Original Article

J Sasang Constitut Med 2020;32(2):10-21
https://doi.org/10.7730/JSCM.2020.32.2.10

유전지표를 활용한 사상체질 분류모델

반효정[1] · 이시우[2] · 진희정[3*]

[1]한국한의학연구원 지능화추진팀 선임연구원, [2]한국한의학연구원 미래의학부 책임연구원
[3]한국한의학연구원 지능화추진팀 책임연구원

Abstract

Predictive Models for Sasang Constitution Types Using Genetic Factors

Hyo-Jeong Ban[1] · Siwoo Lee[2] · Hee-Jeong Jin[3*]

[1]*Intellectual Information Team, Korea Institute of Oriental Medicine, Senior Researcher*
[2]*Future Medicine Division, Korea Institute of Oriental Medicine, Principal Researcher*
[3]*Intellectual Information Team, Korea Institute of Oriental Medicine, Principal Researcher*

Objectives

Genome-wide association studies(GWAS) is a useful method to identify genetic associations for various phenotypes. The purpose of this study was to develop predictive models for Sasang constitution types using genetic factors.

Methods

The genotypes of the 1,999 subjects was performed using Axiom Precision Medicine Research Array (PMRA) by Life Technologies. All participants were prescribed Sasang Constitution-specific herbal remedies for the treatment, and showed improvement of original symptoms as confirmed by Korean medicine doctor. The genotypes were imputed by using the IMPUTE program. Association analysis was conducted using a logistic regression model to discover Single Nucleotide Polymorphism (SNP), adjusting for age, sex, and BMI.

Results & Conclusions

We developed models to predict Korean medicine constitution types using identified genectic factors and sex, age, BMI using Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN). Each maximum Area Under the Curve (AUC) of Teaeum, Soeum, Soyang is 0.894, 0.868, 0.767, respectively. Each AUC of the models increased by 6~17% more than that of models except for genetic factors.

By developing the predictive models, we confirmed usefulness of genetic factors related with types. It demonstrates a mechanism for more accurate prediction through genetic factors related with type.

Key Words *: Sasang typology, Sasang constitution, Genetic variant, Genetic factor, Predictive model*

Received April 20, 2020 **Revised** April 20, 2020 **Accepted** April 27, 2020
Corresponding author 진희정
대전시 유성구 유성대로 1672, 한국한의학연구원 지능화추진팀 책임연구원
Tel : +82-42-868-0555, Fax : +82-42-868-9480, E-mail : hjjin@kiom.re.kr

Ⅰ. 緖論

사상체질은 이제마에 의해 제시된 한국의 체질이론인데, 그는 1901년의 저서 『동의수세보원』에서 체질이 부모로부터 자식으로 유전됨을 제시한 바 있다(天稟之已定 固無可論).

실제 임상자료를 토대로 분석한 사상체질의 유전성에 대한 연구는 이 등[1]의 연구에서 처음 이뤄졌는데, 부모의 체질이 같은 경우, 자녀도 같은 체질을 갖는 비율이 매우 높은 것을 확인하였다. 현대 유전연구 방식을 차용한 연구에서, 2007년 쌍둥이 1462명을 대상으로 최초로 세 가지 체질에 대한 유전율을 확인한 바 있고[2], 2009년 연구에서는 101개 가계 593명을 대상으로 이뤄진 가계연구에서 전문가의 체질 진단 값을 활용하여 태음인 55%, 소음인 41%, 소양인 47%의 유전율을 확인한 바 있다[3]. 2018년 1742명의 쌍생아들을 대상으로 태음인은 남자 71%, 여자 81%였으며, 소음인은 남자는 70%, 여자는 71%, 소양인은 남녀 모두 47%로 높은 유전율을 보였다[4]. 이러한 유전성에 대한 근거들은 사상체질을 진단할 때 유전지표를 이용할 수 있는 가능성을 높여줬지만, 이를 뒷받침할 유전지표를 찾는 것은 쉬운 일이 아니었다.

2012년 체질 유전지표 발굴을 위한 GWAS(Genome-Wide Association Studies) 분석에서 1200여 명을 대상으로, 태음인 15개, 소음인 12개, 소양인 17개를 찾았지만, p-value < 0.05로 전통적인 GWAS 결과에 부합하지 못했다[5]. 2015년 5500명의 샘플을 이용해서 FTO 비만유전자의 Single Nucleotide Polymorphism (SNP) 1개를 확보하는 데 그쳐[6], 유전지표를 이용한 체질분류는 어려워 보였다.

사상체질진단도구는 Questionnaire for the Sasang Constitution Classification Ⅱ (QSCC Ⅱ)[7,8], Korea Sasang constitutional diagnostic questionnaire-15 (KS-15)[9], Sasang Digestive Function Inventory (SDFI)[10], Sasang Constitution Analysis Tool (SCAT)[11] 등이 있으며, 전문가시스템을 이용한 분류도구이긴 하나, 응답자의 편향적인 응답이나 사진 촬영 시의 조도, 얼굴의 기울기나 숙임 정도, 방향 등에 따라 획득되는 특징점이 달라지는 문제가 있어 실제 의료 환경에서 활용되기에 여전히 부족하며, 체질현상의 생물학적 근거를 찾기 어렵다는 단점이 있다.

유전지표에 의한 사상체질 분류모델이 만들어진다면, 이러한 점을 극복하고, 사상체질의 과학적 근거확보와 이론의 확장에 더욱 도움이 될 것이다. 본 연구에서는 사상체질처방을 복용한 결과를 토대로, 체질전문가의 진단을 거쳐 체질이 임상적으로 명확히 진단된 대상자 집단을 이용하여, 새로운 유전체 분석 방법을 활용, 사상체질 연관 유전지표를 찾고, 해당 유전지표와 기본적 임상정보를 활용하여 사상체질을 분류하는 모델을 개발함으로써, 사상체질을 기반으로 한 정밀의료 실현의 가능성을 높이고자 한다.

Ⅱ. 硏究方法

1. 연구대상자

본 연구는 한국한의학연구원 한의임상정보은행(Korean medicine Data Center, KDC)[12,13] 자료를 활용하였으며, 보다 명확한 체질진단을 위해 사상체질처방 정보와 전문가의 진단 값이 있는 2008년에서 2013년 사이에 다기관 한방 의료 기관에서 수집한 사상 체질 확진자 3,891명 중에서 20세 이상이고, 동일한 SNP chip 정보가 있는 성인 2,000명을 대상으로 하였다(IRB No. 1-0910/02-001). 분석 시에는 지노타이핑 결측값 1명을 제외한 1,999명을 대상으로 하였다.

2. 연구도구

1) 사상체질진단

사상체질진단은 사상체질과 전문의 및 인정의 등 사상체질 임상경력 5년 이상의 한의사에 의해 판단되었는데, 사상체질약리 기준을 바탕으로 피험자가 사상체질처방을 복약 후 주증상의 호전 정도가 평가된

의무기록을 기반으로 진행되었다[14].

2) 지노타이핑 정보

대상자의 지노타이핑은 써모피셔 사이언티픽(ThermoFisher Scientific)사의 Axiom Asia Precision Medicine Research Array (PMRA)로 생산하였다. Asia PMRA chip에는 ClinVar[15]나 GWAS Catalog[16] 내에 포함되어 있는 질병 관련 희귀 또는 공통 변이를 포함하고, human genome version 19 (build 37) 기준으로 동남아시아인을 포괄할 수 있는 새로운 50,000개의 유전변이를 포함하여, 총 750,000개 이상의 유전마커를 포함하고 있다.

3) 정량지표

체질분류모델에서 대상자의 성별, 나이, Body Mass Index (BMI) 정보를 정량적인 정보로 활용하여, 유전지표와 함께 사용하였다. 체질의 특성을 반영하는 다양한 임상정보들이 있지만, 본 연구에서는 유전지표의 활용도를 확인하기 위해, 최소한의 정량지표 BMI를 사용하고, 성별과 나이에 따라 BMI의 차이를 보정하기 위해서 성별, 나이를 추가하여 사용하였다.

4) 전장유전체 분석

실험에서 사용된 SNP 칩의 경우 결측이 발생할 수 있기 때문에 1000 genome project[17,18]의 아시안 패널을 레퍼런스로 하고, IMPUTE2[19,20] 프로그램을 사용하여 결측치를 추정하였다. 결측치를 추정한 SNP 칩을 대상으로 Plink 툴[21]을 사용하여 태음인, 소음인, 소양인에 대해 전장 유전체 분석(Genome-Wide Association Study, GWAS)을 진행하였다. GWAS 분석은 사상체질 처방을 복약 후 전문가 진단 값을 기준으로 유전형과 연관 분석(logistic regression)을 수행하고, 이 때, 대상자의 성별, 연령, BMI 정보를 사용하여 보정하여, 체질값 자체와 연관된 유전지표만을 선택하였다. PMRA chip 정보가 있는 대상자의 수를 고려하여 전체를 대상으로 분석을 수행하였으며, GWAS 결과의 검증을 위해 별도의 Whole Genome Sequencing (WGS, 30X) 정보가 있는 체질 대상자 120명의 정보를 활용하였다. WGS는 GWAS 분석을 통해 얻어진 체질 연관 SNP들을 검증하기 위해 체질 확진자 중 질환을 가지지 않은 60대 이하의 건강한 소음인 30명, 소양인 30명, 태음인 과체중 30명 정상체중 30명을 대상으로 하였다. 이때, 각 그룹별 남녀의 성비는 1:1로 동일하며, 태음인은 비교적 과체중인 사람이 많아서, 비만에 치우쳐진 유전정보를 얻는 것을 피하고자, 과체중과 정상체중 그룹을 나누어서 생산하였다.

5) 머신러닝 분석

사상체질 분류 모델을 개발하기 위해서 Random Forest(RF), Support Vector machine(SVM), Neural Network (NN) 알고리즘을 활용하였으며, R[22]의 caret[23], randomForest[24], kernlab[25], nnet[26] 라이브러리를 사용하여 '태음 vs. 비태음', '소음 vs. 비소음', '소양 vs. 비소양' 알고리즘으로 각각 개발하였다. Training과 test 데이터는 7:3으로 100번 랜덤하게 생성하여 사용하였다. 또한 분류를 할 때, 분류하고자 하는 데이터의 수가 반반이 아닌 경우, 훈련 데이터 내 비율이 높은 분류 쪽으로 결과를 내놓을 수 있다. 이를 방지하기 위해, 업 샘플링과 다운 샘플링 방법을 추가하여 사용하였다[27].

Ⅲ. 結果

1. 연구대상자의 특성

연구대상자는 전체 1,999명으로 태음인 793명(남자 294명, 여자 499명), 소음인 521명(남자 171명, 여자 350명), 소양인 685명(남자 268명, 여자 417명)이었다. BMI는 평균 23.4±3.3, 연령은 평균 48.6±16.1세로 나타났다 (Table 1).

Table 1. General Characteristic of the Subjects

	TE (n=793)	SE (n=521)	SY (n=685)	Total (n=1999)	F-value	p value* (post-hoc)
Sex						
M	294(37.1%)	171(32.8%)	268(39.1%)	799(36.7)		
F	499(62.9%)	350(67.2%)	417(63.3%)	1266(63.3)		
Age (yr)	50.0±16.9	45.9±16.1	49.0±15.0	48.6±16.1	10.718	<0.001 (SE<TE<SY)
BMI (kg/m^2)	25.3±3.2	21.3±2.5	22.9±2.7	23.4±3.3	347.565	<0.001 (SE<SY<TE)

Data shown as n (%) or means standard ± deviation
TE=Taeeumin, SE=Soeumin, SY=Soyangin
* p value: ANOVA analysis and Scheffe post-hoc

2. 전장유전체 분석 결과

GWAS 수행 후 P< 1.0×10^{-5}을 기준으로 모든 샘플 내에서 genotype missing이 존재하는 SNP을 필터링하였다. WGS 120명을 대상으로 체질별 대상자 모두에게 GWAS 분석을 통해 얻어진 체질 연관 SNP들이 나타나는 것만을 검증된 SNP로 선정하였다. 검증된 SNP들은 소음 25개, 태음은 29개, 소양은 24개였다(Table 2). 선정된 SNP들의 보고된 기능을 알기 위해 dbSNP ID를 이용하여 기능 어노테이션을 수행하였으며 그 결과 연관 질환 클래스를 확인할 수 있었다(Table 3).

Table 2. SNPs Related with Sasang Constitution Type

Type	CHR	POS	SNPID	REF	ALT	OR	P
TE	chr1	204374063	rs113151027	A	G	3.93429	1.22802E-05
	chr2	1037581	rs4246557	C	T	1.35217	4.50512E-05
	chr2	10200316	rs12692398	G	A	1.46411	1.75061E-05
	chr2	101673519	rs62155477	G	A	1.63317	3.38612E-05
	chr3	28752749	rs72911400	G	T	0.570687	1.15514E-05
	chr3	85449656	rs9849542	A	G	1.42587	1.24138E-06
	chr3	97692957	rs854581	A	G	0.645714	4.16072E-05
	chr3	193733621	rs150669894	C	T	0.152256	1.05398E-05
	chr4	172507993	rs11940312	C	A	3.67013	4.24202E-05
	chr5	17404483	rs2173759	A	G	1.74676	1.69676E-05
	chr6	42230031	rs73424328	C	T	1.97923	1.91024E-05
	chr6	152949395	rs769918744	C	T	3.54353	0.000046713
	chr7	17644035	rs76717104	C	T	0.694173	3.12025E-05
	chr7	22098226	rs1419772	C	T	3.28152	2.08641E-05
	chr7	126933202	rs62468909	T	C	2.06984	1.48508E-05
	chr10	16575093	rs11254064	A	G	1.71534	2.35844E-05
	chr10	18386599	rs117738954	C	A	2.13361	3.85345E-05
	chr10	79237895	rs582273	A	G	2.80578	6.32739E-06
	chr11	13876119	rs4756771	G	T	0.735872	4.97711E-05
	chr12	21727209	rs201157731	G	A	2.53033	2.14646E-05
	chr12	26629370	rs74798154	C	T	0.343967	1.01336E-05
	chr13	101524245	rs9518188	C	T	4.15277	1.09527E-05
	chr14	66331900	rs55861984	G	A	0.710151	0.000029203
	chr16	24310100	rs2158245	A	G	4.65659	5.24399E-06
	chr16	73439355	rs9934948	T	C	0.71237	0.000015094
	chr17	13124092	rs76994622	C	A	1.81322	3.55926E-05

Type	CHR	POS	SNPID	REF	ALT	OR	P
	chr19	3365483	rs76049683	G	A	3.44726	4.86563E-05
	chr20	17422063	rs2021785	C	T	0.736699	4.77956E-05
	chr22	19593301	rs5748346	A	G	0.688881	3.43471E-06
SE	chr1	39248701	rs2151266	T	C	3.40955	4.09155E-05
	chr1	88345118	rs3008439	T	G	0.682644	1.93532E-05
	chr1	88382412	rs12141842	G	T	0.673146	1.09873E-05
	chr1	104457967	rs61815922	G	A	1.42877	4.98735E-05
	chr1	215136314	rs17023904	A	G	1.73576	1.76661E-05
	chr2	117681102	rs12478182	A	C	1.59587	4.15658E-05
	chr4	25417244	rs3816587	C	T	1.46307	0.000012934
	chr4	25419283	rs9174	T	C	1.45955	1.51213E-05
	chr4	181947443	rs17070473	C	T	3.66279	1.87104E-05
	chr5	67015614	rs73110943	C	A	1.43709	1.27685E-05
	chr6	152786487	rs140135976	G	A	3.93094	4.66345E-05
	chr8	56144612	rs117586513	C	T	5.41953	1.08787E-06
	chr8	134811897	rs150704178	G	A	2.80556	2.71099E-05
	chr8	144130643	rs78140118	G	A	3.24103	4.29581E-07
	chr8	144139278	rs145521790	G	A	3.1283	1.2999E-06
	chr12	43002837	rs35625089	A	G	3.78096	4.74261E-05
	chr14	54920423	rs11850320	T	G	0.549482	4.18031E-05
	chr14	54970215	rs75356305	G	A	0.512861	8.68319E-06
	chr14	54996872	rs11555279	G	A	0.512294	8.7701E-06
	chr14	55026435	rs61975112	C	G	0.607503	4.95083E-05
	chr16	277912	rs8045291	A	G	1.67081	2.26491E-05
	chr16	49986462	rs7199622	C	T	0.636673	0.000024511
	chr17	37278090	rs75491618	C	A	2.53092	1.50199E-05
	chr19	58199102	rs199582159	A	T	3.0778	0.00002616
	chr22	33828432	rs73166278	C	T	3.56121	2.44106E-05
SY	chr1	6147407	rs4908763	C	T	0.497328	1.45275E-06
	chr1	167484098	rs34322294	C	A	3.90611	3.27205E-05
	chr1	188860684	rs182247528	G	A	2.96587	2.55683E-05
	chr2	152247421	rs10178027	C	T	0.717498	1.24149E-05
	chr3	2799942	rs12494784	C	A	0.737044	1.82099E-05
	chr5	9644727	rs72731192	A	C	1.43921	3.91701E-05
	chr5	16690033	rs2303704	A	C	0.65221	2.74483E-05
	chr5	83074951	rs35125267	G	T	1.48623	3.04379E-05
	chr5	83151160	rs10072920	T	C	1.48863	3.39765E-05
	chr6	31643522	rs9378164	G	A	1.58794	2.70239E-05
	chr6	139860412	rs3888326	T	G	1.32813	3.59461E-05
	chr6	160858188	rs2292334	G	A	1.33438	4.35059E-05
	chr7	17644035	rs76717104	C	T	1.37429	3.24011E-05
	chr7	17662517	rs7788582	A	G	1.33293	2.82059E-05
	chr7	38716734	rs35959763	G	A	1.89487	4.81705E-05
	chr9	20175694	rs7870612	A	G	0.451677	1.27203E-05
	chr9	28768187	rs1590671	T	C	1.80429	2.29289E-05
	chr10	134300395	rs73393404	G	T	1.76357	4.90028E-05
	chr12	76803799	rs77512784	A	G	3.89003	1.02063E-06
	chr13	52081846	rs9646005	C	T	0.747808	3.04719E-05
	chr14	54805719	rs17726338	G	A	1.58501	2.71135E-05
	chr16	13881535	rs59923549	G	T	0.72029	2.17838E-06
	chr19	40541039	rs61737024	C	T	1.86919	3.78786E-05
	chr19	40582565	rs149500464	C	T	1.96198	3.09503E-05

TE=Taeeumin, SE=Soeumin, SY=Soyangin

Table 3. Annotated Disease Class of Sasang Constitution Related Genes

Type	CHR	Associated Disease Class	Gene Symbol	dbSNP IDs
SE	chr1	PSYCH	AGBL3	rs3008439, rs12141842, rs61815922
	chr1	METABOLIC	EPHB2	rs2151266, rs3008439, rs12141842, rs61815922
	chr1	METABOLIC	MTF1	rs2151266, rs3008439, rs12141842, rs61815922
	chr1	CARDIOVASCULAR	PTP4A2	rs2151266
	chr1	CARDIOVASCULAR	SPEN	rs2151266, rs3008439, rs12141842, rs61815922
	chr2	CARDIOVASCULAR	PTPRN2	rs12478182
	chr2	CARDIOVASCULAR	TPO	rs12478182
	chr4	IMMUNE, METABOLIC	ANAPC4	rs3816587, rs9174
	chr4	METABOLIC	HTRA3	rs3816587, rs9174
	chr4	HEMATOLOGICAL	NRAS	rs3816587, rs9174
	chr5	METABOLIC	C7orf63	rs73110943
	chr6	DEVELOPMENTAL, HEMATOLOGICAL, CARDIOVASCULAR, METABOLIC, PSYCH	SYNE1	rs140135976
	chr14	METABOLIC	ABCB1	rs11850320, rs75356305, rs11555279, rs61975112
	chr14	NEUROLOGICAL	ANG	rs11850320, rs75356305, rs11555279, rs61975112
	chr14	DEVELOPMENTAL	DGKE	rs11850320
	chr16	IMMUNE	TNP2	rs7199622
	chr17	CARDIOVASCULAR	PEMT	rs75491618
	chr19	METABOLIC	ZNF107	rs199582159
	chr19	CHEMDEPENDENCY	ZNF138	rs199582159
	chr19	CHEMDEPENDENCY	ZNF92	rs199582159
	chr22	CARDIOVASCULAR, METABOLIC	LARGE	rs73166278
TE	chr2	CARDIOVASCULAR	PTPRN2	rs62155477
	chr2	METABOLIC	TBC1D8	rs62155477
	chr2	CARDIOVASCULAR	TPO	rs12692398, rs62155477
	chr3	CARDIOVASCULAR	CADM2	rs9849542
	chr3	CHEMDEPENDENCY, DEVELOPMENTAL, METABOLIC	COLQ	rs72911400
	chr3	CARDIOVASCULAR, METABOLIC, REPRODUCTION	LRRN1	rs72911400, rs9849542, rs854581
	chr3	METABOLIC, CARDIOVASCULAR	TKT	rs9849542, rs854581
	chr6	INFECTION	C7orf44	rs73424328
	chr6	METABOLIC	CALN1	rs73424328
	chr6	VISION	CAP2	rs73424328
	chr6	IMMUNE, NEUROLOGICAL	DHFRP2	rs73424328
	chr6	CARDIOVASCULAR	NRAS	rs73424328
	chr6	IMMUNE	PHTF2	rs73424328
	chr6	INFECTION	PKD1L1	rs73424328
	chr6	IMMUNE	SEMA3D	rs73424328
	chr6	DEVELOPMENTAL, HEMATOLOGICAL, CARDIOVASCULAR, METABOLIC, PSYCH	SYNE1	rs769918744
	chr7	METABOLIC	EIF2AK1	rs76717104, rs1419772
	chr10	CARDIOVASCULAR, MATABOLIC, RENAL, PSYCH	ANK3	rs582273
	chr10	DEVELOPMENTAL, HEMATOLOGICAL, CARDIOVASCULAR, METABOLIC, NEUROLOGICAL	KCNMA1	rs582273
	chr10	DEVELOPMENTAL	RET	rs582273
	chr10	METABOLIC	ZNF32	rs582273
	chr12	NEUROLOGICAL, CHEMDEPENDENCY, CARDIOVASCULAR, METABOLIC	ITPR2	rs74798154

Type	CHR	Associated Disease Class	Gene Symbol	dbSNP IDs
SY	chr13	CARDIOVASCULAR	CNTNAP2	rs9518188
	chr14	METABOLIC	ABCB1	rs55861984
	chr14	NEUROLOGICAL	ANG	rs55861984
	chr16	NEUROLOGICAL	CACNG3	rs2158245
	chr16	METABOLIC	CDH3	rs9934948
	chr16	IMMUNE	TNP2	rs2158245, rs9934948
	chr20	REPRODUCTION	LRRN4	rs2021785
	chr20	METABOLIC, CHEMDEPENDENCY, NEUROLOGICAL,CARDIOVASCULAR,	PCSK2	rs2021785
	chr20	NEUROLOGICAL	TGM6	rs2021785
	chr1	IMMUNE	CD247	rs34322294
	chr1	METABOLIC	MTF1	rs34322294, rs182247528
	chr1	METABOLIC	SLC39A1	rs34322294, rs182247528
	chr2	CARDIOVASCULAR, NEUROLOGICAL, METABOLIC	MBD5	rs10178027
	chr2	CARDIOVASCULAR	PTPRN2	rs10178027
	chr2	CARDIOVASCULAR	TPO	rs10178027
	chr3	METABOLIC, HEMATOLOGICAL, CARDIOVASCULAR, NEUROLOGICAL	CNTN4	rs12494784
	chr5	CARDIOVASCULAR, NEUROLOGICAL, METABOLIC	ACTBP2	rs35125267, rs10072920
	chr5	METABOLIC	C7orf63	rs35125267, rs10072920
	chr5	CARDIOVASCULAR, IMMUNE	MYO10	rs2303704
	chr6	IMMUNE, PHARMACOGENOMIC	BAT1	rs9378164
	chr6	METABOLIC	CALN1	rs9378164
	chr6	VISION	CAP2	rs9378164
	chr6	IMMUNE, NEUROLOGICAL	DHFRP2	rs9378164
	chr6	CARDIOVASCULAR	NRAS	rs9378164
	chr6	IMMUNE	PHTF2	rs9378164
	chr6	IMMUNE	SEMA3D	rs9378164
	chr6	CHEMDEPENDENCY, CARDIOVASCULAR, METABOLIC, IMMUNE	SLC22A3	rs2292334
	chr6	IMMUNE, METABOLIC	VARS2	rs9378164
	chr7	METABOLIC	EIF2AK1	rs76717104, rs7788582
	chr9	DEVELOPMENTAL, CARDIOVASCULAR	MPDZ	rs7870612, rs1590671
	chr9	METABOLIC, CARDIOVASCULAR, INFECTION	TDPX2	rs7870612, rs1590671
	chr10	METABOLIC	AGK	rs73393404
	chr10	METABOLIC	CPN1	rs73393404
	chr10	CARDIOVASCULAR	LIPA	rs73393404
	chr12	METABOLIC	SLC26A5	rs77512784
	chr12	METABOLIC	TPH2	rs77512784
	chr13	CARDIOVASCULAR	CNTNAP2	rs9646005
	chr14	METABOLIC	ABCB1	rs17726338
	chr14	NEUROLOGICAL	ANG	rs17726338
	chr16	CARDIOVASCULAR	DNAH5	rs59923549
	chr16	IMMUNE	TNP2	rs59923549
	chr19	METABOLIC	ZNF107	rs61737024, rs149500464
	chr19	CHEMDEPENDENCY	ZNF138	rs61737024, rs149500464
	chr19	CHEMDEPENDENCY	ZNF92	rs61737024, rs149500464

3. 체질분류모델 개발

체질분류모델을 개발하기 위해, 각 체질별 선정된 SNP와 태음, 소음, 소양 연관 SNP 중 WGS 검증을 통과한 유전지표와 정량지표로 성별, 나이, BMI를 사용하였다. Table 4에서는 각 체질별 연관 SNP 정보만을 활용하여 체질분류모델을 개발한 결과이며, Table 5는 SNP 정보와 정량지표를 함께 사용하여 개발된 분류모델의 결과이다.

SNP만을 활용하여 체질을 분류한 모델의 체질별 최대 Area Under the Curve (AUC)는 세 체질 모두 다운 샘플링 데이터를 사용하고, SVM 알고리즘으로 개발한 모델로 태음인에서는 최대 AUC 값이 0.716이었으며, 소음인에서는 0.709, 소양인에서는 0.702이었다. SNP 정보와 정량지표를 함께 사용한 모델에서는 태음인은 원 데이터를 그대로 사용하고, SVM 알고리즘을 활용한 모델(0.894)이었고, 소음인과 소양인에서는 원 데이터를 그대로 사용하고, RF 알고리즘을 활용한 모델이었다(각 0.868, 0.767).

Table 4. Maximum AUC of Each Predictive Model for Sasang Constitution Type Using Genotyping Data

Model	TE vs. non-TE	SE vs. non-SE	SY vs. non-SY
RF_UP	0.68268	0.690037	0.667104
RF_DOWN	0.691678	0.692239	0.67276
RF	0.69219	0.699495	0.679366
SVM_UP	0.71087	0.677134	0.693377
SVM_DOWN	**0.716078**	**0.709385**	**0.702389**
SVM	0.708792	0.67265	0.692636
NN_UP	0.676766	0.629683	0.603245
NN_DOWN	0.65678	0.646639	0.600676
NN	0.645463	0.632865	0.602789

Data shown as maximum AUC

* TE=Taeeumin, Non-TE: non- Taeeumin, SE=Soeumin, Non-SE: non- Soeumin, SY=Soyangin, Non-SY: non- Soyangin, RF: Random Forest, RF_UP: RF up sampling, RF_DOWN: RF down sampling, SVM: Support Vector Machine, SVM_UP: SVM up sampling, SVM_DOWN: SVM down sampling, NN: Neurnal Network, NN_UP: NN up sampling, NN_DOWN: NN down sampling

Table 5. Maximum AUC of Each Predictive Model for Sasang Constitution Type Using Genotyping and Clinical Data (Sex, Age, BMI)

Model	TE vs. non-TE	SE vs. non-SE	SY vs. non-SY
RF_UP	0.873964	0.849682	0.76128
RF_DOWN	0.886476	0.846939	0.751676
RF	0.878307	**0.868203**	**0.766644**
SVM_UP	0.893538	0.853485	0.761229
SVM_DOWN	0.887347	0.842382	0.755727
SVM	**0.894456**	0.839792	0.75604
NN_UP	0.859287	0.84541	0.725565
NN_DOWN	0.865301	0.829683	0.74421
NN	0.868209	0.843991	0.735819

Data shown as maximum AUC

* TE=Taeeumin, Non-TE: non- Taeeumin, SE=Soeumin, Non-SE: non- Soeumin, SY=Soyangin, Non-SY: non- Soyangin, RF: Random Forest, RF_UP: RF up sampling, RF_DOWN: RF down sampling, SVM: Support Vector Machine, SVM_UP: SVM up sampling, SVM_DOWN: SVM down sampling, NN: Neurnal Network, NN_UP: NN up sampling, NN_DOWN: NN down sampling

IV. 考察

본 연구는 유전지표를 활용한 체질분류 모델을 개발하고, 이를 통해 사상체질의 과학적 근거확보와 이론의 확장에 더욱 도움이 되고자 하였다. 2008 ~ 2013년 다기관 한방의료기관에서 수집한 20세 이상으로 사상체질처방을 복용한 결과를 토대로, 체질전문가의 진단을 거쳐 체질이 임상적으로 명확히 진단된 1999명의 대상자의 지노타이핑 정보와 기본적인 정량지표(성별, 나이, BMI)를 사용하여 체질분류모델을 개발하였다.

체질별 GWAS 분석을 통해, 태음인 29개, 소음인 25개, 소양인 24개의 SNP를 선정하였다. 태음 연관 29개의 유전지표 중 22개는 유전자 주변 영역에 위치했다. rs582273는 BMI, Triglycerides, Schizophrenia 및 LDL 콜레스테롤과 같이 다양한 질환 연관 SNP으로 보고되어 있었다. 소음 연관 SNP 중 14개가 유전자 영역 주변에 위치해 있었고, 그 중, rs2151266의 경우 심부전, 인슐린 저항성에 관여하는 SNP로 보고되어

있었으며, rs3008439의 경우 혈압 연관 SNP으로 보고되어 있었다. 소양인의 연관 SNP 중 rs10072920의 경우 루게릭병과 관련된 SNP로 신경계 기능에 관여하는 것으로 보인다(Table 3).

각 체질별 연관 SNP만을 활용하여 개발한 체질별 분류 모델의 AUC 값은 0.702~0.716이었다. 최근 실제 임상에서 활용 가능한 위암 진단 모델[28]의 AUC 값은 0.851~0.981이었던 것에 비하면, 낮은 수치값이라 할 수 있다. 지금까지 사상체질과 같은 인간의 복잡한 형질을 유전지표를 활용하여 예측하고자 하는 연구는 지속되어 왔었지만, 복합 형질에 대한 개별 유전지표의 영향력은 매우 미비하며 아직 일반적인 연구라고 볼 수는 없다[29-32]. 가장 최근에 이루어진 대규모 연구로는 Louis Lello의 연구로, UK Biobank[33,34], eMERGE[35], adjacent ancestry(AA)[36]의 데이터를 활용하여 고혈압, 갑상선 기능 항진증, 제 2형 당뇨, 유방암, 녹내장 등을 예측했다[37]. Louis Lello의 연구에서는 복합질환과 연관된 전체 유전지표를 활용하여 0.58~0.71의 AUC를 얻을 수 있었다. 현재 수집된 가장 큰 데이터 집합인 UK biobank를 활용하여 복합질환 또는 형질을 예측했다는 점에서, 현재 유전지표의 복합형질에 대한 설명력의 한계를 엿볼 수 있지만, 복합 형질에 대한 유전적 중요성을 보여준 연구라 볼 수 있다. 이러한 점에서 본 연구에서 유전지표만으로 예측한 체질 분류모델의 AUC 값은 신뢰성이 있다고 할 수 있겠다. 또한 AUC가 0.7 이상으로 계산된 경우, 그 결과 점수의 예측력은 수용 가능한 것으로 결정되었지만 0.8보다 큰 AUC는 우수하다고 간주[38,39]하는 이전 연구들을 비추어보면 예측력을 갖추었다고 볼 수 있다.

성별, 나이, BMI로 이루어진 정량지표를 추가하여 개발한 분류모델은 0.767~0.894로 높은 AUC를 얻을 수 있었으며, 개발된 체질분류모델의 AUC 값이 모두 0.7을 넘어, 예측력이 수용가능한 것으로 볼 수 있다. Louis Lello[37]의 연구에서도 각 복합질환을 성별과 나이를 추가하여 예측 모델을 만들었으며, 이 때 AUC 값은 0.651~0.864를 얻어, 본 연구 결과와 유사하였다. SNP만을 활용하여 개발한 분류 모델에 비하여 약 0.6~0.17로 AUC 값이 높아지는 것을 볼 수 있는데, 이는 분류모델에 사용한 정량지표가 체질과 연관되어 있음을 알 수 있다. 특히 BMI는 이전의 많은 연구에서 체질과 연관성이 높은 정보로, KS-15[9]설문지에서는 BMI를 체질 진단을 위한 정량지표로 활용하고 있으며, QSCCⅡ[7,8]에서는 '체격이 큰 편이다/작은 편이다, 뚱뚱한 편이다/마른 편이다'로 활용되는 등 통상적으로 체형의 정량화된 대표적인 지표라 할 수 있다. 성별과 연령은 사용한 대상자의 수의 한계로 인하여, 유전지표나 BMI 정보를 보정하기 위해 사용하였다. 정량지표로 활용한 분류모델에서 더욱 좋은 성능을 보인 결과를 비추어보아, 체질과 연관된 보다 많은 정량지표를 사용한다면 분류모델의 성능을 더욱 높일 수 있을 것이라 생각된다.

본 연구에서 사용한 데이터마이닝 모델은 Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN)이다. 이들은 데이터마이닝에서 자주 활용되는 모델들로, 세가지 기법을 활용하여 체질분류 모델을 개발한 것은 어느 하나의 모델 특성에 따라 체질분류 모델이 정해지는 것을 방지하고, 다양한 모델에서의 성능을 비교해보기 위해서이다. 본 연구 결과에서 유전지표인 SNP만을 활용한 연구에서는 SVM 모델이 가장 좋은 성능을 보였고, 정량지표를 추가한 모델에서는 RF와 SVM 모델에서 가장 좋은 성능을 보였다. 전체적으로 성능의 차이는 크지 않았지만, RF, SVM 모델에서 좋은 결과를 얻었고, NN 모델에서 가장 낮은 성능을 보인 것으로 나타났다. 이는 NN 모델의 특성 상, 실제 상황에 정확히 부합하는 작동을 학습하기 위해서 즉, 여러 가지 새로운 케이스에 정확히 동작하는 근본적인 구조를 잡기 위해서는 수많은 훈련 데이터들이 필요한데, 본 연구에서 사용한 데이터 수가 이에 미치지 못했기 때문이라 생각된다. 앞으로 체질 유전정보가 더욱 확충된다면, NN 모델에서도 좋은 결과를 얻을 수 있을 것이다.

샘플링 방법에서는 유전지표만을 사용한 분류 모델에서 다운 샘플링에서 결과가 가장 좋은 것으로 나타났지만, 업 샘플링과 원데이터를 사용한 결과와 비

교하여 큰 차이가 없었다. 이는 체질분류모델에 사용한 데이터가 각 체질 내에서 극단적인 값을 보이는 정보가 없었다는 것을 뜻한다.

하지만, 본 연구에는 몇 가지 제한점이 있다. 첫째, 사상체질처방을 복용한 결과를 토대로, 체질전문가의 진단을 거쳐 체질이 임상적으로 명확히 진단된 대상자이기는 하지만, 기존의 유전체 연구에 비하여 대상자 수가 적다. 최근 유전체 연구에 활용이 많이 되고 있는 UK Biobank를 활용한 연구들과 비교하면[40], 굉장히 적은 수라고 볼 수 있다. 앞으로 한의학을 기반으로 유전 연구를 하기 위해서 보다 많은 수의 데이터를 확보할 필요가 있다. 둘째, 체질분류모델에서 사용한 정량지표가 성별, 나이, BMI로 한정되어 있다. BMI는 체질에 유의한 임상지표라고 알려져 있으나, 체형정보, 안면정보, 음성정보와 같이 체질과 유의한 정량지표가 존재한다. 이에 따라 추후 정량지표를 더욱 보강하여 체질분류모델을 개발할 예정이다.

V. 結論

체질을 객관적으로 진단하기 위한 연구들이 진행되고 있으나, 여전히 체질의 유전성에 대한 근거들을 활용하는 연구는 부족했다. 본 연구는 유전 정보를 이용하여 사상체질 연관 SNP를 확인하고, 이를 통해 체질분류모델을 개발하였다. 개발된 체질분류모델을 실제 임상에서 활용할 수 있도록 성능을 확보하기 위해서 성별, 나이, BMI를 추가하여 체질분류모델을 개선하여, 기존의 대규모 선행연구결과와 유사한 AUC를 얻었고, 체질 연관 유전지표를 활용하여 사상체질을 분류하는 모델을 개발하는 것에 대한 가능성을 보인 것이라 할 수 있다. 이러한 결과를 바탕으로 추후 더 많은 연구가 뒷받침되어 사상체질뿐만 아니라 한의 변증을 기반으로 하는 한의정밀의료가 실현되기를 기대한다.

VI. 감사의 글

이 연구는 2020년도 한국한의학연구원의 '빅데이터 기반 한의 예방 치료 원천기술 개발'(KSN2021120)의 지원을 받아 수행되었습니다.

VII. References

1. Lee SH, Yoon YS, Kim HG, Kim JY. Clinical Study on the Distribution of Sasang Constitutions between Parents and their Offsprings. Journal of Physiology & Pathology in Korean Medicine. 2004;18(6):1904-7. (Korean)
2. Lee SW, Hur YM, Park HY, Kim JY. A validation study on Sasang constitutions and genetic influences. FACT: Focus on Alternative and Complementary Therapies. 2007;2007(12). DOI: 10.1111/j.2042-7166.2007.tb05894.x. Korean.
3. Lee MK, Jang ES, Sohn HY, Park JY, Koh BH, Sung J, et al. Investigation of Genetic Evidence for Sasang Constitution Types in South Korea. Genomics & Informatics. 2009;7(2):107-10. DOI: https://doi.org/10.5808/gi.2009.7.2.107.
4. Hur YM, Lee SW, Jin HJ. Genetic and environmental overlaps among sasang constitution types: a multivariate twin study. Twin Research and Human Genetics. 2018;21(6):518-26. DOI: 10.1017/thg.2018.56.
5. Kim BY, Jin HJ, Kim JY. Genome-wide association analysis of Sasang constitution in the Korean population. The Journal of Alternative and Complementary Medicine. 2012;18(3):262-9.
6. Cha SW, Yu HJ, Park AY, Oh SA, Kim JY. The obesity-risk variant of FTO is inversely related with the So-Eum constitutional type: genome-wide association and replication analyses. BMC complementary and alternative medicine. 2015;15(1):120. DOI: 10.

1089/acm.2010.0764.

7. Kim SH, Ko BH, Song IB. A study on the standardization of QSCC II (Questionnaire for the Sasang Constitution Classification II). The Journal of Korean Medicine. 1996;17(2):337-93. Korean.
8. Lee SG, Kwak CK, Lee EJ, Koh BH, Song IB. The Study on the Upgrade of QSCC (II). J of Sasang Const Med. 2003;15(1):39-49. Korean.
9. Baek YH, Jang ES, Park KH, Yoo JH, Jin HJ, Lee SW. Development and validation of brief KS-15 (Korea Sasang Constitutional Diagnostic Questionnaire) based on body shape, temperament and symptoms. Journal of Sasang Constitutional Medicine. 2015;27(2): 211-21. Korean.
10. Lee MS, Bae NY, Hwang MW, Chae H. Development and validation of the digestive function assessment instrument for traditional Korean medicine: Sasang digestive function inventory. Evidence-Based Complementary and Alternative Medicine. 2013;2013. DOI: 10.1155/2013/263752.
11. So JH, Kim JW, Nam JH, Lee BJ, Kim YS, Kim JY, et al. The web application of constitution analysis system-SCAT (Sasang Constitution Analysis Tool). Journal of Sasang Constitutional Medicine. 2016;28(1):1-10. Korean.
12. Jin HJ, Baek YH, Kim HS, Ryu JH, Lee SW. Constitutional multicenter bank linked to Sasang constitutional phenotypic data. BMC complementary and alternative medicine. 2015;15(1):1. DOI: 10.1186/s 12906-015-0553-3.
13. Hyun MK, Baek YH, Lee SW. Association between digestive symptoms and sleep disturbance: a cross-sectional community-based study. BMC gastroenterology. 2019;19(1):34. DOI: 10.1186/s12876-019-0945-9.
14. Baek YH, Kim HS, Lee SW, Ryu JH, Kim YY, Jang ES. Study on the ordinary symptoms characteristics of gender difference according to Sasang constitution. 2009;23(1):251-8. Korean.
15. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic acids research. 2016;44(D1):D862-D8. DOI: 10.1093/nar/gkv1222.
16. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014;42(D1):D1001-D6.DOI:10.1093/nar/gkt1229.
17. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81.DIO:10.1038/nature15394.
18. Consortium GP. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.DOI:10.1038/nature15393.
19. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature genetics. 2012;44(8):955-9. DOI: 10.1038/ng.2354.
20. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3: Genes, Genomes, Genetics. 2011;1(6):457-70. DOI: 10.153 4/g3.111.001198.
21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics. 2007;81(3):559-75. DOI: 10.1086/519795.
22. Team RC. R language definition. Vienna, Austria: R foundation for statistical computing. 2000.
23. Kuhn M. Caret: classification and regression training. Astrophysics Source Code Library. 2015.
24. Breiman L. Random forests. Machine learning. 2001;

45(1):5-32.

25. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab-an S4 package for kernel methods in R. Journal of statistical software. 2004;11(9):1-20.
26. Venables WN, Ripley BD. Modern applied statistics with S-PLUS: Springer Science & Business Media; 2013.
27. Kovacevic J, Vetterli M. The commutativity of up/downsampling in two dimensions. IEEE transactions on information theory. 1991;37(3):695-8. DOI: 10.1109/18.79936.
28. Yoon HJ, Kim SH, Kim JH, Keum JS, Oh SI, Jo JI, et al. A Lesion-Based Convolutional Neural Network Improves Endoscopic Detection and Depth Prediction of Early Gastric Cancer. Journal of clinical medicine. 2019;8(9):1310. DOI: 10.3390/jcm8091310.
29. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS one. 2008;3(10). DOI : 10.1371/journal.pone.0003395.
30. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nature Reviews Genetics. 2016;17(7):392. DOI: 10.1038/nrg.2016.27.
31. Janssens ACJ, Ioannidis JP, Bedrosian S, Boffetta P, Dolan SM, Dowling N, et al. Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. European journal of epidemiology. 2011;26(4):313. DOI: 10.1111/j.1365-2362.2011.02493.x.
32. Kraft P, Hunter DJ. Genetic risk prediction—are we there yet? New England Journal of Medicine. 2009; 360(17):1701-3.DOI:10.1093/jnci/djq413.
33. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018; 562(7726):203-9. DOI: 10.1038/s41586-018-0579-z.
34. Barbour V. UK Biobank: a project in search of a protocol? The Lancet. 2003;361(3970):1734-8. DOI: 10.1016/S0140-6736(03)13377-6.
35. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics. 2011;4(1):13. DOI: 10.1186/1755-8794-4-13.
36. Marquez-Luna C, Gazal S, Loh P-R, Furlotte N, Auton A, Price AL, et al. Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. bioRxiv. 2018: 375337. DOI: 10.1101/375337.
37. Lello L, Raben TG, Yong SY, Tellier LC, Hsu SD. Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. Scientific reports. 2019;9(1):1-16. DOI: 10.1038/s41598-019-51258-x.
38. Berliner JL, Brodke DJ, Chan V, SooHoo NF, Bozic KJ. John Charnley Award: preoperative patient-reported outcome measures predict clinically meaningful improvement in function after THA. Clinical Orthopaedics and Related Research®. 2016;474(2): 321-9. DOI: 10.1007/s11999-015-4350-6.
39. Keswani A, Tasi MC, Fields A, Lovy AJ, Moucha CS, Bozic KJ. Discharge destination after total joint arthroplasty: an analysis of postdischarge outcomes, placement risk factors, and recent trends. The Journal of arthroplasty. 2016;31(6):1155-62. DOI: 10.1016/j.arth.2015.11.044.
40. Jones SE, Tyrrell J, Wood AR, Beaumont RN, Ruth KS, Tuke MA, et al. Genome-wide association analyses in 128,266 individuals identifies new morningness and sleep duration loci. PLoS genetics. 2016;12(8). DOI: 10.1371/journal.pgen.1006125.