G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2014;12(4):165-170
http://dx.doi.org/10.5808/GI.2014.12.4.165

ORIGINAL ARTICLE

Short Reads Phasing to Construct Haplotypes in Genomic Regions That Are Associated with Body Mass Index in Korean Individuals

Kichan Lee, Seonggyun Han, Yeonjeong Tark, Sangsoo Kim*

Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea

Genome-wide association (GWA) studies have found many important genetic variants that affect various traits. Since these studies are useful to investigate untyped but causal variants using linkage disequilibrium (LD), it would be useful to explore the haplotypes of single-nucleotide polymorphisms (SNPs) within the same LD block of significant associations based on high-density variants from population references. Here, we tried to make a haplotype catalog affecting body mass index (BMI) through an integrative analysis of previously published whole-genome next-generation sequencing (NGS) data of 7 representative Korean individuals and previously known Korean GWA signals. We selected 435 SNPs that were significantly associated with BMI from the GWA analysis and searched 53 LD ranges nearby those SNPs. With the NGS data, the haplotypes were phased within the LDs. A total of 44 possible haplotype blocks for Korean BMI were cataloged. Although the current result constitutes little data, this study provides new insights that may help to identify important haplotypes for traits and low variants nearby significant SNPs. Furthermore, we can build a more comprehensive catalog as a larger dataset becomes available.

Keywords: genome-wide association study, haplotypes, Korea, NGS, phasing, single-nucleotide polymorphism

Introduction

Genome-wide association studies (GWASs) have been a useful tool to identify genetic variants that affect various traits [1]. Numerous novel important genetic variants associated with disease susceptibility have been identified through GWASs [2]. However, genome-wide association (GWA) results for complex diseases can generally explain only a small part of the genetic variants for complex diseases [3]. Generally, a GWA study identifying an association between a trait and a genetic variation may be limited in understanding complex diseases involving multiple functional loci [4].

Recently, many approaches to detect an association between a trait and one or multiple genetic variants have been proposed to study numerous data from GWASs [5]. Inference about linkage disequilibrium (LD), known as non-independent association of alleles at different loci, provides information for the association of genetic variants affecting complex traits [6]. Using LD analysis, it is possible to characterize multiple genomic variants associated with phenotypes in terms of haplotypes. A haplotype defines a combination of phased alleles in a chromosomal region [4]. Thus, haplotype analysis is useful to understand multiloci genetic associations and to identify susceptibility loci for diseases. In spite of the advantages of haplotype analysis, performing an analysis on a genome-wide scale is not simple, due to the uncertainty and complexity of haplotypes [7].

Whole-genome sequencing (WGS) of a reference population may provide insights into the potential causal variants hidden within the LD block of interest. While the 1000 Genomes Project aims to provide such information in worldwide populations or ethnic groups [8], it would be more informative to focus on a reference population on a national scale than on a global scale. Recently, several Korean

Received October 17, 2014; Revised November 11, 2014; Accepted November 20, 2014

***Corresponding author:** Tel: +82-2-820-0457, Fax: +82-2-824-4383, E-mail: sskimb@ssu.ac.kr

This is 2014 KOBIC best paper awarded.

reference WGS datasets have been published based on next-generation sequencing (NGS) platforms [9]. Here, we tried to construct haplotype blocks affecting body mass index (BMI) in Korean individuals through integrative analysis of NGS and GWA data. We analyzed haplotypes using NGS data only in ranges that were nearby significant single-nucleotide polymorphisms (SNPs) that were identified by GWA results. The analysis may have several advantages in haplotype analysis. The complexity and the time can be reduced by analyzing smaller regions that are known to harbor significant GWA signals. We cataloged all possible 44 haplotype blocks within the LD blocks of significant SNPs for BMI.

Methods

Overall pipeline of the analysis

The NGS raw data were mapped to reference sequences using an alignment tool. The significant SNPs for BMI were selected from the GWA result data. The LD blocks that encompassed the significant SNPs were searched with Haploview [10], and those NGS reads that were mapped to the LD ranges were selected with Samtools. For each LD block, the heterozygous variants from the short reads were phased with Samtools. Finally, the phased haplotypes harboring the significant SNPs were cataloged (Fig. 1).

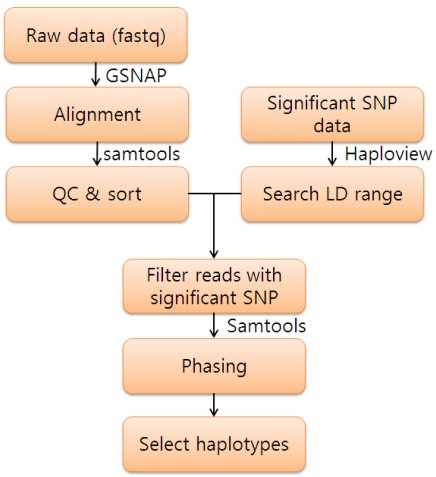

Fig. 1. Pipeline for this study. First, next-generation sequencing raw data (fastq) are aligned using Genomic Short-read Nucleotide Alignment Program (GSNAP), and then, the reads without bad mapping quality are selected. In addition, significant single-nucleotide polymorphisms (SNPs) are learned from the genome-wide association study results, and linkage disequilibrium (LD) ranges with the SNPs are searched by HaploView. Using Samtools, reads within the LD ranges are selected and phased. As a result, a haplotype catalog can be constructed, including haplotypes with significant SNPs.

Genotyping, imputation, and GWA analysis

The Korean samples and genotype data used in this study have been described by Cho *et al.* [11]. Briefly, through the Korea Association Resource (KARE) project, a total of 10,038 participants were recruited from Ansan and Anseong population-based cohorts, aged 40 to 69. The genotypes that had been measured on Affymetrix Genome-Wide Human SNP array 5.0 (Affymetrix, Santa Clara, CA, USA) were filtered for quality control, resulting in 8,842 samples and 352,228 markers. The genotype data were expanded to a total of 1,827,004 SNPs through imputation with IMPUTE [12] using the Japanese in Tokyo, Japan (JPT)/Han Chinese in Beijing, China (CHB) component of HapMap as the reference.

BMI traits were tested for association by linear regression analysis with an additive model after adjustments for recruitment region, age, and sex as covariates using PLINK. The threshold of genome-wide significance was set at $p < 5 \times 10^{-4}$. The significant SNPs were clustered into LD blocks ($r^2 > 0.9$) using HaploView [13].

NGS data analysis

We downloaded previously published whole-genome NGS data from the TIARA database, sequenced at Seoul National University College of Medicine (TIARA; http://tiara.gmi.ac.kr) [14]. The short reads from seven samples (AK3, AK4, AK5, AK6, AK7, AK14, and AK20) were mapped to the human reference genome (hg19) using the alignment algorithm in Genomic Short-read Nucleotide Alignment Program (GSNAP) [15]. From the resulting BAM file, we selected the mapped reads within the LD blocks that encompassed the significant SNPs from the BMI GWA result using the Samtools view function [16], creating a separate BAM file for each LD block.

Heterozygous SNP calling and phasing

The heterozygous SNPs were phased using the phase function in Samtools, version 0.1.19-44428cd [17]. The Samtools phase algorithm calls heterozygous SNPs automatically and phases those that segregate on the same DNA fragment as inferred from paired-end read information in the BAM file. We used the default options for the Samtools phase function. We confirmed the heterozygous SNPs through independent calling using the Samtools mpileup function [17]. Whenever an LD block split into multiple phased haplotype blocks, we kept those that encompassed the significant SNPs from the GWAS.

Table 1. Representative LD blocks from the GWA results[a]

CHR	LD No. from GWAS	SNP from GWAS	SNP ID in LD	Locus	Manimum/Maximum
1	3	rs10753250	rs10753250	31973266	31973266/31977853
			rs10158101	31975641	
			rs10914422	31977853	
1	5	rs7542777	rs7542777	33788884	33789670/33796229
			rs11061	33789670	
			rs1130800	33789968	
			rs11554674	33790496	
			rs7512470	33796229	

LD, linkage disequilibrium; GWA, genome-wide association; CHR, chromosome; GWAS, genome-wide association study; SNP, single-nucleotide polymorphism; BMI, body mass index; KARE, Korea Association Resource.
[a]The LD ranges include SNPs that are significantly associated with BMI; 435 loci significant SNPs were analyzed from the KARE genotype data.

Results

GWA analysis

The genotype data of 8,842 samples in a total of 10,038 participants in the KARE were analyzed to find significant SNPs associated with BMI in Korean individuals. Among the total of 1,827,004 loci, 435 loci were identified using a p-value cutoff ($p < 5 \times 10^{-4}$) from the GWAS. The default algorithm in the HaploView program for identifying LD blocks was taken from confidence intervals [13]. We set a ±20-kb window option and $r^2 > 0.9$ to capture LD blocks of 435 significant SNPs based on the KARE genotype data. For 435 loci that were significant in the association analysis ($p < 5 \times 10^{-4}$), 53 LD blocks were identified using HaploView (Table 1). These 53 LD blocks contained all 435 significant loci. Thus, a typical LD block included multiple significant loci.

Phasing short reads

We downloaded whole-genome paired-end NGS data of seven samples from the TIARA database to discover haplotypes with the SNPs associated with BMI within the 53 LD blocks from the GWA analysis. We mapped the short reads to the human reference genome (hg19), and those reads with bad mapping quality (mapping quality <5) were filtered out. The average mapping rate was about 93%. We separated the mapped reads into the 53 LD blocks for each of seven individuals (Table 1). Each group of separated reads was called for variants, and the heterozygous SNPs were phased using the phase function in Samtools. Not all of the SNPs in an LD block were always phased into a contiguous haplotype block. Whenever multiple haplotype blocks were found for a given LD block, we kept only those that encompassed the significant SNPs identified from the GWAS.

Construction of the haplotype blocks

After phasing heterozygous SNPs using Samtools, we searched for phased haplotypes that harbored significant risk alleles for BMI from the GWA analysis results. We found 44 haplotype blocks in 23 of 53 LD blocks. We could not find haplotypes with significant risk alleles in the rest of the LD blocks. We ordered 44 haplotypes and established a catalog for possible haplotype blocks associated with BMI variations in the Korean population (Table 2). For example, AK4 and AK7 share the TCTGAGCC haplotype, which comprises the variants at bps 246142250, 246142279, 246144436, 24614-6137, 246146178, 246146444, 246148791, and 246149166 in chromosome 1. Bp 57895600 in chromosome 18 has been previously known as rs8089366 (G/T) in the dbSNP database and was a significant SNP for BMI in KARE. On the other hand, bp 57900630, which participated in the same haplotype, has not been registered in the dbSNP database. In Table 2, 16 LD blocks were shared by several individuals, while the rest of the LD blocks were carried by single individuals.

Distribution of haplotype blocks

We calculated the average space between alleles in a haplotype block from our results. For example, AK4, AK5, and AK7 share CGCC at bps 182065282, 182065314, 18206-5341, and 182065351 in chromosome 4 (Table 2). The haplotype block ranges from bp 182065282 through bp 182065351; the length of the haplotype is 69 bp. Since there are four SNPs, the average space is about 17. We plotted the distribution of the average spaces between SNPs within a haplotype (Fig. 2). The number of alleles found for a haplotype is also plotted in Fig. 2.

Table 2. Haplotypes with significant SNPs[a]

CHR	LD	Sample	Haplotype	Locus
1	3	AK20	CT	33820033/**33820134 (rs12026290)**
	5	AK4	GTXT	177820861/**177821366 (rs3131313)**/177822084/177823423/177824537
		AK7	CGTCTX	
	6	AK4	TCTGAGCC	246142250/246142279/**246144436 (rs4654179)**/246146137/**246146178**
		AK6	TCTGAGCC	**(rs1538293)**/246146444/**246148791 (rs4654180)**/246149166
	6	AK4	CTA	**246152107 (rs13376134)**/246152121/246152129
		AK6	CXA	
		AK7	CTA	
	6	AK4	TA	**246141157 (rs12024270)**/246141458
		AK6	TA	
		AK7	TA	
2	8	AK5	TT	**222280602 (rs13013934)**/222281275
4	11	AK6	GT	**53133408 (rs729476)**/53134293
	12	AK7	TC	**133942227 (rs13122167)**/133942537
	16	AK4	CGCC	**182065282 (rs6824854)/182065314 (rs6825217)/182065341**
		AK5	CGCC	**(rs6824888)**/182065351
		AK7	CGCC	
	16	AK5	CCGG	182065152/182065165/**182065180 (rs6824986)/182065232 (rs4637448)**
		AK7	CCGG	
5	18	AK6	AGC	124683338/124683352/**124683815 (rs10060296)**
	18	AK6	TG	**124702595 (rs12654336)**/124702788
	18	AK6	CA	124708465/**124709413 (rs925896)**
	18	AK6	CA	124719059/**124720340 (rs1988043)**
	19	AK7	GG	160212103/**160212114 (rs11135120)**
	21	AK7	AC	**169356966 (rs6896240)**/169357240
7	25	AK5	ACTTC	**135329004(rs6979439)**/135329107/135329269/**135329690**
		AK6	ACTCC	**(rs12540688)**/135332379/
	25	AK6	GC	**135333854 (rs12540273)**/135334111
		AK7	GC	
	25	AK5	TGA	135325910/**135325947 (rs11984203)**/135326104
		AK6	CGG	
	25	AK5	GT	135334991/**135335373 (rs4291211)**
		AK6	XT	
		AK7	XT	
	25	AK7	GT	**135328689 (rs6975251)**/135328873
		AK6	GX	
8	26	AK6	CAA	**5172935 (rs6993835)**/5172992/5173185
	26	AK5	ATTC	5176649/5176753/5176782/**5176920 (rs1004161)**
		AK6	ATTC	
	27	AK6	AAAACGCA	**5211617 (rs7819482)**/5212030/5212240/5212243/**5212413** **(rs2052334)**/5212537/**5212562 (rs7846604)**/5212579
12	33	AK6	GGG	**128182752 (rs10773418)**/128183164/128184294
13	34	AK4	XTCG	55551581/55552706/55552879/**55553277 (rs9536931)**
		AK6	TTCG	
14	36	AK4	GGGG	55158376/55158424/55158851/**55160290 (rs2884684)**
		AK6	GXAG	
	36	AK7	CC	**55157430 (rs10483637)**/55158376
15	38	AK6	GCAT	94581157/94581942/**94581996 (rs2388442)**/94582024
	39	AK6	TCG	94590208/**94590228 (rs1031912)**/94591144

SNP, single-nucleotide polymorphism; CHR, chromosome; LD, linkage disequilibrium; BMI, body mass index; GWA, genome-wide association.

[a]The underlined letters represent alleles without haplotype-phased results data but were filled in manually from the mapped data. The bold letters represent SNPs that are significantly associated with BMI from the GWA analysis results, and the non-bold ones do not exist in the GWA analysis results.

Table 2. Continued

CHR	LD	Sample	Haplotype	Locus
18	41	AK4	TACTC	**4078560 (rs12967692)/4078698**
		AK5	TACCT	**(rs12967842)**/4078802/4078845/4078852
	41	AK4	CCTC	4079990/4081009/4081017/**4081867 (rs6506177)**
	41	AK5	TAG	4084949/**4085221 (rs6506178)**/4085466
	41	AK5	TGG	4087117/4087308/**4087331 (rs7506389)**
	44	AK4	TCCT	57826273/57829899/**57830095 (rs649721)/57831468 (rs633265)**
	44	AK4	CACA	57835406/57836482/**57836715 (rs2051312)**/57837028
	44	AK4	CC	**57843875 (rs559623)**/57844375
	45	AK4	CCGT	57871154/57872075/**57872361 (rs2051312)/57872449 (rs8084515)**
	45	AK7	CG	57872956/**57872989 (rs12955983)**
	46	AK4	XAXC	57903745/**57904011 (rs590654)/57904088 (rs590215)**/57907787
		AK7	GATX	
	46	AK4	TT	**57896038 (rs477181)**/57896924
	46	AK4	TT	**57898600 (rs8089366)**/57900630
	46	AK7	GG	57893330/**57893618 (rs8083289)**
	48	AK4	AC	57945105/**57945953 (rs11152219)**

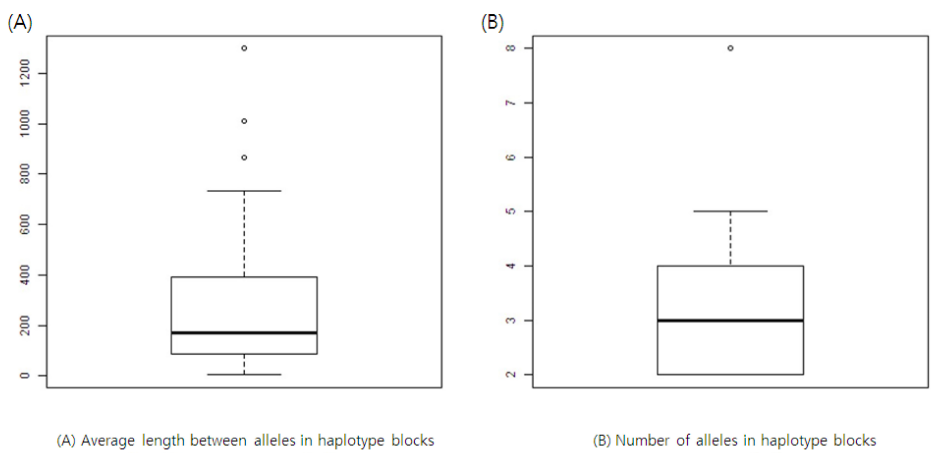

Fig. 2. Boxplots of average space between alleles (A) and the number of alleles in haplotype blocks (B).

Discussion

The GWAS is a powerful method to identify genomic variants affecting traits. Many studies have found important SNPs associated with diseases. However, the study is not well suited to identify rare variants or small effects of several SNPs [18]. Here, we tried to find haplotype blocks in 7 Korea individuals that harbor SNPs that are significantly associated with BMI in a Korean population. The significant SNPs for BMI were detected through GWASs, and the haplotypes were discovered by analyzing the NGS data. We selected short reads within the LD ranges that encompassed the significant SNPs from the GWAS results, and the heterozygous variants that were called from those reads were phased, based on the paired-end information. As a result, we detected 44 haplotype blocks harboring SNPs that are significantly associated with BMI in Koreans. The haplotypes may affect BMI in Korea individuals. However, it is not statically powerful, due to the small number of samples (7 individuals) used in this study. Actually, the alleles of a haplotype need to be compared with those of the haplotype of the same range in other individuals to have high accuracy. We could not follow this strategy. For example, most of the haplotypes found by this analysis were discovered in a single individual. In our small dataset, a significant SNP from a GWAS is not likely to be shared by several individuals, because many SNPs have a minor allele frequency of around 10% in a population. This problem may be solved if a large dataset is used. In addition, read depth also reduces accuracy. In Table 2, several haplotypes were identified in a single GWA LD block. These haplotypes could not be joined, as there were not enough short reads that could bridge them using the phase function in Samtools. In addition, AK7 has the GATX haplotype at bps 57903745, 57904011 (rs590654), 57904088 (rs590215), and 57907787 in chromosome 18, and AK4 has XAXC in the same locus. An X represents a

missing allele, probably due to low coverage. We were able to recover them, as Samtools was still able to phase them. We filled in these alleles manually to build the haplotype. In Table 2, the green letters represent such alleles that have been filtered by read depth and recovered manually. Although the accuracy is not high, this integrative method has several advantages. First, rare variants can be detected. For example, AK4 has the TT risk allele haplotype at bps 57895600 and 57900630 in chromosome 18. The former was associated with BMI and was registered in the dbSNP database, while the latter was not registered in the dbSNP database. Bp 57900630 is possibly a low-frequency risk allele, which is difficult to be detected by GWAS. Second, we can explain a trait systematically. rs559623 is an SNP (C/A: forward strand) that is associated with BMI. It may act with bp 57844375 in chromosome 18 as CC according to our results (Table 2). Traits can be explained properly by multiple variants. Finally, this method is useful for studying phasing. Many SNPs from an array or NGS reads are unphased genotypes. However, the bulk of SNP information needs to be phased for identifying co-located alleles. While GATK HaplotypeCaller, Samtools Phase, and Beagle are used to phase variants, there are several problems, such as the long execution time and the need for large system memory. In fact, it takes too long to phase entire chromosomes or a single chromosome. However, with this method, it is possible to analyze efficiently with regard to time and memory, as it does not consider reference and genotype data of the entire chromosome. This method focuses only on the data within regions that are nearby important variants from the GWA analysis results. We constructed a crucial haplotype catalog for BMI traits in Korean individuals by integratively analyzing NGS data and GWA analysis data.

Acknowledgments

The genotype and phenotype data were kindly provided by the Korea National Institute of Health, Centers for Disease Control and Prevention, the Republic of Korea. Financial support for this work was made available by the National Research Foundation of Korea (NRF-2012M3A9D1054705), funded by the Ministry of Education, Science, and Technology.

References

1. Guo J, Jorjani H, Carlborg Ö. A genome-wide association study using international breeding-evaluation data identifies major loci affecting production traits and stature in the Brown Swiss cattle breed. *BMC Genet* 2012;13:82.
2. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. *BMC Med Genet* 2009;10:6.
3. Cusanovich DA, Billstrand C, Zhou X, Chavarria C, De Leon S, Michelini K, *et al*. The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Hum Mol Genet* 2012;21:2111-2123.
4. Crosslin DR, Qin X, Hauser ER. Assessment of LD matrix measures for the analysis of biological pathway association. *Stat Appl Genet Mol Biol* 2010;9:Article35.
5. Hendricks AE, Dupuis J, Gupta M, Logue MW, Lunetta KL. A comparison of gene region simulation methods. *PLoS One* 2012;7:e40925.
6. Wang M, Jia T, Jiang N, Wang L, Hu X, Luo Z. Inferring linkage disequilibrium from non-random samples. *BMC Genomics* 2010;11:328.
7. Song C, Chen GK, Millikan RC, Ambrosone CB, John EM, Bernstein L, *et al*. A genome-wide scan for breast cancer risk haplotypes among African American women. *PLoS One* 2013;8:e57298.
8. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061-1073.
9. Ju YS, Kim JI, Kim S, Hong D, Park H, Shin JY, *et al*. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat Genet* 2011;43:745-752.
10. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; 21:263-265.
11. 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.
12. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;39:906-913.
13. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, *et al*. The structure of haplotype blocks in the human genome. *Science* 2002;296:2225-2229.
14. Hong D, Lee J, Bleazard T, Jung H, Ju YS, Yu SB, *et al*. TIARA genome database: update 2013. *Database (Oxford)* 2013; 2013: bat003.
15. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010;26: 873-881.
16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078-2079.
17. Osborne OG, Batstone TE, Hiscock SJ, Filatov DA. Rapid speciation with gene flow following the formation of Mt. Etna. *Genome Biol Evol* 2013;5:1704-1715.
18. Bailey KR, Cheng C. Conference Scene: The great debate: genome-wide association studies in pharmacogenetics research, good or bad? *Pharmacogenomics* 2010;11:305-308.