated by Oct4, Sox2, and Nanog, and these studies revealed the combinatorial occupancy of target gene promoters by these core factors, as well as both autoregulatory and feed-forward transcriptional circuits [16-20].

Here, we applied two methods, RNA profiling to uncover gene expression changes, and lectin profiling to survey glycan structure changes, to compare human iPSCs and parental somatic cells (SCs), including 51 cells of nine iPSC lines from four kinds of SCs, from amniotic mesodermal, placental artery endothelial, and uterine endometrium sources, and one available hiPSC line, MRC-5. The changes were computationally analyzed by a network approach [21] in conjunction with information on the gene binding from previous CHIP-seq studies and knowledge of the gene functions. The sum of these analyses uncovered novel expression, network, and lectin signatures that are unique to the hiPSC lines and differ from those of the parental cells. The following correspondence between the three signatures identified a few glycosyltransferases as novel candidates, due to the characteristic changes of their cellular surface states in hiPSCs, which shed light on a possible link between the inner and outer cellular states. Whether the hiPSC signatures described here actually play a functional role in bridging the gap between the two cellular states warrants extensive further investigations.

Results and Discussion

hiPSCs descended from different parent SCs are distinguishable by gene expression

To determine the gene expression signatures of hiPSCs, a detailed genome-wide expression analysis was performed to compare iPSCs and their parental SCs from amniotic mesodermal (AM), placental artery endothelial (PAE), uterine endometrium (UtE), and MRC-5 (MRC) cell sources (see additional file 1: Cell lines and numbers of passages analyzed in the present study, and Methods). In total, 51 cell samples of 13 cell lines (39 hiPSC samples of 9 hiPSC lines [22,23]) were analyzed in the present study, for a statistical comparison of the hiPSCs and the parental SCs (see additional file 2: Generation of iPSCs from human PAE cells).

Unsupervised hierarchical clustering of the gene expression data across the four hiPSC lines (AM, PAE, UtE, MRC) and their corresponding parental SCs revealed interesting patterns in the gene expression heat map (Fig. 1A). First, the hiPSCs were clearly distinguishable from their respective parental SCs. This finding was verified by another clustering method with a distinct technique (see additional file 3: Clustering of the gene expression data with another method). Second, the gene expression profiles of the four hiPSC lines were linked to those of their parental SCs, while these profiles of the

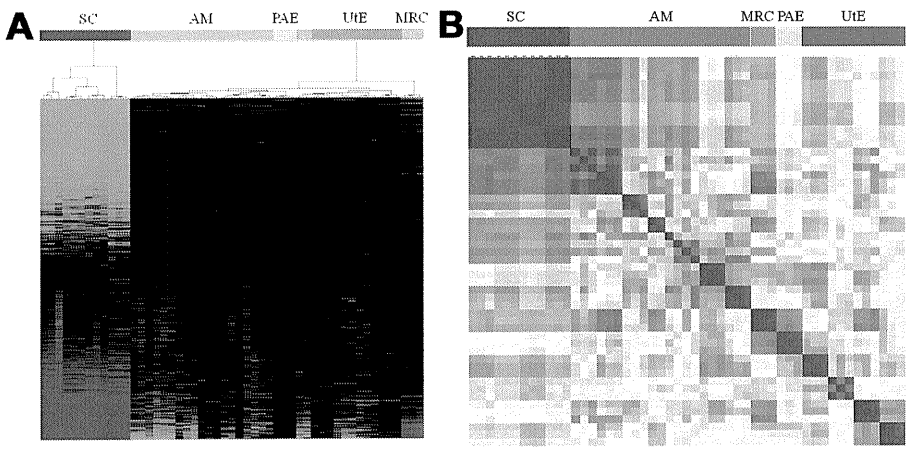


Figure 1 Classification of hiPSCs and SCs from gene expression data (*A*) Heat map and hierarchical clustering for all cells and genes. Cell types are indicated by colored bars, and the following abbreviations are used for the sources of the cell types of the human somatic cells (SCs) and induced pluripotent stem cells (hiPSCs): AM, amniotic membrane cell; PAE, placental artery cell; UtE, uterine endometrium cell; MRC, MRC-5 cell. Details are available in additional file 1. (*B*) Heat map of the correlation coefficient matrix for all of the cells. See also additional file 3.

hiPSCs from different passages were clustered more closely with each other, rather than with those of the hiPSCs from the corresponding parental SCs (Fig. 1A). In support of these findings, a Pearson correlation analysis demonstrated that the gene expression profiles of the hiPSCs from different passages were more closely related to each other than to the hiPSCs from the same parental SCs (Fig. 1B and see also additional file 4: Correlation coefficient matrix for all cells). Furthermore, the above relationship between the hiPSCs and the parental SCs was verified by estimating the classification accuracy by leave-one-out cross-validation (LOOCV) on the nearest-neighbor classifier, based on Pearson's correlation distance (see additional file 5: Cross validation of cell classification).

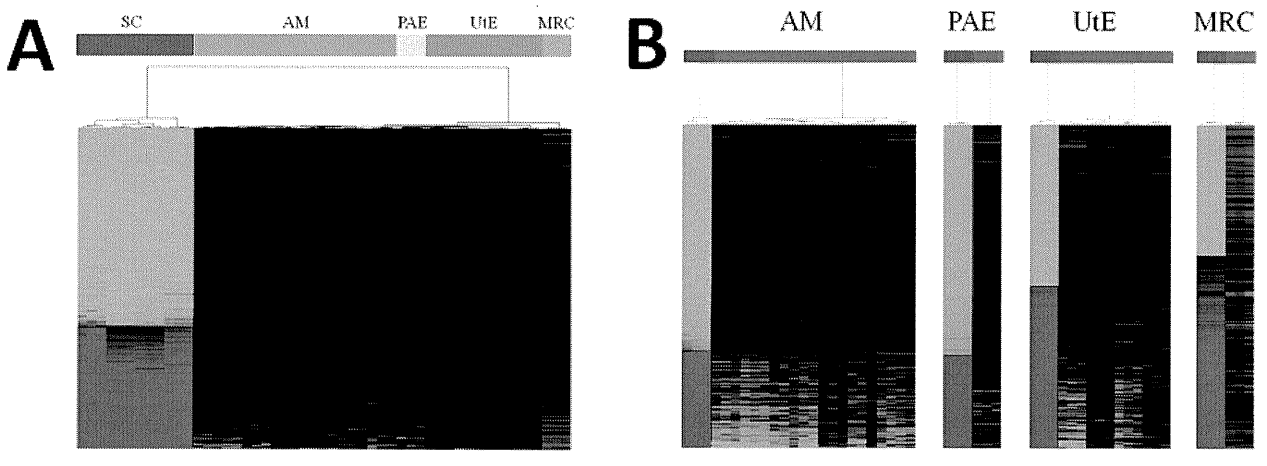
Gene expression signature for hiPSCs descended from different parent SCs

Analyses of the differences in gene expression between the four hiPSC lines and the parental SC lines revealed that 8,287 (out of 16,483) genes in the AM cells, 7,249 genes in the MRC cells, 7,465 genes in the PAE cells, and 6,314 genes in the UtE cells showed significant differences between the hiPSC lines and the corresponding parental lines, as determined using the Student's *t*-test (for a false discovery rate [FDR] < 5% and requiring a ≥2.0-fold change in expression between the cells) (Fig. 2A). In total, 2,502 genes were categorized into a gene expression signature common to the above four gene sets with expression differences (Fig. 2B and see also additional files 6: Number matrix for common genes, and 7: List of 2,502 genes in the expression signature, together with the fold-changes in expression levels and FDR values).

In this expression signature, 62% of the genes (1,549 genes) were upregulated and 38% (953 genes) were downregulated in the hiPSCs, as compared to the parental SCs (Fig. 2B and see also additional file 7). From the 953 genes in the gene signature that were expressed at lower levels, gene ontology analyses revealed 60 terms with significant probability (family-wise error rate [FWER] <0.05), whereas the 1,549 genes that were expressed at higher levels were characterized by 89 terms (see additional file 8: List of enriched GO terms with significant probabilities (FWER < 0.05)). In total, 149 terms were found, and the GO analysis was determined to be inadequate for defining the biological functions of the expression signature in hiPSCs. The 149 terms were summarized into 33 terms as a macroscopic view; these terms shared 9 terms between the higher and lower expression levels (Fig. 2C).

Network signature of hiPSCs by network screening

To elucidate the nature of the expression signature of the hiPSCs, we incorporated information on gene binding and function into a network analysis approach, named network screening [21] (see additional file 9: Schematic representation of the network screening used to obtain the network signature, and Methods). To prepare the network analysis, we identified 146 regulatory networks of 313 genes in the expression signature, which were classified with their functions using the gene sets defined previously [24] (see additional file 10: Reference networks and constituent genes, and Methods), among 519 genes that were identified as being bound by the four factors in ChIP-on-chip experiments [20]. We then analyzed the 146 reference networks, which were regarded as being directly induced by the four factors (OCT3/4, SOX2, KLF4, c-MYC), to define the



C

iPSC>SC		iPSC<SC	
GO TERM	FWER	GO TERM	FWER
multicellular organismal development	2.18E-23	**anatomical structure morphogenesis**	2.03E-08
cell differentiation	4.34E-18	**response to external stimulus**	8.06E-08
regulation of biological process	1.41E-16	**multicellular organismal development**	2.32E-07
cell communication	2.57E-14	response to stress	2.58E-07
cell-cell signaling	8.03E-12	**regulation of biological process**	5.70E-07
anatomical structure morphogenesis	7.75E-10	**cell proliferation**	7.84E-07
signal transduction	8.13E-08	**signal transduction**	3.03E-06
behavior	7.74E-07	cell death	3.15E-06
transport	6.69E-05	death	3.64E-06
cell proliferation	1.52E-04	**cell communication**	3.71E-05
response to external stimulus	4.09E-04	protein modification process	7.08E-04
embryonic development	1.91E-03	cell growth	6.74E-03
ion transport	3.78E-03	protein transport	2.20E-02
reproduction	6.27E-03	**cell differentiation**	3.96E-02
response to endogenous stimulus	1.90E-02	**biological_process**	4.07E-02
growth	2.67E-02	response to abiotic stimulus	4.08E-02
biological_process	3.11E-02		

Figure 2 Gene expression signature (*A*) Heat map of 2,502 genes in the hiPSCs and parental cells. Cell types are indicated by colored bars. (*B*) Heat maps for the MRC, AM, UtE, and PAE hiPSCs and their corresponding parental cells. The hiPSCs and the parental SCs are discriminated by colored bars (red, SCs; purple, hiPSCs). (*C*) GO terms in the Biological Process, with the significance probabilities for the genes with higher and lower levels of expression. GO terms were summarized into 33 terms, as a macroscopic view, and the overlapped terms in the higher and lower expression classes are indicated by bold characters. See also additional file 6.

network signature of the hiPSCs, according to the following two thresholds (Fig. 3): 1) the enrichment probability of the genes in the expression signature for each network; and 2) the consistency of the network structure in relation to the gene expression profile [21]. Thus, as the network signature, we defined 28 networks of 76 genes that fulfilled these conditions (Fig. 3A and see also additional file 11: Details of the network signature).

As expected, the network signature almost completely covered the pathways that were previously implicated in the reprogramming of hiPSC pluripotency (Figs. 3A and B). For example, the relationship between reprogramming for pluripotency and signal transduction was emphasized for the TGF-β [25], Wnt [26], and MAPK pathways [27]. In addition, pathways related to cell-cell interactions were implicated. Although the molecular mechanisms underlying the cell-cell interactions in the inner cellular states are less understood, several studies have highlighted the importance of cellular communication through the extracellular matrix with respect to changes in the cellular states, such as those that occur during development and differentiation [28]. Furthermore, relationships to cancer-related pathways were identified, consistent with the fact that the four factors induce various cancer cells [29]; this finding may be useful in the prevention of cancer induction by hiPSCs. Although several pathways in the network signature remain to be characterized, it provides clues as to the molecular mechanisms underlying the reprogramming for hiPSC pluripotency and self-renewal, in contrast to the information obtained from a characterization based simply on GO terms.

Networks with significant correlation between reprogramming and glycan biosynthesis

Interestingly, two regulatory networks related to the glycome, for linkage of the inner and outer cellular states,

A

Pathway
HSA04010-MAPK_SIGNALING_PATHWAY
HSA04115-P53_SIGNALING_PATHWAY
HSA04310-WNT_SIGNALING_PATHWAY
HSA04340-HEDGEHOG_SIGNALING_PATHWAY
HSA04350-TGF_BETA_SIGNALING_PATHWAY
HSA04630-JAK_STAT_SIGNALING_PATHWAY
HSA01430-CELL_COMMUNICATION
HSA04520-ADHERENS_JUNCTION
HSA04530-TIGHT_JUNCTION
HSA04540-GAP_JUNCTION
HSA04810-REGULATION_OF_ACTIN_CYTOSKELETON
HSA01030-GLYCAN_STRUCTURES_BIOSYNTHESIS_1
HSA01031-GLYCAN_STRUCTURES_BIOSYNTHESIS_2
HSA04916-MELANOGENESIS
HSA05210-COLORECTAL_CANCER
HSA05213-ENDOMETRIAL_CANCER
HSA05215-PROSTATE_CANCER
HSA05216-THYROID_CANCER
HSA05217-BASAL_CELL_CARCINOMA
HSA05218-MELANOMA
APOPTOSIS
APOPTOSIS_GENMAPP
CALCIUM_REGULATION_IN_CARDIAC_CELLS
HSA04060-CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
HSA04110-CELL_CYCLE
HSA04670-LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION
HSA05120-EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION
SMOOTH_MUSCLE_CONTRACTION

B

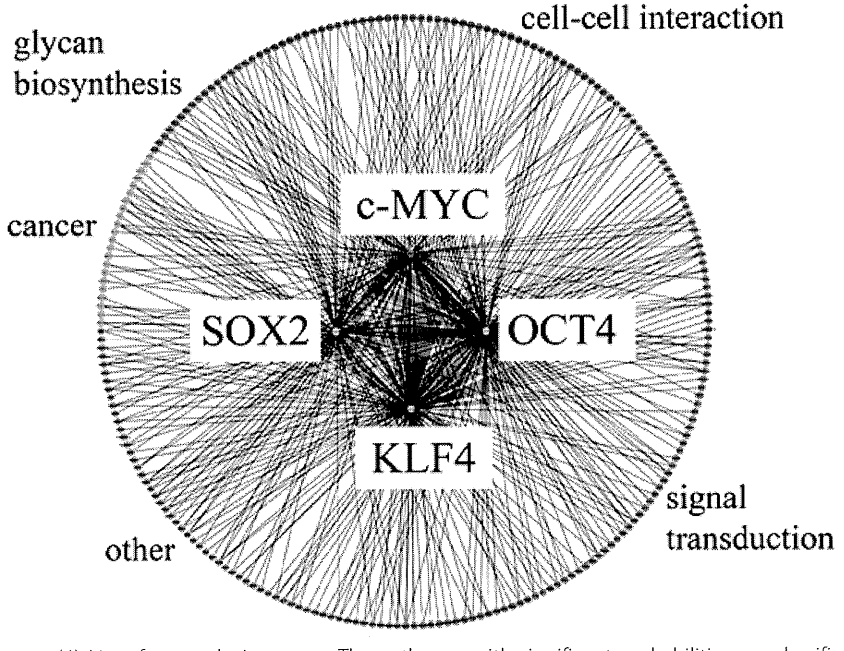


Figure 3 Network signature (*A*) List of network signatures. The pathways with significant probabilities are classified into the following categories: orange, pathways related to signal transduction; blue, pathways related to cell-cell interactions; red, pathways related to glycan biosynthesis; green, pathways related to cancer; and black, unclassified pathways. (*B*) Schematic presentation of networks. The four induced factors are described in the center, and the binding genes, which are colored according to the classification scheme described in (*A*), are connected by thin lines.

HSA01030-GLYCAN_STRUCTURES_BIOSYNTHESIS_1	**ALG3**, **ALG8**, B4GALT3, **DPAGT1**, **EXT1**, GALNT7, HS3ST3B1, **HS6ST2**, **MAN1A2**, MGAT1, **OGT**, **ST6GAL1***, STT3B
HSA01031-GLYCAN_STRUCTURES_BIOSYNTHESIS_2	A4GALT, **B3GNT3***, B3GNT5, B4GALT3, **GCNT2***, ST8SIA1, UGCG

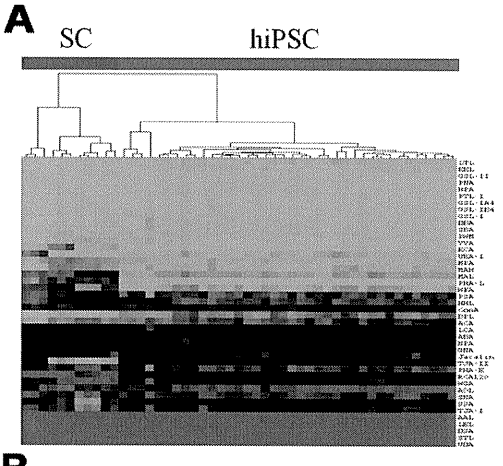
Figure 4 Genes involved in two glycome biosynthesis pathways The genes found in the expression signature are indicated by bold characters. Three genes related to glycan transfer are indicated by asterisks.

appeared in the network signature (Fig. 3A). In general, glycan biosynthesis is a multi-step process that requires a variety of enzymes, i.e., glycosyltransferases and enzymes involved in cytosolic sugar metabolism, and in many cases, glycan biosynthesis follows a glycan-specific, linear pathway. Most glycosyltransferases are regulated at the transcriptional level, thus warranting an assessment of the transcriptional profile of the glycan biosynthesis genes. In the two pathways, we found three genes (*ST6GAL1*, *B3GNT3*, and *GCNT2*) related to glycan transfer and two genes (*EXT1* and *HS6ST2*) related to heparan sulfate biosynthesis that were included in the expression signature (Fig. 4). These findings are consistent with recent studies that revealed the association between *N*-glycans and the maintenance of embryonic stem cell (ESC) pluripotency [5] and that between heparan sulfate and the reprogramming of ESCs [30]. Therefore, the genes identified in the above two pathways are candidates for the maintenance of the outer cellular state of iPSCs.

Glycan signature unique to hiPSCs

In addition to the expression and network signatures of the inner cell state, we examined the differences in the outer cellular states of the hiPSCs and the parental SCs using a lectin array, which detects glycan structures on cell surface proteins, based on glycan-lectin interactions [31]. In this analysis, the hiPSCs were clearly distinct from their parental SCs, and the dendrogram of the lectin microarray generated by unsupervised hierarchical clustering showed a clear separation between the hiPSCs and the parental SCs (Fig. 5A). Although the binding relationships between lectins and glycans and the relationships between the changes in glycan structures and the corresponding glycosyltransferases are redundant [32], we summarized the lectin-glycan-glycosyltransferase relationships using KEGG GLYCAN [33] and by manual curation of previous reports. We found strong correlations between the gene expression profiles of the glycosyltransferases and the corresponding lectin fluorescence intensities (see additional file 12: Lectin-glycan-glycosyltransferase relationships and correlations of lectin array intensities with glycosyltransferase expression patterns). This result indicates that the glycosyltransferases are coordinately expressed with the reprogramming, with the result that the hiPSCs bear glycan structures that are distinct from those of their parental SCs, reflecting the reprogramming of the inner cellular state.

Based on the Student's *t*-test (FDR <0.05) analysis, 28 of the 43 lectins in the lectin microarray showed significant differences between the hiPSCs and the parental SCs (see also additional file 12). For the glycan signature, we assigned 16 lectins, which interacted with the 12 glycosyltransferases that were related to the six patterns of glycan reactions, based on the correspondence with the expression signature (Fig. 5B).



B

Lectin	Reaction	Transferase
AOL	Fuc a1-2 Gal	FUT2
MAL	Neu5Ac a2-3 Gal	ST3GAL1
TJA-II	Fuc a1-2 Gal	FUT2
WFA	GalNAc a1- Ser/Thr	GALNT6, GALNT8, GALNT10, GALNT12, GALNT14, GALNTL2
SNA	Neu5Ac a2-6 Gal	ST6GAL1
BPL	Gal b1-3 GalNAc	B3GALT5, B3GALT1
TJA-I	Neu5Ac a2-6 Gal	ST6GAL1
SSA	Neu5Ac a2-6 Gal	ST6GAL1
ABA	Gal b1-3 GalNAc	B3GALT5
	GalNAc a1- Ser/Thr	GALNT6, GALNT8, GALNT10, GALNT12, GALNT14, GALNTL2
MPA	GalNAc a1- Ser/Thr	GALNT6, GALNT8, GALNT10, GALNT12, GALNT14, GALNTL2
AAL	Fuc a1-2 Gal	FUT2
ACA	Gal b1-3 GalNAc	B3GALT5
Jacalin	Gal b1-3 GalNAc	B3GALT5
	GalNAc a1- Ser/Thr	GALNT6, GALNT8, GALNT10, GALNT12, GALNT14, GALNTL2
VVA	GalNAc a1- Ser/Thr	GALNT6, GALNT8, GALNT10, GALNT12, GALNT14, GALNTL2
STL	GlcNAc b1-3 Gal	B3GNT2
UDA	GlcNAc b1-3 Gal	B3GNT2

Figure 5 Glycan signature (*A*) Heat map and hierarchical clustering of lectins. The hiPSCs and parental SCs are depicted by colored bars (red, SCs; purple, hiPSCs). (*B*) Correspondence between lectin gene expression patterns and glycan signatures. The lectin-glycosyltransferase relationships are described, together with their reactions. The lectins were selected under the condition that the corresponding glycosyltransferases were found in the expression signature.

Candidates of possible linkages between the inner and outer cellular states

Based on the correspondences between the expression and network signatures and between the expression and glycan signatures, we identified a total of 14 glycosyltransferases, since ST6GAL1 appeared in both sets of correspondences. These glycosyltransferases are potential candidates for the linkage between the inner and outer cellular states in hiPSCs. Interestingly, these glycosyltransferases may be related to the biosynthesis of a glycolipid that is characteristic of hiPSCs (see additional file 13: Knowledge-based relationships between glycosyltransferases and their biosynthetic pathways). Indeed, the allocation of the above glycosyltransferases to the pathways of "Glycan Biosynthesis and Metabolism" in KEGG GLYCAN (Table 1 and see also additional file 14: Locations of the glycosyltransferases detected in the present study in the pathways of "Glycan Biosynthesis and Metabolism") revealed that the glycosyltransferases identified in the present study are important in the glycolipid biosynthetic pathway. We identified B3GALT5 in the biosynthetic pathway for the carbohydrate chains of the globo-series of glycosphingolipids bearing the well-known SSEA-3 and SSEA-4 epitopes for ESCs and iPSCs [34,35], and although FUT2 is not directly involved in the synthesis of these glycans, it was found in the neighboring pathway that leads to the type IV H antigen. Furthermore, B3GALT1 and GCNT2, in addition to B3GALT5 and FUT2, were found in the extensive biosynthetic pathway of the carbohydrate chains of the lacto- and neolacto-series glycosphingolipids that carry SSEA-1, which is intensively expressed in ESCs, but is absent in cells that have differentiated from ESCs [36]. In addition, the members of the GALNT family, responsible for the *O*-glycan biosynthetic pathway of sialyl-T antigen, which is the most abundant glycan in several carcinoma cell lines, and ST6GAL1 were only found in the *N*-glycan biosynthetic pathway, which is involved in the generation of cell-surface carbohydrate determinants and the differentiation antigens HB-6, CDw75, and CD76 [37]. These analyses identified the glycosyltransferases that are directly and indirectly related to known glycan epitopes, thereby indicating the key molecules and the marker epitopes involved in reprogramming.

Further remarks on the present study

We analyzed more than 50 hiPSCs that were originally established from parental SCs, and the correspondence between each hiPSC and its parental SC was strictly controlled, which supports the present results based on a comparison with a clear genetic relationship. To further clarify the molecular mechanisms of the pluripotency, embryonic stem cells (ESCs) should be analyzed, following the context of the present study. Indeed, the pluripotency of hiPSCs has been extensively evaluated with reference to that of human ESCs, by various comparisons [38-43]. At present, we have prepared more than 100 hiPSCs with higher passages, and their comparisons with ESCs will be reported in the near future.

As for the experimental measurements, two types of data, gene expression and glycan structure, were analyzed by using microarrays and lectin arrays in the present study. To comprehensively understand the features of hiPSCs, more experimental data should be utilized, such as DNA-methylation and mi-RNA data. In particular, the recent availability of the next-gen sequencer will produce RNA-seq and ChIP-seq data with more

Table 1 Relationships between glycosyltransferase expression, network, and glycan signature

Glycosyltransferase	Functions	Glycan structure
ST6GAL1	N-, O-Glycan and glycolipid biosynthesis	Siaa2,6Galb1,4GlcNAc-R
B3GNT3	O-Glycan biosynthesis	core1 extension
GCNT2	N-, O-Glycan and glycolipid biosynthesis	I antigen Siaa2,3Galb1,3GalNAca1-
ST3GAL1	O-Glycan biosynthesis	Ser/Thr
FUT2	N-, O-Glycan and glycolipid biosynthesis	H antigen
GALNT6	O-Glycan biosynthesis	GalNAca1-Ser/Thr
GALNT8	O-Glycan biosynthesis	GalNAca1-Ser/Thr
GALNT10	O-Glycan biosynthesis	GalNAca1-Ser/Thr
GALNT12	O-Glycan biosynthesis	GalNAca1-Ser/Thr
GALNT14	O-Glycan biosynthesis	GalNAca1-Ser/Thr
GALNTL2	Unknown	
B3GALT5	N-, O-Glycan and glycolipid biosynthesis	Galb1,3GlcNAc-R, SSEA-3
B3GALT1	N-, O-Glycan and glycolipid biosynthesis N- and O-Glycan, keratan sulfate	Galb1,3GlcNAc-R
B3GNT2	biosynthesis	polylactosamine

The fourteen glycosyltransferases with identified correspondences between expression, network, and glycan signatures were allocated to biosynthetic pathways, using the KEGG GLYCAN database with modifications. The names of the pathways are listed. See also additional file 12 for the detailed pathways of notable glycosyltransferases, according to the KEGG GLYCAN database.

accurate measurements of gene expression and concrete information about the regulated genes. In addition, vast amounts of protein interaction data are accumulating. A comprehensive analysis integrating the various data from more hiPSCs will be reported in the near future.

Conclusions

The present study is the first to reveal the relationships between gene expression patterns and cell surface changes in hiPSCs, and it reinforces the importance of the cell surface to identify established iPSCs from SCs. In addition, given the variability of iPSCs, which is related to the characteristics of the parental SCs, a glycosyltransferase expression assay should be established that allows more precise definition of hiPSCs and facilitates their standardization, which are important steps towards eventual therapeutic applications of hiPSCs.

Methods

Cell experiments

Somatic cell pellets were harvested by scraping. The hiPSCs were incubated at 37°C, in a solution containing 1 mg/ml collagenase IV (Invitrogen, Carlsbad, CA), 1 mM $CaCl_2$, 20% KNOCKOUT^{TM} Serum Replacement (KSR), and 10% ACCUMAX (Innovative Cell Technologies, Inc., San Diego, CA). When the edges of the colonies started to dissociate from the bottom of the dish, the collagenase solution was removed and the cells were washed with medium. Colonies were then picked up and collected.

MRC-5 and amniotic mesodermal (AM) cells were maintained in POWEREDBY 10 medium (MED Shiratori Co., Ltd., Tokyo, Japan). The human placental artery endothelial (PAE) cells were harvested from human placenta. To isolate the arterial endothelium, we used the explant culture method, in which the cells were outgrown from pieces of the placenta's arterial vessels. Briefly, arterial vessels were separated from arteries in the chorionic plate, and chopped into approximately 5-mm^3 pieces. The pieces were washed in endothelial basal medium-2 (EBM-2; Cambrex, Walkersville, MD) and cultured in EGM-2MV medium (Cambrex), which consisted of EBM-2, 5% fetal bovine serum (FBS), and the supplemental growth factors VEGF, bFGF, EGF, and IGF. The arterial vessels attached to the substrata of the culture dishes (BD Falcon; Becton Dickinson, San Jose, CA). Cells migrated out from the surface of the tissues after about 20 days of incubation at 37°C in 5% CO_2. The cells were harvested in PBS containing 0.1% trypsin and 0.25 mM EDTA, and were re-seeded at a density of 3×10^5 cells in a 10-cm dish. Confluent monolayers of cells were subcultured. The culture medium was replaced every 3-4 days. Human uterine endometrium (UtE) was harvested from a patient with endometriosis. The endometrium was sterilized in PBS and cut into small pieces with dissection scissors. These pieces were placed in Dulbecco's Modified Eagle's Medium (Sigma Chemical Co. St. Louis, MO), supplemented with 10% FBS and an antibiotic-antimycotic (100×) solution (Invitrogen), and incubated for 10-14 days at 37°C in a humidified 5% CO_2 atmosphere. Subconfluent adherent cells were harvested in PBS containing 0.06% trypsin and 0.005% EDTA, and were subcultured. The culture medium was replaced every 4 days. This study was approved by the Ethical Committee of the National Institute for Child Health and Development. The purpose of this study was explained thoroughly to the patients, who gave their written informed consent.

hiPSCs were cultivated on irradiated MEFs in iPSellon medium (Cardio, Osaka, Japan), supplemented with 10 ng/ml recombinant human bFGF (Wako Pure Chemicals, Osaka, Japan). hiPSCs were established from MRC-5 and AM cells, as previously described [21,22]. In addition, hiPSCs were established from PAE and UtE cells in the present study. Briefly, 1×10^5 cells were infected overnight with pooled viral supernatants, obtained by the transfection of HEK293FT cells (TransIT-293 reagent; Mirus, Madison, WI) with the retroviral vector pMXs, which encodes the cDNAs for OCT3/4, SOX2, KLF4, and c-MYC, together with the packaging plasmids pCLGagPol and pHCMV-VEV-G (a gift from T. Kiyono, National Cancer Center Research Institute, Tokyo, Japan). Four days after infection, the cells were split, plated on irradiated MEFs in 100-mm dishes, and maintained in iPSellon medium until colonies formed.

The immunocytochemical analysis was performed as described previously [22,23]. Human cells were fixed with 4% paraformaldehyde in PBS for 10 min at 4 °C. After washing with 0.1% Triton X-100 in PBS (PBST), the cells were prehybridized in blocking buffer for 1–12 h at 4 °C, and then incubated for 6–12 h at 4°C with the following primary antibodies: anti-SSEA4 (1 : 300 dilution; Chemicon, Temecula, CA), anti-TRA-1–60 (1 : 300; Chemicon), anti-Oct4 (1 : 50; Santa Cruz Biotechnology, Santa Cruz, CA), anti-Nanog (1 : 300; ReproCELL, Tokyo, Japan), and anti-Sox2 (1 : 300; Chemicon). The cells were then incubated with anti-rabbit IgG, anti-mouse IgG or anti-mouse IgM conjugated with Alexa Fluor 488 or Alexa Fluor 546 (1: 500; Molecular Probes, Eugene, OR) in blocking buffer for 1 h at room temperature. The cells were counterstained with DAPI, and then mounted using a SlowFade light antifade kit (Molecular Probes).

Teratoma formation was performed as described previously [22,23]. The 1:1 mixtures of the AM-hiPSC suspension and Basement Membrane Matrix (BD Biosciences, San Jose, CA) were implanted subcutaneously, at 1.0×10^7 cells / site, into immunodeficient,

non-obese diabetic (NOD)/severe combined immunodeficiency (SCID) mice (CREA, Tokyo, Japan). Teratomas were surgically dissected out 6–10 weeks after implantation, and were fixed with 4% paraformaldehyde in PBS and embedded in paraffin. Sections of 10-μm thickness were stained with hematoxylin-eosin.

Gene expression analysis

Total RNA samples were extracted using ISOGEN (NipponGene). The global gene expression patterns and changes in mRNA levels were monitored using Agilent Whole Human Genome Microarray chips (G4112F) with one-color (Cyanine 3) dye. This microarray chip covers 41,000 well-characterized human genes and transcripts. The raw microarray data were submitted to the GEO (Gene Expression Omnibus) microarray data archive (http://www.ncbi.nlm.nih.gov/geo/) at the NCBI (accession number: GSE 20750). After background correction using a Normal plus Exponential convolution model, which adjusts the foreground to the background, we used an offset to dampen the variation of the log-ratios for intensities close to zero.

Among the 41,000 probes, 16,483 representative probes corresponding to MAQC unique genes were used for the following analyses [44]. Global array clustering was performed by the complete linkage method with Euclidean distance, and was visualized using the Java TreeView 1.1.0 software; the gene expression values are displayed as normalized log ratios. Cell line similarities were measured using Pearson correlation coefficients. To further validate whether the global gene expression is different in each origin cell, we evaluated the classification accuracy by leave-one-out cross-validation (LOOCV) on the nearest-neighbor classifier, based on Pearson's correlation distance. To obtain the expression signatures, we performed a differential analysis for each origin cell: differences between the two arbitrary datasets were evaluated by the Student's t-test for the expression of each gene. Thereafter, the false discovery rate (FDR) was estimated using the Benjamini–Hochberg procedure. Differentially expressed genes were selected if they satisfied both FDR <0.05 and a 2.0-fold change in the average values for the cell lines being compared. The gene ontology analysis was performed using the GO Term Finder Perl script [45] (http://go.princeton.edu/cgi-bin/GOTermFinder), with EBI human GO annotations and generic GO slim annotations (http://www.geneontology.org/).

Network screening

Network screening was performed as described previously [21]. This analysis is based on the procedure for estimating the consistency of a network structure (directed acyclic graph) with the measured data for the constituent variables in the graph. The joint density function for a given network (reference network) was recursively factorized into conditional density functions, according to the parent-child relationship in the graph. The conditional functions were quantified into log-likelihoods, using linear regression for the measured data, with the assumption that the data followed a normal distribution. The probability of the log-likelihood for the network structure (graph consistency probability; GCP) was then estimated from the distribution of log-likelihoods for 2,000 networks, generated under the condition that the networks shared the same numbers of nodes and edges as those of the given network. The significance probability of the given network was set at 0.05 in this analysis.

In the present study, the GCP was estimated for the ensemble of reference networks, to extract the candidate activated networks in the hiPSCs, in a process termed 'network screening'. The reference networks were constructed using the ChIP-on-Chip data and the classification scheme for gene function. The genes bound by four factors were cited from a previous report [20], and were divided into sub-networks according to the functional gene sets previously defined in the Molecular Signatures Database (MSigDB) [24]. The sub-networks that included at least one gene of the expression signature were then selected. The set of selected sub-networks was used as the reference network for network screening.

Glycan analysis

We analyzed cell surface glycans with a lectin microarray [31]. The 43 lectins were dissolved at a concentration of 0.5 mg/ml in spotting solution (Matsunami Glass, Osaka, Japan), and were spotted onto epoxysilane-coated glass slides (Nexterion Slide E Epoxysilane-coated Substrate 25 × 75.6 × 1 mm; Schott, Mainz, Germany) attached to a silicone rubber sheet, using a non-contact microarray printing robot (MicroSys 4000; Genomic Solutions, Ann Arbor, MI). The lectins were spotted in triplicate, with a spot diameter of 500 μm. The glass slides were incubated at 25°C for 3 h, to allow lectin immobilization. The lectin-immobilized glass slides were then washed with probing buffer (25 mM Tris-HCl [pH 7.5], 140 mM NaCl, 2.7 mM KCl, 1 mM CaCl2, 1 mM MnCl2, 1% [v/v] Triton X-100), and incubated with the blocking reagent N102 (NOF, Tokyo, Japan) at 20°C for 1 h. Finally, the lectin-immobilized glass slides were flooded with TBS containing 0.1% NaN_3 and stored at 4°C. The cell membrane faction was prepared using the CelLytic MEM Protein Extraction Kit (Sigma-Aldrich, Tokyo, Japan), and the protein concentration was determined using the MicroBCA Protein Assay Reagent kit (Thermo Fisher Scientific, Waltham,

MA). After dilution in PBST (10 mM PBS [pH 7.4], 140 mM NaCl, 2.7 mM KCl, 1% Triton X-100), the cell membrane fraction was labeled with Cy3 NHS ester (GE Healthcare Ltd., Buckinghamshire, England). After dilution in probing buffer to the desired concentration, the Cy3-labeled cell membrane fraction was applied to the lectin microarray and incubated at 20°C overnight. After washing with the probing buffer, fluorescence images were acquired using an evanescent-field activated fluorescence scanner (SC-Profiler; GP BioScience, Kanagawa, Japan). The fluorescence signal of each spot was quantified using the Array Pro Analyzer ver. 4.5 software (Media Cybernetics, Bethesda, MD), and the background value was subtracted. The values shown for the lectin signals represent the average of triplicate spots.

Additional files

There are 14 additional files in the present analysis. For convenience, we provide an overview of the additional files. Additional files 1, 2, 3, 4, 5 are related to the cell classification in Figure 1: the details of the cell lines and their experimental establishment are described in files 1 and 2, and the details of the analyses of the expression data are described in files 3-5. Additional figures 6-8 are related to the gene expression signature in Figure 2: the details of the analyzed data are described in files 6 and 7, and the results obtained by a standard analysis are described in file 8. Additional files 9, 10, 11 are related to the network signature in Figure 3: the methodological aspects of the network screening are described in files 9 and 10, and in file 11, the detailed results are presented. Additional files 12, 13, 14 are related to the glycan signature in Figure 5: all of the information for interpreting the analyzed results is presented in the three files.

Additional material

Additional file 1: Cell lines and numbers of passages analyzed in the present study. The following abbreviations are used for the human somatic cell (SC) and induced pluripotent stem cell (hiPSC) sources: AM, amniotic membrane; PAE, placental artery endothelial; UtE, uterine endometrium; and MRC, MRC-5 cell line. The AM and MRC cell lines were named previously [22,23]. The number of passages for each cell line is indicated by the letter 'p' followed by an Arabic number.

Additional file 2: Generation of iPSCs from human PAE cells. (*A*) PAE cells from the arterial endothelium of a human placenta (a), and generation of hiPSCs through epigenetic reprogramming by retrovirus infection-mediated expression of OCT4, SOX2, KLF4, and c-MYC (b). (*B*) Expression patterns of the pluripotent cell markers, TRA-1-60, SSEA-4, NANOG, OCT3/4, and SOX2. The cell nuclei were stained with DAPI. (*C*) Hematoxylin-eosin staining of sections of teratomas generated by PAE-hiPSC implantation. The histological examination revealed that the tumors contain neural tissues (a: ectoderm), cartilage (b: mesoderm), and a gut-like epithelial tissue (c: endoderm).

Additional file 3: Clustering for all cells by another method. Another clustering was performed by the WARD method, instead of the complete linkage method of Figure 1, with Euclidean distance, and was visualized using the Java TreeView 1.1.0 software. The gene expression values are displayed as normalized log ratios. The abbreviations used are the same as those listed in Figure 1 and additional file 1.

Additional file 4: Correlation coefficient matrix for all cells. Pearson's correlation coefficients between 51 cells for the expression profiles of all genes were calculated. The abbreviations used are the same as those listed in Figure 1 and additional file 1.

Additional file 5: Cross-validation of cell classification. The classification accuracy was evaluated by leave-one-out cross-validation (LOOCV) on the nearest-neighbor classifier, based on the Pearson's correlation distance.

Additional file 6: Number matrix for common genes. The numbers of genes that were different between the iPSCs and SCs are listed on the diagonal of the matrix, and those that were shared between the four gene sets that showed expression differences between the iPSCs are listed above the diagonal. The abbreviations used are the same as those listed in Figure 1.

Additional file 7: List of 2,502 genes in the expression signature, together with the fold-changes in expression levels and FDR values. The fold-change values are listed for the minimum values among the four sets of comparisons between iPSCs and SCs (+, iPSCs>SCs; -, iPSCs<SCs), and the FDR values shown are the maximum values among these sets.

Additional file 8: List of enriched GO terms with significant probabilities (FWER < 0.05).

Additional file 9: Schematic representation of the procedure used to obtain the network signature. The procedure for obtaining the network signature from the expression signature is shown schematically. The detailed procedure is as follows: 1) We first prepare the information for the gene sets to which the transcriptional factors bind, as deduced from the ChIP-on-chip experiments [20]; 2) Next, we prepare the information for the gene sets that were classified using knowledge of biological functions [24]; 3) The large gene sets in step 1 are divided into smaller subsets, according to the classification scheme of the gene sets in step 2; 4) If at least one gene in the expression signature is included in each gene subset in step 3, then the subset is regarded as a reference network; 5) In each reference network, the enrichment probability of the genes in the expression signature is tested with a significance probability of 0.05. Thus, we narrow down the network signature from the reference networks, in terms of gene numbers; 6) The significant reference networks identified in step 5 are further tested by calculating the graph consistency probability, which assesses the consistency between the network structure and the expression data for the constituent genes [24]. In this step, we further refine the network signature, in terms of both the network structure and the extent of gene expression; 7) Finally, we define the network signature, using the reference networks that passed the tests in steps 5 and 6.

Additional file 10: Reference networks and constituent genes.

Additional file 11: Details of the network signature. The characters in the above list are colored, according to the classification of biological function shown in Figure 2A.

Additional file 12: Lectin-glycan-glycosyltransferase relationships and correlations of lectin array intensities with glycosyltransferase expression patterns. Lectins with FDR<0.05 are colored red. The glycosyltransferases in the expression signature are indicated by a circle in the column "Expression signature" in "Gene expression". The Pearson's correlation coefficients between the lectin signal intensities and the expression profiles of the corresponding glycosyltransferases are listed, together with the significance probabilities. The original lectin array data can be obtained by request to HT or JH.

Additional file 13: Knowledge-based relationships between glycosyltransferases and their biosynthetic pathways.

Additional file 14: Locations of the glycosyltransferases detected in the present study in the pathways of "Glycan Biosynthesis and Metabolism". The glycosyltransferases listed in Table 1 were allocated to the pathways in "1.7 Glycan Biosynthesis and Metabolism" of the KEGG GLYCAN program (http://www.genome.jp/kegg/pathway.html#glycan). The glycosyltransferases and epitopes related to differentiation are indicated by red-colored boxes and red lines, respectively, in each pathway (see the text for details).

Acknowledgements
We thank H. Abe, M. Yamazaki–Inoue, M. Machida, and H. Sakaguchi for providing expert technical assistance. This work was supported by the project grant entitled 'Development of Analysis Technology for Induced Pluripotent Stem (iPS) Cell', from The New Energy and Industrial Technology Development Organization (NEDO).
This article has been published as part of *BMC Systems Biology* Volume 5 Supplement 1, 2011: Selected articles from the 4th International Conference on Computational Systems Biology (ISB 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1752-0509/5?issue=S1.

Author details
[1]Computational Biology Research Center, National Institute of Advanced Industrial Science Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan. [2]INFOCOM CORPORATION, Sumitomo Fudosan Harajuku Building, 2-34-17, Jingumae, Shibuya-ku, Tokyo, 150-0001, Japan. [3]Research Center for Stem Cell Engineering, National Institute of Advanced Industrial Science Technology (AIST), Tsukuba Central 4, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8562, Japan. [4]Research Center for Medical Glycoscience, National Institute of Advanced Industrial Science Technology (AIST), Tsukuba Central 2, 1-1-1 Umezono, Ibaraki 305-8568, Japan. [5]Department of Reproductive Biology, National Research Institute for Child Health and Development, 2-10-1 Ookura, Setagaya-ku, Tokyo 157-8535, Japan. [6]Department of Developmental Biology and Pathology, National Research Institute for Child Health and Development, 2-10-1 Okura, Setagaya-ku, Tokyo 157-8535, Japan. [7]Institute for Systems Biology, Shanghai University, Shangda Road 99, Shanghai 200444, China. [8]Department of Life Sciences (Biology), Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan.

Authors' contributions
S.S. (computational analysis and manuscript preparation), Y.O. (cell experiments and DNA microarray), Y.I. (DNA microarray and manuscript preparation), H.T. (lectin microarray and manuscript preparation), M.T. (cell experiments and manuscript preparation), H.A. (cell experiments), K.N. (cell experiments), E.C. (cell experiments), Y.F. (cell experiments), Y.M. (vector construction), H.O. (vector construction), N.K. (vector construction), Y.S. (lectin microarray), A.U. (supervision of cell experiments), J.H. (lectin microarray), K.H. (computational analysis and manuscript preparation), and M.A. (project leader and coordination).

Competing interests
The authors declare that they have no competing interests.

Published: 20 June 2011

References

1. Takahashi K, *et al*: **Induction of pluripotent stem cells from adult human fibroblasts by defined factors.** *Cell* 2007, **131**:861-872.
2. Muramatsu T, *et al*: **Carbohydrate antigens expressed on stem cells and early embryonic cells.** *Glycocon J* 2004, **21**:41-45.
3. Schopperle WM, *et al*: **The Tra-1-60 and Tra-1-81 human pluripotent stem cell markers are expressed on podocaluxin in embryonal carcinoma.** *Stem Cells* 2006, **25**:723-730.
4. Natunen S, *et al*: **The binding specificity of the marker antibodies Tra-1-60 and Tra-1-81 reveals a novel pluripotency associated type 1 lactosamine epitope.** *Glycobiology* 2010.
5. Satomaa T, *et al*: **The N-glycome of human embryonic stem cells.** *BMC Cell Biol* 2009, **2**:10.
6. Toyoda M, *et al*: **Lectin microarray analysis of pluripotent and multipotent stem cells.** *Genes to Cells* 2010, **16**:1-11.
7. Chen T, *et al*: **E-cadherin-Mediated Cell-Cell Contact is Critical for Induced Pluripotent Stem Cell Generation.** *Stem Cells* 2010, **28**:1315-25.
8. Tchieu J, *et al*: **Female human iPSCs retain an inactive X chromosome.** *Cell Stem Cell* 2010, **7**:329-42.
9. Chin MH, *et al*: **Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures.** *Cell Stem Cell* 2009, **5**:111-123.
10. Guenther MG, *et al*: **Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells.** *Cell Stem Cell* 2010, **7**:249-257.
11. Newman AM, *et al*: **Lab-specific gene expression signatures in pluripotent stem cells.** *Cell Stem Cell* 2010, **7**:258-262.
12. Chin MH, *et al*: **Molecular analyses of human induced pluripotent stem cells and embryonic stem cells.** *Cell Stem Cell* 2010, **7**:263-9.
13. Lowry WE, *et al*: **Generation of human induced pluripotent stem cells from dermal fibroblasts.** *Proc Natl Acad Sci USA* 2008, **105**:2883-2888.
14. Maherali N, *et al*: **A high-efficiency system for the generation and study of human induced pluripotent stem cells.** *Cell Stem Cell* 2008, **3**:340-345.
15. Yu J, *et al*: **Human induced pluripotent stem cells free of vector and transgene sequences.** *Science* 2009, **324**:797-801.
16. Boyer LA, *et al*: **Core transcriptional regulatory circuitry in human embryonic stem cells.** *Cell* 2005, **122**:947-56.
17. Loh YH, *et al*: **The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells.** *Nat Genet* 2006, **38**:431-40.
18. Chen X, *et al*: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.** *Cell* 2008, **133**:1106-17.
19. Kim J, *et al*: **An extended transcriptional network for pluripotency of embryonic stem cells.** *Cell* 2008, **132**:1049-61.
20. Sridharan R, *et al*: **Role of the murine reprogramming factors in the induction of pluripotency.** *Cell* 2009, **136**:364-77.
21. Saito S, Aburatani S, Horimoto K: **Network evaluation from the consistency of the graph structure with the measured data.** *BMC Sys Biol* 2008, **2**:84.
22. Makino H, *et al*: **Mesenchymal to embryonic incomplete transition of human cells by chimeric OCT4/3 (POU5F1) with physiological co-activator EWS.** *Exp Cell Res* 2009, **288**:2727-2740.
23. Nagata S, *et al*: **Efficient reprogramming of human and mouse primary extra-embryonic cells to pluripotent stem cells.** *Genes Cells* 2009, **14**:1395-1404.
24. Subramanian A, *et al*: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102**:15545-15550.
25. Lin T, *et al*: **A chemical platform for improved induction of human iPSCs.** *Nat Methods* 2009, **6**:805-808.
26. Sumi T, Tsuneyoshi N, Nakatsuji N, Suemori H: **Defining early lineage specification of human embryonic stem cells by the orchestrated balance of canonical Wnt/beta-catenin, Activin/Nodal and BMP signaling.** *Development* 2008, **135**:2969-2979.
27. Eiselleova L, *et al*: **A complex role for FGF-2 in self-renewal, survival, and adhesion of human embryonic stem cells.** *Stem Cells* 2009, **27**:1847-1857.
28. Discher DE, Mooney DJ, Zandstra PW: **Growth Factors, matrices, and forces combine and control stem cells.** *Science* 2009, **26**:1673-1677.
29. Yamanaka S: **Elite and stochastic models for induced pluripotent stem cell generation.** *Nature* 2009, **460**:49-52.
30. Sasaki N, *et al*: **Heparan sulfate regulates self-renewal and pluripotency of embryonic stem cells.** *J Biol Chem* 2008, **283**:3594-606.
31. Kuno A, *et al*: **Evanescent-field fluorescence-assisted lectin microarray: a new strategy for glycan profiling.** *Nat. Methods* 2005, **2**:851-856.
32. Gabius H-J: **Glycans: bioactive signals decoded by lectins.** *Biochem Soc Trans* 2008, **36**:1491-1496.
33. Hashimoto K, *et al*: **Comprehensive analysis of glycosyltransferases in eukaryotic genomes for structural and functional characterization of glycans.** *Carbohydr Res* 2009, **12**:881-887.
34. Shevinsky L, Knowles BB, Damjanov I, Solter D: **Monoclonal antibody to murine embryos defines a stage-specific embryonic antigen expressed on mouse embryos and human teratocarcinoma cells.** *Cell* 1982, **30**:697-705.
35. Kannagi R, *et al*: **Stage-specific embryonic antigens (SSEA-3 and -4) are epitopes of a unique globo-series ganglioside isolated from human teratocarcinoma cells.** *EMBO J* 1983, **2**:2355-2361.
36. Gooi HC, *et al*: **Stage-specific embryonic antigen involves a1□3 fucosylated type 2 blood group chains.** *Nature* 1981, **292**:156-158.
37. Bert JE, *et al*: **The HB-6, CDw75, and CD76 differentiation antigens are unique cell-surface carbohydrate determinants generated by the ß-galactoside α2,6-sialyltransferase.** *J Cell Biol* 1992, **116**:423-435.
38. Chin MH, *et al*: **Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures.** *Cell Stem Cell* 2009, **5**:111-123.
39. Marchetto MC, *et al*: **Transcriptional signature and memory retention of human-induced pluripotent stem cells.** *PLoS One* 2009, **4(9)**:e7076.

40. Chin MH, Pellegrini M, Plath K, Lowry WE: **Molecular Analyses of Human Induced Pluripo-tent Stem Cells and Embryonic Stem Cells.** *Cell Stem Cell* 2010, **7**:263-269.
41. Newman AM, Cooper JB: **Lab-Specific Gene Expression Signatures in Pluripotent Stem Cells.** *Cell Stem Cell* 2010, **7**:258-262.
42. Guenther MG, *et al*: **Chromatin Structure and Gene Expression Programs of Human Embryonic and Induced Pluripotent Stem Cells.** *Cell Stem Cell* 2010, **7**:249-257.
43. Ghosh Z, *et al*: **Persistent Donor Cell Gene Expression among Human Induced Pluripotent Stem Cells Contributes to Differences with Human Embryonic Stem Cells.** *PLoS One* 2010, **5(2)**:e8975.
44. MAQC Consortium: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.** *Nat Biotechnol* 2006, **24**:1151-1161.
45. Boyle EI, *et al*: **GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.** *Bioinformatics* 2004, **20**:3710-3715.

doi:10.1186/1752-0509-5-S1-S17
Cite this article as: Saito *et al.*: Possible linkages between the inner and outer cellular states of human induced pluripotent stem cells. *BMC Systems Biology* 2011 **5**(Suppl 1):S17.

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 286, NO. 23, pp. 20345–20353, June 10, 2011

Glycome Diagnosis of Human Induced Pluripotent Stem Cells Using Lectin Microarray*[S]

Received for publication, February 15, 2011, and in revised form, March 28, 2011 Published, JBC Papers in Press, April 6, 2011, DOI 10.1074/jbc.M111.231274

Hiroaki Tateno[‡], Masashi Toyota[§], Shigeru Saito[¶‖], Yasuko Onuma[**], Yuzuru Ito[**], Keiko Hiemori[‡], Mihoko Fukumura[‡], Asako Matsushima[‡], Mio Nakanishi[‡‡], Kiyoshi Ohnuma[‡‡], Hidenori Akutsu[§], Akihiro Umezawa[§], Katsuhisa Horimoto[¶], Jun Hirabayashi[‡1], and Makoto Asashima[‡‡]**

From the [‡]Research Center for Medical Glycoscience, National Institute of Advanced Industrial Science and Technology, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan, the [§]Department of Reproductive Biology, National Research Institute for Child Health and Development, 2-10-1 Okura, Setagaya-ku, Tokyo 157-8535, Japan, the [¶]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan, [‖]Infocom Corporation, Sumitomo Fudosan Harajuku Building, 2-34-17, Jingumae, Shibuya-ku, Tokyo 150-0001, Japan, the [**]Research Center for Stem Cell Engineering, National Institute of Advanced Industrial Science and Technology, Tsukuba Central 4, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8562, Japan, and the [‡‡]Department of Life Sciences (Biology), Graduate School of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Induced pluripotent stem cells (iPSCs) can now be produced from various somatic cell (SC) lines by ectopic expression of the four transcription factors. Although the procedure has been demonstrated to induce global change in gene and microRNA expressions and even epigenetic modification, it remains largely unknown how this transcription factor-induced reprogramming affects the total glycan repertoire expressed on the cells. Here we performed a comprehensive glycan analysis using 114 types of human iPSCs generated from five different SCs and compared their glycomes with those of human embryonic stem cells (ESCs; nine cell types) using a high density lectin microarray. In unsupervised cluster analysis of the results obtained by lectin microarray, both undifferentiated iPSCs and ESCs were clustered as one large group. However, they were clearly separated from the group of differentiated SCs, whereas all of the four SCs had apparently distinct glycome profiles from one another, demonstrating that SCs with originally distinct glycan profiles have acquired those similar to ESCs upon induction of pluripotency. Thirty-eight lectins discriminating between SCs and iPSCs/ESCs were statistically selected, and characteristic features of the pluripotent state were then obtained at the level of the cellular glycome. The expression profiles of relevant glycosyltransferase genes agreed well with the results obtained by lectin microarray. Among the 38 lectins, rBC2LCN was found to detect only undifferentiated iPSCs/ESCs and not differentiated SCs. Hence, the high density lectin microarray has proved to be valid for not only comprehensive analysis of glycans but also diagnosis of stem cells under the concept of the cellular glycome.

Increasing attention has been paid to iPSCs[2] and ESCs in their pluripotency and medical applications (1, 2). However, establishment of a robust evaluation system of their properties, including differentiation propensity and risk of possible contamination of xenoantigens and even potential of tumorigenesis, has been hampered by the lack of comprehensive methodology directly applicable to target stem cells, although this is an emerging issue essential for the safe use of iPSCs in regenerative medicine. From many aspects, cell surface glycans are considered to be ideal targets for analyzing or identifying the phenotype of each cell in a direct manner by the following reasons (3, 4). (*a*) Glycans are located at the outermost cell surface. (*b*) The total repertoire of cell surface glycans varies at every level of biological organization (*i.e.* species, tissues, cell types, and molecules). (*c*) Global alterations of the cellular glycome also occur during development, cellular activation, differentiation, malignant transformation, and inflammation. The cell surface glycans are therefore referred to as the "cell signature" that closely reflects cellular backgrounds and conditions, probably because they are actually functioning as cell-to-cell mediators in extensive biological phenomena. This fundamental nature of glycans should be understood with the fact that they are not encoded directly in the genome but are generated by a complex system of a number of glycosidases and glycosyltransferases, whose expressions and activities are significantly affected by both intracellular and extracellular environmental changes. Indeed, cell surface molecules, such as stage-specific embryonic antigens (SSEA1 and -3/4) (5) and tumor rejection antigens (Tra-1-60 and Tra-1-81) (6–8) are glycobiomarkers widely used to evaluate pluripotency. Notably, however, these "representative" glycomarkers have been identified following rather fortuitous development of their specific antibodies, because most carbohydrate structures are poorly antigenic between mammals. In this context, a systematic search is necessary to draw a whole picture of the stem cell glycome and harness its effect on stem cell biology (8, 9). For instance, the growth and directed differentiation of stem cells to specific progeny lineages in cell culture remain problematic. Understanding how stem cells

* This work was supported by the New Energy and Industrial Technology Development Organization in Japan.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Tables S1–S5 and Figs. S1–S4.

[1] To whom correspondence should be addressed: AIST, Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan. Tel.: 81-29-861-3124; Fax: 81-29-861-3125; E-mail: jun-hirabayashi@aist.go.jp.

[2] The abbreviations used are: iPSC, induced pluripotent stem cell; AM, amniotic mesodermal; ESC, embryonic stem cell; MEF, mouse embryonic fibroblast; PAE, placental artery endothelial; SC, somatic cell; UtE, uterine endometrium; rMOA, recombinant MOA; FWER, familywise error rate.

communicate with one another and feeder cells through cell surface glycans may lead to rational design of specific culture systems. However, the glycome is a quite difficult target to predict solely based on any genomic data base because the biosynthetic process of the glycan moieties of glycoproteins is not template-driven and is subject to multiple sequential and competitive enzymatic pathways. In this sense, a rapid and sensitive system enabling direct monitoring of cell surface glycans is essential.

Several methods have been developed for glycan analysis based on physicochemical principles, such as liquid chromatography and mass spectrometry (10–12). Lectin microarray is an alternative technology for structural glycomics, where a panel of lectins with various glycan-binding specificities is printed on a microarray, providing a versatile platform for rapid and high throughput analysis of glycan structures without liberation of glycans (13, 14). Lectins are a class of decoder molecules of cell surface glycans distributed throughout organisms, which mediate various functions through specific glycan recognition. Analytical protocols using lectin microarray have been developed for various sample types: free oligosaccharides (14, 15), tissue sections (16), cell membrane hydrophobic fractions (17, 18), and even whole cells (19, 20). This technology has just begun to be applied to a wide variety of biological researches, including virus profiling (21) and cell profiling (17, 20), and development of cancer glycobiomarkers (22–24). For cell profiling, less than 100 ng of proteins in hydrophobic fractions are sufficient for each analysis (25, 26). Data processing and normalization procedures were optimized to ensure the proper interpretation of the data (25, 26). More recently, we have demonstrated that lectin microarray is also applicable to stem cells (27, 28), although we have yet to reach a clear conclusion as to how the cellular glycome changes upon induction of pluripotency. Moreover, practical applicability of this technology to the quality control of stem cells has not been attained.

Here, we developed an advanced platform of high density lectin microarray with the increased number of probe lectins (96 lectins) to expand the glycome coverage for more precise comparison of various stem cell glycomes. A systematic survey of the cellular glycome was then performed toward 135 cell types in total, including iPSCs (114 cell types) and ESCs (nine cell types). Through this comprehensive analysis, we obtained strong evidence that all of the four SCs with originally distinct glycan profiles have acquired those similar to ESCs upon induction of pluripotency. We also found structural features common to iPSCs and ESCs, which corresponded well to the results of gene expression analysis of glycosyltransferases. Finally, we demonstrate the applicability of lectin microarray in the stem cell diagnosis of multiple factors, including discrimination between undifferentiated and differentiated cells as well as detection of the contamination of the xenoantigen, αGal epitope.

EXPERIMENTAL PROCEDURES

Cells—Endometrium (UtE) (29), placental artery endothelium (PAE) (30), and amnion (AM) (31, 32) were independently established. UtE, AM, and MRC5 cells were maintained in the POWEREDBY10 medium (MED SHIROTORI CO., Ltd.). PAE cells were cultured in EGM-2MV BulletKit (Lonza) containing 5% FBS. Human iPSCs from UtE, PAE, and AM cells were generated according to the procedures described by Yamanaka and colleagues (1) with slight modification (27, 28, 33). The iPSCs derived from MRC5, UtE, PAE, and AM cells were maintained in iPSellon medium (Cardio Inc.) supplemented with 10 ng/ml recombinant human basic fibroblast growth factor (Wako Pure Chemical Industries, Ltd.) on irradiated MEF feeder cells. hiPS201B7 and hiPS253G1 cells were maintained in DMEM/F-12 medium (Invitrogen) supplemented with 20% KSR (Invitrogen), 0.1 mM 2-mercaptoethanol (Sigma-Aldrich), minimum essential medium non-essential amino acids (Invitrogen), and 5–10 ng/ml recombinant human basic FGF (Wako) on mitomycin C-treated mouse embryo fibroblast feeder cells. ESCs were generated and maintained as described previously (34).

Lectins—Lectins from natural sources (58 lectins) were purchased from J-OIL MILLS, Vector Laboratories, EY Laboratories, and Seikagaku Corp. (see the lectin list in supplemental Table S2). Recombinant lectins were prepared as follows. Briefly, genes of carbohydrate recognition domains were cloned into pET27b (Stratagene) and were overexpressed in the *Escherichia coli* BL21-CodonPlus (DE3)-RIL strain under the control of isopropyl-β-D-thiogalactopyranoside (Fermentas Hanover) at appropriate temperatures. All recombinant lectins were purified by affinity chromatography using appropriate sugar-immobilized Sepharose 4B-CL (GE Healthcare) based on the glycan binding specificity of each lectin. They were then dialyzed against diluted PBS (final concentration 2.5 mM phosphate buffer containing 0.015 M NaCl). The protein concentration was determined by a BCA protein assay (Bio-Rad). Lectins were freeze-dried and stored at 4 °C until use. The purity was checked by SDS-PAGE and gel filtration chromatography on Shodex PROTEIN KW-802.5 (Shodex). The glycan binding activity and specificity were analyzed by hemagglutinating activity using 4% rabbit erythrocytes, frontal affinity chromatography (35), and glycoconjugate microarray (36).

Lectin Microarray Production—The lectin microarray was produced as described previously with minor modifications (14, 15). Fifty-eight natural lectins and 38 recombinant lectins (see supplemental Table S2 for a lectin list) were dissolved at a concentration of 0.5 mg/ml in a spotting solution (Matsunami Glass), and spotted onto epoxysilane-coated glass slides (Schott) in triplicate using a non-contact microarray printing robot (MicroSys4000, Genomic Solutions). The glass slides were then incubated at 25 °C overnight to allow lectin immobilization. The lectin-immobilized glass slides were then washed with probing buffer (25 mM Tris-HCl, pH 7.5, 140 mM NaCl (TBS) containing 2.7 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$, and 1% Triton X-100) and incubated with blocking reagent N102 (NOF Co.) at 20 °C for 1 h. Finally, the lectin-immobilized glass slides were washed with TBS containing 0.02% NaN_3 and stored at 4 °C until use. The spot quality and reproducibility of the produced microarrays were checked before use, using a Cy3-labeled test probe containing 250 μg/ml asialofetuin (Sigma-Aldrich), 25 ng/ml Siaα2–3Galβ1–4GlcNAc-BSA (Dextra), 10 ng/ml Fucα1–2Galβ1–3GlcNAcβ1–3Galβ1–4Glc-BSA (Dextra), 10 ng/ml βGlcNAc-BSA (Dextra), 10 ng/ml GalNAcα1–3(Fucα1–2)Gal-BSA (Dextra), 10 ng/ml Galα1–3Galβ1-