DNA methylation marker to estimate the breast cancer cell fraction in DNA samples

doi: 10.1007/s12032-018-1207-3(https://doi.org/10.1007/s12032-018-1207-3)

DNA methylation marker to estimate the breast cancer cell fraction in DNA samples

Hiroki Ishihara[1, 2], Satoshi Yamashita[1], Satoshi Fujii[3], Kazunari Tanabe[2], Hirofumi Mukai[4], Toshikazu Ushijima[1*]

[1]Division of Epigenomics, National Cancer Center Research Institute, Tokyo, Japan

[2]Department of Urology, Tokyo Women's Medical University, Tokyo, Japan

[3]Division of Pathology, Exploratory Oncology Research and Clinical Trial Center, National Cancer Center, Kashiwa, Chiba, Japan

[4]Department of Breast and Medical Oncology, National Cancer Center Hospital East, Kashiwa, Chiba, Japan

***Corresponding author**

Dr. Toshikazu Ushijima

Division of Epigenomics, National Cancer Center Research Institute

5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

Tel: +81 3 3547 5240

FAX: +81 3 5565 1753

E-mail address: tushijim@ncc.go.jp

Short title

Breast cancer cell fraction marker

Abstract

Estimation of the cancer cell fraction in breast cancer tissue is important for exclusion of samples unsuitable for multigene prognostic assays and a variety of molecular analyses for research. Here, we aimed to establish a breast cancer cell fraction marker based on DNA methylation. First, we screened genes unmethylated in non-cancerous mammary tissues and methylated in breast cancer tissues using microarray data from the TCGA database, and isolated 12 genes. Among them, four genes were selected as candidate marker genes without a high incidence of copy number alterations and with broad coverage across patients. Bisulfite pyrosequencing analysis of additional breast cancer biopsy specimens purified by laser capture microdissection (LCM) excluded two genes, and a combination of *SIM1* and *CCDC181* was finally selected as a fraction marker. In further additional specimens without LCM purification, the fraction marker was substantially methylated (≥ 20%) with high incidence (50/51). The cancer cell fraction estimated by the fraction marker was significantly correlated with that estimated by microscopic examination ($p < 0.0001$). Performance of a previously established marker, *HSD17B4* methylation, which predicts therapeutic response of HER2-positive breast cancer to trastuzumab, was improved after the correction of cancer cell fraction by the fraction marker. In conclusion, we successfully established a breast cancer cell fraction marker based on DNA methylation.

Key Words

DNA methylation; cancer cell fraction; breast cancer; trastuzumab; HER2; cancer cell content; *HSD17B4*

Introduction

Accurate molecular analyses of cancer tissues, such as genomic sequencing, gene expression analysis, and epigenetic analysis, can be achieved by taking account of co-existing non-cancerous cells in the cancer tissues [1-4]. A gold standard method to estimate a cancer cell fraction is microscopic cell counting using pathological sections. However, this method is time-consuming, and distinction of cancer cells from co-existing non-cancerous cells is sometimes difficult. To overcome this issue, we established a method to estimate the cancer cell fraction in DNA samples based on DNA methylation [5, 6]. Since DNA methylation patterns are specific to individual cell types [7-13], the cancer cell fraction can be estimated using a small number of genes specifically methylated in cancer cells, not in non-cancerous cells [5]. Because the analysis is conducted using DNA samples, histological sections are unnecessary for this method.

In breast cancer, extensive molecular analyses, including multigene prognostic assays such as Oncotype DX or MammaPrint, are conducted for clinical practice and research. In the multigene prognostic assays, samples with a low cancer cell fraction must be excluded [14, 15]. Among various research, for example, *HSD17B4* methylation predicts pathological complete response (pCR) in patients with HER2-positive breast cancer after trastuzumab therapy [16]. For this prediction, the *HSD17B4* methylation levels need to be corrected by the cancer cell fraction using microscopic examination. Thus, once we can establish the cancer cell fraction in breast cancer, it is expected to reduce the workload of pathologists.

In this study, we aimed to establish a DNA methylation marker to estimate breast cancer cell fractions.

Materials and methods

Breast cancer biopsy specimens and blood samples

Most of the breast cancer specimens (58 of 61) were obtained from our previous study [16], and the remaining three were newly obtained. The clinical study, along with this exploratory study, was approved by the National Cancer Center Ethics Committee (approval no. 2010-250), and was registered at the UMIN Clinical Trial Registry (Registration no. UMIN000007074). Written informed consents were obtained from all participants. All the 61 specimens were collected by needle biopsy from 61 patients with a diagnosis of HER2-positive breast cancer (Supplementary Table 1). The specimens were fixed using the PAXgene Tissue System (Qiagen, Hilden, Germany) and embedded in low-melting paraffin for DNA extraction using 10 slices of 10 μm sections. A certified and experienced pathologist (S.F.) conducted microscopic examination of the specimens to determine the cancer cell fraction. The pathological complete response (pCR) to trastuzumab was defined as no residual cancer cells in the specimens according to the system of American Joint Committee on Cancer. Among the 61 specimens, 10 specimens were purified by laser capture microdissection (LCM). Three peripheral leucocyte samples were obtained from three healthy volunteers.

Breast cancer cell lines and human mammary epithelial cells

A total of 20 human breast cancer cell lines (BT-474, SK-BR-3, MDA-MB-453, HCC38, MDA-MB-231, T-47D, Hs 578T, MCF7, UACC-3199, ZR-75-1, BT-20, MDA-MB-436, HCC1937, MDA-MB-468, HCC1428, BT-549, AU565, HCC1395, MDA-MB-157, and HCC1954) were purchased from the American Type Culture Collection

(Rockville, MD). Human Mammary Epithelial Cells (HMECs) were purchased from Cambrex (East Rutherford, NJ).

Genome-wide DNA methylation analysis

Genome-wide DNA methylation analysis of HMECs, peripheral leucocyte samples, and breast cancer cell lines was conducted using an Infinium HumanMethylation450 BeadChip array (Illumina, San Diego, CA, USA), as previously reported [17]. Additionally, we downloaded HumanMethylation450 data of 27 breast cancer tissues and 15 non-cancerous mammary tissues randomly from the 1,234 breast tissue samples registered in the TCGA database (Supplementary Table 2). The DNA methylation level of an individual probe was obtained as a β value that ranged from 0 (unmethylated) to 1 (fully methylated). From all of 482,421 probes located on CpG sites, we excluded 5,077 probes at genomic positions that could not be identified according to the human genome assembly hg38. The remaining 477,344 probes on autosomes and sex chromosomes were evaluated in this study.

Measurement of DNA methylation levels of specific genomic regions

Gene-specific DNA methylation levels were analyzed by bisulfite pyrosequencing. Specifically, bisulfite modification was conducted using 1 μg of *Bam*HI-digested genomic DNA, as previously reported [18]. The modified DNA was suspended in 40 μl of TE buffer, and an aliquot of 1 μl was used for bisulfite pyrosequencing [19]. A target genomic region was amplified by biotinylated primers. The PCR product labelled with biotin was annealed to a 0.2 μM pyrosequencing primer, and pyrosequencing was carried out using the PSQ 96 Pyrosequencing System (QIAGEN, Valencia, CA, USA). The

methylation level was obtained using PSQ Assay Design software (QIAGEN).

Correction of *HSD17B4* methylation level by the breast cancer cell fraction

The methylation level of *HSD17B4* was corrected by the breast cancer cell fraction estimated by a fraction marker or by microscopic examination as follows: [Corrected *HSD17B4* methylation level = 100 × (*HSD17B4* methylation level)/ (a cancer cell fraction)].

Statistical analysis

The correlation analysis was performed using the Pearson's product-moment correlation coefficients. Differences in corrected methylation level of *HSD17B4* between trastuzumab responders (pCR specimens) and non-responders (non-pCR specimens) were evaluated by the Mann–Whitney *U* test. All the analyses were performed using PASW statistics version 18.0 (SPSS Japan Inc., Tokyo, Japan), and two-sided p-values < 0.05 were considered statistically significant.

Results

Isolation of genomic regions specifically methylated in breast cancer cells

To isolate genomic regions specifically methylated in breast cancer cells, we first selected 136,830 probes unmethylated (β value ≤ 0.2) in the non-cancerous mammary cells (HMECs, peripheral leucocyte samples, and non-cancerous mammary tissues) from the 477,344 probes located on autosomes and sex chromosomes (Figure 1). From the 136,830 probes, we then selected 475 probes methylated in 20 breast cancer cell lines (β ≥ 0.8; ≥ 8/20 cell lines) and 27 cancer tissues (β ≥ 0.3; ≥ 21/27 tissues). We further selected 39 probes from 10 genomic regions (9 genes) that had multiple (≥ 3) flanking probes with consistent values [20, 21]. Alternatively, from the 136,830 probes, we isolated 93 probes more frequently methylated both in cancer cell lines (≥ 16/20) and cancer tissues (≥ 21/27). We further selected 12 probes from 6 genomic regions (3 genes) that had multiple (≥ 2) flanking probes with consistent values. Collectively, a total of 12 genes was isolated as candidate marker genes (Supplementary Table 3).

For the 12 genes, we further evaluated copy number alterations (CNAs) in breast cancer because CNAs could affect the estimation of cancer cell fraction based on the DNA methylation levels [5, 22] (Supplementary Table 3). After the exclusion of one gene with a high incidence of CNAs in breast cancer (≥ 3%) [23], the remaining 11 genes were considered to have minimum influence of CNAs on the estimation of the cancer cell fraction.

Selection of a panel of genes with broad coverage across patients

To identify candidate marker genes methylated in different groups of patients, we

conducted a hierarchical clustering analysis of the 27 breast cancer tissues used for the screening and 48 probes in the 11 candidate marker genes (Supplementary Table 3). The samples were separated into three major clusters, and the probes into four second-level clusters (I, II, III, and IV clusters) (Figure 2). From each of the four second-level clusters, we searched for genes that had broad coverage across the samples and for which high-quality primers for bisulfite pyrosequencing could be designed. Consequently, we selected four genes, namely, *SYCN*, *MIR129-2*, *SIM1*, and *CCDC181* (Table 1 and Figure 3), with high-quality primers (Supplementary Table 4). The four genes covered different groups of patients and collectively had different coverage of patients and breast cancer cell lines (Supplementary Figure 1).

Methylation levels in breast cancer biopsy specimens

To evaluate whether the four candidate marker genes could estimate the breast cancer cell fraction, methylation levels of the four genes were analyzed by bisulfite pyrosequencing in 10 breast cancer biopsy specimens purified by LCM (Supplementary Figure 2). *SYCN* was methylated (≥ 20%), even in LCM-purified non-cancerous cells, too frequently (BC53s, BC57s, BC59s, BC60s, and BC61s), and was excluded from the candidate marker genes. Among the remaining three genes, *MIR129-2* showed consistently lower methylation levels than *SIM1* and *CCDC181* in LCM-purified cancer cells, and was excluded. Resultantly, we adopted the remaining two genes, *SIM1* and *CCDC181*, as final candidate marker genes (Figure 4A).

The methylation levels of the two genes were then analyzed in an additional 51 specimens without LCM purification. Substantial methylation levels (≥ 20%) of at least one of *SIM1* and *CCDC181* were observed in 50 specimens (98.0%) (Figure 4B).

Therefore, when we adopted a higher methylation level of the two genes, their combination was considered to be capable of estimating breast cancer cell fractions with broad coverage across patients.

Correlation between the cancer cell fraction estimated by the final candidate marker genes and that estimated by microscopic examination

To assess how accurately the cancer cell fraction could be estimated by the two marker genes, we evaluated the correlation between the cancer cell fractions estimated by the two genes and that estimated by microscopic examination in the 61 breast cancer biopsy specimens, including the 10 specimens with LCM purification and the 51 specimens without LCM purification (Figure 5). We obtained a significant correlation between the cancer cell fractions estimated by the two methods ($R = 0.48$, $p < 0.0001$). Therefore, the combination of the two genes was considered as a marker that could estimate breast cancer cell fractions.

Application of the cancer cell fraction marker to the correction of *HSD17B4* methylation levels

Finally, we evaluated how the cancer cell fraction marker could correct *HSD17B4* methylation levels by estimating the cancer cell fraction and improve the sensitivity and specificity of *HSD17B4* methylation. For this purpose, we used the 61 breast cancer biopsy specimens in which the pCR was observed in 22 specimens (36.1%). Based upon the raw methylation data, no significant difference of the *HSD17B4* methylation levels was observed between pCR and non-pCR specimens ($p = 0.245$) (Figure 6). In contrast, after the correction, the methylation level was significantly higher in the pCR specimens

than in the non-pCR specimens (microscopic examination: p = 0.0001; fraction marker: p = 0.0004). Regarding the sensitivity and specificity to predict pCR (Table 2), it was 13.6 % and 94.9 %, respectively, before the correction. Those after the correction by the DNA methylation marker (59.1 % and 84.6 %) were equivalent to those corrected by microscopic examination (59.1% and 87.2%).

Discussion

We successfully established a DNA methylation marker using two genes, *SIM1* and *CCDC181*, which could estimate the breast cancer cell fraction in DNA samples. The cancer cell fraction estimated by the DNA methylation marker was significantly correlated with that estimated by microscopic examination. In addition, the performance of the *HSD17B4* methylation to predict pCR was improved after the correction of the cancer cell fraction by the fraction marker to the same degree by the correction using microscopic examination. These findings demonstrated that the DNA methylation marker could be applied to correct the cancer cell fraction in breast cancer.

The estimation of cancer cell fractions using a DNA methylation marker has several advantages. Firstly, DNA methylation can be analyzed using DNA samples without the need for histological sections. Secondly, the DNA methylation marker can not only save pathologists' labor in microscopic cell counting but also improve the quality of estimation of the cancer cell fraction. In histological analysis, only the first or last section of a paraffin-embedded tissue block is stained and used for microscopic cell counting. However, for the middle sections, the cells are not counted, and it results in an unavoidable error in the estimation of the cancer cell fraction. Thirdly, quantitative methylation analysis is more cost-effective compared with single nucleotide polymorphism microarray or next-generation sequencing, which are other molecular technologies to estimate cancer cell fractions [24, 25].

Unexpectedly, we observed that the fraction marker genes had low-level methylation even in the LCM-purified non-cancerous cells. This methylation might be caused by contaminating cancer cells. However, its possibility was considered to be low because the

low-level methylation was observed too frequently (nine of 10 specimens) as a contamination. Alternatively, the low-level methylation could be due to accumulation of aberrant methylation in normal appearing cells, predisposing them to carcinogenesis (i.e., field cancerization) [26, 27]. It is well established in other cancers, such as gastrointestinal cancers, that non-cancerous cells can have aberrant DNA methylation and that the degree of aberrant DNA methylation is correlated with a cancer risk [28-30]. To use the DNA methylation marker established here, we should note a risk of overestimation of the cancer cell fraction in low methylation ranges.

In conclusion, we established a DNA methylation marker to estimate breast cancer cell fractions in DNA samples. We expect that this marker will be useful in many aspects of molecular analyses of breast cancers.

Acknowledgements

The authors are grateful to Drs. K. Ichimura, Y. Matsushita, and M. Kitahara of Division of Brain Tumor Translational Research in the National Cancer Center Research Institute for their technical assistance with the usage of the PSQ 96 Pyrosequencing System.

Funding

This research was supported by the Program for Promoting Platform of Genomics based Drug Discovery (grant number; 18kk0305004h0003) from the Japan Agency for Medical Research and Development, AMED.

Compliance with ethical standards

Conflicts of interest

The authors state no conflicts of interest regarding this work.

Ethical approval

Written informed consent was obtained from all participants.

References

1. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nature reviews Genetics. 2010;11(10):685-96. doi:10.1038/nrg2841.
2. Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics (Oxford, England). 2012;28(1):40-7. doi:10.1093/bioinformatics/btr593.
3. Roma C, Esposito C, Rachiglio AM, Pasquale R, Iannaccone A, Chicchinelli N et al. Detection of EGFR mutations by TaqMan mutation detection assays powered by competitive allele-specific TaqMan PCR technology. BioMed research international. 2013;2013:385087. doi:10.1155/2013/385087.
4. Yau C, Mouradov D, Jorissen RN, Colella S, Mirza G, Steers G et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome biology. 2010;11(9):R92. doi:10.1186/gb-2010-11-9-r92.
5. Takahashi T, Matsuda Y, Yamashita S, Hattori N, Kushima R, Lee YC et al. Estimation of the fraction of cancer cells in a tumor DNA sample using DNA methylation. PloS one. 2013;8(12):e82302. doi:10.1371/journal.pone.0082302.
6. Zong L, Hattori N, Yoda Y, Yamashita S, Takeshima H, Takahashi T et al. Establishment of a DNA methylation marker to evaluate cancer cell fraction in gastric cancer. Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association. 2016;19(2):361-9. doi:10.1007/s10120-015-0475-2.
7. Heller G, Babinsky VN, Ziegler B, Weinzierl M, Noll C, Altenberger C et al. Genome-wide CpG island methylation analyses in non-small cell lung cancer patients. Carcinogenesis. 2013;34(3):513-21. doi:10.1093/carcin/bgs363.
8. Shen J, Wang S, Zhang YJ, Wu HC, Kibriya MG, Jasmine F et al. Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. Epigenetics. 2013;8(1):34-43. doi:10.4161/epi.23062.
9. Wu Y, Davison J, Qu X, Morrissey C, Storer B, Brown L et al. Methylation profiling identified novel differentially methylated markers including OPCML and FLRT2 in prostate cancer. Epigenetics. 2016;11(4):247-58.

doi:10.1080/15592294.2016.1148867.

10. Harada T, Yamamoto E, Yamano HO, Nojima M, Maruyama R, Kumegawa K et al. Analysis of DNA methylation in bowel lavage fluid for detection of colorectal cancer. Cancer prevention research (Philadelphia, Pa). 2014;7(10):1002-10. doi:10.1158/1940-6207.Capr-14-0162.
11. Okamoto Y, Sawaki A, Ito S, Nishida T, Takahashi T, Toyota M et al. Aberrant DNA methylation associated with aggressiveness of gastrointestinal stromal tumour. Gut. 2012;61(3):392-401. doi:10.1136/gut.2011.241034.
12. Yan PS, Venkataramu C, Ibrahim A, Liu JC, Shen RZ, Diaz NM et al. Mapping geographic zones of cancer risk with epigenetic biomarkers in normal breast tissue. Clinical cancer research : an official journal of the American Association for Cancer Research. 2006;12(22):6626-36. doi:10.1158/1078-0432.Ccr-06-0467.
13. Stirzaker C, Zotenko E, Song JZ, Qu W, Nair SS, Locke WJ et al. Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value. Nature communications. 2015;6:5899. doi:10.1038/ncomms6899.
14. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England journal of medicine. 2004;351(27):2817-26. doi:10.1056/NEJMoa041588.
15. Mook S, Schmidt MK, Viale G, Pruneri G, Eekhout I, Floore A et al. The 70-gene prognosis-signature predicts disease outcome in breast cancer patients with 1-3 positive lymph nodes in an independent validation study. Breast cancer research and treatment. 2009;116(2):295-302. doi:10.1007/s10549-008-0130-2.
16. Fujii S, Yamashita S, Yamaguchi T, Takahashi M, Hozumi Y, Ushijima T et al. Pathological complete response of HER2-positive breast cancer to trastuzumab and chemotherapy can be predicted by HSD17B4 methylation. Oncotarget. 2017;8(12):19039-48. doi:10.18632/oncotarget.15118.
17. Shigematsu Y, Niwa T, Yamashita S, Taniguchi H, Kushima R, Katai H et al. Identification of a DNA methylation marker that detects the presence of lymph node metastases of gastric cancers. Oncology letters. 2012;4(2):268-74. doi:10.3892/ol.2012.708.
18. Yamashita S, Takahashi S, McDonell N, Watanabe N, Niwa T, Hosoya K et al. Methylation silencing of transforming growth factor-beta receptor type II in rat prostate cancers. Cancer research. 2008;68(7):2112-21. doi:10.1158/0008-5472.can-07-5282.

19. Yoshida T, Yamashita S, Takamura-Enya T, Niwa T, Ando T, Enomoto S et al. Alu and Satalpha hypomethylation in Helicobacter pylori-infected gastric mucosae. International journal of cancer. 2011;128(1):33-9. doi:10.1002/ijc.25534.
20. Takahashi T, Yamahsita S, Matsuda Y, Kishino T, Nakajima T, Kushima R et al. ZNF695 methylation predicts a response of esophageal squamous cell carcinoma to definitive chemoradiotherapy. Journal of cancer research and clinical oncology. 2015;141(3):453-63. doi:10.1007/s00432-014-1841-x.
21. Gyobu K, Yamashita S, Matsuda Y, Igaki H, Niwa T, Oka D et al. Identification and validation of DNA methylation markers to predict lymph node metastasis of esophageal squamous cell carcinomas. Annals of surgical oncology. 2011;18(4):1185-94. doi:10.1245/s10434-010-1393-5.
22. Robinson MD, Stirzaker C, Statham AL, Coolen MW, Song JZ, Nair SS et al. Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation. Genome research. 2010;20(12):1719-29. doi:10.1101/gr.110601.110.
23. Jung SH, Lee A, Yim SH, Hu HJ, Choe C, Chung YJ. Simultaneous copy number gains of NUPR1 and ERBB2 predicting poor prognosis in early-stage breast cancer. BMC cancer. 2012;12:382. doi:10.1186/1471-2407-12-382.
24. Song S, Nones K, Miller D, Harliwong I, Kassahn KS, Pinese M et al. qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. PloS one. 2012;7(9):e45835. doi:10.1371/journal.pone.0045835.
25. Su X, Zhang L, Zhang J, Meric-Bernstam F, Weinstein JN. PurityEst: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics (Oxford, England). 2012;28(17):2265-6. doi:10.1093/bioinformatics/bts365.
26. Curtius K, Wright NA, Graham TA. An evolutionary perspective on field cancerization. Nature reviews Cancer. 2018;18(1):19-32. doi:10.1038/nrc.2017.102.
27. Cheng AS, Culhane AC, Chan MW, Venkataramu CR, Ehrich M, Nasir A et al. Epithelial progeny of estrogen-exposed breast progenitor cells display a cancer-like methylome. Cancer research. 2008;68(6):1786-96. doi:10.1158/0008-5472.Can-07-5547.
28. Baba Y, Ishimoto T, Kurashige J, Iwatsuki M, Sakamoto Y, Yoshida N et al. Epigenetic field cancerization in gastrointestinal cancers. Cancer letters. 2016;375(2):360-6. doi:10.1016/j.canlet.2016.03.009.

29. Asada K, Nakajima T, Shimazu T, Yamamichi N, Maekita T, Yokoi C et al. Demonstration of the usefulness of epigenetic cancer risk prediction by a multicentre prospective cohort study. Gut. 2015;64(3):388-96. doi:10.1136/gutjnl-2014-307094.
30. Asada K, Ando T, Niwa T, Nanjo S, Watanabe N, Okochi-Takada E et al. FHL1 on chromosome X is a single-hit gastrointestinal tumor-suppressor gene and contributes to the formation of an epigenetic field defect. Oncogene. 2013;32(17):2140-9. doi:10.1038/onc.2012.228.

Figure legends

Figure 1. Isolation flow of genomic regions specifically methylated in breast cancer cells using genome-wide DNA methylation data

The genome-wide DNA methylation data were obtained from i) our own analysis of HMECs, peripheral leucocyte samples, and breast cancer cell lines, and ii) the TCGA database of breast cancer tissues and non-cancerous mammary tissues. Probes unmethylated in non-cancerous mammary cells were isolated using the criteria in the Figure, and then those methylated in breast cancer cells were isolated using three-consecutive- and two-consecutive-probe approaches. Isolated candidate probes were assembled into genes, and 12 genes were finally isolated as candidate marker genes.

Figure 2. Selection of a panel of marker genes with a broad coverage

A hierarchical clustering analysis of the 27 breast cancer tissues used for the screening was conducted using 48 probes in the 11 candidate marker genes. From the four second-level clusters (I, II, III and IV; shown by bars on the left side), four genes with high-quality primers for bisulfite pyrosequencing, *SYCN* (probe ID: cg02863073), *MIR129-2* (cg14416371), *SIM1* (cg27252696), and *CCDC181* (cg24808280), were selected for further analysis.

Figure 3. Genomic structure of the four candidate marker genes

Gene structure and location of a CpG island are shown at the top. A CpG map around the target CpG sites is shown in the bottom. Vertical lines show individual CpG sites. Arrows show the locations of probes in the microarray. A triangle shows the CpG site analyzed

by bisulfite pyrosequencing.

Figure 4. Methylation levels of *SIM1* and *CCDC181* in breast cancer biopsy specimens

Methylation levels of *SIM1* and *CCDC181* were analyzed by bisulfite pyrosequencing. (A) The analysis of 10 breast cancer biopsy specimens with LCM purification showed that at least one of the two genes was specifically methylated in LCM-purified cancer cells. (B) The analysis of an additional 51 specimens without LCM purification showed that substantial methylation levels (≥ 20%) of at least one of the two genes were observed in 50 of 51 [98.0%] specimens (except for BC24).

Figure 5. Correlation between the cancer cell fraction estimated by the *SIM1* and *CCDC181* and that estimated by microscopic examination

There was a significant correlation between the cancer cell fraction estimated by the two genes (*SIM1* and *CCDC181*) and that estimated by microscopic examination ($R = 0.48$, $p < 0.0001$).

Figure 6. Predictive performance of the *HSD17B4* methylation after the correction of cancer cell fraction

The *HSD17B4* methylation level was corrected by the breast cancer cell fraction estimated by microscopic examination and by the fraction marker. *HSD17B4* methylation was significantly higher in pCR specimens compared to that in non-pCR specimens (microscopic examination: $p = 0.0001$; fraction marker: $p = 0.0004$).

Supplementary Figure legends

Supplementary Figure 1. The coverage of breast cancer tissues and cell lines by the four candidate marker genes.

The four candidate marker genes, *SYCN*, *MIR129-2*, *SIM1*, and *CCDC181*, had broad and different coverage groups of (A) breast cancer tissues and (B) cancer cell lines. Red and blue cells show samples with β values ≥ 0.3 and ≥ 0.8, respectively. A black cell shows a gene without signals. Samples and cell lines from HER2-positive breast cancer are shown by asterisks.

Supplementary Figure 2. Individual methylation levels of the four candidate marker genes in 10 LCM-purified breast cancer biopsy specimens

Individual methylation levels of the four candidate genes, *SYCN*, *MIR129-2*, *SIM1*, and *CCDC181*, in 10 LCM-purified breast cancer biopsy specimens are shown. *SYCN* was excluded from the candidates because of its high frequency of methylation (≥ 20%) even in LCM-purified non-cancerous cells (BC53s, BC57s, BC59s, BC60s, and BC61s). *MIR129-2* was also excluded because its methylation levels were consistently lower than those of *SIM1* and *CCDC181* in LCM-purified cancer cells. Resultantly, the remaining two genes, *SIM1* and *CCDC181*, were adopted as final candidate marker genes.

Table 1. Candidate genomic regions for a breast cancer cell fraction marker

No.	Gene symbol	Chr	Nt number	Probe ID	Relation to a CpG island	Position to a TSS	No. of consecutive probes	Incidence of methylation in cancer cell lines	Incidence of methylation in cancer tissues
1	*SYCN*	19	39204191	cg02863073	Island	76	2	16/20	21/27
2	*MIR129-2*	11	43581297	cg14416371	Island	24860;2407;1801;-84	3	12/20	23/27
3	*SIM1*	6	100465064	cg27252696	Island	-1386;-134;-174	3	12/20	21/27
4	*CCDC181*	1	169427630	cg24808280	Island	-155;-212;-167;-155;33040	3	13/20	24/27

Genomic location was based upon human genome assembly hg38. Chr, chromosome; Nt, nucleotide; TSS, transcriptional start site

Table 2. Predictive performance of HSD17B4 methylation before and after the correction by microscopic examination and by the methylation fraction marker

	HSD17B4 **methylation**	**# of samples with pCR**	**# of samples with non-pCR**	**Sensitivity (%)**	**Specificity (%)**	**Positive predictive value (%)**
No correction	High	3	2	13.6	94.9	60
	Low	19	37			
Correted by						
microscopic examination	High	13	5	59.1	87.2	72.2
	Low	9	34			
methylation fraction marker	High	13	6	59.1	84.6	68.4
	Low	9	33			

HSD17B4 methylation levels were divided into high and low using a cutoff value of 50 % previously established [16]. pCR, pathological complete response

Supplementary Table 1. Patient background of breast cancer biopsy specimens

Patient ID	Esterogen receptor status	Progesteron receptor status	HER2 receptor status	Pathological diagnosis	Clinical stage	LCM-purification	Therapeutic response to trastuzumab	In the previous study*
BC01	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	pCR	Used
BC02	Negative	Negative	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC03	Positive	Positive	Positive	Invasive lobular carcinoma	IIB	Non-purified	Non-pCR	Used
BC04	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC05	Positive	Negative	Positive	Invasive ductal carcinoma	IIIC	Non-purified	Non-pCR	Used
BC06	Negative	Negative	Positive	Invasive ductal carcinoma	IIIA	Non-purified	pCR	Not used
BC07	Negative	Negative	Positive	Invasive ductal carcinoma	IIA	Non-purified	pCR	Used
BC08	Negative	Negative	Positive	Invasive ductal carcinoma	IIA	Non-purified	pCR	Used
BC09	Positive	Positive	Positive	Invasive ductal carcinoma	IIIA	Non-purified	Non-pCR	Used
BC10	Negative	Positive	Positive	Invasive ductal carcinoma	IIB	Non-purified	pCR	Used
BC11	Positive	Negative	Positive	Invasive ductal carcinoma	IIIA	Non-purified	Non-pCR	Used
BC12	Negative	Negative	Positive	Invasive ductal carcinoma	IIIA	Non-purified	pCR	Used
BC13	Negative	Negative	Positive	Invasive ductal carcinoma	IIIC	Non-purified	pCR	Used
BC14	Positive	Positive	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC15	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Non-purified	pCR	Used
BC16	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC17	Negative	Negative	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC18	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC19	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC20	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	pCR	Used
BC21	Negative	Negative	Positive	Invasive ductal carcinoma	IIA	Non-purified	pCR	Used
BC22	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Not used
BC23	Negative	Negative	Positive	Apocrine carcinoma	IIA	Non-purified	pCR	Used
BC24	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC25	Positive	Positive	Positive	Invasive lobular carcinoma	IIA	Non-purified	Non-pCR	Used
BC26	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC27	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC28	Positive	Positive	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC29	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC30	Positive	Positive	Positive	Invasive ductal carcinoma	IIIC	Non-purified	Non-pCR	Used
BC31	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC32	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC33	Negative	Negative	Positive	Invasive ductal carcinoma	IIIC	Non-purified	pCR	Used
BC34	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC35	Positive	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC36	Positive	Positive	Positive	Invasive ductal carcinoma	IIB	Non-purified	pCR	Used
BC37	Positive	Positive	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC38	Positive	Positive	Positive	Invasive ductal carcinoma	IIIC	Non-purified	Non-pCR	Used
BC39	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC40	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	Non-pCR	Used
BC41	Negative	Negative	Positive	Invasive ductal carcinoma	IIIA	Non-purified	Non-pCR	Used
BC42	Negative	Negative	Positive	Medullary carcinoma	IIA	Non-purified	Non-pCR	Used
BC43	Positive	Negative	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC44	Negative	Negative	Positive	Invasive lobular carcinoma	IIIA	Non-purified	pCR	Used
BC45	Positive	Negative	Positive	Invasive ductal carcinoma	IIIC	Non-purified	pCR	Used
BC46	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC47	Positive	Positive	Positive	Invasive ductal carcinoma	IIIA	Non-purified	Non-pCR	Used
BC48	Negative	Negative	Positive	Invasive ductal carcinoma	IIIA	Non-purified	Non-pCR	Used
BC49	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Non-purified	Non-pCR	Used
BC50	Negative	Negative	Positive	Invasive ductal carcinoma	IIA	Non-purified	pCR	Used
BC51	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Non-purified	pCR	Used
BC52	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Purified	pCR	Used
BC53	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Purified	Non-pCR	Used
BC54	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Purified	Non-pCR	Used
BC55	Negative	Negative	Positive	Invasive ductal carcinoma	IIIA	Purified	Non-pCR	Used
BC56	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Purified	pCR	Used
BC57	Negative	Negative	Positive	Invasive ductal carcinoma	IIB	Purified	pCR	Used
BC58	Negative	Negative	Positive	Invasive ductal carcinoma	IIA	Purified	pCR	Used
BC59	Positive	Positive	Positive	Invasive ductal carcinoma	IIB	Purified	Non-pCR	Used
BC60	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Purified	pCR	Not used
BC61	Positive	Positive	Positive	Invasive ductal carcinoma	IIA	Purified	Non-pCR	Used

* Used in the previous study with reference number [16]. LCM, laser capture microdissection

Supplementary Table 2. Data of breast cancer tissues and non-cancerous mammary tissues downloaded from the TCGA database

TCGA biospecimen ID	Sample	Sample ID	Esterogen receptor status	Progesteron receptor status	HER2 receptor status	Pathological diagnosis	Pathological Stage	Age
TCGA-AR-A1AM	Cancer tissue	TBC01	Positive	Positive	Negative	Infiltrating Lobular Carcinoma	IIIA	52
TCGA-OL-A6VQ	Cancer tissue	TBC02	Positive	Positive	Negative	Infiltrating Lobular Carcinoma	IIA	49
TCGA-A2-A1G1	Cancer tissue	TBC03	Negative	Negative	Positive	Infiltrating Ductal Carcinoma	IIB	85
TCGA-AR-A24X	Cancer tissue	TBC04	Positive	Positive	Negative	Mixed Histology	IIA	52
TCGA-D8-A1Y3	Cancer tissue	TBC05	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIIA	61
TCGA-BH-A0H7	Cancer tissue	TBC06	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIIA	65
TCGA-A7-A13F	Cancer tissue	TBC07	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIIA	44
TCGA-E2-A1IN	Cancer tissue	TBC08	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	I	60
TCGA-BH-A0HX	Cancer tissue	TBC09	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIB	54
TCGA-E9-A5FL	Cancer tissue	TBC10	Negative	Negative	Negative	Metaplastic Carcinoma	IIB	65
TCGA-AR-A24T	Cancer tissue	TBC11	Positive	Positive	Negative	Infiltrating Lobular Carcinoma	IIIC	46
TCGA-BH-A1EN	Cancer tissue	TBC12	Negative	Negative	Positive	Other specify	IIA	78
TCGA-BH-A0AU	Cancer tissue	TBC13	Positive	Positive	Positive	Infiltrating Ductal Carcinoma	IIA	45
TCGA-A2-A0CR	Cancer tissue	TBC14	Positive	Positive	Negative	Infiltrating Lobular Carcinoma	IIB	54
TCGA-BH-A0DI	Cancer tissue	TBC15	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIB	63
TCGA-A2-A0YD	Cancer tissue	TBC16	Positive	Positive	Negative	Infiltrating Lobular Carcinoma	IIB	63
TCGA-A1-A0SQ	Cancer tissue	TBC17	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIB	45
TCGA-EW-A1J6	Cancer tissue	TBC18	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	I	70
TCGA-S3-AA14	Cancer tissue	TBC19	Positive	Positive	Positive	Infiltrating Ductal Carcinoma	I	47
TCGA-A7-A26G	Cancer tissue	TBC20	Negative	Negative	Negative	Other specify	IIA	50
TCGA-A2-A3XY	Cancer tissue	TBC21	Negative	Negative	Negative	Infiltrating Ductal Carcinoma	IIB	49
TCGA-A7-A4SE	Cancer tissue	TBC22	Negative	Negative	Negative	Infiltrating Ductal Carcinoma	IIA	54
TCGA-BH-A0AW	Cancer tissue	TBC23	Positive	Negative	Positive	Infiltrating Ductal Carcinoma	IIA	56
TCGA-D8-A1JC	Cancer tissue	TBC24	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIIA	59
TCGA-E2-A1IJ	Cancer tissue	TBC25	Positive	Positive	Negative	Infiltrating Lobular Carcinoma	I	57
TCGA-D8-A1XF	Cancer tissue	TBC26	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIA	45
TCGA-EW-A2FR	Cancer tissue	TBC27	Negative	Negative	Positive	Infiltrating Ductal Carcinoma	IIIC	59
TCGA-BH-A1FB	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIB	60
TCGA-E2-A15K	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIA	58
TCGA-BH-A0DV	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIIA	54
TCGA-BH-A0AY	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIA	62
TCGA-BH-A0AZ	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIIA	47
TCGA-BH-A0BV	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIB	78
TCGA-BH-A0BA	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Mixed Histology	IIIC	51
TCGA-BH-A1ES	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIB	35
TCGA-E9-A1RD	Non-cancerous mammary tissue	Not appricable	Unknown	Unknown	Unknown	Infiltrating Ductal Carcinoma	IIA	67
TCGA-E9-A1NA	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Positive	Mixed Histology	IIA	58
TCGA-BH-A1FN	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Unknown	Infiltrating Ductal Carcinoma	IIA	34
TCGA-E2-A1B5	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Infiltrating Lobular Carcinoma	IIA	46
TCGA-AC-A2FB	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Positive	Infiltrating Lobular Carcinoma	IIA	65
TCGA-BH-A0DH	Non-cancerous mammary tissue	Not appricable	Positive	Positive	Negative	Infiltrating Ductal Carcinoma	IIB	63
TCGA-E9-A1RD	Non-cancerous mammary tissue	Not appricable	Unknown	Unknown	Unknown	Infiltrating Ductal Carcinoma	IIA	67

Supplementary Table 3. Twelve genes from 16 genomic regions specifically methylated in breast cancer cells

No.	Gene symbol	Chr	Nt number	Probe ID	Relation to a CpG island	Position to a TSS	No. of consecutive probe	Incidence of methylation in cancer cell lines	Incidence of methylation in cancer tissues	Copy number alterations*
1	*C1orf50*	1	42785402	cg27232866	Island	18109;18095	2	17	22	None
			42785582	cg19254906	S_Shore	18289;18275	2	16	21	
1a	-	1	63319650	cg15617155	Island	-	2	17	26	None
			63319873	cg02283366	Island	-	2	17	22	
2	*SYCN*	19	39204069	cg22290648	Island	198	2	16	22	None
			39204191	cg02863073	Island	76	2	16	22	
2a	-	19	43699597	cg08669447	Island	-	2	16	24	None
			43699761	cg09489306	Island	-	2	16	23	
2b	-	21	36693224	cg00495860	Island	-	2	18	26	None
			36693747	cg10445315	Island	-	2	19	22	
3	*NKX2-6*	8	23706412	cg14428146	Island	187;-730	2	17	21	None
			23706457	cg15854847	Island	142;-685	2	16	21	
4	*CCDC181*	1	169427397	cg00100121	Island	78;21;66;78;33273	3	13	23	None
			169427399	cg13958426	Island	76;19;64;76;33271	3	12	23	
			169427468	cg00002719	Island	7;-50;-5;7;33202	3	12	21	
			169427474	cg08104202	Island	1;-56;-11;1;33196	3	12	22	
			169427547	cg23818870	Island	-72;-129;-84;-72;33123	3	11	24	
			169427596	cg16998150	Island	-121;-178;-133;-121;33074	3	9	23	
			169427620	cg08047907	Island	-145;-202;-157;-145;33050	3	11	24	
			169427630	cg24808280	Island	-155;-212;-167;-155;33040	3	13	24	
5	*HIST3H2BA*	1	228464777	cg26911220	Island	-150	3	14	24	Gain
			228464827	cg07726139	Island	-200	3	14	26	
			228464880	cg13799227	S_Shore	-253	3	11	22	
6	*MIR129-2*	11	43581295	cg15556502	Island	24858;2405;1799;-86	3	12	21	None
			43581297	cg14416371	Island	24860;2407;1801;-84	3	12	24	
			43581307	cg14944647	Island	24870;2417;1811;-74	3	12	23	
			43581329	cg01939477	Island	24892;2439;1833;-52	3	12	23	
			43581364	cg16407471	Island	24927;2474;1868;-17	3	11	23	
			43581370	cg05376374	Island	24933;2480;1874;-11	3	11	22	
7	*PHOX2A*	11	72244357	cg05093169	Island	-180;1308	3	13	26	None
			72244395	cg16922279	Island	-218;1270	3	13	22	
			72244503	cg08876932	Island	-326;1162	3	13	22	
			72244555	cg24530250	Island	-378;1110	3	12	23	
8	*RP11-445F12.1*	17	36934600	cg13677415	Island	2053;2059;2062;2023	3	9	23	None
			36934624	cg16364121	Island	2029;2035;2038;1999	3	13	22	
			36934859	cg23402821	Island	1794;1800;1803;1764	3	10	21	
9	*AC079154.1*	2	124024684	cg03696599	Island	490;-604;-604	3	11	24	None
			124024686	cg13358636	Island	488;-602;-602	3	14	24	
			124025009	cg18582824	Island	165;-279;-279	3	13	22	
10	*GYPC*	2	126656532	cg19484420	Island	598;596;347;389;342;398	3	11	21	None
			126656805	cg17848763	S_Shore	871;869;620;662;615;671	3	10	23	
			126656879	cg13901526	S_Shore	945;943;694;736;689;745	3	9	21	
10a	-	3	171028418	cg07139301	Island	-	3	9	22	None
			171028476	cg14777768	Island	-	3	11	25	
			171028502	cg25203962	Island	-	3	11	25	
11	*CDO1*	5	115816723	cg02792792	Island	232;-65;-65	3	10	24	None
			115816734	cg14470895	Island	221;-76;-76	3	10	23	
			115816788	cg23180938	Island	167;-130;-130	3	14	22	
12	*SIM1*	6	100465030	cg21684012	Island	-1352;-100;-140	3	12	22	None
			100465064	cg27252696	Island	-1386;-134;-174	3	12	22	
			100465070	cg17380661	Island	-1392;-140;-180	3	12	22	

Genomic location was based upon human genome assembly hg38.

*According to a study with reference number [23].

Supplementary Table 4. Conditions for bisulfite pyrosequencing

Gene symbol	Primer	Primer sequence	Length (bp)	Annealing Temprature (°C)	Sequencing primer	Sequencing primer sequence	Sequencing to analyze
SYCN	Forward	GGGTTTTAGATTTAGGTTAGGTAGGT	284	54	Forward	TTAGTGTTTTGAGTTTAGGG	YGTTTGTTTYGTTTT
	Reverse	CCCAACAATTCTCATAATAAAAATC-Biotine					
MIR129-2	Forward	GGAGATAGAGGGATAGGATAGGTAG	274	54	Forward	AGGAGTGGTGAGATTGA	GTYGYGATGGAAYGYGTTGGGGAGATTTAG
	Reverse	ACCCTAAAACCAAACAAACTAAATC-Biotine					
SIM1	Forward	Biotine-GGTTTAGAGGGTAGTAAGATTTAGAGTT	334	54	Reverse	ACCAATAAAACTAAATAACA	CRAATCRACCCCRAACC
	Reverse	AACTACCCCCCCTAACTTCTTTATA					
CCDC181	Forward	GAAGAGAGATAGTTATAAGAGGGAAATTTT	453	54	Forward	GGGAAATTTTATAATTAATA	TAGYGGTATTTYGYGAGTTTTTATAA
	Reverse	ACCCTCTATCCCACCATTAACATCT-Biotine					

Genomic location was based upon human genome assembly hg38.

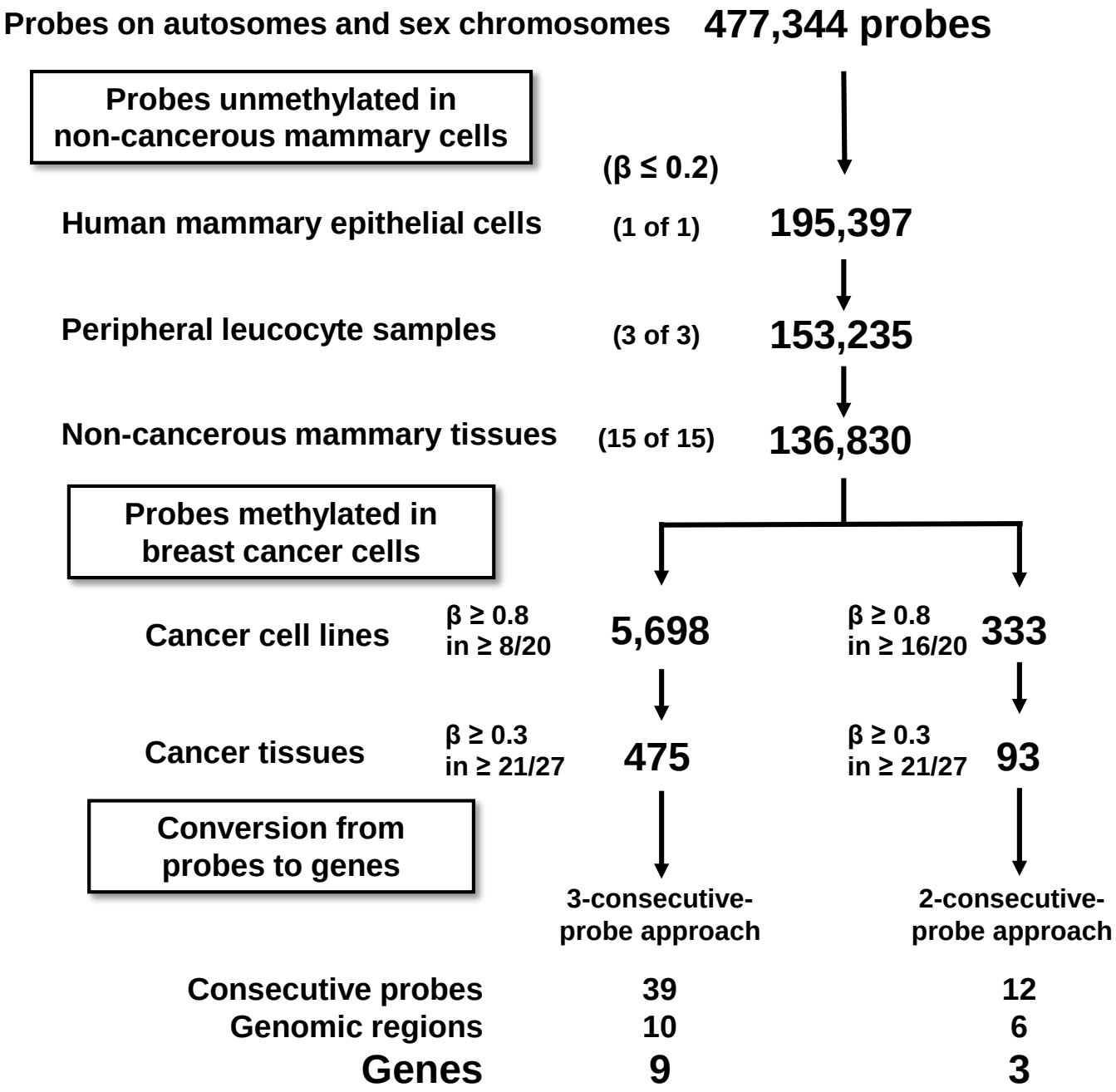

Probes on autosomes and sex chromosomes
477,344 probes
Probes unmethylated in non-cancerous mammary cells
(β ≤ 0.2)
Human mammary epithelial cells
(1 of 1)
195,397
Peripheral leucocyte samples
(3 of 3)
153,235
Non-cancerous mammary tissues
(15 of 15)
136,830
Probes methylated in breast cancer cells
Cancer cell lines
β ≥ 0.8 in ≥ 8/20
5,698
β ≥ 0.8 in ≥ 16/20
333
Cancer tissues
β ≥ 0.3 in ≥ 21/27
475
β ≥ 0.3 in ≥ 21/27
93
Conversion from probes to genes
3-consecutive-probe approach
2-consecutive-probe approach
Consecutive probes
39
12
Genomic regions
10
6
Genes
9
3

Figure 1

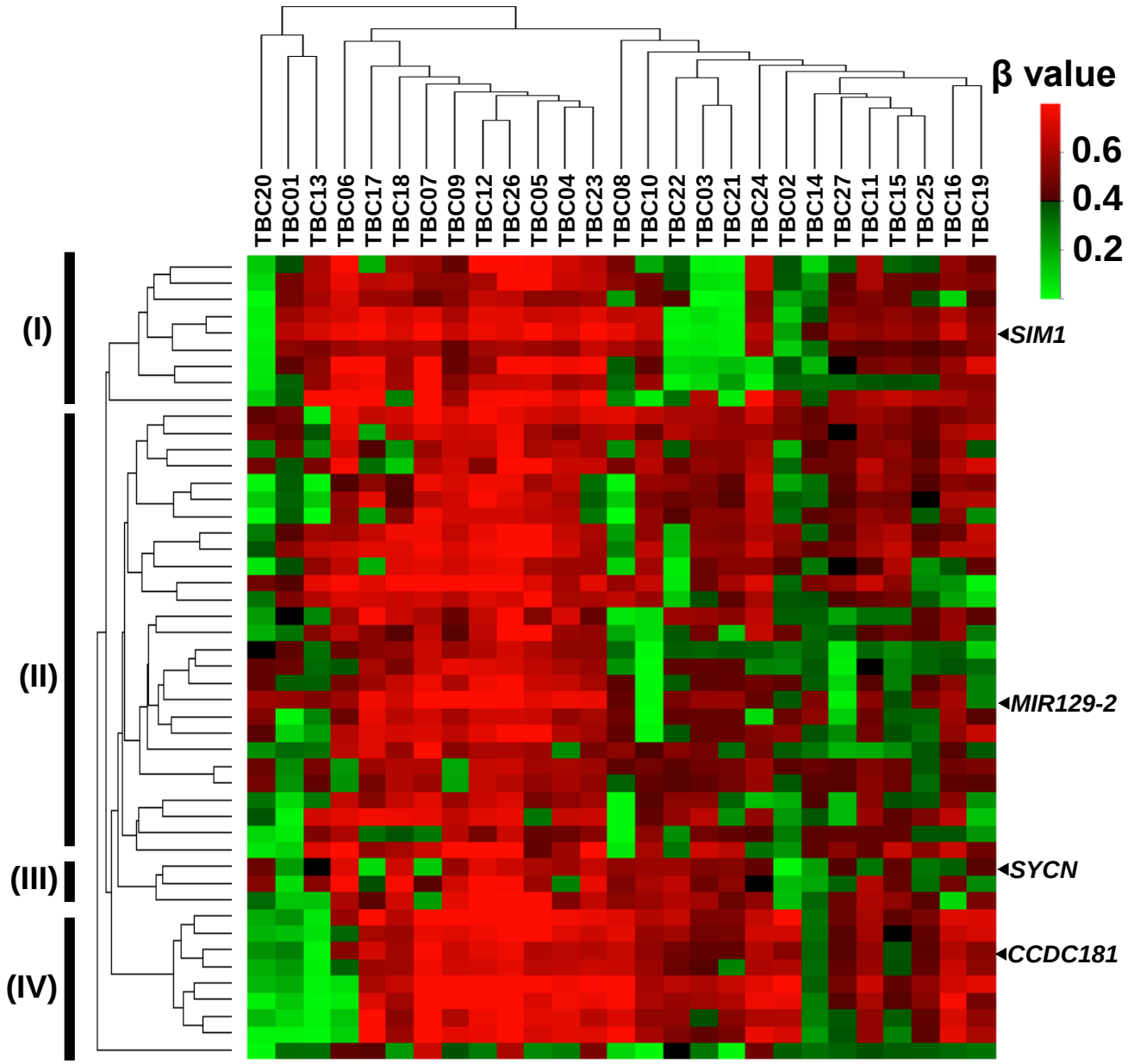

β value
0.6
0.4
0.2
TBC20
TBC01
TBC13
TBC06
TBC17
TBC18
TBC07
TBC09
TBC12
TBC26
TBC05
TBC04
TBC23
TBC08
TBC10
TBC22
TBC03
TBC21
TBC24
TBC02
TBC14
TBC27
TBC11
TBC15
TBC25
TBC16
TBC19
(I)
(II)
(III)
(IV)
SIM1
MIR129-2
SYCN
CCDC181

Figure 2

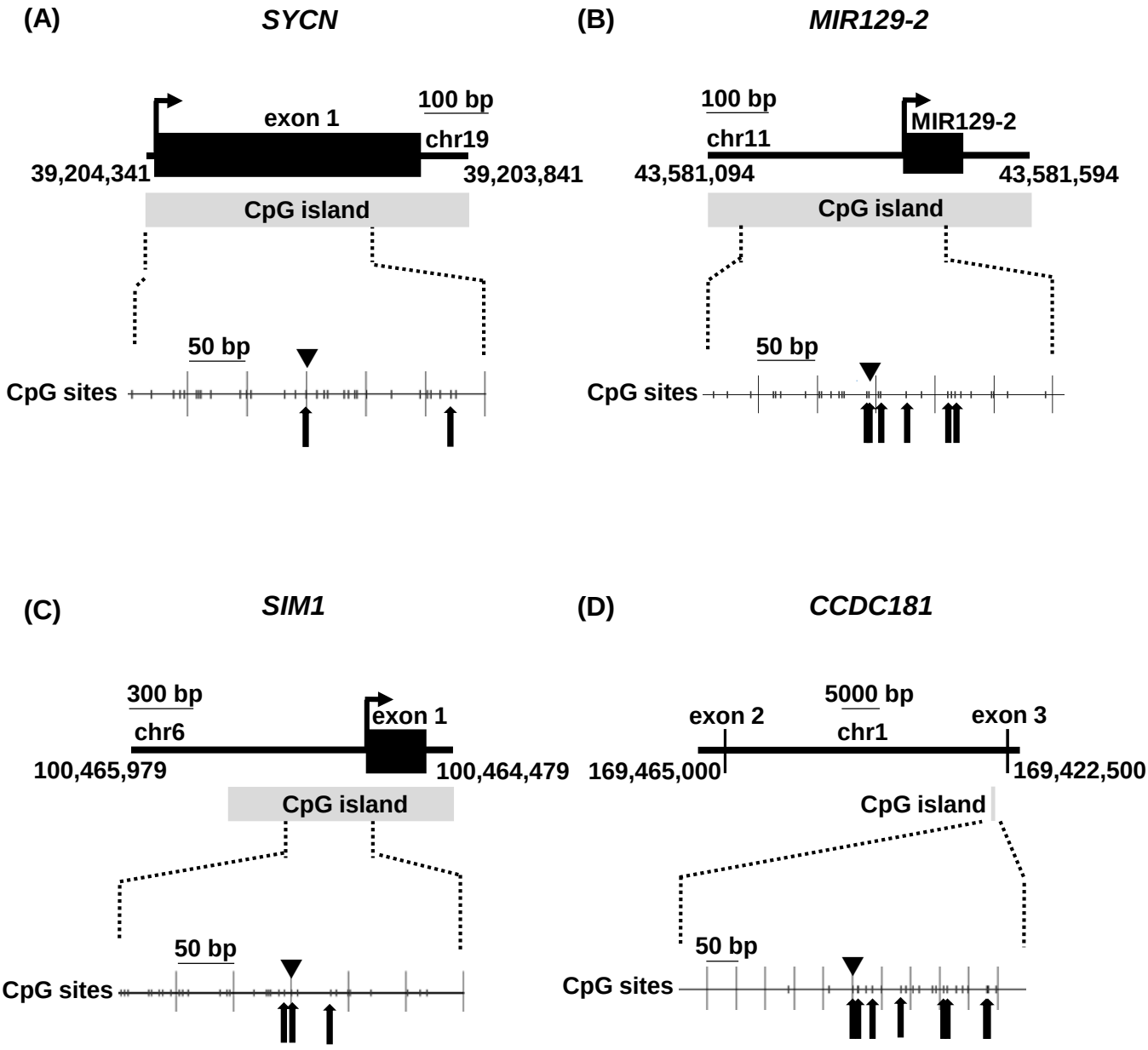
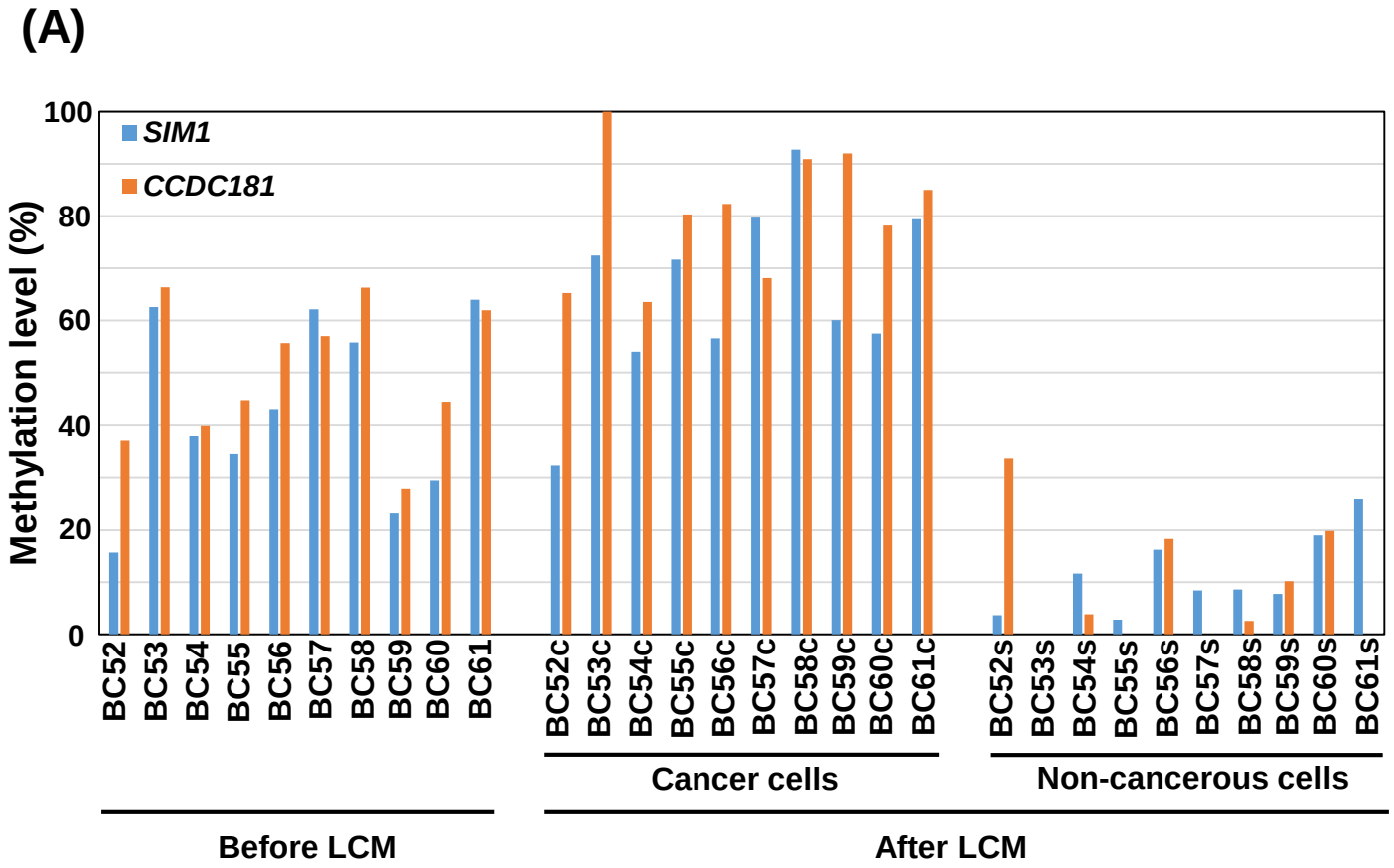

(A)
SYCN
100 bp
exon 1
chr19
39,204,341
39,203,841
CpG island
50 bp
CpG sites
(B)
MIR129-2
100 bp
chr11
MIR129-2
43,581,094
43,581,594
CpG island
50 bp
CpG sites
(C)
SIM1
300 bp
chr6
exon 1
100,465,979
100,464,479
CpG island
50 bp
CpG sites
(D)
CCDC181
5000 bp
exon 2
chr1
exon 3
169,465,000
169,422,500
CpG island
50 bp
CpG sites

Figure 3


(A)
SIM1
CCDC181
Methylation level (%)
100
80
60
40
20
0
BC52
BC53
BC54
BC55
BC56
BC57
BC58
BC59
BC60
BC61
BC52c
BC53c
BC54c
BC55c
BC56c
BC57c
BC58c
BC59c
BC60c
BC61c
BC52s
BC53s
BC54s
BC55s
BC56s
BC57s
BC58s
BC59s
BC60s
BC61s
Cancer cells
Non-cancerous cells
Before LCM
After LCM

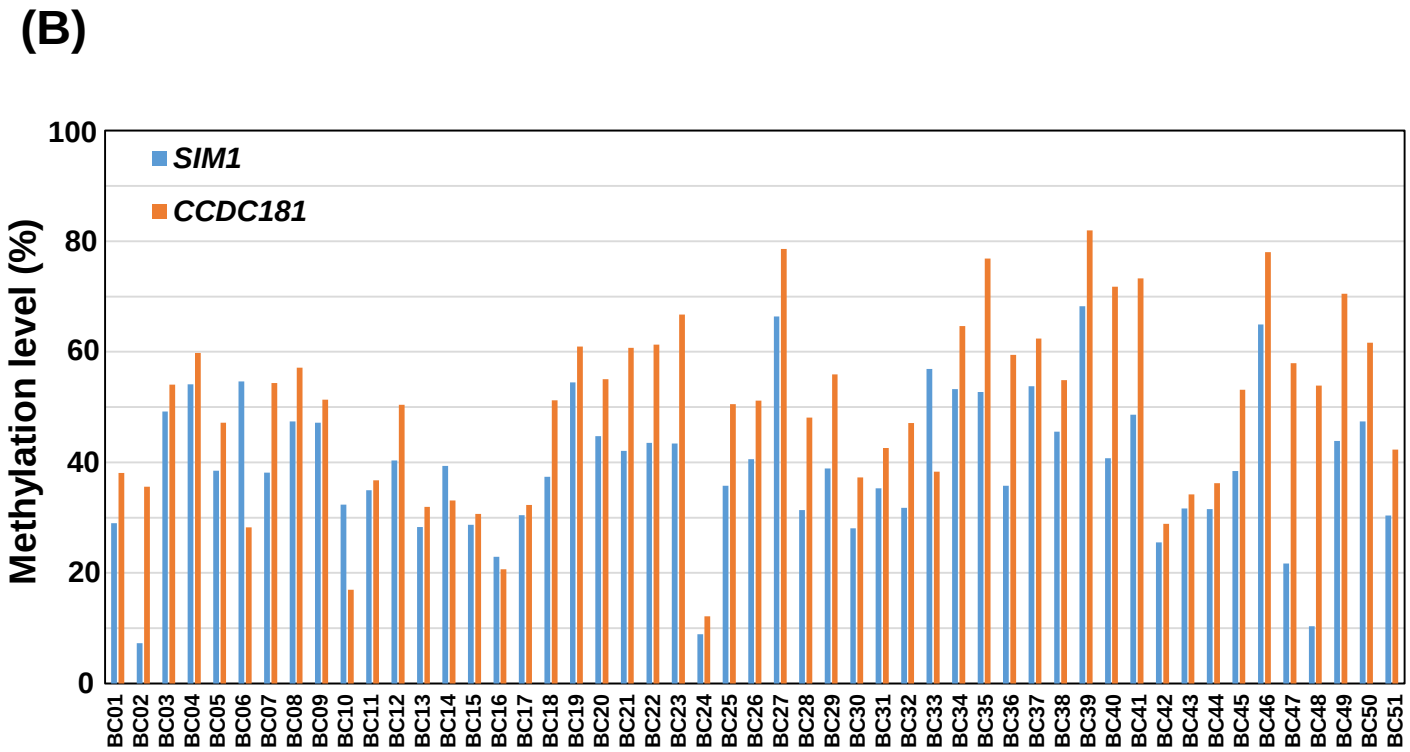

(B)
SIM1
CCDC181
Methylation level (%)
100
80
60
40
20
0
BC01
BC02
BC03
BC04
BC05
BC06
BC07
BC08
BC09
BC10
BC11
BC12
BC13
BC14
BC15
BC16
BC17
BC18
BC19
BC20
BC21
BC22
BC23
BC24
BC25
BC26
BC27
BC28
BC29
BC30
BC31
BC32
BC33
BC34
BC35
BC36
BC37
BC38
BC39
BC40
BC41
BC42
BC43
BC44
BC45
BC46
BC47
BC48
BC49
BC50
BC51

Figure 4

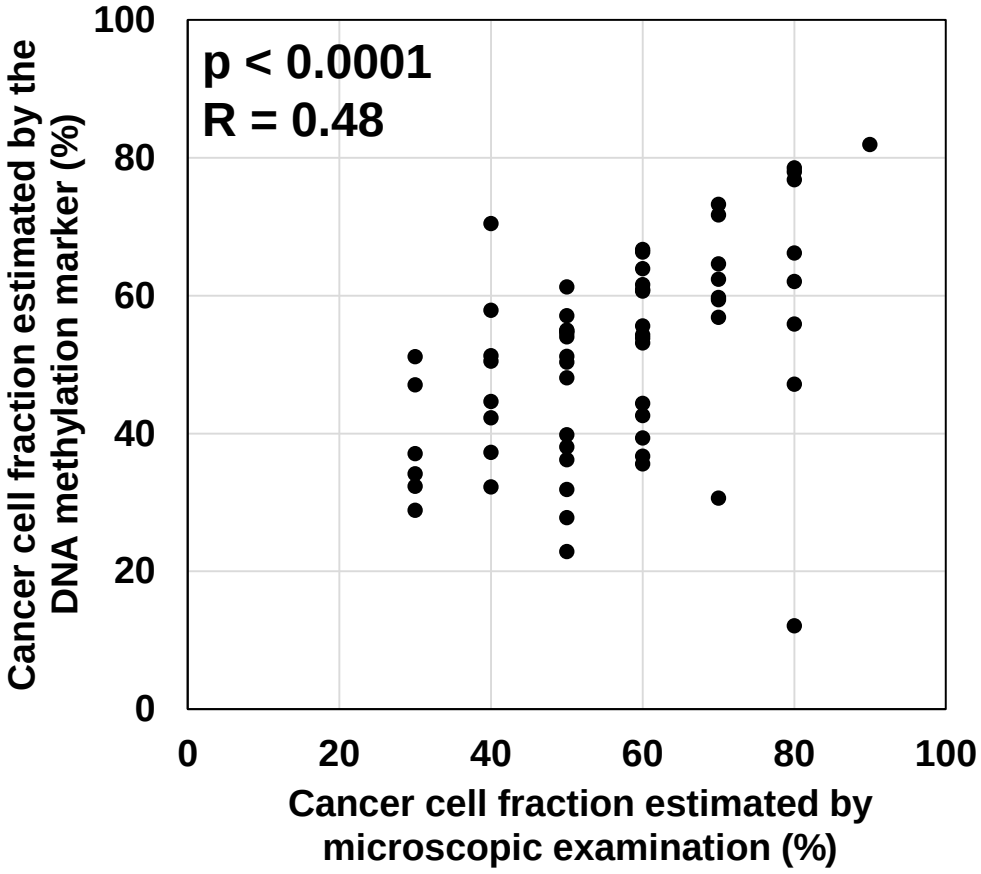

p < 0.0001
R = 0.48
Cancer cell fraction estimated by the DNA methylation marker (%)
Cancer cell fraction estimated by microscopic examination (%)
0
20
40
60
80
100

Figure 5

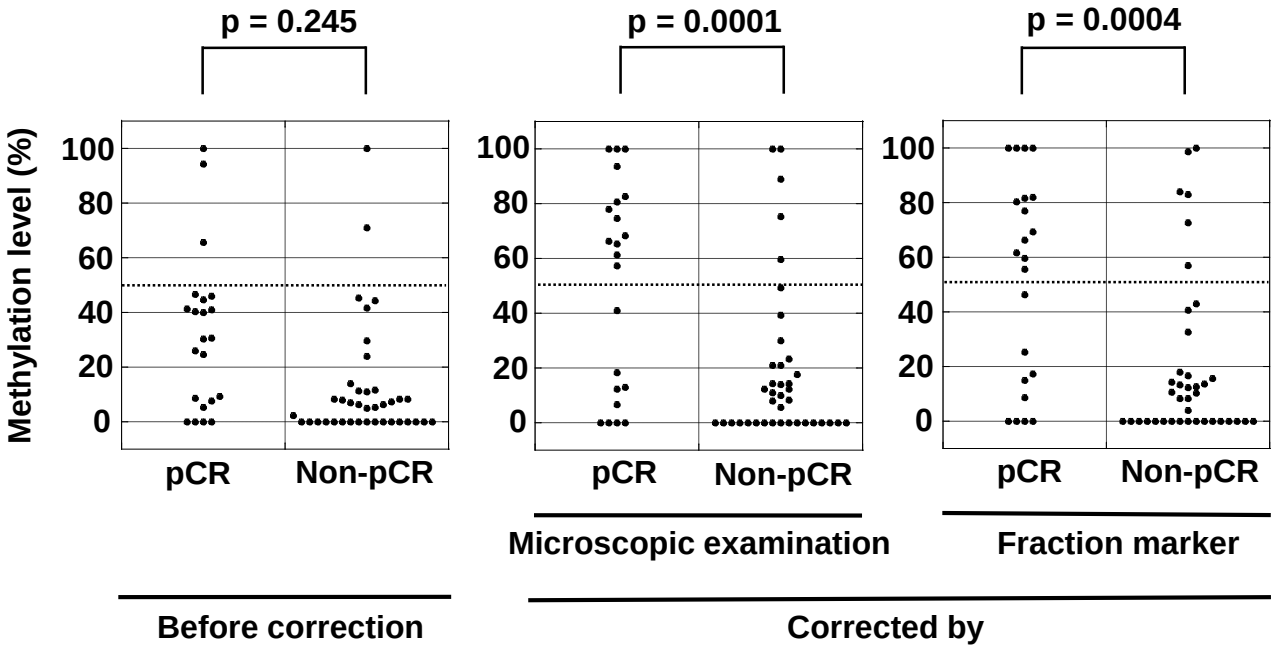
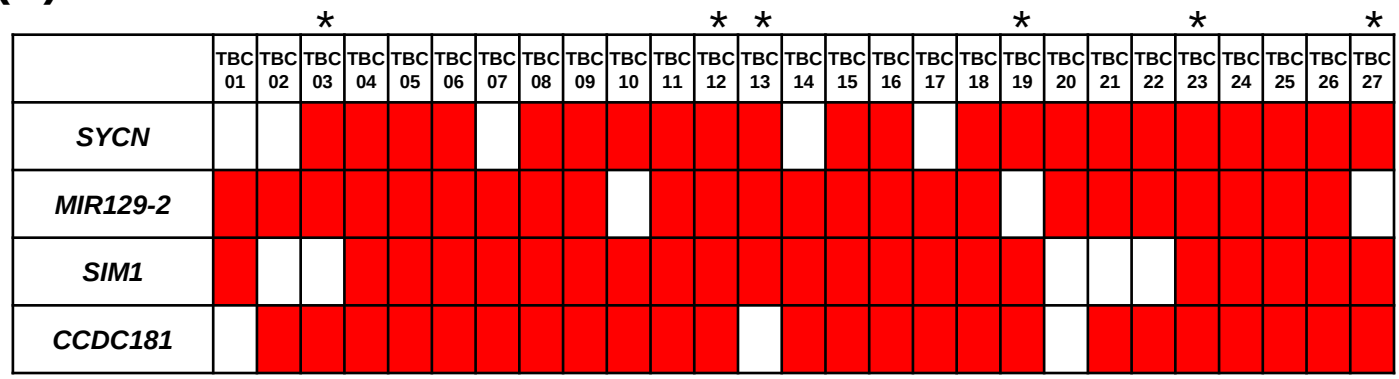
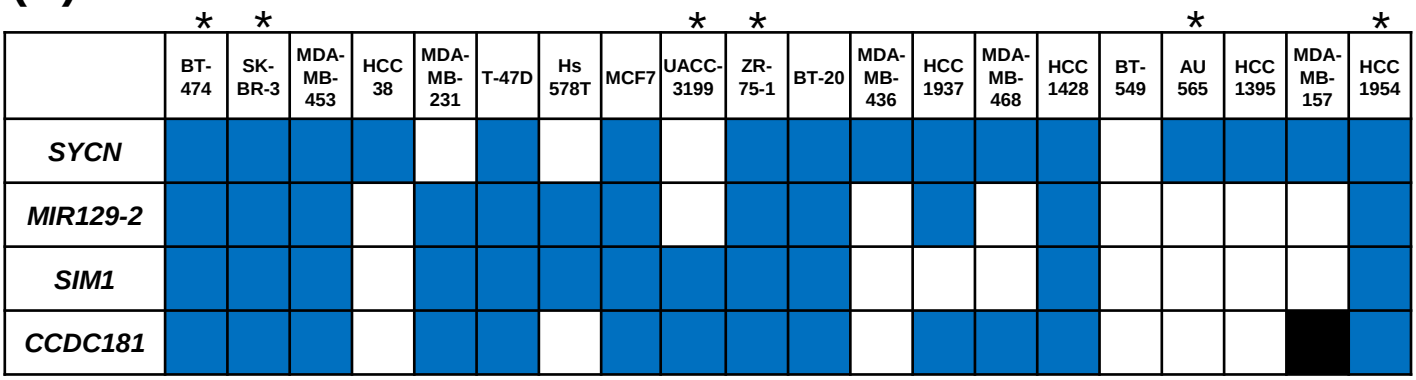

p = 0.245
p = 0.0001
p = 0.0004
Methylation level (%)
100
80
60
40
20
0
pCR
Non-pCR
pCR
Non-pCR
pCR
Non-pCR
Microscopic examination
Fraction marker
Before correction
Corrected by

Figure 6

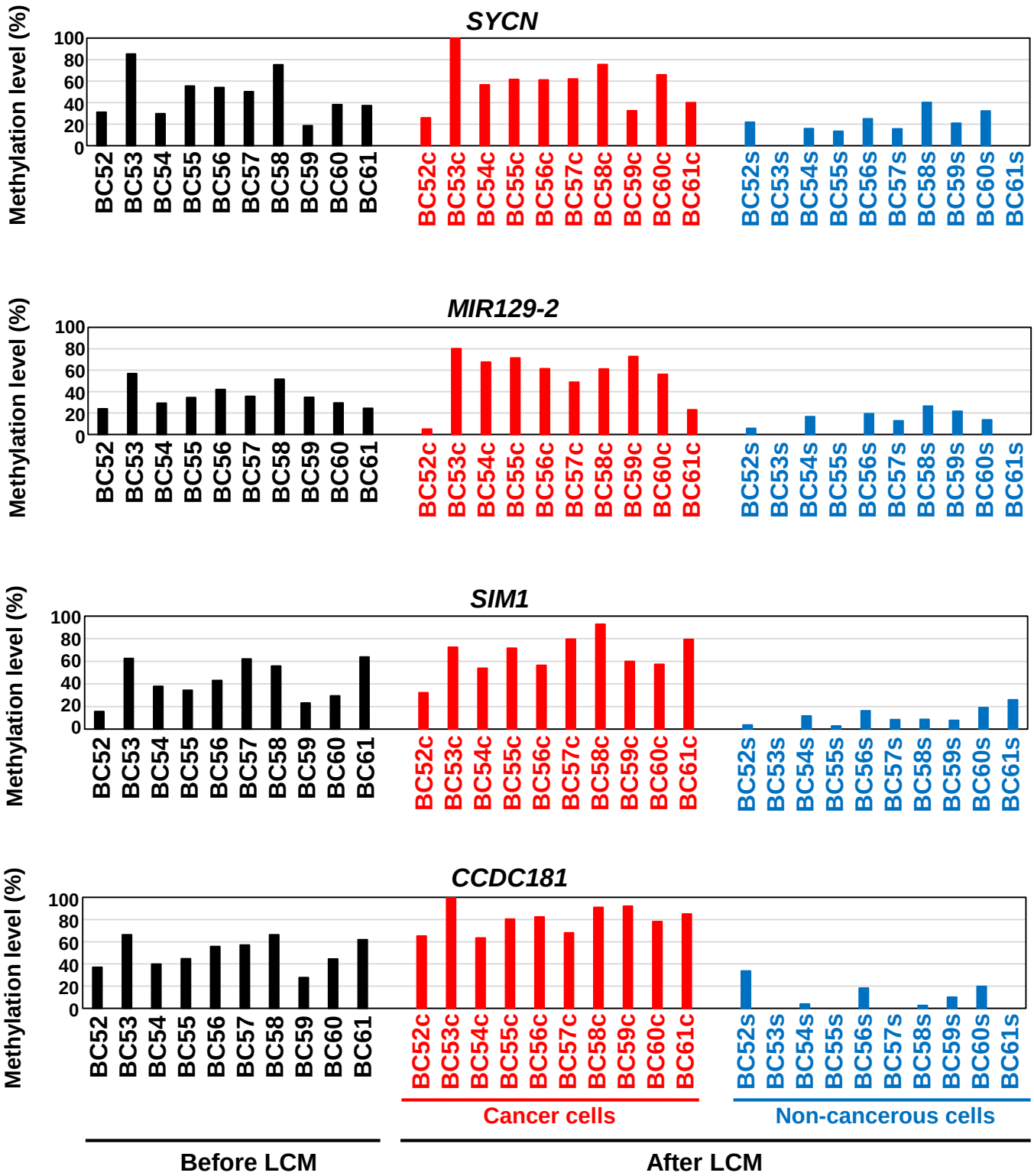
(A)



(B)



Supplementary Figure 1



Supplementary Figure 2