29	Minami-Shimmyo Y, Ohe Y, et al.	Risk Factors for Treatment-Related Death Associated with Chemotherapy and Thoracic Radiotherapy for Lung Cancer	J Thorac Oncol	7(1)	177-182	2012
30	Naito Y, Ohe Y, et al.	Phase II study of nedaplatin and docetaxel in patients with advanced squamous cell carcinoma of the lung	Ann Oncol	22(11)	2471-2475	2011
31	Yano S, Ohe Y, et al.	Hepatocyte Growth Factor Expression in *EGFR* Mutant Lung Cancer with Intrinsic and Acquired Resistance to Tyrosine Kinase Inhibitors in a Japanese Cohort	J Thorac Oncol	6(12)	2011-2017	2011
32	Makino Y, Ohe Y, et al.	Pharmacokinetic and pharmacodynamic study on amrubicin and amrubicinol in Japanese patients with lung cancer	Cancer Chemother Pharmacol	2011 Nov 1. [Epub ahead of print]		
33	Shimada Y, Ohe Y, et al.	Clinical features of unresectable high-grade lung neuroendocrine carcinoma diagnosed using biopsy specimens	Lung Cancer	75(3)	368-373	2012

34	Goto K, Ohe Y, et al.	Epidermal Growth Factor Receptor Mutation Status in Circulating Free DNA in Serum: From IPASS, a Phase III Study of Gefitinib or Carboplatin/Paclitaxel in Non-small Cell Lung Cancer	J Thorac Oncol	7(1)	115-121	2012
35	Nyberg F, Ohe Y, et al.	Proteomic biomarkers for acute interstitial lung disease in gefitinib-treated Japanese lung cancer patients	PLoS One	6(7)	e22062	2011
36	Nyberg F, Ohe Y, et al.	Interstitial lung disease in gefitinib-treated Japanese patients with non-small-cell lung cancer: genome-wide analysis of genetic data	Pharmacogenomics	12(7)	965-975	2011
37	Goto Y, Ohe Y, et al.	Figitumumab combined with carboplatin and paclitaxel in treatment-naïve Japanese patients with advanced non-small cell lung cancer	Invest New Drugs	2011 Jul 13. [Epub ahead of print]		

38	Katori N, Ohe Y, et al.	Genetic variations of orosomucoid genes associated with serum alpha-1-acid glycoprotein level and the pharmacokinetics of paclitaxel in Japanese cancer patients	J Pharm Sci	2011 Jun 2. [Epub ahead of print]		
39	Okamoto I, Ohe Y, et al.	Safety and pharmacokinetic study of nab-paclitaxel plus carboplatin in chemotherapy-naïve patients with advanced non-small cell lung cancer	Invest New Drugs	2011 May 3. [Epub ahead of print]		
40	Niho S, Ohe Y, et al.	Clinical Outcome of Small Cell Lung Cancer with Pericardial Effusion but without Distant Metastasis	J Thorac Oncol	6(4)	796-800	2011
41	Suyama K, Ohe Y, et al.	Development of Cushing's Syndrome During Effective Chemotherapy for Small Cell Lung Cancer	Intern Med	50(4)	335-338	2011
42	Sekine I, Ohe Y, et al.	Phase I Study of Concurrent High-Dose Three-Dimensional Conformal Radiotherapy with Chemotherapy Using Cisplatin and Vinorelbine for Unresectable Stage III Non-Small-Cell Lung Cancer	Int J Radiat Oncol Biol Phys	82(2)	953-959	2012

43	Masuda H, Nakamori S, et al.	Predictive factors for the effectiveness of neoadjuvant chemotherapy and prognosis in triple-negative breast cancer patients	Cancer Chemother Pharmacol	67(4)	911－917	2011
44	Matsubara J, Nakamori S, et al.	Identification of adipophilin as a potential plasma biomarker for colorectal cancer using label-free quantitative mass spectrometry and protein microarray	Cancer Epidemiol Biomarkers Prev	20(10)	2195-2203	2011
45	Yamamoto H, Taniyama K, et al.	OSNA-Based Novel Molecular Testing for Lymph Node Metastases in Colorectal Cancer Patients: Results from a Multicenter Clinical Performance Study in Japan	Surg Oncol	18(7)	1891-1898	2011
46	Yamaji T, Iwasaki M, et al.	Association between Plasma 25-hydroxyvitamin D and Colorectal Adenoma according to Dietary Calcium Intake and Vitamin D Receptor Polymorphism	Am J Epidemiol	175(3)	236-244	2012
47	Takachi R, Iwasaki M, et al.	Red meat intake may increase the risk of colon cancer in Japanese, a population with relatively low red meat consumption	Asia Pac J Clin Nutr	20(4)	603-612	2011

48	Kusano C, Iwasaki M, et al.	Should elderly patients undergo additional surgery after non-curative endoscopic resection for early gastric cancer? Long-term comparative outcomes	Am J Gastroenterol	106(6)	1064-1069	2011
49	Shimazu T, Iwasaki M, et al.	Plasma Isoflavones and the Risk of Lung Cancer in Women: A Nested Case-Control Study in Japan	Cancer Epidemiol Biomarkers Prev	20(3)	419-427	2011
50	Kawano S, Bono H, et al.	Tutorial videos of bioinformatics resources: online distribution trial in Japan named TogoTV	Brief Bioinform	2011 Jul 29. [Epub ahead of print]		
51	Yamamoto Y, Bono H, et al.	Allie: a database and a search service of abbreviations and long forms	Database (Oxford)	2011	bar013	2011
52	Ono H, Bono H, et al.	Gene expression profiling in multipotent DFAT cells derived from mature adipocytes	Biochem Biophys Res Commun	407(3)	562-567	2011

書籍

	著者氏名	論文タイトル名	書籍全体の編集者名	書籍名	出版社名	出版地	出版年	ページ
1	Taniyama K, Morii N, et al.	Topoisomerase II-Alpha Index Predicts the Efficacy of Anthracycline-Based Chemotherapy for Breast Cancers	Williams S. I. et al.	HER2 and Cancer	Nova science publishers	NY	2011	188-200

Breast Cancer (2012) 19:37–45
DOI 10.1007/s12282-011-0263-8

ORIGINAL ARTICLE

Quantitative prediction of tumor response to neoadjuvant chemotherapy in breast cancer: novel marker genes and prediction model using the expression levels

Hiroshi Sano · Satoru Wada · Hidetaka Eguchi · Akihiko Osaki · Toshiaki Saeki · Masahiko Nishiyama

Received: 15 January 2011 / Accepted: 3 March 2011 / Published online: 25 March 2011

Abstract
Background In breast cancer, the identification of accurate predictors of tumor response to neoadjuvant chemotherapy is of key importance, but none of the critical markers have been validated to date. We attempted to identify potent marker genes genome-wide, and we developed a prediction model for individual response to epirubicin (EPI)/cyclophosphamide (CPM) combination chemotherapy (EC).
Methods From 10 human breast cancer cell lines, genes whose expression levels correlated with cytotoxicities of EPI and CPM were chosen through comprehensive gene expression analysis followed by correlation–confirmation study of the quantified expression levels analyzed by real-time reverse transcription polymerase chain reaction (RT-PCR).
Results We finally selected a total of 4 genes (*ANXA1* and *PRKCA* for EPI; *DUSP2* and *SERPINA3* for CPM) as reliable prediction markers. Using quantified expression data of genes in 18 tumor samples, we performed multiple linear regression analysis to establish the best linear model that could convert the quantified expression data to show tumor response to the EC therapy (the ratio of tumor size to the baseline, %). Outliers were identified by referring to the value of AIC (Akaike's information criterion) for each sample (AIC/sample) or checking residuals graphically. The multiple linear regression analysis of the selected genes yielded 2 highly predictive formulae for the tumor response: one used all of the genes except *SERPINA3* ($R = 0.8348$, AIC/sample $= 4.9182$) and the other used all of the 4 genes ($R = 0.8224$, AIC/sample $= 5.0730$).
Conclusions A study to validate the predictive values of the selected 4 genes is now planned, along with research to determine their functional roles.

Keywords Breast cancer · Response prediction · Neoadjuvant chemotherapy · Marker gene

H. Sano · A. Osaki · T. Saeki
Department of Breast Oncology, International Medical Center, Saitama Medical University, 1397-1 Yamane, Hidaka, Saitama 350-1298, Japan

S. Wada · H. Eguchi · M. Nishiyama (✉)
Research Institute for Development Therapeutics, Saitama Medical University, 1397-1 Yamane, Hidaka, Saitama 350-1298, Japan
e-mail: yamacho@saitama-med.ac.jp

Introduction

Neoadjuvant chemotherapy (NAC) is a standard treatment for locally advanced breast cancer, and is also a standard option for patients with primary operable tumors, providing the possibility of increasing rates of breast-conservation surgery and pathological complete response (pCR) [1–4]. Patients receiving neoadjuvant chemotherapy are more likely to undergo breast-conservation surgery, and the recent advent of new-generation agents and the advance of targeted therapy into neoadjuvant therapy offer additional hope for improving the rates [1–4]. However, the survival benefit for such patients has not been validated to date, and the pCR rate remains poor (less than 30%) [5–9]. Furthermore, the response to neoadjuvant chemotherapy varies among individual patients. NAC can give doctors the opportunity to assess the likely outcome in any subsequent adjuvant therapeutic setting, but accurate predictors of response to NAC are still undetermined [10–15]. Some patients, for example, undergo a current regimen with unnecessary toxicity without any standard therapeutic

effect such as downstaging or micrometastasis reduction. These conditions have stimulated research aimed at prior laboratory prediction of individual response to neoadjuvant chemotherapy.

Extensive effort has been directed toward identifying the indicators of host toxicity and treatment efficacy of neoadjuvant chemotherapy, but none of the critical markers have been validated to date [10–15]. Over the years, many biomarkers—including hormone receptors such as estrogen receptors (ER) and progesterone receptors (PgR), ErbB-2/human epidermal growth factor receptor 2 (HER2), and tumor gene expression profiles—have been incorporated in breast cancer management, but most of these have been used mainly for general prognostic assessment and suitability for specific drug therapies [12, 13, 15].

The main obstacle to predicting therapeutic efficacy is the intricate mechanisms of drug sensitivity [16–20]: multifactorial mechanisms limit the prediction of individual drug response using any single marker. Although DNA chip technology enables us to overview a huge number of gene expressions simultaneously, gene expression profiles of drug sensitivity vary considerably even for the same drug. Prediction of a responder for chemotherapy by using "the snapshot expression profile" of microarrays is thus increasingly being recognized as being more challenging than anticipated [17–22].

We therefore attempted to select a set of key marker genes genome-wide using DNA microarrays in vitro, and we developed a prediction system for clinical chemotherapeutic response based on multiple regression analysis using expression data of the selected genes [16–19]. For prediction of response to combination chemotherapy, we used all of the biomarkers selected for each component drug.

The combination of anthracyclines with cyclophosphamide (AC) has been a key regimen in neoadjuvant chemotherapy in breast cancer [1–9]. In this study, we focused on an epirubicin (EPI) plus cyclophosphamide (CPM) combination regimen (EC) and attempted to show powerful prediction marker genes of response as well as a putative prediction model of response to the regimen using the expression data.

Materials and methods

Chemicals

EPI was purchased from Pfizer Pharmaceuticals (New York, USA). 4-Hydroxycyclophosphamide (CPM), an active form for men, was obtained from Shionogi Pharmaceutical Co., Ltd. (Osaka, Japan). All other chemicals were of analytical grade and were purchased from Wako Pure Chemicals (Osaka, Japan) and Nacarai (Kyoto, Japan).

Cells and human tissue samples

The 10 human breast cancer cell lines (BT-20, BT-474, MCF-7, MDA-MB-231, MDA-MB-435S, MDA-MB-453, MDA-MB-468, SK-BR-3, T-47D, and ZR-75-1) were obtained from ATCC (American Type Culture Collection). All cancer cells were cultured in RPMI 1640 containing 10% fetal bovine serum (FBS) and maintained at 37°C in air containing 5% CO_2.

The tumor tissue specimens were collected by needle biopsy from 18 patients in stage II or III (except T0 case) who had pathologically proven breast cancer—5 cases in stage 2A, 10 cases in stage 2B, 1 case each in stage 3A, 3B, and 3C—between October 2008 and December 2009. All of the patients had at least one measurable lesion, and none had received any treatment before tumor sampling. All patients were less than 80 years old (median 51; range 35–67) with performance status 0 to 2, no significant baseline laboratory abnormalities, and life expectancy greater than 3 months. All received epirubicin plus cyclophosphamide combination therapy as preoperative chemotherapy. Both drugs were diluted with 100 mL of 0.9% saline, and epirubicin (90–100 mg/m^2) was administered as a 5-min rapid infusion, followed by a 30-min infusion of cyclophosphamide (600 mg/m^2) on day 1 of each cycle (i.e., every 3 weeks). The patients received 4 cycles of the treatment. Tumor size was measured by ultrasonography in the week preceding treatment, and the measurements were repeated in every chemotherapy cycle to obtain estimates, and response was assessed in accordance with Response Evaluation Criteria in Solid Tumors (RECIST). After the evaluation of response to EC, all of the patients received additional treatment with docetaxel. Written informed consent was obtained from all patients, and the protocol was approved by the institutional ethics committees. The tumor specimens were stored at −80°C before analysis.

Extraction and purification of RNA

For gene expression analysis, frozen tissues were homogenized by a Shake Master NEO (Bio Medical Science, Tokyo, Japan) and exponentially growing cultured cells were collected after washing with phosphate buffered saline (PBS). Total RNA was extracted from tissue homogenates or cell pellets using an RNA Nucleospin RNAII kit (Macherey–Nagel, Duren, Germany) according to the manufacturer's protocols. For microarray analysis, total RNA was checked using an Agilent Technologies 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA). The 2100 Bioanalyzer Expert software program was used

to assign an RNA integrity number (RIN) from 1 to 10, with 1 = poor, 10 = excellent (RNA integrity number). Only RNA samples showing a RIN score greater than 9.0 were used for the further analyses.

Cytotoxic assay

Drug-induced cytotoxicity was evaluated by conventional MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazonium bromide] dye reduction assay. Cells were seeded in 96-MicroWell Plates (Nunclon, Nunc, Roskilde, Denmark) at a density of 4×10^3/well in RPMI 1640 with 10% FBS. After 24-h incubation, the medium was replaced and cells were exposed to the indicated drug concentrations for 72 h, after which 10 μl of 0.4% MTT reagent and 0.1 M sodium succinate were added to each well. After 2 h further incubation, 150 μl of dimethyl sulfoxide (DMSO) was added to dissolve the purple formazan precipitate. The formazan dye was measured spectrophotometrically (570–650 nm) using a MAXline™ microplate reader (Molecular Devices Corp., Sunnyvale, CA) or a ARVO™ MX (Perkin Elmer Inc., MA, USA). The cytotoxic effect of each treatment was assessed by its IC_{50} value (inhibitory drug concentration affording 50% cell growth, i.e., drug concentration affording 50% optical density relative to the control).

DNA microarray analysis

An Agilent 4 × 44 K Whole Human Genome Oligo Microarray (~41,000 transcripts; Agilent Technologies, Tokyo, Japan) was used according to the manufacturer's protocols. Briefly, the first-strand cDNA was generated from 0.5 μg of total RNA using reverse transcriptase and a T7 primer, and then the second-strand cDNA was produced using DNA polymerase mix and RNase H supplied in the Agilent Quick Amp Labeling Kit, One-Color (Agilent Technologies, Santa Clara, USA). cRNA was generated via an in vitro transcription reaction using T7 RNA polymerase, which was quantified by spectrometry and checked using an Agilent Technologies 2100 Bioanalyzer. Then, 1.65 μg of cRNA was fragmented and hybridized to each microarray. After hybridization, the microarrays were rinsed with Agilent Gene Expression Wash Buffer 1 at room temperature and with Buffer 2 at 37°C for 1 min according to the manufacturer's protocol. Finally, the microarrays were scanned using an Agilent DNA Microarray Scanner (Agilent Technologies, Santa Clara, USA), and analyzed with Agilent Feature Extraction software version 9.5. Expression levels were normalized to the 75th percentile expression value of the entire spot using GeneSpring GX (Agilent Technologies, Santa Clara, USA).

The microarray data set was analyzed using the rank products (RP) method via the RankProd package in R version 2.11.1. This method has been shown to be robust in the identification of differentially expressed genes in data sets where there are few replicates and/or large variance. The gene expression data were then further analyzed using Spotfire® software (Tibco Software, CA, USA).

Real-time RT-PCR (reverse transcription polymerase chain reaction)

Total RNA (1 μg) was extracted from each cell line or tumor tissue and converted into cDNA using ReverTra Ace (Toyobo, Osaka, Japan) with oligo $(dT)_{20}$ primer according to the manufacturer's instructions. Primers and Taqman probes for each gene were designed using The Probe Finder software in the Universal Probe Library (UPL) Assay Design Center (Roche Applied Science, Mannheim, Germany). Each reaction was carried out in triplicate using ABI 7900HT Fast Real-Time PCR System (Applied Biosystems). The relative expression levels of each gene were calculated as a ratio to *HPRT1* (hypoxanthine phosphoribosyltransferase 1) expression level.

Development of prediction model using multiple biomarkers

Multiple regression analysis was performed to develop a prediction model of tumor response using multiple biomarkers. The relationship between y (response variable) and $x_{i1}, x_{i2} \ldots x_{ip}$ (explanatory variables) is formulated in the linear model $y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip} + \varepsilon_i$, where θ_0 is a constant and ε_i is an error term following a normal distribution with a mean 0 and variance δ^2, as previously reported [16–19]. Trimmed least squares regression (TLSR) was performed to determine the set of effective genes that would satisfy the value of IC_{50}: $(\theta_0 \ldots \theta_p)$ were estimated from the data $\{y_i; (x_{i1} \ldots x_{ip})\}$ when we used gene expression levels and cellular sensitivity to drugs (IC_{50} value for each drug), as the explanatory and the response variables. TLSR is a robust regression method based on an extended algorithm of least median squares regression (LMSR) by Rousseeuw [23]: it explores models using masked samples with large residuals. We used the NLReg software developed by Ohtaki (http://apollo.rbm.hiroshima-u.ac.jp/), which implemented robust regression analysis [16–19]. Outliers were identified by referring to the value of AIC (Akaike's information criterion) for each sample or checking residuals graphically, and the set of effective genes that satisfied the relative ratio of tumor size to baseline (%) for clinical samples was explored.

Statistical analysis

Mathematical methods to process the microarray data are described above. Other statistical analyses were performed with R, and a comparison of real-time RT-PCR data of drug sensitivity versus drug resistance from cancer cells, and tumor increase versus tumor decrease from cancer tissue specimens, was analyzed using Welch's *t* test which was used to determine the *P* value.

Results

Potent prediction marker genes screened by in vitro comprehensive gene expression analysis

To select candidates for prediction markers, we sorted out the genes which were highly associated with sensitivity to EPI and CPM on the basis of expression levels, using microarray analysis data of 10 breast cancer cell lines.

We first evaluated cellular sensitivity to EPI and CPM by MTT assay in the 10 breast cancer cell lines, divided into drug-sensitive and -resistant groups, according to the 20% trimmed mean value of IC_{50} for each drug: among the 10 cell lines, 5 (BT-20, BT-474, MDA-MB-231, MDA-MB-435S, and SK-BR-3) and 4 cell lines (BT-20, BT-474, MDA-MB-231, and T-47D) were defined to be resistant to EPI and CPM, respectively (Fig. 1). Rank products analysis was then applied to explore differentially expressed genes between the 2 groups. Among 500 top-ranking genes provided by the 1st screening, about 10 genes for each drug were selected as possible candidates through Pearson correlation analysis using the selection criteria of $|R| > 0.5$ and $P < 0.01$. The candidate genes were subjected to real-time RT-PCR analysis in order to confirm any correlation between drug sensitivity and the quantified expression levels, and potent prediction marker genes were determined (Table 1). The selection criterion was determined as $|R| > 0.6$ in Pearson correlation analysis and/or $p < 0.05$ in *t* test for the comparison between sensitive and resistant groups. We eventually selected 2 genes each for EPI, *ANXA1* and *PRKCA*, and 2 genes for CPM, *DUSP2* and *SERPINA3*, as novel prediction markers. The expression levels of *DUSP2* and *SERPINA3* inversely correlated with IC_{50} values for CPM (sensitivity marker), whereas those of *PRKCA* and *ANXA1* positively correlated with IC_{50} values for EPI (resistance marker) (Fig. 2). Interestingly, *DUSP2* was also shown to be closely related to EPI sensitivity ($P = 0.0093$).

Prediction model of clinical response to EC combination chemotherapy in neoadjuvant setting

EPI and CPM appeared to have multiple predictive marker genes for drug sensitivity, and the observed potent predictive value for drug sensitivity suggested that clinical response to EC combination therapy, i.e., tumor response, could be precisely predicted when all of these key genes were used.

All of the 18 patients enrolled in this study were assessed for EC tumor response, in addition to the hormone receptors (ER and PgR) and HER2 status of their tumors (Table 2). We performed real-time RT-PCR analysis of the tumor samples to quantify the expression levels of 4 selected marker genes, and the data were applied to gene expression–clinical response correlation analysis. Although the gene expression levels of *DUSP2* and *SERPINA3* were shown to be significantly higher in partial response (PR) cases than in progressive disease/stable disease (PD/SD) cases (Fig. 3), in the expression levels similarly to the hormonal and HER2 status of the tumors, none of the selected genes alone accurately predicted tumor size after

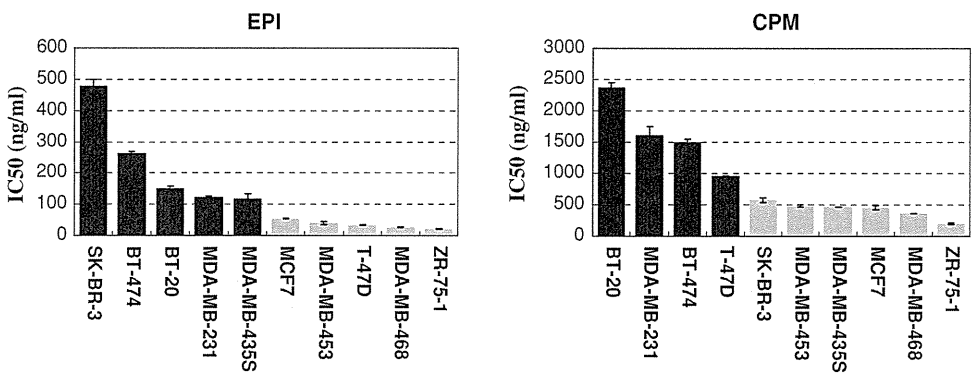

Fig. 1 Evaluation of cellular sensitivity to EPI and CPM by MTT assay. Ten breast cancer cell lines were divided into 2 groups, drug-sensitive or -resistant, using 20% trimmed mean value of IC_{50} for each drug. Five cell lines for EPI (BT-20, BT-474, MDA-MB-231, MDA-MB-435S, and SK-BR-3) and 4 cell lines for CPM (BT-20, BT-474, MDA-MB-231, and T-47D) were categorized in the resistant group and other cell lines were classed in the sensitive group. *EPI* epirubicin, *CPM* cyclophosphamide

Table 1 Correlation between gene expression levels and drug sensitivity

Drug	Gene	Microarray			Real-time RT-PCR		
		Correlation		Sensitive versus resistant	Correlation		Sensitive versus resistant
		R	*P*	*P*	*R*	*P*	*P*
EPI	*KRT19*	−0.6715	0.0335	0.0000	−0.0397	0.9190	0.7404
	SPDEF	−0.6449	0.0441	0.0000	−0.3088	0.4570	0.2291
	CRABP1	−0.6209	0.0554	0.0001	−0.3813	0.2770	0.1664
	LYPD3	−0.5677	0.0870	0.0002	0.0786	0.8290	0.9852
	C10orf116	−0.5664	0.0878	0.0000	−0.0601	0.8690	0.6412
	ST14	−0.5438	0.1040	0.0001	−0.2619	0.4650	0.3146
	C19orf46	−0.5199	0.1230	0.0002	−0.1766	0.6260	0.3949
	CLDN3	−0.5146	0.1280	0.0001	−0.3256	0.3590	0.1628
	PRKCA	**0.5264**	**0.1180**	**0.0002**	0.5667	0.0876	**0.0365**
	ANXA1	**0.5425**	**0.1050**	**0.0000**	**0.7199**	**0.0189**	0.1107
	PDLIM4	0.6752	0.0322	0.0000	0.4811	0.1590	0.1097
CPM	***DUSP2***	**−0.9083**	**0.0003**	**0.0002**	**−0.8580**	**0.0015**	**0.0387**
	SERPINA3	**−0.7718**	**0.0089**	**0.0000**	**−0.6565**	**0.0392**	**0.0186**
	KYNU	−0.7402	0.0144	0.0008	−0.3506	0.3210	0.9191
	PCDHB3	−0.7057	0.0226	0.0009	−0.1766	0.6490	0.7041
	CORO1A	−0.6714	0.0335	0.0009	−0.2065	0.6240	0.8854
	VAMP5	−0.6208	0.0554	0.0004	−0.3181	0.4040	0.3238
	DHRS2	0.6011	0.0660	0.0001	−0.2112	0.5850	0.8039
	ABCC3	0.7555	0.0115	0.0002	−0.1089	0.7800	0.5478

Data for the 4 genes selected as reliable prediction markers are in boldface

EPI epirubicin, *CPM* cyclophosphamide

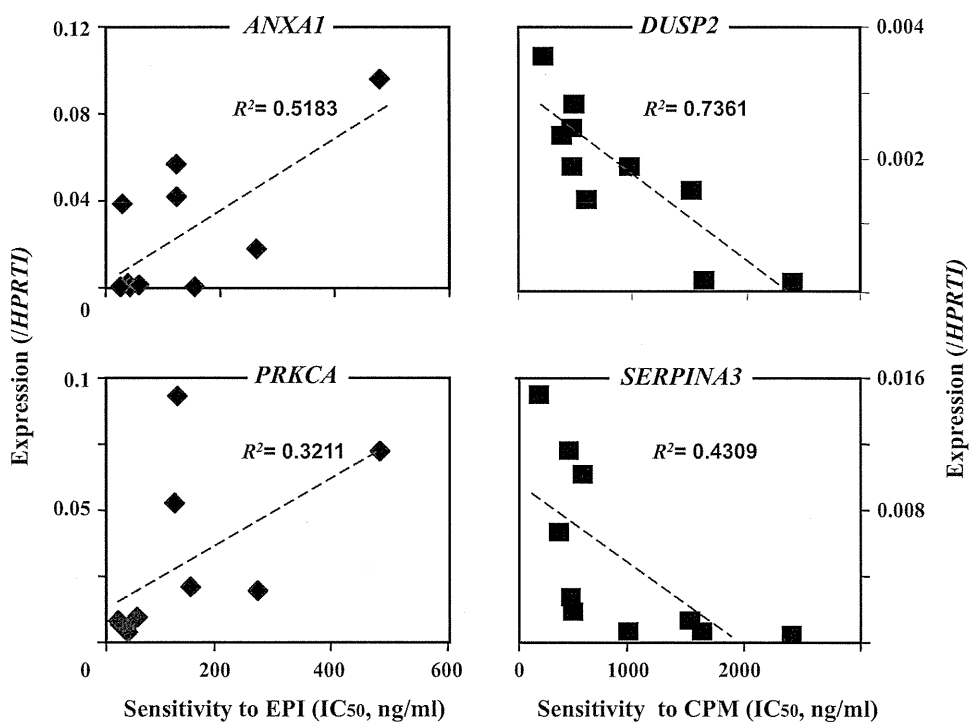

Fig. 2 In vitro correlation between drug sensitivity and expression of 4 genes selected as response predictors. In 10 human breast cancer cell lines, the expression levels of *ANXA1* and *PRKCA* correlated with IC_{50} values for EPI (*filled diamonds*), whereas the expression of *DUSP2* and *SERPINA3* closely related to cellular sensitivity to CPM (*filled squares*)

Table 2 Tumor background and response to EC therapy

Patient no.	Stage	ER	PgR	HER2	Response to EC therapy
1	2A	+	+	−	PR
2	2A	−	−	−	PD
3	2A	+	+	−	PR
4	2A	−	−	−	PR
5	2A	−	−	−	PR
6	2B	−	−	+	PD
7	2B	+	+	−	PR
8	2B	+	−	+	PR
9	2B	−	−	−	PR
10	2B	+	+	−	PR
11	2B	+	−	−	PR
12	2B	+	+	−	PR
13	2B	−	−	+	PR
14	2B	+	+	−	SD
15	2B	+	−	−	PR
16	3A	+	−	−	PR
17	3B	+	+	−	PR
18	3C	+	+	−	PR

EC epirubicin/cyclophosphamide combination, *PR* partial response, *PD* progressive disease, *SD* stable disease

EC therapy (relative ratio of tumor size to baseline, %): *DUSP2*, $R = -0.2820$; *SERPINA3*, $R = -0.2080$; *ANXA1*, $R = 0.0298$; and *PRKCA*, $R = -0.0867$.

In contrast to the findings in the analysis using each of the selected 4 genes alone, analysis using 18 data sets of gene expression and clinical response provided 2 prediction formulae for tumor response that showed the highest fitness for each set of prediction marker genes (Fig. 4). The observed correlation coefficient and Akaike's information criterion per individual sample (AICPS) in the fixed formulae indicated that tumor response to EC could be precisely predicted by these 2 formulae using expression data of at least 3 genes other than *SERPINA3*. We also attempted to fix other prediction formulae using several different sets of marker genes, but none of these predicted response to EC chemotherapy more precisely than the aforementioned formulae.

Discussion

In breast cancer, the identification of accurate predictors of tumor response to neoadjuvant chemotherapy is of key importance to prevent patients from experiencing unnecessary toxicity from ineffective treatments [1]. Several recent studies have demonstrated that various markers, including gene expression profiling, can predict the response, but a clear result is still highly challenging [24, 25]. In this study, using the hypothesis that expression analysis of a set of key drug sensitivity genes for EPI and CPM could allow us to predict therapeutic response to the combination therapy, we proposed 4 genes (*ANXA1* and *PRKCA* for EPI; *DUSP2* and *SERPINA3* for CPM) as novel predictive markers, and constructed 2 prediction formulae using 3 or all 4 of the selected marker genes which could accurately predict clinical tumor response to EC therapy. Obviously, the practical usefulness of our study needs to be evaluated by a larger prospective study, but the indicated advantages in predicting in vitro efficacy of the corresponding drug, and clinical response of the combination

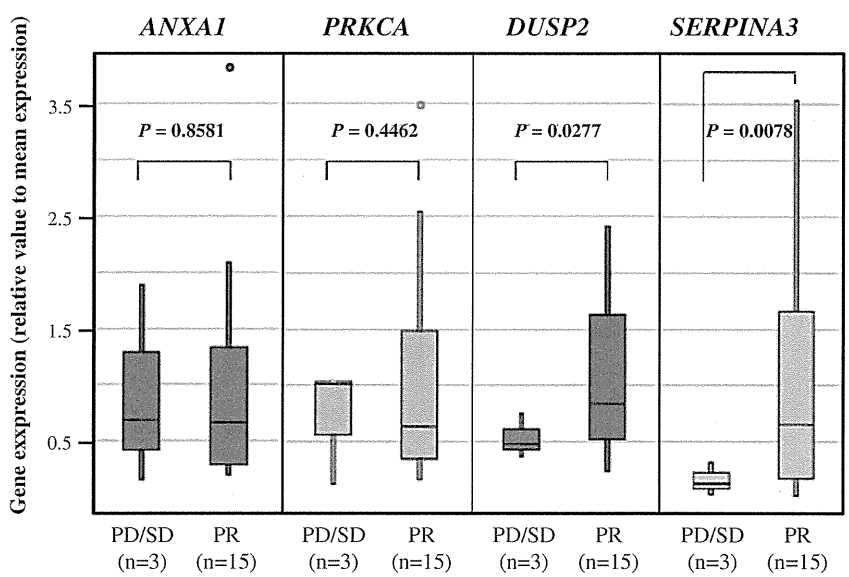

Fig. 3 Clinical response to epirubicin/cyclophosphamide combination (EC) therapy and expression of 4 genes selected as response predictors (box plot analysis). Expression levels of 4 genes in 18 tumor samples were analyzed by real-time RT-PCR and compared with the clinical response to EC therapy evaluated by Response Evaluation Criteria in Solid Tumors (RECIST): *PD* progressive disease, *SD* stable disease, *PR* partial response

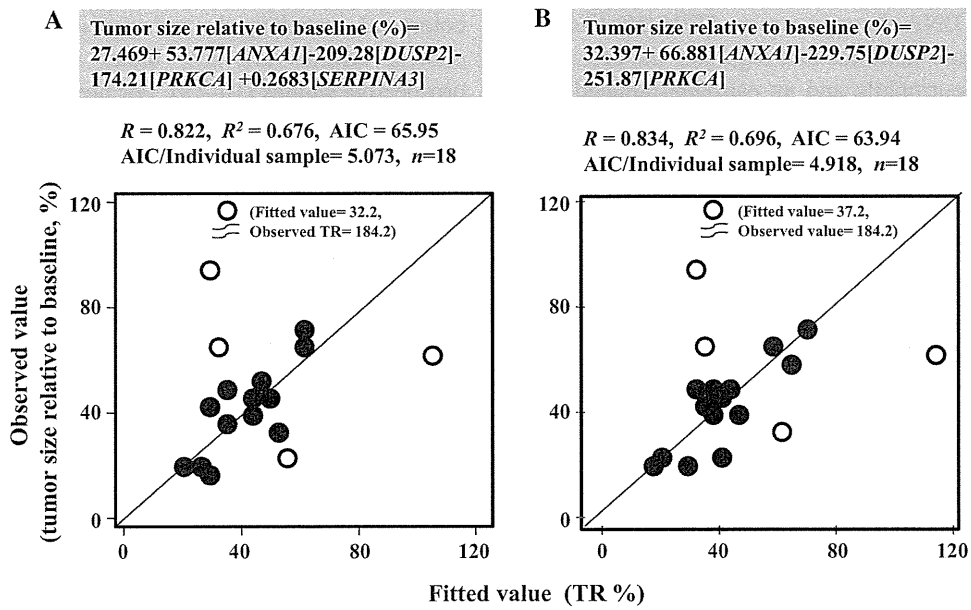

Fig. 4 Predictive fitness of the fixed formulae for therapeutic efficacy of EC therapy. Using multivariate analysis, we developed prediction formulae for therapeutic efficacy of EC therapy and the ratio of tumor size after EC therapy to that before treatment, using the variable expression data of genes selected as response predictors from 18 tumor samples: **a** prediction using 4 genes; **b** prediction using 3 genes. The suitable formulae were fixed by eliminating the outliers using the value of Akaike's information criterion for each sample (AIC/individual sample) or checking residuals graphically. A *closed circle* indicates finally analyzed sample data, whereas an *open circle* indicates a masked outlier. One of the outliers was out of the range of the figure and is indicated by an *open circle underlined with two wavy lines* with the fitted and observed TR values in parenthesis. *R* coefficient of correlation, *AIC* Akaike's information criterion, *TR* tumor size relative to baseline (%)

regimen, suggest that we probably succeeded in selecting powerful candidates for novel prediction marker genes and developing putative prediction models.

The proposed novel 4 genes were sorted out genome-wide as the genes whose expression levels correlated best with corresponding drug sensitivity in vitro, and the ones that would work well in the prediction of clinical response to EC combination chemotherapy. For EPI and CPM, several genes—including *TOP2A*, *ABCB1*, *ABCG2*, *SLC22A16*, *UGT2B7*, *MKI67(Ki67)*, *HER2*, *CYP3A4*, *CYP3A5*, and *CYP2B6*—have been shown to be of predictive benefit [10–12, 14, 15]. Nevertheless, their roles as response predictors for EPI and CPM remain controversial, and our comprehensive gene expression analysis data demonstrated that none of the correlations with drug sensitivity were observed in their expression levels. Although their functions remain little known, our proposed 4 genes are therefore possibly more powerful candidates for prediction markers.

Several reports have found possible roles for these 4 genes in cancer cell biology and interaction with drug action mechanisms: *ANXA1* is a calcium- and phospholipid-binding protein, and recent investigations have demonstrated that negative *ANXA1* expression was significantly associated with the advanced disease stage of breast cancer [26]. *PRKCA*, one of the protein kinase C (PKC) family members, plays an important role in many different cellular processes, such as cell adhesion, cell transformation, cell cycle checkpoint, and cell volume control, and some studies have demonstrated its association with drug resistance in human cancers. *PRKCA*-associated drug resistance is likely mediated by P-gp, which is encoded by the multidrug resistant gene 1 (*ABCB1*) gene [27–29]. *DUSP2* is a member of the dual specificity protein phosphatase subfamily which inactivate their target kinases by dephosphorylating the phosphoserine/threonine and phosphotyrosine residues, and it may participate in the processes critical to the development and progression of human cancer [30]. *DUSP2* acts as a negative regulator of MAP kinase signaling and inhibits extracellular-regulated kinase (ERK) and p38. p38 subfamilies of kinases are activated by stress-related stimuli, including osmotic shock, inhibition of protein synthesis, and formation of oxygen radical species, whereas the ERK subfamily is largely activated by growth factor signals, such as those mediated by receptor tyrosine kinases. Interestingly, this study showed that *DUSP2* could possibly be a response predictor of both EPI and CPM. This might be explained in part by the action on MAP kinase pathways activated by stress-related stimuli. EPI intercalates into DNA and inhibits replication and repair, whereas CPM creates DNA–DNA and DNA–protein interactions

and DNA strand breaks. *DUSP2* may be involved in the response to this DNA damage stress. *SERPINA3* is well known as a protease inhibitor that regulates the activity of cathepsin G in neutrophils and an estrogen-induced gene [31], and its expression was also reported as an indicator of good prognosis in estrogen receptor positive breast cancer [32].

Nevertheless, it will likely be very difficult to predict clinical therapeutic response by using a single marker alone, especially for combination regimens. The potential of finding an in vitro model that precisely reflects clinical response to combination therapy is limited, because individual drug response is driven by complex interactions of molecular pathways rather than one single marker [33]. In fact, none of the selected 4 genes alone accurately predicted clinical response to EC therapy, despite the potent predictive value for each of the component drugs. To overcome these obstacles: (1) we attempted to find a better prediction model using different sets of the selected genes; (2) provided 2 potent models; and (3) showed that tumor response to EC combination might be reliably predictable by using the models. The prediction formulae were developed as the best linear model using multiple linear regression analysis, which embraced the variable expressions of the 3 and 4 component genes and arranged them in order to predict clinical response. Despite the limited number of samples in this study, the indicated advantage in predicting clinical response of the combination regimen does suggest a high potential for the model in practical applications.

These multiple-gene approaches are timely topics in phramacogenomics. Recent studies have increasingly investigated a set of putative biomarkers for each drug used in the combination, based on the hypothesis that knowing key sensitivity markers for each component drug could allow clinicians to predict therapeutic response to the combination therapy [19]. The increasing evidence indicates a prominent role of this approach in various cancers, although these attempts are still in the investigational phase [12, 19, 34–36]. Among them, we believe that our series of attempts are unique because the developed model predicted response to therapy, while providing numerical values of tumor size. Needless to say, expression–sensitivity correlation analyses need to be done in the combination setting in vitro because of the possible synergistic effect. However, the potential of an in vitro sensitivity–evaluation model that precisely reflects clinical response to combination therapy is limited. Even so, our series of pharmacogenomic studies have provided several multi-gene prediction formulae of individual response to anticancer chemotherapies, along with identification of novel marker genes. These attempts are apparently of predictive value in terms of overall and progression-free survival, and/or tumor response in various cancers, including gastric, colorectal, esophageal, and ovarian cancers, and their clinical utilities are now under investigation in larger prospective studies [16–19]. For breast-conserving therapy, tumor size is of key importance, so our prediction model proposed in this study would contribute to personalized medicine for neoadjuvant chemotherapy in breast cancer.

In summary, prediction of tumor response (regression of tumor size) to neoadjuvant chemotherapy is our interest in breast cancer. We attempted to identify possible predictive markers of drug response to EC combination chemotherapy, and found 4 possible marker genes and 2 putative prediction formulae. Although the precise functional mechanisms of the selected genes and their practical significance are still undetermined, the set of novel 3 or 4 genes demonstrated the advantage of predicting tumor response for the EPI/CPM combination. To show the true clinical values, we are now planning a prospective clinical validation study, along with continuing our search for the functional roles of the selected 4 genes in drug sensitivity, and more powerful predictive marker genes for drug sensitivity.

Conflict of interest Toshiaki Saeki received honoraria (such as lecture fees) from Chugai Pharmaceutical Co. Ltd. and research funding from Pfizer Japan Inc. (prediction of chemosensitivity for breast cancer) and from Chugai Pharmaceutical Co. Ltd. (QOL of breast cancer patients).

References

1. Horiguchi J. New trends in primary systemic therapy for breast cancer. Breast Cancer 2010; Epub ahead of print, doi: 10.1007/s12282-010-0243-4.
2. Benson JR, Jatoi I, Keisch M, Esteva FJ, Makris A, Jordan VC. Early breast cancer. Lancet. 2009;373:1463–79.
3. Kinoshita T. Preoperative therapy: recent findings. Breast Cancer 2010; Epub ahead of print doi:10.1007/s12282-010-0227-4.
4. Untch M, von Minckwitz G. Recent advances in systemic therapy: advances in neoadjuvant (primary) systemic therapy with cytotoxic agents. Breast Cancer Res. 2009;11:203.
5. Colleoni M, Viale G, Goldhirsch A. Lessons on responsiveness to adjuvant systemic therapies learned from the neoadjuvant setting. Breast. 2009;18(Suppl 3):S137–40.
6. Beasley GM, Olson JA Jr. What's new in neoadjuvant therapy for breast cancer? Adv Surg. 2010;44:199–228.
7. Gluz O, Liedtke C, Gottschalk N, Pusztai L, Nitz U, Harbeck N. Triple-negative breast cancer—current status and future directions. Ann Oncol. 2009;20:1913–27.
8. Shenoy HG, Peter MB, Masannat YA, Dall BJ, Dodwell D, Horgan K. Practical advice on clinical decision making during neoadjuvant chemotherapy for primary breast cancer. Surg Oncol. 2009;18:65–71.
9. Liu SV, Melstrom L, Yao K, Russell CA, Sener SF. Neoadjuvant therapy for breast cancer. J Surg Oncol. 2010;101:283–91.
10. Munro AF, Cameron DA, Bartlett JM. Targeting anthracyclines in early breast cancer: new candidate predictive biomarkers emerge. Oncogene. 2010;29:5231–40.

11. Bartlett JM, Munro AF, Dunn JA, McConkey C, Jordan S, Twelves CJ, et al. Predictive markers of anthracycline benefit: a prospectively planned analysis of the UK National Epirubicin Adjuvant Trial (NEAT/BR9601). Lancet Oncol. 2010;11:266–74.
12. Marsh S, Liu G. Pharmacokinetics and pharmacogenomics in breast cancer chemotherapy. Adv Drug Deliv Rev. 2009; 61:381–7.
13. Baek HM, Chen JH, Nie K, Yu HJ, Bahri S, Mehta RS, et al. Predicting pathologic response to neoadjuvant chemotherapy in breast cancer by using MR imaging and quantitative 1H MR spectroscopy. Radiology. 2009;251:653–62.
14. Tanioka M, Shimizu C, Yonemori K, Yoshimura K, Tamura K, Kouno T, et al. Predictors of recurrence in breast cancer patients with a pathologic complete response after neoadjuvant chemotherapy. Br J Cancer. 2010;103:297–302.
15. Chuthapisith S, Eremin JM, Eremin O. Predicting response to neoadjuvant chemotherapy in breast cancer: molecular imaging, systemic biomarkers and the cancer metabolome (Review). Oncol Rep. 2008;20:699–703.
16. Tanaka T, Tanimoto K, Otani K, Satoh K, Ohtaki M, Yoshida K, et al. Concise prediction models of anticancer efficacy of 8 drugs using expression data from 12 selected genes. Int J Cancer. 2004;11:617–26.
17. Komatsu M, Hiyama K, Tanimoto K, Yunokawa M, Otani K, Ohtaki M, et al. Prediction of individual response to platinum/ paclitaxel combination using novel marker genes in ovarian cancers. Mol Cancer Ther. 2006;5:767–75.
18. Shimokuni T, Tanimoto K, Hiyama K, Otani K, Ohtaki M, Hihara J, et al. Chemosensitivity prediction in esophageal squamous cell carcinoma: novel marker genes and efficacy-prediction formulae using their expression data. Int J Oncol. 2006;28: 1153–62.
19. Fumoto S, Shimokuni T, Tanimoto K, Hiyama K, Otani K, Ohtaki M, et al. Selection of a novel drug-response predictor in esophageal cancer: a novel screening method using microarray and identification of IFITM1 as a potent marker gene of CDDP response. Int J Oncol. 2008;32:413–23.
20. Nishiyama M, Eguchi H. Pharmacokinetics and pharmacogenomics in gastric cancer chemotherapy. Adv Drug Deliv Rev. 2009;61:402–7.
21. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, et al. Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci U S A. 2001;98:10787–92.
22. McLeod HL, Evans WE. Pharmacogenomics: unlocking the human genome for better drug therapy. Annu Rev Pharmacol Toxicol. 2001;41:101–21.
23. Rousseeuw PJ. Least median of squares regression. J Am Stat Assoc. 1984;79:871–80.
24. Turaga K, Acs G, Laronga C. Gene expression profiling in breast cancer. Cancer Control. 2010;17:177–82.
25. Perou CM, Børresen-Dale AL. Systems biology and genomics of breast cancer. Cold Spring Harb Perspect Biol 2010; Epub ahead of print doi:10.1101/cshperspect.a003293.
26. Ou K, Yu K, Kesuma D, Hooi M, Huang N, Chen W, et al. Novel breast cancer biomarkers identified by integrative proteomic and gene expression mapping. J Proteome Res. 2008;7:1518–28.
27. Bergman PJ, Gravitt KR, Ward NE, Beltran P, Gupta KP, O'Brian CA. Potent induction of human colon cancer uptake of chemotherapeutic drugs by *N*-myristoylated protein kinase C-alpha (PKC-alpha) pseudosubstrate peptides through a P-glycoprotein-independent mechanism. Invest New Drugs. 1997;15:311–8.
28. Gravitt KR, Ward NE, Fan D, Skibber JM, Levin B, O'Brian CA. Evidence that protein kinase C-alpha activation is a critical event in phorbol ester-induced multiple drug resistance in human colon cancer cells. Biochem Phamacol. 1994;48:375–81.
29. Chen Y, Yu G, Yu D, Zhu M. PKCalpha-induced drug resistance in pancreatic cancer cells is associated with transforming growth factor-beta1. J Exp Clin Cancer Res. 2010;29:104.
30. Bermudez O, Pagès G, Gimond C. The dual-specificity MAP kinase phosphatases: critical roles in development and cancer. Am J Physiol Cell Physiol. 2010;299:C189–202.
31. Baker C, Belbin O, Kalsheker N, Morgan K. SERPINA3 (aka alpha-1-antichymotrypsin). Front Biosci. 2007;12:2821–35.
32. Cimino D, Fuso L, Sfiligoi C, Biglia N, Ponzone R, Maggiorotto F, et al. Identification of new genes associated with breast cancer progression by gene expression analysis of predefined sets of neoplastic tissues. Int J Cancer. 2008;123:1327–38.
33. Pohl A, Lurje G, Manegold PC, Lenz HJ. Pharmacogenomics and -genetics in colorectal cancer. Adv Drug Deliv Rev. 2009;61: 375–80.
34. Ruzzo A, Graziano F, Kawakami K, Watanabe G, Santini D, Catalano V, et al. Pharmacogenetic profiling and clinical outcome of patients with advanced gastric cancer treated with palliative chemotherapy. J Clin Oncol. 2006;24:1883–91.
35. Lu JW, Gao CM, Wu JZ, Cao HX, Tajima K, Feng JF. Polymorphism in the 3′-untranslated region of the thymidylate synthase gene and sensitivity of stomach cancer to fluoropyrimidine-based chemotherapy. J Hum Genet. 2006;51:155–60.
36. Ichikawa W, Takahashi T, Suto K, Shirota Y, Nihei Z, Shimizu M, et al. Simple combinations of 5-FU pathway genes predict the outcome of metastatic gastric cancer patients treated by S-1. Int J Cancer. 2006;119:1927–33.

Homozygosity Mapping on Homozygosity Haplotype Analysis to Detect Recessive Disease-Causing Genes from a Small Number of Unrelated, Outbred Patients

Koichi Hagiwara[1]*, Hiroyuki Morino[2], Jun Shiihara[1], Tomoaki Tanaka[1], Hitoshi Miyazawa[1], Tomoko Suzuki[1], Masakazu Kohda[3,4], Yasushi Okazaki[3,4], Kuniaki Seyama[5], Hideshi Kawakami[2]

1 Department of Respiratory Medicine, Saitama Medical University, Moroyama, Saitama, Japan, **2** Department of Epidemiology, Research Institute for Radiation Biology and Medicine, Hiroshima University, Hiroshima, Hiroshima, Japan, **3** Division of Functional Genomics and Systems Medicine, Research Center for Genomic Medicine, Research Center for Genomic Medicine, Saitama Medical University, Hidaka, Saitama, Japan, **4** Division of Translational Research, Research Center for Genomic Medicine, Research Center for Genomic Medicine, Saitama Medical University, Hidaka, Saitama, Japan, **5** Department of Respiratory Medicine, Juntendo University School of Medicine, Bunkyo-ku, Tokyo, Japan

Abstract

Genes involved in disease that are not common are often difficult to identify; a method that pinpoints them from a small number of unrelated patients will be of great help. In order to establish such a method that detects recessive genes identical-by-descent, we modified homozygosity mapping (HM) so that it is constructed on the basis of homozygosity haplotype (HM on HH) analysis. An analysis using 6 unrelated patients with Siiyama-type α1-antitrypsin deficiency, a disease caused by a founder gene, the correct gene locus was pinpointed from data of any 2 patients (length: 1.2–21.8 centimorgans, median: 1.6 centimorgans). For a test population in which these 6 patients and 54 healthy subjects were scrambled, the approach accurately identified these 6 patients and pinpointed the locus to a 1.4-centimorgan fragment. Analyses using synthetic data revealed that the analysis works well for IBD fragment derived from a most recent common ancestor (MRCA) who existed less than 60 generations ago. The analysis is unsuitable for the genes with a frequency in general population more than 0.1. Thus, HM on HH analysis is a powerful technique, applicable to a small number of patients not known to be related, and will accelerate the identification of disease-causing genes for recessive conditions.

Citation: Hagiwara K, Morino H, Shiihara J, Tanaka T, Miyazawa H, et al. (2011) Homozygosity Mapping on Homozygosity Haplotype Analysis to Detect Recessive Disease-Causing Genes from a Small Number of Unrelated, Outbred Patients. PLoS ONE 6(9): e25059. doi:10.1371/journal.pone.0025059

Editor: Kazutaka Ikeda, Tokyo Metropolitan Institute of Medical Science, Japan

Received July 29, 2011; **Accepted** August 26, 2011; **Published** September 20, 2011

Funding: This work is supported in part by the grant-in-aid for scientific research (No. 18390242) from the Japan Society of Promotion of Science, and in part by the grants-in-aid for Health and Labor Science [Nos. H22-Nanchi-Ippan-005 to K.H. and H20-Nanchi-Ippan-023 to K.H.] from the Ministry of Health, labor and Welfare, Japan. 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: hagiwark@saitama-med.ac.jp

Introduction

Identification of susceptible genetic loci is of great importance for understanding the underlying mechanisms of a number of diseases, and thus aiding the development of their treatment. Whole-genome association studies using individuals not known to be related have been very successful for the analysis of common diseases [1], while linkage-based approaches have identified a number of genes with large effect sizes [2]. More lately, greater attention has been directed to diseases that cannot be investigated using these approaches, either because of the difficulty in collecting a large number of samples, or in finding a sizeable family with the disease [3]. Such diseases include those caused by multiple rare genetic variants or by genes with low penetrance or with effects that become apparent only in the elderly [4]. For unraveling the causes of such diseases, there is the need for an approach that is effective in the context of a small number of patients not known to be related.

The homozygosity mapping (HM) method was developed to identify a disease-causing gene through analyses of patients from inbred families [5]. This principle was later expanded and applied to patients from outbred families [6,7]. Moreover, the use of SNP data from genome-wide analyses has increased the sensitivity of the detection [8,9]. However, because the algorithm employed in HM is highly vulnerable to genotyping errors, an appropriate correction for such errors is required [10].

In contrast, the homozygosity haplotype (HH) method [9] is an imputation-free method for determining haplotypes, because it uses only a fraction of SNP genotyping data. When a region of conserved homozygosity haplotype (RCHH) is observed in different individuals, there is a reasonable possibility that these individuals share an identical-by-descent (IBD) fragment in 1 or both strands of the homologous chromosomes. The algorithm is robust to genotyping errors and thus requires very little or no correction for genotyping errors.

During a previous study that aimed to identify a disease-causing gene for amyotrophic lateral sclerosis (MIM 613435) [11], we encountered 2 unrelated patients who shared the same homozygous mutation in the *OPTN* gene (MIM 602432). In addition, the region of DNA encompassing the gene contained a number of SNPs that were homozygous in both patients (a runs of homozygous SNPs [RHS] [10]). Further, the RHS was contained

in a 0.9-Mb region of conserved HH (RCHH) [9]. In contrast, the length of RCHH shared between either of the 2 patients and each of the 85 control subjects was shorter than 0.9 Mb. We therefore concluded that these 2 patients are very likely shared the disease-causing IBD gene [11]. We considered that the reasoning had a general application and the presence of a long RCHH that contains an RHS strongly suggested the presence of an IBD fragment. We then encoded this reasoning into a computer program, thereby establishing HM on HH analysis. Here, we show here that this is a powerful method that can identify susceptible loci by identifying homozygous IBD fragments from a small number of outbred patients.

Methods

Ethics Statement

This study was approved by the Institutional Review Boards of Saitama Medical University, Tokyo University, and Juntendo University. All patients involved in the current study provided written informed consent.

HM on HH analysis

HM on HH analysis is a combination of HM analysis [5,10] employing controls and HH analysis [9] employing controls. The analysis does not presume that the patients are from inbred families, and can be performed on patients from the general population. It searches for an RHS overlap that is contained in an RCHH (see below). A candidate region thus obtained may contain a recessive disease-causing gene.

Most recent common ancestor (MRCA)

For patients sharing a disease-causing gene, the most recent common ancestor (MRCA) is the most recent ancestor from whom they inherited the recessive disease-causing gene (**Figure 1A**). Therefore, in the patients, the disease-causing gene is IBD. HM on HH analysis identifies 2 or more patients who are homozygous for this gene.

Structure formed by the IBD fragments

The IBD fragments generate characteristic regions in the genotyping data both in a single patient and between 2 patients.

In a single patient, the overlap of 2 IBD fragments forms an RHS if its length is greater than the RHS cutoff (**Figure 1B**) [10]. Between 2 patients, RHSs can form an overlap (RHS overlap, hereafter). In the RHS overlap, the genotypes of both subjects are identical, forming an RHS overlap in which 2 subjects share an identical genotype (RHS overlap IG, hereafter) (**Figure 1C**). In addition, the overlap of the "region in which at least 1 fragment is derived from the MRCA" generates an RCHH if its length is greater than the RCHH cutoff (**Figure 1C**) [9]. The RHS overlap IG is contained in the RCHH, and the structure is hereby called the RHS overlap IG-RCHH nest. An RHS overlap IG-RCHH nest may be formed by chance between a patient and a control due to a coincidence in the SNP genotype. However, the RHS overlap IG-RCHH nest between the patients is likely to be longer, both in the size of the RHS overlap IG and in the size of the RCHH, than that formed by chance between a patient and a control (**Figure 1D**). Consequently, if we detect an RHS overlap IG-RCHH nest between 2 patients and it is longer than any of that detected between each patient and each control both in the size of the RHS overlap IG and in the size of the RCHH, the RHS overlap IG-RCHH nest is likely to suggest the presence of the IBD fragments in these 2 patients.

HM on HH analysis

HM on HH analysis searches for the RHS overlap IG-RCHH nest. The analysis is composed of 4 steps. Step 1: HM. The RHSs are obtained, and the RHS overlaps are selected as candidate regions for a disease-causing gene (**Figure 2A**) [10]. Step 2: Intermediate analysis 2 (IM2). RHS overlap IGs are selected as candidate regions (**Figure 2B**). Step 3: Intermediate analysis 3 (IM3). For each SNP position contained in an RHS overlap IG detected in Step 2, the presence of an RHS overlap IG between a patient and a control is investigated. When the RHS overlap IG between the 2 patients is longer in size than any of those between a patient and a control, it is selected as a candidate region (**Figure 2C**). Step 4: HH analysis using controls. The RHS overlap IG-RCHH nest is determined between 2 patients. For each SNP position contained in the RHS overlap IG in the RHS overlap IG-RCHH nest, the presence of an RHS overlap IG-RCHH nest formed between a patient and a control is investigated. When the RHS overlap IG between the 2 patients is longer in length than any of those formed between a patient and a control, and the RCHH between the 2 patients is longer in length than any of those formed between a patient and a control, the RHS overlap IG is selected as a candidate region (**Figure 2D**).

Parameter values

The parameter values used in the current study were as follows. The RHS cutoff was 1.2 centimorgans. At this cutoff, the total length of the regions falsely identified as RHSs was less than 1.5 centimorgans in a genome-wide search [10]. Meanwhile, 8.4% of the total length of RHSs fail to be identified as RHSs when the MRCA occurred 20 generations ago; 25%, 40 generations ago; 42%, 60 generations ago; 57%, 80 generations ago, and 69%, 100 generations ago (**Figure S1A**). Before detecting the RHSs, a genotyping error correction algorithm was applied, with the suspected genotyping error rate set at 0.006 [10]. The RCHH cutoff was 0.0 centimorgans; thus, a match of HH of any length was considered to be an RCHH.

Human subjects

Patients with Siiyama-type α1-antitrypsin deficiency (MIM 107400.0039). Siiyama-type α1-antitrypsin deficiency is a rare recessive disease in Japan [12]. Whole-genome high-density SNP array genotyping data of 6 patients [10], who were not related and lived in different areas of Japan, were used in the current study. All patients provided written informed consent. The maximal likelihood estimates of the generational distance of the MRCA for each pair of patients ranged between 5 and 74 (median 61) generations.

Control subjects. The whole-genome high-density SNP genotyping data of 198 healthy Japanese subjects from the general population were provided by Prof. Tokunaga, Tokyo University. Additionally, the SNP genotyping data of 116 JPT (Japanese in Tokyo) subjects was obtained from the HapMap3 release 28 (http://hapmap.ncbi.nlm.nih.gov/), and data corresponding to the SNPs employed in the Genome-Wide Human SNP Array 6.0 were extracted. From these 314 subjects, we chose 261 subjects based on the number of SNPs genotyped (the number of successfully genotyped SNPs for the selected 261 subjects ranged between 707041 and 903804). These 261 subjects were randomly assigned as controls (200 subjects), as participants in a test population (20, 40, or 60 subjects), and a subject who served as the MRCA. The number of controls used was determined because 200 was the largest round number of controls that could be used. The number of the patients in the test population was determined so that the largest test population had 10 times the number of

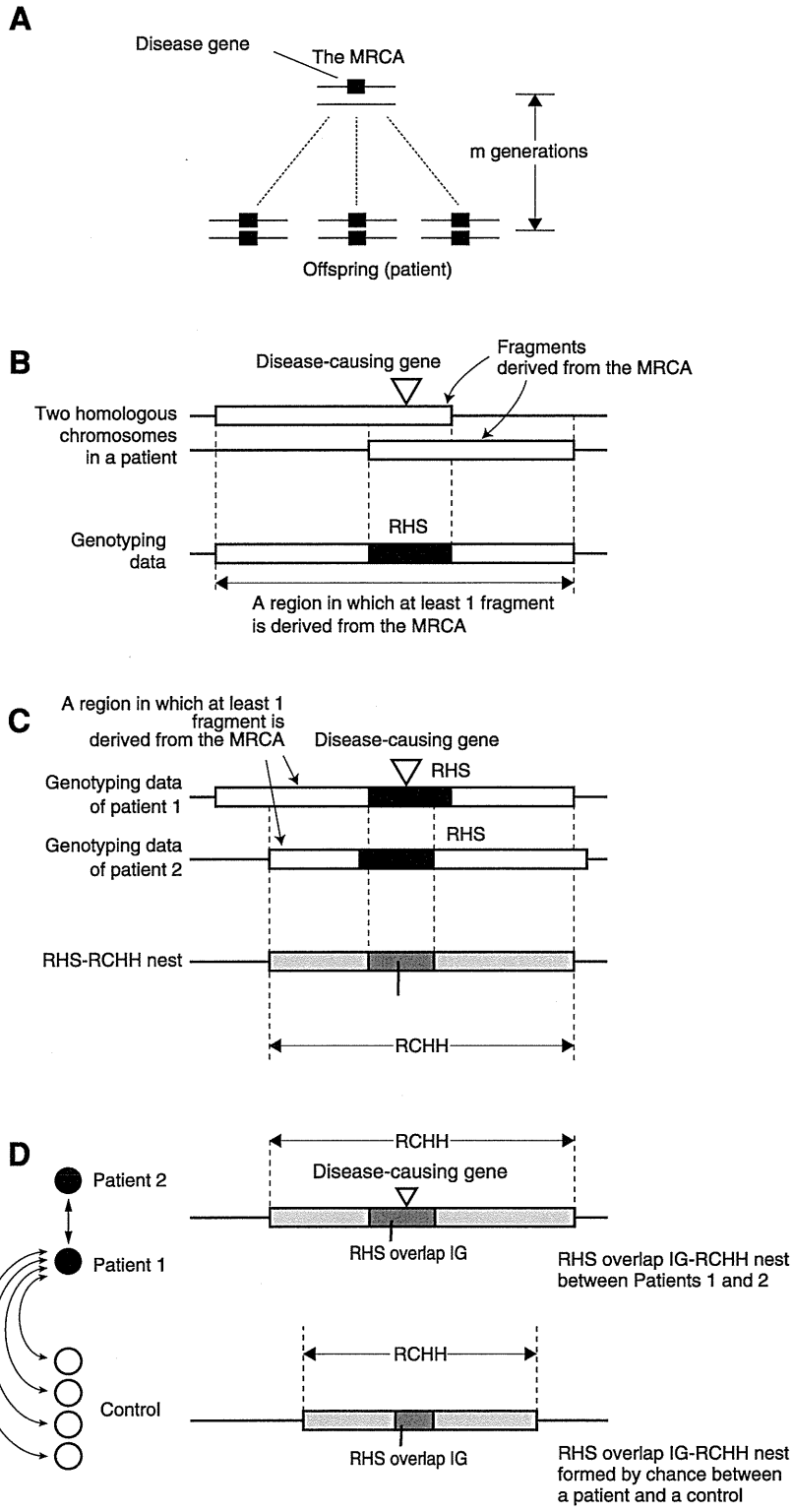

Figure 1. Structures formed by the fragments derived from the MRCA. (A) The MRCA has a single copy of the disease-causing gene. The gene is segregated to the patient through both the maternal and paternal lines, and thus the patients are homozygous for the disease-causing gene. (B) Each of the homologous chromosomes in the patient has a fragment derived from the MRCA. All SNPs in the overlap are homozygous, forming an RHS [10]. The union of the fragments generates "a region in which at least 1 fragment is derived from the MRCA." (C) Assume that there are 2 patients. The genotypes of these patients are identical in the RHS overlap, forming an RHS overlap IG. The overlap of "regions in which at least 1 fragment is derived from the MRCA" forms an RCHH [9]. This RCHH therefore contains the RHS overlap IG. This nested structure is hereby called an RHS overlap IG-RCHH nest. (D) The 2 patients are compared with subjects from a general population (controls). An RHS overlap IG-RCHH nest may be

formed between patients 1 or 2 and each of the controls due to a coincidence in the SNP genotype. However, the RHS overlap IG-RCHH nest between the patients is likely to be longer than any of the RHS overlap IG-RCHH nests accidentally formed between a patient and a control.
doi:10.1371/journal.pone.0025059.g001

patients with Siiyama-type α1-antitrypsin deficiency; it was believed that this number was suited for demonstrating the power of the analysis and for enabling an easy interpretation of the analysis results.

Genotyping

SNP genotyping was performed using the Genome-Wide Human SNP Array 6.0 (Affymetrix).

Synthetic data

The synthetic genotyping data of a patient who shared 2 IBD fragments that contain a disease-causing gene were made as follows: (i) A subject was randomly chosen from the 261 subjects (see above) to serve as the MRCA. (ii) An SNP was randomly chosen from an autosomal region and was considered to mark the position of the disease-causing gene. (iii) The range of the chromosomal region that contained the SNP and was inherited by the patient from the MRCA was calculated according to the Haldane's Poisson process model [13]. (iv) Step (iii) was repeated for the second fragment. (v) The genotyping data of the patient corresponding to the regions that were obtained at steps (iii) and (iv) were replaced with those of the MRCA.

Variables investigated in HM on HH analysis of a population

The variables investigated were the number of subjects in the test population (20, 40, and 60), proportion of patients in the test population (0, 5, 10, 15, 20, 25, and 30%), generational distance of the MRCA (20, 40, 60, 80, and 100 generations), and the gene frequency in the general population (0.0, 0.05, and 0.1). A gene frequency of 0.0 was considered to represent a rare variant, while gene frequencies of 0.05 and 0.1 were considered to represent common variants.

Computer program

The program was written in the Ruby programming language (http://www.ruby-lang.org/en/) with an extension library written in the C programming language (http://gcc.gnu.org/). The program was executed on a MacPro computer that ran on MacOS X 10.6.

Program

HM on HH program is available at Homozygosity Haplotype Analysis Web site, http://www.hhanalysis.com

Results

HM on HH analysis in patients with Siiyama-type α1-antitrypsin deficiency

We tested the performance of HM on HH analysis by using the SNP genotypes of 6 unrelated patients with Siiyama-type α1-antitrypsin deficiency, a rare autosomal recessive disease in Japan caused by a founder mutation of the *SERPINA1* gene (MIM 107400) [12]. As controls, we employed the genotypes of 200 Japanese individuals from the general population. The results obtained after each of the 4 steps that compose HM on HH analysis are shown for a pair of patients (**Figure 3A**). After the completion of the analysis, 2 closely located regions with a total length of 1.4 centimorgans were identified, 1 of which contained *SERPINA1* (**Figure 3A**). The results of the other 14 patient-pair combinations (note that ${}_6C_2 = 15$) were similar: each combination identified candidate regions (total length: 1.2 to 21.8 centimorgans, median: 1.6 centimorgans) that contained *SERPINA1*. Using the genotyping data of only 2 patients, HM on HH analysis was able to narrow the position of the disease-causing gene to a very short chromosomal interval.

HM on HH analysis of a pair of synthetic patients

We further examined the performance of HM on HH analysis of a pair of patients using synthetic data. We investigated the MRCA at 5 different generational distances (20, 40, 60, 80 and 100 generations). For each distance, we employed 60 randomly selected subjects, so that a total of 1770 pairs (${}_{60}C_2 = 1770$) were constructed. Each pair was investigated for 100 randomly selected SNP locations, which were assumed to be the location of a disease-causing gene. The number of trials was thus 177000 (1770 combinations × 100 SNPs) for each generational distance.

Detection of the region containing the RHS overlap depended on the generational distance of the MRCA (**Figure 3B**). This was a major limitation of HM on HH analysis: at the HM step, only RHSs that were longer in length than the RHS cutoff were detected (**Figure S1A**) [10]. The detection will be improved by genotyping more SNPs at a genomewide level, which will allow the use of a smaller RHS cutoff value (**Figure S1B**). Once an RHS overlap was detected at the HM step, HM on HH analysis rarely failed to track it (**Figure 3C**): for the MRCA that occurred 20 generations earlier, the RHS overlap was falsely excluded (false negative) in only 1.5% of the cases, while the falsely included areas (false positive) were reduced from 61.7 centimorgans after the HM step to 0.47 centimorgans after the completion of the HH step, indicating that a small false positive is a prominent feature of HM on HH analysis. Data for the other generations of the MRCA are presented in **Figure S2**.

HM on HH analysis of a population

HM on HH analysis of a population targets a population containing multiple patients sharing an IBD fragment (**Figure 4A**). This simulates a situation in which the population is a collection of patients with the same disease, and some of the patients share an IBD gene. We attempt to identify (1) a patient subgroup sharing an IBD fragment and (2) the chromosomal location of the shared IBD fragment. Here, we defined the analysis level: at analysis level n, the computer program searches for a subgroup consisting of n patients, any pair of which shares an IBD fragment at the same position on the chromosome (**Figure 4B**). To achieve the aims (1) and (2) as stated above, the program identifies (a) the topmost analysis level at which any subgroup is detected, (b) the members that are contained in the subgroup, and (c) the position of the IBD fragment on the chromosome.

First, we investigated the background signal that was detected in the general population (**Figure 4C**). For this purpose, we employed 260 normal subjects. Step (a): 260 normal subjects were randomly divided into a test population (60 subjects) and 200 controls. Step (b): HM on HH analysis of a population was performed. Steps (a) and (b) were repeated 500 times. The histogram of the topmost analysis level, at which any subgroup was detected (**Figure 4D**), demonstrated that a subgroup could be falsely detected (i.e., false positive) in the level 4 analysis and in an earlier analysis level. Conversely, when a positive result was

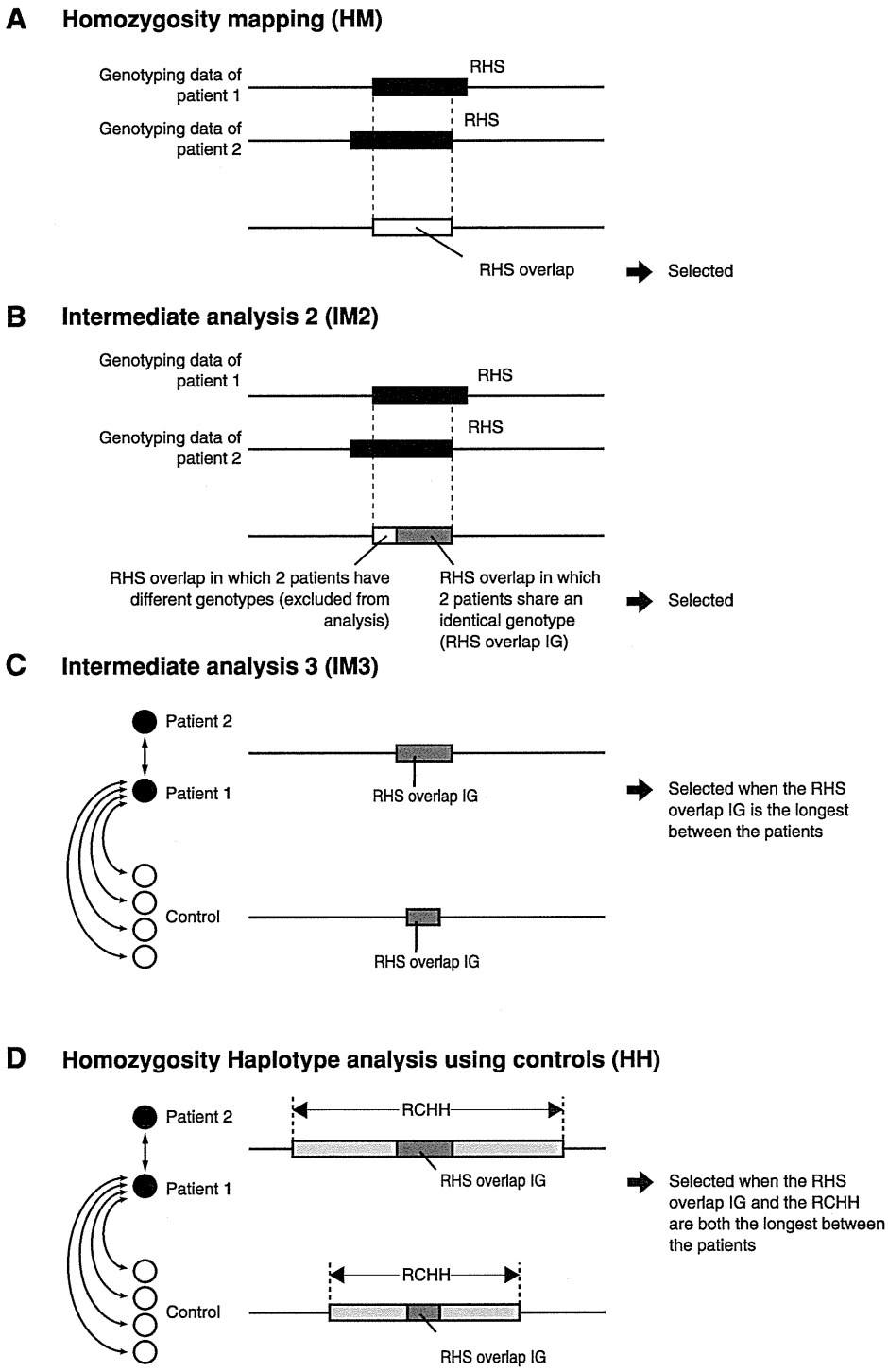

Figure 2. Steps of HM on HH analysis. HM on HH analysis is composed of 4 steps that are serially performed. (A) Homozygosity mapping (HM). The RHSs are determined for each patient, and the RHS overlaps are obtained. (B) Intermediate analysis 2 (IM2). The RHS overlap IGs are determined. (C) Intermediate analysis 3 (IM3). The RHS overlap IGs are compared. The RHS overlap IG is selected as a candidate region when the RHS overlap IG is the longest between the patients. (D) HH analysis using controls. RHS overlap IG-RCHH nests are compared. The RHS overlap IG is selected as a candidate region when the RHS overlap IG and the RCHH are both the longest between the patients.
doi:10.1371/journal.pone.0025059.g002

obtained in the level 5 analysis or in a later analysis, a subgroup sharing an IBD fragment was likely to be detected. Next, we investigated a test population comprising 6 unrelated patients with Siiyama-type α1-antitrypsin deficiency and 54 normal subjects (**Figure 4E**). A subgroup was detected at level 6 (**Figure 4F**); the members of the subgroup were the 6 patients with Siiyama-type

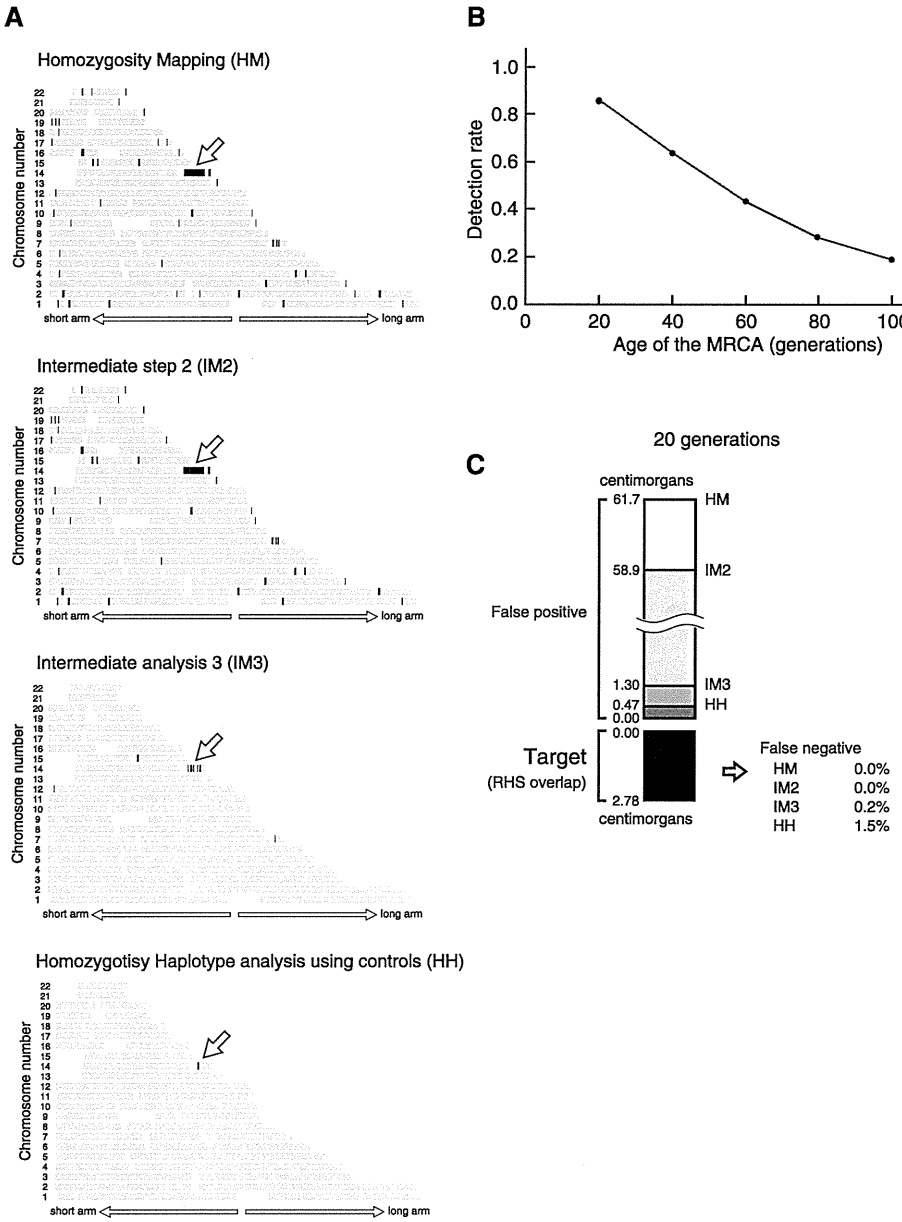

Figure 3. HM on HH analysis of a pair. (A) Analysis of patients 1 and 2 with Siiyama-type α1-antitrypsin deficiency. The position of the disease-causing gene (*SERPINA1*) is indicated by an arrow. HM on HH analysis is composed of 4 steps that are sequentially performed. The regions selected after each step are shown as black bands. The total length of the regions selected at the end was 1.36 centimorgans. (B) The rate at which the RHS overlap was detected by the HM step (i.e., the first step of the analysis) was the major determinant of HM on HH analysis. The detection rate will be improved by genotyping more SNPs genomewide. (C) False positives and false negatives for each analysis. False negatives are decreased with the progression of the analyses. False negatives are very few: 1.5% of the RHS overlap detected by the HM analysis is falsely excluded by HM on HH analysis.
doi:10.1371/journal.pone.0025059.g003

α1-antitrypsin deficiency. The candidate region, 1.2 centimorgans in width, was located on chromosome 14 and contained the *SERPINA1* gene. HM on HH accurately isolated a subpopulation that accounted for only 10% of the population and identified the position of an IBD fragment on the chromosome.

HM on HH analysis of a population containing synthetic patients

To study the performance of HM on HH analysis in more detail, we studied test populations containing synthetic patients. The synthetic patients (5, 10, 15, 20, 25, and 30% of the members of the population) were homozygous for the IBD fragment derived from MRCAs at generational distances of 20, 40, 60, 80, and 100 generations. For each combination of these parameters, the analysis was repeated 100 times by changing the disease-gene location, which was randomly selected from the SNP positions on the autosomes. The analysis was considered successful when (1) only a single candidate region was detected in the topmost level that detected any subgroup, and (2) the candidate region contained the locus of a disease-causing gene. The rates of successful trials (detection rate) were graphed for populations with 60, 40, and 20 subjects (**Figure 5A**). The results demonstrated