



(19) 대한민국특허청(KR)

(12) 등록특허공보(B1)

(45) 공고일자 2022년12월23일

(11) 등록번호 10-2479605

(24) 등록일자 2022년12월16일

(51) 국제특허분류(Int. Cl.)

C12Q 1/6886 (2018.01)

(52) CPC특허분류

C12Q 1/6886 (2022.01)

C12Q 2537/165 (2013.01)

(21) 출원번호 10-2022-0052542

(22) 출원일자 2022년04월28일

심사청구일자 2022년04월28일

(56) 선행기술조사문헌

KR102168801 B1

KR102281058 B1

Lancet Oncol, 21(10): 1378-1386 (2020.09.28.)

(73) 특허권자

주식회사 바스젠바이오

서울특별시 마포구 마포대로 25, 8층 801호(마포동, 신한디엠빌딩)

연세대학교 산학협력단

서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)

(72) 발명자

김호

경기도 광명시 광복로 43번길 8, 삼성빌라 302호 이솔

경기도 광명시 하안로 154, 102동 202호(소하동, 금호어울림아파트)

(뒷면에 계속)

(74) 대리인

류중우, 한윤호, 이내영

전체 청구항 수 : 총 9 항

심사관 : 이준혁

(54) 발명의 명칭 위암 발병의 위험도를 예측하기 위한 조성물 및 이를 이용한 방법

## (57) 요약

단일염기다형성(SNP)의 유전자 분석을 통하여 위암 발병의 위험도를 예측하기 위한 조성물, 마이크로어레이, 키트 및 이를 이용하여 개체의 위암 발병 위험도를 예측하는 방법에 관한 것이다. 이에 의하면, 개체의 위암 발병 위험도를 정확하고 간편하게 예측할 수 있다. 따라서, 본 발명은 위암 발병을 예방하는데 도움을 줄 수 있다.

## 대표도

SNP	CHR	BP	A1	MAF	OR (95% CI)	P
rs377176971	4	56,865,490	T	0.0185	8.80(7.44-10.41)	2.00E-38 (서열번호 1)
rs78381605	4	99,477,699	A	0.0463	2.18(1.95-2.42)	4.30E-13 (서열번호 2)
rs145516817	4	134,816,182	T	0.0175	3.32(2.79-3.93)	2.50E-12 (서열번호 3)
rs114167943	7	100,609,631	T	0.0128	3.17(2.59-3.87)	8.30E-09 (서열번호 4)
rs150274610	8	52,042,955	A	0.0489	1.77(1.60-1.96)	2.30E-08 (서열번호 5)
rs10674898	8	143,780,261	TAA	0.3788	1.46(1.39-1.53)	1.40E-15 (서열번호 6)
rs143172300	9	135,545,995	A	0.0189	8.15(6.90-9.61)	1.00E-36 (서열번호 7)
rs373412232	10	101,120,600	T	0.0211	0.40(0.34-0.47)	6.00E-09 (서열번호 8)
rs2470155	15	51,590,753	T	0.0243	0.42(0.37-0.49)	4.90E-09 (서열번호 9)
rs760077	1	155,178,782	A	0.1236	0.64(0.60-0.69)	6.60E-11 (서열번호 10)

(52) CPC특허분류

C12Q 2600/156 (2013.01)

(72) 발명자

**김정오**

경기도 수원시 팔달구 권광로317번길 12, 3동 402호

**김정은**

서울특별시 서대문구 가재울로 45, 106동 1403호(남가좌동, 남가좌동현대아파트)

**윤상혁**

경기도 김포시 김포한강2로 292, 403동 601호

**박승환**

서울특별시 동작구 상도로61길 8-12, 204호

**차지희**

서울특별시 강서구 강서로 348, 129동 101호(내발산동, 우장산힐스테이트)

**박다현**

서울특별시 관악구 신림로58길 43, 302(신림동, 힐플레이스)

**지선하**

서울 마포구 월드컵북로30길 9-22, 101동 903호

## 명세서

### 청구범위

#### 청구항 1

서열번호 1의 핵산 서열로 이루어진 폴리뉴클레오티드 및 서열번호 7의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드를 포함하는

위암 발병 위험도를 예측하기 위한 조성물.

#### 청구항 2

제1항에 있어서,

서열번호 1의 51번째 뉴클레오티드의 다형성 부위는 C 또는 T이고; 및

서열번호 7의 51번째 뉴클레오티드의 다형성 부위는 G 또는 A인 것인, 위암 발병 위험도를 예측하기 위한 조성물.

#### 청구항 3

제1항 또는 제2항에 따른 조성물을 포함하는 위암 발병 위험도를 예측하기 위한 마이크로어레이.

#### 청구항 4

제1항에 따른 조성물을 포함하는 위암 발병 위험도를 예측하기 위한 키트.

#### 청구항 5

제4항에 있어서,

상기 키트는 PCR 키트 또는 DNA 칩 키트인 것인, 위암 발병 위험도를 예측하기 위한 키트.

#### 청구항 6

개체로부터 수득된 핵산 시료로부터 서열번호 1의 핵산서열로 이루어진 폴리뉴클레오티드 및 서열번호 7의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위의 유전자형(genotype)을 검출하는 단계; 및

검출된 유전자형으로부터 개체의 위암 발병 위험도를 산출하는 단계로서,

서열번호 1의 검출된 유전자형이 T이고 서열번호 7의 검출된 유전자형이 A인 경우, 위암 발병 위험도가 높아질 것으로 결정하는 것인 단계를 포함하는 위암 발병 위험도를 예측하는 방법.

#### 청구항 7

제6항에 있어서,

상기 유전자형을 검출하는 단계는

서열번호 1의 핵산 서열로 이루어진 폴리뉴클레오티드 및 서열번호 7의 핵산 서열로 이루어진 폴리뉴클레오티드에서 상기 다형성부위를 포함하는 5개 이상의 연속 핵산 서열과 동일하거나 또는 상보적인 핵산 서열을 포함하는 프라이머 또는 프로브를 사용하거나, 또는

서열번호 1의 핵산 서열로 이루어진 폴리뉴클레오티드 및 서열번호 7의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 이상의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드가 고정된 마이크로어레이에 상기 핵산 시료를 혼성화하여 수행되는 것인, 위암 발병 위험도를 예측하는 방법.

## 청구항 8

삭제

## 청구항 9

삭제

## 청구항 10

삭제

## 청구항 11

제6항에 있어서,

상기 개체는 60세 미만인 것인, 위암 발병 위험도를 예측하는 방법.

## 청구항 12

제6항에 있어서,

산출된 위암 발병 위험도에 따라 개체의 위암 발병의 위험도를 양호 등급 또는 위험 등급으로 결정하는 단계를 더 포함하는 것인, 위암 발병 위험도를 예측하는 방법.

## 발명의 설명

### 기술 분야

[0001] 단일염기다형성(SNP)의 유전자 분석을 통하여 위암 발병의 위험도를 예측하기 위한 조성물 및 이를 이용한 방법에 관한 것이다.

### 배경 기술

[0002] 위암(gastric cancer)은 가장 흔한 암 중 하나이며 2018년에 783,000명이 사망한 것으로 추정된다. 과거 몇 년 동안, 위암의 발달 및 진행에 영향을 미치는 다중 유전적 또는 후생학적 변화를 설명하기 위한 많은 연구가 있어왔다. 진단 및 치료 기술의 발달은 위암에 대한 탁월한 장기 생존의 결과를 내었지만, 위암은 여전히 전 세계에서 암에 의한 사망에 있어서 그 빈도가 두 번째로 높은 원인이다.

[0003] 인간의 경우 약 1000 bp(base pair) 당 1회 빈도로 변이가 있는데 이것을 단일염기다형성(single nucleotide polymorphism: SNP)이라고 한다. 5% 다형을 보통의 다형성(common polymorphism)이라고 하며, 1% 내지 5% 다형을 드문 다형성(rare polymorphism)이라고 한다. 현재 인간의 전체 염기서열을 분석하기 위해 많은 실험 기법 등이 발달하였고, 그 중 전장 유전체 연관성 분석(genome-wide association study: GWAS) 현재까지 많은 질환 연구에 사용되고 있다. 후보 유전자 접근법 또는 전장 유전체 연관성 분석(GWAS)을 통해 위암과 관련된 SNP로 APE1(rs1760944), DNMT1(rs16999593), ERCC5(rs751402), GSTT1(null/presence), MDM2(rs2278744), PPARG(rs1801282), TLR4(rs4986790), IL-17F(rs763780), CASP8(rs3834129) 등이 보고되었다(Jie Tian et al., Cancer Biol. Med. (2019) 16:361-389). 하지만, 이러한 SNP에 대한 연구는 위암 유형 및 민족성과 관련하여 일관성 없는 결과를 나타냈으며, GWAS로부터 얻은 SNP의 효과 크기는 일반적으로 2.0 미만으로 낮다는 문제가 있다.

[0004] 따라서, SNP 다형성을 이용하여 위암 발병 위험도를 보다 정확하게 예측하기 위한 SNP 바이오마커를 발굴하고, 이를 이용하여 위암의 위험도를 예측할 수 있는 방법을 개발할 필요가 있다.

### 선행기술문헌

#### 비특허문헌

[0005] (비특허문헌 0001) Jie Tian et al. Cancer Biol Med (2019) 16:361-389.

## 발명의 내용

### 해결하려는 과제

- [0006] 위암 발병 위험도를 예측하기 위한 조성물을 제공한다.
- [0007] 위암 발병 위험도를 예측하기 위한 마이크로어레이를 제공한다.
- [0008] 위암 발병 위험도를 예측하기 위한 키트를 제공한다.
- [0009] 위암 발병 위험도를 예측하기 위한 조성물을 이용하여 위암 발병 위험도를 예측하는 방법을 제공한다.

### 과제의 해결 수단

- [0010] 일 양상은 서열번호 1 내지 서열번호 9로 이루어진 군으로부터 선택된 하나 이상의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드를 포함하는 위암 발병 위험도를 예측하기 위한 조성물을 제공한다.
- [0011] 상기 조성물은 서열번호 10의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드를 더 포함할 수 있다.
- [0012] 상기 "다형성 부위(polymorphic site)"는 핵산 서열 중 단일염기다형성(single nucleotide polymorphism: SNP)을 나타내는 부위를 말한다. 개개의 다형성은 또한 당업자에게 공지되어 있고, 예를 들면, NCBI 웹사이트 상에서 이용가능한 뉴클레오티드 염기 변이의 단일염기다형성 데이터베이스(single nucleotide polymorphism database(dbSNP) of nucleotide sequence variation)에서 사용되는 것인, 지정된 독특한 식별자("기준 SNP", "refSNP" 또는 "rs#")이다.
- [0013] 상기 "단일염기다형성(single nucleotide polymorphism, SNP)"은 하나의 유전자 좌위(locus)에 두 가지 이상의 대립유전자(allele)가 존재하는 다형성 부위(polymorphic site) 중에서, 단일 염기만이 다른 것을 의미한다. 예를 들어, 서로 다른 개체의 세 개의 DNA 단편들(예: AAGT[A/A]AG, AAGT[A/G]AG, AAGT[G/G]AG)과 같이 단일염기에서 차이가 있을 때, 이를 두 개의 대립유전자(C 또는 T)라고 부르며 대다수의 SNP는 두 개의 대립유전자(allele)를 포함한다. 상기 "대립유전자(allele)"는 상동염색체의 동일한 유전자좌위에 존재하는 한 유전자의 여러 타입을 의미한다. 대립유전자는 다형성을 나타내는데 사용되기도 하며, 예를 들어, SNP는 두 종류의 대립인자(biallele)를 갖는다.
- [0014] 상기 폴리뉴클레오티드는 서열번호 1 내지 서열번호 10으로 이루어진 군으로부터 선택된 하나 이상의 핵산 서열로 이루어진 폴리뉴클레오티드에서, 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 내지 100개, 10개 내지 100개, 10개 내지 80개, 10개 내지 60개, 10개 내지 40개, 10개 내지 20개, 또는 5개 내지 25개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드일 수 있다.
- [0015] 상기 서열번호 1의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드는 SNP ID rs377176971의 유전자형을 검출하기 위한 폴리뉴클레오티드일 수 있다. 상기 rs377176971은 서열번호 1의 핵산 서열로 구성된 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위일 수 있다. 상기 rs377176971은 인간 4번째 염색체의 56865490번째 위치에 있는 SNP일 수 있다. 상기 rs377176971은 C 또는 T의 유전자형(genotype)을 가질 수 있다. rs377176971의 유전자형이 T인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.
- [0016] 상기 서열번호 2의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드는 SNP ID rs78381605의 유전자형을 검출하기 위한 폴리뉴클레오티드일 수 있다. 상기 rs78381605은 서열번호 2의 핵산 서열로 구성된 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위일 수 있다. 상기 rs78381605은 인간 4번째 염색체의 99477699번째 위치에 있는 SNP일 수 있다. 상기 rs78381605는 G 또는 A의 유전자형을 가질 수 있다. rs78381605의 유전자형이 A인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.

- [0017] 상기 서열번호 3의 핵산 서열로 이루어진 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오타이드는 SNP ID rs145516817의 유전자형을 검출하기 위한 폴리뉴클레오타이드일 수 있다. 상기 rs145516817은 서열번호 3의 핵산 서열로 구성된 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위일 수 있다. 상기 rs145516817은 인간 4번째 염색체의 134816182번째 위치에 있는 SNP일 수 있다. 상기 rs145516817은 G, A 또는 T의 유전자형을 가질 수 있다. rs145516817의 유전자형이 T인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.
- [0018] 상기 서열번호 4의 핵산 서열로 이루어진 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오타이드는 SNP ID rs114167943의 유전자형을 검출하기 위한 폴리뉴클레오타이드일 수 있다. 상기 rs114167943은 서열번호 4의 핵산 서열로 구성된 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위일 수 있다. 상기 rs114167943은 인간 7번째 염색체의 100609631번째 위치에 있는 SNP일 수 있다. 상기 rs114167943은 C 또는 T의 유전자형을 가질 수 있다. rs114167943의 유전자형이 T인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.
- [0019] 상기 서열번호 5의 핵산 서열로 이루어진 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오타이드는 SNP ID rs150274610의 유전자형을 검출하기 위한 폴리뉴클레오타이드일 수 있다. 상기 rs150274610은 서열번호 5의 핵산 서열로 구성된 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위일 수 있다. 상기 rs150274610은 인간 8번째 염색체의 52042955번째 위치에 있는 SNP일 수 있다. 상기 rs150274610은 C 또는 A의 유전자형을 가질 수 있다. rs150274610의 유전자형이 A인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.
- [0020] 상기 서열번호 6의 핵산 서열로 이루어진 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오타이드는 SNP ID rs10674898의 유전자형을 검출하기 위한 폴리뉴클레오타이드일 수 있다. 상기 rs10674898은 서열번호 6의 핵산 서열로 구성된 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위일 수 있다. 상기 rs10674898은 인간 8번째 염색체의 143780261번째 위치에 있는 SNP일 수 있다. 상기 rs10674898은 AAAA 또는 AA 유전자형을 가질 수 있다. rs10674898의 유전자형이 AA인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.
- [0021] 상기 서열번호 7의 핵산 서열로 이루어진 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오타이드는 SNP ID rs143172300의 유전자형을 검출하기 위한 폴리뉴클레오타이드일 수 있다. 상기 rs143172300은 서열번호 7의 핵산 서열로 구성된 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위일 수 있다. 상기 rs143172300은 인간 9번째 염색체의 135545995번째 위치에 있는 SNP일 수 있다. 상기 rs143172300은 G, A 또는 C의 유전자형을 가질 수 있다. rs10674898의 유전자형이 A인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.
- [0022] 상기 서열번호 8의 핵산 서열로 이루어진 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오타이드는 SNP ID rs373412232의 유전자형을 검출하기 위한 폴리뉴클레오타이드일 수 있다. 상기 rs373412232는 서열번호 8의 핵산 서열로 구성된 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위일 수 있다. 상기 rs373412232는 인간 10번째 염색체의 101120600번째 위치에 있는 SNP일 수 있다. 상기 rs373412232는 C 또는 T의 유전자형을 가질 수 있다. rs373412232의 유전자형이 T인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.
- [0023] 상기 서열번호 9의 핵산 서열로 이루어진 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오타이드는 SNP ID rs2470155의 유전자형을 검출하기 위한 폴리뉴클레오타이드일 수 있다. 상기 rs2470155는 서열번호 9의 핵산 서열로 구성된 폴리뉴클레오타이드에서 그의 5'-말단으로부터 51번째 뉴클레오타이드의 다형성 부위일 수 있다. 상기 rs2470155는 인간 15번째 염색체의 51590753번째 위치에 있는 SNP일 수 있다. 상기 rs2470155는 G, C 또는 T의 유전자형을 가질 수 있다. rs2470155의 유전자형이 T인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.



있다.

- [0024] 상기 서열번호 10의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 내지 100개의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드는 SNP ID rs760077의 유전자형을 검출하기 위한 폴리뉴클레오티드일 수 있다. 상기 rs760077은 서열번호 10의 핵산 서열로 구성된 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위일 수 있다. 상기 rs760077은 인간 1번째 염색체의 155178782번째 위치에 있는 SNP일 수 있다. 상기 rs760077은 T 또는 A의 유전자형을 가질 수 있다. rs760077의 유전자형이 A인 경우, 위암 발병 위험도가 높아질 것으로 예상될 수 있다.
- [0025] 상기 폴리뉴클레오티드는 서열번호 1 내지 서열번호 10으로 이루어진 군으로부터 선택된 하나 이상의 핵산 서열로 이루어진 폴리뉴클레오티드와 상동성 또는 동일성이 약 80% 이상, 약 90% 이상, 약 95% 이상, 약 96% 이상, 약 97% 이상, 약 98% 또는 약 99% 이상인 염기서열로 이루어질 수 있으나, 이에 제한되지 않는다.
- [0026] 상기 조성물은 상기 폴리뉴클레오티드를 1종 이상, 2종 이상, 3종 이상, 4종 이상, 5종 이상, 6종 이상, 7종 이상, 8종 이상, 9종 이상 또는 10종을 포함할 수 있다. 상기 폴리뉴클레오티드 또는 이들의 조합을 포함하는 경우 위암 발병 위험도의 정확도를 높일 수 있다. 예를 들어, 서열번호 1에서 다형성 부위의 C 및 서열번호 7에서 다형성 부위의 G가 각각 T 및 A로 모두 변이되는 것일 수 있다.
- [0027] 상기 5 내지 100개의 연속적인 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드는 위암 발병과 유의적 상관관계를 갖는 SNP를 검출하기 위한 프라이머 또는 프로브일 수 있다.
- [0028] 용어 "프라이머(primer)"는 중합효소에 의한 뉴클레오티드의 중합반응에서, 개시점으로 작용할 수 있는 단일가닥의 폴리뉴클레오티드를 말한다. 예를 들면, 상기 프라이머는 적합한 온도 및 적합한 완충액 내에서 적합한 조건, 즉, 4종의 다른 뉴클레오티드 트리포스페이트 및 중합효소의 존재 하에서 주형-지시(template-directed) DNA 합성의 개시점으로 작용할 수 있는 단일가닥의 폴리뉴클레오티드일 수 있다. 프라이머의 적합한 길이는 다양한 인자, 예를 들어, 온도와 프라이머의 용도에 따라 달라질 수 있다. 예를 들어, 프라이머의 길이가 짧을수록, 낮은 어닐링(annealing) 온도에서 주형과 충분히 안정된 혼성화 복합체를 형성할 수 있다. 상기 프라이머의 길이는 약 5 내지 약 100 뉴클레오티드(이하, 'nt'라고 함), 약 10 내지 약 80 nt, 약 10 내지 약 60 nt, 약 10 내지 약 50 nt, 약 10 내지 약 40 nt, 약 10 내지 약 30 nt, 또는 약 10 내지 약 20 nt일 수 있다. 또한, 상기 프라이머는 상기 다형성 부위에 동일하거나 또는 상보적인 서열을 포함할 수 있다. 따라서, 상기 프라이머는 다형성 부위를 포함하는 DNA 서열에 혼성화하여 다형성 부위를 포함하는 DNA 단편을 증폭시킬 수 있다. 상기 프라이머는 대립형질을 검출하여 위암 발병 위험도를 예측하기 위한 키트, 마이크로어레이, 예측 방법 등에 사용될 수 있다.
- [0029] 용어 "프로브(probe)"는 표적 핵산에 특이적으로 결합할 수 있는 올리고뉴클레오티드를 말한다. 상기 프로브는 길이가 약 50 nt 내지 약 400 nt, 약 100 nt 내지 약 350 nt, 약 150 nt 내지 약 300 nt, 또는 약 200 nt 내지 약 250 nt일 수 있다. 또한, 상기 프로브는 상기 다형성 부위에 동일하거나 또는 상보적인 서열을 포함할 수 있다. 상기 프로브는 대립형질을 검출하여 위암 발병 위험도를 예측하기 위한 키트, 마이크로어레이, 예측 방법 등에 사용될 수 있다. 상기 프로브는 검출할 수 있도록 표지될 수 있으며, 예를 들면 방사선 동위원소, 형광 화합물, 바이오 발광 화합물, 화학 발광 화합물, 금속 킬레이트 또는 효소로 표지될 수 있다. 프로브를 표지하는 방법은 당해 분야에서 널리 알려진 기술이며, 통상적인 방법을 통해 수행할 수 있다.
- [0030] 상기 프로브 또는 프라이머는 포스포르아미다이트 고체 지지체 방법, 또는 기타 널리 공지된 방법을 사용하여 화학적으로 합성할 수 있다. 이러한 핵산 서열은 또한 당해 분야에 공지된 많은 수단을 이용하여 변형시킬 수 있다. 이러한 변형의 비-제한적인 예로는 메틸화, 캡핑(capping), 천연 뉴클레오타이드 하나 이상의 동족체로의 치환, 및 뉴클레오타이드 간의 변형, 예를 들면, 하전되지 않은 연결체(예: 메틸 포스포네이트, 포스포트리에스테르, 포스포로아미데이트, 카바메이트 등) 또는 하전된 연결체(예: 포스포로티오에이트, 포스포로디티오에이트 등)로의 변형이 있다.
- [0031] 상기 "위암(gastric cancer)"은 위에 생기는 암을 말한다. 상기 위암은 선암(adenocarcinoma), 위 림프종(lymphoma), 위 간질성 종양(gastrointestinal tumor), 육종(sarcoma), 및 신경내분비암(neuroendocrine tumor)으로 이루어진 군으로부터 선택될 수 있다.
- [0032] 상기 "발병(develop 또는 onset)"은 병의 시작을 말한다. 상기 발병은 복통이나 열감 등 환자 자신의 자각적 증상(증후)나 제3자, 특히 의사에 의해 발견된 타각적 증상(징후)으로 시작되는 경우에는 증상이 없이 각종 임상 검사에 의해서 처음으로 병의 시작을 아는 경우도 포함한다. 병에 따라서는(특히 종양), 상당히 병이 진전된 후

에 처음으로 발병을 알게 되는 경우도 포함한다.

- [0033] 상기 "위험(risk)"은 질병이나 해로움이 생길 우려가 있거나 그런 상태를 말한다. 상기 위험은 특정 다형성 대립유전자를 보유하지 않은 개체의 구성원에서의 질환 또는 상태의 발생 빈도에 비교하여, 특정 다형성 대립유전자를 보유한 개체에서의 질환 또는 상태의 발생 빈도가 통계적으로 높음을 나타낸다. 상기 위험도는 어떤 특정한 위험이 어떤 기간 내에 발생하는 비율로서, 위험발생률이라고도 한다.
- [0034] 상기 "예측(predict)"은 환자가 위암 발병할 가능성이 있는지를 판별하는 것을 말한다. 예측은 화학요법 또는 방사선 치료 등 치료법에 대해 선호적으로 또는 비선호적으로 반응하여 환자의 치료, 예를 들어 특정 치료제, 및/또는 원발성 종양의 수술로 제거, 및/또는 암 재발 없이 특정 시기 동안 화학요법으로 치료된 후 생존할 여부 및/또는 가능성과 관련된다. 상기 예측 방법은 임의의 환자에 대한 위암 발병 위험도가 높은 환자로서, 특별하고 적절한 관리를 통하여 발병 시기를 늦추거나 발병하지 않도록 하거나 또는 위암 발병 환자에 대한 가장 적절한 치료 방식을 선택함으로써 치료 여부를 결정하기 위해 사용될 수 있다.
- [0036] 다른 양상은 일 양상에 따른 조성물을 포함하는 마이크로어레이를 제공한다.
- [0037] 상기 일 양상에 따른 조성물은 전술한 바와 같다.
- [0038] 용어 "마이크로어레이(microarray)"는 기관 표면의 구분된 영역에 상기 폴리뉴클레오티드가 높은 밀도로 고정화되어 있는 것을 말한다. 상기 마이크로어레이는, 상기 영역이 예를 들면  $400\text{개}/\text{cm}^2$  이상,  $10^3\text{개}/\text{cm}^2$  이상, 또는  $10^4\text{개}/\text{cm}^2$  이상의 밀도로 기관 상에 배열되어 있는 것일 수 있다.
- [0039] 상기 마이크로어레이에 고정된 폴리뉴클레오티드는 DNA 또는 RNA일 수 있다. 상기 폴리뉴클레오티드는 단일 가닥 또는 이중 가닥일 수 있다. 상기 폴리뉴클레오티드는 천연 뉴클레오티드, 천연 뉴클레오티드의 유사체, 천연 뉴클레오티드의 당, 염기 또는 인산 부위가 변형되어 있는 뉴클레오티드, PNA(peptide nucleic acid), 및 이들의 조합을 포함할 수 있다. 상기 폴리뉴클레오티드는 검출가능한 표지(예를 들면, Cy3 또는 Cy5 형광성 물질)가 그의 3'-말단, 중간, 또는 5'-말단에 부착된 것일 수 있다.
- [0040] 상기 마이크로어레이에 고정되는 폴리뉴클레오티드는 길이가 약 5 내지 약 100 nt, 약 10 내지 약 80 nt, 약 10 내지 약 60 nt, 약 10 내지 약 50 nt, 약 10 내지 약 40 nt, 약 10 내지 약 30 nt, 또는 약 10 내지 약 20 nt 일 수 있다.
- [0042] 다른 양상은 일 양상에 따른 조성물을 포함하는 키트를 제공한다.
- [0043] 상기 일 양상에 따른 조성물은 전술한 바와 같다.
- [0044] 상기 키트는 PCR 키트 또는 DNA 칩 키트일 수 있으나, 이에 제한되지 않는다.
- [0045] 상기 키트는 필요에 따라 다른 구성성분 조성물(예를 들면 dNTP, 증합효소 및 발색제 등) 또는 장치를 더 포함할 수 있다. 상기 키트는 PCR을 수행하기 위해 필요한 필수 요소를 포함하는 키트일 수 있다. 상기 PCR 키트는 상기 위암 발병 예측용 조성물 외에도 테스트 튜브 또는 다른 적절한 컨테이너, 반응 완충액(pH 및 마그네슘 농도는 다양), 데옥시뉴클레오타이드(dNTPs), Taq-폴리머라아제 및 역전사 효소와 같은 효소, DNase, RNase 억제제, DEPC-수(DEPC-water), 멸균수 등을 포함할 수 있다. 또한 정량 대조군으로 사용되는 유전자에 특이적인 프라이머 쌍을 포함할 수 있다.
- [0046] 상기 키트는 DNA 칩으로 이용하기 위해 필요한 필수 요소를 포함할 수 있다. DNA 칩 키트는, 일반적으로 편평한 고체 지지판, 전형적으로는 현미경용 슬라이드보다 크기 않은 유리 표면에 핵산 종을 격자형 배열(grided array)로 부착한 것으로, 칩 표면에 핵산이 일정하게 배열되어, DNA 칩 상의 핵산과 칩 표면에 처리된 용액 내에 포함된 상보적인 핵산 간에 다중 혼성화(hybridization) 반응이 일어나 대량 병렬 분석이 가능하도록 하는 도구이다. DNA 칩 키트는 유전자 또는 그의 단편에 해당하는 cDNA 또는 올리고뉴클레오티드(oligonucleotide)가 부착되어 있는 기관, 및 형광표지 프로브를 제작하기 위한 시약, 제제, 효소 등을 포함할 수 있다. 또한 기관은 대조군 유전자 또는 그의 단편에 해당하는 cDNA 또는 올리고뉴클레오티드를 포함할 수 있다.
- [0048] 다른 양상은 개체로부터 수득된 핵산 시료로부터 서열번호 1 내지 서열번호 9로 이루어진 군으로부터 선택된 하나 이상의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위의 유전자형을 검출하는 단계; 및 검출된 유전자형으로부터 개체의 위암 발병 위험도를 산출하는 단계를 포함하는 위암 발병 위험도를 예측하는 방법을 제공한다.



- [0049] 상기 폴리뉴클레오티드, 다형성 부위, 유전자형, 검출, 위암, 발병, 위험도, 및 예측은 전술한 바와 같다.
- [0050] 상기 방법은 개체로부터 수득된 핵산 시료로부터 서열번호 1 내지 서열번호 9로 이루어진 군으로부터 선택된 하나 이상의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위의 유전자형을 검출하는 단계를 포함한다.
- [0051] 상기 방법은 서열번호 10의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위의 유전자형을 검출하는 단계; 및 검출된 유전자형으로부터 개체의 위암 발병 위험도를 산출하는 단계를 더 포함할 수 있다.
- [0052] 상기 방법은 개체의 생물학적 시료로부터 핵산 시료를 수득하는 단계를 더 포함할 수 있다.
- [0053] 상기 개체는 인간을 포함한 포유동물일 수 있다. 상기 개체는 80세 이상, 70세 이상, 60세 이상, 50세 이상, 또는 40세 이상의 인간일 수 있다.
- [0054] 상기 생물학적 시료는 위암 발병 위험도를 평가하고자 하는 상기 개체로부터 수득된 시료를 말한다. 상기 생물학적 시료는 예를 들면 조직, 소변, 점액, 타액, 눈물, 혈액, 혈장, 혈청, 객담, 척수액, 흉수, 유두 흡인물, 림프액, 기도액, 장액, 비뇨생식관액, 모유, 림프계 체액, 정액, 뇌척수액, 기관계내 체액, 복수, 낭성 종양 체액, 양수액 또는 이들의 조합일 수 있다. 상기 생물학적 시료로부터 핵산 시료를 수득하는 단계는 통상의 DNA 분리방법에 의하여 수행될 수 있다. 예를 들어, 표적 핵산을 증합효소 연쇄 반응(polymerase chain reaction: PCR), 리가제 연쇄 반응(ligase chain reaction: LCR), 전사 증폭(transcription amplification) 또는 실시간-핵산 서열 기초 증폭(real time-nucleic acid sequence based amplification: NASBA)을 통하여 증폭하고 이를 정제하여 얻을 수 있다. 또한, 분지형 DNA(bDNA) 프로브의 사용이 증폭을 대체할 수 있다.
- [0055] 상기 유전자형을 검출하는 단계는 서열번호 1 내지 9로 이루어진 군으로부터 선택되는 하나 이상의 핵산 서열로 이루어진 폴리뉴클레오티드에서 상기 다형성부위를 포함하는 5개 이상의 연속 핵산 서열과 동일하거나 또는 상보적인 핵산 서열을 포함하는 프라이머 또는 프로브를 사용하여 수행될 수 있다. 상기 유전자형을 검출하는 단계는 서열번호 10의 핵산 서열로 이루어진 폴리뉴클레오티드에서 상기 다형성부위를 포함하는 5개 이상의 연속 핵산 서열과 동일하거나 또는 상보적인 뉴클레오티드 서열을 포함하는 프라이머 또는 프로브를 사용하여 수행되는 단계를 더 포함할 수 있다.
- [0056] 상기 유전자형을 검출하는 단계는 서열번호 1 내지 서열번호 9로 이루어진 군으로부터 선택된 하나 이상의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 이상의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드가 고정된 마이크로어레이에 상기 핵산 시료를 혼성화하여 수행될 수 있다. 상기 유전자형을 검출하는 단계는 서열번호 10의 핵산 서열로 이루어진 폴리뉴클레오티드에서 그의 5'-말단으로부터 51번째 뉴클레오티드의 다형성 부위를 포함하는 5개 이상의 연속 핵산 서열과 동일하거나 또는 상보적인 폴리뉴클레오티드가 고정된 마이크로어레이에 상기 핵산 시료를 혼성화하여 수행되는 단계를 더 포함할 수 있다.
- [0057] 상기 다형성 부위, 프라이머, 프로브, 및 마이크로어레이는 전술한 바와 같다.
- [0058] 상기 유전자형을 검출하는 단계는 증합효소 연쇄 반응(polymerase chain reaction: PCR), 분자 비이콘(molecular beacon), 프라이머 신장법(primer extension) 등으로 수행될 수 있다. 구체적으로, 상기 유전자형을 검출은 상기 프라이머 또는 프로브가 고정된 마이크로어레이에 상기 핵산 시료를 혼성화시키고, 혼성화 결과를 분석함으로써 수행될 수 있다. 이때, 마이크로어레이 상에서의 핵산의 혼성화 및 혼성화 결과의 검출은 당업계에 잘 알려져 있다. 상기 검출은 예를 들면, 핵산 시료를 형광 물질 예를 들면 Cy3 및 Cy5와 같은 물질을 포함하는 검출가능한 신호를 발생시킬 수 있는 표지 물질로 표지한 다음, 마이크로어레이 상에 혼성화하고 상기 표지 물질로부터 발생하는 신호를 검출함으로써 혼성화 결과를 검출할 수 있다.
- [0059] 상기 방법은 검출된 유전자형으로부터 개체의 위암 발병 위험도를 산출하는 단계를 포함한다.
- [0060] 상기 위암 발병 위험도를 산출하는 단계는 다형성 부위의 유전자형에 유전자형 점수를 부여하는 단계를 포함할 수 있다. 상기 유전자형 점수는 1 내지 3의 정수일 수 있다. 예를 들어, 서열번호 3의 폴리뉴클레오티드의 다형성 부위에 대하여 유전자형 GG, GT 및 GA에 각각  $\alpha_1$ ,  $\alpha_2$  및  $\alpha_3$ 의 값을 부여할 수 있다. 상기  $\alpha_1$ ,  $\alpha_2$  및  $\alpha_3$ 의 값은 1 내지 3의 정수일 수 있다. 예를 들어, 유전자형 GG의 점수는 1, 유전자형 GA의 점수는 2, 및 GT의 유전자형 점수는 3일 수 있다.

- [0061] 상기 위암 발병 위험도를 산출하는 단계는 하기 수학식 1에 의해 위암 발병 위험도를 산출하는 단계를 포함할 수 있다:
- [0062] <수학식 1>
- [0063] 위암 발병 위험도 =  $b + \sum(\text{유전자형 점수} \times a)$
- [0064] 상기 수학식 1에서, 상기 유전자형 점수는 개체에서 분석된 폴리뉴클레오티드의 다형성 부위의 유전자형에 부여된 점수로서, 서열번호 1 내지 서열번호 9로 이루어진 군으로부터 선택된 하나 이상의 핵산 서열로 이루어진 폴리뉴클레오티드의 다형성의 유전자형에 따라 부여된 유전자형 점수일 수 있다. 상기 a는 -0.3 내지 0.5의 값이고, 및 상기 b는 -3.0 내지 3.0의 값일 수 있다.
- [0065] 상기 위암 발병 위험도를 산출 시, 서열번호 10의 핵산 서열로 이루어진 폴리뉴클레오티드의 다형성의 유전자형에 따라 부여된 유전자형 점수가 더 추가되어 산출될 수 있다.
- [0066] 상기 수학식 1에서, 상기 유전자형 점수는 전술한 바와 동일하다.
- [0067] 상기 a는 볼트-선형 혼합 회귀 모델(BOLT-linear mixed model)에서 위암군과 정상군의 유전자형을 분석하여 산출된 상수일 수 있다. 상기 정상군은 위암으로 진단되지 않은 대조군일 수 있다. 상기 a는 볼트-선형 혼합 회귀 모델에서 산출된 효과 크기(effect size) 값일 수 있다. 상기 a는 약 -0.3 내지 약 0.5, 약 -0.2 내지 약 0.5, 약 -0.1 내지 약 0.5, 약 0 내지 약 0.5, 약 0.1 내지 약 0.5, 약 0.2 내지 약 0.5, 약 0.3 내지 약 0.5, 약 0.4 내지 약 0.5, 약 -0.3 내지 약 0.4, 약 -0.3 내지 약 0.3, 약 -0.3 내지 약 0.2, 약 -0.3 내지 약 0.1, 약 -0.3 내지 약 0, 약 -0.3 내지 약 -0.1, 또는 약 -0.3 내지 약 -0.2의 값일 수 있다.
- [0068] 상기 b는 볼트-선형 혼합 회귀 모델에서 위암군과 정상군 간의 유전자형을 분석하여 산출된 상수일 수 있다. 상기 b는 볼트-선형 혼합 회귀 모델에서 산출된 회귀계수( 또는 y 절편) 값일 수 있다. 상기 b는 약 -3.0 내지 약 3.0, 약 -3.0 내지 약 2.0, 약 -3.0 내지 약 1.0, 약 -3.0 내지 약 0, 약 -3.0 내지 약 -1.0, 약 -3.0 내지 약 -2.0, 약 -2.0 내지 약 3.0, 약 -1.0 내지 약 3.0, 약 0 내지 약 3.0, 약 1.0 내지 약 3.0, 또는 약 2.0 내지 약 3.0의 값일 수 있다.
- [0069] 상기 방법은 산출된 위암 발병 위험도에 따라 개체의 위암 발병 위험도를 양호 및 위험 중 어느 한 등급으로 결정하는 단계를 더 포함할 수 있다. 상기 방법은 위험 점수를 산출하고, 산출된 값을 정규화할 수 있다. 정규화된 위험 점수를 높은 순으로 나열하여 3:7로 나누고, 상위 30%를 위험 및 하위 70%를 양호로 결정하는 단계를 더 포함할 수 있다.
- [0070] 상기 위암 발병 위험도 예측 방법은 80세 이상, 80세 미만, 70세 미만, 60세 미만, 50세 미만, 또는 40세 미만의 개체에서 수득한 시료를 이용하는 것일 수 있다. 본 발명의 일 구체예에서 상기 10개의 위암 특이적 SNP를 이용하여 산출한 위암 발병 위험도 결과와 rs377176971 및 rs143472300를 이용하여 산출한 위암 발병 위험도 결과에서 모두 60세 미만 그룹에서 보다 더 높은 예측 정확도를 나타내었다(도 7 내지 도 10). 상기 결과를 통해, 일 양상에 따른 방법은 60세 미만의 개체에서 실시하는 것이 가장 정확한 예측 결과를 도출할 수 있음을 예상할 수 있다.

### 발명의 효과

- [0071] 일 양상에 따른 조성물, 마이크로어레이, 키트 및 이를 이용하여 개체의 위암 발병 위험도를 예측하는 방법 또는 위암 발병 위험도 예측에 필요한 정보를 제공하기 위하여 개체의 SNP를 분석하는 방법에 의하면, 개체의 위암 발병 위험도를 정확하고 간편하게 예측할 수 있다. 따라서, 본 발명은 위암 발병을 예방하는데 도움을 줄 수 있다.

### 도면의 간단한 설명

- [0072] 도 1은 전장 유전체 연관성 분석(GWAS) 결과를 맨해튼 플롯(manhattan plot)으로 나타낸 그래프이다.
- 도 2는 전장 유전체 연관성 분석(GWAS) 결과 중 위암 발병과 상관성이 높은 상위 10개의 신규한 위암 SNP 마커를 선별한 결과를 나타낸 도면이다.
- 도 3은 10개의 선별된 위암 SNP 마커에 대해 회귀분석을 통한 위험 점수를 산출하는 방식에 따른 위암 발병 연관성 및 예측 정확도를 평가한 결과를 나타낸 도면이다.

도 4는 10개의 선별된 위암 SNP 마커에 대해 회귀분석을 통한 연령별(60대, 70대, 및 80대)의 위암 발병 연관성 및 위험 점수의 예측 정확도를 평가한 결과를 나타낸 도면이다.

도 5는 10개의 선별된 위암 SNP 마커 중 상위 2개(rs377176971 또는 GC-0001, 및 rs143172300 또는 GC-0002)의 SNP 마커의 유전자 변이에 대해 회귀분석을 통한 위암 발병 연관성을 확인한 결과를 나타낸 도면이다.

도 6은 10개의 선별된 위암 SNP 마커 중 상위 2개(rs377176971 또는 GC-0001, rs143172300 또는 GC-0002)의 SNP 마커의 유전자 변이 조합별에 회귀분석을 통한 위암 발병 연관성을 확인한 결과를 나타낸 도면이다.

도 7은 유전자 위험 점수 산출식에 따라 분류된 유전자 위험도 그룹(risk) 및 유전자 비위험도 그룹(normal)에서 위암 발병률(상단) 및 연령별(60대, 70대, 및 80대)의 위암 발병률을 분석한 결과를 나타낸 도면(하단)이다.

도 8은 유전자 위험 점수 공식에 따라 분류된 유전자 위험도 그룹(risk) 및 유전자 비위험도 그룹(normal)에서 연령별(60대, 70대, 및 80대)의 위암 발병률 분석에 대한 예측 성능 검증을 실시한 결과를 나타낸 도면이다.

도 9는 유전자 위험 점수 공식에 따라 분류된 유전자 위험도 그룹(risk) 및 유전자 비위험도 그룹(normal)에서 연령별(60대, 70대, 80대)의 위암 발병률을 나타낸 그래프(우측) 및 도면이다(중간 및 좌측).

도 10은 유전자 위험 점수 공식에 따라 분류된 유전자 위험도 그룹(risk) 및 유전자 비위험도 그룹(normal)에서 선별된 위암 SNP인 rs377176971(GC-0001)의 유전자 변이별 위암 발병률을 분석한 결과를 나타낸 그래프(우측) 및 도면(좌측)이다.

도 11은 유전자 위험 점수 공식에 따라 분류된 유전자 위험도 그룹 및 유전자 비위험도 그룹에서 선별된 위암 SNP인 rs143472300(GC-0002)의 유전자 변이별 위암 발병률을 분석한 결과를 나타낸 그래프(우측) 및 도면(좌측)이다.

### 발명을 실시하기 위한 구체적인 내용

[0073] 이하, 실시예를 통하여 본 발명을 더욱 상세히 설명하고자 한다. 이들 실시예는 오로지 본 발명을 예시하기 위한 것으로서, 본 발명의 범위가 이들 실시예에 의해 제한되는 것으로 해석되지 않는 것은 당업계에서 통상의 지식을 가진 자에 있어서 자명할 것이다.

#### [0075] 실시예 1. 위암 특이적 SNP 탐색

[0076] 한국인 약 51,000명 코호트에 대한 한국인 칩(Korean chip v1.1\_SNP microarray chip, 자체 제조)을 통하여 유전자 검사를 수행하고 유전체 데이터를 구축하였다. 구축한 유전체 데이터를 위암군과 정상군으로 구분하여 전장 유전체 연관성 분석(GWAS)을 수행하였다(도 1). 이때, 연관성 변이를 구축하는 임계값은  $p\text{-값} < 5 \times 10^{-8}$ 으로 설정하였고, The R Project for Statistical Computing (<https://www.r-project.org/>)을 이용하여 분석을 실시하였다. 그 결과, 위암 전체 결과에서 최적의 위험도 예측을 위한 SNP를 총 636개 선별하였다.

[0077] 선별된 유전자 변이들에 대한 게놈 윈도우(genomic window)를  $\pm 250$  KB로 설정하고 LD block은  $R^2 > 0.2$ 로 설정하여 크럼핑(clumping) 또는 프루닝(pruning)을 진행하였다. 크럼핑은 각 연관비평형(Linkage disequilibrium: LD) 블록에서 가장 중요한 SNP를 식별하고 선택하기 위한 추가 과정을 진행하는 방법을 말하고, 프루닝은 대략적인 연관평형(linkage equilibrium: LE) 관계에 있는 마커의 부분 집합을 선별하는 방법을 말한다. 크럼핑 또는 프루닝된 636개의 SNP 중 데이터에서 머신 러닝(machine learning) 방법을 통하여 위암에 대한 중요도를 구하여 순위를 선정하였다. 상기 선별된 마커 중 강한 신호를 갖고, 게놈 윈도우 당 중복되지 않는 바이오마커들을 선별하여 총 10개의 SNP 바이오마커를 발굴하였다(표 1 및 도 2)(CHR: 염색체 번호, BP: 염색체 중 위치, A1: 대립형질(minor allele), MAF: 대립유전자형빈도(minor allele frequency), OR: 승산비(odd ratio), CI: 신뢰구간(confidence interval)).

표 1

SNP	서열	서열번호
rs377176971	CATTTGTTTTGTCCATTGATTCCTTAGGATAGTGAATGAGCAGCTACAG [C/T] GGTCAGTTGATGACTATCAGCACCGACTTTCCATAAAAAAGAGGTGAACCTT	1
rs78381605	AGATTATTTCACCCCTAAGAAAGAATAAAAAACACCATACCCCTAAACTT [G/A] CGTGTCTCAATAAAATAATAGAGTTATTAACCTAACTGGTTGTGTCCAA	2
rs145516817	AAATGGTGACAGGAACATTTTAAAAAGAAACATGTATTATACCTACACAG [G/T] GTAGTTAGCTCAGTGTCTTTTACTTTAGGTGCTATAAAATACGGAAGAAA	3

rs114167943	CGAGGTGGCCGCTCCACTGGAGGGCGCTGGTCGGGGGCTGACGGCCGGCG CGCGCTGCTGGTGTCTGCTGCTGCTGGCGCTGGGCGTCCGGGCGGTGCGCT	[C/T]	4
rs150274610	ACAAATAAAACAGGAAAAACAGAGTTATCATTTACCGAGATGAGCACAAAC AAGAACGGCTGCCCTGAGGGACAGAGTAACAAAAGATCAGTTTAGCTTG	[C/A]	5
rs10674898	ATTTGGGTGGGGACACAGATCCAACTATATCACAACTGTAAAAA GACAAAAGAAAGCAAAGAAAGGAAAGAGTGTGAAATTCCTGCTCCTG	[AAAA/AA]	6
rs143172300	CGTGGAGCGCCAGGGCGTCCGACCTCTGCACCTGAGAGAAGATGAACACG CCGACCAGGCCCGGGTGGGGCCCGCGGACGACGGGCCTGCGCCGTCTGGG	[G/A]	7
rs373412232	TGATTCTGAGATAGCTGAATCTGTCTCTGTGACCCACAGCTGACAAT GGAAAAAGCAGAGGGTCCCGCAACGGGAGCGGAAGCGGCATGACTTCTCC	[C/T]	8
rs2470155	CCTAAACCATCATCCACTTTACTTTCTCATCTGTAGAAATTTCTAGGGA GGAAAAAGCAGACTAGATGGTTTCTAAGGACTCTTCTAGCTTGGCCATT	[G/T]	9
rs760077	CTGCCGCGCCTTCAGGGGTTTCGGGGCTCCACTTGGACGAGGCGCCGTGAC CTCCGAGGCGCGCCGGCCGACAGCGCTGTCCCGCTACGTGGGCCACCTC	[T/A]	10

[0080] **실시예 2. 선별된 위암 특이적 SNP 마커의 유의성 검증**

[0081] 상기 실시예 1의 방법으로 선별된 10개의 SNP를 토대로 The R Project for Statistical Computing (<https://www.r-project.org/>)을 이용하여 실시예 1에서 상기 SNP 선별에 사용된 유전자 샘플을 제공한 위암군과 정상군에 대하여, 유전자 위험 점수 산출 방식에 따른 정확성을 평가하였다. 통계적 처리는 총 11가지의 알고리즘을 이용하였으며, 그 결과는 도 3에 나타내었다.

[0082] 도 3에 나타낸 바와 같이 총 11개의 분석 모델 중 boosting 분석 모델을 사용한 경우, AUC(area under curve) 값이 0.63으로 가장 정확하게 위암 발병 위험도를 예측 가능할 수 있었다(PPV: 양성예측도, NPV: 음성예측도).

[0083] 또한, 상기 10개의 SNP를 토대로 볼트-선형 혼합 회귀모델을 이용하여 연령별(60대, 70대, 80대) 층화분석을 실시한 결과, 도 4에 나타낸 바와 같이 60대에서 가장 정확하게 위암 발병도를 예측하였다(AUC=0.64).

[0084] 상기 10개의 SNP 중 분석 결과에서 가장 유의하게 나타난 2개의 SNP(rs377176971 또는 GC-0001(서열번호 1), 및 rs143472300 또는 GC-0002(서열번호 7))에 대하여 위암 발병률 및 유전자 위험 점수의 정확성을 확인하였다. 그 결과, rs377176971(GC-0001, 서열번호 1)의 경우, C가 T로 변이된 경우 위암 발병률이 8.8배 증가하였고, rs143472300(GC-0002, 서열번호 7)의 경우, G가 A로 변이된 경우, 위암 발병률이 8.1배 증가하였음을 확인하였다(도 5). 또한, 상기 2개의 SNP의 유전자 변이의 조합별로 위암 발병률을 예측한 결과, rs377176971의 C 및 rs143472300의 G가 각각 T 및 A로 모두 변이된 경우, 위암 발병률이 3.7배로 증가하여 가장 높았다(도 6).

[0086] **실시예 3. 선별된 위암 특이적 SNP 마커를 이용한 위암 발병 위험도 공식 구축 및 검증**

[0087] 상기 실시예 2에서 가장 높은 예측 정확도를 나타내었던 boosting 예측 모델을 바탕으로 다유전자 위험 점수 산출 방식을 공식화하여 위암 발병 위험도를 예측하는 수학적식을 구축하여 하기 수학적식 2에 나타내었다.

[0088] <수학적식 2>

[0089] 위암 발병 위험도(유전자 위험 점수) =  $-2.651 + [(-0.1608 \times \text{rs760077에서 다형성 부위 A 점수}) + (0.4053 \times \text{rs377176971에서 다형성 부위 T의 점수}) + (0.1038 \times \text{rs78381605에서 다형성 부위 A의 점수}) + (0.1584 \times \text{rs145516817에서 다형성 부위 T의 점수}) + (0.3743 \times \text{rs114167943에서 다형성 부위 T의 점수}) + (0.0309 \times \text{rs150274610에서 다형성 부위 A의 점수}) + (0.1342 \times \text{rs10674898에서 다형성 부위 TAA의 점수}) + (0.3643 \times \text{rs143172300에서 다형성 부위 A의 점수}) + (-0.2912 \times \text{rs373412232에서 다형성 부위 T의 점수}) + (-0.2592 \times \text{rs2470155에서 다형성 부위 T의 점수})]$

[0090] 상기 수학적식 2에서 10개의 SNP에 대해 유전자형에 따라 0, 1 또는 2의 점수를 부여하였다. 각각의 유전자형이 비위험/비위험 인 경우 1점을 부여하였고, 비위험/위험인 경우 2점을 부여하였고, 위험/위험인 경우 3점을 부여하였다.

[0091] 상기 수학적식 2를 토대로 산출한 유전자 위험 점수를 개인에게 부여하고 이를 토대로 유전자 위험도 그룹을 분류하였다.

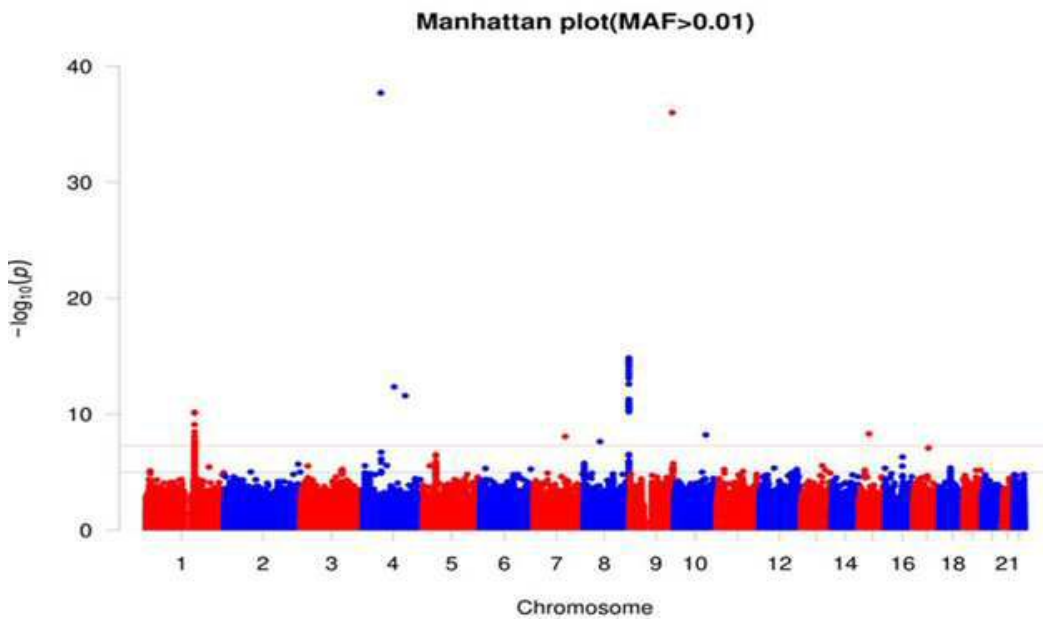
[0092] 상기 유전자 위험도 그룹(risk)을 위암 환자군에서 시간 경과에 따른 위암 발병 위험도 변화량을 제시하는 모델을 확립하고 그 결과를 도 7 내지 도 10에 나타내었다(HR: 위험비(Hazard Ratio), C-index: 일치지수(concordance index)).

[0093]

그 결과, 유전자 비위험도 그룹(normal)에 비하여 유전자 위험도 그룹(risk)로 분류된 군에서 위암 발병률이 높게 예측되었고, 연령별 위암 발병률(도 8) 및 사망 위험도도 증가하였다(도 7). 특히, 상기 10개의 SNP를 사용한 유전자 위험 점수에 따른 위암 발병 위험도 예측 결과와 rs377176971(GC-0001) 및 rs143472300(GC-0002)를 사용한 유전자 위험 점수에 따른 위암 발병 위험도 예측 결과(도 10 및 11)에서 모두 60세 미만 그룹에서 보다 더 높은 예측 정확도를 나타내었다. 상기 결과를 통해, 위암의 경우 50세 내지 60세에서의 유전자 검사를 통한 예측 정확도가 가장 높을 것으로 예측된다.

도면

도면1





도면2

SNP	CHR	BP	A1	MAF	OR (95% CI)	P
rs377176971	4	56,865,490	T	0.0185	8.80(7.44-10.41)	2.00E-38 (서열번호 1)
rs78381605	4	99,477,699	A	0.0463	2.18(1.95-2.42)	4.30E-13 (서열번호 2)
rs145516817	4	134,816,182	T	0.0175	3.32(2.79-3.93)	2.50E-12 (서열번호 3)
rs114167943	7	100,609,631	T	0.0128	3.17(2.59-3.87)	8.30E-09 (서열번호 4)
rs150274610	8	52,042,955	A	0.0489	1.77(1.60-1.96)	2.30E-08 (서열번호 5)
rs10674898	8	143,780,261	TAA	0.3788	1.46(1.39-1.53)	1.40E-15 (서열번호 6)
rs143172300	9	135,545,995	A	0.0189	8.15(6.90-9.61)	1.00E-36 (서열번호 7)
rs373412232	10	101,120,600	T	0.0211	0.40(0.34-0.47)	6.00E-09 (서열번호 8)
rs2470155	15	51,590,753	T	0.0243	0.42(0.37-0.49)	4.90E-09 (서열번호 9)
rs760077	1	155,178,782	A	0.1236	0.64(0.60-0.69)	6.60E-11 (서열번호 10)

도면3

Model		AUC	민감도	특이도	PPV	NPV	예측 정확도
GRS	추정값	0.5721	0.9486	0.1353	0.6757	0.4600	0.6614
	95% 신뢰구간	0.5144-0.6298	0.9345-0.9627	0.0893-0.1813	0.6566-0.6948	0.3781-0.5420	0.6406-0.6817
	추정값	0.6243	0.9585	0.0894	0.4913	0.700	0.5051
GBLUP	95% 신뢰구간	0.5694-0.6792	0.9442-0.9727	0.0699-0.1089	0.4660-0.5165	0.6130-0.7871	0.4804-0.5288
	추정값	0.5721	0.9486	0.1353	0.6757	0.4600	0.6614
	95% 신뢰구간	0.5144-0.6298	0.9345-0.9627	0.0893-0.1813	0.6566-0.6948	0.3781-0.5420	0.6406-0.6817
SBLUP	추정값	0.6276	0.9573	0.0902	0.5186	0.6735	0.5288
	95% 신뢰구간	0.5738-0.6814	0.9434-0.9712	0.0702-0.1102	0.4932-0.5440	0.5843-0.7627	0.5040-0.5536
Lasso	추정값	0.6233	0.9566	0.0929	0.5233	0.6613	0.5324
	95% 신뢰구간	0.5696-0.6770	0.9424-0.9707	0.0716-0.1141	0.4986-0.5480	0.5730-0.7496	0.5080-0.5567
	추정값	0.6319	0.9573	0.0921	0.5292	0.6687	0.5385
Ridge	95% 신뢰구간	0.5778-0.6880	0.9434-0.9712	0.0716-0.1126	0.5040-0.5544	0.5790-0.7584	0.5137-0.5631
	추정값	0.6301	0.9570	0.0909	0.5280	0.6659	0.5371
	95% 신뢰구간	0.5764-0.6838	0.9432-0.9709	0.0706-0.1112	0.5027-0.5534	0.5761-0.7556	0.5123-0.5618
Lasso + boosting	추정값	0.5272	0.9374	0.2008	0.9777	0.0788	0.9183
	95% 신뢰구간	0.5014-0.5530	0.9254-0.9495	0.0773-0.3242	0.9703-0.9852	0.0282-0.1294	0.9037-0.9312
	추정값	0.5104	-	0.2059	-	-	0.7547
Randomforest	95% 신뢰구간	0.4961-0.5247	-	0.0315-0.3804	-	-	0.7413-0.7672
	추정값	0.6273	0.9561	0.0910	0.5367	0.6534	0.5444
	95% 신뢰구간	0.5732-0.6814	0.9423-0.9701	0.0705-0.1115	0.5114-0.5620	0.5627-0.7441	0.5196-0.5691
XGboost	추정값	0.6299	0.9565	0.0913	0.5377	0.6553	0.5454
	95% 신뢰구간	0.5757-0.6841	0.9427-0.9703	0.0708-0.1118	0.5123-0.5630	0.5649-0.7459	0.5206-0.5701
	95% 신뢰구간	0.5757-0.6841	0.9427-0.9703	0.0708-0.1118	0.5123-0.5630	0.5649-0.7459	0.5206-0.5701

risk score  
예측모형

도면4

Group	AUC	민감도	특이도	PPV	NPV	예측 정확도
60y's group	추정값	0.6486	0.9551	0.1163	0.6018	0.6246
	95% 신뢰구간	0.5829-0.7142	0.9384-0.9717	0.0830-0.1496	0.5724-0.6313	0.5200-0.7292
70y's group	추정값	0.6444	0.9567	0.1089	0.5927	0.6269
	95% 신뢰구간	0.5880-0.7006	0.9425-0.9709	0.0814-0.1363	0.5671-0.6183	0.5348-0.7191
80y's group	추정값	0.6335	0.9552	0.1014	0.5948	0.602
	95% 신뢰구간	0.5799-0.6872	0.9416-0.9688	0.076-0.1269	0.5707-0.619	0.5141-0.6899

도면5

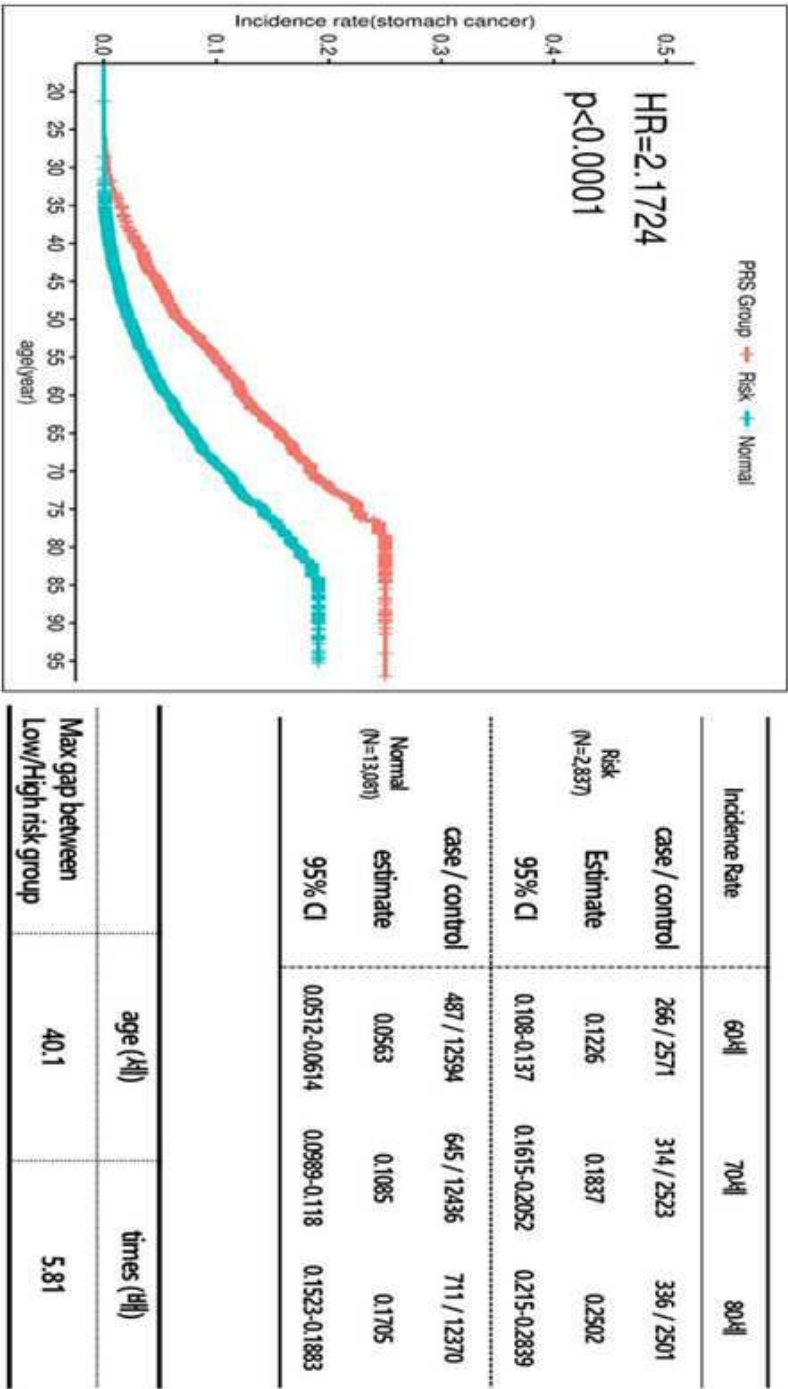
SNP	Control (n=14,865)	Case (n=1,053)	OR (95% CI)	p-value
rs <sup>***</sup> (BAS-GC-0001)				
CC	14,398 (96.9)	932 (88.5)	1.000 (ref.)	
CT	467 (3.1)	121 (11.5)	8.80 (7.44 - 10.41)	2.0e-38
rs <sup>***</sup> (BAS-GC-0002)				
GG	14382 (96.8)	933 (88.6)	1.000 (ref.)	
GA	483 (3.2)	120 (11.4)	8.15 (6.90 - 9.61)	1.0e-36
rs <sup>***</sup> (BAS-GC-0001)				
⇒ Consequence : Missense variant				
⇒ Gene : CEP135				
⇒ Gene Ontology (GO) : GO:0005515 & GO:0008022				
⇒ Pathways : Anchoring of the basal body to the plasma membrane AURKA Activation by TPX2 Cell Cycle, Mitotic Centrosome maturation				
⇒ Previous study : breast cancer / cancer cell phosphoproteome				
rs <sup>***</sup> (BAS-GC-0002)				
⇒ Consequence : Missense variant				
⇒ Gene : GTF3C4				
⇒ Gene Ontology (GO) : GO:0003677 & GO:0004402 & GO:0005515 GO:0008047 & GO:0016740				
⇒ Pathways : Gene Expression RNA Polymerase I RNA Polymerase III Mitochondrial Transcription				
⇒ Previous study : breast cancer / colorectal cancer / gastric cancer cancer cell phosphoproteome				

도면6

Genotype combination	Control (n=14,865)	Case (n=1,053)	OR (95% CI)	p-value
BAS-GC-0001 / 002				
CC - GG	15139 (96.4)	1125 (90.1)	1.000 (ref.)	
CC - GA	76 (0.5)	1 (0.1)	0.177 (0.025-1.274)	0.086
CT - GG	62 (0.4)	2 (0.2)	0.434 (0.106-1.777)	0.246
CT - GA	430 (2.7)	120 (9.6)	3.755 (3.040-4.639)	1.2e-34



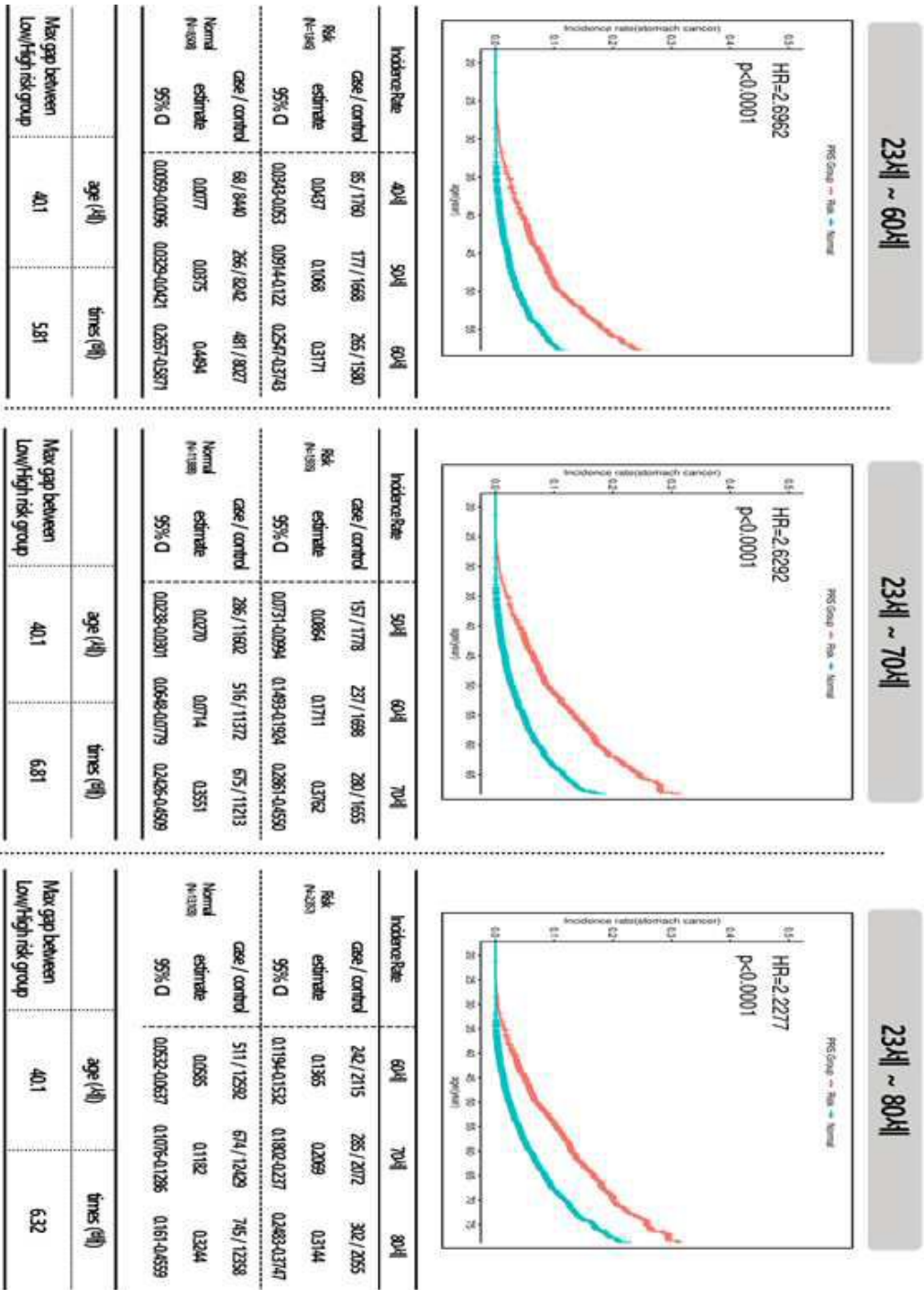
도면7



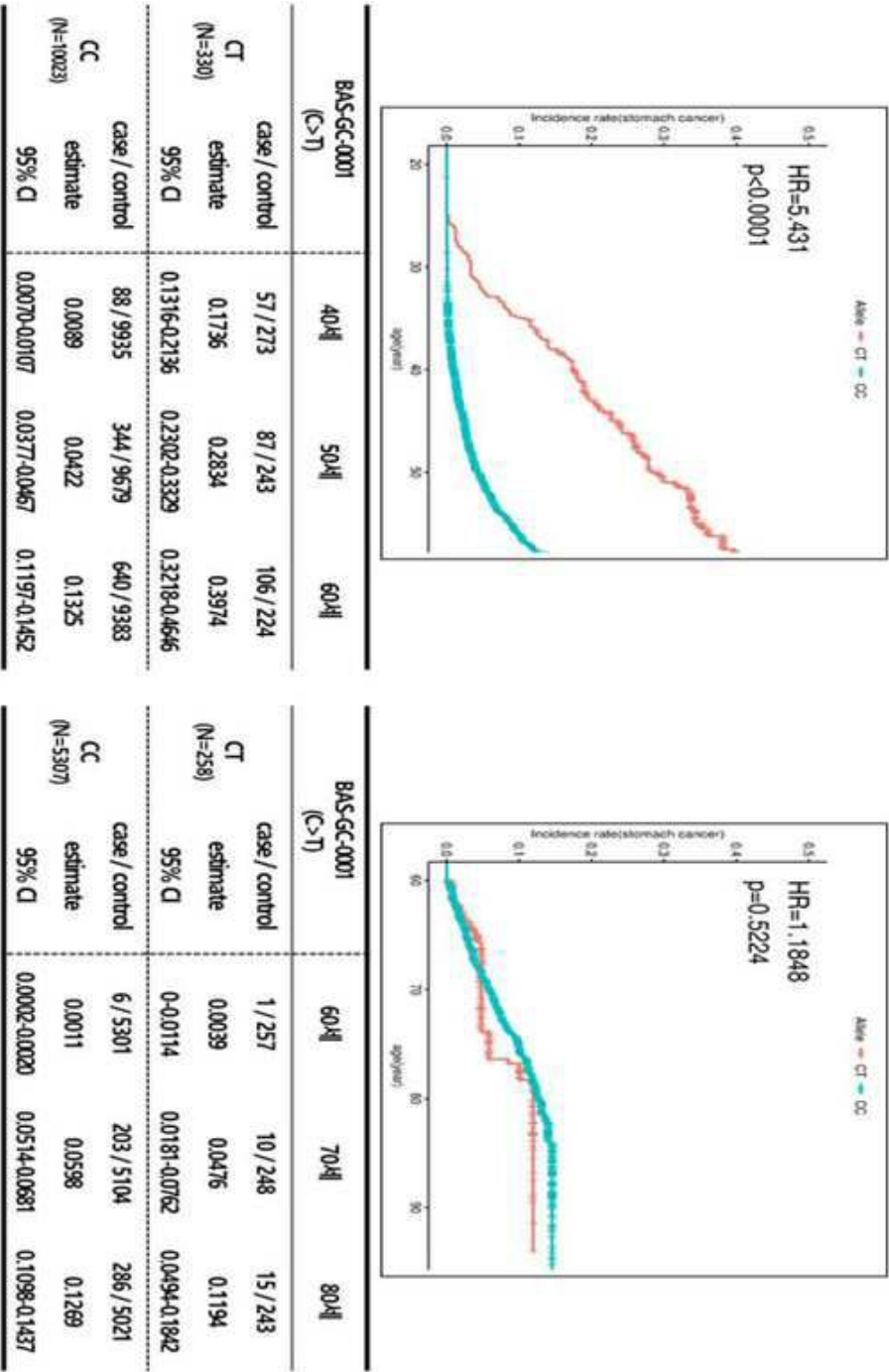
도면8

Group		Result		
60y's group	C-index		0.6642	
	HR	p-value	2.6527	0.0010
70y's group	C-index		0.6529	
	HR	p-value	2.7031	0.0025
80y's group	C-index		0.6472	
	HR	p-value	2.4410	0.0030

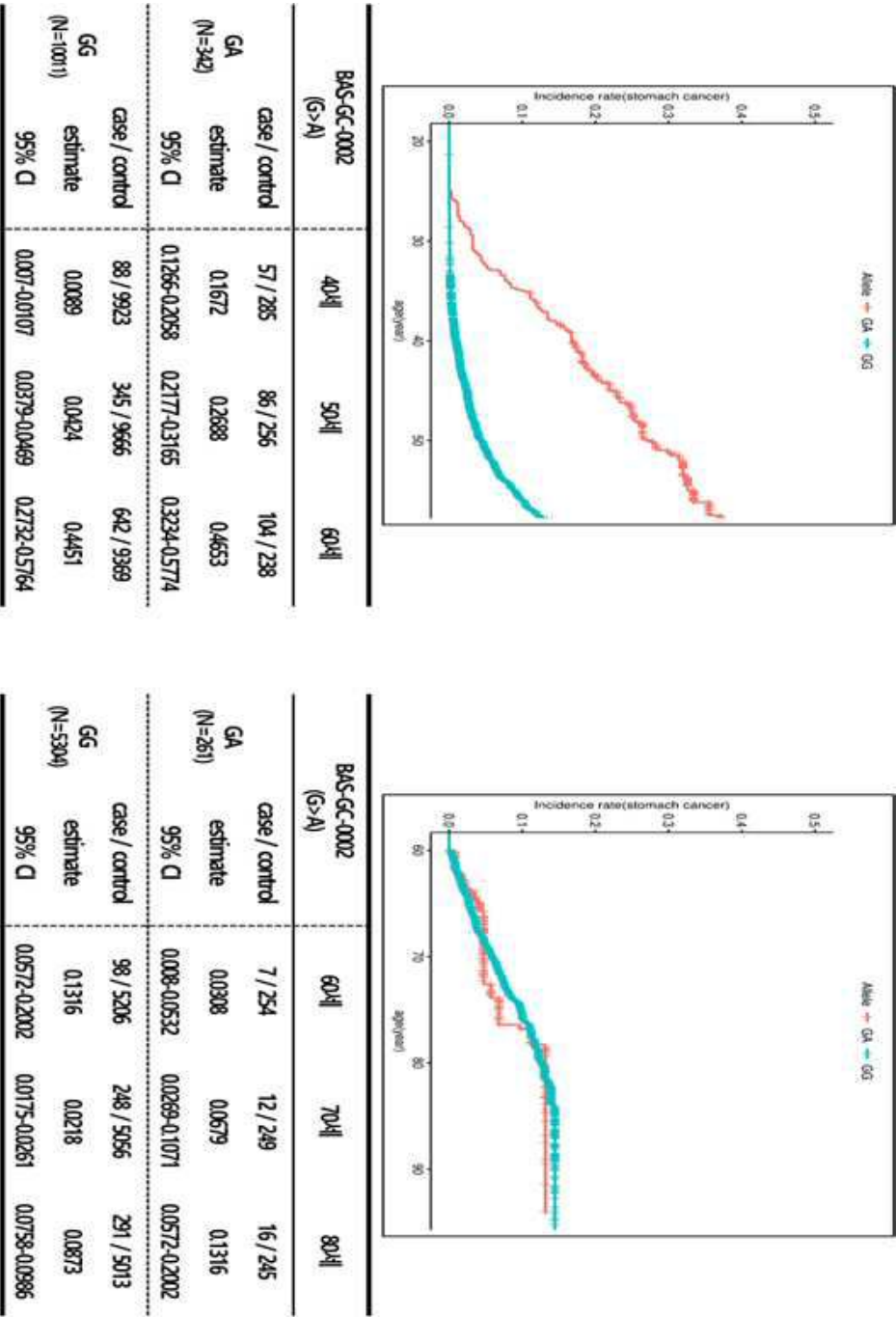
도면9



도면10



도면11



서열 목록

- <110> basgenbio Co., Ltd.
- <120> Composition for predicting a risk of developing gastric cancer and method using the same
- <130> SPD21-088/BSG
- <160> 10
- <170> KoPatentIn 3.0
- <210> 1



<211> 101  
 <212> DNA  
 <213> Artificial Sequence  
 <220><223> rs377176971  
 <220><221> variation  
 <222> (51)  
 <223> C at position 51 is a single nucleotide polymorphism of T.  
 <400> 1  
 catttgtttt tgtccattga ttcttttagga tagtgaatga gcagctacag cggtcagttg 60  
  
 atgactatca gcaccgactt tccataaaaa gaggtgaact t 101  
 <210> 2  
 <211> 101  
 <212> DNA  
 <213> Artificial Sequence  
 <220><223> rs78381605  
 <220><221> variation  
 <222> (51)  
 <223> G at position 51 is a single nucleotide polymorphism of A.  
 <400> 2  
 agattatttc aaccctaaga aagaataaaa aaacaccata ccctaaactt gcgtgttctc 60  
 aataaaataa tagagttatt aaactaactg gttgtgtcca a 101  
 <210> 3  
 <211> 101  
 <212> DNA  
 <213> Artificial Sequence  
  
 <220><223> rs145516817  
 <220><221> variation  
 <222> (51)  
 <223> G at position 51 is a single nucleotide polymorphism of T.  
 <400> 3  
 aaatgggtgac aggaaacatt ttaaaagaaa catgtattaa tcacctacag ggtagttagc 60  
 tcagtgtcat ttactttagg tgctataaaa tacggaagaa a 101  
 <210> 4

<211> 101  
 <212> DNA  
 <213> Artificial Sequence  
 <220><223> rs114167943  
 <220><221> variation  
 <222> (51)  
 <223> C at position 51 is a single nucleotide polymorphism of T.  
  
 <400> 4  
 cgaggtggcc gtccactgga gggcgctggt cgggggcctg acggccggcg ccgcgtgct 60  
 ggtgctgctg ctgctggcgc tgggcgtccg ggcggtgcgc t 101  
 <210> 5  
 <211> 101  
 <212> DNA  
 <213> Artificial Sequence  
 <220><223> rs150274610  
 <220><221> variation  
 <222> (51)  
 <223> C at position 51 is a single nucleotide polymorphism of A.  
 <400> 5  
 acaataaaaa caggaataac agagttatca ttaccgaga tgagcacaac caagaacggc 60  
 ctgccctgag ggacagagta acaaaagatc agtttagctt g 101  
  
 <210> 6  
 <211> 104  
 <212> DNA  
 <213> Artificial Sequence  
 <220><223> rs10674898  
 <220><221> variation  
 <222> (51)..(55)  
 <223> AAAA at positions 51 to 55 is a single nucleotide polymorphism of  
 AA.  
 <400> 6  
 atttgggtgg ggacacagat ccaactata tcacaactgt aaaaaaaaaa aaaagacaaa 60  
 gaagaaagca aagaaaggaa aagagtgtga aattcctgct cctg 104

<210> 7  
 <211> 101  
 <212> DNA  
 <213> Artificial Sequence  
 <220><223> rs143172300

<220><221> variation  
 <222> (51)  
 <223> G at position 51 is a single nucleotide polymorphism of A.  
 <400> 7  
 cgtggagcgc cagggcgtcc gacctctgca cctgagagaa gatgaacacg gccgaccagg 60  
 cccgggtggg gcccgcggaac gacgggcctg cgccgtcttg g 101

<210> 8  
 <211> 101  
 <212> DNA  
 <213> Artificial Sequence  
 <220><223> rs373412232  
 <220><221> variation  
 <222> (51)  
 <223> C at position 51 is a single nucleotide polymorphism of T.  
 <400> 8  
 tgattctgag atagctgtaa tctgtcctct gtgacccac agctgacaat cggaaaaagc 60  
 agagggtccc gcaacgggag cggaagcggc atgacttctc c 101

<210> 9  
 <211> 101  
 <212> DNA  
 <213> Artificial Sequence  
 <220><223> rs2470155  
 <220><221> variation  
 <222> (51)  
 <223> G at position 51 is a single nucleotide polymorphism of T.  
 <400> 9  
 cctaaacat catccacttt actttctcat ctgtagaat tttctaggga gggaaaaagc 60  
 agactagatg gtttctaagg actcttctag cttggccatt t 101

<210> 10  
 <211> 101  
 <212> DNA  
 <213> Artificial Sequence

<220><223> rs760077

<220><221> variation

<222> (51)

<223> T at position 51 is a single nucleotide polymorphism of A.

<400> 10

ctgccgcgcc ttcaggggtt cggggctcca ctggacgag gcgccgtgac tctccgaggc 60

gcgccgggcc gacagcgctg tccgctacg tgggccacct c 101