Molecular Profiling of Clinical Features in Breast Cancer Using Principal Component Analysis

주성분 분석 방법을 이용한 유방암의 임상적 특징과 관련된 유전자 분석

Mi-Ryung Han[1], Seokho Lee[1], Wonshik Han[2,3], Mihyeon Kim[1], Ju Han Kim[1*], Dong-Young Noh[2,3*]

[1]Seoul National University Biomedical Informatics (SNUBI), [2]Department of Surgery, and [3]Cancer Research Institute, Seoul National University College of Medicine, Seoul 110 -799, Korea

*To whom correspondence s hould be addressed. E-mail: dynoh@plaza.snu.ac.kr, juhan@snu.ac.kr

Abstract

유방암 환자의 임상정보(clinical features)와 cDNA microarray 기술을 이용하여 얻은 유전자 발현 프로파일은 유방암 예후 인자를 찾는 데에 매우 중요하다. 본 논문에서는 임상정보와 유전자 발현 정보를 접목해서 분석하는 방법으로써 주성분 분석(Principal Component Analysis)을 이용하였다. 이 방법은 다변량 자료의 차원을 줄이는 방법으로써, 대용량 실험 데이터로 인해 발생하는 문제점을 해결하기 위하여 많이 쓰이고 있다. 본 연구에서는 주성분 분석을 이용하여 먼저 한국인 유방암 환자 73명의 cDNA microarray 데이터 차원을 줄이고, 이를 통해 얻어진 주성분(Principal Components)과 임상정보 데이터와의 상관관계를 보았다. One-way ANOVA를 이용한 상관관계 분석 결과의 *P*-value는 permutation test를 통해 검증하였다. 동일한 방법을 estrogen receptor (ER) (+) 환자 20명과 ER (-) 환자 31명에 적용해본 결과, ER (-) 환자 중에서 재발과 관련된 유전자를 찾을 수 있었다. 주성분 분석을 molecular phenotypic profiles of clinical features에 이용한 결과 발견된 유전자는 유방암의 재발과 관련된 예후 인자로서 의미가 있다.

Introduction

ER status는 유방암의 biological behavior를 정의하는 데에 중요한 지표가 된다. 특히 ER (+) 환자는 adjuvant hormonal therapy를 적용하였을 때 효과적인 반면, ER (-) 환자의 경우는 효과적이지 못하다[3, 10]. 따라서 ER (-) 환자이면서 재발과 관련된 유전자를 찾

는 것은 postoperative prognosis를 예측하는 데에 매우 의미가 있다[1, 4, 7, 8].

본 논문에서는 cDNA microarray 기술을 이용하여 대량의 유전자 발현 프로파일을 얻었다. DNA에 있는 유전자 정보들을 분석해 내기 위한 cDNA microarray 기술은 현대 생명과학에서 점점 더 중요해지고 있으며 한번에 수만 혹은 수 십 만개의 유전자 발현 프로파일을 얻을 수 있다는 장점을 가지고 있다. cDNA microarray 데이터는 유전자의 개수가 환자의 수(sample)보다 현저하게 많아서 기존의 방법으로는 분석이 곤란하다. 따라서 차원축약이 필요하게 되며 이 방법 중의 하나로써 주성분 분석을 사용할 수 있다[11]. 주성분 분석은 본래의 변수들의 변이(variation)를 적은 수의 변환된 변수로 설명하는 것을 그 목적으로 한다. 이러한 변환된 변수는 본래의 변수들의 선형결합 중에서 분산이 큰 몇 개로 택하게 된다. 여기서는 유방암 환자로부터 얻은 유전자 발현 프로파일을 이 방법을 이용하여 분석하였다.

주성분 분석은 복잡한 genetic dataset으로부터 독립적인 주성분을 얻고 이것을 이용하여 molecular phenotypic profiles of clinical features를 분석하는 데에 이용되었다[2, 5]. 즉 유방암의 지표가 되는 age, tumor size, tumor stage, lymph node metastasis, estrogen receptor status, c_erbB2 status, 재발여부와 같은 임상적 특징들을 주성분과의 상관관계를 이용하여 설명할 수 있는 것이다. 본 연구에서는 독립적인 주성분에 영향을 미치는 gene loading value중에서 extreme loading value 값을 가지는 유전자를 유방암의 재발에 관련된 예후 인자로 정의하였다.

본 논문에서는 먼저 한국인 유방암 환자 73명의 molecular phenotypic profiles of clinical features를 주성분 분석을 이용하여 얻어진 주성분과의 상관관계로 분석하였다. 그리고 동일한 방법으로 ER status에 따라 두 그룹으로 나누어(ER (+), ER (-)) 분석하였다.

Materials and Methods

Patients and tissue samples

1996년 4월부터 2002년 9월까지 서울대학교병원에서 유방암 수술을 받은 73명의 한국인 primary invasive 유방암 환자를 대상으로 하였다. 유방암 수술 후 대부분의 환자는 chemotherapy (77.5%), radiotherapy (36.6%), endocrine therapy (40.8%)의 세가지 adjuvant treatments를 받았다. 유방암 환자 73명의 임상정보는 표 1에 기술되어 있다.

유방암 조직은 맥관 절제(devascularization) 후 20분 이내에 liquid nitrogen에서 얼린 뒤 -80℃에서 냉동 보관하였다. 적어도 40%이상 암세포를 가진 specimen만 실험에 사용되었고, 전체 RNA는 TRIzol solution (Invitrogen, Carlsbad, CA)을 사용하여 추출하였다. 전체 RNA의 농도는 GeneSpec I spectrophotometer(Hitachi, Yokohama, Japan)를 이용하여 측정하였고, amplified tumor RNA는 Cy5로, amplified Universal Human Reference RNA(Stratagene, La Jolla, CA)는 Cy3로 labeling하였다. Labeling한 amplified RNA는 43,200개의 human cDNA clone을 가지는 cDNA microarray slide에 hybridization하였다[9]. Labeling과 hybridization은 이미 기술된 방법대로 수행하였다[12].

이미지 분석과 데이터 전처리(preprocessing)

각각의 cDNA microarray slide는 Axon scanner로 scanning한 후, GenePix Pro3.0 (Axon Instruments, Foster City, CA)을 사용하여

		Number (%)			Number (%)
Age (years)	<30	4 (5.50%)	**Tumor size (cm)**	< 1 cm	2 (2.70%)
	30~39	24 (32.90%)		1~2 cm	15 (20.50%)
	40~49	26 (35.60%)		2~5 cm	47 (4.40%)
	50~59	11 (15.10%)		> 5 cm	9 (2.30%)
	60 >	8 (11.00%)			
Stage	Stage I	10 (13.70%)	**Lymph node positive (number)**	0	27 (37.00%)
	Stage IIa	20 (27.40%)		1~3	19 (26.00%)
	Stage IIb	27 (37.00%)		4~9	13 (17.80%)
	Stage IIIa	13 (17.80%)		≥ 10	13 (17.80%)
	Stage IV	3 (4.10%)		Unknown	1 (1.40%)
c-erbB2 status	Positive	37 (50.70%)	**Estrogen receptor status**	Positive	30 (41.10%)
	Negative	33 (45.20%)		Negative	40 (54.80%)
	Unknown	3 (4.10%)		Unknown	3 (4.10%)
			Recurrence	Disease-free	48 (65.75%)
				Recurrence	25 (34.25%)

표 1. 한국인 유방암 환자 73명의 임상정보

이미지 분석하였다. 이미지 분석 도중에, spot quality가 낮은 것은 직접 눈으로 보면서 제거하였다. Missing value를 가지는 유전자는 분석에서 제외하였고, 43200개의 유전자 중에서 23806개의 유전자를 분석에 이용하였다. Bioconductor R package를 이용하여 VSN transformation을 한 뒤, LOWESS normalization을 하여 각 slide의 평균은 0, 표준편차는 1이 되도록 하였다.

주성분과 임상정보 데이터의 상관관계 분석

먼저 73명의 환자 데이터를 가지고 Bioconductor R package로 주성분 분석을 하여 70개의 독립적인 주성분을 추출하였다. 각 주성분은 전체변동에 대한 상대적인 기여도에 따라 그 값이 내림차순으로 순위가 매겨진다. 주성분을 추출할 때에는 표준편차가 현저히 작은 것은 제외하였다.

다음으로 주성분의 상대적인 기여도와 임상정보 데이터와의 상관관계를 보기 위해 one-way ANOVA가 사용되었다. R package를 사용하여 각각의 주성분과 다음에 해당하는 임상정보 데이터와의 상관관계가 *P*-value로 계산되었다; age, tumor size, tumor stage, lymph node metastasis, estrogen receptor status, c_erbB2 status, 재발여부.

위와 동일한 방법으로 ER (+) 환자 20명과 ER (-) 환자 31명을 각 그룹에 대하여 분석하여 재발여부와의 상관관계를 보았다. 총 51명의 환자는 36개월을 기준으로 재발을 일으키지 않은 환자 26명과 재발을 일으킨 환자 25명으로 구분된다. 이 때에는 주성분 분석을 사용하여 ER (+) 그룹으로부터 19개의 독립적인 주성분을, ER (-) 그룹으로부터 29개의 독립적인 주성분을 각각 추출하였다.

Permutation test를 이용한 상관관계 검증

주성분과 임상정보 데이터와의 one-way ANOVA test후에 얻어진 *P*-value가 유의한 것인지를 검증하기 위해 R package를 이용하여 permutation test를 하였다[6]. 우선 각 환자에 번호를 매기고 번호를 임의로 섞은 후, 다시 임상정보 데이터와 one-way ANOVA test를 하여 *P*-value를 얻는다. 이때 서로 다른 70개의 *P*-value중에서 가장 높은 *P*-value 1개가 기록된다. 같은 방법을 999번 반복한 후에, 가장 높은 999개의 *P*-value와 기록해 놓은 1개의 *P*-value를 비교하여 순위를 매긴다. 그 결과, 기록해 놓은 기존의 *P*-value를 우연히 얻을 수 있을 확률이 구해지고 이것은 P로 나타낸다.

동일한 방법으로 19명의 환자와 29명의 환자에 대하여 각각 permutation test를 한 뒤, 두 그룹에서 얻은 *P*-value가 각각 유의한 것인지를 검증하였다.

Results

주성분 분석

70개의 주성분이 임상·생물학적으로 어떤 의미가 있는지 알아보기 위하여 임상정보 데이터와 one-way ANOVA test를 한 결과는 표 2에 기술되어 있다. 2번째 주성분이 ER status와 가장 유의한 상관관계가 있고 ($P < .00001$, one-way ANOVA), 17번째 주성분이 재발여부와 가장 유의한 상관관계가 있음을 알 수 있었다($P = 0.0008$, one-way ANOVA). 이는 ER statue와 재발에 관련된 임상정보 데이터가 본 연구에 사용된 유방암 환자 데이터와 유의한 연관성이 있음을 보여준다. 통계적인 유의성 검증(permutation test) 결과 나온 확률 값은 기존의 *P*-value의 유의성이 유지됨을 증명하였다(ER; P = 0.001, 재발여부; P = 0.004).

다음으로 ER status (ER (+), ER (-))와 재발여부와의 상관관계를 알아보았다. 20명의 ER (+) 환자와 31명의 ER (-) 환자 데이터에서 각각 추출한 19개, 29개의 주성분이 분석에 이용되었다. ER (+) 환자에서는 16번째 주성분이 재발여부와 가장 유의한 상관관계를 보이고($P = 0.072$, one-way ANOVA), ER (-) 환자에서는 첫 번째 주성분이 재발여부와 가장 유의한 상관관계를 보였다($P < .00001$, one-way ANOVA). ER (-) 환자 데이터로 통계적인 유의성 검증(permutation test)을 하여 얻은 확률 값은 기존의 *P*-value의 유의성이 유지됨을 증명하였다(재발여부; P = 0.001).

주성분과 관련된 유전자 분석

주성분에 대한 특정 유전자의 기여도는 loading value로 측정된다. 따라서 각각의 주성분에 관련된 의미 있는 유전자는 gene loading value로 찾아낼 수 있다.

본 연구에서는 첫 번째 주성분에서 재발과 유의한 상관관계를 보인 ER (-) 환자 데이터를 이용하여 재발관련 진단 유전자를 찾았다. 첫 번째 주성분은 ER (-) 환자이면서 재발을 일으킨 경우와(b) 재발을 일으키지 않은 경우(a)에 molecular phenotype이 서로 다르다는 것을 보여준다(그림 1).

ER (-) 환자에서 재발을 일으키는 유전자는 첫 번째 주성분에서 positive loading value 값을 가지는 것임을 알 수 있다.

이 중에서도 재발과 관련이 깊은 extreme positive loading value를 가지는 몇 가지 유전자를 찾아보았다(표 3). GeneBank accession 번호가 AA233079인 insulin-like growth factor binding protein 1(IGFBP-1)은 유방암 환자의 재발과 관련된 중요한 유전자로 알려져 있

< Clinical features >		ER	c_erbB2	rec	age	stage	size	Np
ER(+) vs ER(-)	PC2	<.00001						
c_erbB2(+) vs c_erbB2(-)	PC9		0.0001					
rec=1 vs rec=0	PC17			0.0008				
age>50 vs age≤50 (year)	PC62				0.003			
stage2 vs stage3	PC23					0.007		
size>5 vs size≤5 (cm)	PC34						0.022	
Np=0 vs Np>0	PC65							0.023

표 2. 주성분과 임상정보 데이터와의 상관관계

(ER, estrogen receptor; rec; recurrence; Np, lymph node positive개수)

다[13]. GeneBank accession 번호가 AA865573인 GRB2-associated binding protein 2은 유방암에서 estrogen 조절인자와 관련이 있다[14].

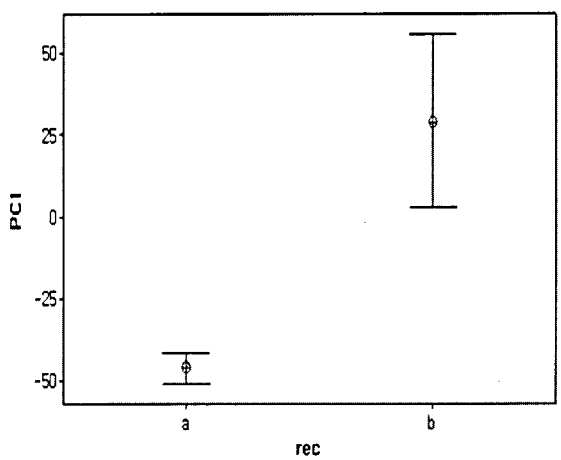

그림 1. ER (-) 환자 31명 데이터에서 추출한 첫 번째 주성분(Principal Component)과 재발여부와의 상관관계 (PC1, Principal Component 1; rec, recurrence; a, 비재발; b, 재발)

이 외에도 Tissue factor pathway inhibitor 2, Lymphocyte cytosolic protein 1 (L-plastin), LIM and SH3 protein 1, Solute carrier family 39(Zinc transporter) member 6, B lymphoid tyrosine kinase는 유방암에 있어서 종양 형성과정에 영향을 미친다.

Discussion

본 논문에서는 임상적 특징들을 유전자 발현 프로파일과의 상관관계로 설명하기 위해 주성분 분석을 이용하였다. 주성분 분석은 대량의 데이터를 분석할 때 차원을 줄이는 방법으로 쓰이고 있다.

본 연구에서는 cDNA microarray 기술을 이용하여 얻은 대량의 유전자 발현 프로파일을 분석하기 위하여 주성분 분석을 이용하였다. ER (-) 환자 데이터 분석에서 추출한 주성분과 임상정보 데이터와의 상관관계는 재발에 영향을 미치는 유전자를 찾는 데에 기여한다. 특히 ER (-) 환자는 adjuvant hormonal therapy로 치료되기가 어려우므로 ER (-)이면서 재발과 관련된 유전자는 예후 인자로써 의미가 있다.

먼저 73명의 유방암 환자 데이터 전체를 분석하였을 때 ER status와 재발여부가 각각의 주성분과 유의한 상관관계를 보였고, 다음으로 두 그룹(ER (+), ER (-))에서 각각 재발여부와의 상관관계를 분석하였다. ER (-) 환자 데이터로 주성분 분석을 하였을 때 가장 데이터를 잘 설명할 수 있는 첫 번째 주성분이 재발여부와 가장 유의한 상관관계를 보였으므로($P < .00001$, one-way ANOVA) 본

Loading value	Accession N	Gene name
0.039864	AA973805	chromosome 6 open reading frame 110
0.039475	W56522	dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)
0.038201	H29315	solute carrier family 39 (zinc transporter), member 6
0.036653	AA779752	T-cell activation WD repeat protein
0.036113	R96780	apolipoprotein A-I
0.035086	AA133167	KIAA1644 protein
0.035067	AA496795	intersectin 1 (SH3 domain protein)
0.034693	AA425719	hypothetical protein BC016861
0.034341	AA670430	G-protein coupled receptor protein signaling pathway
0.033465	AA233079	insulin-like growth factor binding protein 1: IGFBP-1
0.032781	AA486284	DORA reverse strand protein 1
0.032622	T59043	alpha-fetoprotein
0.032569	AA682807	Rhesus blood group-associated glycoprotein
0.032347	AA251784	galactosidase, alpha,GLA (GALA)
0.032269	T71042	echinoderm microtubule associated protein like 4
0.031934	H64379	angiotensinogen (serine (or cysteine) proteinase inhibitor, clade A
0.031525	AA670357	solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11
0.031494	AA994689	natriuretic peptide receptor B/guanylate cyclase B (atrionatriuretic peptide receptor B)
0.031265	AI084074	B lymphoid tyrosine kinase
0.030862	H93332	apolipoprotein B (including Ag(x) antigen)
0.030476	AA001290	spectrin, alpha, erythrocytic 1 (elliptocytosis 2)
0.030148	R60169	guanine deaminase; Guanine deaminase (GDA)
0.029921	H73941	hemoglobin, zeta
0.029886	AA919020	transketolase-like 1
0.029576	R59086	LIM and SH3 protein 1
0.029459	W73144	lymphocyte cytosolic protein 1 (L-plastin)
0.02837	T61078	carbamoyl-phosphate synthetase 1, mitochondrial
0.028366	AA865573	GRB2-associated binding protein 2
0.028317	AI261686	dopa decarboxylase (aromatic L-amino acid decarboxylase)
0.028228	AA679218	hypothetical protein MGC14421
0.028193	AA701860	follistatin

표 3. 첫 번째 주성분과 상관관계가 있는 유전자

연구에서 밝혀진 유전자는 예후 인자로써 의미가 있다. 이 때에 통계적인 유의성 검증(permutation test) 결과 나온 확률 값도 유의하게 나와 기존의 *P*-value가 유지됨을 증명하였다(재발여부; P = 0.001). 첫 번째 주성분에서 positive gene loading value를 통해 밝혀진 유전자 중에서 특히 유방암의 재발, 형성과 관련이 깊은 유전자는 다음과 같다; Insulin-like growth factor binding protein 1, Tissue factor pathway inhibitor 2, Lymphocyte cytosolic protein 1 (L-plastin), Solute carrier family 39 (zinc transporter), Apolipoprotein A -I, LIM and SH3 protein 1, GRB2 -associated binding protein 2, B lymphoid tyros ine kinase

이와 같이 주성분 분석을 이용하여 환자의 유전자 발현 프로파일을 임상정보 데이터와의 상관관계로 설명할 수 있는 것은, molecular phenotypic profile을 분석할 수 있다는 점에서 의의가 있다.

Acknowledgements

본 연구는 Korea Health 21 R&D Project. Ministry of Health & Welfare, R.O.K (01-PJ3-PG6-01GN07-0004)에 의해 지원되었음.

References

[1] C. Jones *et al.*, Expression Profiling of Purified Normal Human Luminal and Myoepithelial Breast Cells: Identification of Novel Prognostic Markers for Breast Cancer, *Cancer Research, 64*, 2004, 3037-3045

[2] S. Bicciato, A. Luchini and C. D. Bello, PCA disjoint models for multiclass cancer analysis using gene expression data, *Bioinformatics, 19*, 2003, 571-578

[3] T. Nagahata *et al.*, Expression profiling to predict postoperative prognosis for estrogen receptor-negative breast cancers by analysis of 25,344 genes on a cDNA microarray, *Cancer Science, 95*, 2004, 218-225

[4] E. Huang *et al.*, Gene expression predictors of breast cancer outcomes, *THE LANCENT, 361*, 2003

[5] F. M. Selaru *et al.*, An Unsupervised Approach to Identify Molecular Phenotypic Components Influencing Breast Cancer Features, *Cancer Research, 64*, 2004, 1584-1588

[6] J. Landgrebe, W. Wurst and G. Welzl, Permutation-validated principal components analysis of microarray data, *Genome Biology, 3*(4), 2002

[7] M. Onda *et al.*, Gene expression patterns as marker for 5-year postoperative prognosis of primary breast cancers, *J Cancer Res Clin Oncol, 130*, 2004, 537-545

[8] A. Barnes *et al.*, Expression of p27kip1 in breast cancer and its prognostic significance, *Journal of Pathology, 201*, 2003, 451-459

[9] H. Zhao *et al.*, Different Gene Expression Patterns in Invasive Lobular and Ductal Carcinomas of the Breast, *Molecular Biology of the Cell, 15*, 2004, 2523-2536

[10] N. Yoshida *et al.*, Prediction of prognosis of estrogen receptor-positive breast cancer with combination of selected estrogen-regulated genes, *Cancer Science, 95*, 2004, 496-50

[11] S. Raychaudhuri, J. M. Stuart and R. B. Altman, Principal Components Analysis to Summarize Microarray Experiments : Application to Sporulation Time Series, *Pac. Symp. Biocomput.*, 2000, 455-466

[12] Zhao *et al.*, Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis, *BMC Genomics, 3*(1), 2002, 31

[13] Pamela J. *et al.*, Insulin-like growth factor binding proteins 1 and 3 and breast cancer outcomes, *Breast Cancer Research and Treatment, 74*, 2002, 65-76

[14] Daly RJ et al., The docking protein Gab2 is overexpressed and estrogen regulated in human breast cancer, *Oncogene, 21*(33), 5175-81, 2002