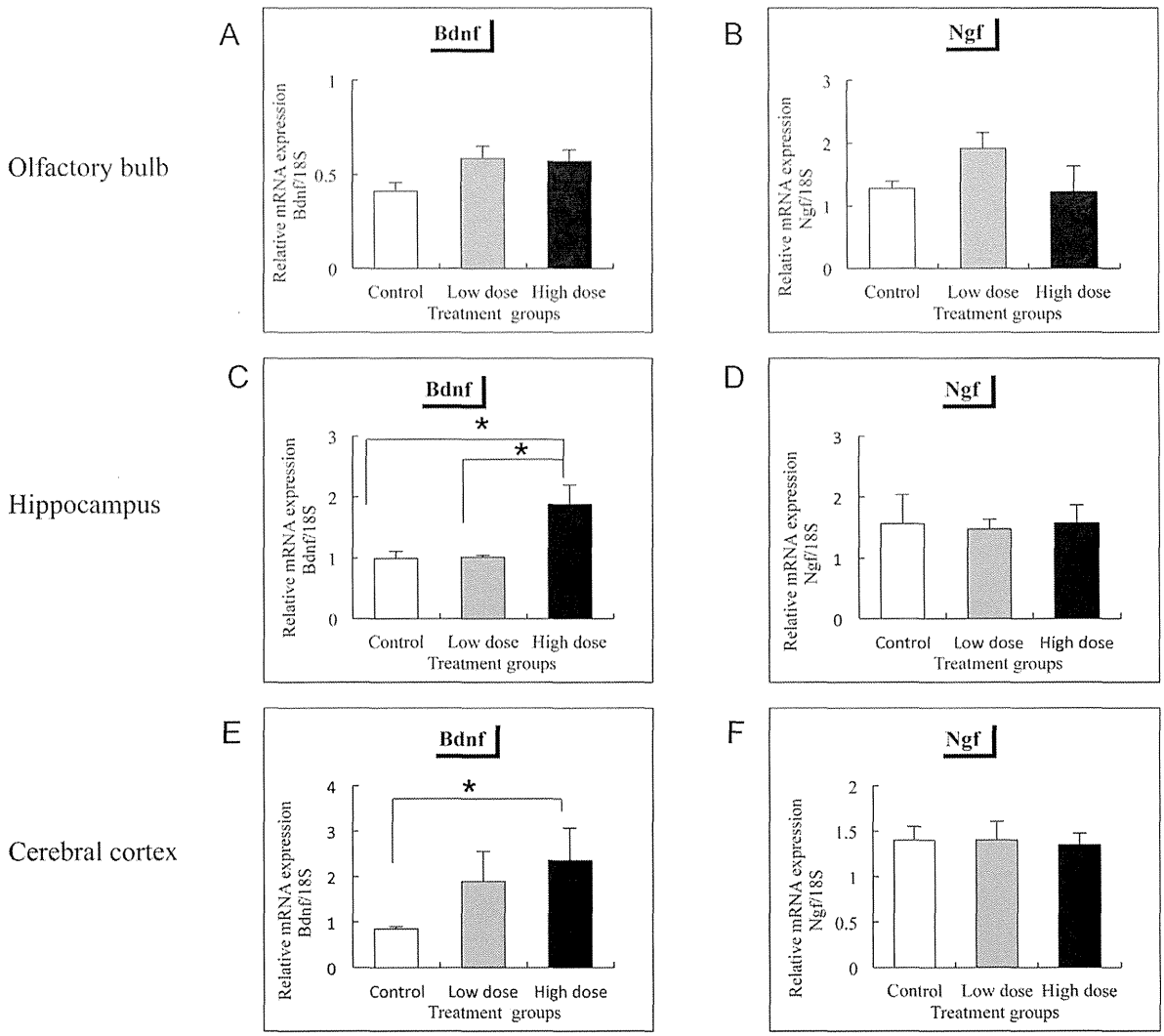


Fig. 5. mRNA expression of neurotrophins in dendrimer-treated mice (A–F). Each bar represents the mean ± SE ($n = 5$ from each group). *Abbreviations*: Brain-derived neurotrophic factor, *Bdnf*; nerve growth factor, *Ngf*. (A) and (B) show the expressions in the olfactory bulb; (C) and (D) show the expressions in the hippocampus; and (E) and (F) show the markers in the cerebral cortex.

by rats for 6 h in a whole-body exposure chamber led to a significant and persistent increase in the accumulation of ^{13}C NPs in the olfactory bulb on day 1 and that the NP concentration continued to increase up until day 7 (Oberdörster et al., 2004). The same study also showed that the concentrations of ^{13}C NPs were significantly increased in the cerebrum and cerebellum, but that the increase was inconsistent; that is, a significant difference was only observed on one additional day of the post-exposure period (day 1). Regarding the present study, we have no evidence that dendrimers translocate to the systemic circulation. However, Nemmar et al. (2002) reported that inhaled ultrafine technetium (^{99m}Tc)-labeled carbon particles translocate into the systemic circulation within 5 min by diffusion. The transport of nanoparticles across the blood–brain barrier (BBB) is reportedly possible by either passive diffusion or by carrier-mediated endocytosis (Hoet et al., 2004). A recent study using a BBB *in vitro* model showed that G4 PAMAM dendrimers were able to cross the BBB and induced CD11b and CCR2 overexpression in primary murine microglia (Bertero et al., 2014). In the present study, we observed fluorescent signals thought to represent dendrimers in the hippocampus. Thus, the intranasally instilled 4-nm PAMAM used in the present study might translocate to the brain *via* an olfactory nerve or the systemic circulation. The potential translocation routes for intranasally instilled dendrimers in the mouse brain are shown in Fig. 6. Although we could not identify the exact mechanism for the translocation of the dendrimers from the nose to the brain, drugs or molecules can reportedly be transported by two possible routes: a transporter-mediated route and paracellular transport (Frey, 2002; Thorne et al., 2004; Dhanda et al., 2005; Pardridge, 2005). Recently, an intranasal route *via* an epithelial permeabilizer has been shown to be capable of delivering drugs to the CNS (Krishan et al., 2014).

We have measured the size of PAMAM dendrimers in the original commercial solution [100% methanol (3.4 ± 0.9 nm)], in the ultrapure water (mean size is 5.7 ± 1.4 nm in first peak; 976 ± 391 nm in second peak) after removing methanol and after 24 h in the ultrapure water (5.6 ± 2.3 nm) to remove second peak of aggregation. Our results indicated that no much change of size and not much aggregation were observed before administration to animals.

Neurotrophins have been identified as targets for neurotoxicants and are known to play a role in bidirectional signaling between the cells of the immune and nervous systems. Bdnf is a primary neurotrophin in the hippocampus (Karaoulanis and Angelopoulos, 2010) and plays a key role in neuro-immune responses (Aloe et al., 1994). Moreover, multiple neurotransmitter systems, such as dopamine, glutamate, acetylcholine, and

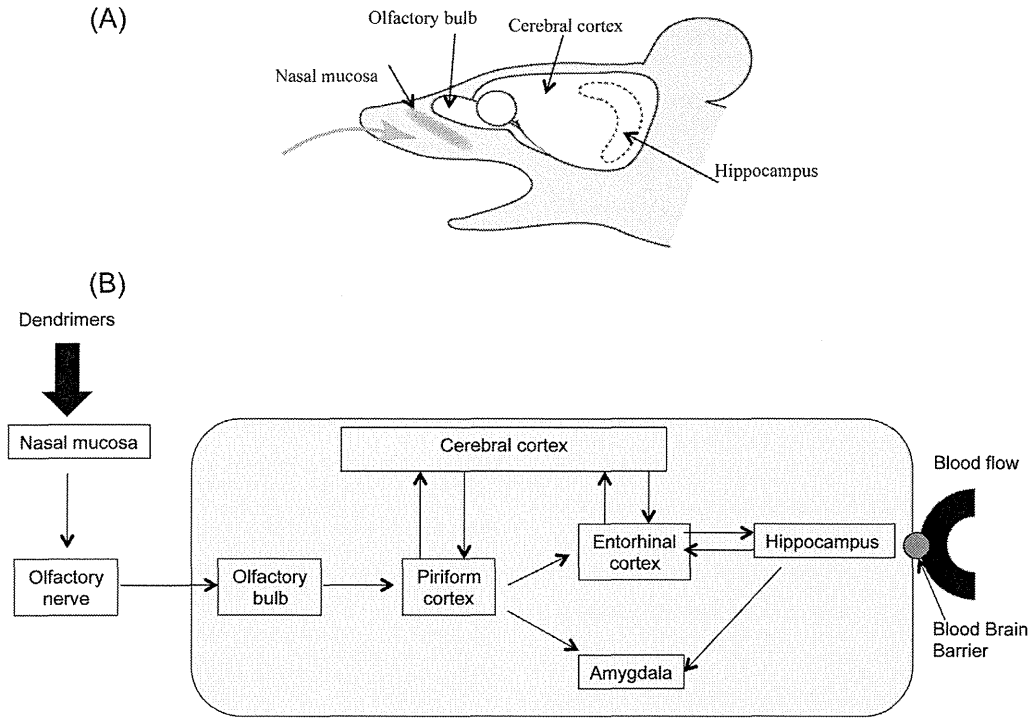


Fig. 6. Diagram showing the target brain regions of the mouse brain examined in the present study (A) and the potential translocation routes for intranasally instilled dendrimers in the mouse brain (B).

serotonin, also take part in proper neuronal functions. On the other hand, according to requirements, body homeostasis signals induce neurotrophin production in the brain. The balance between toxic substances and trophic substances secreted from microglia may influence neurotoxicity and neuroprotection. Finally, our present study indicates that dendrimers induce minor toxicity in the brain but that BDNF production compensates for this toxicity.

Based on our present findings, we suggest that PAMAM dendrimers appeared to be not toxic in general; however, the expressions of some neurological-related genes were induced by high-dose treatment. Although we observed the effects of dendrimers at one time point in the present study, time points that are relevant to the PAMAM pharmacokinetics should also be examined. The cytotoxicity and cell permeability of PAMAM dendrimers depend upon the concentration and generation of the dendrimers (Jevprasesphant et al., 2003). Furthermore, cytotoxicity may be related to the radius of gyration, the molecular shape, and the dimensions of a particular dendrimer (Metullio et al., 2004). A recent report indicated that dendrimer toxicity was mainly due to its outer surface layer and that the toxicity could be manipulated by modifying the surface layer (Chauhan et al., 2010). We suggest that the route of administration (*e.g.*, intranasal, intravenous, intratracheal, or intraperitoneal), dosage (low or high dose), and duration of treatment (acute, subchronic, chronic, or intermittent) may also influence the toxicological and pharmacological effects of dendrimers *in vivo*. Further studies are needed to explore the time course effects of dendrimers on biodistribution and the effects of various exposure routes and durations on dendrimer toxicity.

Conflict of interest

The authors declare that there are no conflicts of interest.

Transparency document

The Transparency document associated with this article can be found in the online version.

Acknowledgment

This work was funded by a by Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Culture and Sports of Japan (24241013) to H.S. We thank Ms. Hiroko Nansai, Ryoko Yanagisawa and Naoko Ueki for their technical assistance.

References

Albertazzi, L., Gherardini, L., Brondi, M., Sulis, Sato, S., Bifone, A., Pizzorusso, T., Ratto, G.M., Bardi, G., 2013. In vivo distribution and toxicity of PAMAM dendrimers in the central nervous system depend on their surface chemistry. Mol. Pharm. 10, 249–260.

Aloe, L., Skaper, S.D., Leon, A., Levi-Montalcini, R., 1994. Nerve growth factor and autoimmune diseases. Autoimmunity 19, 141–150.

Bertero, A., Boni, A., Gemmi, M., Gagliardi, M., Bifone, A., Bardi, G., 2014. Surface functionalisation regulates polyamidoamine dendrimer toxicity on blood–brain barrier cells and the modulation of key inflammatory receptors on microglia. Nanotoxicology 8, 158–168.

Buhleier, E., Wehner, W., Vogtle, F., 1978. Cascade and nonskid-chain-like synthesis of molecular cavity topologies. Synthesis 2, 155–158.

Chauhan, A.S., Jain, N.K., Diwan, P.V., 2010. Pre-clinical and behavioural toxicity profile of PAMAM dendrimers in mice. Proc R Soc A, http://dx.doi.org/10.1098/rspa.2009.0448.

Cho, W.S., Kang, B.C., Lee, J.K., Jeong, J., Che, J.H., Seok, S.H., 2013. Comparative absorption, distribution, and excretion of titanium dioxide and zinc oxide nanoparticles after repeated oral administration. Part Fibre Toxicol. 10, 9.

De Lorenzo, A.J.D., 1970. The olfactory neuron and the blood–brain barrier. In: Wolstenholme, G.E.W., Knight, J. (Eds.), Taste and Smell in Vertebrates. J. & A. Churchill, London, UK, pp. 151–176.

Dear, J.W., Kobayashi, H., Brechbiel, M.W., Star, R.A., 2006. Imaging acute renal failure with polyamine dendrimer-based MRI contrast agents. Nephron. Clin. Pract. 103, c45–c49.

Dhanda, D.S., Frey, I.I.W.H., Leopold, D., et al., 2005. Approaches for drug deposition in the human olfactory epithelium. Drug Del. Tech. 5, 64–72.

Elder, A., Gelein, R., Silva, V., Feikert, T., Opanashuk, L., Carter, J., Potter, R., Maynard, A., Ito, Y., Finkelstein, J., Oberdorster, G., 2006. Translocation of inhaled ultrafine manganese oxide particles to the central nervous system. Environ. Health Perspect. 114, 1172–1178.

Fattori, V.AbeS., Kobayashi, K., Costa, L.G., Tsuji, R., 2008. Effects of postnatal ethanol exposure on neurotrophic factors and signal transduction pathways in rat brain. J. Appl. Toxicol. 28, 370–376.

Frey, I.I.W.H., 2002. Bypassing the blood–brain barrier to delivery therapeutic agents to the brain and spinal cord. Drug Del. Tech. 2, 46–49.

Funk, J.A., Gohlke, J., Kraft, A.D., McPherson, C.A., Collins, J.B., Jean Harry, G., 2011. Voluntary exercise protects hippocampal neurons from trimethyltin injury: possible role of interleukin-6 to modulate tumor necrosis factor receptor-mediated neurotoxicity. Brain Behav. Immun. 25, 1063–1077.

Heiden, T.C., Dengler, E., Kao, W.J., Heideman, W., Peterson, R.E., 2007. Developmental toxicity of low generation PAMAM dendrimers in zebrafish. Toxicol. Appl. Pharmacol. 225, 70–79.

Henriksson, J., Tjälve, H., 2000. Manganese taken up into the CNS via the olfactory pathway in rats affects astrocytes. Toxicol. Sci. 55, 392–398.

Hoet, P.H., Brüske-Hohlfeld, I., Salata, O.V., 2004. Nanoparticles – known and unknown health risks. J. Nanobiotechnol. 2, 12.

Hong, S., Bielinska, A.U., Mecke, A., Keszler, B., Beals, J.L., Shi, X., Balogh, L., Orr, B.G., Baker Jr., J.R., Banaszak Holl, M.M., 2004. Interaction of poly(amidoamine) dendrimers with supported lipid bilayers and cells: hole formation and the relation to transport. Bioconjug. Chem. 15, 774–782.

Jevprasesphant, R., Penny, J., Jalal, R., Attwood, D., McKeown, N.B., D'Emanuele, A., 2003. The influence of surface modification on the cytotoxicity of PAMAM dendrimers. Int. J. Pharm. 252, 263–266.

Karaoulanis, S.E., Angelopoulos, N.V., 2010. The role of immune system in depression. Psychiatrike 21, 17–30.

Krishan, M., Gudelsky, G.A., Desai, P.B., Genter, M.B., 2014. Manipulation of olfactory tight junctions using papaverine to enhance intranasal delivery of gemcitabine to the brain. Drug Deliv. 21, 8–16.

Lee, C.C., MacKay, J.A., Fréchet, J.M., Szoka, F.C., 2005. Designing dendrimers for biological applications. Nat. Biotechnol. 23, 1517–1526.

Lee, J.H., Cha, K.E., Kim, M.S., Hong, H.W., Chung, D.J., Ryu, G., Myung, H., 2009. Nanosized polyamidoamine (PAMAM) dendrimer-induced apoptosis mediated by mitochondrial dysfunction. Toxicol. Lett. 190, 202–207.

Leroueil, P.R., Berry, S.A., Duthie, K., Han, G., Rotello, V.M., McNerny, D.Q., Baker Jr., J.R., Orr, B.G., Holl, M.M., 2008. Wide varieties of cationic nanoparticles induce defects in supported lipid bilayers. Nano. Lett. 8, 420–424.

Mukherjee, S.P., Davoren, M., Byrne, H.J., 2010a. In vitro mammalian cytotoxicological study of PAMAM dendrimers – towards quantitative structure activity relationships. Toxicol. In Vitro 24, 169–177.

Mukherjee, S.P., Lyng, F.M., Garcia, A., Davoren, M., Byrne, H.J., 2010b. Mechanistic studies of in vitro cytotoxicity of poly(amidoamine) dendrimers in mammalian cells. Toxicol. Appl. Pharmacol. 248, 259–268.

Metullio, L., Ferrone, M., Coslanich, A., Fuchs, S., Fermeglia, M., Paneni, M.S., Pricl, S., 2004. Polyamidoamine (Yet Not PAMAM) dendrimers as bioinspired materials for drug delivery: structure-activity relationships by molecular simulations. Biomacromolecules 5, 1371–1378.

Na, M., Yiyun, C., Tongwen, X., Yang, D., Xiaomin, W., Zhenwei, L., Zhichao, C., Guanyi, H., Yunyu, S., Longping, W., 2006. Dendrimers as potential drug carriers Part II. Prolonged delivery of ketoprofen by in vitro and in vivo studies. Eur. J. Med. Chem. 41, 670–674.

Naha, P.C., Davoren, M., Casey, A., Byrne, H.J., 2009. An ecotoxicological study of poly(amidoamine) dendrimers-toward quantitative structure activity relationships. Environ. Sci. Technol. 43, 6864–6869.

Naha, P.C., Davoren, M., Lyng, F.M., Byrne, H.J., 2010. Reactive oxygen species (ROS) induced cytokine production and cytotoxicity of PAMAM dendrimers in J774A.1 cells. Toxicol. Appl. Pharmacol. 246, 91–99.

Nemmar, A., Hoet, P.H., Vanquickenborne, B., Dinsdale, D., Thomeer, M., Hoylaerts, M.F., Vanbilloen, H., Mortelmans, L., Nemery, B., 2002. Passage of inhaled particles into the blood circulation in humans. Circulation 105, 411–414.

Newkome, G.R., Yao, Z.Q., Baker, G.R., Gupta, V.K., 1985. Cascade molecules: a new approach to micelles A [27]-arborol. J. Org. Chem. 50, 2003–2006.

Oberdörster, G., Sharp, Z., Atudorei, V., Elder, A., Gelein, R., Kreyling, W., Cox, C., 2004. Translocation of inhaled ultrafine particles to the brain. Inhal. Toxicol. 16, 437–445.

Paino, I.M., Marangoni, V.S., de Oliveira Rde, C., Antunes, L.M., Zucolotto, V., 2012. Cyto and genotoxicity of gold nanoparticles in human hepatocellular carcinoma and peripheral blood mononuclear cells. Toxicol. Lett. 215, 119–125.

Pardridge, W.M., 2005. The blood–brain barrier and neurotherapeutics. NeuroRx 2, 1–2.

Persson, E., Henriksson, J., Tjälve, H., 2003. Uptake of cobalt from the nasal mucosa into the brain via olfactory pathways in rats. Toxicol. Lett. 145, 19–27.

Pushkar, S., Philip, A., Pathak, K., Pathak, D., 2006. Dendrimers: nanotechnology derived novel polymers in drug delivery. Indian J. Pharm. Educ. Res. 40, 153–158.

Swanson, S.D., Kukowska-Latallo, J.F., Patri, A.K., Chen, C., Ge, S., Cao, Z., Kotlyar, A., East, A.T., Baker, J.R., 2008. Targeted gadolinium-loaded dendrimer nanoparticles for tumor-specific magnetic resonance contrast enhancement. Int. J. Nanomed. 3, 201–210.

Tallkvist, J., Henriksson, J., d'Argy, R., Tjälve, H., 1998. Transport and subcellular distribution of nickel in the olfactory system of pikes and rats. Toxicol. Sci. 43, 196–203.

Tekade, R.K., Kumar, P.V., Jain, N.K., 2009. Dendrimers in oncology: an expanding horizon. Chem. Rev. 109, 49–87.

Thorne, R.G., Pronk, G.J., Padmanabhan, V., Frey, W.H.I.I., 2004. Delivery of insulin-like growth factor-I to the rat brain and spinal cord along olfactory and trigeminal pathways following intranasal administration. Neuroscience 127, 481–496.

Tjälve, H., Henriksson, J., Tallkvist, J., Larsson, B.S., Lindquist, N.G., 1996. Uptake of manganese and cadmium from the nasal mucosa into the central nervous system via olfactory pathways in rats. Pharmacol. Toxicol. 79, 347–356.

Tjälve, H., Henriksson, J., 1999. Uptake of metals in the brain via olfactory pathways. Neurotoxicology 20, 181–196.

Tomalia, D.A., Baker, H., Dewald, J., Hall, M., Kallos, G., Martin, S., Roeck, J., Ryder, J., Smith, P., 1985. A new class of polymers: starburst-dendritic macromolecules. Polym. J. 17, 117–132.

Win-Shwe, T.T., Yamamoto, S., Ahmed, S., Kakeyama, M., Kobayashi, T., Fujimaki, H., 2006. Brain cytokine and chemokine mRNA expression in mice induced by intranasal instillation with ultrafine carbon black. Toxicol. Lett. 163, 153–160.

Win-Shwe, T.T., Yamamoto, S., Fujitani, Y., Hirano, S., Fujimaki, H., 2008a. Spatial learning and memory function-related gene expression in the hippocampus of mouse exposed to nanoparticle-rich diesel exhaust. Neurotoxicology 29, 940–947.

Win-Shwe, T.T., Mitsushima, D., Yamamoto, S., Fukushima, A., Funabashi, T., Kobayashi, T., Fujimaki, H., 2008b. Changes in neurotransmitter levels and proinflammatory cytokine mRNA expressions in the mice olfactory bulb following nanoparticle exposure. Toxicol. Appl. Pharmacol. 15, 192–198.

Win-Shwe, T.T., Mitsushima, D., Yamamoto, S., Fujitani, Y., Funabashi, T., Hirano, S., Fujimaki, H., 2009. Extracellular glutamate level and NMDA receptor subunit expression in mouse olfactory bulb following nanoparticle-rich diesel exhaust exposure. Inhal. Toxicol. 21, 828–836.

Win-Shwe, T.T., Fujimaki, H., 2011. Nanoparticles and neurotoxicity. Int. J. Mol. Sci. 12, 6267–6280.

Win-Shwe, T.T., Yamamoto, S., Fujitani, Y., Hirano, S., Fujimaki, H., 2012a. Nanoparticle-rich diesel exhaust affects hippocampal-dependent spatial learning and NMDA receptor subunit expression in female mice. Nanotoxicology 6, 543–553.

Win-Shwe, T.T., Fujimaki, H., Fujitani, Y., Hirano, S., 2012b. Novel object recognition ability in female mice following exposure to nanoparticle-rich diesel exhaust. Toxicol. Appl. Pharmacol. 262, 355–362.

Ze, Y., Zheng, L., Zhao, X., Gui, S., Sang, X., Su, J., Guan, N., Zhu, L., Sheng, L., Hu, R., Cheng, J., Cheng, Z., Sun, Q., Wang, L., Hong, F., 2013. Molecular mechanism of titanium dioxide nanoparticles-induced oxidative injury in the brain of mice. Chemosphere 92, 1183–1189.

Zhou, J., Wu, J., Hafdi, N., Behr, J.P., Erbacher, P., Peng, L., 2006. PAMAM dendrimers for efficient siRNA delivery and potent gene silencing. Chem. Commun. (Camb). 22, 2362–2364.

OPEN ACCESS Freely available online

PLOS ONE

Tracking Difference in Gene Expression in a Time-Course Experiment Using Gene Set Enrichment Analysis

Pui Shan Wong[1]*, Michihiro Tanaka[2], Yoshihiko Sunaga[3,4], Masayoshi Tanaka[3], Takeaki Taniguchi[5], Tomoko Yoshino[3,4], Tsuyoshi Tanaka[3,4], Wataru Fujibuchi[1,2], Sachiyo Aburatani[1]

1 CBRC, National Institute of AIST, Tokyo, Japan, **2** Center for iPS Research and Application, Kyoto University, Kyoto, Japan, **3** Institute of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan, **4** JST, CREST, Sanbancho 5, Chiyoda-ku, Tokyo, Japan, **5** Mitsubishi Research Institute, Inc., Tokyo, Japan

Abstract

Fistulifera sp. strain JPCC DA0580 is a newly sequenced pennate diatom that is capable of simultaneously growing and accumulating lipids. This is a unique trait, not found in other related microalgae so far. It is able to accumulate between 40 to 60% of its cell weight in lipids, making it a strong candidate for the production of biofuel. To investigate this characteristic, we used RNA-Seq data gathered at four different times while *Fistulifera* sp. strain JPCC DA0580 was grown in oil accumulating and non-oil accumulating conditions. We then adapted gene set enrichment analysis (GSEA) to investigate the relationship between the difference in gene expression of 7,822 genes and metabolic functions in our data. We utilized information in the KEGG pathway database to create the gene sets and changed GSEA to use re-sampling so that data from the different time points could be included in the analysis. Our GSEA method identified photosynthesis, lipid synthesis and amino acid synthesis related pathways as processes that play a significant role in oil production and growth in *Fistulifera* sp. strain JPCC DA0580. In addition to GSEA, we visualized the results by creating a network of compounds and reactions, and plotted the expression data on top of the network. This made existing graph algorithms available to us which we then used to calculate a path that metabolizes glucose into triacylglycerol (TAG) in the smallest number of steps. By visualizing the data this way, we observed a separate up-regulation of genes at different times instead of a concerted response. We also identified two metabolic paths that used less reactions than the one shown in KEGG and showed that the reactions were up-regulated during the experiment. The combination of analysis and visualization methods successfully analyzed time-course data, identified important metabolic pathways and provided new hypotheses for further research.

Citation: Wong PS, Tanaka M, Sunaga Y, Tanaka M, Taniguchi T, et al. (2014) Tracking Difference in Gene Expression in a Time-Course Experiment Using Gene Set Enrichment Analysis. PLoS ONE 9(9): e107629. doi:10.1371/journal.pone.0107629

Editor: Cynthia Gibas, University of North Carolina at Charlotte, United States of America

Received November 7, 2013; **Accepted** August 21, 2014; **Published** September 30, 2014

Funding: This study was supported by JST-CREST. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Takeaki Taniguchi is employed by Mitsubishi Research Institute, Inc. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

* Email: shan.wong@aist.go.jp

Introduction

The search for sustainable and environmentally-friendly fuel is a burgeoning field in biology because organic waste products and organisms are abundant and renewable sources of biofuel compounds. There is strong focus on producing biofuel from food crops, such as corn and soy, as well as oleaginous algae, such as *Chlamydomonas reinhardtii* and *Nannochloropsis oceanica*. One of the big advantages of algae over terrestrial crops is that they require less land to grow on while producing more biomass [1]. This characteristic is important in large-scale production to minimize competition with the production of food or with the preservation of neighboring habitats. Algae can be farmed in open tanks or closed columns and does not deplete soil for agricultural use. Most oleaginous algae accumulate biofuel compounds in low nitrogen conditions at the expense of cell growth [2] [3] [4]. For that reason, we have focused our analysis on a newly sequenced strain of microalgae, *Fistulifera* sp. strain JPCC DA0580, which is able to accumulate lipids while undergoing logarithmic growth [5]. *Fistulifera* sp. strain JPCC DA0580 is a pennate diatom that is possibly an allodiploid, sharing many of its genes with the diatoms, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. It demonstrates a high growth rate concurrently with achieving high lipid content (40–60% w/w) [6]. There have been 20,618 genes sequenced from the nuclear, chloroplast and mitochondrion genomes. Although the *Fistulifera* sp. strain JPCC DA0580 genome contains some genes that are homologous to the ones involved in lipid metabolism, the cellular mechanisms for its ability to simultaneously grow and accumulate lipids is unknown.

In our analysis, we utilized RNA-Sequencing (RNA-Seq) data gathered from *Fistulifera* sp. strain JPCC DA0580 while it was grown in oil accumulating and non-accumulating conditions at four time points, from 0 to 60 hours. RNA-Seq is a high-throughput sequencing method that produces a large amount of data per experiment and can be used to investigate differences in gene expression between several conditions. The method produces count data of RNA sequences which can be normalized using Reads Per Kilobase Per Million (RPKM). The normalization corrects for the varying coverage a sequence may get due to its length. Most analyses that involve comparisons in gene expression focus on identifying differentially expressed genes, especially methods that use linear modeling which take advantage of preexisting microarray analyses [7] [8]. Another type of method

that is less stringent is gene set enrichment analysis (GSEA), which is more focused on relating the results with previous knowledge. GSEA approaches the data analysis by looking for associations between predefined groups of genes, a gene set, and a phenotype of interest. This type of method is better at detecting small but coordinated differences in gene expression than linear modeling and is less interested in differentially expressed genes and more focused on a group of genes being expressed differently from the background expression. GSEA generally has simple requirements for the data to be analyzed. The important elements are sets of genes that can be compared to the data and data values that can be distilled into one value per gene, usually gene expression or fold change. This makes GSEA more suitable for analysing our data.

There are a variety of GSEA tools available for analyzing high-throughput sequencing data from experiments investigating two conditions with a robust number of replicates on a model organism [9]. For example, online services such as DAVID [10] [11], FuncAssociate [12] and GOEAST [13], statistical packages for R such as SPIA [14] and standalone scripts such as PAGE [15]. Unfortunately, our data was not suitable for these methods. When investigating multiple time points with a new organism, it is sometimes not feasible to have enough replicates, even with the decreasing cost of RNA-Seq experiments. There are some methods that can accommodate these data but they still depend on variance estimation which is inadequate for our data. Therefore, we proposed a new approach to analyse data from a new organism that takes into account the change in gene expression through time in order to avoid reducing our data as done by some existing tools.

We demonstrate a modified approach to GSEA that is able to analyse one sampled data with multiple time points, and custom annotations in an investigation on the difference in gene expression between two conditions through four time points. We then use the results to identify a sequence of reactions starting with a compound such as glucose, and ending with a compound of interest such as triacylglycerol. To create gene sets for a genome with custom annotations, we associate our genes with known KEGG pathways and make each metabolic pathway a gene set. In order to fully utilize the time-course data, each time point is treated as a variable so that GSEA is performed in multiple dimensions, and gene expression variation across time can be conserved. We use re-sampling to address the low replicate number issue and create an empirical cumulative distribution that is then used to calculate the enrichment p-value on multidimensional data without the need to assume multivariate normality. Finally, we visualize and interpret the results using graphs that join the enriched gene sets. The graphs also let us calculate a hypothesized pathway of reactions from one compound to another. In the interest of learning about oil accumulation, we chose to focus our demonstration on the reactions involved in turning glucose into the target biofuel lipid, triacylglycerol (TAG).

Results and Discussion

Gene Set Enrichment Analysis

Using the modified GSEA method on our data, we identified 9 significantly enriched pathways (Table 1). These pathways contain genes whose difference in gene expression was significantly different, as a group, to the general background level of gene expression of the whole data set.

The photosynthesis and photosynthesis antenna protein pathways were two related pathways that were significantly enriched with p-values <0.0001. The gene expression in the photosynthesis pathway showed a positive relationship between log fold change and time, indicating that there was increased energy synthesis via photosynthesis during oil accumulation. Although a similar relationship was present in the photosynthesis antenna proteins pathway, the log fold change values at 60 hours was higher than in the photosynthesis pathway. Further investigation reveals that the values came from the expression of light-harvesting complex I chlorophyll a/b binding proteins; LHCA1, LHCA2 and LHCA4. Additionally, the general difference in expression of proteins in light-harvesting complex II is lower than in light-harvesting complex I. The preference of light-harvesting complex I may be due to the highly efficient nature of photosystem I [16] even though *Fistulifera* sp. strain JPCC DA0580 is using both systems simultaneously in this case.

The other prominent pathways are related to cellular energy metabolism; glycolysis, the pentose phosphate pathway and oxidative phosphorylation were significantly enriched in our analysis. The glycolysis and pentose phosphate pathways are fundamental to the conversion of glucose to fatty acids while oxidative phosphorylation is essential for providing the energy needed to power metabolic reactions. Some of the proteins in the oxidative phosphorylation pathway form the membrane protein V-type ATPase. It is a proton pump responsible for ATP turnover in mitochondria and was up-regulated in our data. There is some evidence of a relationship between increased C16-C18 length fatty acids, which are used in TAG production, and increased hydrolytic activity of V-ATPase [17]. Along with a gradual down-regulation of NADH dehydrogenase, it would seem that *Fistulifera* sp. strain JPCC DA0580 focuses on recycling ATP instead of reducing NADP+ for its energy requirements during oil accumulation. Predictably, most glycolysis genes were up-regulated during the experiment, although there were notable exceptions; phosphoglucomutase (PGM), phosphoglycerate kinase (PGK) and glyceraldehyde 3-phosphate dehydrogenase (GAPDH). PGM transfers a phosphate group to and from the 1' position to the 6' position in α-D-glucose so its down-regulation suggests that *Fistulifera* sp. strain JPCC DA0580 is getting its source of α-D-glucose 6-phosphate elsewhere. PGK and GAPDH are used in two reversible reactions to make glycerate 3-phosphate which is a key molecule for TAG production [18]. However, this reaction can be done in one irreversible step by glyceraldehyde-3-phosphate dehydrogenase (NADP) which was up-regulated in our data. The substrate for that reaction, glyceraldehyde 3-phosphate, is used in the pentose phosphate shunt to make nucleic and amino acids like deoxyribose, 2-Deoxy-D-ribose 1-phosphate and D-ribulose 5-phosphate. The genes involved in those reactions were found to be up- regulated in our data; they were ribokinase (rbsK), phosphopentomutase (PGM2), 6-phosphogluconate dehydrogenase (PGD) and 3-hexulose-6-phosphate synthase (hxlA). So it seems that *Fistulifera* sp. strain JPCC DA0580 relies on glucose to produce TAG, and nucleic and amino acids to achieve accumulation and growth at the same time while using a proton pump to power the reactions under low nitrogen conditions.

The other significant pathways are related to synthesizing the materials for TAG and growth; they are fatty acid biosynthesis and amino sugar and nucleotide sugar metabolism. Expectedly, the difference in gene expression in fatty acid biosynthesis shows a general up-regulation of the genes in the pathway as *Fistulifera* sp. strain JPCC DA0580 accumulates TAG and continues cell growth. Gene expression in the amino sugar and nucleotide sugar metabolism pathway also had a positive trend through time. The up-regulation of genes in this pathway suggests that sugars are being metabolised for growth during oil accumulation. Two of the up-regulated genes are glucokinase (glk) and glucose-6-phosphate isomerase (GPI) which are involved in reversible reactions that

Table 1. Results of GSEA Method.

Pathway Name	P-value
Photosynthesis	0*
Photosynthesis - antenna proteins	0*
Pentose phosphate pathway	0*
Carbon fixation in photosynthetic organisms	0*
Fatty acid biosynthesis	0*
Amino sugar and nucleotide sugar metabolism	0.013
Methane metabolism 00680	0.013
Oxidative phosphorylation	0.026
Glycolysis	0.026

The enriched pathways identified using GSEA and their enriched p-values. There were 9 pathways enriched out of 39 pathways tested.
*P-value <0.0001.
doi:10.1371/journal.pone.0107629.t001

convert glucose into fructose and eventually lead to the production of nucleotide sugars. As the reactions are reversible, we are unable to discern whether the forward or backward reaction was dominant without further data but their up-regulation means that there was a considerable amount of converting occurring.

The next significantly enriched pathway, carbon fixation in photosynthetic organisms, has several genes that are also present in pyruvate metabolism, glycolysis and the pentose phosphate pathway. The genes that exhibit varied differences in gene expression are the ones associated with pyruvate metabolism. During the experiment, malate dehydrogenase (decarboxylating) up-regulated the reaction that turns malate into pyruvate. In contrast malate dehydrogenase (oxaloacetate-decarboxylating) was down-regulated. The preference for the decarboxylating reaction could be due to the reactant, NADP, being used in other reactions, such as photosynthesis. Notably, the pyruvate metabolism pathway was not significantly enriched as a gene set however it only shares seven reactions with the carbon fixation in photosynthetic organisms pathway and is directly linked to 13 other pathways. It is likely that the process of oil accumulation uses the reactions in the carbon fixation pathway as a whole, instead of pyruvate specifically.

The remaining significantly enriched pathway was unexpectedly the methane pathway. Upon further investigation, it was discovered that many genes expressed in the methane pathway were also expressed in other pathways. For example, both glycolaldehyde dehydrogenase (ALDA) and 6-phosphofructokinase 1 (pfkA) are in the pentose phosphate pathway while (2R)-3-sulfolactate dehydrogenase (comC) is also found in the cystein and methionine metabolism pathway where it takes part in reactions that make pyruvate. The overlap of genes between gene sets can cause problems with detection, especially if some of the genes has a particularly strong signal. In this case, the genes in the pentose phosphate pathway have strongly defined differences in gene expression that may be masking the difference in gene expression of other genes. Although it is fairly reasonable for some genes to be present in multiple pathways, it should be checked if the overlapping genes are making biased contributions. The effect is further amplified in our data as the number of annotated genes are few.

Enriched Pathway Plots

To better visualize the results from GSEA, we plotted the enriched pathways as graphs (Figure 1). The graph's nodes were set up as compounds as we wanted to focus on compounds and reactions instead of the usual approach using genes. As such, the glycerolipid pathway was added so that the key compound, TAG, was included. The graph consisted of 353 compounds and 661 reactions. Most compounds were unique to their pathway but there were 18 compounds that were found in two pathways and 13 compounds that were found in three pathways. These included pyruvate, oxaloacetate and ADP and were found in glycolysis, pentose phosphate metabolism and other related processes.

Once the graph was constructed, the shortest path between glucose and TAG was calculated. As the graph was created using pathways that showed a significant relationship with oil accumulation, it can be considered a hypothesized path of metabolic reactions that metabolises glucose to produce TAG. We found two shortest paths with a length of 11 compounds (Figures 2 and 3); the conventional path found in KEGG contains 15 compounds. Our two shortest paths were very similar to each other, mainly differing between the use of glycerol or glycerone. Although it is possible to produce TAG in a smaller number of steps, it is unknown where the reactions take place in the cell. If the proteins are located close to each other, the path that was identified could be how *Fistulifera* sp. strain JPCC DA0580 produces TAG from glucose. Future experiments on metabolite quantity could also provide adequate evidence for the hypothesis.

In the final step, we showed that the genes along the hypothesized paths were up-regulated by plotting the direction of the difference in gene expression on the edges of the graph. When viewed next to each other, the direction of the difference in gene expression at each time point shows which reactions change from up-regulation to down-regulation and vice versa (Figure 4). We observed that genes along the identified shortest paths were up-regulated during the 60 hours of the experiment. However, the up-regulation occurs in sections along the path instead of being concerted. This suggests that the gene expression of a phenotype does not change for every gene along the reaction path at a single time point. Instead, the change in gene expression occurs in sections which eventually leads to the up-regulation of the full path. This visual presentation also brings to attention the possibility of time lag effects where there could be little difference in expression in earlier time points and not others. As our method does not address this issue directly, the testing may be underpowered at detecting true signals. The testing could be improved by applying a restriction on the difference in fold change between time points or restricting time points to those where fold

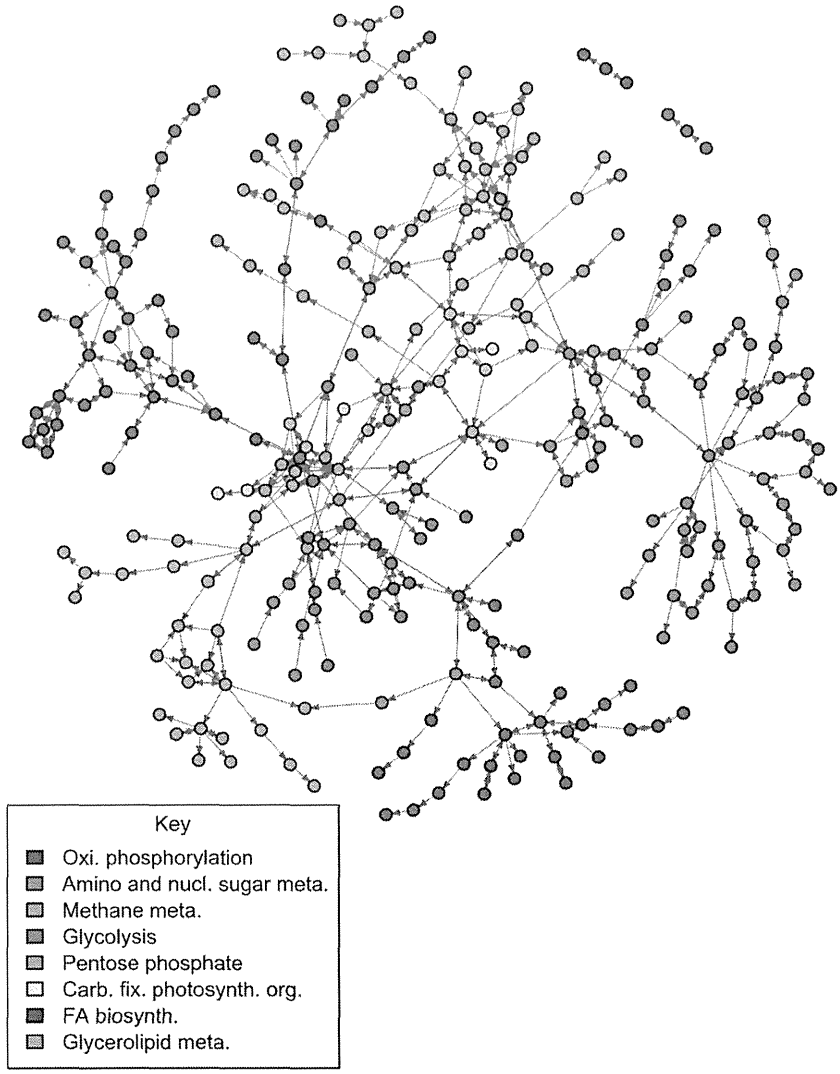


Figure 1. The graph of the significantly enriched pathways found using our GSEA method combined with the glycerolipid pathway. The full network contains 307 compounds and 558 reactions but compounds without reaction data were not drawn to reduce clutter. The graph is plotted with compounds drawn as nodes and reactions drawn as edges. The compounds are colored by their pathway membership; compounds belonging to 2 or more pathways are a mixture of the pathway colors. There were 7 compounds belonging to three pathways, 15 compounds belonging to two pathways and 117 compounds that were unique to their pathway. Many of the shared compounds are concentrated in the center of the graph and are related to glycolysis and pentose phosphate metabolism.
doi:10.1371/journal.pone.0107629.g001

change differences exist. However, this would require more knowledge about the organism than we currently have available.

Conclusion

GSEA is a useful tool for exploring data when there is a preconceived area of interest such as oil accumulation for our data. The way it can be used to analyse data more broadly is a big advantage when the data set is limited. As the cost of high-throughput sequencing experiments is decreasing, investigations with new organisms and time-course experiments can be utilized more often. For our expression data, we wanted to include time as a variable in our analysis so we modified GSEA to use it instead of removing it by averaging them. Although the number of replicates in our data caused issues with accurately isolating experimental and biological effects, we were still able to extract meaningful information through our use of resampling and GSEA. Being able to keep the time variable is an important step for future investigations. Drawbacks observed during our analysis included overlapping elements between gene sets, the reliance on pre-existing knowledge of our organism and as a consequence, the inability to assign meaning to unannotated data and improve our method's accuracy.

The results from GSEA were then graphed to produce a clear visualization of the results that is easier to interpret and grants access to other approaches for understanding the data. By plotting the direction of the difference in gene expression on our graph, we were able to observe the change in direction of the difference in gene expression as they occurred during the experiment. Using graphs in this way makes existing graph tools available, extending the investigation beyond the initial GSEA. In this analysis we looked at the shortest path of reactions between two compounds

but betweenness indexes can also be investigated to identify bottleneck compounds that are important in the network. These methods can be used to help generate hypotheses as a basis for further investigations.

Methods

Data preparation

The expression data was gathered from *Fistulifera* sp. strain JPCC DA0580 grown in two substrates; the treatment substrate was artificial sea water where oil accumulation took place, and the control substrate was a 10 fold dilution of the treatment substrate where oil was not accumulating [19]. The RNA-Seq data was obtained at four time points (0, 24, 48 and 60 hours) when *Fistulifera* sp. strain JPCC DA0580 was grown in the two substrates. Sequences with RPKM values of 0 for all time points were discarded leaving a remainder of 22,550 sequences. We used Ssearch with MIQS [20] to annotate the sequences so that 7,822 sequences were annotated with a KEGG Orthology identifier (K ID). The unannotated sequences either did not have a match in the KEGG database or the match did not have a KEGG Orthology identifier. The gene expression of the annotated sequences were then averaged if their matching K ID was shared among several sequences, by using the following equation

$$\mathbf{RPKM}_x = \frac{\Sigma_{\forall i}\mathbf{v}_i}{n} \quad (1)$$

where $\mathbf{RPKM}_x$ is a vector of RPKM values at each time point for K ID x, $\mathbf{v}_i$ is the i th vector of RPKM values for K ID x and n is the number of RPKM vectors with K ID x. For our data, this resulted in 2,873 $\mathbf{RPKM}_x$'s where each vector had a length of four that corresponded to the four time points, 0, 24, 48 and 60 hours.

As RNA-Seq data often have a disproportionate amount of small RPKM values, they are usually not normally distributed, even with the use of log transformation. The resulting fold changes calculated from them can follow the same non-normality. We corrected the RPKM values by implementing a threshold of 0.1 to minimize the influence of small read numbers [21]. This was done using the sRAP R package which also performed a log transform during the normalization process [22]. The normalized RPKM vectors, $\mathbf{sRAP}_x$, were then used to calculate the log fold change for each K ID x by the following equation

$$\mathbf{FC}_x = \mathbf{sRAP}_{x_{treatment}} - \mathbf{sRAP}_{x_{control}} \quad (2)$$

where $\mathbf{FC}_x$ is the log fold change vector of K ID x, $\mathbf{sRAP}_{x_{control}}$ is the vector of control RPKM values of K ID x and $\mathbf{sRAP}_{x_{treatment}}$ is the vector of treatment RPKM values of K ID x.

Gene Set Enrichment Analysis

We first established the gene sets which would be used in the analysis. Generally, gene sets are lists of gene identifiers that share an attribute of interest. For our analysis, these were K IDs divided into each metabolic pathway in the KEGG database. The

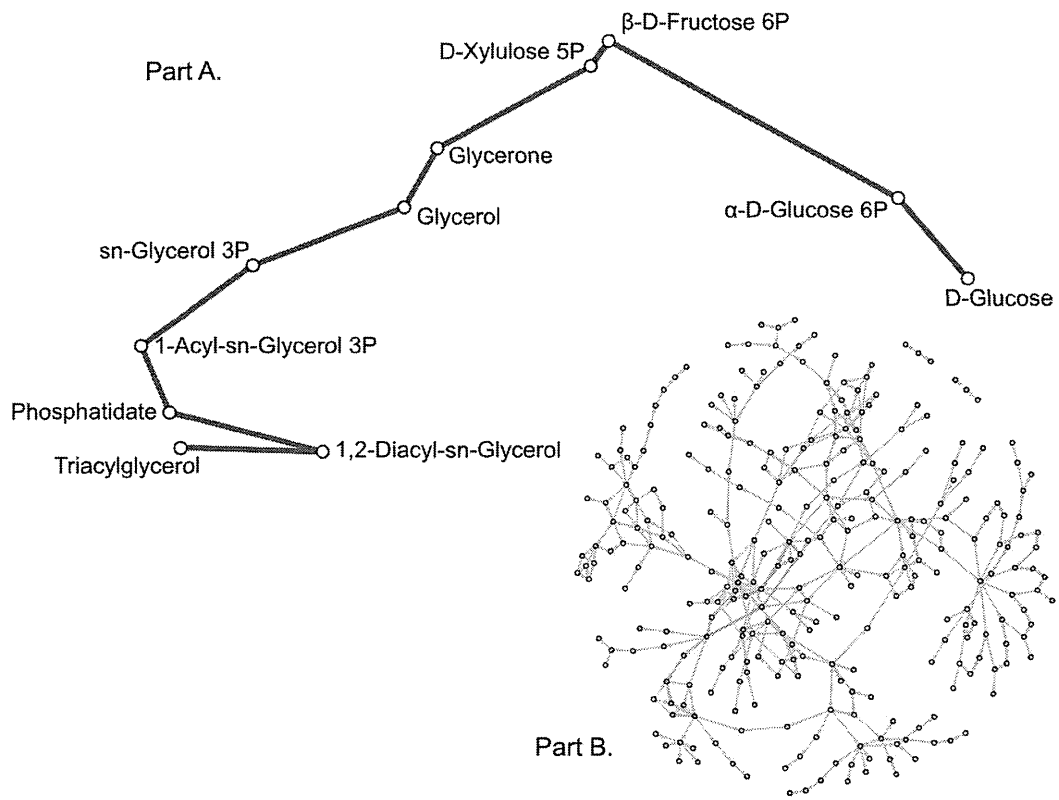


Figure 2. The first shortest path found in our graph between glucose and triacylglycerol using breadth-first search. A. This is the detailed view of the path showing the names of the compounds involved at each step. B. The shortest path is highlighted in green on the full graph to show its location. In contrast, the path presented in KEGG is highlighted in orange. The shortest path contains 11 compounds while the KEGG path contains 15 compounds.
doi:10.1371/journal.pone.0107629.g002

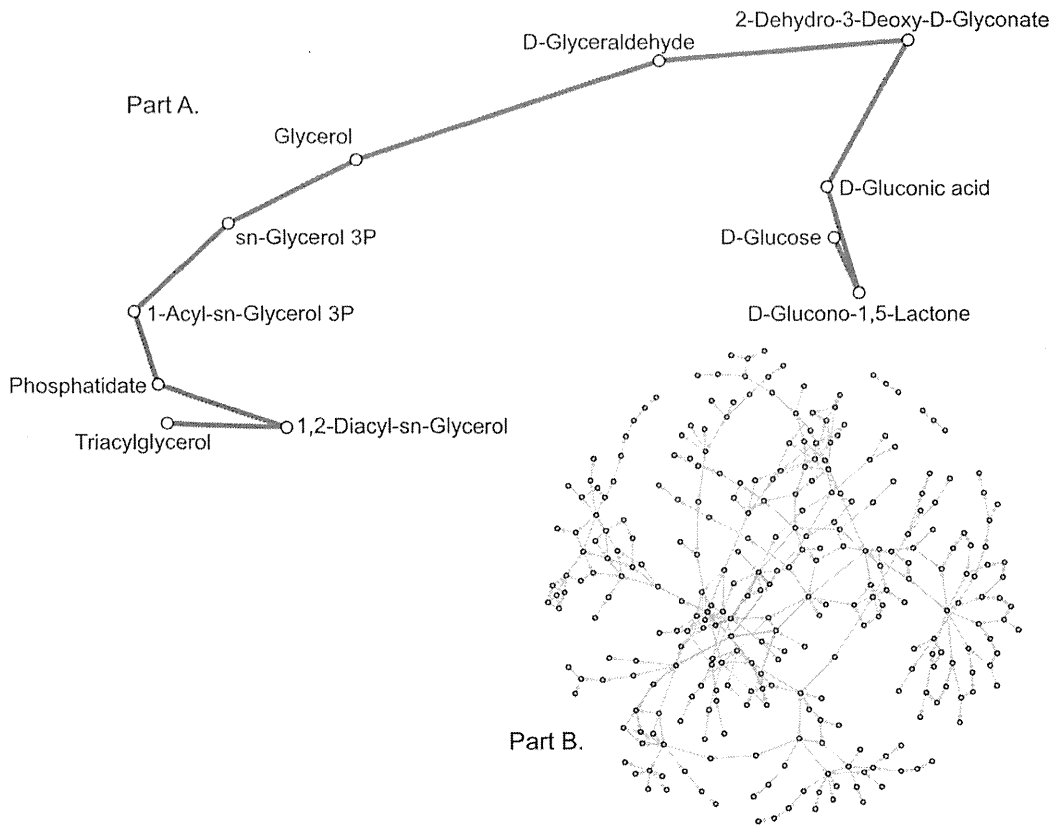


Figure 3. The second shortest path found in our graph between glucose and triacylglycerol using breadth-first search. A. This is the detailed view of the path showing the names of the compounds involved at each step. B. The shortest path is highlighted in green on the full graph to show its location. In contrast, the path presented in KEGG is highlighted in orange. The shortest path contains 11 compounds while the KEGG path contains 15 compounds.
doi:10.1371/journal.pone.0107629.g003

pathways we chose to investigate were associated with carbohydrate (15 pathways), energy (8 pathways) and lipid metabolism (17 pathways). The Secondary Bile Acid Biosynthesis gene set was removed as our data contained no data for it, thus our analysis used a total of 39 gene sets [23] [24]. Importantly, these 39 gene sets included the glycolysis and glycerolipid metabolic pathways which contains the compounds central to oil accumulation, glucose and TAG.

The following steps of the algorithm were carried out for each gene set which produces a test statistic and p-value that describes the significance of the gene expression of the gene set compared to the overal gene expression.

Step 1: Create a matrix of fold change data of genes present in gene set s.

$$\mathbf{FCM}_s = \begin{pmatrix} FC_{x,0} & FC_{x,24} & FC_{x,48} & FC_{x,60} \\ .. & .. & .. & .. \end{pmatrix} \quad (3)$$

where $\mathbf{FCM}_s$ is a n x 4 matrix, s denotes gene set s, n is the number of genes in the set and 4 is the number of time points in our data. Each row of $\mathbf{FCM}_s$ corresponds to a fold change vector $\mathbf{FC}_x$ (Equation 2). This vector consists of $FC_{x,t}$ which is the fold change of K ID x at time t. In our data, t takes a value from time point 0, 24, 48 or 60 (hours).

Step 2: Calculate the column mean of $\mathbf{FCM}_s$.

$$\overline{\mathbf{FCM}_s} = \left(\frac{\Sigma_{\forall i} FC_{i,0}}{n} \quad \frac{\Sigma_{\forall i} FC_{i,24}}{n} \quad \frac{\Sigma_{\forall i} FC_{i,48}}{n} \quad \frac{\Sigma_{\forall i} FC_{i,60}}{n} \right) \quad (4)$$

where $\overline{\mathbf{FCM}_s}$ is a column mean vector of matrix $\mathbf{FCM}_s$ (Equation 3). This is used to represent the fold change of gene set s through the 4 time points.

Step 3: Resample n rows from the whole fold change data matrix to construct a new matrix, $\mathbf{RSM}_i$. The resulting matrix, $\mathbf{RSM}_i$, is the ith matrix created from randomly resampling fold change vectors without replacement [25]. It has the same dimensions as $\mathbf{FCM}_s$ (Equation 3) but the rows of $\mathbf{RSM}_i$ do not necessarily overlap with rows in FCM_s.

Step 4: Calculate the column mean of $\mathbf{RSM}_i$. The column mean $\overline{\mathbf{RSM}_i}$ is used to represent the background fold change of n genes and is calculated in a similar manner as equation 4.

Step 5: Repeat steps 3 and 4 6000 times. The $\overline{\mathbf{RSM}_i}$ from iteration i are stored as rows in a 6000 x 4 matrix, **ECD**.

Step 6: Calculate the enrichment p-value of gene set s by using an empirical cumulative distribution derived from the 6000 x 4 matrix ECD. The empirical cumulative distribution is defined by the following function

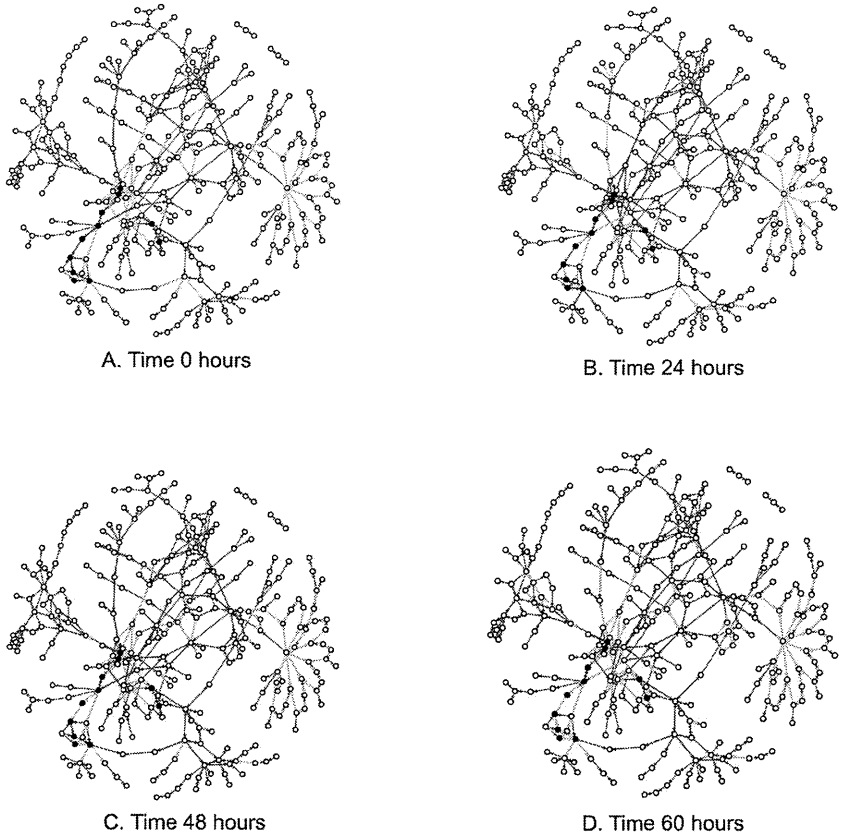


Figure 4. These graphs highlight the fold change direction of known genes in our data in response to oil accumulating conditions at each time point. A gene involved in a reaction is represented by an edge while the compounds in a reaction are represented by the nodes. Genes that were up-regulated during oil accumulation are drawn as green edges while red edges represent genes that were down-regulated. Genes for which data was unknown were drawn as gray edges. The compounds colored in black are part of the first shortest path found between glucose and triacylglycerol (Figure 2). The edges that connect those compounds shift from red to green during the 60 hour course of the experiment.
doi:10.1371/journal.pone.0107629.g004

$$\hat{F}_s(\mathbf{u}) = \frac{\Sigma_{\forall i} \mathbb{I}(ECD_{i,0} \leq u_0, ECD_{i,24} \leq u_{24}, ECD_{i,48} \leq u_{48}, ECD_{i,60} \leq u_{60})}{n} \quad (5)$$

where $\hat{F}_s$ is the empirical cumulative distribution of gene set s, $\mathbf{u}$ is a fold change vector with a length equal to the number of columns of **ECD** (Step 5), u_t is a value in $\mathbf{u}$ at time t which takes the values 0, 24, 48 and 60 in our data, $\mathbb{I}$ is the indicator matrix, $ECD_{i,t}$ is the fold change value of the ith row at time t in the **ECD** matrix and n is the size of gene set s.

The enrichment p-value of gene set s is calculated by substituting $\mathbf{u}$ with $\overline{\mathbf{FCM}_s}$ (Equation 4).

The algorithm detailed above was implemented in R [22], and the empirical cumulative distribution and enrichment p-value was calculated using the mecdf package [26].

Enriched Pathway Plots

The significantly enriched gene sets selected from the GSEA results are metabolic pathways which were plotted to display the GSEA results and visualise reactions of the compounds within them. The generic pathway and enzyme KGML files were downloaded from KEGG and read into R. They were parsed using the KEGGgraph package [27] using the default data structure where nodes represent KEGG orthologs and edges represent reactions. This was restructured so that the nodes represent compounds and the edges represent KEGG orthologs. The graphs were then merged into one and converted into an igraph object for plotting and access to network analyses such as *get.all.shortest.paths* [28]. Unconnected nodes were removed to reduce clutter in the final plot.

Author Contributions

Conceived and designed the experiments: YS Masayoshi Tanaka TY T. Tanaka. Analyzed the data: PSW SA Michihiro Tanaka WF T. Taniguchi. Wrote the paper: PSW SA.

References

1. Mata TM, Martins AA, Caetano NS (2010) Microalgae for biodiesel production and other applications: A review. Renewable and Sustainable Energy Reviews 14: 217–232.
2. Rodolfi L, Zittelli GC, Bassi N, Padovani G, Biondi N, et al. (2009) Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. Biotechnology and Bioengineering 102: 100–112.

3. Radakovits R, Jinkerson RE, Fuerstenberg SI, Tae H, Settlage RE, et al. (2012) Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropsis gaditana. Nature Communications 3.
4. Rismani-Yazdi H, Haznedaroglu BZ, Hsin C, Peccia J (2012) Transcriptomic analysis of the oleaginous microalga Neochloris oleoabundans reveals metabolic insights into triacylglyceride accumulation. Biotechnology for Biofuels 5.
5. Satoh A, Ichii K, Matsumoto M, Kubota C, Nemoto M, et al. (2013) A process design and productivity evaluation for oil production by indoor mass cultivation of a marine diatom, Fistulifera sp. JPCC DA0580. Bioresource Technology 137: 132–138.
6. Muto M, Fukuda Y, Nemoto M, Yoshino T, Matsunaga T, et al. (2013) Establishment of a Genetic Transformation System for the Marine Pennate Diatom Fistulifera sp. Strain JPCC DA0580—A High Triglyceride Producer. Marine Biotechnology 15: 48–55.
7. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biology 11: R106.
8. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140.
9. Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Computational Biology 8.
10. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc 4: 44–57.
11. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13.
12. Berrix GF, Beaver JE, Cenik C, Tasan M, Roth FP (2009) Next generation software for functional trend analysis. Bioinformatics 25: 3043–3044.
13. Zheng Q, Wang X (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res 36.
14. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, et al. (2009) A Novel Signaling Pathway Impact Analysis (SPIA). Bioinformatics 25: 75–82.
15. Kim SY, Volsky DJ (2005) PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics 6.
16. Croce R, van Amerongen H (2013) Light-harvesting in photosystem I. Photosynthesis Research.
17. Grasso EJ, Scalambro MB, Calderón RO (2011) Differential response of the urothelial V-ATPase activity to the lipid environment. Cell Biochemistry and Biophysics 61: 157–168.
18. Ettema TJ, Ahmed H, Geerling AC, van der Oost J, Siebers B (2008) The non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN) of Sulfolobus solfataricus: a key-enzyme of the semi-phosphorylative branch of the Entner-Doudoroff pathway. Extremophiles 12: 75–88.
19. Nojima D, Yoshino T, Maeda Y, Tanaka M, Nemoto M, et al. (2013) Proteomics Analysis of Oil Body-Associated Proteins in the Oleaginous Diatom. Journal of Proteome Research.
20. Yamada K, Tomii K (2013) Revisiting amino acid substitution matrices for identifying distantly related proteins. Bioinformatics.
21. Warden CD, Yuan YC, Wu X (2013) Optimal Calculation of RNA-Seq Fold-Change Values. International Journal of Computational Bioinformatics and In Silico Modeling 2: 285–292.
22. R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org.
23. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res 40: D109–D114.
24. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28: 27–30.
25. Ripley BD (1987) Stochastic Simulation. Wiley-Interscience Paperpack Series.
26. Maia C (2011) mecdf: Multivariate Empirical Cumulative Distribution Functions.
27. Zhang JD, Wiemann S (2009) KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor. Bioinformatics 25: 1470–1471.
28. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems: 1695.

3　iPS細胞からのビッグデータの情報セキュリティと創薬，医療への活用

藤渕　航*

3.1　はじめに

2006年にマウス，2007年にヒトのiPS細胞の作製が発表されて以来，目覚ましい勢いでiPS細胞やiPS細胞由来の細胞の医療や創薬への応用への準備が進んでいる。iPS細胞は正に夢の生体ツールとして大きな期待が膨らんでおり，将来には総国民iPS細胞ストック時代を迎える可能性も秘めている。本節執筆の現時点では，コスト的にも技術的にも難しく総国民iPS細胞ストックには否定的な見解も多いが，既に2013年度よりJSTから「再生医療実現拠点ネットワークプログラム」が開始され，その中核拠点に選出された京都大学iPS細胞研究所で，10年の歳月をかけて日本人の９割をカバーする推定約140名の組織適合性遺伝子HLAがA，B，DR三座ホモドナーのiPS細胞化ストック構想が始まった。少なくとも本構想は，未来においてiPS細胞を医療の中核に据えるための基本科学技術を発展させることは間違いなく，10年間で質の高いiPS細胞を低コストで作製できる時代を到来させるであろう。思い起こすことは，米国NIHで2003年に1,000ドルでヒトゲノムを解読する構想のロードマップが提示されてから，僅か10年以内の2012年に1,000ドルゲノムを達成したと報道が出るまでになった。これは当時，ヒトゲノム１人分の解読に100万ドル以上かかっていた事実からすれば，コストが1,000分の１になるとは誰も想像がつかなかったことである。このような経験を踏まえて本節では，現在から10年後の未来にわたってiPS細胞から生じるいわゆる「ビッグデータ」について，現状を踏まえた上でできる限り正確なこれからの動向について予測を含め，記載したいと思う。

3.2　iPS細胞がもたらすビッグデータ

3.2.1　iPS細胞の品質管理

１個人の体細胞からiPS細胞を作るのに，熟練した技術者なら数ヶ月程度しかかからない。簡易な実験用iPS細胞であればこのようなiPS細胞でも問題ないが，医療や創薬用となると厳重な品質管理が必要とされ，この期間は最低でも６ヶ月はかかると考えられる。例えば，表１に示したように，iPS細胞の特性は，ドナー自身が持つ内在的な性質であるbiological characteristicsとiPS細胞の作製過程で生じる様々な技術的な性質であるtechnical characteristicsに大別される。biological characteristicsの方は，性別や血液型や年齢などの基本情報の他に，このドナー由来のiPS細胞は免疫寛容性が高いか，また，遺伝的疾患を持っていないか，AIDSウイルスなど後天的要因の病気に罹患していないか，など他者の再生医療に用いることが可能かを判定する基準となる。一方，technical characteristicsの方はiPS細胞を生成した時の完成度であり，幹細胞マーカーの発現，細胞分裂能力，分化能力などの他に，作製過程で用いた外来遺伝子の残存度など作製方法にも依存するものである。

＊　Wataru Fujibuchi　京都大学　iPS細胞研究所　増殖分化機構研究部門　教授

表1　iPS細胞の品質評価で必要とされる情報例（文献1）より改修）

特徴分類	カテゴリー	例
biological characteristics	基礎情報	性別，血液型，年齢，人種など
	免疫情報	HLAタイプなど
	先天的疾患	既往症，遺伝病など
	後天的疾患	HIV，A/C型肝炎，HTLVなど
	由来組織	皮膚，脂肪，歯髄，血液など
technical characteristics	細胞形態	コロニー形状，数，密度など
	幹細胞マーカー	Nanog，Oct3/4など
	増殖能力	分裂速度，分裂形式など
	分化能力	外，内，中胚葉層形成など
	外来遺伝子残存度	導入遺伝子検査
	体細胞突然変異	SNPs，CNVsなど
	細胞同一性	STR解析など
	微生物汚染	マイコプラズマ検査など
	ゲノム安定性	核型解析など
	エピゲノム	アレイ，シーケンサー解析など

3.2.2　遺伝的リスク

特に，今後は，biological characteristicsであるSNPs情報からわかる遺伝的なリスクが重用視されると考えられる。既に世界的に有名なGWAS研究などによるSNPsとフェノタイプを結ぶdbGaPなど情報データベースが蓄積されている。近年，SNPsは親由来だけでなく，卵細胞から成体に至るまでに高度に蓄積された体細胞突然変異が原因となっていることもわかっており，これがさらにiPS細胞作製過程でも生じる変異と合わせて最終的に生じたiPS細胞での遺伝的リスクを調べる必要がある。表2に主なSNPsやCNVsなどゲノム変異に関わる情報データベースを記した。

現在，このようなゲノム上でわかっている疾患関連サイトについて，潜在的な遺伝的リスクを調べる高性能なツールの開発も世界中で進められていると考えられる。例えば，dbSNPなどのデータベースを用いてSNPsを検索してアノテーションするSNPdat，SNPnexus，Snap，SNP Function Portal，SNPper，Fans，FunctSNP，Annovarなどがある。一方，データベースにないrare variantsの場合には既知の疾病パスウェイや遺伝子発現やデータなどとassociateさせるVAASTやBioBinなどの様々なソフトウェアが使用されている。

3.3　ゲノム情報産業と我が国での個人情報保護

3.3.1　ゲノム情報を活用した産業

個人から得られるゲノム配列には膨大な情報が含まれている。これを利用して，個人の疾病リスクの予想のみならず，性格や適正などまで検査する産業が発展してきている。既に欧米では，

表2　代表的なゲノム変異関連のデータベース例

名称	開発機関	主な特徴	アドレス
dbSNP	米国National Center for Biotechnology Information/NIH	1塩基置換および短い欠失・挿入，レトロポゾン，STR多型のデータレポジトリ	http://www.ncbi.nlm.nih.gov/snp
dbVar	米国National Center for Biotechnology Information/NIH	1kb以上の逆位，転座，欠失，挿入などゲノム構造変異のデータレポジトリ	http://www.ncbi.nlm.nih.gov/dbvar
dbGaP	米国National Center for Biotechnology Information/NIH	医療以外のフェノタイプを含むジェノタイプとの関連データベース／アクセス制限データ有り	http://www.ncbi.nlm.nih.gov/gap
ClinVar	米国National Center for Biotechnology Information/NIH	疾病に関係する塩基やアミノ酸変異のデータベース	http://www.ncbi.nlm.nih.gov/clinvar
DGV	カナダThe Centre for Applied Genomics	健康人における50b以上の逆位，転座，欠失，挿入などゲノム構造変異のデータレポジトリ	http://dgv.tcag.ca/dgv/app/home
GWAScentral/HGVBase	スウェーデンKarolinska Institute，英国European Bioinformatics Institute，ドイツEuropean Molecular Biology Laboratory	ヒトGWAS研究のレポジトリ	http://www.gwascentral.org
PharmGKB	米国Stanford University	ゲノム変異と薬物反応に関するデータベース	http://www.pharmgkb.org
HGMD	英国Cardiff University	遺伝子欠失と遺伝病のデータベース	http://www.hgmd.cf.ac.uk
JSNP	東京大学医科学研究所，JST	日本人のSNPsデータベース	http://snp.ims.u-tokyo.ac.jp
DECIPHER	英国ウェルカム・トラスト　サンガー研究所	Ensemblを利用したヒト染色体不均衡とフェノタイプデータベース	http://decipher.sanger.ac.uk

数年前から民間で有料の遺伝子検査サービスが始まっており，脳梗塞，心筋梗塞，糖尿病，乳がんリスク，などのDNA多型情報に基づく情報を蓄積している。2012年3月には，米国立衛生研究所（NIH）から，企業などが提供した遺伝子検査情報を提供する遺伝子検査レジストリ（GTR）というデータベースが公開された（図1）。このデータベースの目的は，乱立する遺伝子検査企業に透明性を持たせ，どこでどのような検査を行っているか，また，遺伝子データを研究者が共有する仕組みを提供して科学的研究を促進させようとするものである。

一方，これに伴う個人情報の保護が問題となってきている。欧米では個人情報保護は随分と進んでおり，1980年のOECD（経済協力開発機構）によるプライバシーガイドラインを受けて，早くから強固なデータ保護やプライバシー保護の法律ができ，DPA（欧州Data Protection Authority）やFTC（米連邦取引委員会）などの公的組織が情報保護の取り締まりを行っている。また，両者の間では2001年のSafe Harbor合意に基づく国際間での情報保護の取り扱いまで法的に整備されている。2013年6月に米国立衛生研究所（NIH）のNCBIを訪問し，実際にどのようにしてプラ

図１　遺伝子検査レジストリ（GTR）
http://www.ncbi.nlm.nih.gov/gtr/

イバシーと健康情報提供を両立しているのか調査した。その結果，①個人を特定できるDNA情報についてはインフォームドコンセントなどを含む規定の整備，②情報漏洩を想定したデータの暗号化や自動削除技術開発，③ユーザの階層的利用権限化などを行って保護する方法の開発を進行中であるとの情報を得ている。

3.3.2　日本における個人情報保護

我が国は個人情報保護においては欧米に比べてやや後進国である。日本では2003年にようやく個人情報保護法に基づく規制が制定されたばかりで，急増する個人の遺伝情報の取り扱いについて国内の法整備は遅れている。また，日本人から得られた遺伝情報を国外へ持ち出すことについて何ら法的規制がないのが現状である。本節の執筆中に「特定秘密保護法」が国会で可決されたが，今後，これがどのように日本人の遺伝情報の漏洩に影響するのかはまだ予想がつかない段階である。我が国でも国内向けに遺伝子検査の企業も増加している中，国立バイオサイエンスデータベースセンター（NBDC）では，2013年４月に「ヒトデータ共有ガイドライン」および「ヒトデータ取り扱いセキュリティガイドライン」の策定を始めている。また，遺伝子から得られる個人情報保護の規定や法令については，各省庁で様々に検討が開始されつつある。どのように統廃合され，国全体として整備されるのかにはまだ少し時間がかかるであろう。

さらに，バイオ関連情報漏洩をさせないためのセキュリティ技術としても，化合物の秘密検索（産総研＆東大 2011年），臨床データの秘密計算（日本成人白血病治療共同研究グループ&NTT 2012年），クラウド上遺伝子情報データでの秘密検索（日立ソリューションズ 2012年）などと，

いずれも計算機上のデータを暗号化したまま計算が可能な技術が報告されている。しかしながら全体像としては，今後，莫大な量に膨れ上がる個人の細胞や遺伝子情報の保護だけでなく，転送，利用におけるセキュリティの標準化などの包括的な取り組みや技術開発がなされていないままである。

3.4　高度医療情報時代における創薬と再生医療

3.4.1　高度医療情報時代の到来

我が国は国策の一環としてiPS細胞を基軸として再生医療・医療情報大国を目指している。この過程においてこれまで見過ごされてきた個人情報の問題が浮き彫りになってくるのは必定である。今後は，全国の病院・研究所などからiPS細胞情報データベースにアクセス可能であるが，同時に高いセキュリティを維持するシステムが必要である。その上で，国民がiPS細胞を個人で作製する上で不安となるプライバシーの問題を解決しつつ医療上重要な遺伝情報を提供するという，真に医療情報大国として発展するために必要なインフラを構築する必要がある。具体的には，国内の病院やデータセンターなどの拠点間でiPS細胞の異なる情報を相互交換するため，セキュアなデータ検索・表示とデータのダウンロードアップロードのための計算機技術が必要である。基本技術として，一部先述したが，データの暗号化や第三者に情報が漏洩した場合の自動削除，ユーザの階層的利用権限化，分散データなどが必要な開発項目と想定される。

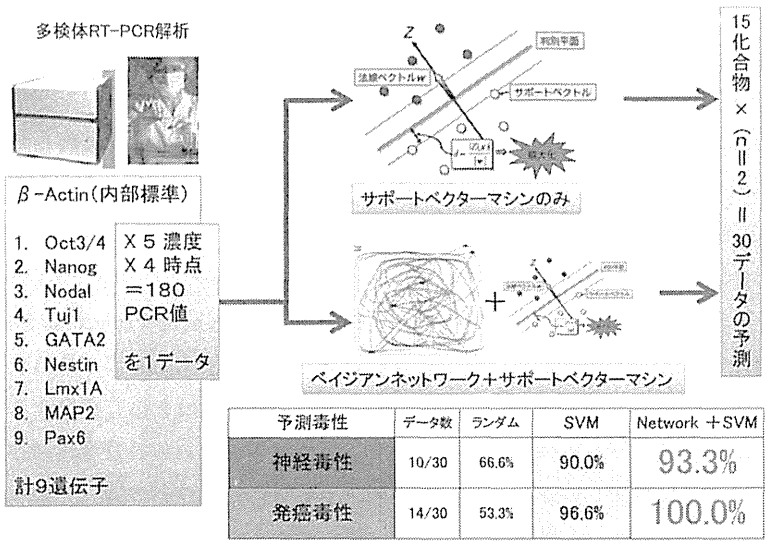
3.4.2　iPS細胞の創薬・毒性評価からの情報

これまでマウス，ラットなどで行われていた創薬や毒性評価の研究が急激な勢いでヒトiPS細胞を用いたシステムに置き換わりつつある。これは，サリドマイドなどに代表されるようにマウスで無毒性が保障された化合物でもヒトでは毒性を示すなどヒト特有の効果や副作用を示すことが明らかになったからである。例えば，ヒトES細胞から神経細胞を誘導する過程でメチル水銀を投与すると，ヒトでは神経樹状突起が縮退するが，マウスでは縮退せず，替わりに細胞死が観察された[2]。このように，より実態を反映したヒトiPS細胞システムが今後は創薬や毒性評価の主力となると考えられる。また，米国NIHのNCATS（National Center for Advancing Translational Sciences）では，iPS細胞チップの開発も行っており，これまで臨床試験でドロップアウトした化合物を再利用できないか研究する計画もある。今後はこのような化合物の投与データも多く蓄積するが，そこから有益な情報を取り出して活用することも必要となる。例えば，我々は少ない毒性のデータを学習させ，新規化合物の毒性を予測するための手法を開発し報告した[3]。そこでは，遺伝子発現情報から遺伝子ネットワークを構築し，これを従来の遺伝子発現データだけのサポートベクターマシン予測と比較するもので，遺伝子ネットワークを学習に加えると予測精度が向上する結果が得られている（図2）。

3.4.3　再生医療情報のデータマイニング

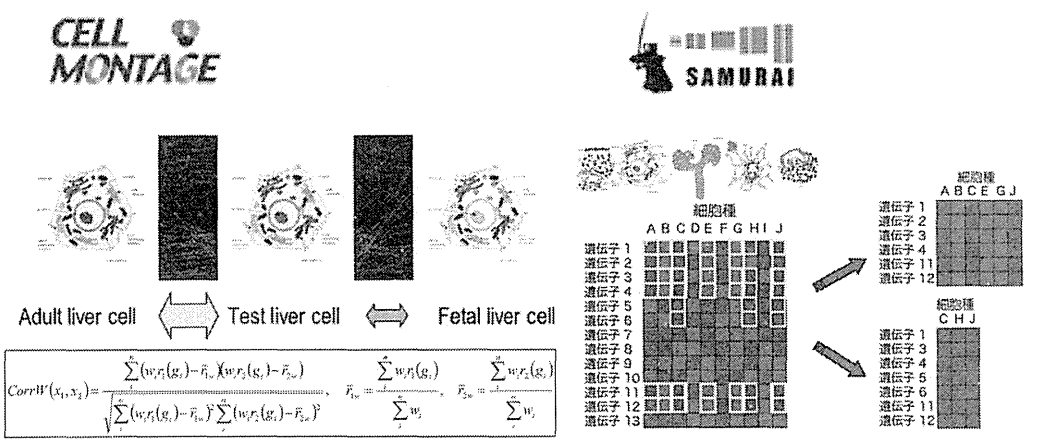
今後，iPS細胞を含め，蓄積する細胞ビッグデータから有益な情報を引き出すためのツール開発を我々の研究室では進めてきた。例えば，高速類似細胞検索CellMontage[4]，網羅的遺伝子モジュ

ール探索システムSAMURAI[5]などがある（図３）。CellMontageでは，手持ちの細胞や人工作製した細胞がどのタイプの細胞に近いかを瞬時に検索することが可能である。現在，開発されている計算システムでは１秒間に数千件の細胞データを検索することができる[6]。例では膵臓の細胞



予測毒性	データ数	ランダム	SVM	Network ＋SVM
神経毒性	10/30	66.6%	90.0%	93.3%
発癌毒性	14/30	53.3%	96.6%	100.0%

図２　幹細胞を用いた毒性化合物予測
（データ提供：平成21～23年度厚生労働科研費「化学物質リスク事業：大迫班」による）



重み付き順位相関係数に基づく高速検索アルゴリズムRaPiDS

Horton *et al.*, *Genome Informatics* (2006)
Fujibuchi *et al.*, *Bioinformatics* (2007)

飽和アイテム集合マイニングとノイズを許したモジュール合成

Okada and Fujibuchi, *CAMDA* (2007)
Fujibuchi *et al.*, *Methods Mol. Biol.* (2009)

図３　高速類似細胞検索CellMontage（左）と網羅的遺伝子モジュール探索システムSAMURAI（右）

Query >NR7f5 !VALUE!iAFLP!single!iAFLP!Homo sapiens!adult endocrine/exocrine pancreas pancreas NR7f5 , (total genes)
Platform: GPL999999(129 entries), sampling genes: 32512(GO-bound: all), Probability: 0.01, Correlation: 0.0
Found: 100 entries. Start: Tue Sep 26 18:22:46 2006 End: Tue Sep 26 18:22:47 2006

Top100	Sample	DataSet	Platform	Type	Channel	Organism	Description	Probability(Correlation,#Genes)	Detail
1	!VALUE!iAFLP!single!iAFLP!Homo sapiens!adult endocrine/exocrine pancreas pancreas NR7f5	iAFLP	iAFLP	single VALUE	Homo sapiens	adult endocrine/exocrine pancreas pancreas NR7f5	0.000e+00(1.00,32512)	
2	!VALUE!iAFLP!single!iAFLP!Homo sapiens!adult endocrine/exocrine pancreas pancreas(caudal) NR8d5	iAFLP	iAFLP	single VALUE	Homo sapiens	adult endocrine/exocrine pancreas pancreas(caudal) NR8d5	2.413e-820(0.33,32512)	
3	!VALUE!iAFLP!single!iAFLP!Homo sapiens!adult endocrine/exocrine pancreas pancreas(caput) NR8c5	iAFLP	iAFLP	single VALUE	Homo sapiens	adult endocrine/exocrine pancreas pancreas(caput) NR8c5	2.085e-610(0.29,32512)	
4	!VALUE!iAFLP!single!iAFLP!Homo sapiens!adult endocrine/exocrine pancreas pancreas(caput) NR8b5	iAFLP	iAFLP	single VALUE	Homo sapiens	adult endocrine/exocrine pancreas pancreas(caput) NR8b5	7.868e-583(0.28,32512)	
5	!VALUE!iAFLP!single!iAFLP!Homo sapiens!adult stomach/colon stomach stomach NR7f4	iAFLP	iAFLP	single VALUE	Homo sapiens	adult stomach/colon stomach stomach NR7f4	4.721e-439(0.24,32512)	
6	!VALUE!iAFLP!single!iAFLP!Homo sapiens!adult endocrine/exocrine pancreas pancreas(caudal) NR8f5	iAFLP	iAFLP	single VALUE	Homo sapiens	adult endocrine/exocrine pancreas pancreas(caudal) NR8f5	1.410e-436(0.24,32512)	
7	!VALUE!iAFLP!single!iAFLP!Homo sapiens!adult endocrine/exocrine pancreas pancreas(corpus) NR8e5	iAFLP	iAFLP	single VALUE	Homo sapiens	adult endocrine/exocrine pancreas pancreas(corpus) NR8e5	3.498e-300(0.20,32512)	
8	!VALUE!iAFLP!single!iAFLP!Homo sapiens!adult stomach/colon small_intestine duodenum NR7d5	iAFLP	iAFLP	single VALUE	Homo sapiens	adult stomach/colon small_intestine duodenum NR7d5	4.833e-266(0.19,32512)	
9	!VALUE!iAFLP!single!iAFLP!Homo sapiens!adult liver liver liver NR7c3	iAFLP	iAFLP	single VALUE	Homo sapiens	adult liver liver liver NR7c3	5.979e-204(0.17,32512)	

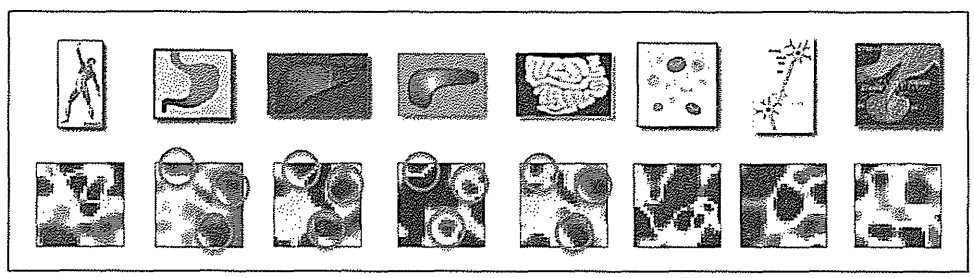


図4　CellMontageで膵臓の細胞をクエリーとして検索

をクエリーとして検索しているが，膵臓に近い細胞は胃や肝臓などである（図4）。これらの臓器は全て内胚葉性であり，発生上，近い関係にある。また，機械学習の手法と組み合わせることによって人工作製したiPS細胞の質も検索可能である。例えば，iPS細胞がどの細胞に由来していたかについての試験的な結果では，10種の由来細胞を含む73マイクロアレイデータ検索の順番間違いがわずか6％と大変に有効な結果が得られている。また，iPS細胞コロニーの写真画像から良質なiPS細胞を推定する研究では，17日目以降の初期化をほぼ完了したiPS細胞においては100％の正確さで判定できた。

また，SAMURAIシステムでは，買い物をする時のassociation ruleという考えから生まれたLCM：Linear Time Closed Itemset Minerと呼ばれるアルゴリズムを用いて，大量の細胞データに共通する遺伝子の組み合わせ（飽和アイテム集合）を網羅的に取り出すものである（図5）。現在，2,912件の多様なヒト細胞マイクロアレイデータから遺伝子モジュールの辞書を作成する研究が当ラボで進行中である。これから将来においても，益々，大量の細胞関連データからのデータマイニングツールのアイデアが必要となるであろう。

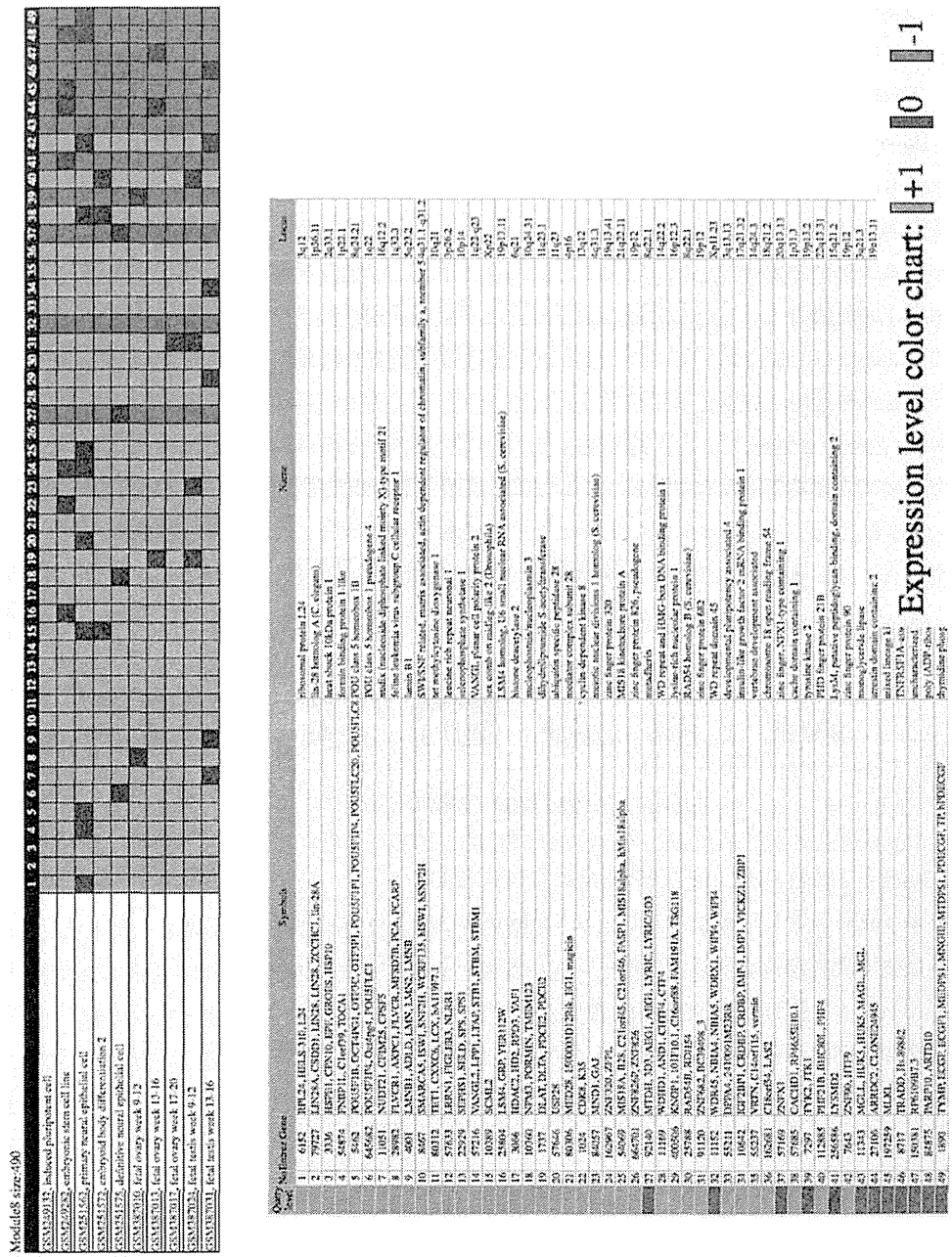


図５　iPS細胞，ES細胞，胚葉体，神経表皮，胎児の生殖細胞に共通な遺伝子モジュールの例
POU5F1，LIN28などの49遺伝子が抽出されている。

3.5　今後必要とされる解析技術について

今後はiPS細胞の情報だけでなく，ヒト細胞全てについての最新情報を全国の医療関連機関を結んで相互に共有できることが，未来の細胞医療には必要となってくると考えられる。その時に

大量のデータを首尾よく保存し，検索可能にするには，従来のようなデータベーススキーマに基づく関係データベースのような変化に弱いシステムでは追い付かないであろう。今後必要とされるのは，乱雑な大量のデータを記憶装置上にただ単に置いただけで，計算機が自動的にデータを整理し，関連付けを行い，検索やデータマイニング，機械学習を行うことを可能にする，より一般性の高いシステムを開発することが必要である。特に関連する報告として特筆したいのは，IBMのWatsonシステムが百科事典を機械学習し，「Jeopardy!」という米国の人気クイズ番組で人間の優勝者に勝利したことである。今後は，このWatsonをさらに進化させて，自動整理，自動学習するシステムを開発することがiPS細胞からのビッグデータマイニングには必要であるかも知れない。

文　　献

1) G. Stacey, *Prog. Brain Res.*, **200**, 41 (2012)
2) X. He *et al.*, *Toxicol. Lett.*, **212**, 1 (2012)
3) W. Fujibuchi *et al.*, Prediction of Chemical Toxicity by Network-based SVM on ES-cell Validation System, The Proceedings of the 2011 Joint Conference of CBI-Society and JSBi, Kobe (2011)
4) W. Fujibuchi *et al.*, *Bioinformatics*, **23**, 3103 (2007)
5) W. Fujibuchi *et al.*, *Methods Mol. Biol.*, **577**, 55 (2009)
6) P. Horton *et al.*, *Genome Informatics*, **17**, 67 (2006)