Genomics & Informatics Vol. 8(3) 131-137, September 2010

Identification of Causal and/or Rare Genetic Variants for Complex Traits by Targeted Resequencing in Population-based Cohorts

Yun Kyoung Kim, Chang Bum Hong and Yoon Shin Cho*

Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Seoul 122-701, Korea

Abstract

Genome-wide association studies (GWASs) have greatly contributed to the identification of common variants responsible for numerous complex traits. There are, however, unavoidable limitations in detecting causal and/or rare variants for traits in this approach, which depends on an LD-based tagging SNP microarray chip. In an effort to detect potential casual and/or rare variants for complex traits, such as type 2 diabetes (T2D) and triglycerides (TGs), we conducted a targeted resequencing of loci identified by the Korea Association REsource (KARE) GWAS. The target regions for resequencing comprised whole exons, exon-intron boundaries, and regulatory regions of genes that appeared within 1 Mb of the GWA signal boundary. From 124 individuals selected in population-based cohorts, a total of 0.7 Mb target regions were captured by the NimbleGen sequence capture 385K array. Subsequent sequencing, carried out by the Roche 454 Genome Sequencer FLX, generated about 110,000 sequence reads per individual. Mapping of sequence reads to the human reference genome was performed using the SSAHA2 program. An average of 62.2% of total reads was mapped to targets with an average 22X-fold coverage. A total of 5,983 SNPs (average 846 SNPs per individual) were called and annotated by GATK software, with 96.5% accuracy that was estimated by comparison with Affymetrix 5.0 genotyped data in identical individuals. About 51% of total SNPs were singletons that can be considered possible rare variants in the population. Among SNPs that appeared in exons, which occupies about 20% of total SNPs, 304 nonsynonymous singletons were tested with Polyphen to predict the protein damage caused by mutation. In total, we were able to detect 9 and 6 potentially functional rare SNPs for T2D and triglycerides, respectively, evoking a further step of replication genotyping in independent populations to prove their *bona fide* relevance to traits.

Keywords: genome-wide association study, type 2 diabetes (T2D), triglycerides (TG), targeted resequencing, exon

Introduction

Since the completion of the Human Genome Project (Venter *et al.*, 2001), millions of single nucleotide polymorphisms (SNPs) have been available for understanding the pattern of the human genome structure, resulting in the construction of a high-density haplotype map (International HapMap Consortium 2005). These advances in genomics, along with the development of high-throughput and high-density microarray-based genotyping platforms, have led genome-wide association (GWA) studies to seek genetic variations related to human diseases. Indeed, GWA studies have been highly effective in identifying common variants (MAF$>$5%), showing strong evidence of association with complex diseases and traits ($p<5\times10^{-7}$) (Cirulli *et al.*, 2010). Most common variants identified by GWA studies, however, have shown small effects on disease risk (e.g., odds ratio of 1.2) and low levels of heritability (Aulchenko *et al.*, 2009; Cirulli *et al.*, 2010; Zeggini *et al.*, 2008). These limitations in GWA studies imply that additional genetic variants (such as rare variants) remain to be detected to explain the missing heritability of complex traits. GWA studies have shown an additional limitation in identifying causal variants for complex traits as well, since this approach depends highly on LD-based tagging SNP microarray chips.

Next-generation sequencing (NGS) technologies, such as the Illumina Genome Analyzer (GA), Roche/454 FLX system, and ABI SOLiD system, recently have significantly improved throughput and reduced the cost remarkably as compared to capillary-based electrophoresis systems (Shendure *et al.*, 2004). For example, in a single experiment using the Roche/454 FLX, the sequences of approximately 100 million reads of up to 350 bases in length can be determined (Rothberg *et al.*,

*Corresponding author: E-mail yooncho33@korea.kr
Tel +82-2-380-2262, Fax +82-2-354-1063
Accepted 2 September 2010

2008). Therefore, NGS technologies are regarded as one of many revolutionary methods of large-scale personal genome sequencing. In this regard, NGS technologies are highly attractive for overcoming the limitations of GWASs by providing a way of fine-mapping to identify causal and/or rare variants. Some rare sequence variants revealed to date by whole genome sequencing or targeted sequencing using NGS technologies have been reported for their association with rare diseases, such as Mendelian disorders (Ng *et al.*, 2010), while a few are associated with complex diseases, such as type 2 diabetes (T2D) in the large-scale population.

Recently, we conducted the Korea Association REsource (KARE) project, the first Korea GWA study for complex diseases, including T2D, as well as for various quantitative traits, including anthropometric and blood biochemical traits (Cho *et al.*, 2009). From the KARE GWA study, we were able to identify numerous significant genome-wide common loci for diverse complex traits. Most of those loci, however, showed small effects on the traits of interest-even no functional relevance to the traits. Thus, in the hope of understanding missing heritability for complex traits of interest, we extended our findings from the KARE GWA study to the detection of causal and/or rare variants by performing targeted resequencing using the Roche 454 Genome Sequencer (GS) FLX instrument in 124 individuals, selected from Korean population-based cohorts.

Methods

Samples

For targeted resequencing, samples were selected from two population-based cohorts, the Ansung and Ansan cohorts, which were initiated as a part of the Korean Genome Epidemiologic Study (KoGES) in 2001. Sample selection was stratified to three groups in order to increase the chances of identifying causal and/or rare variants for T2D and plasma triglycerides. Group 1 samples included 40 T2D patients, diagnosed by 2006 WHO guidelines. Group 2 samples included 42 nondiabetic individuals, selected from the lower quartile of fasting plasma triglycerides (TGs) (44-56 mg/dl) in the study cohorts. Group 3 samples comprised 42 individuals, selected from the upper quartile of fasting plasma TGs (321-395 mg/dl) in the study cohorts. Thus, Group 2 samples were considered as common controls for Group 1 T2D cases as well as for Group 3 samples, which had high levels of plasma TGs. About 5 μg of genomic DNA per individual was used for subsequent targeted sequence capture and sequencing experiments.

Targeted capture and sequencing

Target regions for resequencing comprised the whole exon, exon-intron boundary, and regulatory region (10 kb upstream from the transcription start site) of genes that appeared within 1 Mb of the GWA signal boundary for complex traits, such as T2D and TG. Adding genes for BMI, waist-hip ratio (WHR), and bone density, based on the KARE study results, a total of 45 genes, covering about 670 kb, were selected as targets for resequencing (Table 1). To capture the targeted regions, the NimbleGen Sequence Capture 385K array method was applied (NimbleGen, Madison, WI, USA). Sequencing of capture regions was carried out by the Roche 454 GS FLX sequencer on a large PicoTiterPlate (Roche, Basel, Switzerland) according to the manufacturer's protocols. Image analysis and base calling were performed using the GS FLX pipeline software (version 2.3) with default parameters.

Read mapping

Human genome UCSC hg18 was used as the reference for the mapping of reads produced from sequencing experiments. Mapping of sequence reads to the human reference genome was performed using the SSAHA2 (version 2.5.2) program. The reference human genome was formatted by SSAHA2 using parameters *-454* for indexing/hashing before mapping. The mapping of sequence reads to the formatted reference was performed using SSAHA2. The best mapping per read was determined using the parameters *-454 -best 1*. Mapped reads were changed to Binary Sequence Alignment/Map (BAM) format by samtools (version 0.1.8) using the parameters *view -bt*. Mapping results were recalibrated by the GATK program using the parameters *-cov QualityscoreCovariate -cov DinucCovariate*.

Variant calling

The GATK module, unified genotyper, was used for variant calling via a two-step procedure. In the first step, a raw variant call was generated with the parameters --*platform ROCHE454 -confidence 50 --heterozygosity 1.000000e-03 -stand_call_conf 30.0*. In the second step, a filtered variant call was produced with the parameters --*filterExpression "QUAL<30.0 || AB>0.75 && DP>40 || QD<5.0 || HRun>5 || SB>−0.10*".

Accuracy test for sequence calls

As an accuracy test, some of the total sequence calls generated from 124 individuals were compared with

Table 1. Basic information of target regions for sequencing

Gene	NCBI Reference Sequence	Chromosome	Start	End	Size (bp)	No. of exon	Size of exon (bp)	Target region (bp)	Traits
IGF2BP2	NM_006548	chr3	186844220	187025521	181,301	16	3,671	20,171	T2D
CDKAL1	NM_017774	chr6	20642666	21339743	697,077	16	2,408	18,908	T2D
CDKN2A	NM_058195	chr9	21957750	21984490	26,740	3	1,151	17,087	T2D
CDKN2B	NM_078487	chr9	21992901	21999312	6,411	2	3,984	19,084	T2D
KCNQ1	NM_000218	chr11	2422796	2826916	404,120	16	3,246	20,107	T2D
DKFZp686O24166	NR_026750	chr11	17329892	17355445	25,553	5	6,375	21,775	T2D
KCNJ11	NM_000525	chr11	17363371	17366782	3,411	1	3,411	18,411	T2D
ABCC8	NM_000352	chr11	17371007	17455025	84,018	39	4,978	23,777	T2D
CCDC63	NM_152591	chr12	109769193	109829721	60,528	12	1,945	18,045	T2D
GDAP1L1	NM_024034	chr20	42309321	42342427	33,106	6	2,244	17,744	T2D
R3HDML	NM_178491	chr20	42399039	42413289	14,250	5	1,377	16,777	T2D
HNF4A	NM_001030004, NM_000457	chr20	42417854	42493444	75,590	10	3,239	19,674	T2D
GCKR	NM_001486	chr2	27573209	27600052	26,843	19	2,186	18,979	TG
ZNF512	NM_032434	chr2	27659396	27699467	40,071	14	3,426	19,726	TG
LPL	NM_000237	chr8	19840861	19869050	28,189	10	3,747	19,647	TG
ZNF462	NM_021224	chr9	108665198	108813617	148,419	13	8,295	24,495	TG
BUD13	NM_032725	chr11	116124100	116148914	24,814	10	2,191	18,091	TG
ZNF259	NM_003904	chr11	116154485	116163949	9,464	14	1,778	18,078	TG
KIAA0999	NM_025164	chr11	116219327	116474203	254,876	24	6,069	23,369	TG
MAP2K1	NM_002755	chr15	64466264	64570936	104,672	11	2,586	18,586	BMI
FTO	NM_001080432	chr16	52295375	52705882	410,507	9	4,294	20,094	BMI
MC4R	NM_005912	chr18	56189543	56190981	1,438	1	1,438	16,438	BMI
CTNNBL1	NM_030877	chr20	35755847	35933934	178,087	16	1,888	18,388	BMI
C12orf51	NM_001109662	chr12	111082493	111228421	145,928	69	14,089	35,880	WHR
PTPN11	NM_002834	chr12	111340918	111432100	91,182	16	6,283	22,783	WHR
HSD11B1	NM_181755	chr1	207926172	207974918	48,746	7	1,415	17,086	Bone density
ADAMTS16	NM_139056	chr5	5193442	5373412	179,970	23	4,974	22,174	Bone density
CTNND2	NM_001332	chr5	11024951	11957110	932,159	22	5,436	22,536	Bone density
HIVEP2	NM_006734	chr6	143114296	143308031	193,735	10	9,724	25,624	Bone density
TXNDC3	NM_016616	chr7	37854723	37906527	51,804	18	2,311	18,984	Bone density
SFRP4	NM_003014	chr7	37912059	37923050	10,991	6	2,973	29,165	Bone density
EPDR1	NM_017549	chr7	37926687	37958067	31,380	3	2,598	7,100	Bone density
FAM3C	NM_001040020	chr7	120776140	120823658	47,518	10	2,502	18,442	Bone density
PLXNA4	NM_020911, NM_181775	chr7	131458630	131911863	453,233	36	13,061	33,912	Bone density
VDR	NM_001017535	chr12	46521586	46585081	63,495	11	4,775	20,775	Bone density
NT5DC3	NM_001031701	chr12	102690210	102759105	68,895	14	7,214	23,514	Bone density
TMEM132B	NM_052907	chr12	124377114	124709542	332,428	9	7,578	23,378	Bone density
FLT3	NM_004119	chr13	27475410	27572729	97,319	24	3,842	21,115	Bone density
TNFSF11	NM_033012	chr13	42034871	42080149	45,278	7	2,359	18,412	Bone density
GPC5	NM_004466	chr13	90848935	92317486	1,468,551	8	2,878	18,578	Bone density
OSCAR	NM_133169	chr19	59289744	59295960	6,216	5	1,410	17,336	Bone density
Total					7,108,313	570	171,349	844,245	

TG: Triglyceride, T2D: Type 2 Diabetes, WHR: Waist Hip Ratio, BMI: Body Mass Index.

SNPs that were genotyped from the same individuals with the Affymetrix Genome-Wide Human SNP array 5.0 in the KARE project. Sequence calls for comparison were extracted by Samtools (parameter, *view -b*) based on genotype information of the SNPs that appeared in the targeted region.

Variant annotation

Variant annotation was performed by the GATK module GenomicAnnotator using information, including Human genome UCSC hg18, NCBI dbSNP build 130, and reference genes provided by UCSC. Annotated information

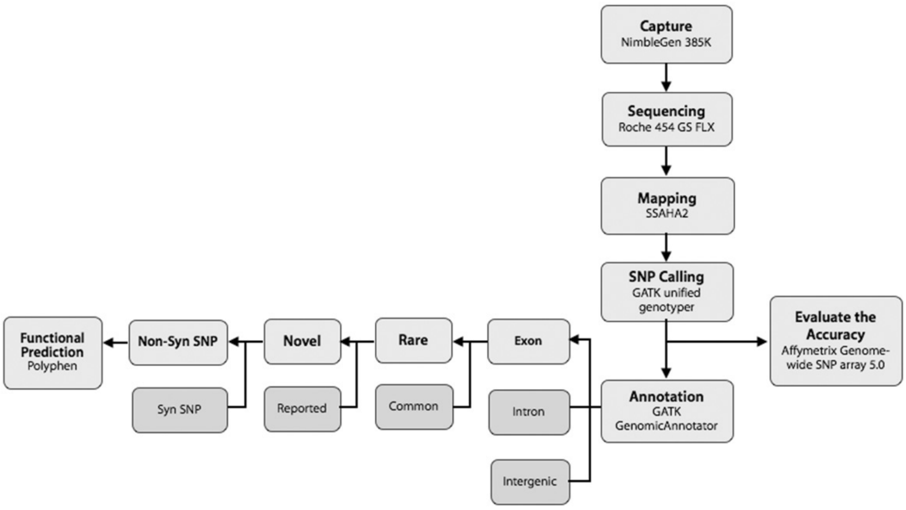

Fig. 1. Overall work flow to identify causal variants and rare variants for complex traits.

Table 2. The status of targeted sequence reads in 124 individuals

Group*	Sample number (N)	Total number of reads	Total number of reads aligned	Number of reads mapped in target region	% of reads mapped in target region
Group 1	40	110,145	106,334	66,021	63.2
Group 2	42	108,329	104,657	60,862	59.1
Group 3	42	106,565	103,194	65,142	64.4
Average		108,317	104,702	63,976	62.2

*Group 1, Group 2, and Group 3 correspond to T2D, shared normal controls, and individuals with high TGs, respectively.

included the dbSNP's rs IDs, SNP locations (eg, exon (3'-utr, 5'-utr) intron and intergenic regions), and SNP functions (eg, nonsynonymous SNPs and synonymous SNPs).

Results and Discussion

The GWA study has been considered a highly reliable method of identifying genetic markers associated with complex traits. However, it is usually agreed that GWA studies have limitations in understanding the total heritability for a given complex trait (Manolio *et al*., 2009). In an effort to explain the missing heritability of complex traits, the identification of causal variants and rare variants is a choice, along with the identification of structural variants, such as copy number variation (CNV) and epigenetic modifications.

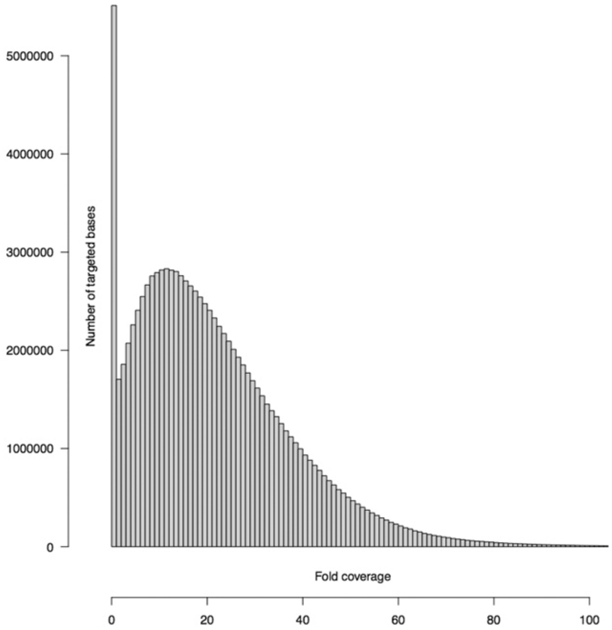

Fig. 2. The distribution of fold-coverage of mapping-targeted bases (670 kb) summed across the 124 individuals. The average fold-coverage for each individual was 22X.

Currently, we conducted a KARE GWA study for numerous complex traits, including T2D and fasting plasma triglycerides (TGs). To detect causal variants and rare variants for T2D and TGs, we performed targeted resequencing for genes located within a 1 Mb region of SNPs identified in the KARE GWA study for their strong evidence of association with T2D or TGs. Exons, exon-intron boundaries, and regulatory regions (10 kb upstream from the first exon) were targeted for sequencing. In addition to T2D- and TG-associated regions, we also included regions for BMI, WHR, and bone density for targeted resequencing. Samples for sequencing included 40 T2D cases (Group 1), 42 nondiabetic individuals with low TGs (Group 2), and 42 individuals with high TGs (Group 3). Group 2 samples were used as controls for T2D case samples as well as high-TG samples in this study (see METHODS).

The overall study involved several steps, such as target region capture, sequencing, mapping, SNP calling, SNP annotation, and functional prediction of identified SNPs with an *in silico* tool (Fig. 1). Target regions for the sequence capture included 41 genes covering a total size of 670 kb (Table 1). Sequencing for captured regions with the Roche 454 Genome Sequencer FLX generated an average of about 108,317 reads (one read=300-350 bp) per individual (Table 2). Sequence reads that passed the initial quality criteria were aligned to the reference human genome (hg18) using the SSAHA2 (version 2.5.2) program. It is estimated that 62.2% of the reads were mapped to the target region (Table 2), with a fold coverage of 22X (Fig. 2). For example, the mean depth information of the targeted gene, *HSD11B1*, on chromosome 1 in all 3 groups is demonstrated in Fig. 3.

Sequence variant (single nucleotide polymorphism, SNP) calling was carried out by the GATK unified genotyper module. The quality criteria for variant calling were 10X coverage by reads and base calls with a Phred-like quality score greater than 30 (Ewing and Green. 1998). The GATK Genomic Annotator module was applied to SNP annotations, based on UCSC hg18 and dbSNP build 130, generating information, such as dbSNP rs ID, SNP location, and SNP function. A total of 5,983 variants (average 846 variants per individual) were called and successfully passed our variation quality criteria. Seventeen percent of the annotated SNPs have not

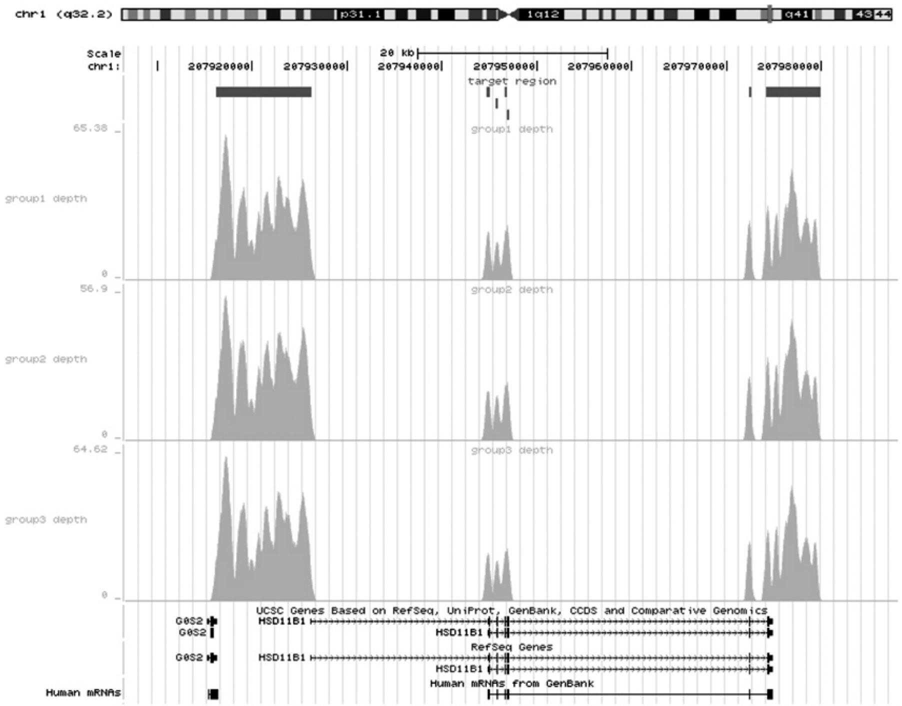

Fig. 3. Average depth of target regions for *HSD11B1*.

Table 3. Summary statistics of SNP calls per individual, annotated in target regions from 124 individuals

Group	Average number of SNP calls	number of SNPs reported in dbSNP DB	Percentage of novel SNPs	Type		Location			Function of exon SNPs			
				Number of heterozygous	Number of homozygous	Exon	Intron	Regulatory region	Non-syn	syn	utr3	utr5
Group 1	852	711	16%	557	295	166	165	521	30	41	87	7
Group 2	851	702	17%	536	315	172	162	517	31	43	91	7
Group 3	835	692	17%	517	318	166	159	510	31	42	87	7
Average	846	702	17%	536	310	168	162	516	31	42	88	7

been reported in the dbSNP database and are regarded as novel SNPs (Table 3). An average of 168 of the 846 SNP calls per individual appeared in exon regions, of which 31 SNPs (18.21%) were nonsynonymous SNPs that can lead amino acid changes (Table 3).

The accuracy of sequence calls was estimated by comparing them with array-based genotype data (Affymetrix Genome-Wide Human SNP array 5.0) in the identical individual. Concordance rates were 97.95% in the homozygous reference, 95.99% in the homozygous nonreference, and 95.46% in the heterozygous genotype (Table 4).

We counted genetic variants that were shared in each individual. About 51% of the total SNP calls (3,049 SNPs) were singletons (Fig. 4). To detect novel SNPs that can influence T2D and TGs, we selected 304 nonsynonymous SNPs, identified from candidate genes for T2D and TGs, and tested their relevance to the function of proteins-the expression products of candidate genes-by *in silico* functional analysis. Analyses by PolyPhen (Sunyaev *et al.*, 2001) predicted 9 novel nonsynonymous singleton SNPs for T2D and 6 for TGs that showed potentially damaging effects on candidate gene products for T2D and TGs (Table 5). Thus, it is considered that these SNPs are potentially causal and rare variants, possibly influencing T2D and fasting plasma TG levels. Validation of our findings in this study requires a repli-

Table 4. Concordance of sequence calls with genotype data obtained from Affymetrix 5.0 array chip experiments

Group	Concordance with Affymetrix 5.0 array chip data		
	Homozygous reference	Heterozygous	Homozygous nonreference
Group 1	58/58 (97.52)	31/32 (94.55)	23/22 (95.21)
Group 2	56/57 (98.04)	29/29 (96.06)	23/23 (96.11)
Group 3	58/58 (98.26)	30/30 (95.73)	23/23 (96.61)
Average	57/58 (97.95)	30/30 (95.46)	23/23 (95.99)

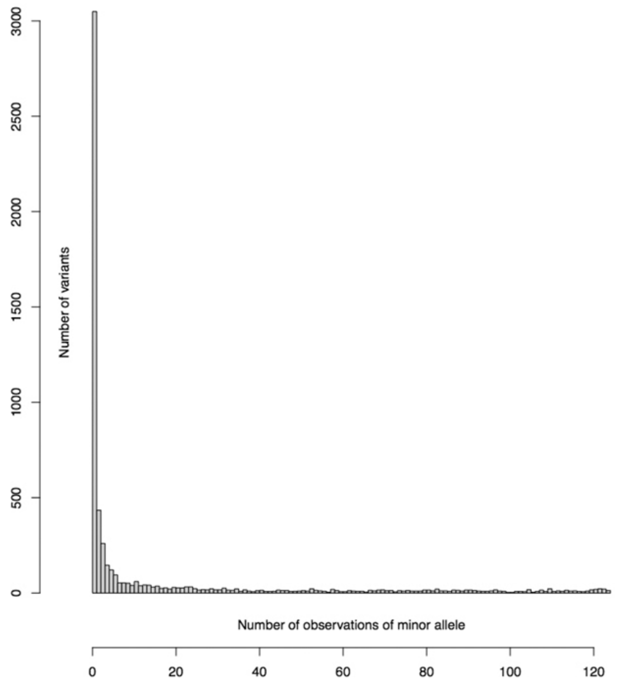

Fig. 4. Distribution of variant count.

Table 5. Results of in silico functional assay for selected novel variants

Chr	Position	Reference genotype	Observed genotype	Amino acidsubstitution	AA1	AA2	Gene	Trait	Functional prediction*
3	186847550	T	C	Q9Y6M1	E	G	*IGF2BP2*	T2D	Possibly damaging
6	21108512	G	A	Q5VV42	V	I	*CDKAL1*	T2D	Benign
9	21961192	T	C	P42771	S	G	*CDKN2A*	T2D	Benign
11	17365424	T	G	B2RC52	N	T	*KCNJ11*	T2D	Benign
11	17383680	A	C	Q09428	F	L	*ABCC8*	T2D	Probably damaging
11	2505772	T	C	P51787	L	P	*KCNQ1*	T2D	Probably damaging
12	109796040	G	A	Q8NA47	M	I	*CCDC63*	T2D	Possibly damaging
20	42399306	C	T	Q9H3Y0	A	V	*R3HDML*	T2D	Benign
20	42463539	A	C	P41235	N	T	*HNF4A*	T2D	Benign
2	27575135	C	T	Q14397	P	L	*GCKR*	TG	Benign
2	27684345	G	A	Q96ME7	A	T	*ZNF512*	TG	Benign
8	19862727	T	C	P06858	L	P	*LPL*	TG	Probably damaging
11	116133754	T	C	Q9BRD0	E	G	*BUD13*	TG	Probably damaging
11	116155748	T	C	O75312	M	V	*ZNF259*	TG	Benign
11	116166106	T	C	Q6Q788	D	G	*APOA5*	TG	Probably damaging

*Prediction of functional effect of mutation, derived from the substitution effect prediction algorithm PolyPhen. T2D: Type 2 Diabetes, TG: Triglyceride.

cation study for these SNPs in large samples of independent populations. Furthermore, a functional and clinical verification of 15 SNPs ultimately will be required to apply these findings to the treatment of T2D and hypertriglyceridemia.

Acknowledgments

This work was supported by grants from the Korea Centers for Disease Control and Prevention, Republic of Korea (4845-301, 4851-302, 4851-307).

References

Aulchenko, Y.S., Ripatti, S., Lindqvist, I., Boomsma, D., Heid, I.M., Pramstaller, P.P., Penninx, B.W., Janssens, A.C., Wilson, J.F., Spector, T., Martin, N.G., Pedersen, N.L., Kyvik, K.O., Kaprio, J., Hofman, A., Freimer, N.B., Jarvelin, M.R., Gyllensten, U., Campbell, H., Rudan, I., Johansson, A., Marroni, F., Hayward, C., Vitart, V., Jonasson, I., Pattaro, C., Wright, A., Hastie, N., Pichler, I., Hicks, A.A., Falchi, M., Willemsen, G., Hottenga, J.J., de Geus, E.J., Montgomery, G.W., Whitfield, J., Magnusson, P., Saharinen, J., Perola, M., Silander, K., Isaacs, A., Sijbrands, E.J., Uitterlinden, A.G., Witteman, J.C., Oostra, B.A., Elliott, P., Ruokonen, A., Sabatti, C., Gieger, C., Meitinger, T., Kronenberg, F., Döring, A., Wichmann, H.E., Smit, J.H., McCarthy, M.I., van Duijn, C.M., Peltonen, L., and ENGAGE Consortium. (2009). Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat. Genet.* 41, 47-55.

Cho, Y.S., Go, M.J., Kim, Y.J., Heo, J.Y., Oh, J.H., Ban, H.J., Yoon, D., Lee, M.H., Kim, D.J., Park, M., Cha, S.H., Kim, J.W., Han, B.G., Min, H., Ahn, Y., Park, M.S., Han, H.R., Jang, H.Y., Cho, E.Y., Lee, J.E., Cho, N.H., Shin, C., Park, T., Park, J.W., Lee, J.K., Cardon, L., Clarke, G., McCarthy, M.I., Lee, J.Y., Lee, J.K., Oh, B., and Kim, H.L. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* 41, 527-534.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Ozen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., and Lifton, R.P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. USA* 106, 19096-19101.

Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* 11, 415-425.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186-194.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F., McCarroll, S.A., and Visscher, P.M. (2009) Finding the missing heritability of complex diseases. *Nature* 461, 747-753.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., Shendure, J., and Bamshad, M.J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* 42, 30-35.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., Bamshad, M., Nickerson, D.A., and Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272-276.

Rothberg, J.M., and Leamon, J.H. (2008). The development and impact of 454 sequencing. *Nat. Biotechnol.* 26, 1117-1124.

Shendure, J., Mitra, R.D., Varma, C., and Church, G.M. (2004). Advanced sequencing technologies: Methods and goals. *Nat. Rev. Genet.* 5, 335-344.

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W. 3rd, Kondrashov, A.S., and Bork, P. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.* 10, 591-597.

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* 437, 1299-1320.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. *Science* 291, 1304-1351.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., Ardlie, K., Boström, K.B., Bergman, R.N., Bonnycastle, L.L., Borch-Johnsen, K., Burtt, N.P., Chen, H., Chines, P.S., Daly, M.J., Deodhar, P., Ding, C.J., Doney, A.S., Duren, W.L., Elliott, K.S., Erdos, M.R., Frayling, T.M., Freathy, R.M., Gianniny, L., Grallert, H., Grarup, N., Groves, C.J., Guiducci, C., Hansen, T., Herder, C., Hitman, G.A., Hughes, T.E., Isomaa, B., Jackson, A.U., Jørgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F.G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C.M., Lyssenko, V., Marvelle, A.F., Meisinger, C., Midthjell, K., Mohlke, K.L., Morken, M.A., Morris, A.D., Narisu, N., Nilsson, P., Owen, K.R., Palmer, C.N., Payne, F., Perry, J.R., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N.W., Rees, M., Roix, J.J., Sandbaek, A., Shields, B., Sjögren, M., Steinthorsdottir, V., Stringham, H.M., Swift, A.J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N.J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R.M., Weedon, M.N., Willer, C.J.; Wellcome Trust Case Control Consortium, Illig, T., Hveem, K., Hu, F.B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N.J., Barroso, I., Hattersley, A.T., Collins, F.S., Groop, L., McCarthy, M.I., Boehnke, M., and Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* 40, 638-645.