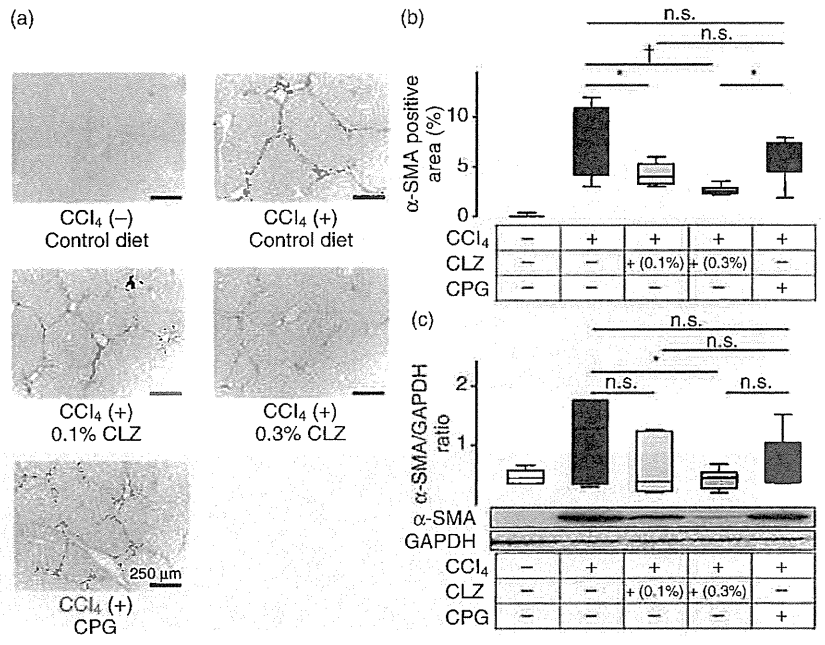

Figure 4 Cilostazol attenuated the expression of α-SMA protein in the liver. (a) α-SMA immunostaining of liver sections in each group. Treatment with CCl_4 for six weeks remarkably increased α-SMA expression. Among CCl_4-treated groups, the liver in cilostazol-administrated groups has a reduced α-SMA-positive area compared with that in the control diet or clopidogrel-administrated group (original magnification ×100). (b) Quantification of the α-SMA positive area in each group. Cilostazol-administrated groups had significantly decreased α-SMA positive areas compared with control diet and clopidogrel-administrated groups. (c) Measurement of α-SMA protein in the liver by immunoblotting. Administration of 0.3% cilostazol reduced α-SMA levels in CCl_4-treated mice. The box plots present the median and 25th–75th percentiles. Upper and lower lines represent the minimum and maximum values ($n = 10$). $^*P < 0.05$; $\dagger P < 0.001$ vs CCl_4-treated control diet group. α-SMA, α-smooth muscle actin; CCl_4, carbon tetrachloride; CLZ, cilostazol; CPG, clopidogrel; GAPDH, glyceraldehyde 3-phosphate dehydrogenase.

primary cultures of HSC. Generally, isolated HSC undergo autonomous activation in culture, and the activation is associated with a depletion of retinoid droplets, morphological change, cell proliferation, and expression of several activation markers such as α-SMA and collagen α1 (I). During 6 days of culture, control HSC gradually lost retinoid droplets and showed myofibroblast-like activated morphology, whereas cilostazol-supplemented HSC maintained retinoid droplets and retained quiescent morphology (Fig. 5a). In addition, cilostazol suppressed HSC proliferation in a dose-dependent manner, without showing cell toxicity (Fig. 5b). The expression of α-SMA protein was dose-dependently suppressed in the presence of cilostazol (Fig. 5c). Because Kupffer cells have been shown to be implicated in liver fibrosis as well as HSC,[11,38–40] we examined the effect of cilostazol on Kupffer-cell activation *in vivo* and *in vitro*. Pathological examination revealed a weak tendency for decrease in the F4/80 positive (Kupffer cell) area in the liver of cilostazol-administrated mice, however, we could not detect significant changes in our experimental setting (Fig. 5d,e), and no significant change was observed in mRNA levels of tissue TNF-α and TGF-β1 in the liver by cilostazol treatment (Fig. 5f). In fact, cilostazol did not affect the mRNA expression of F4/80 in isolated Kupffer cells (Fig. 5g), suggesting the minimal effect of cilostazol *in vivo* may be simply explained by the secondary effect of the resolution of fibrosis. Likewise, cilostazol exhibited an insignificant effect on the Kupffer cell production of TNF-α, IL-1β, MCP1 and TGF-β1 (Fig. 5g). These data together propose the notion that the *in vivo* therapeutic efficacy of cilostazol is mediated, at least in part, by its direct effects on HSC. If so, why did HSC respond well to cilostazol? One possible explanation is that HSC are more sensitive to cilostazol than other cell types (e.g., Kupffer cells). Actually, cilostazol-induced cAMP accumulation, which is an indicator for cilostazol

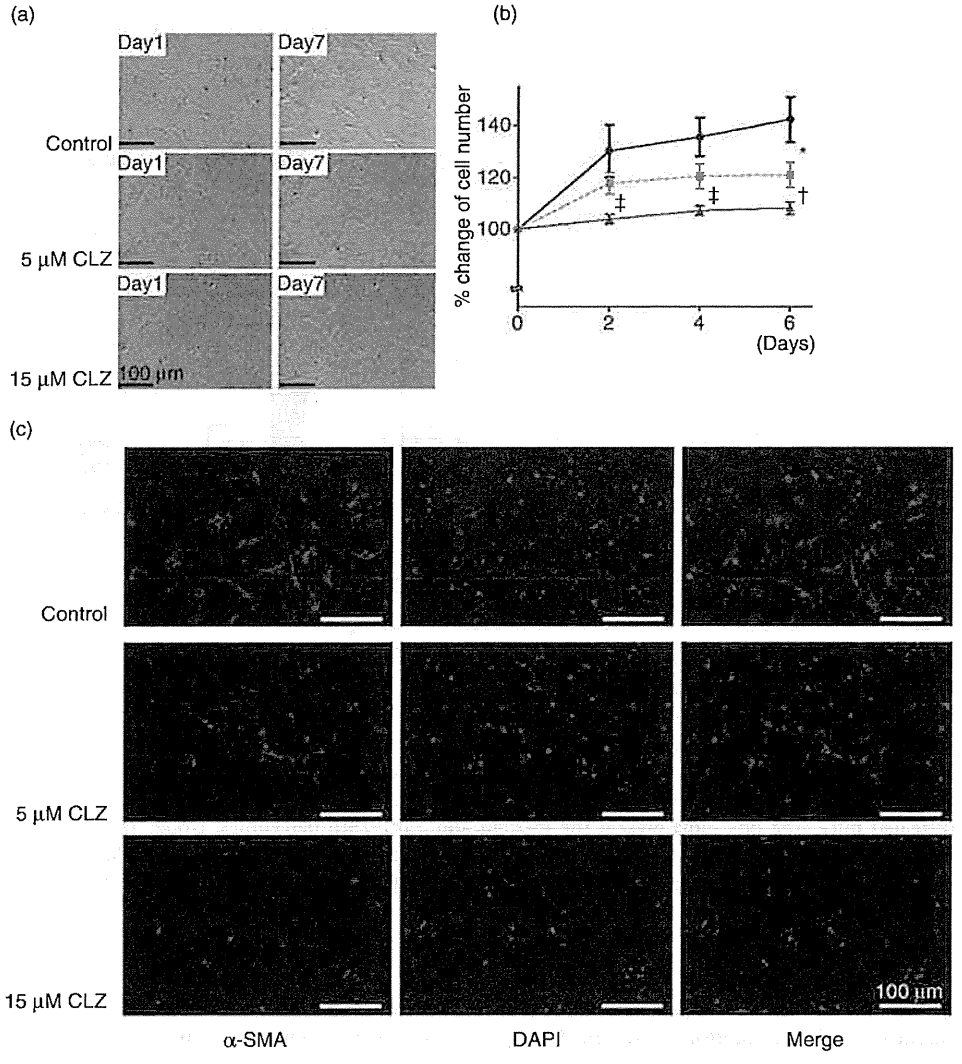

Figure 5 Cilostazol suppressed the proliferation and activation of HSC, but did not affect Kupffer-cell activation. (a) Morphological changes in HSC from 0–6 days were viewed on a phase contrast microscope (original magnification ×200). HSC supplemented with 15 μM cilostazol resulted in visible short cytoplasmic dendritic processes and perinuclear vacuoles containing retinoids. (b) HSC proliferation was determined by direct count of the cell numbers. Cilostazol supplementation slowed the increase in cell numbers compared with control. (c) Immunofluorescent staining of α-SMA (red) in HSC on the second day of culture (original magnification ×200). The protein expression of α-SMA was decreased in cilostazol-supplemented HSC in a dose-dependent manner. (d) F4/80 immunostaining of liver sections in each group. (original magnification ×100). (e) Quantification of the F4/80-positive area in each group. Cilostazol-administrated groups tended to show fewer F4/80-positive areas than the control, but no significant differences were observed among CCl_4-treated groups. (f) mRNA expression levels of TNF-α and TGF-β1 in the liver were not affected by cilostazol. (g) Expression of Kupffer cell marker (F4/80) and inflammation-related genes (TNF-α, IL-1β, MCP1 and TGF-β1) in primary Kupffer cells on the second day of culture was not altered by cilostazol. (h) Accumulation of cAMP in primary cultured HSC and Kupffer cells. Cilostazol supplementation significantly elevated the cAMP level only in HSC. The box plots present the median and 25th–75th percentiles. Upper and lower lines represent the minimum and maximum values ($n = 4$). *$P < 0.05$; †$P < 0.001$; ‡$P < 0.01$ vs control group. α-SMA, α-smooth muscle actin; cAMP, cyclic adenosine monophosphate; CCl_4, carbon tetrachloride; CLZ, cilostazol; DAPI, 4′,6′-diamidino-2-phenylindole dihydrochloride; HSC, hepatic stellate cell; IL-1β, interleukin-1β; MCP1, monocyte chemotactic protein-1; TNF-α, tumor necrosis factor-α; TGF-β1, transforming growth factor-β1. -●-, control; -■-, 5 μM CLZ; -▲-, 15 μM CLZ.

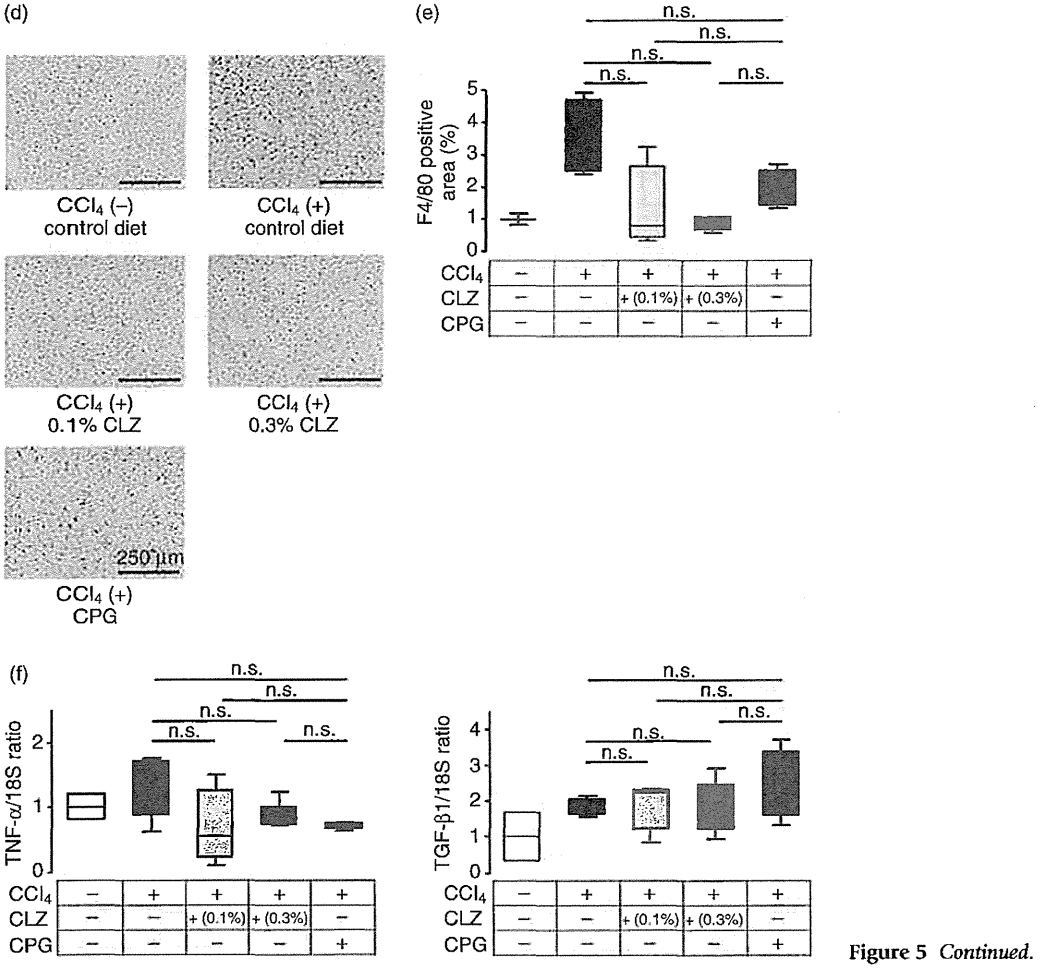

Figure 5 *Continued.*

inhibition of the PDE3 enzyme, was significantly higher only in HSC supplemented with cilostazol (2.283; 95% CI = 1.45–3.12; $P < 0.01$) but not in Kupffer cells (1.363; 95% CI = 0.4374–2.289; Fig. 5 h).

Cilostazol suppressed PDGFR expression in HSC

To further delineate the effect of cilostazol on the activation of HSC, we characterized the cilostazol-affected gene expression profiles during the activation phase of HSC. First, to confirm the direct effects of cilostazol on the gene activation mechanism of HSC, we examined the α-SMA and collagen α1 (I) gene induction. As suggested by the previous data (Figs 2,4,5), mRNA induction of α-SMA was lower in cells supplemented with 5 μM cilostazol (0.555; 95% CI = 0.085–1.024) and 15 μM cilostazol (0.221; 95% CI = 0.086–0.356) when compared with the control (2.53; 95% CI = 1.01–4.05; $P < 0.01$; Fig. 6a). Similarly, collagen α1 (I) mRNA expression was lower in cells supplemented with 5 μM cilostazol (0.411; 95% CI = 0.010–0.833) and 15 μM cilostazol (0.059; 95% CI = 0.042–0.159) as compared with the control cells (2.20; 95% CI = 0.31–4.08; $P < 0.01$; Fig. 6b). Then, to gain further mechanistic insight into the action of cilostazol on HSC, the mRNA expression of PDGF-B, PDGFR-β and TGF-βR1, an important cytokine and cytokine receptors for HSC activation, was determined. The expression of PDGF-B, one of the most important mitogens for HSC, was unaffected by cilostazol treatment (Fig. 6c), however, PDGFR-β mRNA expression in the 5 μM cilostazol-supplemented cells (0.282; 95% CI = 0.104–0.460) and

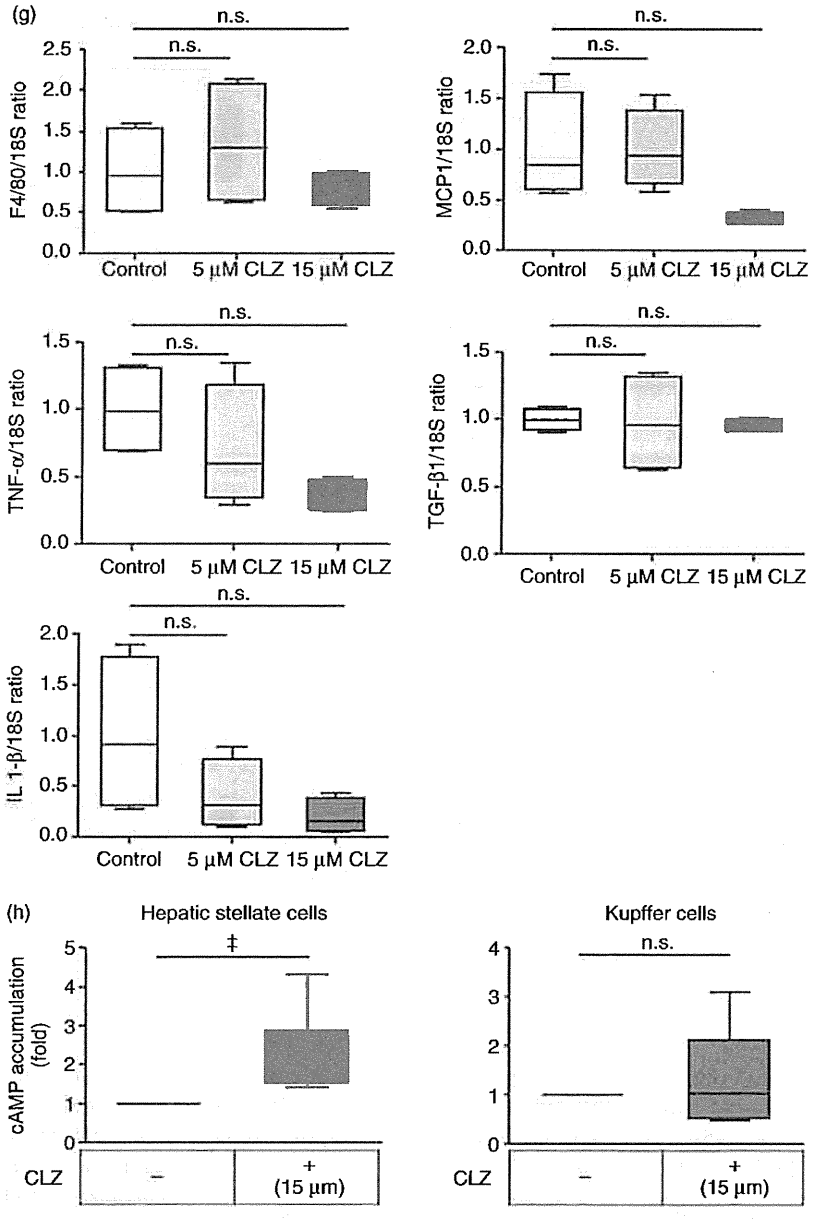

Figure 5 *Continued.*

15 μM cilostazol-supplemented cells (0.336; 95% CI = 0.036–0.636) was significantly decreased compared with that in control cells (0.749; 95% CI = 0.290–1.210; $P < 0.001$; Fig. 6d). In contrast, TGF-βR1 mRNA expression was not affected by cilostazol treatment (Fig. 6e). These results indicate the possibility that cilostazol attenuates the activation-induced proliferation of HSC through the abrogation of PDGF-autocrine signaling by limiting the receptor (PDGFR-β) signaling regardless of the ligand (PDGF) availability.

DISCUSSION

THE P.O. ADMINISTRATION of cilostazol effectively prevents the development of CCl_4-induced liver fibrosis in mice. In agreement with the previous study,

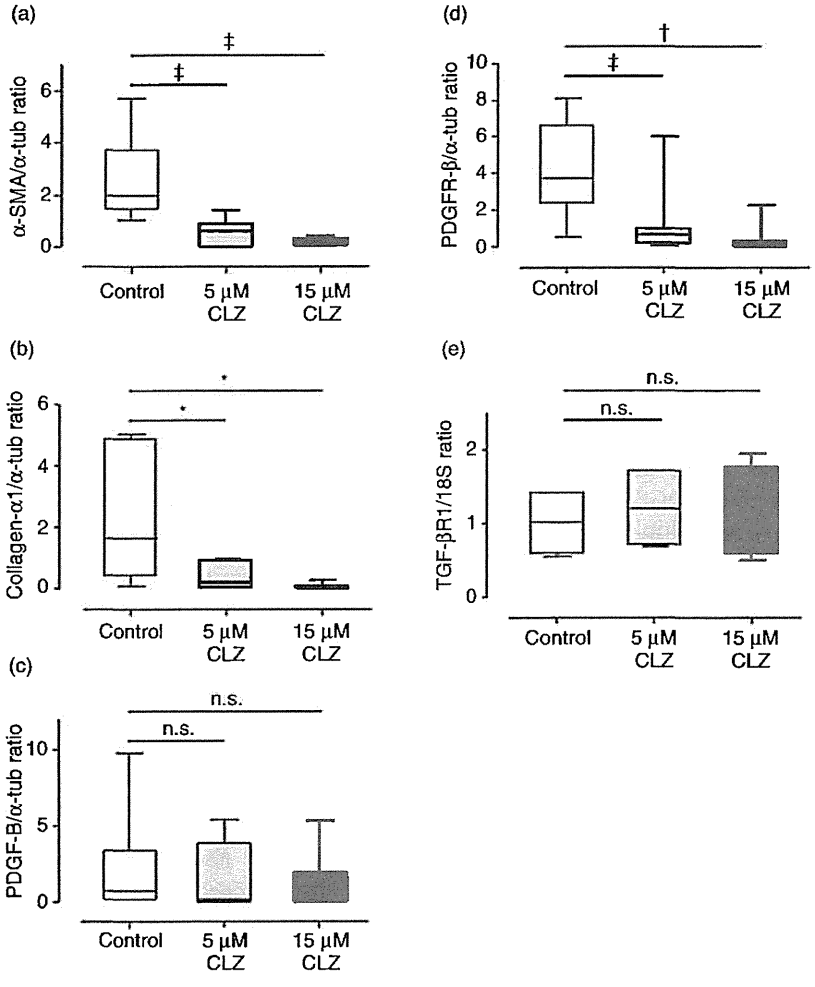

Figure 6 Expression of genes associated with HSC activation was assessed by reverse transcription quantitative polymerase chain reaction. (a) α-SMA, (b) collagen α1 (I) and (d) PDGFR-β expression levels in HSC supplemented with 5 μM and 15 μM cilostazol for 2 days were significantly suppressed compared with control (*$P < 0.05$; ‡$P < 0.01$; †$P < 0.001$ vs control, respectively). The expression of (c) PDGF-B and (e) TGF-βR1 had no difference between control and cilostazol-supplemented HSC. The box plots present the median and 25th–75th percentiles. Upper and lower lines represent the minimum and maximum values ($n = 7$). α-SMA, α-smooth muscle actin; CLZ, cilostazol; HSC, hepatic stellate cell; PDGF-B, platelet growth factor-B; PDGFR-β, platelet growth factor receptor-β; TGF-βR1, transforming growth factor-β receptor 1.

cilostazol was not toxic to HSC as indicated by the morphology and proliferation of the cells (Fig. 5a,b).[25] It is noteworthy that unlike many other candidate medications, cilostazol is already widely used as an antiplatelet agent in clinical practice with proven long-term safety. For this reason, cilostazol holds potential to become an antifibrotic agent for chronic liver diseases in humans.

In the present study, we employed clopidogrel as an alternative antiplatelet agent for the comparison. Although both cilostazol and clopidogrel showed minimal side-effects (Fig. 3), only cilostazol attenuated liver fibrosis (Figs 2,4), suggesting that cilostazol may have distinct antifibrotic mechanisms apart from its antiplatelet action. This notion is consistent with the results of the present *in vitro* study in which the treatment of primary HSC with cilostazol attenuated the HSC proliferation (Fig. 5b) and the expression of α-SMA and collagen α1 (I) (Figs 5c,6), indicating the direct effect of cilostazol on HSC. On the other hand, our *in vitro* and *in vivo* studies showed only a minimal effect of cilostazol on Kupffer cells (Fig. 5d,e,g,h), and no significant change was detected in inflammatory and fibrogenic genes (such as TNF-α and TGF-β1) by cilostazol administration (Fig. 5f). These results lend support to the concept that cilostazol exerts its antifibrotic effect(s) via the suppression of HSC activation *in vivo*.

As reported,[6,39,41] PDGFR-β was absent in quiescent HSC, but was upregulated in an early stage of liver injury. Activating factors from autocrine or paracrine sources such as TGF-β1 stimulate the transcriptional induction of PDGFR-β in quiescent HSC, thereby

rendering them responsive to PDGF-B chain molecules. Among several activating pathways, the autocrine loop exerted by PDGF–PDGFR signaling is regarded as one of the most potent mitogenic pathways for HSC.[42] Although PDGF itself seemed unaffected in our study, our quantitative analyses showed that cilostazol significantly suppressed PDGFR-β in HSC (Fig. 6c,d). Because the PDGF–PDGFR signaling not only promotes myofibroblast proliferation but also participates in other fibrogenic actions, including stimulation of collagen production and promotion of cell adhesion, it has been speculated that the activated PDGF–PDGFR signaling pathway is a candidate target for antifibrotic therapy in liver diseases.[43] Indeed, focusing on PDGFR, recent studies have shown attenuation of hepatic fibrosis by a PDGFR tyrosine kinase inhibitor.[44–46] In the same sense, a blockade of the autocrine loop of PDGF–PDGFR signaling by cilostazol may also have multiple benefits for preventing the development of hepatic fibrosis.

Cilostazol is a selective inhibitor of PDE3, and PDE3 inhibition in platelets exhibits antithrombotic effects by preventing platelet aggregation. Recently, increased intracellular cAMP levels and activation of protein kinase A (PKA) were reported to reduce PDGF-stimulated cellular proliferation.[47,48] Interestingly, cilostazol has been shown to be effective against the development of non-alcoholic fatty liver disease through the activation of the cAMP/PKA signaling pathway *in vivo*.[26] Although the exact mechanism remains to be determined, there may be a link between PDGFR downregulation and cAMP/PKA signaling in HSC.

In conclusion, orally bioavailable cilostazol attenuates HSC activation, possibly through the suppression of PDGFR expression in HSC, and thereby alleviates hepatic fibrogenesis. Further studies may yield a future intervention strategy against liver diseases.

ACKNOWLEDGMENTS

WE ARE GRATEFUL for technical assistance provided by the Central Laboratory of Osaka City University Medical School. We also thank Kimie Takafuji (Hamanaka), Eri Hikawa, Tadashi Mizutani, Kenta Aohara, Kazushi Yamasato and Marika Horiuchi. This work was supported in part by a Grant-in-Aid for Scientific Research (Grant Numbers 20372435 [M. A.], 21689035 [K. H.], 24689056 [M. A.]) from the Japan Society for the Promotion of Science; a grant from the Japanese Ministry of Health, Labour and Welfare (K. Iwaisako., K. Ikeda, M. A and S. U.); and grants from Otsuka Pharma (K. H and S. U.), the Uehara Memorial Foundation (M. A.) and Kanae Foundation for the Promotion of Medical Science (M. A.).

REFERENCES

1 Friedman SL. Mechanisms of hepatic fibrogenesis. *Gastroenterology* 2008; 134: 1655–69.
2 Gabele E, Brenner DA, Rippe RA. Liver fibrosis: signals leading to the amplification of the fibrogenic hepatic stellate cell. *Front Biosci* 2003; 8: d69–77.
3 Afdhal NH, Nunes D. Evaluation of liver fibrosis: a concise review. *Am J Gastroenterol* 2004; 99: 1160–74.
4 Tsukada S, Parsons CJ, Rippe RA. Mechanisms of liver fibrosis. *Clin Chim Acta* 2006; 364: 33–60.
5 Inagaki Y, Higashiyama R, Higashi K. Novel anti-fibrotic modalities for liver fibrosis: molecular targeting and regenerative medicine in fibrosis therapy. *J Gastroenterol Hepatol* 2012; 27 (Suppl 2): 85–8.
6 Bonner JC. Regulation of PDGF and its receptors in fibrotic diseases. *Cytokine Growth Factor Rev* 2004; 15: 255–73.
7 Tsukamoto H. Cytokine regulation of hepatic stellate cells in liver fibrosis. *Alcohol Clin Exp Res* 1999; 23: 911–6.
8 Novobrantseva TI, Majeau GR, Amatucci A *et al.* Attenuated liver fibrosis in the absence of B cells. *J Clin Invest* 2005; 115: 3072–82.
9 Bataller R, Brenner DA. Hepatic stellate cells as a target for the treatment of liver fibrosis. *Semin Liver Dis* 2001; 21: 437–51.
10 Friedman SL. Liver fibrosis – from bench to bedside. *J Hepatol* 2003; 38 (Suppl 1): S38–53.
11 Zhang D, Utsumi T, Huang HC *et al.* Reticulon 4B (Nogo-B) is a novel regulator of hepatic fibrosis. *Hepatology* 2011; 53: 1306–15.
12 Kambayashi J, Liu YG, Sun B, Shakur Y, Yoshitake M, Czerwiec F. Cilostazol as a unique antithrombotic agent. *Curr Pharm Des* 2003; 9: 2289–302.
13 Okuda Y, Mizutani M, Ikegami T, Ueno E, Yamashita K. Hemodynamic effects of cilostazol on peripheral artery in patients with diabetic neuropathy. *Arzneimittelforschung* 1992; 42: 540–2.
14 Wang S, Cone J, Fong M, Yoshitake M, Kambayashi J, Liu YG. Interplay between inhibition of adenosine uptake and phosphodiesterase type 3 on cardiac function by cilostazol, an agent to treat intermittent claudication. *J Cardiovasc Pharmacol* 2001; 38: 775–83.
15 Kimura Y, Tani T, Kanbe T, Watanabe K. Effect of cilostazol on platelet aggregation and experimental thrombosis. *Arzneimittelforschung* 1985; 35: 1144–9.
16 Kohriyama T, Tanaka E, Katayama S, Yamamura Y, Nakamura S. [Antiplatelet therapy in patients with cerebral thrombosis at the chronic phase – assessment of its effect on coagulation and fibrinolytic parameters]. *Rinsho Shinkeigaku* 1994; 34: 771–6.

17 Liu Y, Cone J, Le SN *et al.* Cilostazol and dipyridamole synergistically inhibit human platelet aggregation. *J Cardiovasc Pharmacol* 2004; 44: 266–73.
18 Liu Y, Shakur Y, Yoshitake M, Kambayashi Ji J. Cilostazol (pletal): a dual inhibitor of cyclic nucleotide phosphodiesterase type 3 and adenosine uptake. *Cardiovasc Drug Rev* 2001; 19: 369–86.
19 Mizutani M, Okuda Y, Yamashita K. Effect of cilostazol on the production of platelet-derived growth factor in cultured human vascular endothelial cells. *Biochem Mol Med* 1996; 57: 156–8.
20 Bender AT, Beavo JA. Cyclic nucleotide phosphodiesterases: molecular regulation to clinical use. *Pharmacol Rev* 2006; 58: 488–520.
21 Hayashi S, Morishita R, Matsushita H *et al.* Cyclic AMP inhibited proliferation of human aortic vascular smooth muscle cells, accompanied by induction of p53 and p21. *Hypertension* 2000; 35: 237–43.
22 Kawada N, Kuroki T, Kobayashi K, Inoue M, Kaneda K. Inhibition of myofibroblastic transformation of cultured rat hepatic stellate cells by methylxanthines and dibutyryl cAMP. *Dig Dis Sci* 1996; 41: 1022–9.
23 Mallat A, Preaux AM, Serradeil-Le Gal C *et al.* Growth inhibitory properties of endothelin-1 in activated human hepatic stellate cells: a cyclic adenosine monophosphate-mediated pathway. Inhibition of both extracellular signal-regulated kinase and c-Jun kinase and upregulation of endothelin B receptors. *J Clin Invest* 1996; 98: 2771–8.
24 Houglum K, Lee KS, Chojkier M. Proliferation of hepatic stellate cells is inhibited by phosphorylation of CREB on serine 133. *J Clin Invest* 1997; 99: 1322–8.
25 Shimizu E, Kobayashi Y, Oki Y, Kawasaki T, Yoshimi T, Nakamura H. OPC-13013, a cyclic nucleotide phosphodiesterase type III, inhibitor, inhibits cell proliferation and transdifferentiation of cultured rat hepatic stellate cells. *Life Sci* 1999; 64: 2081–8.
26 Fujita K, Nozaki Y, Wada K *et al.* Effectiveness of antiplatelet drugs against experimental non-alcoholic fatty liver disease. *Gut* 2008; 57: 1583–91.
27 Hase Y, Okamoto Y, Fujita Y *et al.* Cilostazol, a phosphodiesterase inhibitor, prevents no-reflow and hemorrhage in mice with focal cerebral ischemia. *Exp Neurol* 2012; 233: 523–33.
28 Hasegawa SI, Inomata J, Tuji Y, Saitoh E. Assessment of buccal absorption of cilostazol 100-mg orally disintegrated tablets in healthy adult male subjects. *Jpn Pharmacol Ther* 2009; 37: 813–9.
29 Hasegawa SI, Inomata J, Tuji Y, Saitoh E. Bioequivalence study of cilostazol 100-mg orally disintegrated tablets and cilostazol 100-mg conventional tablets in healthy adult male subjects. *Jpn Pharmacol Ther* 2009; 37: 805–12.
30 Isayama F, Hines IN, Kremer M *et al.* LPS signaling enhances hepatic fibrogenesis caused by experimental cholestasis in mice. *Am J Physiol Gastrointest Liver Physiol* 2006; 290: G1318–28.
31 Toda Y, Kono K, Abiru H *et al.* Application of tyramide signal amplification system to immunohistochemistry: a potent method to localize antigens that are not detectable by ordinary method. *Pathol Int* 1999; 49: 479–83.
32 Seki E, De Minicis S, Osterreicher CH *et al.* TLR4 enhances TGF-beta signaling and hepatic fibrosis. *Nat Med* 2007; 13: 1324–32.
33 Schnabl B, Kweon YO, Frederick JP, Wang XF, Rippe RA, Brenner DA. The role of Smad3 in mediating mouse hepatic stellate cell activation. *Hepatology* 2001; 34: 89–100.
34 Siegmund SV, Uchinami H, Osawa Y, Brenner DA, Schwabe RF. Anandamide induces necrosis in primary hepatic stellate cells. *Hepatology* 2005; 41: 1085–95.
35 Weiskirchen R, Gressner AM. Isolation and culture of hepatic stellate cells. *Methods Mol Med* 2005; 117: 99–113.
36 Friedman SL, Roll FJ. Isolation and culture of hepatic lipocytes, Kupffer cells, and sinusoidal endothelial cells by density gradient centrifugation with Stractan. *Anal Biochem* 1987; 161: 207–18.
37 Mou L, Xing Y, Kong Z, Zhou Y, Chen Z, Wang R. The N-terminal domain of human hemokinin-1 influences functional selectivity property for tachykinin receptor neurokinin-1. *Biochem Pharmacol* 2011; 81: 661–8.
38 Tomita K, Tamiya G, Ando S *et al.* Tumour necrosis factor alpha signalling through activation of Kupffer cells plays an essential role in liver fibrosis of non-alcoholic steatohepatitis in mice. *Gut* 2006; 55: 415–24.
39 Friedman SL, Arthur MJ. Activation of cultured rat hepatic lipocytes by Kupffer cell conditioned medium. Direct enhancement of matrix synthesis and stimulation of cell proliferation via induction of platelet-derived growth factor receptors. *J Clin Invest* 1989; 84: 1780–5.
40 Tsukamoto H, Gaal K, French SW. Insights into the pathogenesis of alcoholic liver necrosis and fibrosis: status report. *Hepatology* 1990; 12: 599–608.
41 Wong L, Yamasaki G, Johnson RJ, Friedman SL. Induction of beta-platelet-derived growth factor receptor in rat hepatic lipocytes during cellular activation in vivo and in culture. *J Clin Invest* 1994; 94: 1563–9.
42 Borkham-Kamphorst E, van Roeyen CR, Ostendorf T, Floege J, Gressner AM, Weiskirchen R. Pro-fibrogenic potential of PDGF-D in liver fibrosis. *J Hepatol* 2007; 46: 1064–74.
43 Melton AC, Yee HF. Hepatic stellate cell protrusions couple platelet-derived growth factor-BB to chemotaxis. *Hepatology* 2007; 45: 1446–53.
44 Yasuda Y, Shimizu M, Sakai H *et al.* (-)-Epigallocatechin gallate prevents carbon tetrachloride-induced rat hepatic fibrosis by inhibiting the expression of the PDGFRbeta and IGF-1R. *Chem Biol Interact* 2009; 182: 159–64.
45 Gonzalo T, Beljaars L, van de Bovenkamp M *et al.* Local inhibition of liver fibrosis by specific delivery of a platelet-derived growth factor kinase inhibitor to hepatic stellate cells. *J Pharmacol Exp Ther* 2007; 321: 856–65.

46 Yoshiji H, Noguchi R, Kuriyama S *et al.* Imatinib mesylate (STI-571) attenuates liver fibrosis development in rats. *Am J Physiol Gastrointest Liver Physiol* 2005; 288: G907–13.
47 Graves LM, Bornfeldt KE, Raines EW *et al.* Protein kinase A antagonizes platelet-derived growth factor-induced signaling by mitogen-activated protein kinase in human arterial smooth muscle cells. *Proc Natl Acad Sci U S A* 1993; 90: 10300–4.
48 Mallat A, Gallois C, Tao J *et al.* Platelet-derived growth factor-BB and thrombin generate positive and negative signals for human hepatic stellate cell proliferation. Role of a prostaglandin/cyclic AMP pathway and cross-talk with endothelin receptors. *J Biol Chem* 1998; 273: 27300–5.

SUPPORTING INFORMATION

ADDITIONAL SUPPORTING INFORMATION may be found in the online version of this article at the publisher's web-site:

Video Clip S1–S3 Time-lapse motion pictures of cultured hepatic stellate cells (HSC) with or without cilostazol treatment.

PLOS ONE

Principal Component Analysis Based Feature Extraction Approach to Identify Circulating microRNA Biomarkers

Y-h. Taguchi[1]*, Yoshiki Murakami[2]

1 Department of Physics, Chuo University, Tokyo, Japan, 2 Department of Hepatology, Graduate School of Medicine, Osaka City University, Osaka, Japan

Abstract

The discovery and characterization of blood-based disease biomarkers are clinically important because blood collection is easy and involves relatively little stress for the patient. However, blood generally reflects not only targeted diseases, but also the whole body status of patients. Thus, the selection of biomarkers may be difficult. In this study, we considered miRNAs as biomarker candidates for several reasons. First, since miRNAs were discovered relatively recently, they have not yet been tested extensively. Second, since the number of miRNAs is relatively limited, selection is expected to be easy. Third, since they are known to play critical roles in a wide range of biological processes, their expression may be disease specific. We applied a newly proposed method to select combinations of miRNAs that discriminate between healthy controls and each of 14 diseases that include 5 cancers. A new feature selection method is based on principal component analysis. Namely this method does not require knowledge of whether each sample was derived from a disease patient or a healthy control. Using this method, we found that hsa-miR-425, hsa-miR-15b, hsa-miR-185, hsa-miR-92a, hsa-miR-140-3p, hsa-miR-320a, hsa-miR-486-5p, hsa-miR-16, hsa-miR-191, hsa-miR-106b, hsa-miR-19b, and hsa-miR-30d were potential biomarkers; combinations of 10 of these miRNAs allowed us to discriminate each disease included in this study from healthy controls. These 12 miRNAs are significantly up- or downregulated in most cancers and other diseases, albeit in a cancer- or disease-specific combinatory manner. Therefore, these 12 miRNAs were also previously reported to be cancer- and disease-related miRNAs. Many disease-specific KEGG pathways were also significantly enriched by target genes of up−/downregulated miRNAs within several combinations of 10 miRNAs among these 12 miRNAs. We also selected miRNAs that could discriminate one disease from another or from healthy controls. These miRNAs were found to be largely overlapped with miRNAs that discriminate each disease from healthy controls.

Citation: Taguchi Y-h, Murakami Y (2013) Principal Component Analysis Based Feature Extraction Approach to Identify Circulating microRNA Biomarkers. PLoS ONE 8(6): e66714. doi:10.1371/journal.pone.0066714

Editor: Rossella Rota, Ospedale Pediatrico Bambino Gesù, Italy

Received September 7, 2012; **Accepted** May 9, 2013; **Published** June 24, 2013

Funding: This study was supported by KAKENHI (23300357) (http://www.jsps.go.jp/english/e-grants/index.html). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: tag@granular.com

Introduction

Specific and sensitive noninvasive biomarkers for the detection of human diseases, including malignancies, are urgently required to reduce worldwide morbidity and mortality caused by cancer [1–3]. Although some successful use of transcriptome components as biomarkers was reported [4,5], circulating microRNAs have also recently been identified as new clinical biomarker candidates [6–12]. MicroRNAs are post-transcriptional regulators that are involved in many physiological and pathophysiological conditions. A recent study by Keller *et al.* [13] compared the expression profiles of hundreds of blood-borne microRNAs across a variety of nonmalignant and malignant diseases to identify disease-specific expression patterns. The resulting microRNA expression data could be used to discriminate disease samples with a high level of accuracy, demonstrating the potential use of microRNA signatures for blood-based diagnosis of disease. Using extensive bioinformatics research, Keller *et al.* demonstrated that a wide range of cancers and other diseases could be discriminated from healthy controls by only miRNA expression. The data set Keller *et al.* used was the most extensive data set ever reported, i.e., it included various types of diseases (14 diseases plus normal controls) and a large number of patient (n = 384) and control (n = 70) data sets from a large number of blood-based miRNA biomarker studies.

In spite of this, selecting a biomarker based on feature-extraction techniques remains challenging. Although Keller *et al.* [13] successfully discriminated cancers and other diseases from healthy controls by using the expression of only 10 miRNAs (see Supplementary Table 6 on page 14 of their Supplementary Materials), they did not state which 10 miRNAs were selected because of the problem of stability, raised by Abeel *et al.* [14]. Stability is the measure of how stable feature selections are. For example, suppose we have 2 set of samples, each of which consists of 2 categories. When features are independently extracted so as to discriminate 2 categories for each sample, if the majority of selected features are common between the 2 samples, it can be considered a stable feature extraction. If not, it is unstable. That is, if selected features fluctuate depending on the sample, the stability of feature selection is poor. Conversely, if most features are selected independent of the sample, the stability of feature selection is high.

Although Keller *et al.* [13] employed 10-fold cross-validation, the selection of 10 miRNAs fluctuated between trials (see demonstration in the "Stability" subsection below). This prevented them from presenting 10 specific miRNAs as biomarkers to

discriminate patients with cancers and other diseases from healthy controls. This is a significant disadvantage of their research if it is to be applied for clinical use, since it is impossible to decide in advance which miRNAs should be employed as biomarkers.

In order to overcome this problem, we propose a new feature selection technique to select miRNAs as biomarkers. This method is based on principal component analysis (PCA), more specifically, sparse PCA [15–19].

PCA [20] is a type of dimensional reduction or ordination analysis. Ordination analysis attempts to embed objects distributed in high dimensional space into lower dimensional space. In PCA, dimensional reduction is achieved by projection to lower dimensional space using linear transformation. Although PCA is a simple and classical method, it can often effectively reduce redundant information.

Sparse PCA is defined as follows. In contrast to ordinary PCA, which employs all features to express lower dimensional space, sparse PCA tries to express lower dimensional space by a smaller number of features, even if the accuracy decreases. That is, sparse PCA is a feature extraction method that eliminates unnecessary features through a method that is not uniquely defined, but varies depending on the implementation.

Some similar trials of this kind using clustering-based feature extraction have been reported [21,22]. For example, Liu *et al.* [23,24] proposed gene selection using spectral biclustering, Dy *et al* [25] used hierarchical clustering for feature selection of lung cancer image classification, and Modha *et al* [26] made use of k-means for feature extraction. However, these are feature selection methods that require prior knowledge of class partitions or labeling. At minimum, prior to feature selection, the previous methods require knowledge or inference of the number of clusters, which our current approach does not require. Moreover, there has been no discussion of the stability of feature selection when using these cluster-based feature selection criteria. Such stability problems have started to be discussed only very recently [14].

In contrast to both of these above-mentioned general feature extraction approaches and several previously proposed feature extraction methods especially designed for gene expression analysis, e.g., significance analysis of microarrays (SAM) [27], gene selection based on a mixture of marginal distributions (gsMMD) [28], and ensemble recursive feature elimination (RFE) [14], our approach is free from stability problems. These types of classification-independent and stability-problem-free approaches were invented only very recently (e.g., unsupervised feature filtering (UFF) [29]) and are still very rare.

For each pair of diseases and normal controls, our approach enabled the selection of a set of 10 strict (confident) miRNA biomarker candidates. Each set of miRNAs could not only accurately discriminate patients with each disease from normal controls, but could also accurately discriminate one disease from another. Moreover, most of the sets shared the majority of miRNAs, which would allow for simplification of the measurement of a biomarker since a limited number of sets of miRNA measurements would permit the discrimination of several diseases.

The reason why we tried to employ blood-based biomarkers in spite of the drawbacks that miRNAs in blood inevitably reflect the whole body status and thus have less of a relationship with targeted diseases is because previous studies have demonstrated that the use of several circulating miRNAs can work well in a practical sense. Thus unification and consideration of our analysis results will improve the ability of miRNAs as biomarkers.

Materials and Methods

Feature Extraction Methods

PCA-based feature extraction. Suppose we have the miRNA profiles $x_{ij},(i=1,\ldots,N,j=1,\ldots,M)$, each corresponding to the ith miRNA in the jth sample. N and M are the total number of miRNAs and samples, respectively. Samples were classified into L clinical sets, $G_l,(l=1,\ldots,L)$. We applied PCA to the $\{x_{ij}\}$ set in 2 ways:

1. Method 1 (miRNA-based): Substitute $K_s(<M)$ principal component (PC) score $\{x_{ik},k=1,\ldots,K_s\}$ to $\{x_{ij}j=1,\ldots,M\}$. In this case, PCA was applied to a matrix $\{x_{ij}\}$.
2. Method 2 (sample-based): Substitute $K_m(<N)$ PC score $\{x_{kj},k=1,\ldots,K_m\}$ to $\{x_{ji},i=1,\ldots,N\}$. In this case, PCA was applied to a transverse matrix $\{x_{ji}\}$.

PCA-based feature extraction was performed as follows.

1. Choose a pair of clinical sets, l and l'.
2. Compute x_{ik} with Method 1 PCA from $\{x_{ij}|j\in G_l\cup G_{l'}\}$.
3. Compute distance r_i,

$$r_i\equiv\sqrt{\sum_{k=1}^{K_s^0}x_{ik}^2},$$

where $K_s^0(<K_s)$ is the number of components to be used for feature selection.

4. Select miRNAs i' with top $N_1(<N)$ r_is.

N_1 miRNAs are a set of selected features to distinguish clinical sets l and l'. Throughout this paper, K_s^0 was assumed to be 2 unless stated otherwise. PCA was computed by the prcomp function in the R base package [30].

It should be noted that our method did not use any classification information. This enabled us to obtain stable feature extractions.

t-test based feature extraction. The P-value for the ith miRNA between $\{x_{ij}|j\in G_1\}$ and $\{x_{ij}|j\in G_2\}$ was computed using a t test. The top N_1 miRNAs with smaller P-values were selected.

SAM-based feature extraction. The P-value for the ith miRNA between $\{x_{ij}|j\in G_1\}$ and $\{x_{ij}|j\in G_2\}$ was computed using SAM. The top N_1 miRNAs with smaller P-values were selected.

gsMMD-based feature extraction. Significantly up- or downregulated miRNAs were selected by gsMMD, which was implemented in Bioconductor software. The P-values for up- or downregulation were considered separately, and the top N_1 miRNAs were selected for both up- and downregulated miRNAs.

RFE- and ensemble RFE-based feature extraction. As described by Abeel *et al.* [14], the support vector machine with a linear kernel was applied to 40 independent resampled sets (with replacements) for ensemble RFE. After 100 independent cross-validations with 10% test samples and 90% training samples, the top N_1 miRNAs with better accuracy were extracted. For simplicity, we employed only complete linear aggregation for weighting among each resampled set. No resamplings were conducted for simple RFE, and only cross-validations were performed. The top N_1 miRNAs with better accuracy were selected.

UFF. As described by Varshavsky *et al.* [29], differential singular value decomposition (SVD)-entropy ΔH_i,

$$\Delta H_i \equiv H - H_i$$

$$H \equiv -\frac{1}{\log(N)} \sum_{i'=1}^{N} \rho_{i'} \log(\rho_{i'})$$

$$H_i \equiv -\frac{1}{\log(N)} \sum_{i' \neq i} \rho_{i'} \log(\rho_{i'})$$

$$\rho_i \equiv \frac{s_i^2}{\sum_{i'=1}^{N} s_{i'}^2}$$

where s_i is a singular value and N is the total number of miRNAs, was attributed to each miRNA. After 100 independent cross-validations with 10% test samples and 90% training samples, the top N_1 miRNAs with larger ΔH_i were selected.

PCA-based Linear Discriminant Analysis (LDA)

PCA-based LDA was conducted as follows:

1. Choose a pair of clinical sets, l and l'.
2. If necessary, apply feature extraction and reduce the number of miRNAs used for LDA.
3. Compute x_{kj},$(k=1,\ldots,K_m)$ using Method 2 PCA.
4. Divide samples into training and test sets.
5. Apply LDA to training set.
6. Validate performance of LDA using test set.
7. Repeat steps 4–6 the specified number of times depending on the employed cross-validation method.
8. Compute performance with averaged values.
9. Estimate the optimal value of K_m by repeating steps 3–8 as K_m changes.

It should be noted that the division between training and test sets was carried out **AFTER** the computation of PCA (and feature extraction if necessary). Thus, x_{kj} includes the test set information as well. Feature extraction, if applied, was also conducted before division was performed; thus, it was sampling free. One may suggest that this was erroneous since we do not know the classification of the test set. However, we can compute the PCA even if we do not have prior knowledge of the classification because we do not need classification information to compute x_{kj}. This is explained in the "Why did PCA-based feature selection work so well?" subsection in Results and Discussion section as well. The LDA was computed by the lda function in the R base package [30].

miRNA Expression and Normalization

The miRNA expression used in this study was obtained from the Gene Expression Omnibus (GEO) accession number GSE31568, which was used by Keller *et al.* [13]. We downloaded GSE31568_raw and normalized miRNA expression within each sample to obtain the mean and standard deviation (SD).

Stability Test

As in Abeel *et al.* [14] and Varshavsky *et al* [29], we evaluated whether the selection of miRNAs for the discrimination between patients with diseases and healthy controls was stable [13]. The procedure was as follows:

1. Choose a pair of clinical sets, each including one cancer or other disease sample and one healthy control sample.
2. Pick 90% samples randomly, independent of classification.
3. Apply feature selection to select 10 miRNAs as biomarkers.
4. Repeat steps 2 and 3 a total of 100 times and count the frequency of each miRNA selection.
5. Repeat steps 2–4 for all pairs of diseases and healthy controls.

The above procedures were applied to all feature extraction methods, i.e., those based on t-tests, PCA, SAM, gsMMD, RFE, RFE ensemble, and UFF.

Amount of Contribution from each miRNA to Discrimination

Suppose we obtained x_{kj} by PCA analysis after PCA-based feature extraction was applied. Then

$$x_{kj} = \sum_{i=1}^{Km} a_{ik} x_{ij}$$

If we applied LDA to discriminate one cancer or disease from one healthy control using x_{kj}, we obtained the discriminant function LD_j as

$$LD_j = \sum_{k=1}^{PC} b_k x_{kj} = \sum_{k=1}^{PC} b_k \sum_{i=1}^{Km} a_{ik} x_{ij} = \sum_{i=1}^{Km} \left(\sum_{k=1}^{PC} b_k a_{ik} \right) x_{ij},$$

for the jth sample, where PC is the number of PCs used for discrimination. Typically, a positive (negative) LD_j indicates that the sample j represents a patient with cancer or another disease (healthy control) sample. Then, the amount of contribution, C_i, of miRNA i to the discriminant function is

$$C_i = \sum_{k=1}^{PC} b_k a_{ik}.$$

KEGG Pathway Analysis of miRNA Target Genes

DIANA-mirPath. DIANA-mirPath [31] is a web tool. We used this software version for the implementation of multiple miRNAs (http://diana.cslab.ece.ntua.gr/pathways/index_multiple.php). DIANA-mirPath accepts a set of miRNAs, estimates union of miRNA target genes, and finally computes P-values that describe KEGG pathway enrichment of the target genes. For the data set of our selection, we extracted up- or downregulated sets of miRNAs and uploaded them onto DIANA-mirPath. Using default settings, DIANA-mirPath employed a list of target genes estimated by DIANA-microT v4.0. When we needed to infer P-values attributed to KEGG pathways for other studies, we uploaded a set of miRNAs identified as biomarkers in the relevant research.

Starbase. Starbase [32] is another web tool. Among several tools provided by Starbase, we used miRPathway(http://starbase.sysu.edu.cn/miRPathway.php) to infer *P*-values attributed to KEGG pathways. Instead of target gene tables inferred computationally, Starbase employed cross-linking immunoprecipitation (CLIP)-Seq data. This has both advantages and disadvantages. One advantage was the certainty of the miRNA target, while one disadvantage was the range of targets. If some genes are listed as targets of some miRNAs, it is very likely true. On the other hand, if no CLIP-Seq data exists for targeted diseases, there are likely no disease-specific miRNA target genes listed. Thus, we employed Starbase to support mirPath software. The lack of detection of KEGG pathways listed by mirPath is not discussed here. All parameters were kept as default values.

Results and Discussion

Simulation

Before performing biomarker selection of miRNAs for real data sets, we performed numerical simulations that compared our proposed method with 2 other methods, i.e., SAM-based and *t*-test based feature extraction (for details, see §1 Simulation in Text S2). In this simulation, we prepared 100 miRNAs with 200 samples. The first 100 samples belonged to category 1, while the second 100 samples belonged to category 2. Among 100 miRNAs, only the first 10 miRNAs exhibited distinct expression between the 2 categories. The task is to select 10 correct miRNAs among the 100 miRNAs and achieve better performance for the discrimination between the 2 categories. We tested 3 scenarios. In the scenario I, expression differences of the 10 miRNAs were kept constant, while noise added to these 10 miRNAs was varied. In scenario II, the expression differences of the 10 miRNAs varied, while the noise added to these 10 miRNAs was kept constant. In scenario III, both expression differences of the 10 miRNAs and noise added to the 10 miRNAs were varied simultaneously (Table 1, for more detailed discussion, see Text S2). Then, we found that our method outperformed the other 2 methods over a wide range of parameters. Thus, we concluded that our method can achieve both better performance of discrimination and more ability to select features that differ between the 2 categories. R core that generates simulation data set used in this study can be found in Text S3.

Biomarker Identification for the Discrimination of Patients with Cancers and other Diseases from Healthy Controls

Based on the findings in the previous section, we employed and applied PCA-based feature extraction to biomarker identification for cancers and other diseases [13]. As we will explain in the "Why did PCA-based feature selection work so well?" subsection in this section, since our PCA-based feature extraction was free from sampling, we could strictly define the top 10 miRNAs that were distinct between pairs of clinical samples and healthy control samples (Table 2). The reasons we employed 10 miRNAs as biomarkers were as follows. First, a previous study [13] extensively studied a situation in which 10 miRNAs were employed as biomarkers, making it easy for us to compare our findings with their discrimination performances, e.g., accuracy, sensitivity, and specificity. Second, as can be seen below, using 10 miRNAs as biomarkers allowed us to achieve sufficiently good performance. Third, measurement of 10 miRNAs is practical for clinical use. Finally, as stated in the previous section, 10 miRNAs were sufficient to achieve a performance comparable to that of all (100) miRNAs.

Table 1. Performance of several feature extraction methods for scenario III.

		Accuracy			# of miRNAs		
D_μ	D_σ	*t* test	PCA	SAM	*t* test	PCA	SAM
2.0	0.5	0.99	0.99	0.99	9.0	8.7	8.0
2.0	1.0	0.95	0.95	0.95	8.1	7.9	7.5
2.0	1.5	0.88	0.88	0.88	7.7	9.2	7.5
2.0	2.0	0.83	0.82	0.82	6.8	9.3	6.9
1.5	0.5	0.98	0.98	0.98	8.7	8.0	7.2
1.5	1.0	0.90	0.88	0.90	7.9	8.0	7.1
1.5	1.5	0.82	0.81	0.82	7.1	8.5	6.7
1.5	2.0	0.77	0.76	0.77	6.4	8.9	6.5
1.0	0.5	0.95	0.89	0.94	8.2	6.2	6.5
1.0	1.0	0.82	0.75	0.81	6.9	6.4	6.0
1.0	1.5	0.71	0.66	0.71	6.2	7.1	5.9
1.0	2.0	0.66	0.63	0.66	5.4	8.1	5.4
0.5	0.5	0.82	0.50	0.80	6.9	0.0	4.5
0.5	1.0	0.66	0.50	0.65	5.3	0.1	3.9
0.5	1.5	0.59	0.50	0.57	3.9	2.1	3.5
0.5	2.0	0.55	0.51	0.55	3.5	4.4	3.4

Accuracy and the number of correctly selected miRNAs among 10 miRNAs with distinct expression between the 2 classes (averaged over 100 trials) for *t*-test-, PCA-, and SAM-based feature extractions. Scenario III was employed. Upper rows indicate easier classification problems. D_μ and D_σ represent the amplitudes of mean and standard deviation of the first 10 miRNAs that exhibit distinct expression between the 2 categories.

doi:10.1371/journal.pone.0066714.t001

We can also make use of these 10 selected miRNAs for discrimination between patients with diseases and healthy controls. The performance of PCA-based LDA between patients with diseases and healthy controls using only these 10 miRNAs is summarized in Table 3. In contrast to Keller *et al.* [13], we successfully identified 10 miRNAs as biomarkers. Keller *et al* could not do this because *t*-test-based feature extraction is highly dependent on divisions between training and test sets. Since they carried out 100 division trials, it would have been impossible for them to create a definite set of 10 miRNAs (see the "Stability" subsection in this section).

Instead of a list of 10 miRNAs used for discrimination, they listed miRNAs that were deregulated in at least 6 diseases (Keller *et al.* [13] Supplementary Table 2). Surprisingly, there was very little overlap between the miRNAs reported by Keller *et al* and the miRNAs reported in our Table 2 in the present study. In fact, the only overlapping miRNA was hsa-miR-16. Even if we took Figure 1 in the study by Keller *et al.* [13] into account, where upregulated miRNAs were considered together, no other miRNAs were selected both in their paper and in the present study.

Recently, Keller *et al.* [33] attempted similar research with next-generation sequencing. They renewed a list of significant miRNAs in the supplementary information of their original study, but again, there were only 2 overlaps with the current study, i.e., miR-425 (for gastric cancer and Wilm's tumor) and miR-140-3p (for melanoma, ovarian cancer, and periodontitis).

Comparison with previous studies. In order to validate our selections independent of the research by Keller *et al*, we have reviewed the literature for previous reports to support our findings that these miRNAs are closely related to cancers and other

Table 2. miRNAs selected to distinguish patients with cancers or other diseases from healthy controls by PCA-based feature extraction.

		lung cancer	other pancreatic tumors and diseases	pancreatitis	ovarian cancer	COPD (chronic obstructive pulmonary disease)	ductal pancreatic cancer	gastric cancer	sarcoidosis	prostate cancer	acute myocardial infarction	periodontitis	multiple sclerosis	melanoma	Wilm's tumor
A	hsa-miR-425	+	+	+	−	+	+	−	+	+	−	−	−	−	−
A	hsa-miR-191	+	+	+	−	+	+	+	-	+	−	−	*	+	−
	hsa-miR-185	−	−	−	−	−	−	−	+	−	−	−	−	+	−
	hsa-miR-140-3p	+	+	+	+	+	−	+	+	+	−	+	+	+	+
B	hsa-miR-15b	−	−	−	+	+	+	−	−	−	+	−	−	−	+
B	hsa-miR-16	−	+	+	+	+	+	+	−	+	−	+	+	−	−
	hsa-miR-320a	+	−	−	−	+	−	−	+	+	+	+	+	+	+
	hsa-miR-486-5p	−	+	+	−	−	+	−	+	−	+	−	−	−	−
C	hsa-miR-92a	±	+	−	+	−	±	+	+	−	+	−	+	±	+
C	hsa-miR-19b	±	*	*	*	*	±	*	*	*	*	*	-	±	*
	hsa-miR-106b	*	+	+	−	+	*	−	*	−	−	−	−	*	−
	hsa-miR-30d	*	*	*	*	*	*	*	+	*	*	*	*	*	*

+ (−) indicates that the miRNA was expressed in patients with cancers or other diseases (healthy controls).
*indicates that the miRNA was not selected within the top 10 most significant miRNAs contributing to discrimination. A–C: miRNAs belonging to common clusters, which were defined by an inter-miRNA distance of 1 kbp. Coincidence within clusters A and C are underlined. See Fig. 5 for actual amount of expression/suppression.
doi:10.1371/journal.pone.0066714.t002

Table 3. Performance of PCA-based LDA for discrimination between patients with cancers or other diseases and healthy controls.

cancer or other disease	PC	Accuracy	Specificity	Sensitivity	Precision
Lung cancer	5	0.784 (+)	0.800 (+)	0.750 (+)	0.632
Other pancreatic tumors and diseases	7	0.814 (+)	0.771	0.875 (+)	0.724
Pancreatitis	8	0.833 (+)	0.786 (−)	0.921 (+)	0.700
Ovarian cancer	6	0.800	0.786 (−)	0.867 (+)	0.464
COPD	2	0.713 (−)	0.671 (−)	0.833 (+)	0.465
Ductal pancreatic cancer	2	0.765 (−)	0.743 (−)	0.800 (+)	0.667
Gastric cancer	9	0.855 (+)	0.857 (+)	0.846	0.524
Sarcoidosis	10	0.835 (−)	0.800 (+)	0.889 (−)	0.741
Prostate cancer	5	0.806 (+)	0.800 (+)	0.826 (+)	0.576
Acute myocardial infarction	7	0.789 (−)	0.900	0.757 (−)	0.964
Periodontitis	10	0.807 (+)	0.814 (+)	0.778 (−)	0.519
Multiple sclerosis	10	0.892 (+)	0.871 (+)	0.957 (+)	0.710
Melanoma	10	0.867 (−)	0.857 (+)	0.886 (−)	0.756
Wilm's tumor	7	0.867	0.886	0.600	0.273

+ (−) indicates that the performance was better (worse) than that of Keller *et al* [13]. PC is the number of PCs used for PCA-based LDA. LOOCV was applied. See the Table in the study by Keller *et al* [13] on page 14 of the Supplementary Materials.

doi:10.1371/journal.pone.0066714.t003

diseases. We discovered a large number of previously published reports supporting the relationship between diseases and the miRNAs observed in this study (Table 2 and Text S1). Although the reports were not always consistent, miR-15b, miR-185, miR-140-3p, miR-320a, miR-486-5p, miR-16, and miR-30d were found to function generally as tumor suppressors, and miR-425, miR 92a, miR 191, miR 106b, and miR-19b were primarily oncogenic. In order to confirm the validity of our evaluation, we listed the reported up- and downregulated miRNAs in several cancers in Table 4. However, since not all miRNAs have been reported to be up- or downregulated, the fact that most of the miRNAs in Table 2 were also included in Table 4 (with the exception of miR-320, miR-486, and miR-191) supports the notion that our findings agree with those of previous studies. Their up- and downregulation patterns are essentially consistent with what we have described above, since a tumor suppressor (oncogene) should be suppressed (expressed) in cancers. Among these, some miRNAs exhibited slightly more complicated functionalities. For example, miR-185 was frequently upregulated in cancers (see Table 4) while its expression sometimes suppressed cell proliferation (see Text S1). Another example of an miRNA with complicated features is miR-15b, which was not always suppressed in tumors. As shown in Table 4 this miRNA was upregulated in colon cancer, but sometimes inhibited tumor function (see Text S1). This somewhat difficult-to-understand situation can be observed in expression profiles as well. Even when reviewing a heat map (Fig. S1), one can discern that no specific expression of miRNAs was associated with cancers and other diseases. Thus, we need to develop approaches that are more sophisticated than observing individual miRNA expression one at a time.

Probability of different disease sharing same miRNA subsets. If we also consider the fact that our list was common for most of the comparisons between healthy control samples and disease samples, we believe that our list of miRNAs as biomarkers to distinguish between patients with cancers or other diseases and healthy controls was accurate. Such a trend would rarely occur only because of simple accidental/coincidental agreement; there are too many miRNAs for this to occur by chance. Suppose that there are N miRNAs and we select N_1 among them. Assuming the selection of 10 miRNAs as biomarkers from a total 862 miRNAs are independent of each other, the expected number of miRNAs being always selected for 14 selections is 8×10^{-27}, when $N = 862$ and $N_1 = 10$. This is much less than the number of common miRNAs in Table 2, i.e., 8 (miR-425, miR-15b, miR-185, miR-92a, miR-140-3p, miR-320a, miR-486-5p, and miR-16). Thus, our list is plausible even if it does differ dramatically from Supplementary Table 2 reported by Keller *et al.* [13]. Nevertheless, there are no theoretical/biological reasons that a set of 10 representative miRNAs used to discriminate between patients with cancers or other diseases and healthy controls must be unique.

Disease-specific co-expression of miRNAs. In order to understand more deeply how each miRNA cooperatively discriminates between cancers or other diseases and healthy controls, we visualized the contribution of each miRNA to discrimination (Fig. 1 and Table 2). Since LDA is a linear method, it allowed us to do this easily (see Materials and Methods).

Interestingly, miRNAs that belong to the same cluster, defined by a inter-miRNA distance of 1 kbp, often share combinations of positive/negative contributions. For example, in Table 2, there are remarkable coincidences between miR-92a and miR-19b in the rows labeled "C" in the left column and those in the same row that are underlined. Three (lung cancer, ductal pancreatic cancer, and melanoma) out of 4 cancers or other diseases for which contributions of miR-19b are listed shared the same outcomes, although they were not significant ($P = 0.3125$). Similarly, miR-425 and miR-191 (rows "A" and underlined in the same row in Table 2) had the same positive/negative contributions for 10 ($P = 0.046$) out of 13 cancers or other diseases, whereas miR-191 made non-zero contributions (3 exceptions: gastric cancer, sarcoidosis, and melanoma). However, since this does not hold true for miR-15b and miR-16 (rows "B" but not underlined in the

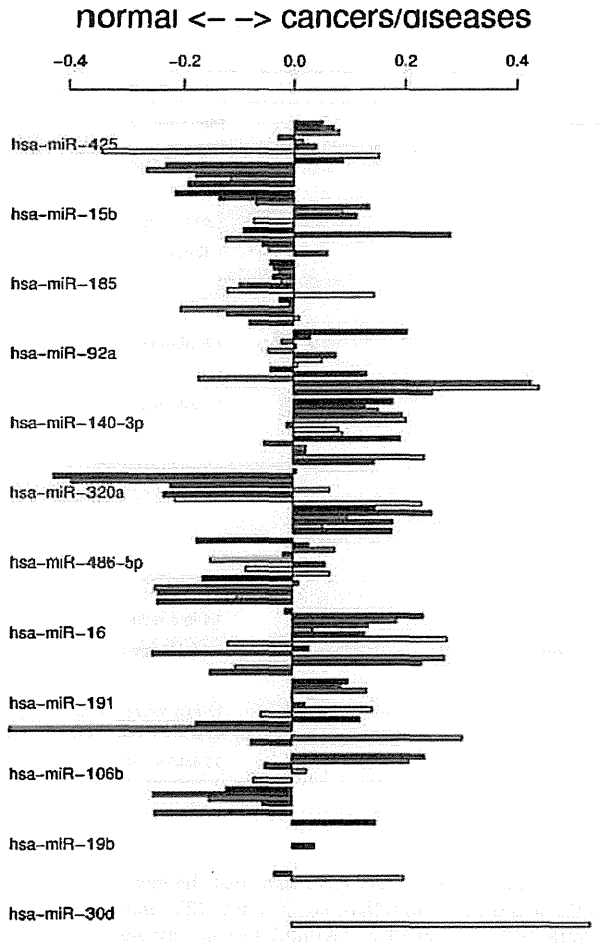

Figure 1. Individual contributions of miRNAs to discrimination between patients with cancers or other diseases and normal controls. The height of the bars indicates the amount of contribution from each miRNA in discriminating patients with cancers or other diseases from healthy controls. A positive (negative) value indicates that the miRNA was expressed in patients with cancer and other diseases (healthy controls). The order of cancers or other diseases is the same as that in Table 2 (top to bottom): lung cancer (black), other pancreatic tumors and diseases (red), pancreatitis (green), ovarian cancer (blue), COPD (cyan), ductal pancreatic cancer (pink), gastric cancer (yellow), sarcoidosis (grey), prostate cancer (black), acute myocardial infarction (red), periodontitis (green), multiple sclerosis (blue), melanoma (cyan), and Wilm's tumor (pink).
doi:10.1371/journal.pone.0066714.g001

same row in Table 2 because of the small number of coincidences), this is again not as straightforward as expected.

Some miRNAs appeared to be consistent with their known functions. For example, miR-486-5p is known to be a tumor-suppressive miRNA (see above and Text S1). As can be seen in Fig. 1, miR-486-5p was more highly expressed in normal controls. On the other hand, miR-92a was more highly expressed in cancers and other diseases, which was consistent with the previous belief that the miR-17-92 cluster is oncogenic.

Moreover, some miRNAs exhibit features contrary to previous findings. For example, miR-106b and miR-425 are believed to be oncogenic miRNAs but are expressed mainly in normal controls (Fig. 1). These apparent discrepancies may result from the measurement of miRNAs from blood samples. If we examine the PhenomiR database [34], we would find many cases in which expression in blood differs from that in tissues. For example, miR-140 is reported to be downregulated in lung cancer tissues (database IDs 132 and 134), but is overexpressed in serum from patients with lung cancer (database ID 503). miR-92a-1 is reported to be downregulated in lung cancer tissues (database IDs 530 and 543), but is overexpressed in serum from patients with lung cancer (database ID 503). These findings in blood are in agreement with those of the present study, demonstrating that miR-140 and miR-92a are expressed in the blood of lung cancer patients (Table 2 and Fig. 1). Similarly, mir-92a is highly expressed in hepatocellular carcinoma (HCC), but is decreased in plasma from HCC patients compared with that from healthy donors [35].

One may still wonder whether miRNAs in the blood can function as useful biomarkers in spite of these disagreements with tissue miRNAs. However, there are many studies that have reported inconsistencies between miRNAs in the blood and tissue miRNAs; these studies have still concluded that miRNAs in the blood can function as useful biomarkers [36–40]. For more detailed discussions of these studies, see §3 Frequent disagreement between blood and tissue miRNAs in Text S2.

KEGG Pathway Analysis for miRNA Target Genes

Although a substantial number of studies have supported that the miRNAs selected in this paper are associated with several cancers and diseases, relating these miRNAs with specific diseases directly and biologically would be a more effective approach. One such method is to check whether any KEGG pathways were enriched with sets of miRNA target genes. As can be seen in the following results, our findings were validated as biologically meaningful. Up- or downregulated sets of miRNAs are selected from Table 2 and uploaded them onto DIANA-mirPath.

Cancer-related pathways. Some pathways directly related to specific cancers were included in the KEGG pathways. For these cancers and cancer-related diseases, it was not difficult to validate whether the up- or downregulated miRNA target genes are related to cancer. In Table 5, we list target-gene enrichment of KEGG pathways annotated as cancers investigated in our study (Table 2). For lung cancer, ductal pancreatic cancer, pancreatitis, other pancreatic tumors and diseases, prostate cancer, and melanoma, corresponding cancer-specific pathways were enriched with miRNA target genes that were up- or downregulated between patients with cancer or cancer-related diseases and healthy controls. Thus, our selection of miRNAs as biomarkers in this study was biologically validated.

Other pathways. Although there were no other KEGG pathways directly related to diseases, many previously known disease-related pathways are enriched with miRNA target genes. For more details about KEGG pathway enrichments related to other diseases, i.e., ovarian cancer, gastric cancer, chronic obstructive pulmonary disease (COPD), acute myocardial infraction, Wilm's tumor, and periodontitis, see §2 KEGG pathway analysis in Text S2.

Stability

In order to confirm the findings above, i.e., commonness of miRNAs that can discriminate patients with cancers or other diseases from healthy controls, we evaluated the stability of the selection (for methodological details, see Materials and Methods).

The concept of stability was defined as follows:

Table 4. miRNAs in Table 2 whose up- and/or downregulation in any cancer was reported in the study by Bandyopadhyay *et al.* [41].

miRNA	Cancer type	Expression	Mean fold change
hsa-miR-425	Central nervous system	Downregulated	13.6-fold reduction
hsa-miR-15b	Colon	Upregulated	1.5-fold increase
hsa-miR-185	Bladder (urothelial)	Upregulated	1.30-fold increase
hsa-miR-185	Kidney	Upregulated	1.42-fold increase
hsa-miR-92-2	Pancreas	Upregulated	
hsa-miR-92-2	Prostate	Upregulated	
hsa-miR-140	Central nervous system	Downregulated	2.7-fold reduction
hsa-miR-140	Colon	Downregulated	11.4-fold reduction
hsa-miR-140	Hematologic	Downregulated	3.5-fold reduction
hsa-miR-140	Lung	Downregulated	
hsa-miR-140	Ovary	Downregulated	3.51-fold reduction
hsa-miR-16-1	Uterus/endometrial _cancer	Upregulated	At least 2-fold increase
hsa-miR-16-2	B cell CLL	Downregulated/Deleted	
hsa-miR-16a	B cell CLL	Downregulated	
hsa-miR-191	Breast	Upregulated	
hsa-miR-191	Central nervous system	Downregulated	4.4-fold reduction
hsa-miR-191	Colon	Upregulated	1.4-fold increase
hsa-miR-191	Lung	Upregulated	
hsa-miR-106b	Lung	Upregulated	12-fold increase in small lung cancer cell line SKLC-2.
hsa-miR-30d	Central nervous system	Downregulated	3.2-fold reduction

Any miRNAs listed in the Additional File in the study by Bandyopadhyay *et al.* [41].
doi:10.1371/journal.pone.0066714.t004

Suppose we have a set of samples, generate subsamplings many times, and apply feature extraction to each subsampling. Stability is defined as the amount of overlapping features over all subsamplings.

If there are no features selected for all subsamplings, we can define stability alternatively by the average frequency that each feature is selected. The importance of this concept was not recognized until very recently. Abeel *et al* [14], pointed out this issue and proposed a new method that grants better performance regarding stability, RFE, and ensemble RFE. However, their method still requires classification information as prior knowledge. Earlier, Varshavsky *et al* [29] described UFF, to our knowledge, the first classification-free feature extraction method.

Thus, the concept of stability was not newly introduced by the present authors, but has already been discussed by others [14,29]. According to our results, our implementation is the suitable that resolves the difficulty of stability of feature extractions.

Although UFF and our method are sampling-independent, as mentioned above, we checked the stability of our method for discrimination between diseased individuals and healthy controls, following the methods of Abeel *et al.* [14] or Varshavsky *et al* [29]. In our case, there were 14 diseases represented in the samples. Thus, we attempted to discriminate the normal control from each of the 14 diseases. Since 10 miRNAs were selected for each discrimination event, a total of 140 miRNAs were selected as biomarkers, although each miRNA could be selected more than once if it was selected in a different discrimination event. These 140 miRNAs could change at every subsampling. Next, we checked whether each miRNA could always be selected for all of subsamplings. If the number of miRNAs that was always selected as biomarkers was large, this means that the method was stable. We found that 8 miRNAs, i.e., hsa-miR-425, hsa-miR-15b, hsa-miR-185, hsa-miR-92a, hsa-miR-140-3p, hsa-miR-320a, hsa-miR-486-5p, and hsa-miR-16, were always selected by our method with 100% probability as biomarkers, independent of cancers and other diseases. Hsa-miR-191 and hsa-miR-106b were selected with 100% probability as biomarkers for 9 and 5 out of 14 cancers or other diseases, respectively. In addition to this, hsa-miR-19b was selected as a biomarker with 100% probability 3 times. Thus, in total, $8\times14+9+5+3=129$ miRNAs were selected as biomarkers with 100% probability. Furthermore, all miRNAs selected as biomarkers for any cancer and diseases (Table S1) have also appeared in Table 2. Thus, it is clear that our miRNA selection was highly independent of sampling.

We also examined the stability by *t*-test-based feature extraction, as proposed by Keller *et al.* [13]. As expected, their results were too scattered to allow for the proposal of well-defined biomarkers consisting of 10 specific miRNAs. By our evaluation, it is very rare for an miRNA to be selected as a biomarker for cancer or other diseases with 100% probability in *t*-test-based feature extraction. In fact, there were only a total of 40 miRNAs selected as biomarkers with 100% probability (Table S1) in *t*-test-based feature extraction; our method identified 129 miRNAs. Based on this, it is almost a certainty that the reason Keller *et al.* [13] could not present 10 specific miRNAs as biomarkers was this heavy fluctuation. We also examined SAM, up- and downregulation by gsMMD, RFE, ensemble RFE, and UFF. We identified 30, 5, 1, 1, 0, and 111 miRNAs, respectively, that were selected as biomarkers with 100% probability (Table S1). Although UFF was as good as our novel method, UFF requires execution of SVD as many times

Table 5. Cancer-specific KEGG pathways enriched in miRNA target genes.

		DIANA-mirpath		Starbase	
		$-\log_{10} P$		adjusted P-value (<0.1)	
		up	down	up	Down
KEGG ID	description	Lung cancer			
hsa05223	Non-small cell lung cancer	3.55	7.31	4.26e-02	-
hsa05222	Small cell lung cancer	2.9	3.64	-	-
hsa05200	Pathways in cancer	-	-	-	5.69e-02
KEGG ID	description	Ductal pancreatic cancer			
hsa05212	Pancreatic cancer	9.71	4.66	1.12e-02	-
hsa05200	Pathways in cancer	-	-	1.10e-02	-
KEGG ID	description	Pancreatitis			
hsa05212	Pancreatic cancer	13.28	12.36	2.14e-05	1.10e-02
hsa05200	Pathways in cancer	-	-	5.11e-05	1.27e-03
KEGG ID	description	Other pancreatic tumors and diseases			
hsa05212	Pancreatic cancer	11.93	15.51	2.20 e-04	1.01e-03
hsa05200	Pathways in cancer	-	-	3.67 e-05	2.90e-03
KEGG ID	description	Prostate cancer			
hsa05200	Pathways in cancer	-	1.53e-04	1.44e-02	
hsa05215	Prostate cancer	8.12	8.98	-	4.04e-04
KEGG ID	description	Melanoma			
hsa05218	Melanoma	6.32	9.21	-	-

A list of cancer-specific KEGG pathways enriched in up- and/or downregulated miRNA target genes between normal controls and corresponding cancer patients. DIANA-mirpath gave $-\log_{10}P$-values while Starbase gave adjusted P-values.
doi:10.1371/journal.pone.0066714.t005

as the number of features (in this case, the number of human miRNAs, i.e., 862), while our method requires execution of PCA only once; thus, there was no need for us to employ UFF instead of our proposed method. Therefore, it is clear that our method outperformed others from the perspective of stability of feature selection.

However, it is also true that UFF, which is the only other classification-free feature extraction method, is the second best method and is comparable to our method. This definitely demonstrates the usefulness of classification-free extraction methods for the identification of blood-borne miRNAs as biomarkers to discriminate between diseased individuals and normal controls.

Discrimination between Diseases

One may wonder whether miRNAs can be used to discriminate not only between normal controls and diseased individuals, but also between diseases. In order to answer this question, we applied our methods to discriminate between diseases. As can be seen in Fig. S2, discrimination between diseases was also quite good. Thus, we may conclude that our methods can discriminate between diseases.

Table 6 lists the miRNAs that were used for discrimination between diseases. The miRNAs in Table 6 are almost identical to those listed in Table 2. Additionally, most of the miRNAs in Table 6 were also included in the miRNA list in Table 2. The miRNAs that were included in Table 6 but were not included with the miRNAs in Table 2, i.e., miR-103, miR-22, and miR-720, were selected only twice each among the total of $140 \times 14 = 1960$ selections (see Table S2). Thus, the miRNAs selected for the discrimination between patients with cancers or other diseases and healthy controls can also discriminate between diseases well, except for discrimination between closely related diseases, e.g., pancreatitis, ductal pancreatic cancer, and other pancreatic tumors and diseases, or lung cancer and melanoma, etc (for details, see Fig. S3 and Table S3).

Why did PCA-based Feature Selection Work So Well?

Finally, we would like explain why PCA-based feature extraction could select biomarkers that could be used to discriminate patients with cancers or other diseases from healthy controls, or to discriminate between patients with distinct diseases, without the knowledge of classification. As can be seen in the Materials and Methods, our PCA-based feature extraction did not consider classification information, even in the training set. One may find this odd because biomarkers should represent the maximum difference between more than 2 distinctive groups. The selection of useful biomarkers should not be possible without the knowledge of classification. However, from our point of view, sample selection itself contains important information about the maximum difference between distinctive groups. If we attempt to gather as many samples as possible belonging to each considered and distinctive group, any features not directly related to classification should be averaged out. For example, if one does not consider gender at all, the male to female ratio should converge to 1 to 1 when a large enough number of samples is collected for each group. This should hold true for any other feature not apparently considered. When PCA was applied to the data set in this situation, the maximum distinctive feature detected should be generated by considering any classification, as the others should have been phased out. We believe that this is the reason our PCA-based feature extraction could detect biomarkers well enough to discriminate between patients with cancers or other

Table 6. miRNAs used for discrimination between diseases.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Lung cancer	140-3p	15b	16	185	191	19b	30d	320a	425	486-5p	92a			
Multiple selerosis	106b	140-3p	15b	16	185	191	19b	30d	320a	425	486-5p	92a		
Other pancreatic tumors and diseases	106b	140-3p	15b	16	185	191	19b	320a	425	486-5p	92a			
Pancreatitis	106b	140-3p	15b	16	185	191	19b	30d	320a	425	486-5p	92a		
Ductal pancreatic cancer	106b	140-3p	15b	16	185	191	19b	320a	425	486-5p	92a			
Gastric cancer	106b	140-3p	15b	16	185	191	19b	22	30d	320a	425	486-5p	92a	
Sarcoidosis	140-3p	15b	16	185	191	19b	30d	320a	425	486-5p	720	92a		
Melanoma	140-3p	15b	16	185	191	19b	30d	320a	425	486-5p	92a			
Wilm's tumor	103	106b	140-3p	15b	16	185	191	19b	30d	320a	425	486-5p	720	92a
Prostate cancer	106b	140-3p	15b	16	185	191	19b	30d	320a	425	486-5p	92a		
Acute myocardial infarction	106b	140-3p	15b	16	185	191	19b	30d	320a	425	486-5p	92a		
Periodontitis	15b	185	140-3p	320a	486-5p	16	92a	425	106b	191	19b	103	30d	
Ovarian cancer	106b	140-3p	15b	16	185	191	19b	22	30d	320a	425	486-5p	92a	
COPD	106b	140-3p	15b	16	185	191	19b	30d	320a	425	486-5p	92a		

Each row lists the miRNAs used for discrimination between the diseases. A set of 10 miRNAs among the miRNAs listed in each row was used for discrimination between the disease shown in the left most column and any of other 13 diseases or normal control. Since 10 miRNAs were selected for each of 14 discrimination analyses, a total of 140 miRNAs were selected as biomarkers. However, there are at most 14 miRNAs listed in each row. In addition to this, miRNAs shown in each row overlapped significantly with each other. This means that miRNAs to be used as biomarkers to discriminate between diseases are highly disease-independent. More detailed information about which 10 miRNAs discriminated between each pair of diseases or control/disease can be found in Table S2. All miRNAs excluding the mRNAs underlined are also in Table 2.

doi:10.1371/journal.pone.0066714.t006

diseases and healthy controls in spite the lack of classification information being considered explicitly.

One may wonder whether other unsupervised clustering methods, e.g., hierarchical clustering and k-means, could have worked as well. However, for these methods, well-defined clusters must exist. This requirement is not always fulfilled. For example, although we tried to apply the feature extraction methods proposed by Chaussabel *et al* [4,5] and based on k-means clustering of transcriptome, which was successfully applied for their purposes, we could not get enough clusters within each disease (at most 10 clusters, often only a few). That is, k-means failed to converge when a large number of clusters of miRNAs was assumed. This prevents us from effectively applying the method of Chaussabel *et al* to our data set of miRNAs. From this point of view, PCA, which does not require any clustering, is better than other unsupervised clustering methods.

Validation Analysis

In order to demonstrate that our proposed method is not specific to present data set treated here, we added two small scale validation analyses for independent data sets (See Tables S10 and S11 in Text S4). The good performance in the validation analysis suggests the robustness of our methodology.

Conclusion

In this paper, we proposed a new feature extraction method based on PCA for biomarker identification from miRNAs in the blood. With simulation data, our method outperformed conventional methods in detecting informative components from a mixture of informative components and noise. When our method was applied to miRNA expression in the blood of patients with cancers or other diseases and normal controls, we identified 10 common miRNAs independent of the cancer or other disease considered. PCA-based LDA using these 10 miRNAs could discriminate patients with cancers or other diseases from healthy controls as well as or slightly better than discrimination using 10 miRNAs selected by *t*-test-based feature selection. We have shown for the first time that the most distinctive feature of cancers and other diseases was not the expression of specific miRNAs, but that of common miRNAs in a cancer- or disease-specific manner.

However, this conclusion may change if more samples are considered and with cost of any technology coming down and the highthroughputs methods getting smaller to fit benchtops, detecting 100s of miRNA biomarkers identified through miRNAome studies might be much efficient and cost effective clinical application soon.

Supporting Information

Figure S1 Heat map of miRNA expression for miRNAs shown in Table 2. Heat map of miRNA expression for miRNAs selected to discriminate between patients with diseases or cancers and normal controls. No miRNAs were shown to be specific to any disease. Thus, it is clear that we need a combination of miRNAs to discriminate between controls and patients with cancers or other diseases.
(EPS)

Figure S2 Percentages of performances shown in Fig. S3. Percentages of discriminant performances (i.e., either of precision, sensitivity, specificity or accuracy shown in Fig. S3) for all pairs of diseases and pairs of normal controls and diseases. The total number of pairs was $15 \times 14/2 = 105$. Percentages are based on the classifications, greater than 0.9 (magenta), between 0.9 and 0.8 (blue), 0.8 and 0.7 (cyan), 0.7 and 0.6 (green), 0.6 and 0.5 (yellow) and less than 0.5 (red).
(EPS)

Figure S3 Performances of discrimination between diseases using the optimal number of PCs. Accuracy, sensitivity, specificity, and precision of each discrimination between diseases using the optimal number of PCs. The 15 columns correspond to, from left to right, lung cancer, control, multiple sclerosis, other pancreatic tumors and diseases, pancreatitis, ductal pancreatic cancer, gastric cancer, sarcoidosis, melanoma, Wilm's tumor, prostate cancer, acute myocardial infarction, periodontitis, ovarian cancer, and COPD. Actual values can also be found in Table S3.
(EPS)

Table S1 Frequency of miRNAs selected by several feature selections. Frequency of miRNA selection within 90% sampling by feature selection based on PCA (100), *t*-test (100), SAM (100), gsMMD_up (100), gsMMD_down (100), RFE (100), RFE ensemble (100), and UFF (100). Numbers in parentheses are the numbers of subsamplings. Cells filled with "100" indicate that the miRNA was always selected by feature extraction for discrimination between patients with the disease denoted at the top of column and healthy controls.
(XLSX)

Table S2 Frequency of miRNAs selected for discrimination between diseases. The pair discriminated is the intersection of the table and column names, e.g., if an miRNA was selected by feature extraction for discrimination between lung cancer and control, 1 was substituted in the cell located in the row named for the miRNA and in the column named as the control in the table named as "lung cancer".
(XLSX)

Table S3 Performance of discrimination between diseases using the optimal number of PCs. Accuracy, specificity, sensitivity, and precision of discrimination between diseases using the optimal number of PCs. The pair discriminated is the intersecting table and row names, e.g., the accuracy in the row named control in the table named "lung cancer" is the accuracy between control and lung cancer.
(XLSX)

Text S1 Previous reports describing miRNAs selected by the proposed feature selection method. Selected papers/reports describing the relationship between the miRNAs reported here and several diseases. PubMed IDs, if available, are reported with brief descriptions of the findings.
(PDF)

Text S2 Supplementary analysis. Detail discussions about simulation, disagreement between tissue miRNA and miRNAs in blood, and KEGG pathway analysis. Fig. S4 and Tables S4 to S9 are included in Text S2.
(PDF)

Text S3 R code. R code that generates simulation data used in this study.
(R)

Text S4 Validation Analyses. Two small scale validation analyses using two independent data sets. It includes Tables S10 and S11. **Table S10, Validation analysis for breast cancer.** Validation of our method using independent sets for breast cancer. **Table S11, Validation analysis for carcinoma *in situ*/squamous cell carcinoma.** Validation of our method using independent sets for carcinoma *in situ*/squamous cell carcinoma.
(PDF)

Author Contributions

Conceived and designed the experiments: YHT YM. Performed the experiments: YHT. Analyzed the data: YHT. Contributed reagents/materials/analysis tools: YHT. Wrote the paper: YHT.

References

1. Van't Westeinde SC, van Klaveren RJ (2011) Screening and early detection of lung cancer. Cancer J 17: 3–10.
2. Hant FN, Silver RM (2011) Biomarkers of scleroderma lung disease: recent progress. Curr Rheumatol Rep 13: 44–50.
3. Nair VS, Krupitskaya Y, Gould MK (2009) Positron emission tomography 18F-uorodeoxyglucose uptake and prognosis in patients with surgically treated, stage I non-small cell lung cancer: a systematic review. J Thorac Oncol 4: 1473–1479.
4. Chaussabel D, Pascual V, Banchereau J (2010) Assessing the human immune system through blood transcriptomics. BMC Biol 8: 84.
5. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, et al. (2008) A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity 29: 150–164.
6. Sauter E, Patel N (2011) Body uid micro(mi)rnas as biomarkers for human cancer. J Nucleic Acids Inv 2: e1.
7. Zen K, Zhang CY (2012) Circulating microRNAs: a novel class of biomarkers to diagnose and monitor human cancers. Med Res Rev 32: 326–348.
8. Yu DC, Li QG, Ding XW, Ding YT (2011) Circulating MicroRNAs: Potential Biomarkers for Cancer. Int J Mol Sci 12: 2055–2063.
9. Scholer N, Langer C, Kuchenbauer F (2011) Circulating microRNAs as biomarkers - true blood? Genome Med 3: 72.
10. Brase JC, Wuttig D, Kuner R, Sultmann H (2010) Serum microRNAs as non-invasive biomarkers for cancer. Mol Cancer 9: 306.
11. Pritchard CC, Kroh E, Wood B, Arroyo JD, Dougherty KJ, et al. (2012) Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies. Cancer Prev Res (Phila) 5: 492–497.
12. Debey-Pascher S, Chen J, Voss T, Staratschek-Jox A (2012) Blood-based miRNA preparation for noninvasive biomarker development. Methods Mol Biol 822: 307–338.
13. Keller A, Leidinger P, Bauer A, Elsharawy A, Haas J, et al. (2011) Toward the blood-borne miRNome of human diseases. Nat Methods 8: 841–843.
14. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26: 392–398.
15. Moghaddam B, Weiss Y, Avidan S (2005) Spectral bounds for sparse pca: Exact and greedy algorithms. In: NIPS. Available: http://www.merl.com/reports/docs/TR2006-007.pdf. Accessed 2013 May 16.
16. Guan Y, Dy JG (2009) Sparse probabilistic principal component analysis. Journal of Machine Learning Research - Proceedings Track 5: 185–192.
17. d'Aspremont A, El Ghaoui L, Jordan MI, Lanckriet GRG (2007) A direct formulation for sparse PCA using semidefinite programming. SIAM Review 49: 434–448.
18. Journée M, Nesterov Y, Richtárik P, Sepulchre R (2010) Generalized power method for sparse principal component analysis. J Mach Learn Res 11: 517–553.
19. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. *Journal of Computational and Graphical Statistics* 15: 265–286.
20. Everitt BS, Dunn G (2010) Applied Multivariate Data Analysis. Wiley.
21. Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5: 845–889.
22. Dy J, Brodley C (2000) Feature subset selection and order identification for unsupervised learning. In: Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 247–254.
23. Liu B, Wan C, Wang L (2006) An efficient semi-unsupervised gene selection method via spectral biclustering. NanoBioscience, IEEE Transactions on 5: 110–114.
24. Liu B, Wan C, Wang L (2004) Unsupervised gene selection via spectral biclustering. In: Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. IEEE, volume 3, 1681–1686.
25. Dy J, Brodley C, Kak A, Broderick L, Aisen A (2003) Unsupervised feature selection applied to content-based retrieval of lung images. Pattern Analysis and Machine Intelligence, IEEE Transactions on 25: 373–378.
26. Modha D, Spangler W (2003) Feature weighting in k-means clustering. Machine learning 52: 217–237.
27. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98: 5116–5121.
28. Qiu W, He W, Wang X, Lazarus R (2008) A marginal mixture model for selecting differentially expressed genes across two types of tissue samples. Int J Biostat 4: Article 20.

29. Varshavsky R, Gottlieb A, Horn D, Linial M (2007) Unsupervised feature selection under perturbations: meeting the challenges of biological data. Bioinformatics 23: 3343–3349.
30. R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available: http://www.R-project.org/. Accessed 2013 May 16. ISBN 3-900051-07-0.
31. Papadopoulos GL, Alexiou P, Maragkakis M, Reczko M, Hatzigeorgiou AG (2009) DIANAmirPath: Integrating human and mouse microRNAs in pathways. Bioinformatics 25: 1991–1993.
32. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, et al. (2011) starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res 39: D202–209.
33. Keller A, Backes C, Leidinger P, Kefer N, Boisguerin V, et al. (2011) Next-generation sequencing identifies novel micrornas in peripheral blood of lung cancer patients. Mol BioSyst 7: 3187–3199.
34. Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, et al. (2010) PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. Genome Biol 11: R6.
35. Shigoka M, Tsuchida A, Matsudo T, Nagakawa Y, Saito H, et al. (2010) Deregulation of miR-92a expression is implicated in hepatocellular carcinoma development. Pathol Int 60: 351–357.
36. Wulfken LM, Moritz R, Ohlmann C, Holdenrieder S, Jung V, et al. (2011) MicroRNAs in renal cell carcinoma: diagnostic implications of serum miR-1233 levels. PLoS ONE 6: e25787.
37. Wu Q, Wang C, Lu Z, Guo L, Ge Q (2012) Analysis of serum genome-wide microRNAs for breast cancer detection. Clin Chim Acta 413: 1058–1065.
38. Wang LG, Gu J (2012) Serum microRNA-29a is a promising novel marker for early detection of colorectal liver metastasis. Cancer Epidemiol 36: e61–67.
39. Kurashige J, Kamohara H, Watanabe M, Tanaka Y, Kinoshita K, et al. (2012) Serum microRNA-21 is a novel biomarker in patients with esophageal squamous cell carcinoma. J Surg Oncol 106: 188–192.
40. Lee Y, Cho H, Lee S, Yun S, Kim J, et al. (2011) Microrna-23a: A novel serum based diagnostic biomarker for lung adenocarcinoma. Tuberc Respir Dis 71: 8–14.
41. Bandyopadhyay S, Mitra R, Maulik U, Zhang MQ (2010) Development of the human cancer microRNA network. Silence 1: 6.