Vol.2013-MPS-96 No.23
Vol.2013-BIO-36 No.23
2013/12/12

Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as candidate drug targets

Ryoichi Kinoshita[1] Mitsuo Iwadate[2,a)] Hideaki Umeyama[2,b)] Y-h. Taguchi[1,c)]

Abstract:
Background: Aberrant DNA methylation is often associated with cancers. Thus, screening genes with cancer-associated aberrant DNA methylation is a useful method to identify candidate cancer-causing genes. Aberrant DNA methylation is also genotype dependent. Thus, the selection of genes with genotype-specific aberrant DNA methylation in cancers is potentially important for tailor-made medicine. The selected genes are important candidate drug targets.
Results: The recently proposed principal component analysis based selection of genes with aberrant DNA methylation was applied to genotype and DNA methylation patterns in squamous cell carcinoma measured using single nucleotide polymorphism (SNP) arrays. SNPs that are frequently found in cancers are usually aberrantly methylated, and the genes that were selected using this method were reported previously to be related to cancers. Thus, genes with genotype-specific DNA methylation patterns will be good therapeutic candidates. The tertiary structures of the proteins encoded by the selected genes were successfully inferred using two profile-based protein structure servers, FAMS and Phyre2. Candidate drugs for three of these proteins, tyrosine kinase receptor (ALK), EGLN3 protein, and NUAK family SNF1-like kinase 1 (NUAK1), were identified by ChooseLD.
Conclusions: We detected genes with genotype-specific DNA methylation in squamous cell carcinoma that are candidate drug targets. Using *in silico* drug discovery, we successfully identified several candidate drugs for the ALK, EGLN3 and NUAK1 genes that displayed genotype-specific DNA methylation.

Keywords: genotype, DNA methylation, principal component analysis, protein tertiary structure, cancer, *in silico* drug discovery, gene selection

1. Introduction

Promoter methylation is widely recognized as an important factor that regulates gene expression, especially in cancers[1], [2]. Many genes with tumor-specific methylated promoters have been identified. For example, the promoters of the PAK3, NISCH, KIF1A, and OGDHL genes are specifically methylated in several cancers, including breast, esophagus, lung, pancreas, colon, prostate, gastric, cervix, thyroid, kidney, head and neck, ovary, and bladder cancers [3]. Because genes with methylated promoters are believed to be suppressive, genes with tumor-specific hypermethylated promoters were assumed to be tumor suppressors. Similarly, genes with tumor-specific hypomethylated promoters were supposed to be oncogenic (i.e., expressed in tumors) and potential oncogene targets. Identification of promoter methylation in cancer genes is important in helping to find critical genes that can cause cancer formation.

Genotype, on the other hand, is another critical factor that can affect cancer formation [3]. Many genotypes are known to be associated with cancers. Currently, there are no established mechanisms that can relate gene mutations to cancer formation. For example, a cancer-specific single nucleotide polymorphism (SNP) is often associated with specific cancers [4], but this SNP is located in an intron of the gene. It is still unclear how intronic SNPs affect gene expression. Typically, cancer-associated genotypes work solely as biomarkers.

Despite of the known importance of DNA methylation and genotype on cancer formation, how DNA methylation and genotype cooperatively mediate cancer formation has rarely been discussed. An exception is the recent association study reported by Scherf et al. [5] who found that genotype-specific promoter DNA methylation of the oncogene CHRNB4 was related to lung cancer. Opavsky et al. [6] also found that the P53, E2f2 and Pten genes in a mouse model of lymphoma were methylated in a genotype-specific manner. Thus, genotype and DNA methylation may contribute cooperatively to cancer formation in many other cancers.

In this paper, we sought to detect genotype-specific DNA methylation in esophageal squamous cell carcinoma (ESCC). Many previous studies have reported ESCC-specific genotypes. For example, Abnet et al. [7] found that genotypic variants at position 2q33 on the human chromosome were related to risk of ESCC. Maeng et al. [8] found that phosphoinositide-3-kinase and BRAF mutations were associated with metastatic ESCC and Wang et al. [9] found that ESCC was related to polymorphisms

[1] Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan
[2] Department of Biological Science, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan
[a)] iwadate@bio.chuo-u.ac.jp
[b)] umeyama@bio.chuo-u.ac.jp
[c)] tag@granular.com

in ALDH2 and ADH1B in Chinese females. Thus, genotype-specific DNA methylation is expected to exist widely in ESCC. In this study, we used two publicly available distinct SNP microarray data sets to identify genotype-specific DNA methylation in ESCC.

2. Results

2.1 Estimation of genotype-specific DNA methylation

There is no unique criterion that can estimate genotype-specific DNA methylation. Aberrant methylation itself can be estimated by various criteria; for example, using the ratio or the difference of mean values between normal and tumor tissues or using *P*-values obtained by a statistical test such as a t-test. Each of the criterion may give a different genotype-specific DNA methylation set of genes. In addition, some genotypes are either heavily demethylated or methylated in tumor tissue compared with normal tissue. If this genotype is very rare in the tumor tissue, it is clearly unreasonable to regard this genotype-specific DNA methylation as being the cause of the tumor. Ideally, to be sure that a particular genotype-specific DNA methylation could cause the tumor, the following conditions should be satisfied:

(1) The genotype is specifically demethylated/methylated in the tumor tissue compared with other genotypes (strength of aberrant DNA methylation).
(2) The genotype is abundant in the tumor tissue (abundance of aberrant DNA methylation).

The best balance between these two conditions is not easy to estimate, because there is no standard understanding about the kind of gene abnormalities that generally cause tumors. In this study, we used three kinds of samples, and blood, normal and tumor tissues. This made the comparisons more difficult than a comparison between only normal and tumor tissues, because we are not sure if normal tissue is an expected intermediate between blood and tumor. To avoid uncertainties that this complicated situations might cause when estimating genotype-specific DNA methylation, we employed a recently proposed PCA-based unsupervised feature selection method[10]. This procedure does not require the user to select the criterion that is used to estimate genotype-specific DNA methylation. It is necessary simply to select the suitable PC by which the SNPs with genotype-specific DNA methylation are selected. Over all design of selecting gene with genotype specific DNA methylation is shown in Fig. 1.

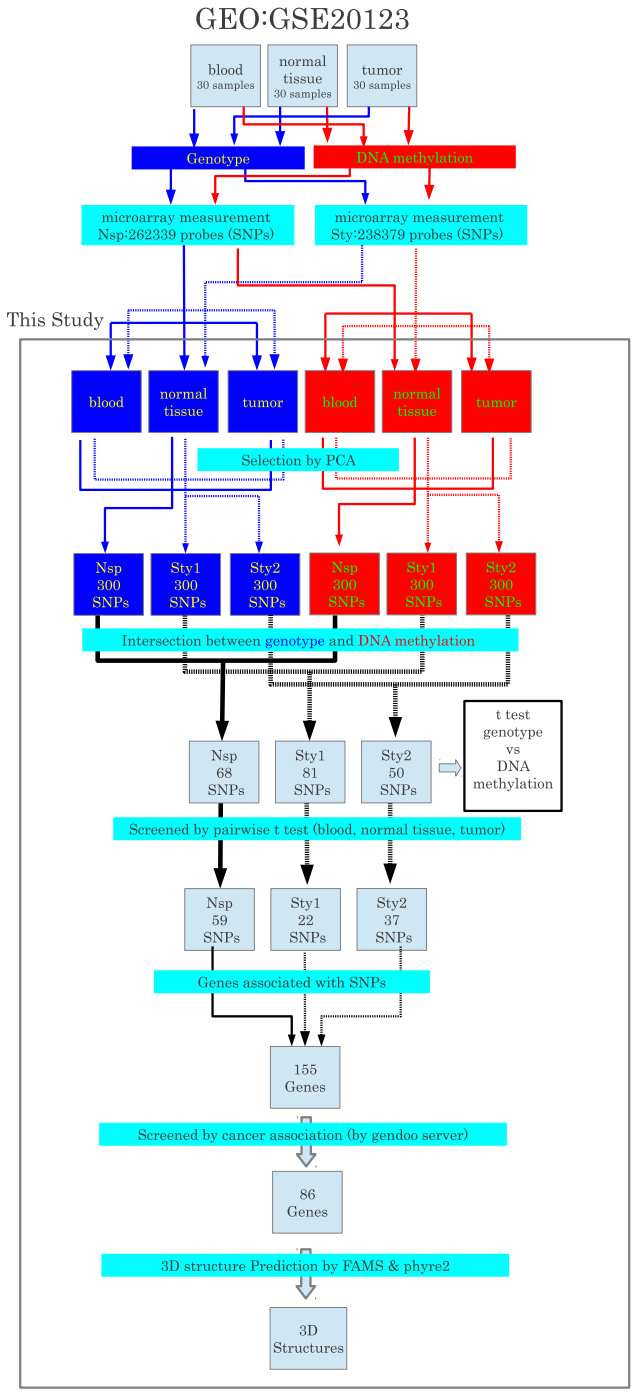

Fig. 1 Schematic illustration of the gene screening process. The grey rectangle indicates the processes performed in this study. The red (blue) boxes indicate the data processing flow for the genotype (DNA methylation) data. The solid (dotted) lines indicate data processing flow for the Nsp (Sty) measurements. Sty1 and Sty2 indicate the two combinations of PCs that were used; PC4 for genotype /PC3 for DNA methylation, and PC3 for genotype /PC4 for DNA methylation.

2.2 Genotype-specific DNA methylation estimated using the Nsp microarray data

The PCs obtained when PCA was applied to the Nsp microarray measurements of genotype are shown in Fig. 2. Although the first PC (PC1; Fig. 2a) had the dominant contribution (80%), no significant differences between blood, and the normal and tumor tissues were seen. On the other hand, the second PC (PC2; Fig. 2b) clearly distinguished between blood, and normal and tumor tissues. Therefore, we used PC2 to select probes (SNPs) that exhibited significant differences between the blood, and normal and tumor tissues. Because PC3 (not shown here) exhibited no significant differences between the blood, normal and tumor samples and had very little contribution, we did not use the third PC (PC3) to select SNPs.

The PCs obtained when PCA was applied to the Nsp microarray measurements of DNA methylation are shown in Fig. 3. PC2 (Fig. 3b) was again the PC that clearly distinguished between blood, and normal and tumor tissues. PC2 was, therefore, used to

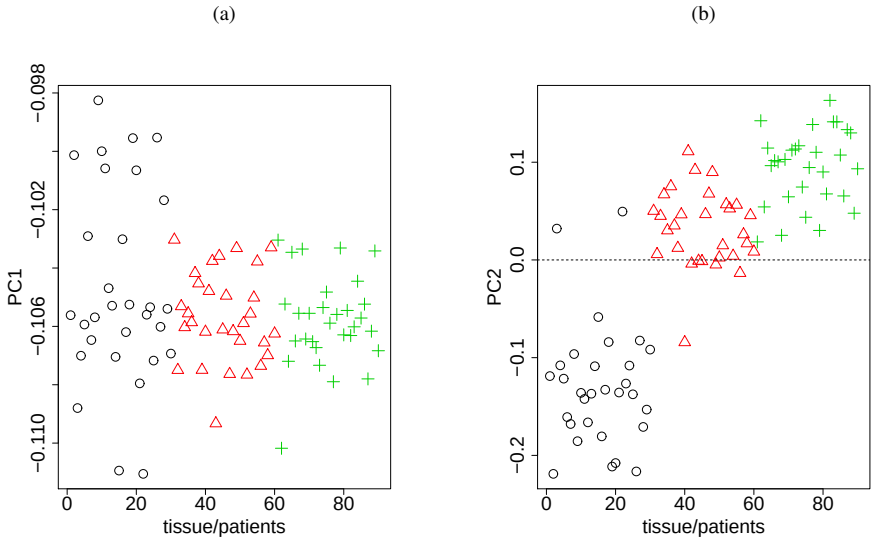

Fig. 2 PCs for genotypes measured by Nsp microarray. (a) PC1 (81%). (b) PC2 (3%). Black circle, blood; red triangle, normal tissue; green cross, tumor tissue. The horizontal axes indicate the subjects and their samples. The order of the 30 subjects in the 1—30, 31—60, and 61—90 sections are the same; i.e., 1, 31, and 61 are samples from the same patient.

select the SNPs that exhibited significant differences between the three samples.

The two dimensional (PC1 and PC2) embedding of SNPs (probes) for DNA methylation and genotype are shown in Fig. 4. Because PC2 showed significant differences between the blood, and normal tissues and tumor tissues, we selected the 300 topmost outliers along the PC2 axis for both DNA methylation and genotype. To see if genotype-specific methylated SNPs were selected correctly, we filtered the selected SNPs based on the following criteria:

(1) Intersection between top N outliers between DNA methylation and genotype.
(2) All three associated P-values adjusted by the BH criterion are less than 0.05, when three pairwise one-sided t-tests (tumor tissue vs normal tissue, normal tissue vs blood, tumor tissue vs blood) are applied.

A total of 68 SNPs were selected in common from the top 300 outliers between genotype and DNA methylation after applying the first criterion. Because there were more than 250,000 SNPs on the Nsp microarray, the P-value for 68 SNPs being selected in common from 300 is less than 1×10^{-16} . After applying the P-value filtering (the second criterion) 59 SNPs were filtered as SNPs with genotype specific DNA methylation.

2.3 Genotype-specific DNA methylation estimated using the Sty microarray data

After repeating similar procedures to Sty microarray, using the third and fourth PCs, we identified two sets of SNPs with genotype specific DNA methylation, having 81 and 50 genes, respectively (the first criterion, not shown here). Then applying p-value based screening (the second criterion), finally 22 and 37 SNPs were filtered as those with genotype specific DNA methylation (not shown here).

3. Discussion

3.1 Properties of the selected SNPs

Almost all selected SNPs were located outside protein cording regions of the genes excluding four exceptions. Thus, the majority of the SNPs are presumably related to the regulation of gene expression. The SNPs that were not located in protein coding regions were located in the promoters, and also in introns and in the downstream regions of genes. Thus, the effect of genotype-specific DNA methylation on gene expression is not straightforward.

In addition, some of the selected SNPs have not been reported in Chinese populations, although all patients in the microarray data sets that we used in this study were Chinese. This finding indicates that we have correctly selected mutation that may cause cancer formation.

3.2 Screening of cancer-related genes

To determine if the selected SNPs are biologically related to cancers, the genes containing the SNPs were annotated using Gendoo[11], [12]. The RefSeq mRNA IDs of the genes were extracted from GEO and mapped to gene symbols. The gene symbols were uploaded to the Gendoo server and the diseases that were reported to be associated with each of the gene symbols were listed. We found that 86 of the 155 genes associated with selected SNPs were also associated with at least one cancer-related disease. In addition, we performed a literature search to find papers that reported the relationship between any of the 86 selected genes and cancers, because the Gendoo server annotation is based on automated text-mining and may include some misinterpretations. We found that most of 86 genes were mentioned in at least one published paper that described their relationship with cancer. Thus, we confirmed that more than half (86) the 155 genes screened by our method were cancer-related genes. In particular, twelve genes (CCND1, CCNL1, CKAP4,

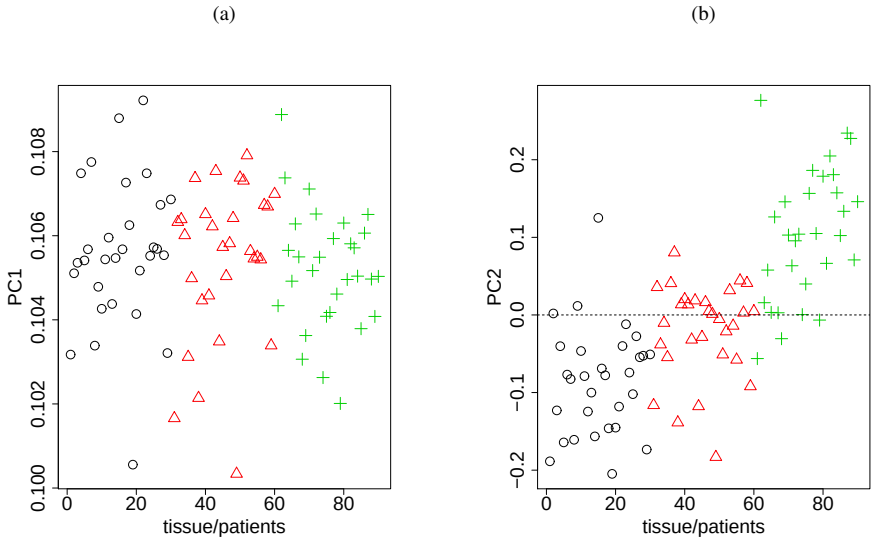

Fig. 3 PCs for DNA methylation measured by Nsp microarray. (a) PC1 (80%). (b) PC2 (3%). Other notations are the same as those in Fig. 2.

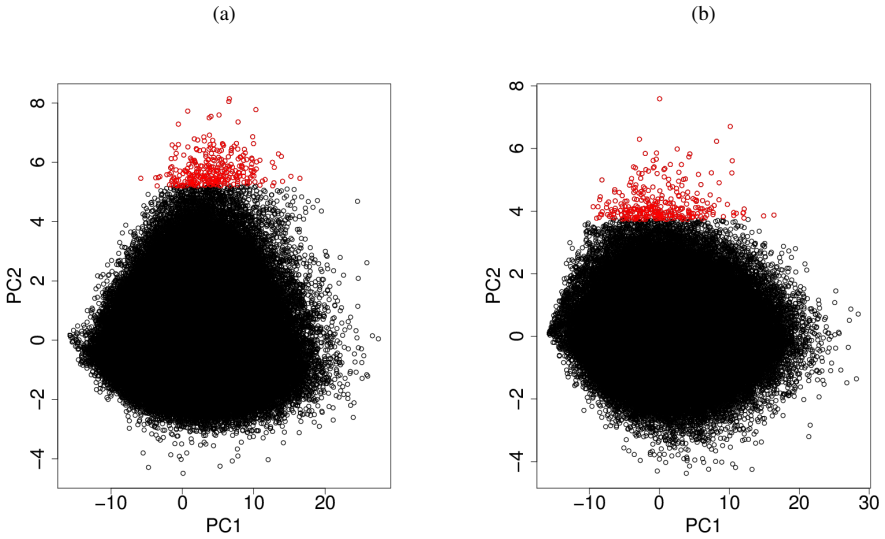

Fig. 4 Two dimensional embedding of SNPs with PC1 and PC2 for the Nsp microarray measurements (a) Genotype (Fig. 2). (b) DNA methylation (Fig. 3). The top 300 outliers are shown in red.

CRABP1, FGF3, GRHL2, MYEOV, PKP4, RAP2B, RPL14, SMAD3, ZNF639) were associated with "Carcinoma, Squamous Cell" and eleven genes (CCND1, CKAP4, CRABP1, EVI1, FGF3, MYEOV, PKP4, RPL14, SMAD3,TMEM16A,ZNf639) were associated with "Esophageal Neoplasms". Among them, nine genes are associated with both. Because this study used data sets for ESCC (esophageal squamous cell carcinoma), this association is reasonable and demonstrates the reliability of our method.

3.3 Genes with genotype-specific DNA methylation are less methylated than expected

We compared the microarray measurements between genotype and DNA methylation of the probes selected in common and found that the microarray DNA methylation measurements were always less than the genotype measurements. This observation is interesting, because a less methylated promoter usually indicates a more expressive genes, although not all the selected SNPs with DNA methylation were in the promoter region of the genes. To check that the demethylation was not because of inaccurate microarray measurement normalization, we randomly sampled the same number of SNPs as those in Tables 1, 2, and 3 1,000 times, and computed *P*-values adjusted by the BH criterion[13]. We found that typically less than 1 % of the trials had adjusted *P*-vales < 0.05. Thus, we determined that there were no normalization biases in the data sets and the low observed *P*-values shown in Table 6 were not obtained because of fluctuations.

3.4 Structure prediction of the proteins associated with selected genes

Although we selected genes with genotype-specific DNA methylation, for therapeutic purposes, we need to design drugs for the proteins that are encoded by these genes. To identify candidate drugs computationally, the tertiary structures of the target proteins are required as templates. However, the structures of many of the encoded proteins have not been reported.

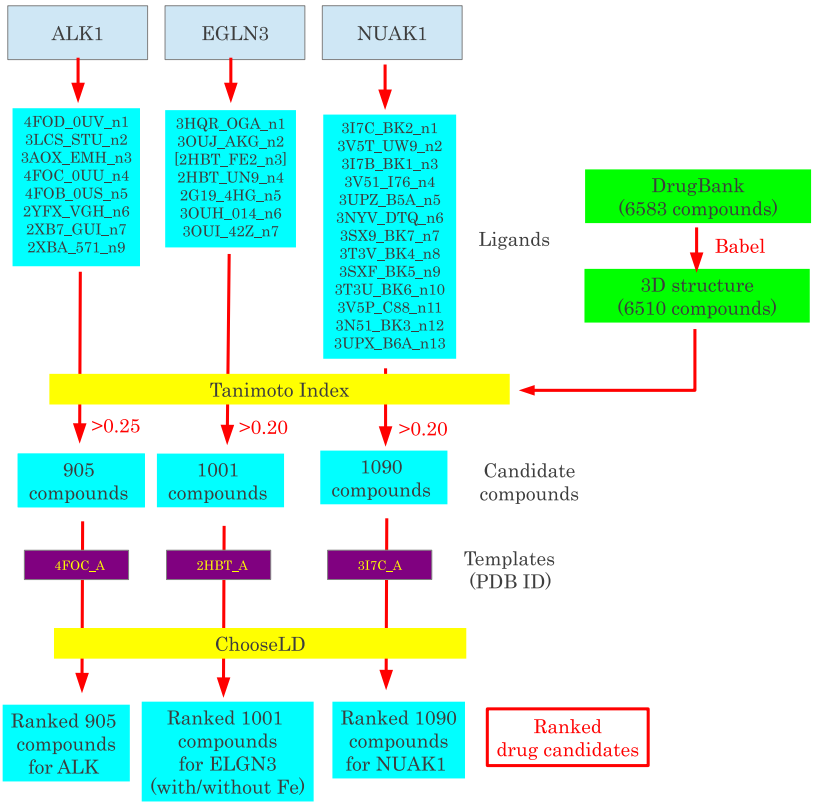

Fig. 5 Schematic illustration of the drug discovery process. For the proteins encoded by the selected genes (ALK, EGLN3 and NUAK1), about 1,000 compounds, selected based on the Tanimoto index from DrugBank, were tested by ChooseLD using template protein structures from PDB. The templates are specified by their PDB IDs. The ligands are specified by the PDB ID, ligand name and a sequential number. For example, 3I7C_BK2_n1 indicates ligand BK2 (1-tert-butyl-3-naphthalen-2-yl-1H-pyrazolo[3,4-d]pyrimidin-4-amine) included in PDB entry 3I7C [PDB: 3I7C], and n1 means no.1. The drug discovery process for EGLN3 was performed twice, with and without Fe as a ligand. When Fe was excluded as a ligand, it was regarded as a mediator. That is, Fe bounds to the protein during docking simulation, but was excluded from the Tanimoto index computation.

To obtain the tertiary structure of these proteins, we used two protein structure prediction servers FAMS[14], [15] and phyre2[16], [17] to predict the structure using only the amino acid sequence of the protein.

Some protein structures were already in the protein data bank (PDB)[18], if not, they were modeled using the structure of a suitable reference protein. These structures were then used as templates to predict drug candidates *in silico*.

For the proteins that were not in the PDB, for the reference proteins that were used for the structure prediction, we sought cancer-related papers that cited the reference proteins. Most of reference proteins used for structure prediction were cancer-related. This finding also suggests that our gene selection process and protein structure prediction are plausible.

3.5 *In silico* drug discovery

We tried to design drugs that could bind to some of the protein templates using an *in silico* drug discovery method in which chemical compounds that potentially bind to proteins and suppress protein functions were sought computationally. For this purpose, we selected the three proteins encoded by ALK, EGLN3, and NUAK1 as drug targets, based upon a literature search and the gene annotations that indicated that these genes were expressed in cancer and had potentially functional binding pockets (e.g., protein kinase) for ligands. The drug discovery process that we used is illustrated in Fig. 5.

After the FPAScores that represent binding affinities of compounds to proteins were estimated, to check if three independent trials were feasible, we tested coincidence between three trials in two ways. First, we computed the correlation coefficients between three independent trials. For all pairwise computations for ALK, EGLN3, and NUAK1, the correlation coefficients were greater than 0.9. This suggests that the FRAScores computed by ChooseLD were highly reproducible. However, the correlation coefficients represent the overall reproducibilities of FPAScores for the candidate drug compounds. It is more important that the compounds with higher FPAScores, i.e., those regarded as being highly reliable, were reproducible. Therefore, we checked how often the highly ranked compounds were selected between the three trials and found that the selection of the highly ranked compounds was also highly reproducible.

Among the 10 top-ranked compounds for ALK (Table 1), eight compounds targeted cancer genes, and two out of the eight targeted ALK. Among the 10 top-ranked compounds for ELGN3

Table 1 The 10 top-ranked compounds as drug targets for ALK. The compounds were ranked based on FPAScores averaged over three independent trials and their representative target cancer genes.

DrugBank ID	Compound name	Representative target cancer genes
	ALK	
DB01933	7-Hydroxystaurosporine	PDK1
DB08700	3-[(1R)-1-(2,6-dichloro-3-fluorophenyl)ethoxy] -5-(1-piperidin-4-yl-1H-pyrazol-4-yl)pyridin-2-amine	**ALK**, c-MET, LCK, TRKA, TRKB, TIE2, ABL
DB04651	BIOTINOL-5-AMP	—
DB02491	4-[4-(1-Amino-1-Methylethyl)Phenyl]-5-Chloro-N -[4-(2-Morpholin-4-Ylethyl)Phenyl]Pyrimidin-2-Amine	FGFR2
DB07006	9-HYDROXY-6-(3-HYDROXYPROPYL)-4 -(2-METHOXYPHENYL)PYRROLO[3,4-C] CARBAZOLE-1,3(2H,6H)-DIONE	WEE1
DB02010	Staurosporine	ITK, SYK, MAPKAPK2, GSK3, CSK, CDK, PIK3CG, ZAP-70
DB02654	6-Hydroxy-Flavin-Adenine Dinucleotide	—
DB07460	2-({5-CHLORO-2-[(2-METHOXY-4-MORPHOLIN -4-YLPHENYL)AMINO]PYRIMIDIN-4 -YL}AMINO)-N-METHYLBENZAMIDE	**ALK**, PTK2
DB07186	4-(4-METHYLPIPERAZIN-1-YL)-N-[5 -(2-THIENYLACETYL)-1,5-DIHYDROPYRROLO [3,4-C]PYRAZOL-3-YL]BENZAMIDE	AURKA, PLK1
DB03247	Riboflavin Monophosphate	RPS6KA4, POR(P450), SGK1, NOS1, DPYD, DHODH

(not shown here), including Fe as a ligand, eight compounds targeted cancer genes and two out of the eight targeted EGLN1, which is paralog of EGLN3. Among the 10 top-ranked compounds for ELGN3 (not shown here), without including Fe as a ligand but as a mediator, six were in common with the top-ranked compounds for EGLN3 when Fe was included as a ligand. Among the other four compounds, one targeted EGLN1. Of the 10 of the top-ranked compounds for NUAK1 (not shown here), most target more than 100 other genes and thus lack specificity. All of these findings suggested that the top-ranked compounds for each of the proteins were feasible candidate drugs.

Note

Full paper version has been accepted to be published in the supplement of BMC Sys. Biol. as the APBC2014 proceedings at Jan. 2014 after this presentation was submitted to SIGBIO36.

Acknowledgments We would like to thank Dr. Katsuichiro Komatsu who helped with the *in silico* drug screening using ChooseLD. This research was supported by KAKENHI, 23300357 and Chuo University Joint Research Grant.

References

[1] Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R. A. and Issa, J. P.: Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters, *PLoS Genet.*, Vol. 3, No. 10, pp. 2023–2036 (2007).

[2] McCabe, M. T., Brandes, J. C. and Vertino, P. M.: Cancer DNA methylation: molecular mechanisms and clinical implications, *Clin. Cancer Res.*, Vol. 15, No. 12, pp. 3927–3937 (2009).

[3] Pasche, B. and Yi, N.: Candidate gene association studies: successes and failures, *Curr. Opin. Genet. Dev.*, Vol. 20, No. 3, pp. 257–261 (2010).

[4] Zhou, W.: Mapping genetic alterations in tumors with single nucleotide polymorphisms, *Curr Opin Oncol*, Vol. 15, No. 1, pp. 50–54 (2003).

[5] Scherf, D. B., Sarkisyan, N., Jacobsson, H., Claus, R., Bermejo, J. L., Peil, B., Gu, L., Muley, T., Meister, M., Dienemann, H., Plass, C. and Risch, A.: Epigenetic screen identifies genotype-specific promoter DNA methylation and oncogenic potential of CHRNB4, *Oncogene* (2012).

[6] Opavsky, R., Wang, S. H., Trikha, P., Raval, A., Huang, Y., Wu, Y. Z., Rodriguez, B., Keller, B., Liyanarachchi, S., Wei, G., Davuluri, R. V., Weinstein, M., Felsher, D., Ostrowski, M., Leone, G. and Plass, C.: CpG island methylation in a mouse model of lymphoma is driven by the genetic configuration of tumor cells, *PLoS Genet.*, Vol. 3, No. 9, pp. 1757–1769 (2007).

[7] Abnet, C. C., Wang, Z., Song, X. et al.: Genotypic variants at 2q33 and risk of esophageal squamous cell carcinoma in China: a meta-analysis of genome-wide association studies, *Hum. Mol. Genet.*, Vol. 21, No. 9, pp. 2132–2141 (2012).

[8] Maeng, C. H., Lee, J., van Hummelen, P., Park, S. H., Palescandolo, E., Jang, J., Park, H. Y., Kang, S. Y., MacConaill, L., Kim, K. M. and Shim, Y. M.: High-throughput genotyping in metastatic esophageal squamous cell carcinoma identifies phosphoinositide-3-kinase and BRAF mutations, *PLoS ONE*, Vol. 7, No. 8, p. e41655 (2012).

[9] Wang, Y., Ji, R., Wei, X., Gu, L., Chen, L., Rong, Y., Wang, R., Zhang, Z., Liu, B. and Xia, S.: Esophageal squamous cell carcinoma and ALDH2 and ADH1B polymorphisms in Chinese females, *Asian Pac. J. Cancer Prev.*, Vol. 12, No. 8, pp. 2065–2068 (2011).

[10] Ishida, S., Umeyama, H., Iwadate, M. and Taguchi, Y. H.: Bioinformatic screening of autoimmune disease genes and protein structure prediction with FAMS for drug discovery. in press.

[11] Nakazato, T., Bono, H., Matsuda, H. and Takagi, T.: Gendoo: functional profiling of gene and disease features using MeSH vocabulary, *Nucleic Acids Res.*, Vol. 37, No. Web Server issue, pp. W166–169 (2009).

[12] : Gendoo, , available from ⟨http://gendoo.dbcls.jp/⟩

[13] Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300 (1995).

[14] : FAMS, , available from ⟨http://fams.bio.chuo-u.ac.jp/fams/⟩

[15] Umeyama, H. and Iwadate, M.: FAMS and FAMSBASE for protein structure, *Curr Protoc Bioinformatics*, Vol. Chapter 5, p. Unit5.2 (2004).

[16] : Phyre2, , available from ⟨http://www.sbg.bio.ic.ac.uk/phyre2/⟩

[17] Kelley, L. A. and Sternberg, M. J.: Protein structure prediction on the Web: a case study using the Phyre server, *Nat Protoc*, Vol. 4, No. 3, pp. 363–371 (2009).

[18] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E.: The Protein Data Bank, *Nucleic Acids Res.*, Vol. 28, No. 1, pp. 235–242 (2000).