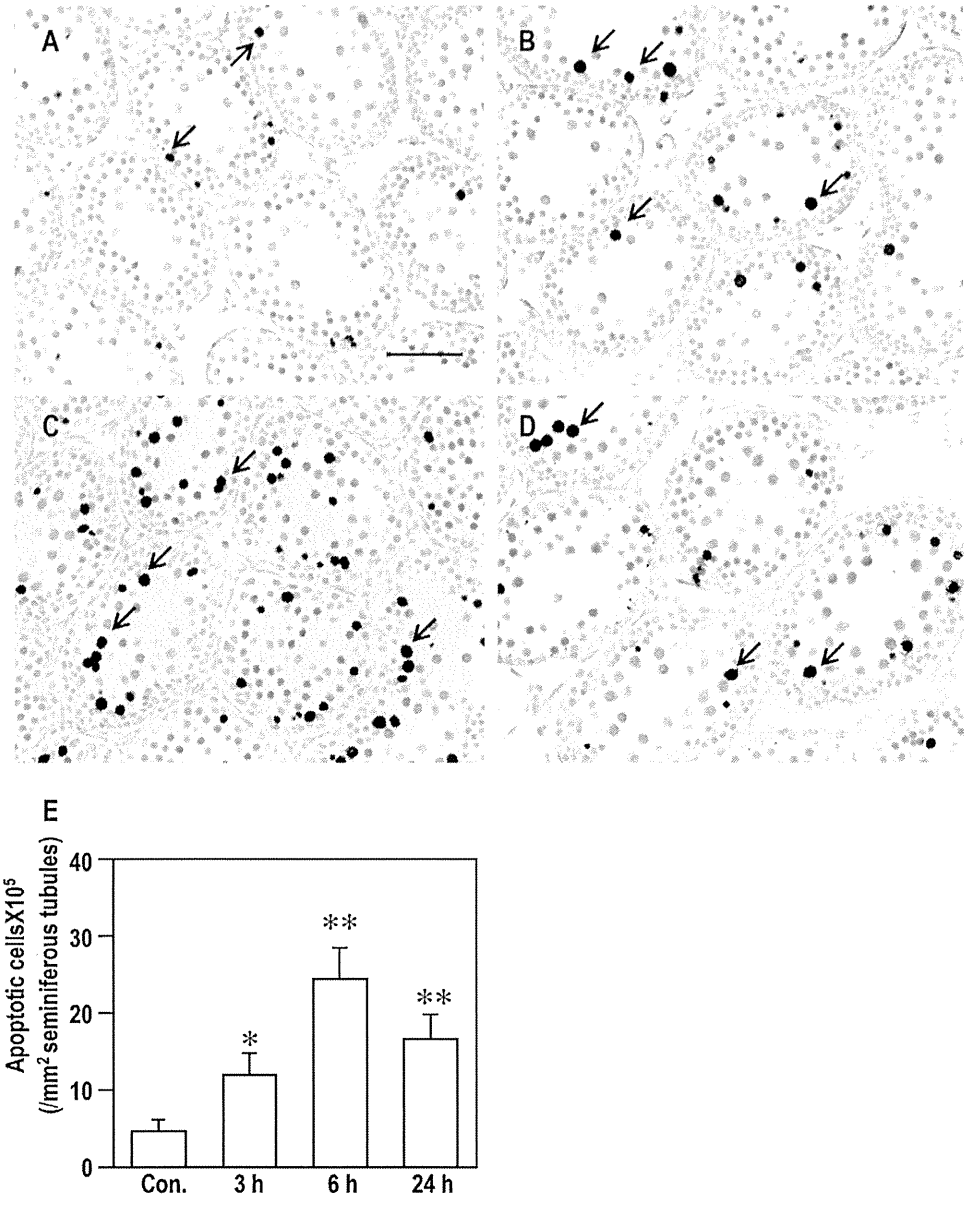

Fig. 2. TUNEL assay of testicular sections for apoptotic spermatogenic cells by single administration of butylparaben. Vehicle-treated control (A), and at 3 h (B), at 6 h (C), and at 24 h (D) after administration. Arrows indicate apoptotic cells. Note the maximal apoptotic cell number at 6 h after administration. Scale bar = 50 μm. Quantification of apoptotic cells (E). Values representing the number of apoptotic cells/mm^2 cross sectional area of seminiferous tubules are expressed as the eans ± SE ($n = 8$). Statistically significant differences between means from the control and treated groups were determined by ANOVA followed by Fisher's PLSD test (*$P < 0.05$; ** $P < 0.01$, *versus* control).

2000). Spermatogenic cell apoptosis is also induced by many factors including hormonal deprivation, heat, radiation and environmental endocrine disruptors (Koji and Hishikawa, 2003; Shaha, 2007). Excessive and/or abnormal apoptosis of spermatogenic cells is one of the main reasons for oligozoospermia and azoospermia (Tesarik et al., 1998).

There is much evidence that butylparaben can interact with estrogen receptors (ERs). For example, several studies employing an ER-mediated yeast growth assay or a reporter assay with human breast cancer cell line MCF-7 have demonstrated the interaction between butylparaben and ERs (Routledge et al., 1998; Pedersen et al., 2000; Okubo et al., 2001). Routledge et al. (1998) reported that estrogenic activity was inhibited by 4-hydroxy tamoxifen. In addition, proliferative effects of parabens on MCF-7 cells were completely suppressed by the anti-estrogen ICI 182,780 *in vitro* (Okubo et al., 2001). Moreover, butylparaben has also been shown to increase uterine weight in both immature rats and mice and in adult ovariectomized mice in the rat uterotrophic assay (Routledge et al., 1998), thus indicating its estrogenic activity. Decreased testicular testosterone biosynthesis, as well as decreased serum LH and serum FSH levels, occurs after exogenous estrogen exposure, together with increased spermatogenic cell apoptosis (D'Souza et al., 2005; Alam et al., 2010b). Similarly, butylparaben has been shown to decrease serum testosterone levels, resulting in decreased counts of round and elongated spermatids (Byford et al., 2002; Oishi, 2002; Taxvig et al., 2008). As far as we are aware, the present study has shown for the first time that butylparaben induces increased apoptosis of spermatogenic cell shortly after treatment. These findings agree well with our previous data on di(*n*-butyl) phthalate, a suspected estrogenic compound, which significantly increases spermatogenic cell apoptosis in prepubertal rats (Alam et al., 2010a,b). The mechanism how butylparaben induces spermatogenic cell apoptosis is not known at present and further exploration is needed. The development of the male reproductive system and

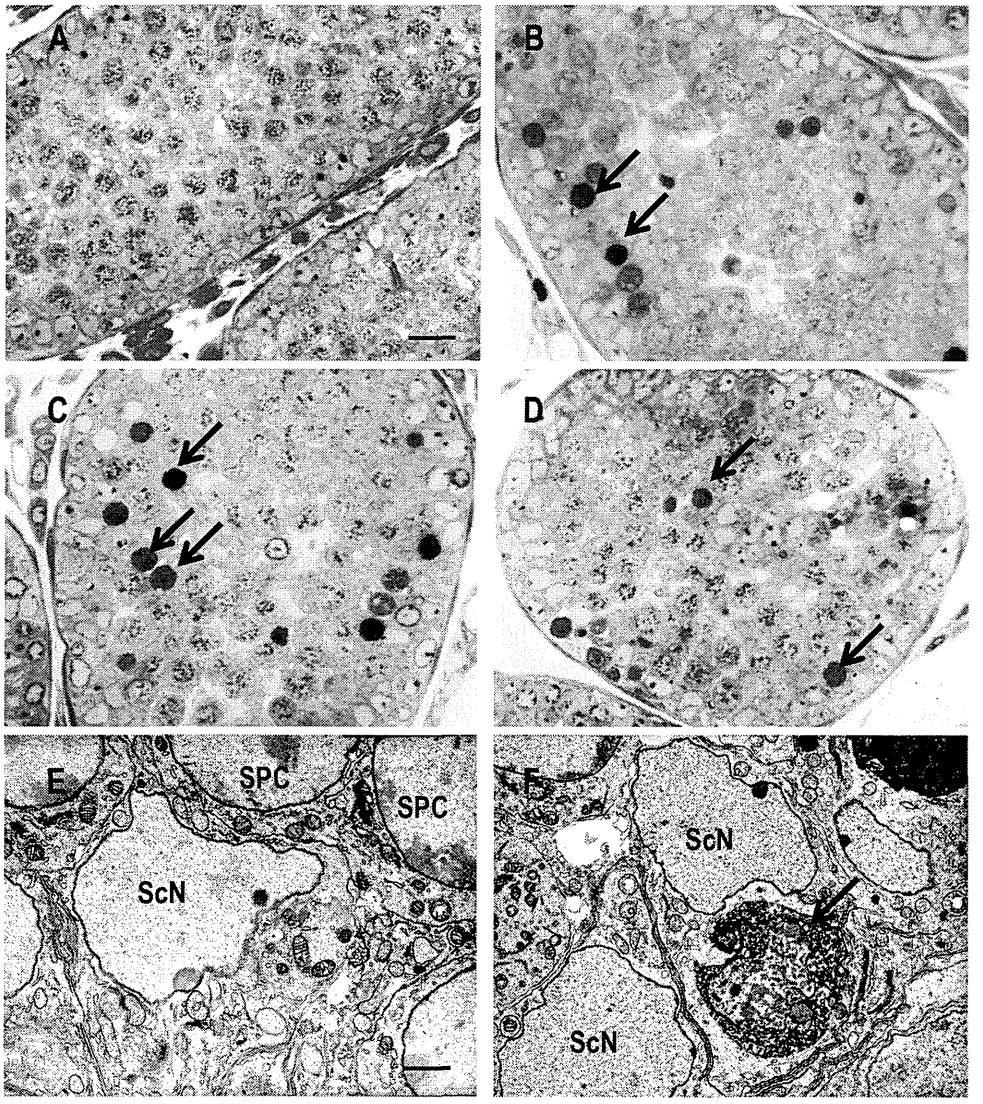

Fig. 3. Semithin sections (A–D) and transmission electron micrographs (E–F) of seminiferous tubules after single butylparaben administration. Vehicle-treated control (A), and at 3 h (B), at 6 h (C), and at 24 h (D) after treatment. Arrows indicate apoptotic spermatogenic cells. Scale bar = 20 μm. Control rats showing normal Sertoli cell nucleus (ScN), spermatogonium (SG), and spermatocyte (SPC) (E). An apoptotic spermatogenic cell characterized by heterochromatin condensation and shrinkage of cytoplasm surrounded by a normal Sertoli cell nucleus is distinctly observed at 24 h (F). Scale bar = 2 μm.

spermatogenesis are controlled by testosterone. Although the possibility that decreased testosterone levels may result in the increase in spermatogenic cell apoptosis cannot be ruled out, it seems to be due to direct cytotoxic effects of parabens on spermatogenic cells or estrogenic action. For example, our previous study has demonstrated that an estrogenic compound, such as di(*n*-butyl) phthalate/estradiol-3-benzoite-induced spermatogenic cell apoptosis was not associated with testicular steroidogenesis (Alam et al., 2010b).

It has also been reported that estrogen directly induces spermatogenic cell apoptosis by cytochrome c release from mitochondria and FasL up-regulation in *in vitro* model with isolated spermatogenic cells (Mishra and Shaha, 2005). However, this apoptosis was inhibited by tamoxifen, indicating that an estrogen-induced change occurs through hormone receptor interaction in spermatogenic cells. Therefore, it is possible to postulate that the ERs present in testes, probably in spermatocytes and spermatids, have a role in inducing spermatogenic cell apoptosis when binding to exogenous estrogenic compounds including butylparaben. Indeed, both ERα and ERβ are present in rat spermatocytes and round spermatids (Saunders et al., 1998; Pelletier et al., 2000). In the present study, we also observed that butylparaben-induced apoptosis was commonly found in spermatocytes, less frequently in spermatids and not found in spermatogonia and somatic cells. Moreover, butylparaben adversely affects spermatozoa by an inhibitory effect on acrosin and impairment of sperm membrane function (Song et al., 1991), suggesting that spermatogenic cells are a direct target of parabens in testicular toxicity (Tavares et al., 2009). In addition, our previous studies have reported that disruption of Sertoli cell-spermatogenic cell physical interactions leads to detachment and sloughing of spermatogenic cells from the seminiferous epithelium after exposure to phthalates, and these detached spermatogenic cells lost the support and nurture provided by Sertoli cells and eventually undergo apoptosis (Alam et al., 2010c). Similarly, in the present study, we also observed that apoptotic cells become detached from their neighbors, probably due to collapse of Sertoli cell vimentin filaments.

The maximal number of apoptotic spermatogenic cells was found at 6 h after butylparaben administration, and at 24 h, the number of apoptotic cells began to decline, though it was still significantly greater than that in the control group. This is most probably due to the rapid elimination of apoptotic cells by phagocytosis, a

common fate of cells undergoing apoptosis (Maeda et al., 2002; Tay et al., 2007). In the present finding, in transmission electron microscopy, apoptotic cells appeared to be engulfed by neighboring Sertoli cells. Sertoli cells play an important role in clearing apoptotic spermatogenic cells by the process of phagocytosis. It is likely to be a self-defense mechanism (Maeda et al., 2002). During spermatogenic cell differentiation, although more than half of differentiating spermatogenic cells die by apoptosis before they mature into spermatozoa (Dym, 1994), only a limited number of apoptotic cells are detected when testicular sections are examined by TUNEL assay (Koji et al., 2001; Maeda et al., 2002). Unfortunately, the role of phagocytosis in butylparaben administration cannot be ascertained at present, and further exploration is needed.

In conclusion, the present results of single butylparaben administration in prepubertal rats demonstrated histopathological changes in the seminiferous tubules and loss of spermatogenic cells by apoptosis. Because of the importance of these effects, more detailed studies on the mechanism of toxicity induced by parabens in the male reproductive organs are necessary. This study is now underway in our laboratory to elucidate the mechanism of action of testicular dysfunction induced by parabens.

Acknowledgments

We thank Dr. Andriana Bibin for his technical support during TEM handling in this study. This work was supported in part by Grants-in-Aid from the Ministry of Education, Science, and Culture, Japan.

References

Alam MS, Andriana BB, Tay TW, Tsunekawa N, Kanai Y, Kurohmaru M. Single administration of di(*n*-butyl) phthalate delays spermatogenesis in prepubertal rats. Tissue Cell 2010a;42:129–35.

Alam MS, Ohsako S, Matsuwaki T, Zhu XB, Tsunekawa N, Kanai Y, et al. Induction of spermatogenic cell apoptosis in prepubertal rat testes irrespective of testicular steroidogenesis: a possible estrogenic effects of di(*n*-butyl) phthalate. Reproduction 2010b;139:427–37.

Alam MS, Ohsako S, Tay TW, Tsunekawa N, Kanai Y, Kurohmaru M. Di(*n*-butyl) phthalate induces vimentin filaments disruption in rat Sertoli cells: a possible relation with spermatogenic cell apoptosis. Anat Histol Embryol 2010c;39:186–93.

Billig H, Furuta I, River C, Tapanainen J, Parvinen M, Hsueh AJ. Apoptosis in testis germ cells: developmental changes in gonadotropin dependence and localization to selective tubule stages. Toxicology 1995;1:189–95.

Byford JR, Shaw LE, Drew MG, Pope GS, Sauer MJ, Darbre PD. Oestrogenic activity of parabens in MCF7 human breast cancer cells. J Steroid Biochem Mol Biol 2002;80:49–60.

Derache R, Gourdon J. Metabolism of a food preservation: parahydroxybenzoic acid and its esters. Food Cosmetic Toxicol 1963;1:189–95.

D'Souza R, Gill-Sharma MK, Pathak S, Kedia N, Kumar R, Balasinor N. Effect of high intratesticular estrogen on the seminiferous epithelium in adult male rats. Mol Cell Endocrinol 2005;241:41–8.

Dym M. Spermatogonial stem cells of the testis. Proc Natl Acad Sci U S A 1994;91:11287–9.

Koji T, Hishikawa Y. Germ cell apoptosis and its molecular trigger in mouse testes. Arch Histol Cytol 2003;66:1–16.

Koji T, Hishikawa Y, Ando H, Nakanishi Y, Kobayashi N. Expression of Fas and Fas ligand in normal and ischemia-reperfusion testes: involvement of the Fas system in the induction of germ cell apoptosis in the damaged mouse testis. Biol Reprod 2001;64:946–54.

Kondo T, Shono T, Suita S. Age-specific effect of phthalate ester on testicular development in rats. J Pediatr Surg 2006;41:1290–3.

Maeda Y, Shiratsuchi A, Namiki M, Nakanishi Y. Inhibition of sperm production in mice by annexin V microinjected into seminiferous tubules: possible etiology of phagocytic clearance of apoptotic spermatogenic cells and male infertility. Cell Death Differ 2002;9:742–9.

Mishra DP, Shaha C. Estrogen-induced spermatogenic cell apoptosis occurs via the mitochondrial pathway: role of superoxide and nitric oxide. J Biol Chem 2005;280:96–6181.

Oishi S. Lack of spermatotoxic effects of methyl and ethyl esters of *p*-hydroxybenzoic acid in rats. Food Chem Toxicol 2004;42:54–1845.

Oishi S. Effects of butyl paraben on the male reproductive system in mice. Arch Toxicol 2002;76:423–9.

Okubo T, Yokoyama Y, Kano K, Kano I. ER-dependent oestrogenic activity of parabens assessed by proliferation of human breast cancer MCF-7 cells and expression of ERa and PR. Food Chem Toxicol 2001;39:32–1225.

Pedersen KL, Pedersen SN, Christiansen LB, Korsgaard B, Bjerregaard P. The preservatives ethyl-, propyl- and butyl paraben are estrogenic in an in vivo fish assay. Pharmacol Toxicol 2000;86:110–3.

Pelletier G, Labrie C, Labrie F. Localization of oestrogen receptor alpha, oestrogen receptor beta and androgen receptors in the rat reproductive organs. J Endocrinol 2000;165:359–70.

Print CG, Loveland KL. Germ cell suicide: new insights into apoptosis during spermatogenesis. BioEssays 2000;22:423–30.

Prusakiewicz JJ, Harville HM, Zhang Y, Ackermann C, Voorman RL. Parabens inhibit human skin estrogen sulfotransferase activity: possible link to paraben estrogenic effects. Toxicology 2007;232:248–56.

Pugazhendhi D, Pope GS, Darbre PD. Oestrogenic activity of p-hydroxybenzoic acid (common metabolite of paraben esters) and methylparaben in human breast cancer cell lines. J Appl Toxicol 2005;25:301–9.

Routledge EJ, Parker J, Odum J, Ashby J, Sumpter JP. Some alkyl hydoroxybenzoate preservatives (parabens) are estrogenic. Toxicol Appl Pharmacol 1998;153:12–9.

Routledge EJ, Sumpter JP. Structural features of alkylphenolic chemicals associated with estrogenic activity. J Biol Chem 1997;272:8–3280.

Saikumar P, Dong Z, Mikhailov V, Denton M, Weinberg JM, Venkatachalam MA. Apoptosis: definition, mechanisms, and relevance to disease. Am J Med 1999;107:489–506.

Saunders PT, Fisher JS, Sharpe RM, Millar MR. Expression of oestrogen receptor beta (ER beta) occurs in multiple cell types, including some germ cells, in the rat testis. J Endocrinol 1998;156:13–7.

Shaha C. Modulators of spermatogenic cell survival. Soc Reprod Fertil Suppl 2007;63:173–86.

Song BL, Peng DR, Li HY, Zhang GH, Zhang J, Li KL, et al. Evaluation of the effect of butyl *p*-hydroxybenzoate on the proteolytic activity and membrane function of human spermatozoa. J Reprod Fertil 1991;91:435–40.

Soni MG, Carabin IG, Burdock GA. Safety assessment of esters of phydroxybenzoic acid (parabens). Food Chem Toxicol 2005;43:985–1015.

Tavares RS, Martins FC, Oliveira PJ, Ramalho-Santos J, Peixoto FP. Parabens in male infertility – isthere a mitochondrial connection? Reprod Toxicol 2009;27:1–7.

Taxvig C, Viggaard AM, Hass U, Axelstad M, Boberg J, Hansen PR, et al. Do parabens have the ability to interfere with steroidogenesis. Toxicol Sci 2008;106:206–13.

Tay TW, Andriana BB, Ishii M, Choi EK, Zhu XB, Alam MS, et al. Phagocytosis plays an important role in clearing dead cells caused by mono(2-ethylhexyl) phthalate administration. Tissue Cell 2007;39:241–6.

Tesarik J, Greco E, Cohen-Bacrie P, Mendoza C. Germ cell apoptosis in men with complete and incomplete spermiogenesis failure. Mol Hum Reprod 1998;4:757–62.

vom Saal FS, Cooke P, Buchaman DL, Palanza PP, Thayer KA, Nagel SC, et al. A physiologically based approach to the study of bisphenol A and other estrogenic chemicals on the size of reproductive organs, daily sperm production, and behaviour. Toxicol Ind Health 1998;14:239–60.

Yan W, Suominen J, Samson M, Jegou B, Toppari J. Involvement of Bcl-2 family proteins in germ cell apoptosis during testicular development in the rat and pro-survival effect of stem cell-factor on germ cells in vitro. Mol Cell Endocrinol 2000;165:115–29.

生体の科学　65(2)：154-158，2014

特集　細胞の少数性と多様性に挑む―シングルセルアナリシス
C．シングルセルアナリシスで見えること

単細胞技術に基づくiPS細胞の標準化

山根順子　丸山徹　藤渕航

これまでの生物学の常識を大きく覆した人工多能性幹細胞(iPS細胞；induced pluripotent stem cell)の発見がなされたのが2006年のことである[1]。iPS細胞はES細胞と異なり作製段階にヒト胚を破壊する必要がないことから，ES細胞を用いた研究において大きな障壁となっていた倫理問題が生じず，再生医療を一気に加速させる夢の細胞として登場した。体細胞にわずか数因子を導入するのみで多能性を持った細胞を生み出すことができるという報告はあまりにセンセーショナルであり，それ以降様々な細胞種由来のiPS細胞の樹立や，より安全かつ効率的な樹立法が次々に見いだされた。また，iPS細胞を用いた幹細胞生物学としての基礎研究や再生医療，創薬へ向けた応用研究など，多数の報告がなされている。世界中で樹立が試みられ報告されているiPS細胞は，樹立された数だけ質の異なる細胞になっている可能性が指摘され，今度は質の良いiPS細胞を選別する手法を開発するという新たな研究の方向性も生まれた。

iPS細胞は通常コロニー(細胞集団)として維持培養される。しかしながら，コロニーのなかでも均一な状態ではなく細胞の個性があることがわかっており，集団レベルでiPS細胞の解析を続けていくだけでは標準化を目指すことは難しい。そこでわれわれはより解像度を上げた解析が必要になると考え，従来のような“細胞集団”として遺伝子発現レベルを調べるのではなく，“個”としての細胞，つまり“シングルセルレベル”での遺伝子発現を調べ，標準化に向けた試みを行った。

Ⅰ．シングルセルトランスクリプトーム解析

2009年にTangら[2]によってシングルセルレベルでRNA-seqを行う方法が発表されてから，これまでに幾つかの手法が報告されてきた。われわれが行ったシングルセルトランスクリプトーム解析の手法は各細胞を判別するためのDNAバーコードをtemplate switchingにより導入するSTRT法[3]であり，同時に大量のシングルセルを解析できる利点がある。この方法では，mRNAの5'末端にDNAバーコードを導入するため，次世代シークエンサーで読まれたリードは5'末端側の配列であるという特徴がある。その他の手法で最近報告されたものでは，SMART-seq[4]やCEL-seq[5]，Quartz-seq[6]などがあり，そのうちCEL-seq，Quartz-seqは3'末端側のバイアスがあることが知られている。加えて，これまで問題とされてきたデータの精度についても，Quartz-seqでは細胞周期が区別できるまで改善されており，この先シングルセルトランスクリプトーム解析が大幅に普及することが予想される。

Standardization of iPS cells by single-cell transcriptome analysis

Yamane Junko：京都大学iPS細胞研究所(特定研究員：〒606-8507 京都市左京区聖護院川原町53)
Maruyama Toru：早稲田大学大学院 先進理工学研究科 生命医科学専攻(修士課程)，京都大学iPS細胞研究所(特別研究学生)
Fujibuchi Wataru：京都大学iPS細胞研究所(教授)

0370-9531/04/¥500/論文/JCOPY

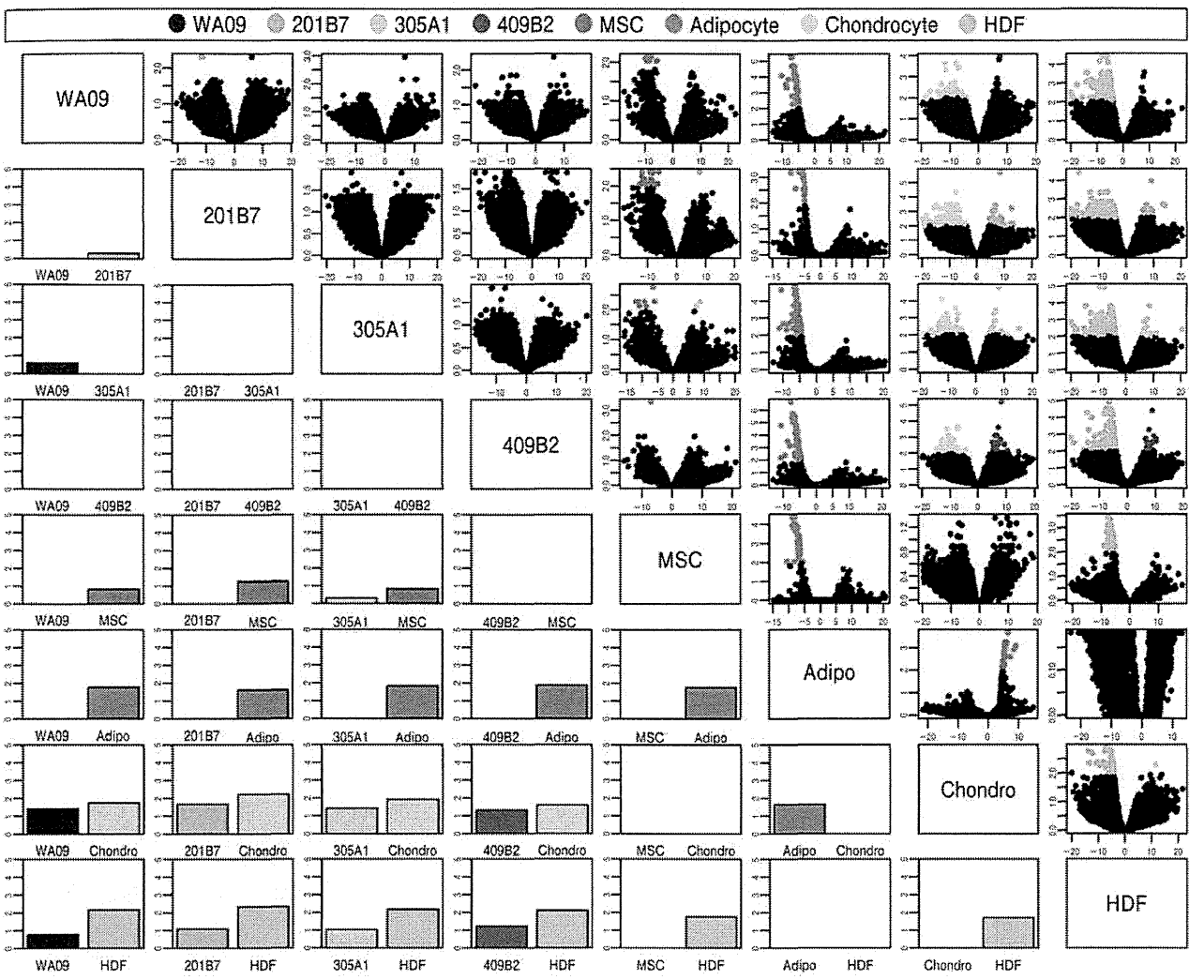

図1　DEG解析

対角線より上：Volcano plot(横軸：$\log_2$(FC)，縦軸：$\log_{10}$(FDR))．色がついたプロットは有意差(FDR<0.01)があった遺伝子。対角線より下：DEGの数を示す棒グラフ。プロット・棒の色は細胞種に対応している。例えば赤いプロットはAdipocyteで発現が高い遺伝子を，赤い棒はAdipocyteで発現が高い遺伝子の数を表す。

Ⅱ．シングルセルトランスクリプトーム解析により見えてきた特徴

われわれは異なる手法によって樹立されたiPS細胞3種類(201B7，305A1，409B2)に加えてES細胞1種類(WA09)，ヒトの体細胞4種類(MSC：間葉系幹細胞，Chondro：軟骨細胞，adipo(誘導させた)脂肪細胞，HDF：皮膚線維芽細胞)の計8細胞種を対象にしてシングルセルレベルのトランスクリプトーム解析を行った。

まず，BioconductorのedgeRパッケージを用いて2細胞種間で真の発現変動遺伝子(DEG)の数を調べたところ，ES細胞とiPS細胞間で発現変動がある遺伝子はほとんど見つからなかった(図1)。しかし，本解析は細胞集団を対象にした手法を適用しているので，シングルセルレベルでの解析とは結果が異なる可能性がある。

そこでシングルセルレベルでの各種細胞における遺伝子発現パターンの特徴をより詳細に調べるため，8細胞種の主成分分析(PCA)を行ったところ，おおよそ細胞種ごとにクラスターを形成していることが確認できた(図2)。この結果から軟骨細胞とHDFのクラスターが重なっていることがわかったが，これらの細胞種間では遷移が起こることが知られている[7]。よって，単一細胞のトランスクリプトームには細胞系譜における細胞間の関係性や，細胞間で遷移が起こりうる可能性といった新たな情報が包含されている可能性が示唆

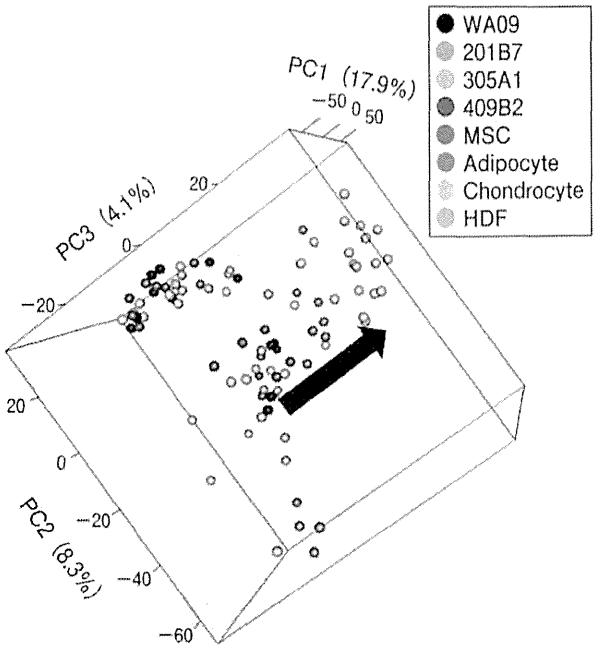

図 2　8細胞種の主成分解析(PCA)
細胞ごとにある程度クラスターができている。括弧内は各種成分の寄与率を示す。HDFとChondrocyteが混在しているが，これはHDFとChondrocyteの間でトランジションが起こりやすいことを示唆していると考えられる。また，矢印の方向に相転移が起こりうることを示唆している。

された。これまでの集団として解析してきた方法では平均化されてしまい見えていなかった細胞本来のばらつきの様相の一端が明らかになった。

また，遺伝子発現の揺らぎの大きさを評価するために，細胞種ごとに全遺伝子の変動係数(CV)を調べた(図3)。細胞集団での解析と一致して，すべての細胞でハウスキーピング遺伝子のCVは他の遺伝子と比較して遺伝子発現の揺らぎが小さいという結果が得られた。微量なサンプルを対象とするシングルセル解析を行ううえで，得られた結果が細胞由来の揺らぎによるものか，テクニカルな問題により生じる差を見ているのかを見極めることは非常に重要である。よって，指標の一つとなるハウスキーピング遺伝子の発現の揺らぎが他の遺伝子と比べて小さいという結果はシングルセルを扱う研究を進めるにあたり精度の良さを確認するための大きな情報となり得る。

最後に，全遺伝子の変動係数の解析において今回調べた細胞間でヒストグラム分布の異なる遺伝子について一部報告する。まず，DNAメチル化にかかわるDNMT3Bは既報のように[8]，ES細胞，iPS細胞の多能性幹細胞群ではほぼ共通し発現レ

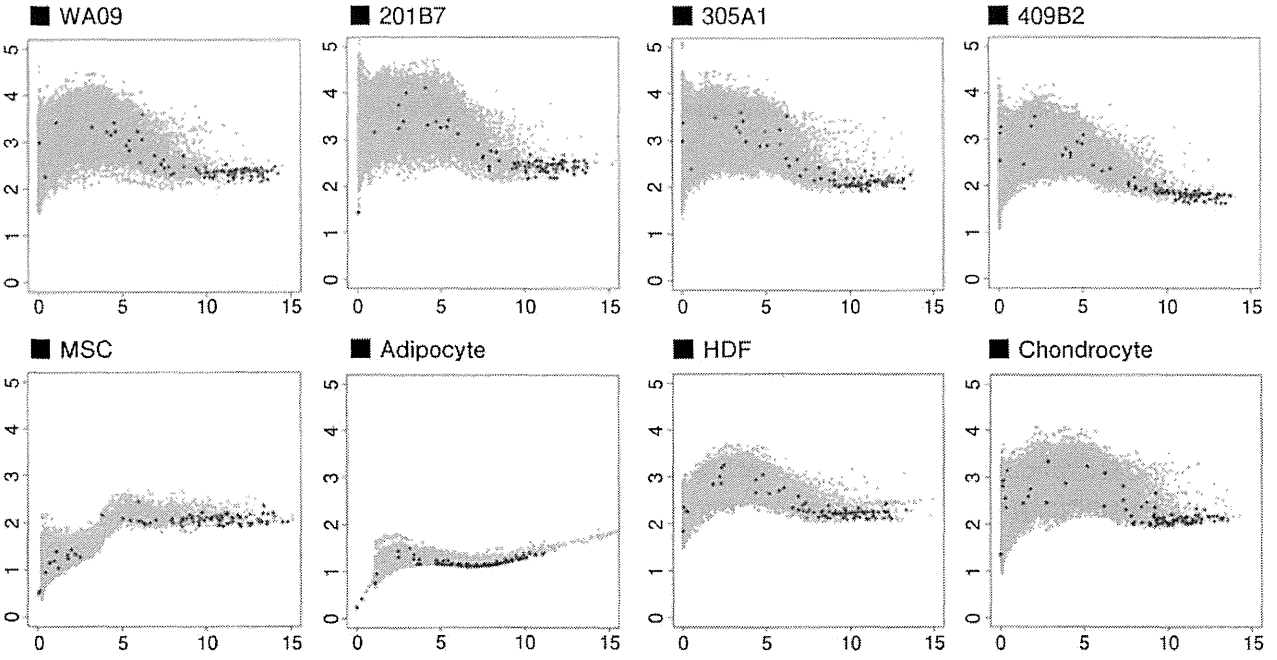

図 3　全遺伝子の変動係数(CV)
各点は遺伝子に対応しており，赤点はハウスキーピング遺伝子に対応している。横軸は$\log_2$(RPM)の平均，縦軸はedgeRによって算出した発現の変動(biological coefficient of variation)を表している。下に位置する遺伝子ほど安定に発現していることを意味する。

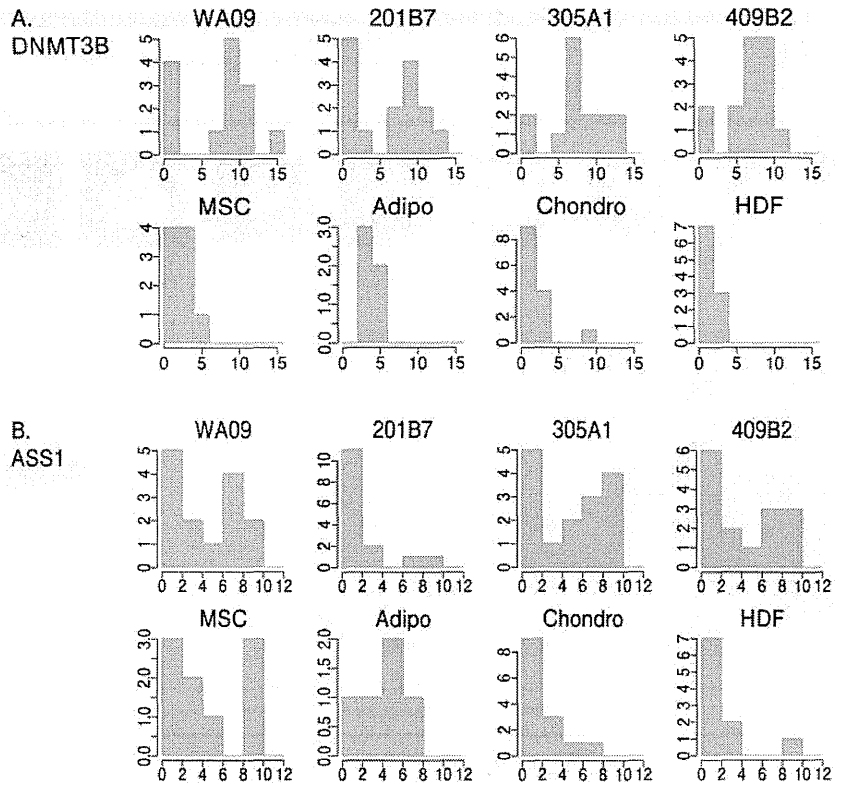

図 4　ヒストグラムの分布で差が見られた遺伝子例(横軸は $\log_2$(RPM + 1))
A. は多能性幹細胞と分化細胞の間で顕著な差が見られた。B. は多能性幹細胞の中の比較で一部の株(201B7)で異なる分布が見られた。

ベルの高い分布が見られ，逆に分化細胞ではいずれも発現レベルの低い分布が傾向として見られた(図 4A)。また，アルギニン生合成経路にかかわる酵素の一つである ASS1 は多能性幹細胞群の中で比較をした場合，201B7 株でのみ低発現の分布を示している傾向があった。この遺伝子は分化細胞でも比較的低発現の分布を示している(図 4B)。このように平均で解析していると見えてこなかった遺伝子発現の分布が，個々のレベルで見ると差が見えてくるケースがある。

以上のように，われわれはシングルセルトランスクリプトーム解析によって，これまでの細胞集団を対象にする解析では見えなかった事象を確認することができた。

おわりに

シングルセル RNA-seq 解析が普及するに伴い，これまで 250 種類程度と言われてきた細胞がより細かく分類されるようになる可能性がある。そこで，われわれの研究室では細胞のデータベースである“SHOGoiN”を開発しており，細胞の遺伝子発現プロファイル，画像，形態計測情報，実験条件などのメタデータ，文献情報を貯蔵している(図 5)。ユーザーはウェブブラウザを通して本データベースにアクセスし，細胞に関する情報を検索することができる。

今後，われわれはシングルセル RNA-seq 解析と体系的に貯蔵された細胞のデータを用いて，細胞に普遍的に存在する理論を見つけることを目指している。その一つとして次のような問題に取り組んでいる。

iPS 細胞を作製する際には，元となる体細胞に *Oct4*, *Sox2*, *Klf4*, *c-Myc* などの遺伝子を導入し，強制的に発現させる必要がある。しかし，すべての細胞に同じ刺激を与えているにもかかわらず，なぜ iPS 細胞の作製効率は悪く，質の良し悪しがあるのだろうか。われわれはこの問題に対して以下のようなモデルを考えている。遺伝子の発現は同じ種類の細胞であっても均一でなく，揺らぎがある。揺らぎは転写因子の結合や転写反応，

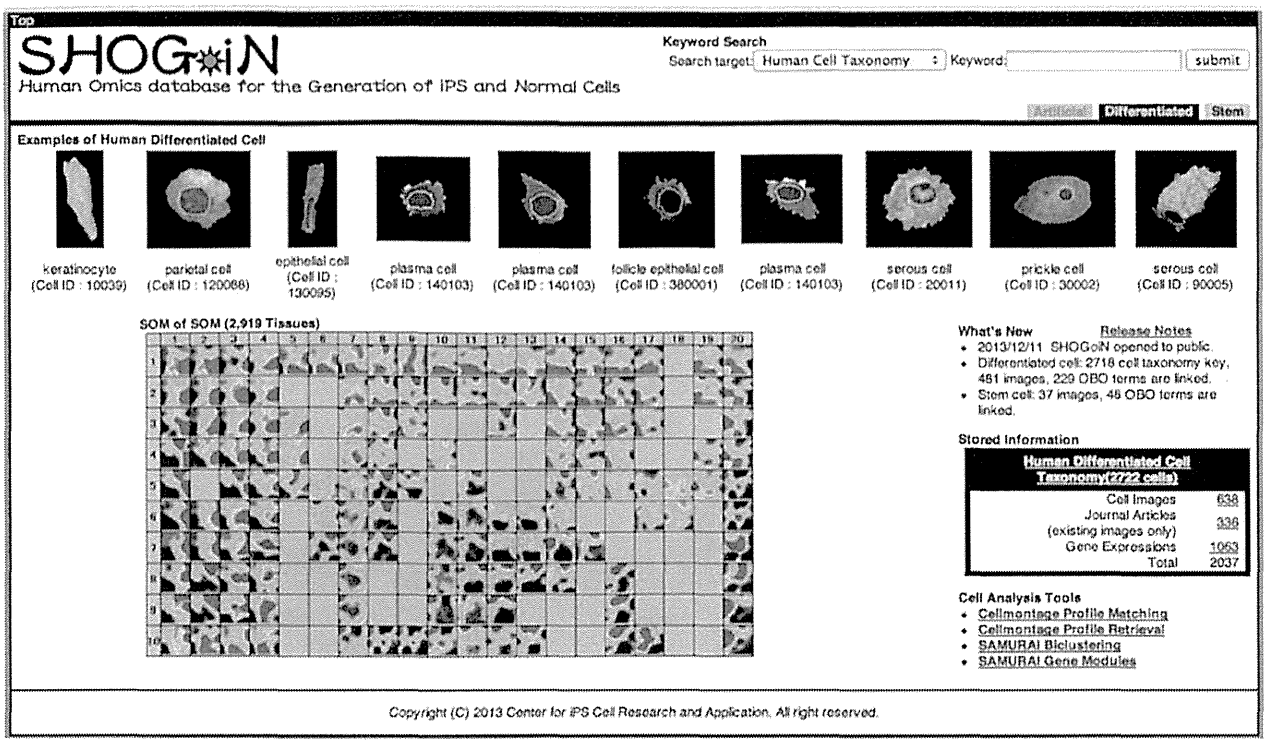

図5　ヒト細胞情報統合データベース“SHOGoiN”
http://shogoindb.cira.kyoto-u.ac.jp

RNAの分解といったイベントが確率的であることに起因している[9]。Huangらの論文[10]にもあるように，われわれはこういった遺伝子発現の揺らぎが細胞状態の変化に影響を与えているため，細胞分化やiPS細胞の作製が確率的なものになっているのではないかと考えている。よって，揺らぎが細胞分化に与える影響を解き明かすことができれば，質の良い細胞をより効率的に作製する手法が明らかになるかもしれない。また，シングルセル解析により得られた詳細な情報をデータベースに集約し，広く研究者らが利用することで，iPS細胞の標準化や，そこから分化させた各種分化細胞の標準化に向けた足がかりとなることを期待している。

謝辞　本研究は『内閣府最先端研究開発支援プログラム（FIRSTプログラム）』の支援のもと行いました。

●文　献

1) Takahashi K, Yamanaka S：*Cell* **126**：663-676, 2006
2) Tang F, Barbacioru C, Wang Y et al：*Nat Methods* **6**：377-382, 2009
3) Islam S, Kjällquist U, Moliner A et al：*Genome Res* **21**：1160-1167, 2011
4) Ramsköld D, Luo S, Wang Y-C et al：*Nat Biotech* **30**：777-782, 2012
5) Hashimshony T, Wagner F, Sher N et al：*Cell Rep* **2**：666-673, 2012
6) Sasagawa Y, Nikaido I, Hayashi T et al：*Genome Biol* **14**：R31, 2013
7) Outani H, Okada M, Yamashita A et al：*ProS One* **8**：e77365, 2013
8) Takahashi K, Tanabe K, Ohnuki M et al：*Cell* **131**：861-872, 2007
9) Wills QF, Livak KJ, Tipping AJ et al：*Nat Biotechnol* **31**：748-752, 2013
10) Chang HH, Hemberg M, Barahona M et al：*Nature* **453**：544-547, 2008

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

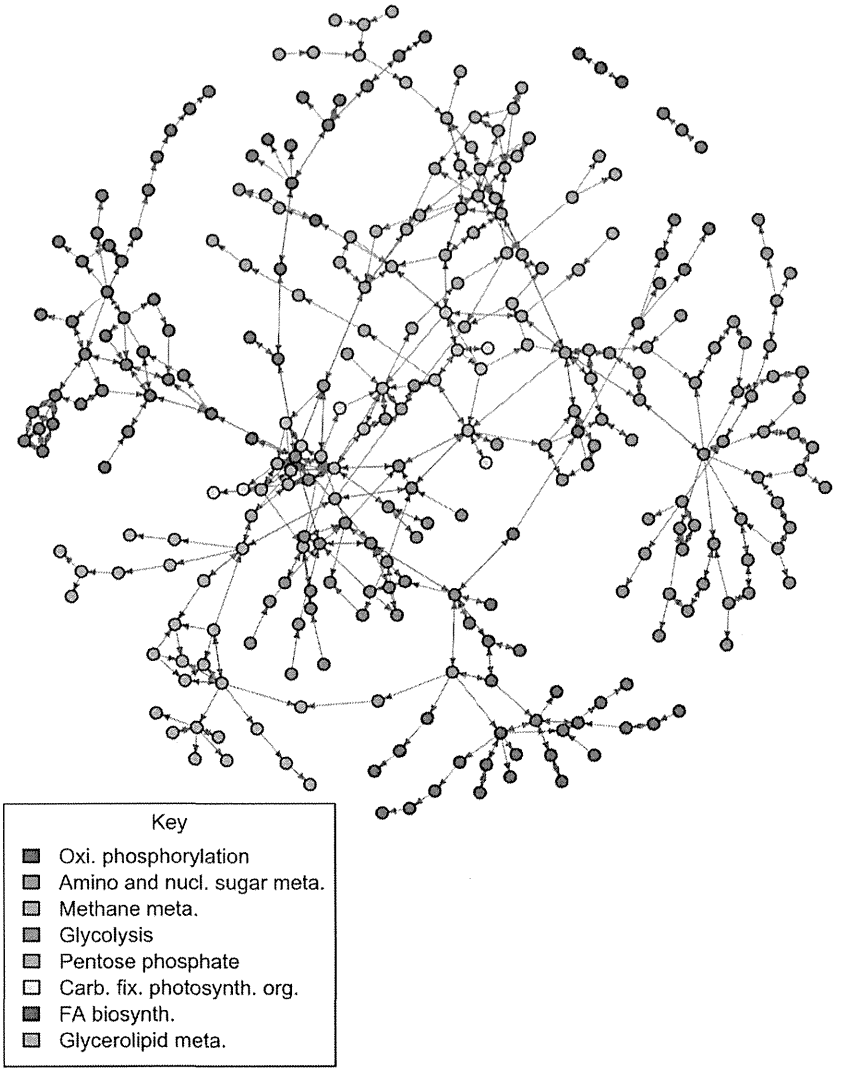

Figure 1. The graph of the significantly enriched pathways found using our GSEA method combined with the glycerolipid pathway. The full network contains 307 compounds and 558 reactions but compounds without reaction data were not drawn to reduce clutter. The graph is plotted with compounds drawn as nodes 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

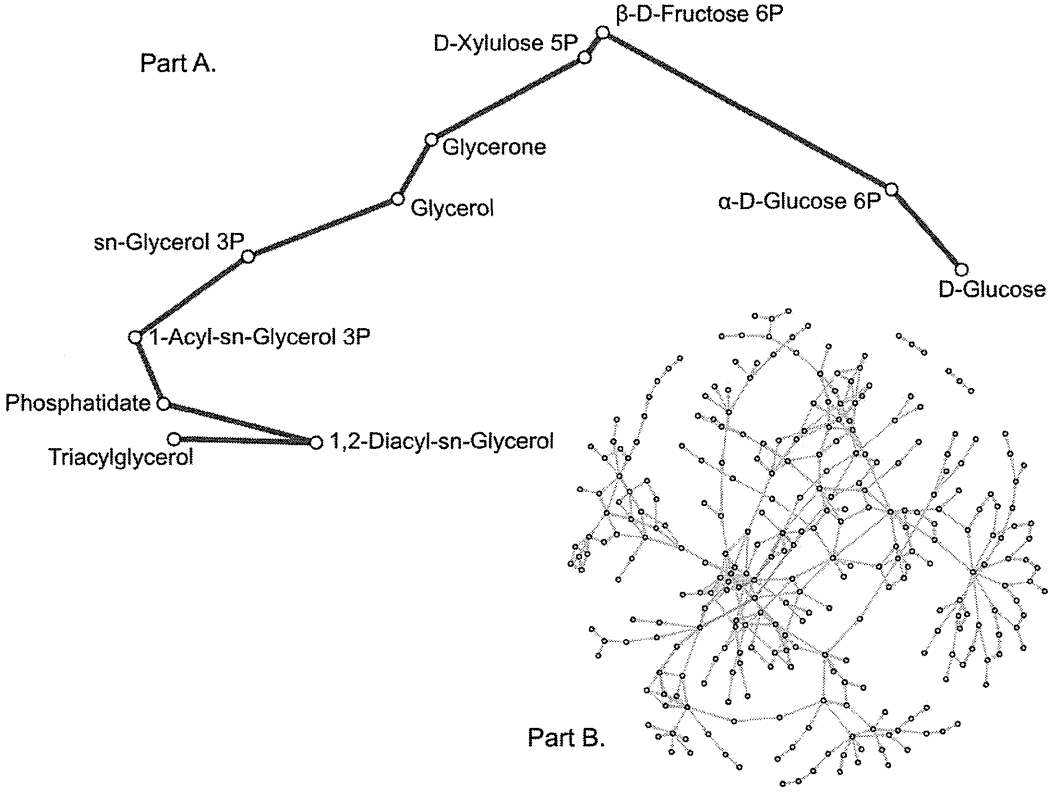

Figure 2. The first shortest path found in our graph between glucose and triacylglycerol using breadth-first search. A. This is the detailed view of the path showing the names of the compounds involved at each step. B. The shortest path is highlighted in green on the full graph to show its location. 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

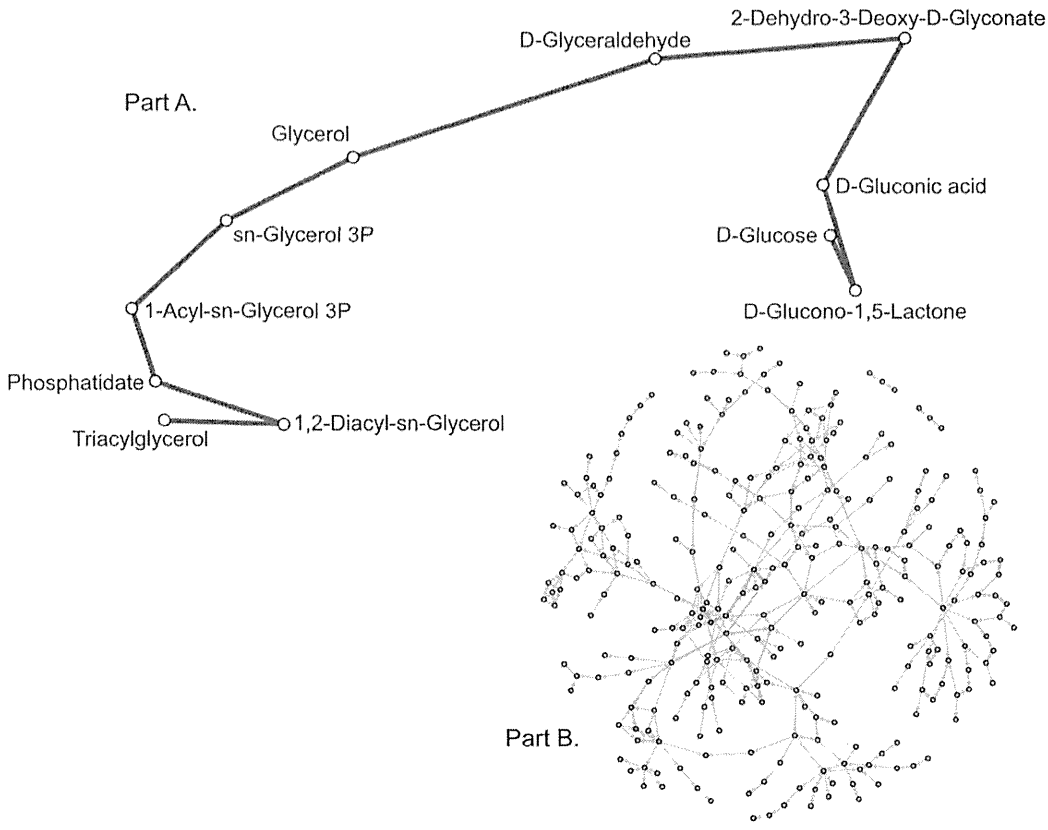

Figure 3. The second shortest path found in our graph between glucose and triacylglycerol using breadth-first search. A. This is the detailed view of the path showing the names of the compounds involved at each step. B. The shortest path is highlighted in green on the full graph to show its location. 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 $\mathbf{FCM}_s$.

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

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

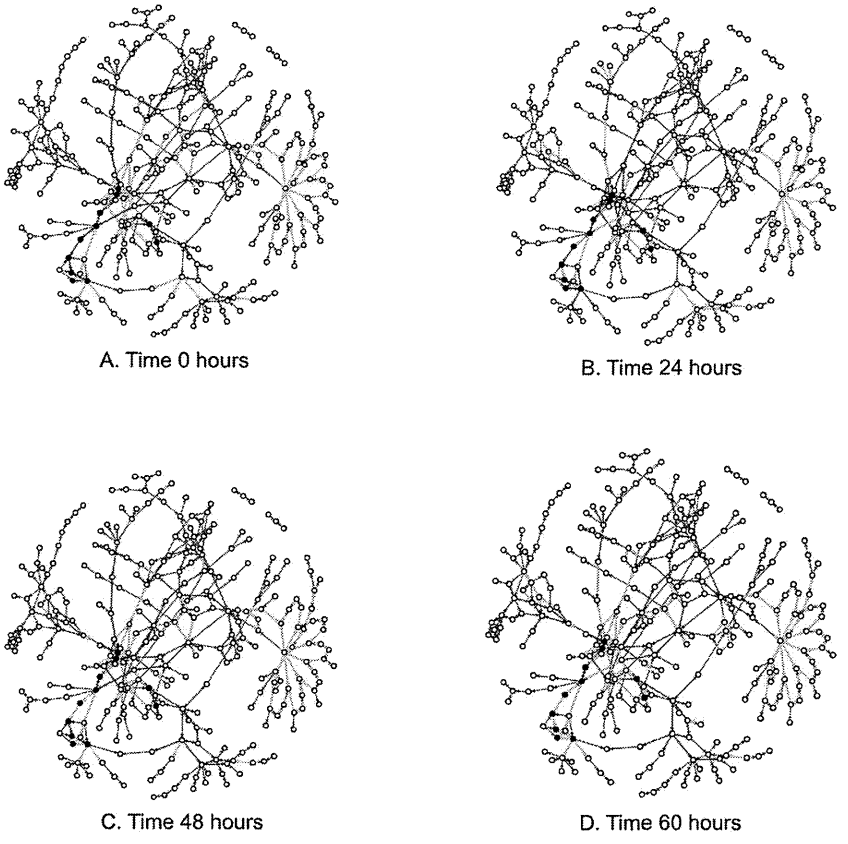

Figure 4. These graphs highlight the fold change direction of known genes in our data in response to oil accumulating conditions at each time point. A gene involved in a reaction is represented by an edge while the compounds in a reaction are represented by the nodes. Genes that were up-regulated during 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.

Analytical Biochemistry 472 (2015) 75–83

Contents lists available at ScienceDirect

Analytical Biochemistry

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

A set of external reference controls/probes that enable quality assurance between different microarray platforms

Hideo Akiyama [a,*], Yoji Ueda [a], Hitoshi Nobumasa [a], Hiroyuki Ooshima [b], Yohei Ishizawa [c], Koji Kitahiro [d], Isao Miyagawa [d], Kazufumi Watanabe [e], Takazumi Nakamura [f], Ritsuka Tanaka [g], Nobuko Yamamoto [h], Hiroki Nakae [h], Mitsuo Kawase [i], Nobuhiro Gemma [j], Yuji Sekiguchi [k], Wataru Fujibuchi [l], Ryo Matoba [c,*]

[a] New Projects Development Division, Toray Industries, Inc., Kamakura, Kanagawa 248-8555, Japan
[b] Yokohama Research Laboratories, Mitsubishi Rayon Co. Ltd., Yokohama, Kanagawa 230-0053, Japan
[c] DNA Chip Research Inc., Yokohama, Kanagawa 230-0045, Japan
[d] Technical Research Laboratory, Kurabo Industries Ltd., Neyagawa, Osaka 572-0823, Japan
[e] Hokkaido System Science Co. Ltd., Sapporo, Hokkaido 001-0932, Japan
[f] S-BIO Business Division, Sumitomo Bakelite Co. Ltd., Amagasaki, Hyogo 661-8588, Japan
[g] Department of Bio Research, Kamakura Techno-Science, Inc., Kamakura, Kanagawa 248-0036, Japan
[h] Japan Multiplex bio-Analysis Consortium (JMAC), Chiyoda-ku, Tokyo 102-0083, Japan
[i] Graduate School of Biomedical Engineering, Tohoku University, Sendai, Miyagi 980-8579, Japan
[j] Ricoh Institute of Technology, Ricoh Company, Ltd., Yokohama, Kanagawa 224-0035, Japan
[k] Bio-Measurement Research Group, Biomedical Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba Central 6, Ibaraki 305-8566, Japan
[l] Computational Biology Research Center, Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo 135-0064, Japan

ARTICLE INFO

Article history:
Received 18 August 2014
Received in revised form 16 November 2014
Accepted 19 November 2014
Available online 4 December 2014

Keywords:
External RNA standards
DNA microarray
Dynamic range
Cross-platform
Multiplatform calibration

ABSTRACT

RNA external standards, although important to ensure equivalence across many microarray platforms, have yet to be fully implemented in the research community. In this article, a set of unique RNA external standards (or RNA standards) and probe pairs that were added to total RNA in the samples before amplification and labeling are described. Concentration–response curves of RNA external standards were used across multiple commercial DNA microarray platforms and/or quantitative real-time polymerase chain reaction (RT–PCR) and next-generation sequencing to identify problematic assays and potential sources of variation in the analytical process. A variety of standards can be added in a range of concentrations spanning high and low abundances, thereby enabling the evaluation of assay performance across the expected range of concentrations found in a clinical sample. Using this approach, we show that we are able to confirm the dynamic range and the limit of detection for each DNA microarray platform, RT–PCR protocol, and next-generation sequencer. In addition, the combination of a series of standards and their probes was investigated on each platform, demonstrating that multiplatform calibration and validation is possible.

Recent advances in DNA microarray technology have opened up new applications in both basic and clinical research [1–4]. Consequently, new tests in many areas of biomedical science, including clinical pharmacogenetics, cancer genotyping, and cancer prognosis, have been developed [5–7].

Clinical applications of DNA microarray technology include gene expression analysis for early disease detection, disease classification and diagnosis, selection of treatment protocol, determination of changes in disease status, and the monitoring of therapeutic effects and side effects. A clinical application in which DNA microarray gene expression analysis has already been applied is the "MammaPrint," developed in the United States and Europe, used to select the optimal breast cancer treatment [5]. In addition, OncoType DX, a product based on quantitative real-time polymerase chain reaction (RT–PCR)[1], has also been used for analyzing the expression of multiple RNA targets as an indicator in the selection of optimal breast cancer treatment [6].

* Corresponding authors.
E-mail addresses: hideo_akiyama@nts.toray.co.jp (H. Akiyama), matoba@dna-chip.co.jp (R. Matoba).

[1] *Abbreviations used:* RT–PCR, real-time polymerase chain reaction; HURR, human universal reference total RNA; HBRR, human brain reference total RNA; JMAC, Japan Multiplex bio-Analysis Consortium; cDNA, complementary DNA; 3D, three-dimensional; aRNA, antisense amplified RNA; SSC, sodium saline citrate; SDS, sodium dodecyl sulfate; PBS, phosphate-buffered saline; mRNA, messenger RNA.

http://dx.doi.org/10.1016/j.ab.2014.11.012
0003-2697/

However, if DNA microarray data are to be routinely used for clinical applications, it is vital that the data are both reliable and reproducible and that errors or ambiguities in the interpretation of results are eliminated [8–10]. In particular, because gene expression is highly variable, quality assurance in the handling of specimens—storage conditions, transport conditions, and pretreatment protocols—must be robust (Fig. 1).

We report here the development of a set of unique RNA external standards (or RNA standards) and probe pairs that may be spiked into test samples to ensure equivalence across many microarray platforms. This suite of synthetic nucleotides is derived from unique non-mammalian sequences and designed to minimize cross-hybridization with common transcripts from humans, mice, and rats. Six microarray platforms were evaluated using this set of standards: 3D-Gene (Toray Industries, Tokyo), Agilent SurePrint (Agilent Technologies, Santa Clara, CA, USA), Genopal (Mitsubishi Rayon, Tokyo), GeneSQUARE (Kurabo Industries, Osaka, Japan), S-Bio (Sumitomo Bakelite, Tokyo), and NimbleGen (Roche NimbleGen, Basel, Switzerland). An RT–PCR protocol (Life Technologies, Foster City, CA, USA) and a next-generation sequencer GAII (Illumina, San Diego, CA, USA) were also tested. We compared performance across DNA microarray platforms and/or RT–PCR and next-generation sequencing by spiking a set of our standards into a commonly available commercial total RNA sample. A variety of standards can be added in a range of concentrations spanning high and low abundances, thereby enabling the evaluation of assay performance across the expected range of concentrations found in a clinical sample.

Using this approach, we show that we are able to confirm the dynamic range and the limit of detection for each DNA microarray platform, RT–PCR protocol, and next-generation sequencer. In addition, the combination of a series of standards and their probes was investigated on each platform, demonstrating that multiplatform calibration and validation is possible (Fig. 2).

Materials and methods

RNA external standard transcripts

Ten candidate external RNA standard clones (in pUC19 plasmid) were synthesized from artificial sequences designed to have the following characteristics: (i) a unique sequence that exhibits low similarity with any eukaryotic genome and EST sequence known to date, (ii) no nucleic acid homopolymer longer than three bases, (iii) a G+C content in the range of 40 to 60%, (iv) no repeated sequences such as a motif, and (v) no strong secondary structure within the sequence. The standard sequences were designed by using our original program software. Inserts for the clones are 500 to 1000 bp with a 30-bp polyadenylated tail and T7 promoter sequence. All candidate standards were prepared by in vitro transcription of linearized plasmids using a T7 RNA polymerase (MEGAScript Kit, Life Technologies, Carlsbad, CA, USA) according to the manufacturer's instructions. Ten transcripts corresponding to the RNA external standards were purified using TURBO DNase (Life Technologies) and further purified by phenol–chloroform extraction and ethanol precipitation. The 10 standard transcripts were dissolved in RNase-free water and then quantified using a Quant-iT RNA Assay Kit (Life Technologies). The sequences of the external standards (R001-500 to R010-1000) have been deposited in the DDBJ/GenBank/EMBL databases under the accession numbers AB610939 to AB610950.

RNA external standard spiked total RNA cocktail

Human universal reference total RNA (HURR, Agilent Technologies) and human brain reference total RNA (HBRR, Agilent Technologies) controls were used. Ten external RNA standards were diluted using HURR or HBRR RNA solution at 50 ng/ml. The standard spiked total RNA cocktail (see Supplementary Tables S1 and S2 in online supplementary material) was prepared at the Japan Multiplex bio-Analysis Consortium (JMAC) central laboratory and delivered to each test site.

Design of probe for RNA external standards

For probe design, each external standard was divided into two regions as follows: 1- to 300-nt and 301- to 500-nt regions for 500-nt RNA and 1- to 500-nt and 501- to 1000-nt regions for 1000-nt RNA, numbering from their 3′ ends. All candidate sequences from the sense strand were extracted by moving 60-nt windows in each region.

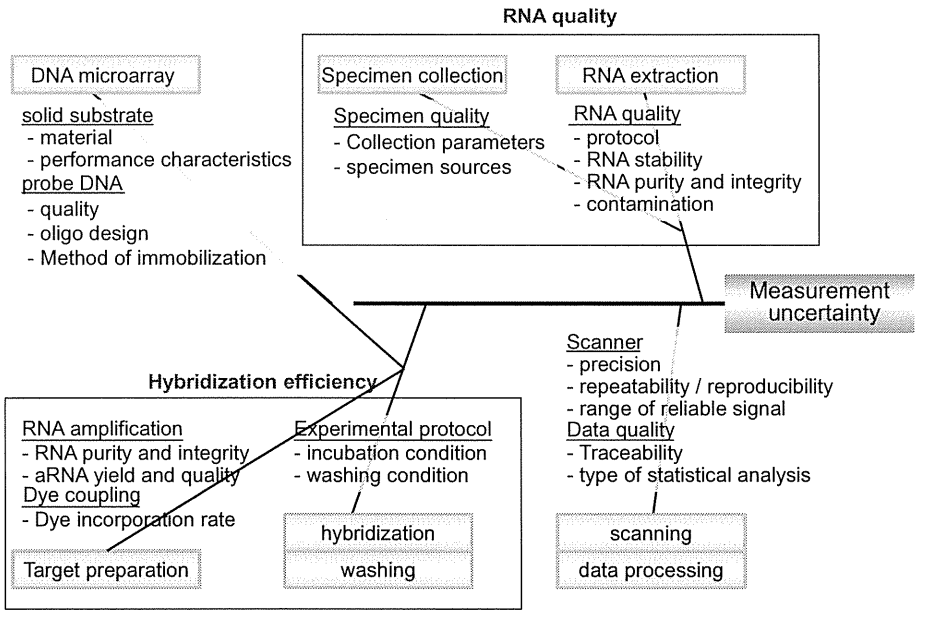

Fig.1. Measurement uncertainty of DNA microarray analysis. Unless the uncertainties of a measurement are being evaluated and stated, the fitness for the purpose of measurement cannot be judged properly. The uncertainties of a measurement using microarray are complicated and intertwined. The sources of uncertainties come from mainly the platform material, RNA quality, and hybridization efficiency and during data acquisition and processing.

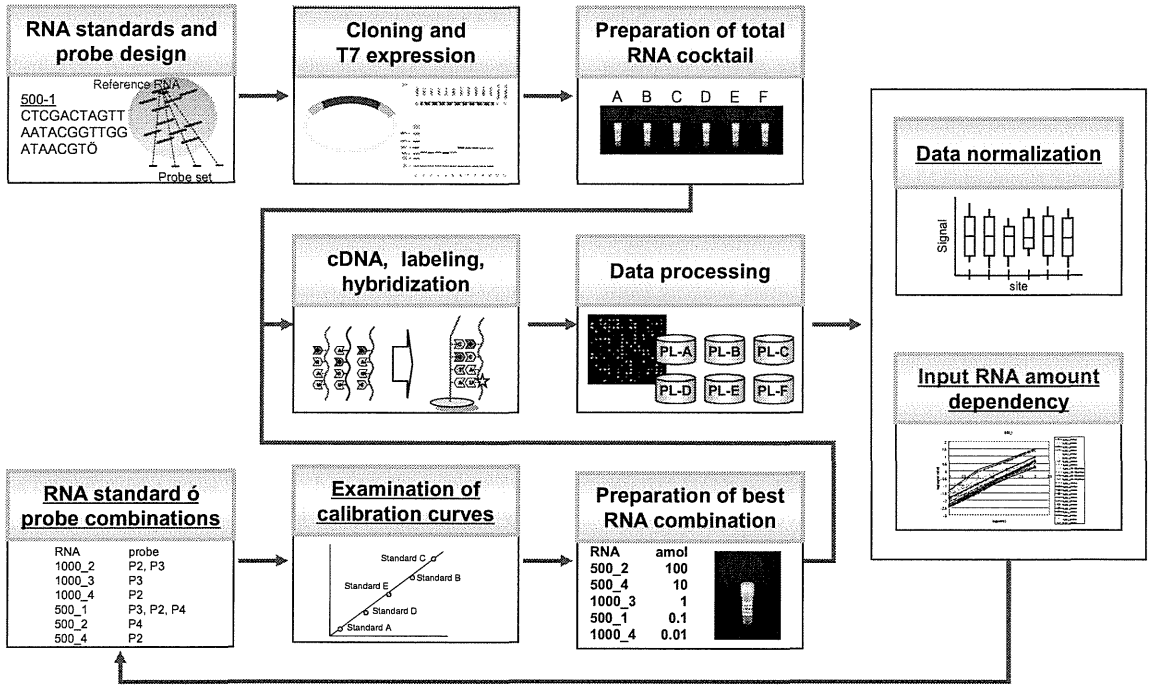

Fig.2. Experimental design of data comparison and quality assurance among different microarray platforms. Spiking RNA external standards into testing samples is an essential and effective method to monitor the quality of a microarray experiment, starting from sample preparation, hybridization, to data analysis.

First, the cross-hybridization potentials of candidate probes against off-target RNAs were analyzed. Because exact calculation based on thermodynamics requires a large computational cost, the search program in the FASTA package [11] (version 3.5) with default parameters to align candidate probes against both strands of 56,155 human complementary DNA (cDNA) sequences from the Ensembl database (http://www.ensembl.org, release 49) was used, and the top 100 off-target cDNA sequences that have the closest similarity to each probe based on the alignment scores were selected for potential cross-hybridization targets. To refine these results, the hybrid-min program in the UNAFold package [12] was performed to calculate the free energy change of hybridization, and then a program to calculate the cross-hybridization ratios for each probe based on Ref. [13] was coded. A cross-hybridization ratio $\geqslant$10–3 was removed from the candidate probes.

Second, T_m values using the nearest neighbor method [14] were calculated, and four candidate probes for each RNA standard that had the closest T_m to 80 °C were selected (Supplementary Table S3A). Potentials for dimerization and secondary structure formation were also calculated by hybrid-min and hybrid-ss in UNAFold. For thermodynamic calculations, 0.5 μM of primers, 2 mM Mg^{2+}, and 50 mM Na^+ parameters were used.

DNA microarray platform analyses

3D-Gene

The custom DNA microarray was constructed using the 3D-Gene platform (Toray Industries) [15] and spotted with the DNA probes (140 probes) shown in Supplementary Tables S3A and S3B. The 3D-Gene platform has a three-dimensional (3D) array that is constructed within a well with the oligonucleotide probes on the top. A total RNA cocktail (0.5 μg) was amplified and labeled using an Amino Allyl MessageAmp II aRNA Amplification Kit (Life Technologies) according to the manufacturer's instructions. Each sample of aRNA (antisense amplified RNA) labeled with Cy5 was hybridized with 3D-Gene at 37 °C for 16 h. After hybridization, the DNA microarray was washed and dried. Hybridization signals derived from Cy5 were scanned using Scan Array Lite (PerkinElmer, Waltham, MA, USA). The scanned image was analyzed using GenePix Pro 6.0 software (Molecular Devices, Sunnyvale, CA, USA). Spots that might be associated with artifacts were eliminated using software- and visual-guided flags. In this study, the background (blank) average was subtracted from the median values of the foreground signals that are higher than the background (blank) average + 2 standard deviations to give a feature intensity.

Agilent SurePrint

The custom microarray used in this study was designed using the Agilent e-Array platform (Agilent Technologies). Total RNA cocktail (0.5 μg) was used as a starting material to prepare Cy3-labeled aRNA. Fluorescently labeled aRNA was produced using the Quick Amp Labeling Kit (Agilent Technologies) and purified using the RNeasy Mini Kit (Qiagen, Hilden, Germany). The Cy3-labeled 600-ng aRNA was fragmented and hybridized at 65 °C for 17 h to microarray platform slides using the Agilent Gene Expression Hybridization Kit (Agilent Technologies). The microarray platform slides were washed and scanned with an Agilent scanner. The fluorescent intensities of individual spots were obtained with Feature Extraction (version 10.5.1.1, Agilent Technologies).

Genopal

The custom oligonucleotide microarray, Genopal (Mitsubishi Rayon), was made in the following manner. Plastic hollow fibers were bundled in an orderly arrangement, and hardened with resin to form a block. Oligonucleotide capture probes (140 probes) were chemically bonded inside each hollow fiber with hydrophilic gel [16]. The block was then sliced to make thin microarray platforms, each of which was set into a holder (for details, see http://www.mrc.co.jp/genome/e).

Total RNA cocktail was amplified using the MessageAmp II Biotin-Enhanced Amplification Kit (Life Technologies) according to the manufacturer's instructions, and was column purified. Biotinylated RNA (5 μg) was fragmented by incubation with fragmentation reagents (Life Technologies) at 94 °C for 7.5 min. Hybridization was carried out with DNA microarray in 150 μl of hybridization buffer (0.12 M Tris–HCl, 0.12 M NaCl, and 0.05% Tween 20) and