G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics Inform 2012;10(4):244-248
http://dx.doi.org/10.5808/GI.2012.10.4.244

ORIGINAL ARTICLE

Association Analysis of Reactive Oxygen Species-Hypertension Genes Discovered by Literature Mining

Ji Eun Lim[1], Kyung-Won Hong[2], Hyun-Seok Jin[3], Bermseok Oh[1]*

[1]Department of Biomedical Engineering, Kyung Hee University School of Medicine, Seoul 130-701, Korea,
[2]Division of Epidemiology and Health Index, Center for Genome Science, Korea National Institute of Health,
Korea Centers for Disease Control and Prevention, Cheongwon 363-951, Korea,
[3]Department of Medical Genetics, Ajou University School of Medicine, Suwon 443-721, Korea

Oxidative stress, which results in an excessive product of reactive oxygen species (ROS), is one of the fundamental mechanisms of the development of hypertension. In the vascular system, ROS have physical and pathophysiological roles in vascular remodeling and endothelial dysfunction. In this study, ROS-hypertension-related genes were collected by the biological literature-mining tools, such as SciMiner and gene2pubmed, in order to identify the genes that would cause hypertension through ROS. Further, single nucleotide polymorphisms (SNPs) located within these gene regions were examined statistically for their association with hypertension in 6,419 Korean individuals, and pathway enrichment analysis using the associated genes was performed. The 2,945 SNPs of 237 ROS-hypertension genes were analyzed, and 68 genes were significantly associated with hypertension ($p < 0.05$). The most significant SNP was rs2889611 within *MAPK8* ($p = 2.70 \times 10^{-5}$; odds ratio, 0.82; confidence interval, 0.75 to 0.90). This study demonstrates that a text mining approach combined with association analysis may be useful to identify the candidate genes that cause hypertension through ROS or oxidative stress.

Keywords: genetic association study, hypertension, literature mining, reactive oxygen species

Introduction

Hypertension is defined as blood pressure measurement consistently higher than 140 mm Hg systolic blood pressure (SBP) and/or 90 mm Hg diastolic blood pressure (DBP) [1]. It is a complex syndrome determined by genetic and environmental factors and affected by multiple genetic factors to 30% to 50% of blood pressure variability in human hypertension [2]. Although hypertension is a leading cause of cardiovascular disease, ischemic heart disease, and stroke, the exact cause of hypertension is unclear [3].

Oxidative stress, which results in an excessive product of reactive oxygen species (ROS), is one of the fundamental mechanisms of the development of hypertension. In the vascular system, ROS has physical and pathophysiological roles that are important in vascular remodeling and endothelial dysfunction associated with hypertension [4]. Since 1960, when the association between free radicals and hypertension was reported [5], plenty of data supporting a role of oxidative stress in hypertension have been published. However, the evidence of whether oxidative stress causes hypertension is weak, and a few clinical studies have shown the relationship between blood pressure and ROS. Nonetheless, oxidative stress has an important role in vascular biology and a potential role in hypertension.

In this study, ROS-hypertension-related genes were collected by the biological literature-mining tools, such as SciMiner and gene2pubmed, in order to identify the genes that would cause hypertension through ROS. Further, single nucleotide polymorphisms (SNPs) located within these gene regions were examined statistically for their association with hypertension in 6,419 Korean individuals, and pathway enri-

Received November 2, 2012; Revised November 13, 2012; Accepted November 15, 2012
*Corresponding author: Tel: +82-2-961-0617, Fax: +82-2-6008-0647, E-mail: ohbs@khu.ac.kr

chment analysis using the associated genes was performed.

Methods

Study participants and genotyping

The Korea Association Resource (KARE) study recruited 10,038 participants aged 40 years to 69 years from the rural Ansung and urban Ansan cohorts and has been previously described in detail [6]; 1,196 subjects were excluded due to poor genotyping data, and we also excluded subjects with prehypertensive status (120 mm Hg $<$ SBP $<$ 140 mm Hg and/or 80 mm Hg $<$ DBP $<$ 90 mm Hg). In total, 6,420 participants—1,968 hypertensive cases with hypertensive therapy or SBP $\geq$ 140 mm Hg or DBP $\geq$ 90 mm Hg and 4,452 controls with SBP $\leq$ 120 mm Hg and DBP $\leq$ 80 mm Hg—were examined for a hypertension case control study.

The Affymetrix Genome-Wide Human SNP array 5.0 (Affymetrix, Inc., Santa Clara, CA, USA) was used to genotype KARE study individuals. The accuracy of the genotyping was examined by Bayesian Robust Linear Modeling using the Mahalanobis distance (BRLMM) genotyping algorithm [7]. The sample and SNP quality control criteria have been described [6]. In brief, samples with accuracies that were lower than 98%, high missing genotype call rates ($\geq$4%), high heterozygosity ($>$30%), or gender biases were excluded. SNPs were excluded according to filter criteria as follows: SNP call rate $>$ 5%, minor allele frequency $<$ 0.01, and Hardy-Weinberg equilibrium $p < 1 \times 10^{-6}$. After quality control, 8,842 individuals and 352,228 markers remained.

Ascertaining ROS- and hypertension-related genes

The SciMiner [8] web-based literature mining tool was used to obtain gene sets associated with ROS and hypertension. SciMiner was run on a query of "Reactive Oxygen Species" [MeSH] AND "Hypertension" [MeSH], identifying ROS-hypertension articles and genes as of April 24, 2012. We also retrieved genes for these ROS-hypertension articles from NCBI gene2pubmed (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA) data. The newly found genes from gene2pubmed were added to the ROS-hypertension gene set. The positions of genes in the human genome were downloaded from the Ensembl Biomart database (NCBI build 36). Some gene symbols were different from the results of SciMiner and Biomart, such as *NOS2A* → *NOS2* and *STN* → *EEF1A2*. The functional analysis tools, such as SciMiner, WebGestalt [9, 10], and DAVID [11, 12], were used for enrichment analysis to find the pathway with ROS-hypertension-associated genes, and the statistical significance of biological functions was calculated with Benjamini and Hochberg-adjusted $p < 0.05$ as the cutoff.

Table 1. Distribution of number of genes according to number of papers

No. of papers	No. of genes (%)
>10	58 (10.1)
2-10	201 (34.9)
1	317 (55.0)

Statistical analyses

PLINK version v1.07 (http://pngu.mgh.harvard.edu/~purcell/plink) was used to perform the association analysis, and the hypertension case control study was tested by logistic regression analysis. The association tests were based on an additive genetic model and adjusted for recruitment area, age, sex, and body mass index.

Results

Ascertaining ROS and hypertension candidate genes

With the results of SciMiner, queried with "'Reactive Oxygen Species' [MeSH] AND 'Hypertension' [MeSH]", 574 genes were obtained from 903 ROS-hypertension-related articles; 49 genes were found through the NCBI gene2pubmed data with these 903 papers, and only 2 genes out of 49 genes were new to the 574 SciMiner genes. Three hundred seventeen genes (55%) among the 576 ROS-hypertension genes were referenced in only 1 article (Table 1) and were excluded for further analysis, with 259 genes remai-

Table 2. Frequent reactive oxygen species (ROS)-hypertension genes (number of papers > 40)

No. of papers	Symbol	Chr no.	Start	End
320	*AGT*	1	228,884,897	228,936,564
223	*NOX5*	15	66,989,918	67,156,127
223	*SOD1*	21	31,933,806	31,983,115
135	*NOS2*	17	23,087,922	23,171,682
112	*REN*	1	202,370,571	202,422,088
101	*NOS3*	7	150,299,080	150,362,608
90	*CAT*	11	34,397,054	34,470,176
85	*NOS1*	12	116,100,497	116,303,965
79	*ACE*	17	58,888,166	58,972,937
79	*CYBA*	16	87,217,199	87,264,958
71	*AGTR1*	3	149,878,348	149,963,478
65	*XDH*	2	31,390,692	31,511,115
63	*INS*	11	2,104,432	2,159,027
62	*EDN1*	6	12,378,599	12,424,286
46	*PRKCA*	17	61,709,216	62,257,324
43	*SOD2*	6	160,000,141	160,054,343
42	*NOX4*	11	88,679,163	88,884,301

ning.

Using Ensembl Biomart (NCBI Build 36), we then extracted the position information of 259 genes, and the genes located on chromosomes X, Y, and MT were also excluded. Finally, 237 genes that included SNPs genotype information from KARE data within the gene boundary (±20 kb upstream and downstream of the gene) were selected, and 2,945 SNPs were tested for hypertension association analysis. The frequently mentioned genes (number of papers > 40) in the ROS-hypertension papers are shown in Table 2.

Association analysis of hypertension

We examined 2,945 SNPs of 237 genes for a hypertension case control study by logistic regression analysis; 68 genes were significantly associated with hypertension ($p < 0.05$) (Table 3). The most significant SNP was rs2889611 within mitogen-activated protein kinase 8 (*MAPK8*; $p = 2.70 \times 10^{-5}$; odds ratio [OR], 0.82; confidence interval [CI], 0.75 to 0.90), and rs1356415 from *PROM1* and rs4536994 from *KDR* were strongly associated with hypertension ($p = 3.45 \times 10^{-4}$; OR, 1.18; CI, 1.08 to 1.29 and $p = 3.73 \times 10^{-4}$; OR, 1.19; CI, 1.08 to 1.31, respectively).

Functional analysis of ROS-hypertension gene set

The 68 targets that were significantly associated with ROS and hypertension were tested for functional enrichment analysis. Three functional analysis tools, SciMiner, WebGestalt, and DAVID, identified 34 significantly over-represented biological functions in the Kyoto Encyclopedia of Genes and Genomes pathway [13, 14]. The most significant biological pathway from the 3 functional analysis tools was focal adhesion, involved in the cell communication pathway group (Table 4). The most frequent pathway group was cancer pathways (n = 9), such as glioma and pancreatic

Table 3. Associated genes of hypertension ($p < 0.01$)

Chr no.	Start	End	Gene	Lowest p-value	No. of SNP	No. of papers
10	49,164,739	49,337,409	*MAPK8*	2.70×10^{-5}	19	13
4	15,554,385	15,714,766	*PROM1*	3.45×10^{-4}	48	2
4	55,619,416	55,706,519	*KDR*	3.73×10^{-4}	7	2
4	23,382,742	23,520,798	*PPARGC1A*	0.002	19	4
17	55,305,225	55,402,564	*RPS6KB1*	0.002	5	2
18	58,921,559	59,158,341	*BCL2*	0.004	38	5
13	77,347,625	77,411,751	*EDNRB*	0.004	10	5
14	92,439,198	92,491,389	*CHGA*	0.004	7	2
6	24,516,384	24,617,829	*GPLD1*	0.004	22	8
4	106,829,390	107,008,331	*GSTCD*	0.005	6	4
1	157,928,704	157,971,003	*CRP*	0.006	5	10
3	180,329,005	180,455,189	*PIK3CA*	0.007	9	15
4	24,386,183	24,431,564	*SOD3*	0.007	9	24
9	94,395,287	94,492,368	*IPPK*	0.007	11	2
10	6,061,835	6,164,278	*IL2RA*	0.007	26	2
12	10,182,171	10,236,057	*OLR1*	0.008	6	3
20	8,041,296	8,833,547	*PLCB1*	0.009	153	4

SNP, single-nucleotide polymorphism.

Table 4. Top 10 most significant pathways through pathway enrichment analysis with 68 genes

Pathways	SciMiner	WebGestalt	DAVID	Group of pathway (level 1)	Group of pathway (level 2)
Focal adhesion	9.95×10^{-18}	4.23×10^{-20}	5.80×10^{-7}	Cellular processes	Cell communication
ErbB signaling pathway	1.76×10^{-15}	2.49×10^{-17}	5.50×10^{-7}	Environmental information processing	Signal transduction
Glioma	3.51×10^{-13}	1.31×10^{-14}	5.90×10^{-6}	Human diseases	Cancers
Pancreatic cancer	1.02×10^{-12}	2.77×10^{-14}	1.30×10^{-5}	Human diseases	Cancers
Fc epsilon RI signaling pathway	1.15×10^{-12}	5.46×10^{-14}	2.10×10^{-5}	Organismal systems	Immune system
Renal cell carcinoma	2.84×10^{-11}	1.45×10^{-12}	9.10×10^{-5}	Human diseases	Cancers
Colorectal cancer	1.22×10^{-10}	5.85×10^{-12}	2.50×10^{-4}	Human diseases	Cancers
Non-small cell lung cancer	2.56×10^{-10}	1.49×10^{-11}	2.10×10^{-4}	Human diseases	Cancers
Prostate cancer	2.60×10^{-10}	8.45×10^{-12}	3.30×10^{-4}	Human diseases	Cancers
T cell receptor signaling pathway	2.40×10^{-10}	3.18×10^{-11}	7.80×10^{-4}	Organismal systems	Immune system

cancer. Eight other pathways related to signal transduction and 11 organismal system pathways (level 1) related to the immune system, endocrine system, and nervous system were significantly identified.

Discussion

Oxidative stress due to excess production of ROS is one of the reasons for the development of hypertension [4]. To identify genetic risk factors that induce hypertension through ROS, this study extracted ROS-hypertension-related genes using text-mining tools and investigated the association of genes with hypertension in 6,419 unrelated Koreans. *MAPK8*, *PROM1*, and *KDR* had strong association signals with hypertension ($p < 4 \times 10^{-4}$). Especially, *MAPK8* was published 13 times in ROS-hypertension articles, while most genes strongly associated with hypertension ($p < 0.01$) were published an average of 6.29 times.

MAPK8, known as *JNK1*, included 19 SNPs in the KARE genotype data, and 14 SNPs among the 19 SNPs were significantly associated with hypertension, ranging in p-value from 2.7×10^{-5} to 1.3×10^{-3}—moderate in comparison with a genome-wide association study (GWAS)-significant p-value (5.0×10^{-8}); thus *MAPK8* was not considered as the candidate gene of hypertension in previous GWAS studies [15, 16]. *MAPK8* plays a key role in T cell proliferation, apoptosis, and differentiation through the studies of Jnk1-deficient mice [17, 18]. *MAPK8* was included on HumanCVD Beadchip, a customized cardiovascular disease (CVD) SNP chip containing more than 2,100 CVD candidate genes [19]. However, previous cardiovascular disease GWASs regarding high-density lipoprotein particle features, lipids, and apolipoproteins did not report the association of the *MAPK8* gene [20, 21]. Therefore, it needs replication to make it sure whether *MAPK8* is indeed involved in the development of hypertension through ROS.

Using the text-mining tool, we found 237 ROS-hypertension-related genes. The most frequent gene was *AGT* (angiotensinogen [serpin peptidase inhibitor, clade A, member 8]), which was reported on 320 ROS-hypertension articles, but it was not associated with hypertension in this study or our previous report [22]. Most of the genes that were published in more than 40 articles were not associated with hypertension or showed weak associations; 6 of 17 genes were significant, and the lowest p-value was 0.014 (nitric oxide synthase 1 [neuronal], *NOS1*). The average number of articles for genes with strong signals ($p < 0.01$) was 6.29 articles, and that for those with moderate signals ($0.01 \le p < 0.05$) was 18.90 articles.

Two large GWASs, the International Consortium for Blood Pressure Genome-Wide Association Studies (IC-BPGWAS) [23] and Asian Genetic Epidemiology Network Blood Pressure (AGEN-BP) [16], reported 33 blood pressure candidate loci in 2011. Among 66 genes within the 33 blood pressure candidate loci, 6 genes were included in the ROS-hypertension gene set as follows: *NPPA*, *NPPB*, *PTPN11*, *CYP1A1*, *GNAS*, and *EDN3*. We examined their association with hypertension by case control study, and *NPPA*, *NPPB*, and *CYP1A1* were associated with hypertension with $p < 0.05$. The weakly associated SNP rs1023252 ($p = 0.047$) overlapped with *NPPA* and *NPPB*, and rs2472299 within the *CYP1A1* locus was previously mentioned for the oxidative stress pathway from WikiPathway (http://www.wikipathways.org).

In conclusion, we listed ROS-hypertension genes that were extracted by a text-mining approach and tested their association with hypertension in Korean population. Several genes, including the *MAPK8* gene, were identified as potential genes causing hypertension through ROS. This study demonstrates that a text-mining approach combined with association analysis may be useful to identify candidate genes that cause hypertension through ROS or oxidative stress.

Acknowledgments

This research was performed within the Consortium for Large-Scale Genome-Wide Association Study III (no. 2011 E7300400), which was supported by the genotyping data (the Korean Genome Analysis Project no. 4845-301) and the phenotypic data (the Korean Genome Epidemiology Study no. 4851-302) from the Korea Center for Disease Control. This work was supported by the Basic Science Research Program through a National Research Foundation of Korea (NRF) grant, funded by the Korean government (MEST) (no. 2010-0012080) and (MEST) (no. 2012-0009384).

References

1. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL Jr, *et al.* Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension* 2003;42:1206-1252.
2. Saavedra JM. Studies on genes and hypertension: a daunting task. *J Hypertens* 2005;23:929-932.
3. Fields LE, Burt VL, Cutler JA, Hughes J, Roccella EJ, Sorlie P. The burden of adult hypertension in the United States 1999 to 2000: a rising tide. *Hypertension* 2004;44:398-404.
4. Touyz RM, Briones AM. Reactive oxygen species and vascular biology: implications in human hypertension. *Hypertens Res* 2011;34:5-14.
5. Romanowski A, Murray JR, Huston MJ. Effects of hydrogen peroxide on normal and hypertensive rats. *Pharm Acta Helv* 1960;35:354-357.

6. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, *et al*. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 2009;41:527-534.
7. Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* 2006;22:7-12.
8. Hur J, Schuyler AD, States DJ, Feldman EL. SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics* 2009;25:838-840.
9. Duncan D, Prodduturi N, Zhang B. WebGestalt2: an updated and expanded version of the Web-based Gene Set Analysis Toolkit. *BMC Bioinformatics* 2010;11(Suppl 4):P10.
10. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* 2005;33:W741-W748.
11. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009;37:1-13.
12. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4:44-57.
13. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27-30.
14. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;40:D109-D114.
15. Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, *et al*. Genome-wide association study of blood pressure and hypertension. *Nat Genet* 2009;41:677-687.
16. Kato N, Takeuchi F, Tabara Y, Kelly TN, Go MJ, Sim X, *et al*. Meta-analysis of genome-wide association studies identifies common variants associated with blood pressure variation in east Asians. *Nat Genet* 2011;43:531-538.
17. Dong C, Yang DD, Wysk M, Whitmarsh AJ, Davis RJ, Flavell RA. Defective T cell differentiation in the absence of Jnk1. *Science* 1998;282:2092-2095.
18. Dong C, Yang DD, Tournier C, Whitmarsh AJ, Xu J, Davis RJ, *et al*. JNK is required for effector T-cell function but not for T-cell activation. *Nature* 2000;405:91-94.
19. Keating BJ, Tischfield S, Murray SS, Bhangale T, Price TS, Glessner JT, *et al*. Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS One* 2008;3:e3583.
20. Talmud PJ, Drenos F, Shah S, Shah T, Palmen J, Verzilli C, *et al*. Gene-centric association signals for lipids and apolipoproteins identified via the HumanCVD BeadChip. *Am J Hum Genet* 2009;85:628-642.
21. Kaess BM, Tomaszewski M, Braund PS, Stark K, Rafelt S, Fischer M, *et al*. Large-scale candidate gene analysis of HDL particle features. *PLoS One* 2011;6:e14529.
22. Song SB, Jin HS, Hong KW, Lim JE, Moon JY, Jeong KH, *et al*. Association between renin-angiotensin-aldosterone system-related genes and blood pressure in a Korean population. *Blood Press* 2011;20:204-210.
23. International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, *et al*. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 2011;478:103-109.