we did not observe statistically significant trends in means of $R^2$ values (P = 0.99) or in means of $|\beta|$ (P = 0.92).

## Multi-regulatory eQTLs

A *cis*-eQTL that is associated with expression of multiple genes might indicate the existence of a long-range enhancer/repressor that influences the expression of a cluster of genes in a region. We identified 6 *cis*-eQTLs that were each associated with expression levels of three or more mRNA-coding genes (Table 3). These multi-regulatory *cis*-eQTLs were each associated with the regulated transcripts in the same direction (Figure S4A).

A *trans*-eQTL that is associated with the expression of multiple genes is a potential master regulator. Our *trans*-eQTL map indicates that there are some *trans*-eQTL hotspots that were involved in multiple genes across the genome (Figure S5). We identified 5 *trans*-eQTLs that were each associated with three or more mRNA-coding genes (Table 3). Rs7801498 was also identified as a *cis*-eQTL for two genes (*LRWD1* and *ORAI2*). Notably, again, these multi-regulatory *trans*-eQTLs were each associated with the regulated transcripts in the same direction (Figure S4B).

## Replication analysis with independent studies

We compared our eQTLs to a meta-analysis of eQTL studies of whole blood samples conducted by Westra *et al.* [20]. They analyzed samples from 5,311 individuals from European populations. We focused on 15,733 genes that were commonly tested in both studies. At FDR<0.05, 10.6% of the genes were found *cis*-regulated in both studies; 60.9% of 2,750 *cis*-regulated genes in this study were replicated; and the concordance rate (*i.e.*, consistently *cis*-regulated or non-*cis*-regulated in both studies) was 68.8%. The concordance rate increased as FDR thresholds became more stringent up to 74.4% at FDR<1E-06. The replication rate of our *cis*-regulated genes was significantly associated with median non-adjusted expression levels (logistic regression $P$<2E-16, log OR = 0.14), but not with SD ($P$ = 0.14). 45.2% of 3,106 pairs of our *cis*-eQTLs (including SNPs in $r^2$>0.8) and genes tested in the meta-analysis were replicated. For replication of *trans*-eQTLs we found 978 distant SNP-transcript pairs in our results that corresponded to *trans*-eQTL-gene pairs identified in the meta-analysis. Six pairs were significant at $P$<5.1E-05, which corresponds to Bonferroni-corrected $P$ = 0.05 for 978 tests (Table S3). Particularly, *trans*-eQTL for *CALD1* was replicated at the original significance level ($P$ = 5.30E-16). Regarding that only the limited number of SNP-gene pairs were tested in common with the meta-analysis for *trans*-eQTLs, we also compared our *trans*-eQTLs with those identified for whole blood samples obtained from 76 Japanese individuals [17]. Over 8.6 billion tests for SNP-gene pairs were performed in both studies. Of the common tests, 41 and 2 pairs were identified as *trans*-eQTLs in the current and previous studies, respectively. We identified 1 *trans*-eQTL-gene pair exactly consistent between the studies (rs4487686 for *POLR2J4*). Regarding the number of performed tests, identifying one consistent result by chance is extremely unlikely (Fisher's exact test $P$<5E-08).

## Application of the eQTL map to interpretation of GWAS results

eQTL maps improve interpretation of GWAS results by linking SNPs and genes whose expressions are actually altered. We used previously published GWAS of Crohn's disease to comprehensively illustrate how our eQTL map improves interpretation of GWAS results. We identified 12 records for which our eQTL maps were informative for interpretation among all 220 records for Crohn's disease obtained from the NHRGI GWAS Catalog (http://www.genome.gov/gwastudies/) (Table 4). We define the following four informative cases for results of applying our eQTL map to GWAS results; a GWAS result is classified into Case 1 when the eQTL map may suggest different possible interpretation for GWAS, Case 2 when the eQTL map supports the interpretation provided by GWAS, Case 3 when the eQTL map helped to prioritize multiple genes inconclusively reported by the GWAS, or Case 4 when a *trans*-effect of GWAS-identified SNP was suggested (see supplementary note in File S3 for detailed definition).

For an example of Case 1, an intergenic SNP, rs694739, was identified in a GWAS of Crohn's disease (record 3 in Table 4); the study reported *PRDX5* and *ESRRA* as putative causative genes [21]. The GWAS-identified SNP was found in LD ($r^2$ = 0.85) with a *cis*-eQTL (rs600377) for *CCDC88B* in our eQTL map ($\beta$ = –0.26, $P$ = 1.0E-06). A *cis*-eQTL was identified for *PRDX5*, but the *cis*-eQTLs for *PRDX5* and *CCDC88B* were not in LD ($r^2$ = 0.01); and after correcting for the genotypes of the GWAS-identified SNP, the *cis*-effect on *CCDC88B* expression was not significant ($P_c$ = 0.51). Therefore, given the eQTL map, the most likely causative gene was *CCDC88B*. Based on our analyses, 6 of the 12 records were classified into Case 1, and thus, in each of these records, the eQTL-suggested gene should also be considered as another candidate gene.

Four of the 12 records were classified into Case 2. Three intergenic SNPs (rs7714584, rs11747270, rs13361189) were each reported in GWAS [21–23]; and in each study, *IRGM* was suggested as the candidate gene (records 8–10 in Table 4). All of these SNPs were each in perfect LD ($r^2$ = 1.00) with a *cis*-eQTL (rs1428554) that influenced expressions of *IRGM* ($\beta$ = –0.40, $P$ = 3.4E-13). None of the three SNPs were in LD ($r^2$>0.8) with any other *cis*-eQTLs that affected any other gene. Therefore, our eQTL analysis supported the conclusions of the GWAS.

As an example of Case 3, a GWAS (record 11 in Table 4) identified rs4656940 (in the intron of *CD244*) reporting two candidate causative genes (*CD244 and ITLN1*) [21]. The reported SNP was in perfect LD with a *cis*-eQTL (rs11265498) that influenced expressions of *ITLN1* ($\beta$ = −0.67, $P$ = 2.4E-17), where no *cis*-eQTL was identified for *CD244*. Therefore, our eQTL map indicated that *ITLN1* was the most likely causative gene. Our eQTL map helped to prioritize candidate genes for two of 12 records.

Any records for Crohn's disease were not classified into Case 4. In all GWAS records, we identified 13 Case-4 records (File S2). For instance, rs1354034 (in the intron of *ARHGEF3*, on chr3) was reportedly associated with platelet counts and mean platelet volume [24,25], and *ARHGEF3* was identified as a putative causative gene. In our eQTL map, the reported SNP was not associated with the expression of *ARHGEF3* or any other tested gene on the same chromosome, but with *CALD1* on a different chromosome, chr7 ($\beta$ = –0.48, $P$ = 5.3E-16). For another example, rs2517713 (intergenic, on chr6) was identified in a study of nasopharyngeal carcinoma [26] and *HLA-A* was reported as a putative causative gene. In our eQTL map, the reported SNP was not associated with the expression of *HLA-A* or any other tested gene on the same chromosome, but of *NRSN2* on a different chromosome, chr20 ($\beta$ = –0.21, $P$ = 2.2E-15). Notably, decreased expression of *NRSN2* was reported to be associated with hepatocellular carcinoma [27].

Similarly, we analyzed 10,076 (8,069) GWAS records (unique SNPs). We identified 386 cases in which *cis*- or *trans*-effects were identified for the reported SNPs, and classified each into one of the four cases; we found 191 (148) Case-1 records, 97 (80) Case-2 records, 85 (60) Case-3 records, and 13 (6) Case-4 records. We

**Table 3.** Multi-regulatory *cis*-eQTLs and *trans*-eQTLs.

| | | | | | LD block | | | |
|---|---|---|---|---|---|---|---|---|
| **eQTL** | **Chr** | **Position** | **MAF** | **HWE-P** | **Start** | **End** | **Length** | **Gene Symbol** |
| *cis* | | | | | | | | |
| rs7522860 | 1 | 156,275,281 | 0.49 | 0.644 | 156,208,230 | 156,314,627 | 106,398 | *TMEM79;SMG5;C1orf85;PAQR6* |
| rs6464103 | 7 | 150,478,385 | 0.37 | 0.711 | 150,476,888 | 150,478,385 | 1,498 | *TMEM176B;TMEM176A;ABP1* |
| rs4390300 | 10 | 60,144,207 | 0.47 | 0.817 | 60,144,207 | 60,168,003 | 23,797 | *IPMK;UBE2D1;TFAM* |
| rs2416549 | 12 | 11,325,804 | 0.24 | 0.116 | 11,045,512 | 11,349,454 | 303,943 | *TAS2R14;TAS2R30;PRB1* |
| rs35969491 | 12 | 11,339,020 | 0.24 | 0.084 | 11,045,512 | 11,349,454 | 303,943 | *TAS2R10;PRR4;PRH2;PRB4* |
| rs7226263 | 17 | 44,814,884 | 0.32 | 0.111 | 44,788,310 | 44,853,872 | 65,563 | *WNT3;ARL17B;ARL17A;NSF* |
| *trans* | | | | | | | | |
| rs116711766 | 1 | 160,093,165 | 0.075 | 0.3909 | 160,093,165 | 160,093,165 | 1 | *ITGA7;MC1R;FAM22G* |
| rs11718621 | 3 | 40,362,122 | 0.288 | 1.0000 | 40,362,122 | 40,463,063 | 100,942 | *DIRC1;MAB21L2;PRSS36; HIST2H2BF;KRTAP19–2;FSD1;LRRD1* |
| rs6773917 | 3 | 40,469,254 | 0.492 | 0.4881 | 40,373,259 | 40,498,845 | 125,587 | *DIRC1;MAB21L2;PRSS36; HIST2H2BF;NEURL;KRTAP19–2;FSD1;LRRD1* |
| rs7801498 | 7 | 102,089,595 | 0.368 | 0.8039 | 102,089,595 | 102,089,595 | 1 | *MUC4;GFRA1;MIOX;GYPA* |
| rs10873415 | 14 | 92,558,171 | 0.380 | 0.0097 | 92,434,957 | 92,558,171 | 123,215 | *GADD45GIP1;SOX13;TFEB;EIF2C1* |

Chr, Position: chromosomal positions of eQTLs; MAF: minor allele frequency; HWE-P: Hardy-Weinberg Equilibrium test *P* value; LD block: range in which SNPs in LD ($r^2$>0.8) with the eQTLs exist.
doi:10.1371/journal.pone.0100924.t003

**Table 4.** Summary of GWAS records associated with Crohn's disease and eQTL mapping results.

| | | Suggested genes | | SNPs | | | eQTL statistics | | | Top local SNP for GWAS gene | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Case | Record | GWAS | eQTL | GWAS | eQTL | $r^2$ | β | *P* | $P_c$ | SNP | β | *P* | $r^2$ |
| Case 1 | 1[21] | *CCR6*[a] | *RNASET2* | rs415890 | rs400837 | 0.99 | −0.36 | 2.7E-39 | 0.87 | Not tested | | | |
| | 2[21] | *FADS1* | *FADS2* | rs102275 | rs108499 | 0.97 | 0.16 | 3.2E-10 | 0.74 | rs174570 | 0.17 | 6.2E-07 | 0.99 |
| | 3[21] | *PRDX5* | *CCDC88B* | rs694739 | rs600377 | 0.85 | −0.26 | 1.0E-06 | 0.51 | rs2286614 | 0.42 | 4.5E-23 | 0.01 |
| | | *ESRRA* | | | | | | | | rs641811 | 0.06 | n.s. | 0.01 |
| | 4[21] | *IKZF3* | *GSDMB* | rs2872507 | rs1008723 | 0.98 | −0.38 | 6.9E-38 | 0.81 | rs56030650 | 0.05 | n.s. | 0.01 |
| | | *ZPBP2* | | | | | | | | rs62065216 | −0.09 | n.s. | 0.01 |
| | | *ORMDL3* | | | | | | | | rs1054609 | −0.18 | 3.6E-14 | 0.98 |
| | | *GSMDL*[a] | | | | | | | | Not tested | | | |
| | 5[22] | *ORMDL3* | *GSDMB* | rs2872507 | rs1008723 | 0.98 | −0.38 | 6.9E-38 | 0.81 | rs1054609 | −0.18 | 3.6E-14 | 0.98 |
| | 6[21] | *RTEL1* | *ZGPAT* | rs4809330 | rs6011058 | 1.00 | 0.09 | 2.9E-07 | 1.00 | rs2252258 | −0.05 | n.s. | 0.002 |
| | | *SLC2A4RG* | | | | | | | | rs310609 | −0.07 | n.s. | 0.02 |
| | | *TNFRS-F6B*[a] | | | | | | | | Not tested | | | |
| Case 2 | 7[21] | *PLCL1* | *PLCL1* | rs6738825 | rs1866664 | 0.98 | −0.25 | 3.0E-07 | 0.81 | rs1866664 | −0.25 | 3.0E-07 | 1 |
| | 8[21] | *IRGM* | *IRGM* | rs7714584 | rs1428554 | 1.00 | −0.40 | 3.4E-13 | 0.98 | rs1428554 | −0.40 | 3.4E-13 | 1 |
| | 9[22] | *IRGM* | *IRGM* | rs11747270 | rs1428554 | 1.00 | −0.40 | 3.4E-13 | 0.98 | rs1428554 | −0.40 | 3.4E-13 | 1 |
| | 10[23] | *IRGM* | *IRGM* | rs13361189 | rs1428554 | 1.00 | −0.40 | 3.4E-13 | 0.98 | rs1428554 | −0.40 | 3.4E-13 | 1 |
| Case 3 | 11[21] | *ITLN1*[b] | *ITLN1* | rs4656940 | rs11265498 | 1.00 | −0.67 | 2.4E-17 | 1.00 | rs11265498 | −0.67 | 2.4E-17 | 1 |
| | | *CD244* | | | | | | | | rs574610 | −0.12 | n.s. | 0.16 |
| | 12[46] | *RNASET2*[b] | *RNASET2* | rs2149085 | rs400837 | 0.99 | −0.36 | 2.7E-39 | 0.87 | rs400837 | −0.36 | 2.7E-39 | 1 |
| | | *FGFR1OP* | | | | | | | | rs73039162 | 0.68 | 7.5E-45 | 0.078 |
| | | *CCR6*[a] | | | | | | | | Not tested | | | |
| | | *MIR3939*[a] | | | | | | | | Not tested | | | |

[a]The GWAS-reported gene was not included in our study.

[b]GWAS-reported genes that match the eQTL-suggested genes in Case 3.

$r^2$: correlation of genotypes for linkage disequilibrium between the GWAS-identified SNP and *cis*-eQTL (in the "SNPs" column), or between the top local SNP for GWAS gene and *cis*-eQTL (in the "Top local SNP for GWAS Gene" column).

$P_c$: *P* value of a conditional regression on genotypes of GWAS-identified SNP.

Genes suggested by GWAS and our eQTL map are listed in the "Suggested genes" column; eQTL statistics are listed in the "eQTL statistics" column; most significant local SNP for the GWAS-reported gene is shown in the "Top local SNP for GWAS gene" column.

n.s: not significant.

doi:10.1371/journal.pone.0100924.t004

identified 6 lincRNAs in the Case-1 records that were most significantly associated with GWAS-reported SNPs. In summary, our eQTL map was informative for 3.8% of the GWAS records, each of which was classified into one of the four cases; 1.9% into Case 1, 1.0% into Case 2, 0.8% into Case 3, and 0.1% into Case 4. We provide the results of our application of our eQTL map to the GWAS records in File S2.

## Discussion

This study identified the largest number of eQTLs for East Asian whole blood samples to our knowledge. We identified 3,804 *cis*-eQTLs and 165 *trans*-eQTLs. *Cis*-effects were previously found for 44% (6,418 genes) of tested genes [20] for Caucasian whole blood samples. In the current study, *cis*-effects were found for 16.9% of the tested genes, which is in line with estimated powers in a previous study [28].

We identified 74 genes with *trans*-effects, which constituted 0.4% of tested genes. We believe that we underestimated the proportion of true *trans*-effects because we used the most stringent corrections for multiple testing. In fact, the smallest $R^2$ for any of the *trans*-eQTLs ($R^2 = 0.16$) was 2.4-fold greater than the smallest $R^2$ for any identified *cis*-eQTLs ($R^2 = 0.065$).

We analyzed and characterized our eQTLs in various aspects; 1) *cis*-eQTLs in terms of gene structure, epigenetic factors, and distance from genes; 2) multi-regulatory eQTLs; 3) eQTLs for mRNA as compared to those for lincRNAs; 4) application of eQTL maps to GWAS results; and 5) replication with independent samples.

### 1) *Cis*-eQTL analyses

The comparison between the genic and intergenic *cis*-eQTLs suggested that factors involved in expression levels are more enriched and stronger in genic regions (those located within a gene or within 1 kb of a gene) than intergenic regions (>1 kb from genes). All genic subcategories were each overrepresented compared to the intergenic regions (Table 2). We also showed that upstream and 5′-UTR regions particularly had strong effects compared to other genic regions. It would be reasonable to consider that upstream regions are important because transcription factor binding sites and transcription regulatory modules are enriched in 5′ flanking regions of genes. Strong effects in 5′ UTRs would imply that post-transcriptional regulation via 5′ UTRs has a particularly strong impact on expression levels. The significant association between $R^2$ of *cis*-eQTLs and epigenetic classification indicated that epigenetic factors (e.g., transcription regulatory modules) have influences on transcription that depend upon nucleotide sequences. Interestingly, the trend was not observed for $|\beta|$.

92% of *cis*-eQTLs were within their target genes or in 100 kb flanking regions, which is consistent with previous studies [3,7]; and it was also consistent that most of large-effect eQTLs were located within 20 kb [29].

### 2) Multi-regulatory eQTLs

We identified 6 and 5 multi-regulatory *cis*- and *trans*-eQTLs, respectively. We note that a pair of multi-regulatory *cis*-eQTLs on chr12, rs2416549 and rs35969491, and another pair of multi-regulatory *trans*-eQTLs on chr3, rs11718621 and rs6773917, each are likely to indicate the same locus because they were each close ($r^2 = 0.99$ and 0.39, respectively) and the regulated gene sets are similar. Multi-regulatory eQTLs may comprise two types of eQTLs; some may be true master regulators, while others may each comprise a group of eQTLs in strong LD, each of which regulates one gene. Further studies are needed to identify more multi-regulatory eQTLs so that they would be further analyzed in terms of LD structure and effect sizes comparing with eQTLs regulating one gene. Presence of *trans*-acting master regulators has been increasingly suggested[30–32]. However, it is very challenging to identify master regulators because statistical power to detect *trans*-eQTLs is low because of multiple testing corrections. Interestingly, with the stringent threshold of this study, *trans*-regulated genes were often associated with multi-regulatory *trans*-eQTLs (Figure S5), which may suggest multi-regulatory *trans*-eQTLs tend to have large effects.

### 3) mRNA and lincRNA transcripts

The importance of lincRNAs to phenotypic variation is increasingly recognized; nevertheless, previous eQTL studies focused only on coding genes, and did not include analyses of lincRNA transcripts. Here, we examined the genetic causes of variation in expression of coding genes and of lincRNAs. Coding genes and lincRNAs exhibited different characteristics; for example, the proportion of *cis*-regulated transcripts was 3 times larger for mRNAs (15.1% vs. 4.8%, Table 1); sequence variations influence coding gens more than lincRNAs. Nevertheless, eQTLs for lincRNAs should not be ignored because still 5.3% of lincRNAs were regulated by either *cis*- or *trans*-eQTLs, and the mean $R^2$ values of *cis*- or *trans*-eQTLs regulating lincRNAs were as large as those regulating mRNAs (Table 1, Wilcoxon's rank-sum $P = 0.094$), and $|\beta|$ values were even larger (Table 1, Wilcoxon's rank-sum $P = 3.2E\text{-}14$), which might indicate that lincRNAs are more variable than mRNAs, while the eQTL effects were similar in terms of $R^2$. These differences and similarities between coding transcripts and lincRNAs may indicate interesting mechanisms underlying the expressional regulations.

### 4) Application to GWAS results

The rationales behind utilizing eQTL mapping to interpret GWAS are that evidence from GWAS supports that transcriptional alterations contribute to risks of complex diseases; 1) a substantial fraction of GWAS-identified SNPs fell intergenic regions; and 2) eQTLs identified in previous study are enriched in GWAS-reported SNPs. Indeed, our eQTLs were also enriched in GWAS-reported SNPs: 1.7-fold for *cis*-eQTLs (one-sample proportion test $P<2.2E\text{-}16$) and 3.7-fold for *trans*-eQTLs (one sample proportion test $P = 3.5E\text{-}15$). Interestingly, *trans*-eQTLs were more enriched than *cis*-eQTLs. We identified 386 records for which our eQTL map may provide another evidence to interpret GWAS results. We emphasize that our results of applying our eQTL map to GWAS interpretation can only suggest another possibilities for candidate causative genes based on expressional variations and that the significant association with expression does not necessarily indicate the gene is causative (an example was shown for *RPS26* and type I diabetes [33]). Thorough and close assessment is required for each case to conclude what gene is truly causative. Still, reviewing previous GWAS results while referring to eQTL maps, not only regarding *cis*-eQTLs but also *trans*-eQTLs, would be worthwhile, and eQTL maps will provide useful information for interpreting and understanding future GWAS results as well.

### 5) Replication

*Cis*-regulated genes identified in our study were in a good concordance with those identified by Westra *et al.* [20]: 60.9% of our *cis*-regulated genes were replicated. The 60% replication rate seems reasonable for whole blood samples because, in the current study, we replicated 56% of 112 *cis*-regulated genes identified in a

previous study [17] for whole blood samples from 76 Japanese individuals. On the other hand, replication of *trans*-eQTLs was challenging; only <1% of *trans*-eQTLs identified by Westra et al. [20] were replicated in the current study. Variation between different populations might be important for *trans*-eQTLs because we could replicate one of two *trans*-eQTLs in the previous study for the Japanese population [17]. We speculate the reason of low replication for *trans*-eQTLs as follows: Mechanisms of *trans*-effects of many sequence variations are considered as that a variant induces transcriptional alteration in a *cis* manner or functional change by substituting amino acids of proteins that involve in transcriptional regulation of other genes, and then, the locally induced change causes changes in expression levels of other genes [34]. Although *trans*-regulatory mechanisms are largely unknown, such a regulatory system may depend on a network of genes in which the genes interactively and cooperatively work in the same biological process; consequently, individual out-put gene expression levels are a cumulative result of a net effect of the whole network which could involve complex feedback mechanisms. The state of such a network should change dynamically with cell types, environmental conditions, and time. This is one of the reasons for the low reproducibility of *trans*-eQTLs. It should be noted that our *trans*-eQTLs were identified under just one set of conditions; therefore, the validity of applying our results to situations that represent different conditions needs to be carefully evaluated. However, we believe that our *trans*-eQTL analysis provides general insights into *trans*-effects, such as how effect magnitudes, $\beta$ or $R^2$, are distributed.

## Methods

### Subjects and ethics statement

The study subjects were 301 apparently healthy individuals residing in Nagahama City, Japan. All participants provided written informed consent. The study protocol was approved by the Ethics Committee of Kyoto University Graduate School and Faculty of Medicine.

### SNP genotyping and quality control

We extracted DNA from leukocytes and carried out genome-wide SNP genotyping with the Infinum HumanOmni5Exome BeadChip (Illumina, Inc., San Diego, CA, USA). We excluded any SNP with a missing rate >1%, Hardy-Weinberg equilibrium test *P* value<1E-07, minor allele frequency <5%, or that mapped to a sex chromosome. Ultimately, we examined a final set of 1,425,832 autosomal SNPs in the analysis. We excluded three samples from the analysis; one was excluded because of unsuccessful DNA extraction, and two others were excluded because of kinship with other sample. The snpStats package (http://www.bioconductor.org/packages/release/bioc/html/snpStats.html) in Bioconductor [35] was used to conduct the principal component analysis, and no subjects were identified as outliers relative to the HapMap JPT (Figure S2).

### Gene expression profiles

Whole blood was collected from each participant when in a non-stimulated state; PAXgene Blood RNA Kits (QIAGEN, Hilden, Germany) were then used to collect samples of total RNA. For each participant, we used the Low Input Quick Amp Labeling Kit (Agilent Technologies, Inc., Santa Clara CA, USA) according to the manufacturer's protocol and 100 ng of total RNA to synthesize each labeled cRNA sample. We used Gene Expression Hybridization kits (Agilent Technologies, Inc.) to hybridize labeled cRNA to arrays from SurePrint G3 Human Gene Expression 8×60 K Microarray Kits (Agilent Technologies, Inc., design ID: 028004); Gene Expression Wash Packs (Agilent Technologies, Inc.) were then used according to the manufacturer's protocols to wash each microarray. Each microarray was scanned with a DNA Microarray Scanner (Agilent Technologies, Inc.), and Feature Extraction Ver.9.5.3 (Agilent Technologies, Inc.) was used to measure signal intensity.

### Normalization and exclusion of expression data

The data were processed using the GeneSpringGX11 as follows. For each set of duplicated probes, the mean signal intensity was calculated. Signal intensities less than 1 were each set to 1, and each signal intensity value was transformed by taking the binary logarithm. Normalization was carried out by a 75th percentile shift; this normalization procedure was recommended by Agilent. After this normalization, the 75th percentile signal intensity of each chip was set to 0, at which point the signal values ranged from −7.3 to 12.3 with a median (mean) of −2.6 (−2.1).

We excluded 5,550 probes for which we were not able to obtain specific positions on the chromosomes of their target genes and 1,488 probes that were mapped on the sex chromosomes. We did not filter any probes based on expression abundance because the information that the transcript is not expressed might be of biological importance. However, signal values for low or non-expressed genes are often unreliable; therefore, we show median expression values for our eQTL-regulated transcripts provided in File S1; and to interpret the expression values Figure S3 shows how expression values were distributed for expressed or non-expressed transcripts.

### Annotation of expression microarray probes

Annotation for gene expression probes of our chip (Agilent Technologies, Inc., design ID: 028004) was obtained from eArray (release date: 2012/04/11, build version: hg19:GRCh37:Feb2009, available online https://earray.chem.agilent.com/earray/). We defined three groups of probes: probes for *mRNA* transcripts, probes for *lincRNA* transcripts, and probes for *other* transcripts. Probes were classified into the *mRNA* group if they had assigned RefSeq NM accession numbers. *lincRNA* probes were indicated as such in Agilent's annotation. All the other probes were classified into the *other* group. The transcription start and end sites of genes represented by the probes that were classified into *mRNA* or *other* were obtained from a seq_gene.md file downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/accessed on 2013/02/20); and those represented by *lincRNA* probes were obtained from either Agilent's annotation or lincRNAsTranscripts table downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/accessed on 2013/04/09).

### Annotation of SNPs

BLAST was used to map probes from the SNP genotyping array into GRCh37; a rsID was assigned to each SNP based on its mapped chromosomal position on GRCh37. We defined a distance between a SNP and a gene as base pairs between the chromosomal position of the SNP and the position of the nearest transcription start/end site of the gene. If the SNP was located within the gene, then the distance was set to 0. Directions of genes were considered, and the sign associated with each distance indicated that the SNP was located upstream (negative) or downstream (positive) of the gene. ANNOVAR version 2013-05-09 [36] (http://www.openbioinformatics.org/annovar/) was used to annotate SNPs for classification into gene-structure-based categories; the RefSeq Gene (build version 19) was used as the reference. We annotated SNPs with ANNOVAR's default

definitions and precedence of SNP functional categories if a SNP was located within its target gene or within 1 kb-flanking regions of its target gene, and the gene name in the ANOVVAR annotation matched the target gene (if the gene name did not match, no specific functions were assigned); and otherwise, the SNP was categorized into *intergenic* (see supplementary note in File S3 for details). Using this method, we would classify an eQTL as intergenic if it was located outside its target gene even though it was located within another gene; in a different example, an eQTL in an intron of its target gene was classified as intronic even though it was located in any other category of another gene.

We classified each intergenic SNP into one of the regulatory potential classes as defined based on epigenetic information available in public databases by RegulomeDB [19] for dbSNP132 (downloaded from http://regulome.stanford.edu/on 2013/07/24). We were able to assign a regulatory classification to each of 1,396,242 SNPs (97.9% of the tested SNPs). We considered seven categories (Category 1–7) of regulatory classes as defined by the RegulomeDB, but we did not use the 15 subcategories (1a–f, 2a–c, 3a–b, 4–7). Briefly, lower scores indicated more evidence for the SNP being located in a regulatory region. Each known eQTL with known additional epigenetic functional annotation was assigned to Category 1. Category 2 requires direct evidence of binding through ChIP-seq and DNase. Category 3 requires a less complete set of evidence of binding. Categories 4–6 each comprised SNPs with minimal evidence of effects on transcription factor binding; Category 4 SNPs had DNase and ChIP-seq evidence; Category 5 SNPs had DNase or ChIP-seq evidence; and Category 6 had any single annotation not categorized above. Finally, Category 7 SNPs had no known evidence of TF binding.

## eQTL mapping

We performed surrogate variable analysis [37] to identify unmodeled latent factors that cause heterogeneity in expression data. We identified two significant surrogate variables with age and gender used as known covariates using sva package (http://bioconductor.org/packages/release/bioc/html/sva.html) in Bioconductor [35,38]. We corrected expressions of each transcript for age, gender, and the two surrogate variables by fitting a multiple linear model in R version 3.0.2 (http://www.R-project.org/). We further excluded 4,972 probes that were mapped to regions with SNPs that was found polymorphic in the HapMap JPT samples or our study subjects because polymorphisms in such regions can alter hybridization efficiency; consequently, signal intensities may not reflect the actual amount of RNA [39–42]. The remaining 30,395 probes were included in the analysis. We assumed an additive model for all SNPs, and we coded each SNP genotypes as 0, 1, or 2, to represent the number of minor alleles in each individual. PLINK v1.07 [43] (http://pngu.mgh.harvard.edu/purcell/plink/) was used to perform the association analysis between each adjusted transcriptional phenotype and each of 1,425,832 autosomal SNPs with 298 individuals.

We define a *local* SNP as a SNP located on the same chromosome and within 500 kb from the nearest transcription start/end site of the gene that encodes the transcript, and a *distant* SNP, otherwise. We defined a *cis*-eQTL as a local SNP that significantly affects expression of a gene; similarly we defined a *trans*-eQTL as a distant SNP that significantly affects expression of a gene. We examined 16,986,695 local SNP-transcript pairs (11,028,260 for mRNAs, 3,485,407 for lincRNAs, and 2,473,028 for other transcripts). The mean number of local SNPs per probe was 560 (minimum 1, maximum 4,630). We examined about 43 billion distant SNP-transcript pairs. To identify *cis*-eQTLs, we estimated FDR with the permutation approach as described by Westra *et al.* [20]. Briefly, sample identifiers were permuted for 10 times, and only the local SNP with the smallest $P$ value for each transcript was used to simulate the null distribution. With this approach we estimated FDR only for the SNP with the smallest $P$ value for each transcript, and local SNPs with the FDR smaller than 5% were identified as *cis*-eQTLs. Therefore, no more than one *cis*-eQTL was identified for each transcript. If multiple SNPs in perfect LD ($r^2 = 1$) were the most significant with the same $P$ value, the middle SNP was used to represent the eQTL. To exclude possible false discoveries caused by outliers or violation of normality assumptions, we performed Kruskal-Wallis test [44], a non-parametric test, and excluded *cis*-eQTL-transcript pairs with $P$ value$>$0.00015 (see supplementary note in File S3).

To identify *trans*-eQTLs, we used the Bonferroni correction for multiple comparisons among the approximately 43 billion tests; only distant SNPs with nominal $P$ values smaller than 1.15E-12, which corresponds to a family-wise error rate of 5%, were considered significant. We applied intensive exclusion criteria to obtain reliable *trans*-eQTLs. First, we excluded *trans*-eQTLs that may only capture *cis*-effects because of LD by a conditional regression on *cis*-eQTL genotypes (i.e., excluded when residuals of fitting *cis*-eQTL genotypes were not significantly associated with *trans*-eQTL genotypes by $P<0.05$). This analysis was performed when a *trans*-eQTL and its target transcript were located on the same chromosome, and a *cis*-eQTL was also identified for the transcript (*cis*-eQTLs excluded by Kruskal-Wallis tests were also considered). Second, we excluded redundant *trans*-eQTLs because of LD with other *trans*-eQTLs by sequential conditional regressions. For each transcript, *trans*-eQTLs on the same chromosome were iteratively tested starting from the *trans*-eQTL of the smallest $P$ value for the transcript. If significant ($P<0.05$), the *trans*-eQTL is kept and residuals were used for the next iteration. If not, the *trans*-eQTL was excluded as redundant, and residuals were not taken for the next iteration. After this procedure, we tested *trans*-eQTLs that were found significant in the sequential conditional regression all together with a multiple linear regression, and non-significant *trans*-eQTLs ($P>0.05$) were further removed. Third, in order to confirm that the *trans*-eQTLs were not false positives because of cross-hybridization of probes to unexpected transcripts near the *trans*-eQTLs, we mapped the probe sequence to the flanking region (±500 kb) of its *trans*-eQTL by SHRiMP v.2.2.3 [45] for each probe-*trans*-eQTL combination. The human reference DNA sequence (GRCh37.p5) was downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/). We used the same relaxed settings as Westra *et al.* [20] (match score of 10, mismatch score of 0, gap open penalty of −250, gap extension penalty of −100, and minimal Smith-Waterman score of 30%); -m 10 -i 0 -q -250 -f -100 -h 30%. We excluded a *trans*-eQTL if its associated probe was mapped to its flanking region. Fourth, we excluded low expression transcripts whose median expression levels were lower than −4.5 because we observed deviation from the distribution of median expression levels of *cis*-regulated transcripts (Figure S6). The cutoff was defined as the 5th percentile of the median expression levels of the *cis*-regulated transcripts. Kruskal-Wallis tests for the remaining SNP-transcript pairs were all significant ($P<0.00015$). We used the remaining *trans*-eQTLs in the further analyses.

The approach we used to correct for multiple testing with local SNPs differed from that used with distant SNPs because the high peak at low $P$ values observed with local SNPs indicated that a substantial fraction of local SNPs were truly associated with the expression phenotype of one or more transcripts, whereas the uniform distribution of $P$ values observed with distant SNPs indicated that the null hypothesis was true for most of the tests (Figure S1).

### Identifying multi-regulatory eQTLs

We defined a multi-regulatory *cis*-eQTL as a *cis*-eQTL that is associated with expression levels of at least three different local protein-coding genes (assigned RefSeq NM accessions). For this, we did not count probes that cross-hybridize to other local genes associated with the same *cis*-eQTL by mapping probe sequences to the exon sequences with SHRiMP v.2.2.3 [45] using the same set of options used for detecting cross-hybridization for *trans*-eQTLs above.

Similarly, we defined a multi-regulatory *trans*-eQTL as a *trans*-eQTL that is associated with expression levels of at least 3 different distant protein-coding genes, after excluding cross-hybridized probes in the same procedure as used for multi-regulatory *cis*-eQTLs.

### Statistical analysis for eQTLs

For comparison of mean effects of gene-based functional categories, we excluded SNPs that we were not able to assign to a specific category; we also excluded categories that comprised fewer than 5 eQTLs. Values of $|\beta|$ and $R^2$ were log-transformed and then subjected to the ANOVA; the ANOVA was followed by Tukey's HSD test (which performs all pairwise comparisons between two subcategories for multiple testing correction). The trend of log-transformed $|\beta|$ and $R^2$ values with seven RegulomeDB classes (Classes 1a–f, 2a–c and 3a–b were grouped as 1, 2 and 3, respectively) was tested with Jonckheere-Terpstra permutation test (one-sided, 100,000 permutations) provided in clinfun package (http://cran.r-project.org/web/packages/clinfun/index.html) in R version 3.0.2 (http://www.R-project.org/). $r^2$ of LD between SNPs were computed with PLINK v1.07 [43].

### Replication analysis

We downloaded the eQTL map by Westra *et al.* [20] from their browser (http://genenetwork.nl/bloodeqtlbrowser/), and annotation files for HT12v3 and Agilent Human Genome 4×44 K array from the GEO (http://www.ncbi.nlm.nih.gov/geo/). We matched Entrez GeneIDs to compare with the replication studies. We referred to GWAS catalog to obtain SNPs tested for *trans*-eQTL in [20] (SNPs reported by 16, July, 2011). We found 978 distant SNP-transcript pairs in our study that corresponded to rs IDs and Entrez Gene IDs tested in [20].

### Matching eQTLs with GWAS-identified SNPs

We downloaded 16,541 public GWAS records from the NHRGI GWAS Catalog (http://www.genome.gov/gwastudies/ accessed on 2014/04/23). We excluded 323 records with reported *P* values that were not significant (reported as NS or Pending); we excluded another 6,142 records because the reported SNPs were not included in our tested SNPs. Ultimately, we examined 10,076 records for 8,069 unique SNPs reported by 1,436 GWAS. We matched GWAS-reported SNPs to our eQTLs when they exactly matched or were in LD ($r^2>0.8$). We excluded records if conditional regression on genotypes of a GWAS-identified SNP was significant $P<0.05$ (File S2) because they might be false discoveries of trait-eQTL association where eQTL and GWAS-identified SNP are two different genetic factors [33]. When matching gene symbols, we also searched their aliases downloaded from the HGNC BioMart version 0.7 (http://www.genenames.org/biomart/accessed on 2013/09/28).

### Accession numbers

Our expression microarray data are available at the NCBI's Gene Expression Omnibus under accession number GSE53351.

## Supporting Information

**Figure S1 Histogram of *P* values of all association tests.** A) Histogram of *P* values obtained from the 16,986,695 association tests between all autosomal transcripts and local SNPs. The excess of smaller *P* values indicates that a substantial fraction of associations are truly positive. B) Histogram of *P* values obtained from about 43 billion association tests between all autosomal transcripts and distant SNPs. The almost uniformly distributed *P* values suggests that most of distant SNPs have no effects on transcriptional regulation, though a slight increase at the low *P* values in frequency indicates a tiny fraction of distant SNPs are truly positive. Also see File S3 for a comment about influence of surrogate variable analysis on the distribution.
(TIF)

**Figure S2 Principal component analysis of study population in comparison with HapMap samples.** The first and second principal components are shown. CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; YRI: Yoruba in Ibadan, Nigeria; JPT: Japanese in Tokyo; CHB: Han Chinese in Beijing, China; Sample: samples of the current study.
(TIF)

**Figure S3 Distribution of normalized expression data.** Distribution of normalized expression data for all 42,405 probes and 298 samples are shown. "A" (absent) if a foreground signal is <2.6 SD of background signal; "M" (marginal) if it was saturated, not uniform in a spot, or not uniform among replicated probes, or "P" (present) otherwise. The number following each class name is the number of data classified into the class.
(TIF)

**Figure S4 Regression coefficients of multi-regulatory eQTLs.** Regression coefficients, β, of each multi-regulatory *cis*-eQTLs (A) or *trans*-eQTLs (B) are shown. Directions of effects of each multi-regulatory eQTL are consistent.
(TIF)

**Figure S5 *Trans*-eQTL map.** A) Chromosomal positions of *trans*-eQTLs are plotted against chromosomal positions of associated transcripts. B) $-\log_{10}$ *P* values of *trans*-eQTLs are plotted against the respective chromosomal positions. (C) $-\log_{10}$ *P* values of *trans*-eQTLs are plotted against the chromosomal positions of associated transcripts. The horizontal and vertical dashed lines separate chromosomes; the diagonal dashed line indicates that the *trans*-eQTL is located at the same chromosomal positions as transcripts. mRNA transcripts are shown in red; lincRNA transcripts are shown in green; and other transcripts are shown in black. $-\log_{10}$ *P* values are truncated at 50, and a triangle indicate truncation.
(TIF)

**Figure S6 Median expression levels of *cis*-regulated or *trans*-regulated genes.**
(TIF)

**Table S1 Demographic characteristics of study subjects.**
(DOCX)

**Table S2 *P* values of Tukey's HSD test.**
(DOCX)

**Table S3 Replicated *trans*-eQTLs identified by Westra *et al.* [20].**
(XLSX)

**File S1 Annotations and statistics of *cis*-eQTLs and *trans*-eQTLs.**
(XLS)

**File S2 Results of our application of eQTL map to GWAS records.** Sheet "Case1–3" shows records classified into Case 1, 2, or 3; sheet "Case 4" shows records classified into Case 4; sheet "Excluded" shows excluded records because GWAS SNP and eQTL are not likely to colocalize.
(XLSX)

**File S3 Supplementary notes.**
(PDF)

## References


1. Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39: 1208–1216.
2. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, et al. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430: 743–747.
3. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. Nat Genet 39: 1202–1207.
4. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365–1369.
5. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78.
6. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.
7. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–1224.
8. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, et al. (2007) A survey of genetic human cortical gene expression. Nat Genet 39: 1494–1499.
9. Mehta D, Heim K, Herder C, Carstensen M, Eckstein G, et al. (2013) Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. Eur J Hum Genet 21: 48–54.
10. Spizzo R, Almeida MI, Colombatti A, Calin GA (2012) Long non-coding RNAs and cancer: a new frontier of translational research? Oncogene 31: 4577–4587.
11. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. Annu Rev Biochem 81: 145–166.
12. Risch N, Merikangas K (1996) The Future of Genetic Studies of Complex Human Diseases. Science 273: 1516–1517.
13. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, et al. (2007) Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet 39: 226–231.
14. Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, et al. (2008) Evaluation of genetic variation contributing to differences in gene expression between populations. Am J Hum Genet 82: 631–640.
15. Bushel PR, McGovern R, Liu L, Hofmann O, Huda A, et al. (2012) Population differences in transcript-regulator expression quantitative trait loci. PLoS One 7: e34286.
16. Duan S, Huang RS, Zhang W, Bleibel WK, Roe CA, et al. (2008) Genetic architecture of transcript-level variation in humans. Am J Hum Genet 82: 1101–1113.
17. Sasayama D, Hori H, Nakamura S, Miyata R, Teraishi T, et al. (2013) Identification of single nucleotide polymorphisms regulating peripheral blood mRNA expression with genome-wide significance: an eQTL study in the Japanese population. PLoS One 8: e54967.
18. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10: 184–194.
19. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22: 1790–1797.
20. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 45: 1238–1243.
21. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42: 1118–1125.
22. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40: 955–962.
23. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, et al. (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 39: 830–832.


## Acknowledgments

We would like to offer our special thanks to Non-profit Organization "Zero-ji Club for Health Promotion", the staff and associates of the Nagahama City Office, Kohoku Medical Association, Nagahama City Hospital, Nagahama Red Cross Hospital, Nagahama Kohoku City Hospital, and the participants of the Nagahama Study for data collection.

## Author Contributions

Conceived and designed the experiments: FM YT. Performed the experiments: SN MI KM. Analyzed the data: MN RY. Contributed reagents/materials/analysis tools: YT TK. Wrote the paper: MN RY. Developed the Human Genome Variation Browser: KH.


24. Li J, Glessner JT, Zhang H, Hou C, Wei Z, et al. (2013) GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children. Hum Mol Genet 22: 1457–1464.
25. Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, et al. (2011) New gene functions in megakaryopoiesis and platelet formation. Nature 480: 201–208.
26. Tse K-P, Su W-H, Chang K-P, Tsang N-M, Yu C-J, et al. (2009) Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. Am J Hum Genet 85: 194–203.
27. Ma H-Q, Liang X-T, Zhao J-J, Wang H, Sun J-C, et al. (2009) Decreased expression of Neurensin-2 correlates with poor prognosis in hepatocellular carcinoma. World J Gastroenterol 15: 4844–4848.
28. Brown CD, Mangravite LM, Engelhardt BE (2013) Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLoS Genet 9: e1003649.
29. Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4: e1000214.
30. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, et al. (2012) Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet 44: 502–510.
31. Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, et al. (2011) Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat Genet 43: 561–564.
32. Fehrmann RSN, Jansen RC, Veldink JH, Westra H-J, Arends D, et al. (2011) Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet 7: e1002197.
33. Plagnol V, Smyth DJ, Todd JA, Clayton DG (2009) Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics 10: 327–334.
34. Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, et al. (2010) Polymorphic cis- and trans-regulation of human gene expression. PLoS Biol 8: e1000480.
35. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
36. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.
37. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3: 1724–1735.
38. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28: 882–883.
39. Alberts R, Terpstra P, Li Y, Breitling R, Nap J-P, et al. (2007) Sequence polymorphisms cause many false cis eQTLs. PLoS One 2: e622.
40. Benovoy D, Kwan T, Majewski J (2008) Effect of polymorphisms within probe-target sequences on oligonucleotide microarray experiments. Nucleic Acids Res 36: 4417–4423.
41. Walter NAR, McWeeney SK, Peters ST, Belknap JK, Hitzemann R, et al. (2007) SNPs matter: impact on detection of differential expression. Nat Methods 4: 679–680.
42. Sliwerska E, Meng F, Speed TP, Jones EG, Bunney WE, et al. (2007) SNPs on chips: the hidden genetic code in expression arrays. Biol Psychiatry 61: 13–16.
43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
44. Kruskal WH, Wallis WA (1952) Use of Ranks in One-Criterion Variance Analysis. J Am Stat Assoc 47: 583–621.
45. David M, Dzamba M, Lister D, Ilie L, Brudno M (2011) SHRiMP2: sensitive yet practical SHort Read Mapping. Bioinformatics 27: 1011–1012.
46. Yang S-K, Hong M, Zhao W, Jung Y, Baek J, et al. (2014) Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. Gut 63: 80–87.

*Human Molecular Genetics, 2014, Vol. 23, No. 24* **6659–6667**
*doi:10.1093/hmg/ddu374*
*Advance Access published on July 17, 2014*

# Identification of three novel genetic variations associated with electrocardiographic traits (QRS duration and PR interval) in East Asians

**Kyung-Won Hong[1,†], Ji Eun Lim[2,†], Jong Wook Kim[3], Yasuharu Tabara[4], Hirotsugu Ueshima[5,6], Tetsuro Miki[7], Fumihiko Matsuda[4], Yoon Shin Cho[8], Yeonjung Kim[1,*] and Bermseok Oh[2,*]**

[1]Division of Epidemiology and Health Index, Center for Genome Science, Korea Centers for Disease Control and Prevention, Chungcheongbuk-do 363-951, Korea, [2]Department of Biomedical Engineering, School of Medicine, Kyung Hee University, Seoul 130-701, Korea, [3]Department of Internal Medicine, Inje University Ilsan Paik Hospital, Gyeonggi-do 411-706, Korea, [4]Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto 606-8507, Japan, [5]Center for Epidemiologic Research in Asia and, [6]Department of Health Science, Shiga University of Medical Science, Otsu 520-2121, Japan, [7]Department of Geriatric Medicine, Ehime University Graduate School of Medicine, Toon 791-0295, Japan and [8]Department of Biomedical Science, Hallym University, Chuncheon, Gangwon-do 200-702, Korea

Received November 14, 2013; Revised June 25, 2014; Accepted July 14, 2014

**The electrocardiogram has several advantages in detecting cardiac arrhythmia—it is readily available, non-invasive and cost-efficient. Recent genome-wide association studies have identified single-nucleotide polymorphisms that are associated with electrocardiogram measures. We performed a genome-wide association study using Korea Association Resource data for the discovery phase (Phase 1, *n* = 6805) and two consecutive replication studies in Japanese populations (Phase 2, *n* = 2285; Phase 3, *n* = 5010) for QRS duration and PR interval. Three novel loci were identified: rs2483280 (*PRDM16* locus) and rs335206 (*PRDM6* locus) were associated with QRS duration, and rs17026156 (*SLC8A1* locus) correlated with PR interval. *PRDM16* was recently identified as a causative gene of left ventricular non-compaction and dilated cardiomyopathy in 1p36 deletion syndrome, which is characterized by heart failure, arrhythmia and sudden cardiac death. Thus, our finding that a *PRDM16* SNP is linked to QRS duration strongly implicates *PRDM16* in cardiac function. In addition, C allele of rs17026156 increases PR interval (beta ± SE, 2.39 ± 0.40 ms) and exists far more frequently in East Asians (0.46) than in Europeans and Africans (0.05 and 0.08, respectively).**

## INTRODUCTION

The electrocardiogram (ECG) has several advantages in detecting cardiac diseases; it is readily available, noninvasive and cost-efficient. QRS complex represents ventricular depolarization, and QRS duration represents the conduction time from atrioventricular node (AV node) to His-Purkinje system and ventricular myocardium (1). PR interval is the time between the onset of atrial depolarization (P-wave) and the onset of ventricular depolarization (R-wave).

QRS duration and PR interval are believed to reflect patient outcomes in several heart diseases (2–4). A diseased ventricular conduction system can lead to life-threatening bradyarrhythmias and tachyarrhythmias (5). Longer QRS duration is a predictor of mortality and sudden death in the general population (6) and in those with hypertension and coronary artery disease (7).

ECG measurements are believed to be complex traits with multiple genetic and environmental determinants (8). The heritability of each ECG measurement ranges from 30 to 50% in

*To whom correspondence should be addressed at: Department of Biomedical Engineering, School of Medicine, Kyung Hee University, #26 Kyungheedaero, Dongdaemun-gu, Seoul, 130-701, Korea. Tel: +82 29610617; Fax: +82 260080647; Email: ohbs@khu.ac.kr (B.O.); Division of Epidemiology and Health Index, Center for Genome Science, KNIH, KCDC #200 Osong-eup, Gangoe-myeon, Cheongwon-gun, Chungbuk-do, 363-951, Korea. Tel: +82 437186720; Fax: +82 437196759; Email: yeonmaru@gmail.com (Y.K.)
Full list of the Japanese study group is given in Appendix.
†K.-W.H. and J.E.L. contributed equally to this work.

© *The Author 2014. Published by Oxford University Press. All rights reserved.*
*For Permissions, please email: journals.permissions@oup.com*

Downloaded from http://hmg.oxfordjournals.org/ at Kyoto University on March 21, 2015

several ethnic groups (8–13). Recent genome-wide association studies (GWASs) have identified single-nucleotide polymorphisms (SNPs) that are associated with PR interval (14,15), QRS duration (14,16) and QT interval (17,18). In particular, QT interval has been studied extensively in European descendants by the QTGEN (17) and QTSCD (18) consortiums, and we have also reported a GWAS in East Asians (19). With regard to PR interval, two GWASs reported eight loci in European descendants, five of which (*SCN5A-SCN10A*, *NKK2–5*, *CAV1/CAV2*, *SOX5* and *TBX5*) were linked to atrial fibrillation (AF) (14,15). Twenty-two loci were correlated with QRS duration in two GWASs of European descendants (14,16), and greater number of risk alleles for prolonged QRS duration was also associated with the risk of ventricular conduction defects (16).

No GWAS on PR interval or QRS duration has been performed in the Asian population. To determine the genetic architecture of PR interval and QRS duration in Asians, we conducted a GWAS using Korea Association Resource (KARE) data during the discovery phase (Phase1, $n = 6805$) and two consecutive replication studies in Japanese populations (Phase 2, $n = 2285$; Phase 3, $n = 5010$) (Supplementary Material, Fig. S1).

## RESULTS

### Discovery GWASs

The clinical characteristics of subjects in the discovery GWAS and two replication studies are described in Table 1. The genomes of discovery subjects were scanned to identify genetic variations that were associated with QRS duration and PR interval. The genotypes in this study consisted of experimentally genotyped SNPs and computationally imputed SNPs. In total, 2.1 million SNPs were examined in the linear regression model as independent variables of ECG traits, controlling for age, sex, recruitment area, body mass index, systolic blood pressure and height as covariates. Q–Q plots of the GWAS in Koreans are shown in Supplementary Material, Figure S2.

All *P*-values are charted in Figure 1A (QRS duration) and B (PR interval), plotting $-\log_{10}(p)$ against the chromosomal position on Manhattan plots and are shown in Supplementary Material, Tables S1 and S2. The red line in Figure 1 indicates a genome-wide significance level ($P < 5 \times 10^{-8}$), and the blue line indicates a suggestive level ($P < 1 \times 10^{-4}$). Two loci (*CDKN1A* and *SETBP1*) for QRS duration and three loci (*SLC8A1*, *SCN5A/SCN10A* and *CAV1/CAV2*) for PR interval met the genome-wide significance level in the discovery phase. The 323 SNPs for QRS duration and 341 SNPs for PR intervals had *P*-values that were lower than the suggestive level, encompassed by 21 and 20 loci, respectively. The lead SNP for each locus is listed in Supplementary Material, Table S3.

The *SETBP1*, *CDKN1A*, *SCN5A* and *HAND1* regions for QRS duration and *CAV1*, *SCN10A* and *TBX5* regions for PR interval have previously been reported (14–16). The remaining suggestive loci shown in Supplementary Material, Table S3 were examined in replication studies of Japanese populations.

**Table 1.** Clinical characteristics of subjects in each phase

| Variables | Phase 1 KARE | Phase 2 Japanese | Phase 3 Japanese |
|---|---|---|---|
| *n* (% male) | 6805 (50.4%) | 2285 (31.9%) | 5010 (33.3%) |
| Age, years | 51.6 (8.7) | 49.8 (13.9) | 56.7 (13.4) |
| BMI, kg/m$^2$ | 24.6 (3.1) | 22.2 (3.2) | 22.6 (3.2) |
| SBP, mm Hg | 116.4 (17.9) | 120.1 (16.8) | 126.1 (19.1) |
| Height, cm | 160.6 (8.7) | 160.4 (8.3) | 158.6 (8.7) |
| PR interval, ms | 163.2 (35.9) | 158.1 (21.4) | 158.7 (22.0) |
| QRS duration, ms | 90.3 (10.3) | 97.0 (8.2) | 95.5 (9.0) |

Data are presented as mean (standard deviation).
BMI, body mass index; SBP, systolic blood pressure.

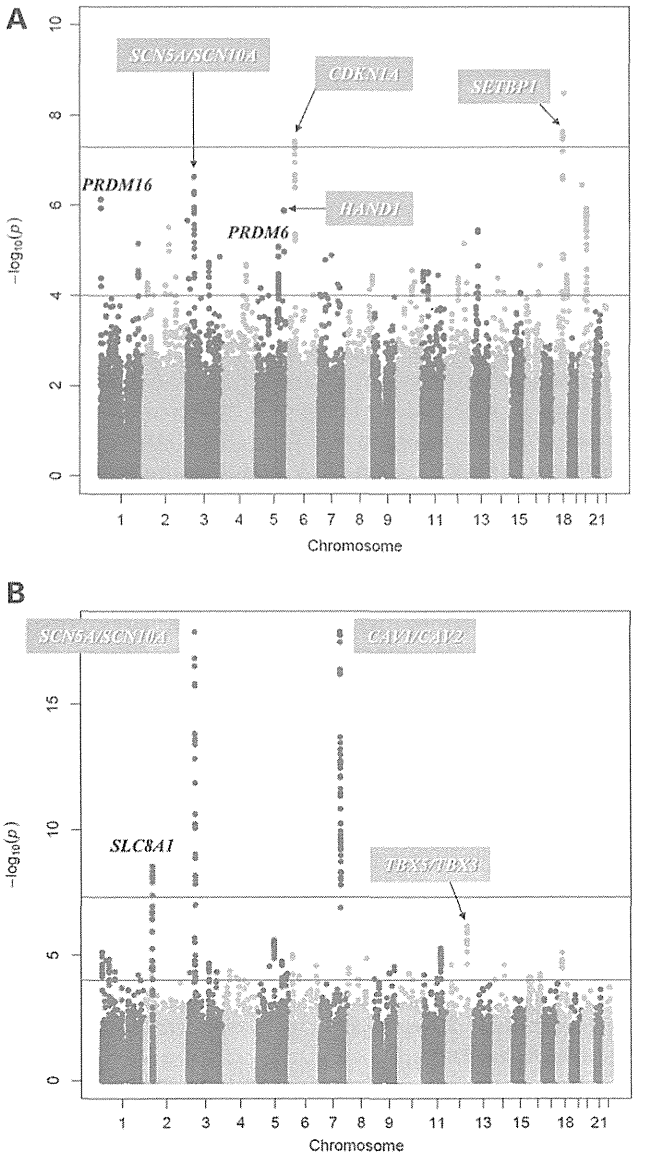


**Figure 1.** Manhattan plot of genome-wide association signals from Phase 1 Study. $-\log10(p)$ values are plotted against chromosomal base-pair positions. Green label indicates previously reported loci for QRS duration and PR interval, and red dots indicate previously unreported loci showing associations with QRS and PR in the Phase 1 study and tested for replication. The red line represents the genome-wide significance level ($P = 5 \times 10^{-8}$), and the blue line represents a *P*-value of $1 \times 10^{-4}$. (**A**) QRS duration and (**B**) PR interval.

Downloaded from http://hmg.oxfordjournals.org/ at Kyoto University on March 21, 2015

### Novel genetic variant of QRS duration and PR interval in East Asians

Seventeen suggestive loci for QRS duration and 17 suggestive loci for PR interval were identified in the discovery GWAS. To confirm these findings, two consecutive replication tests were conducted in two independent Japanese populations.

We carried out Phase 2 follow-up *in silico* genotyping for 34 SNPs in 2285 Japanese individuals (Supplementary Material, Table S4). In the meta-analysis of Phases 1 and 2, two SNPs (rs2483280 and rs17026156) reached genome-wide significance *P*-value. These two SNPs and five additional SNPs that became better *P*-values than Phase 1 in meta-analysis of Phase 1 + 2 were genotyped in a subsequent de novo replication study in 5010 Japanese population (Phase 3). Three SNPs (rs2483280 and rs335206 for QRS duration and rs17026156 for PR interval) reached genome-wide significance in the meta-analyses of Phases 1, 2 and 3 (Table 2).

The genetic regions of these three SNPs and their association results are depicted as signal plots in Figure 2. rs2483280 lies in the third intron of *PRDM16* (based on the NM_022114.3 transcript), rs335206 resides in the fifth intron of *PRDM6* (based on the NM_001136239.1 transcript) and rs17026156 is located 21 kb upstream of *SLC8A1*.

### Extension of variants identified in European descendants to Koreans

To compare the genetic architecture of QRS duration and PR interval between Europeans and Koreans, the SNPs that were previously identified in European descendants were examined in discovery GWAS (Table 3 and Supplementary Material, Table S5). Three GWASs reported 21 SNPs for QRS duration and 10 SNPs for PR intervals (14–16). In the KARE genotype data of discovery GWAS, there were only five SNPs that matched the reported SNPs. Thus, we added SNPs with linkage disequilibrium (LD) ($r^2 > 0.8$ and $D' > 0.9$) of the lead SNPs in the European studies. A total of 22 SNPs were examined for their association in Koreans (Supplementary Material, Table S5).

Seven of 14 QRS-related loci, 3 of 5 PR-related loci and all 3 loci for both traits were associated with Koreans, based on $P < 0.05$ (Fig. 3). Of the seven QRS duration-associated loci, three (*HAND1-SAP30L*, *CDKN1A* and *SETBP1*) had *P*-values of $<1 \times 10^{-5}$. The three loci (*EXOG-SCN5A-SCN10A*, *CAV1-CAV2* and *TBX5-TBX3*) that were linked to both traits were also significantly associated with Koreans. Further, the *EXOG-SCN5A-SCN10A* and *CAV1-CAV2* regions had large effect sizes compared with other loci (beta ± SE = 4.33 ± 0.56, $P = 1.45 \times 10^{-14}$ and beta ± SE = 3.21 ± 0.42, $P = 3.33 \times 10^{-14}$, respectively).

## DISCUSSION

### *In silico* annotation of novel SNP sites

Our discovery GWAS in Koreans and two replication studies in Japanese identified three novel loci for QRS duration and PR interval: rs2483280 (*PRDM16* locus) and rs335206 (*PRDM6* locus) for QRS duration and rs17026156 (*SLC8A1* locus) for PR interval. Because the three SNPs lay in noncoding regions, we examined their function in regulating gene expression

**Table 2.** Replication results of novel SNPs in each phase and meta-analysis

| SNP ID | CHR | BP | Gene[a] | Coded allele | Phase 1 ($n = 6805$) | | | Phase 2 ($n = 2285$) | | | Meta-analysis (Phase 1 + 2) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | AF | Beta ± SE | *P*-value | AF | Beta ± SE | *P*-value | Beta | *P*-value | *Q* | $I^2$ | |
| QRS | | | | | | | | | | | | | | | |
| rs2483280 | 1 | 3 245 399 | *PRDM16* | A | 0.26 | −0.91 ± 0.18 | $7.47 \times 10^{-7}$ | 0.28 | −0.61 ± 0.24 | 0.010 | −0.80 | $\mathbf{3.83 \times 10^{-8}}$ | 0.33 | 0.00 | |
| rs335206 | 5 | 122 532 465 | *PRDM6* | T | 0.33 | −0.76 ± 0.17 | $8.38 \times 10^{-6}$ | 0.34 | −0.50 ± 0.23 | 0.026 | −0.67 | $9.63 \times 10^{-7}$ | 0.37 | 0.00 | |
| *PR* | | | | | | | | | | | | | | | |
| rs17026156 | 2 | 40 614 469 | *SLC8A1* | C | 0.39 | 2.39 ± 0.40 | $\mathbf{2.85 \times 10^{-9}}$ | 0.28 | 2.00 ± 0.68 | 0.003 | 2.29 | $\mathbf{3.79 \times 10^{-11}}$ | 0.62 | 0.00 | |
| rs7948943 | 11 | 97 642 455 | Intergenic | T | 0.29 | 1.96 ± 0.43 | $5.21 \times 10^{-6}$ | 0.28 | 1.00 ± 0.68 | 0.139 | 1.68 | $3.45 \times 10^{-6}$ | 0.23 | 30.47 | |
| rs3136189 | 16 | 13 942 202 | *ERCC4* | C | 0.24 | 1.82 ± 0.46 | $7.60 \times 10^{-5}$ | 0.24 | 0.96 ± 0.71 | 0.176 | 1.56 | $4.97 \times 10^{-5}$ | 0.31 | 3.86 | |
| rs6499591 | 16 | 71 464 505 | *ZFHX3* | C | 0.21 | −1.97 ± 0.49 | $5.58 \times 10^{-5}$ | 0.24 | −1.12 ± 0.70 | 0.108 | −1.69 | $2.39 \times 10^{-5}$ | 0.32 | 0.00 | |
| rs17744182 | 18 | 28 286 523 | *GAREM* | A | 0.39 | 1.84 ± 0.41 | $7.91 \times 10^{-6}$ | 0.39 | 0.88 ± 0.62 | 0.157 | 1.55 | $6.51 \times 10^{-6}$ | 0.20 | 39.97 | |

| SNP ID | CHR | BP | Gene[a] | Coded allele | Phase 3 ($n = 5010$) | | | Meta-analysis (Phase 1 + 2 + 3) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | AF | Beta ± SE | *P*-value | Beta | *P*-value | *Q* | $I^2$ |
| QRS | | | | | | | | | | | |
| rs2483280 | 1 | 3 245 399 | *PRDM16* | A | 0.28 | −0.68 ± 0.17 | $8.19 \times 10^{-5}$ | −0.75 | $\mathbf{1.51 \times 10^{-11}}$ | 0.55 | 0.00 |
| rs335206 | 5 | 122 532 465 | *PRDM6* | T | 0.35 | −0.64 ± 0.16 | $8.06 \times 10^{-5}$ | −0.66 | $\mathbf{3.19 \times 10^{-10}}$ | 0.66 | 0.00 |
| *PR* | | | | | | | | | | | |
| rs17026156 | 2 | 40 614 469 | *SLC8A1* | C | 0.29 | 1.74 ± 0.45 | $9.44 \times 10^{-5}$ | 2.08 | $\mathbf{2.58 \times 10^{-14}}$ | 0.55 | 0.00 |
| rs7948943 | 11 | 97 642 455 | Intergenic | T | 0.29 | −1.18 ± 0.45 | $9.12 \times 10^{-3}$ | 0.56 | $4.16 \times 10^{-2}$ | 0.00 | 92.25 |
| rs3136189 | 16 | 13 942 202 | *ERCC4* | C | 0.24 | −0.82 ± 0.47 | $8.35 \times 10^{-2}$ | 0.61 | $4.02 \times 10^{-2}$ | 0.00 | 87.72 |
| rs6499591 | 16 | 71 464 505 | *ZFHX3* | C | 0.24 | −0.11 ± 0.47 | $8.10 \times 10^{-1}$ | −1.03 | $7.25 \times 10^{-4}$ | 0.02 | 73.24 |
| rs17744182 | 18 | 28 286 523 | *GAREM* | A | 0.40 | 1.26 ± 0.42 | $2.40 \times 10^{-3}$ | 1.43 | $6.25 \times 10^{-8}$ | 0.38 | 0.00 |

[a]Genes are defined as the gene closest to the SNP within a 200-kb window (HaploReg v2). Bold indicates genome-wide significant *P*-values ($5 \times 10^{-8}$).
CHR, chromosome; BP, base pair; AF, coded allele frequency; *Q*, *P*-value for Cochrane's Q statistic; $I^2$, heterogeneity index.

Downloaded from http://hmg.oxfordjournals.org/ at Kyoto University on March 21, 2015

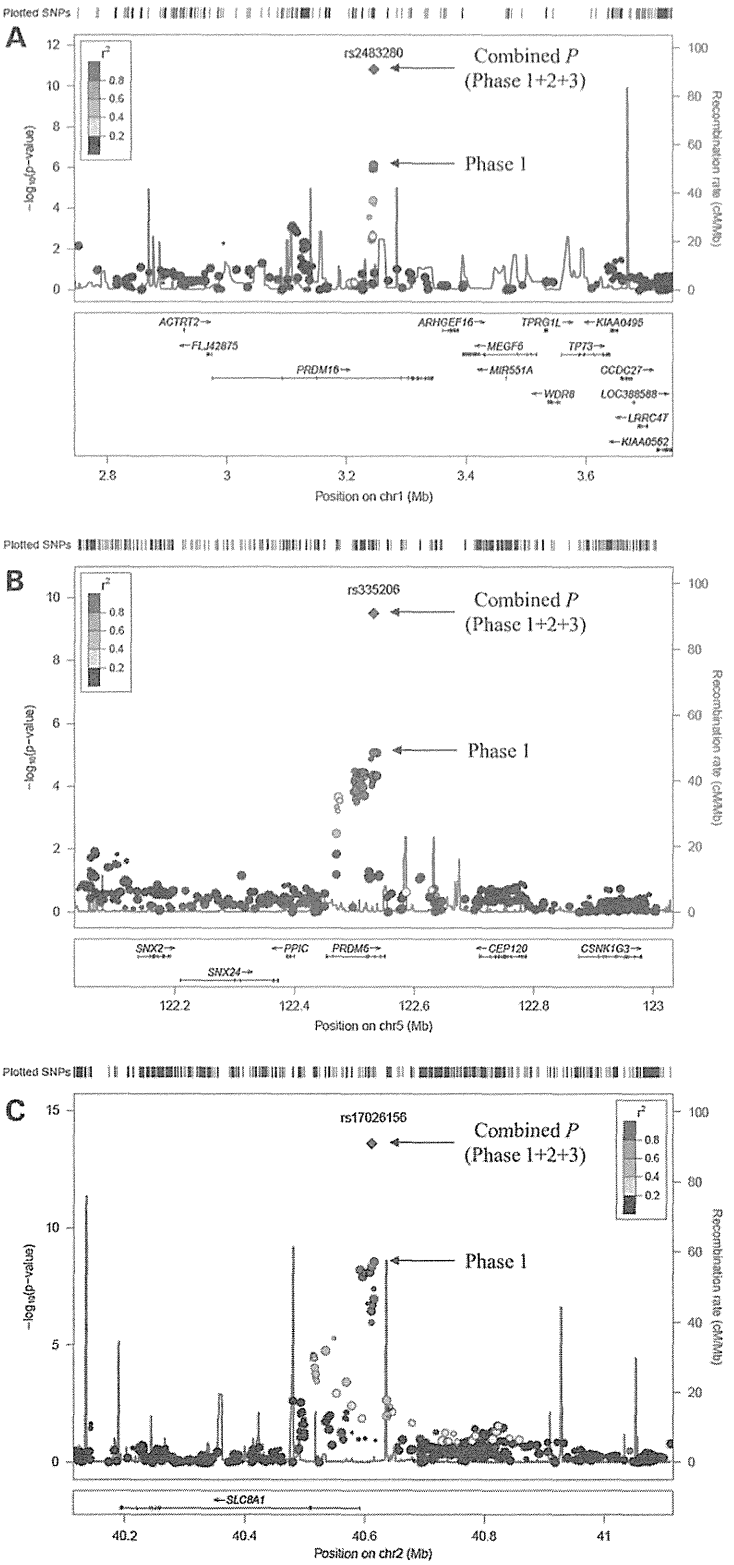


**Figure 2.** Signal plots for three novel loci across a 1-Mb window. Association of individual SNPs in the Phase 1 study plotted as $-\log10(p)$ against chromosomal base-pair position. The *y*-axis on the right shows the recombination rate, estimated from the HapMap CHB and JPT populations. All *P*-values are from the discovery phase. The purple diamond represents the meta-analysis results of the Phase 1, 2 and 3 studies. (**A**) rs2483280 of PRDM16, (**B**) rs335206 of PRDM6 and (**C**) rs17026156 of SLC8A1.

Downloaded from http://hmg.oxfordjournals.org/ at Kyoto University on March 21, 2015

**Table 3.** Extension of variants identified in European descendants to Koreans

| Reported gene | Ref. | European GWAS SNP | Coded allele | AF | Beta ± SE | *P*-value | Korean GWAS SNP | Coded allele | AF | Beta ± SE | *P*-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| QRS | | | | | | | | | | | |
| *NFIA* | 16 | rs9436640 | G | 0.46 | −0.59 ± 0.07 | $4.57 \times 10^{-18}$ | rs2103883 | C | 0.35 | −0.57 ± 0.17 | $6.85 \times 10^{-4}$ |
| *CRIM1* | 16 | rs7562790 | G | 0.40 | 0.39 ± 0.07 | $8.22 \times 10^{-9}$ | rs7562790 | G | 0.65 | 0.64 ± 0.17 | $1.52 \times 10^{-4}$ |
| *HEATR5B-STRN* | 16 | rs17020136 | C | 0.21 | 0.51 ± 0.08 | $1.90 \times 10^{-9}$ | rs2160411 | T | 0.55 | 0.55 ± 0.16 | $6.76 \times 10^{-4}$ |
| *HAND1-SAP30L* | 16 | rs13165478 | A | 0.36 | −0.55 ± 0.07 | $7.36 \times 10^{-14}$ | rs6580083 | G | 0.30 | −0.85 ± 0.18 | $1.29 \times 10^{-6}$ |
| *CDKN1A* | 14,16 | rs9470361 | A | 0.25 | 0.87 ± 0.08 | $3.00 \times 10^{-27}$ | rs9470366 | T | 0.15 | 1.16 ± 0.23 | $4.02 \times 10^{-7}$ |
| *VTI1A* | 16 | rs7342028 | T | 0.27 | 0.48 ± 0.08 | $4.95 \times 10^{-10}$ | rs10885378 | G | 0.47 | 0.34 ± 0.16 | $3.55 \times 10^{-2}$ |
| *SETBP1* | 16 | rs991014 | T | 0.42 | 0.42 ± 0.07 | $6.2 \times 10^{-10}$ | rs4890489 | A | 0.32 | 0.97 ± 0.17 | $2.32 \times 10^{-8}$ |
| PR | | | | | | | | | | | |
| *MEIS1* | 15 | rs11897119 | C | 0.39 | 1.36 ± 0.21 | $4.62 \times 10^{-11}$ | rs4430933 | G | 0.76 | −1.54 ± 0.47 | $1.02 \times 10^{-3}$ |
| *ARHGAP24* | 14,15 | rs7692808 | A | 0.31 | −2.01 ± 0.22 | $5.99 \times 10^{-20}$ | rs10012090 | A | 0.92 | −1.87 ± 0.76 | $1.33 \times 10^{-2}$ |
| *SOX5* | 15 | rs11047543 | A | 0.15 | −2.09 ± 0.29 | $3.34 \times 10^{-13}$ | rs4246224 | A | 0.14 | −1.89 ± 0.58 | $1.16 \times 10^{-3}$ |
| QRS, PR | | | | | | | | | | | |
| *EXOG-SCN5A-SCN10A* | 14–16 | rs9851724 | C | 0.33 | −0.66 ± 0.07 | $1.91 \times 10^{-20}$ | rs7633988 | T | 0.29 | −0.86 ± 0.18 | $1.50 \times 10^{-6}$ |
| | | rs6800541 | C | 0.40 | 3.77 ± 0.21 | $2.10 \times 10^{-74}$ | rs7433306 | C | 0.15 | 4.33 ± 0.56 | $1.45 \times 10^{-14}$ |
| *CAV1-CAV2* | 14,15 | rs3807989 | A | 0.40 | 3.3 | $1.10 \times 10^{-4}$ | rs11773845 | C | 0.34 | 0.60 ± 0.17 | $5.36 \times 10^{-4}$ |
| | | | | | 2.30 ± 0.21 | $3.66 \times 10^{-28}$ | | | | 3.21 ± 0.42 | $3.33 \times 10^{-14}$ |
| *TBX5-TBX3* | 15,16 | rs10850409 | A | 0.27 | −0.49 ± 0.08 | $3.06 \times 10^{-10}$ | rs3914956 | T | 0.49 | −0.41 ± 0.16 | $1.11 \times 10^{-2}$ |
| | | rs1896312 | C | 0.28 | 1.95 ± 0.23 | $3.13 \times 10^{-17}$ | rs10744836 | T | 0.52 | 1.29 ± 0.40 | $1.41 \times 10^{-3}$ |

Ref, References; AF, coded allele frequency.

using ENCODE data and the web-based program RegulomeDB. Further, their evolutional conservation was studied by comparing the allelic sequences with primate genome sequences using Ensembl data.

The DNA sequence that encompassed the rs2483280 SNP was predicted to be a ZBTB3 transcription factor-binding motif and an open chromatin region (DNase I-hypersensitive region). However, the SNP sequence was not conserved in primate DNA (Fig. 4A). Thus, we searched for high-LD SNPs near the lead SNP and identified rs2255212 1.5 kb away from the lead SNP (LD score $r^2 = 0.98$ and $D' = 1.00$) (Fig. 4A). rs2255212 was predicted to be a TCF4 transcription factor-binding site and an open chromatin region, and the SNP was highly conserved in all primates. However, the association *P*-value of rs2255212 ($1.16 \times 10^{-6}$) was not better than that of the lead SNP (rs2483280, $P = 7.47 \times 10^{-7}$). rs335206 was predicted to be a ZNF263-binding site and conserved in all primates, but not an open chromatin region (Fig. 4B).

rs17026156 did not match any functionally conserved sequence, although it was conserved in all primates. Thus, we searched for LD SNPs near the lead SNP and identified a high-LD ($r^2 = 0.93$ and $D' = 1.00$) SNP (rs13017846) that was predicted to be a PIT-1-binding site (Fig. 4C). However, the association *P*-value of rs13017846 ($4.47 \times 10^{-9}$) was also not better than that of the lead SNP (rs17026156, $P = 2.85 \times 10^{-9}$).

Based on simulation study, the functional variant in a GWAS locus may not have the most significance owing to random sampling. Therefore, functional validation is required to implicate or dismiss rs13017846 and rs2255212, although they did not exceed the significance *P*-values of the lead SNPs.

Notably, the allele frequency of rs17026156 varies widely between ethnicities. The allele frequencies of the ancestral C allele are as low as 0.05 in Europeans (HapMap-CEU) and 0.08 in Africans (HapMap-YRI), whereas they reach as high as 0.57 in Chinese (HapMap-HCB) and 0.35 in Japanese (HapMap-JPT). The rs17026156 was not identified in previous GWASs in European descendants, possibly due to its low allelic frequency in Europeans. Individuals with the C allele of rs17026156 increased PR intervals (beta ± SE, 2.39 ± 0.40 ms) in East Asian population.

## Annotation of proximal genes

rs2483280, associated with QRS duration, lies in the third intron of *PRDM16* (PR domain-containing 16), which encodes a protein with a zinc finger DNA-binding domain and PR domain. PRDM16 regulates brown adipose tissue differentiation (20), and its genetic region translocates frequently chromosome 3q21, causing acute myeloid leukemia and myelodysplastic syndrome (21). Recently, fine mapping analysis of 1p36 deletion syndrome implicated a mutation in *PRDM16* as a cause of cardiomyopathy with left ventricular non-compaction and dilated cardiomyopathy, both of which are characterized by progressive cardiac dysfunction, resulting in heart failure, arrhythmia and sudden cardiac death (22).

*PRDM16*, expressed in the nuclei of cardiomyocytes, potentiates cardiomyocyte proliferation. Haploinsufficiency of *PRDM16* in zebrafish results contractile dysfunction and the reduction of ventricular conduction velocity (22), supporting our finding that rs2483280 in *PRDM16* is associated with QRS duration.

Downloaded from http://hmg.oxfordjournals.org/ at Kyoto University on March 21, 2015

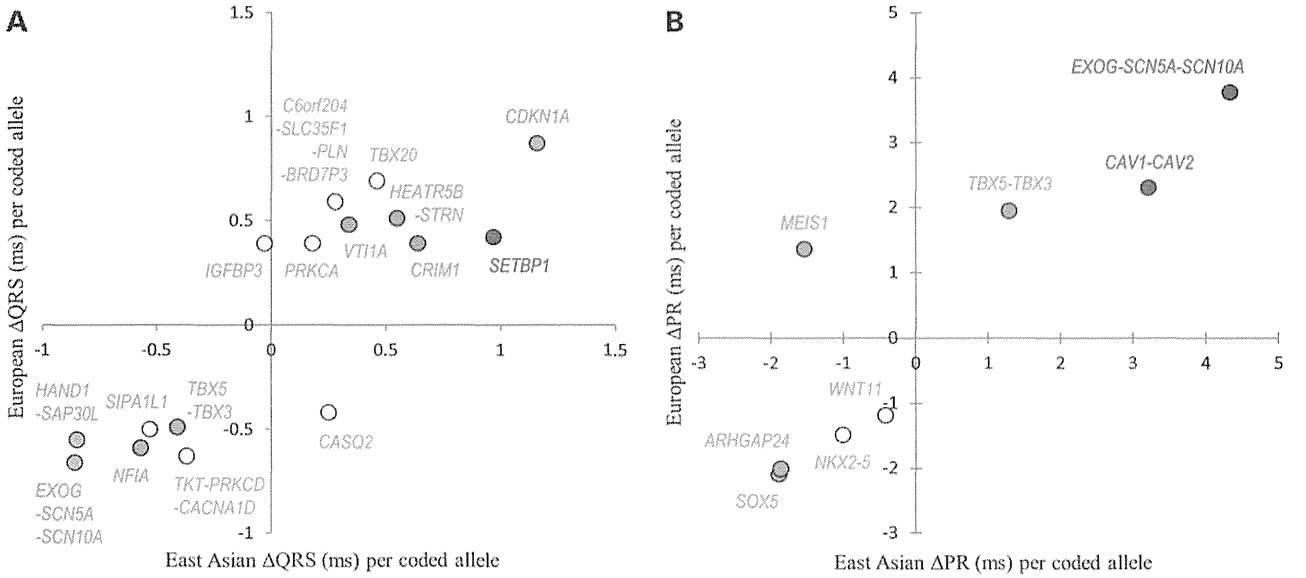


**Figure 3.** Comparison of effect size ($\beta$) of SNPs previously identified in Europeans with those in East Asians. Each dot refers to an association signal, with colors (red, $P < 5 \times 10^{-8}$; orange, $5 \times 10^{-8} \leq P < 10^{-5}$; green, $10^{-5} \leq P < 0.05$; white, $P \geq 0.05$). *CAV1-CAV2* locus was not included in QRS plot because it did not reach genome-wide significance in European GWAS. Effect size was presented beta per copy of the coded allele. (**A**) QRS duration and (**B**) PR interval.

Another novel SNP was associated with QRS duration—rs335206, in the fourth intron of *PRDM6* (PR domain-containing 6), a transcriptional repressor in smooth muscle cells. *PRDM6* regulates the development, differentiation and proliferation of blood vessels (23), and rs335206 was recently linked to systolic blood pressure (24). It is recently reported in mice that *Prdm6* knockout embryos die during development, displaying signs of cardiac insufficiency including a thinning of the myocardial walls (25).

rs17026156, associated with PR interval, lies ~21 kb upstream of *SLC8A1* (sodium/calcium exchanger 1 precursor). SLC8A1 extrudes calcium from cardiac myocytes during relaxation and returns the myocardium to its resting state after excitation (26). Targeted disruption of *SLC8A1* causes defects in heartbeat—*SLC8A1-/-* mouse embryos experience slow and arrhythmic heart contractions (27). We have also reported this locus to correlate with QT interval traits in East Asians (19). Based on the previous reports, *SLC8A1* appears to mediate electrophysiological conductivity during heart.

In conclusion, we have identified three novel loci for QRS duration and PR interval and confirmed 13 previously reported loci. These data will increase our understanding in the genetic architecture that underlies the mechanisms of electrocardiographic traits, QRS duration and PR interval.

## MATERIALS AND METHODS

### Subjects

The study subjects have been described in a QT interval GWAS (19). Briefly, 6805 subjects from KARE were selected from an ongoing population-based cohort, as part of the Korean Genome and Epidemiology Study (KoGES). Subjects without a self-reported history of cardiac disease, concurrent use of medications that interfere with the ECG measurements and abnormal electrolyte values at the ECG were included. Written informed consent was obtained from all participants, and this project was approved by the institutional review board of the Korea National Institute of Health.

The Phase 2 Japanese subjects were part of the Nagahama Prospective Genome Cohort for Comprehensive Human Bioscience (The Nagahama Study). The Nagahama Study cohort was recruited from 2008 to 2010 from the general population in Nagahama City, a largely rural city of 125 000 inhabitants in Shiga Prefecture that lies in the center of Japan. Of the 9804 participants, persons whose genome-wide SNP was analyzed ($n$ = 3710) and who were free of symptomatic cardiovascular disease and abnormal ECG readings ($n$ = 2285) were used in the second GWAS panel. All clinical measurements and sampling of blood were performed on enrollment. Genomic DNA was extracted from peripheral blood samples with phenol–chloroform.

The Phase 3 replication panel comprised Japanese from three independent subcohorts. First, the Anti-Aging Center (AAC) cohort included consecutive participants in the medical check-up program at Ehime University Hospital, which was designed specifically to evaluate age-related disorders, including atherosclerosis, cardiovascular disease, physical function and mild cognitive impairment. All clinical data in this study were obtained through the check-up process.

With regard to the second subcohort, the Takashima Study is an ongoing longitudinal study, based on community residents in Takashima City. Takashima City is a semiurban area in Shiga Prefecture, with a population of ~54 000. Study subjects were recruited between 2002 and 2003 from participants of the annual medical check-up program, held by Takashima City. The basic clinical parameters in this study were obtained from the personal medical check-up records of the subjects. The third subcohort of the replication analysis comprised the remaining sample of the Nagahama Study.

All study procedures in Japan were approved by the ethics committee of Ehime University Graduate School of Medicine,

Downloaded from http://hmg.oxfordjournals.org/ at Kyoto University on March 21, 2015

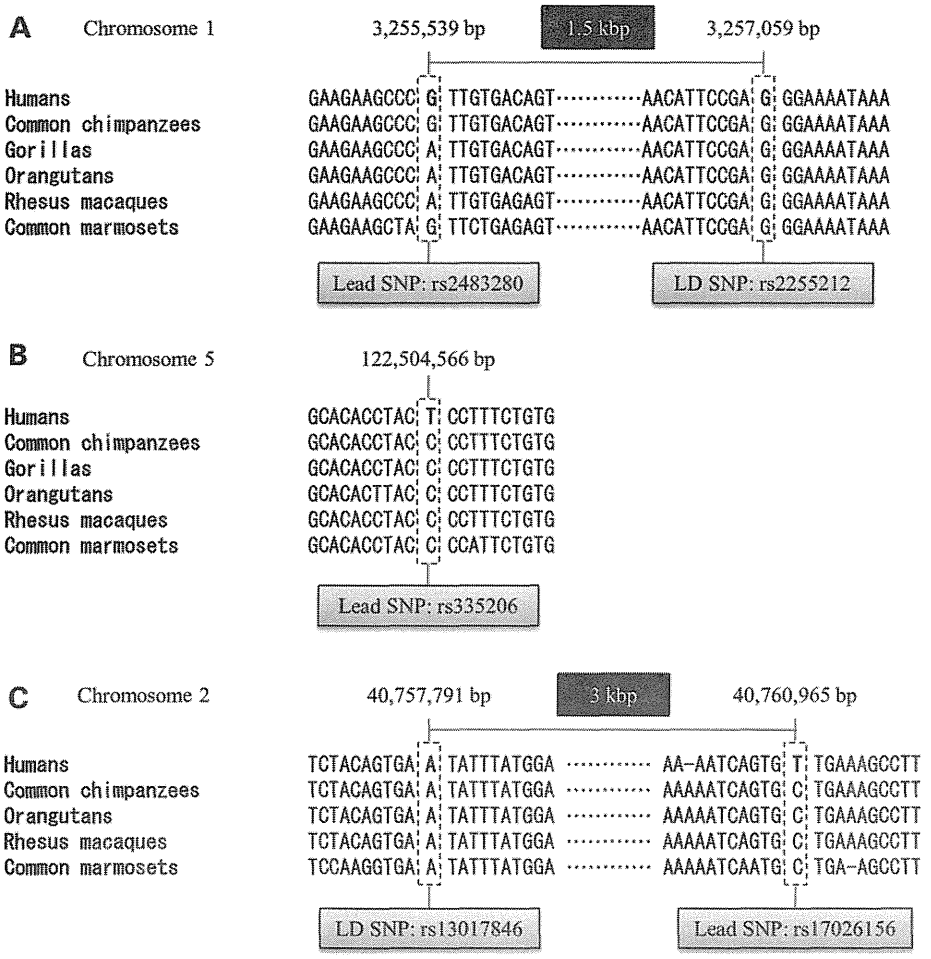


**Figure 4.** Evolutionary conservation of three novel genetic variants. The human and primate sequences of the SNP ± 10 bp were obtained from Ensembl website. (**A**) rs2483280 of PRDM16, (**B**) rs335206 of PRDM6 and (**C**) rs17026156 of SLC8A1.

Shige University of Medical Science and Kyoto University Graduate School of Medicine.

### ECG measurements

PR interval and QRS duration values were obtained from a supine 12-lead ECG using digital electrocardiographic recorders—Phase 1, MAC5000 (GE Medical System, CT, USA); Phase 2, FCP-7411 and FCP-7431 (Fukuda Denshi, Tokyo, Japan); Phase 3, AAC, ECG-1500 (Nihon Kohden, Tokyo, Japan) and Takashima, FCP-4720 (Fukuda Denshi, Tokyo, Japan). ECGs with insufficient quality (e.g., owing to baseline drift or missing leads) and those with rhythms other than sinus rhythm or AF were excluded. PR interval was measured from the onset of the P-wave to the onset of ventricular depolarization. QRS duration was measured from the onset of ventricular depolarization to the J point.

### Genotyping

The genotyping data were obtained from KARE, which used the Affymetrix Genomewide Human SNP Array 5.0. The genotype quality control criteria have been reported in a previous GWAS study (28). Briefly, the criteria for the inclusion of SNPs were genotype call rate of $>0.98$, minor allele frequency (MAF) of $>0.01$ and Hardy–Weinberg equilibrium (HWE) ($P > 1 \times 10^{-6}$). The related individuals were excluded from the KARE genotype dataset, whose computed average pairwise identity-by-state value was higher than that estimated from first-degree relatives of Korean sib-pair samples ($>0.80$, $n = 601$). Ultimately, 352 228 SNPs passed the quality control process and were subsequently used in the GWASs for PR interval and QRS duration. SNP imputation was performed with IMPUTE (29) using the JPT and CHB sets of HapMap Phase 2 as references. After removing SNPs with MAF of $<0.01$ and SNP missing rate of $>0.05$, we combined the remaining 1.8 million imputed SNPs with the SNPs that were typed directly in KARE for the association analysis.

Genome-wide SNP genotyping of the Nagahama sample was performed using a series of BeadChip DNA arrays (Illumina, San Diego, CA, USA). Genotyping quality was controlled by excluding SNPs with call rates of $<99\%$, with an MAF of $<0.1\%$, and deviating significantly from HWE ($P < 1 \times 10^{-7}$).

Downloaded from http://hmg.oxfordjournals.org/ at Kyoto University on March 21, 2015

Individuals who met the following criteria were excluded from analysis: average genotype call rate <95%, high degree of kinship (Pi-hat >0.35 [PLINK version 1.07 (30)]), and identified as an ancestry outlier by principal component analysis with the HapMap Phase 2, release 28 JPT dataset as the reference [EIGENSTRAT version 2.0 (31)]. Genotype imputation was performed using MACH, version 1.0.16 (32). Imputed SNPs for which the MAF was <0.01 or R-square value was <0.5 were excluded from the association analysis.

Replication genotyping of the Phase 3 sample was performed using a TaqMan probe assay and commercially available primer and probe sets (Life Technologies Corporation, Carlsbad, CA, USA). The fluorescence level of the PCR products was measured on a 7900HT Fast Real-Time PCR System (Applied Biosystems, Foster, CA, USA).

### Statistical analysis

The effect of a genotype was analyzed by linear regression. The effect size (beta) and standard error (SE) of coded alleles were calculated on PR interval and QRS duration. All analyses were adjusted for age, sex, recruitment area, BMI, systolic blood pressure and height. PLINK (30) was used for all statistical tests. All tests were based on an additive model, and Phase 1 SNPs for replication test were selected, based on $P < 1 \times 10^{-4}$. We combined Phase 1 and Phase 2 data by inverse-variance meta-analysis under the assumption of fixed effects using Cochran's Q test to determine between-study heterogeneity (33). Phase 1 + Phase 2 SNPs were selected, based on meta-analyses $P$-values that were more significant than Phase 1 $P$-values. Finally, Phase 1 + Phase 2 + Phase 3 meta-analyses were conducted, and through which we identified significant genome-wide-level variants. All meta-analysis calculations were implemented in PLINK (30) (version 1.07).

### *In silico* functional analysis of novel SNPs

Proximal SNP and LD were computed using SNAP, a web-based software program (http://www.broadinstitute.org/mpg/snap/ldsearchpw.php) (34). Evolutional conservation was confirmed using the Ensembl Genome browser (http://www.ensembl.org/index.html), comparing the SNP ± 10 bp in primates. The functional elements that were linked to the associated SNPs were analyzed using the RegulomeDB (http://regulome.stanford.edu/), which was developed by the ENCODE project (35).

## SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

*Conflict of Interest statement*. None declared.

## FUNDING

The genotype and epidemiological data were provided by the Korean Genome Analysis Project (4845-301) and the Korean Genome and Epidemiology Study (4851-302), funded by the Ministry for Health and Welfare, Republic of Korea. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A2012069). This work was supported by a National Research Foundation of Korea (NRF) grant, funded by the Korean government (MSIP)(NRF-2011-0030072).

## APPENDIX

Principal investigators of the Japanese study cohorts are as follows:

Nagahama Study: Fumihiko Matsuda (chairperson), Yasuharu Tabara, Takahisa Kawaguchi, Yoshimitsu Takahashi, Kazuya Setoh, Chikashi Terao, Ryo Yamada, Akihiro Sekine, Shinji Kosugi and Takeo Nakayama (Kyoto University Graduate School of Medicine, and School of Public Health); the AAC study: Yasuharu Tabara (chairperson), Katsuhiko Kohara, Michiya Igase and Tetsuro Miki (Ehime University Graduate School of Medicine)

Takashima study: Yoshikuni Kita (chairperson), Hirotsugu Ueshima and Naoyuki Takashima (Shiga University of Medical Science).

## REFERENCES

1. Saksena, S. and Camm, J.A. (2011) *Electrophysiological Disorders of the Heart*. Elsevier Saunders.
2. Algra, A., Tijssen, J.G., Roelandt, J.R., Pool, J. and Lubsen, J. (1991) QTc prolongation measured by standard 12-lead electrocardiography is an independent risk factor for sudden death due to cardiac arrest. *Circulation*, **83**, 1888–1894.
3. Desai, A.D., Yaw, T.S., Yamazaki, T., Kaykha, A., Chun, S. and Froelicher, V.F. (2006) Prognostic significance of quantitative QRS duration. *Am. J. Med.*, **119**, 600–606.
4. Benjamin, E.J., Chen, P.S., Bild, D.E., Mascette, A.M., Albert, C.M., Alonso, A., Calkins, H., Connolly, S.J., Curtis, A.B., Darbar, D. *et al.* (2009) Prevention of atrial fibrillation: report from a national heart, lung, and blood institute workshop. *Circulation*, **119**, 606–618.
5. Fang, F., Sanderson, J.E. and Yu, C.M. (2013) Potential role of biventricular pacing beyond advanced systolic heart failure. *Circ. J.*, **77**, 1364–1369.
6. Aro, A.L., Anttonen, O., Tikkanen, J.T., Junttila, M.J., Kerola, T., Rissanen, H.A., Reunanen, A. and Huikuri, H.V. (2011) Intraventricular conduction delay in a standard 12-lead electrocardiogram as a predictor of mortality in the general population. *Circ. Arrhythm. Electrophysiol.*, **4**, 704–710.
7. Marijon, E., Trinquart, L., Otmani, A., Waintraub, X., Kacet, S., Clementy, J., Chatellier, G. and Le Heuzey, J.Y. (2009) Competing risk analysis of cause-specific mortality in patients with an implantable cardioverter-defibrillator: the EVADEF cohort study. *Am. Heart J.*, **157**, 391–397 e391.
8. Li, J., Huo, Y., Zhang, Y., Fang, Z., Yang, J., Zang, T. and Xu, X. (2009) Familial aggregation and heritability of electrocardiographic intervals and heart rate in a rural Chinese population. *Ann. Noninvasive Electrocardiol.*, **14**, 147–152.
9. Havlik, R.J., Garrison, R.J., Fabsitz, R. and Feinleib, M. (1980) Variability of heart rate, P-R, QRS and Q-T durations in twins. *J. Electrocardiol.*, **13**, 45–48.
10. Mathers, J.A., Osborne, R.H. and Degeorge, F.V. (1961) Studies of blood pressure, heart rate, and the electrocardiogram in adult twins. *Am. Heart J.*, **62**, 634–642.
11. Moller, P., Heiberg, A. and Berg, K. (1982) The atrioventricular conduction time - a heritable trait? III. Twin studies. *Clin. Genet.*, **21**, 181–183.
12. Russell, M.W., Law, I., Sholinsky, P. and Fabsitz, R.R. (1998) Heritability of ECG measurements in adult male twins. *J. Electrocardiol.*, **30**(Suppl), 64–68.
13. Singh, J.P., Larson, M.G., O'Donnell, C.J., Tsuji, H., Evans, J.C. and Levy, D. (1999) Heritability of heart rate variability: the Framingham Heart Study. *Circulation*, **99**, 2251–2254.
14. Holm, H., Gudbjartsson, D.F., Arnar, D.O., Thorleifsson, G., Thorgeirsson, G., Stefansdottir, H., Gudjonsson, S.A., Jonasdottir, A., Mathiesen, E.B., Njolstad, I. *et al.* (2010) Several common variants modulate heart rate, PR interval and QRS duration. *Nat. Genet.*, **42**, 117–122.
15. Pfeufer, A., van Noord, C., Marciante, K.D., Arking, D.E., Larson, M.G., Smith, A.V., Tarasov, K.V., Muller, M., Sotoodehnia, N., Sinner, M.F. *et al.* (2010) Genome-wide association study of PR interval. *Nat. Genet.*, **42**, 153–159.

Downloaded from http://hmg.oxfordjournals.org/ at Kyoto University on March 21, 2015

16. Sotoodehnia, N., Isaacs, A., de Bakker, P.I., Dorr, M., Newton-Cheh, C., Nolte, I.M., van der Harst, P., Muller, M., Eijgelsheim, M., Alonso, A. *et al.* (2010) Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nat. Genet.*, **42**, 1068–1076.
17. Newton-Cheh, C., Eijgelsheim, M., Rice, K.M., de Bakker, P.I., Yin, X., Estrada, K., Bis, J.C., Marciante, K., Rivadeneira, F., Noseworthy, P.A. *et al.* (2009) Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nat. Genet.*, **41**, 399–406.
18. Pfeufer, A., Sanna, S., Arking, D.E., Muller, M., Gateva, V., Fuchsberger, C., Ehret, G.B., Orru, M., Pattaro, C., Kottgen, A. *et al.* (2009) Common variants at ten loci modulate the QT interval duration in the QTSCD Study. *Nat. Genet.*, **41**, 407–414.
19. Kim, J.W., Hong, K.W., Go, M.J., Kim, S.S., Tabara, Y., Kita, Y., Tanigawa, T., Cho, Y.S., Han, B.G. and Oh, B. (2012) A common variant in SLC8A1 is associated with the duration of the electrocardiographic QT interval. *Am. J. Hum. Genet.*, **91**, 180–184.
20. Borensztein, M., Viengchareun, S., Montarras, D., Journot, L., Binart, N., Lombes, M. and Dandolo, L. (2012) Double Myod and Igf2 inactivation promotes brown adipose tissue development by increasing Prdm16 expression. *FASEB J.*, **26**, 4584–4591.
21. Mochizuki, N., Shimizu, S., Nagasawa, T., Tanaka, H., Taniwaki, M., Yokota, J. and Morishita, K. (2000) A novel gene, MEL1, mapped to 1p36.3 is highly homologous to the MDS1/EVI1 gene and is transcriptionally activated in t(1;3)(p36;q21)-positive leukemia cells. *Blood*, **96**, 3209–3214.
22. Arndt, A.K., Schafer, S., Drenckhahn, J.D., Sabeh, M.K., Plovie, E.R., Caliebe, A., Klopocki, E., Musso, G., Werdich, A.A., Kalwa, H. *et al.* (2013) Fine mapping of the 1p36 deletion syndrome identifies mutation of PRDM16 as a cause of cardiomyopathy. *Am. J. Hum. Genet.*, **93**, 67–77.
23. Davis, C.A., Haberland, M., Arnold, M.A., Sutherland, L.B., McDonald, O.G., Richardson, J.A., Childs, G., Harris, S., Owens, G.K. and Olson, E.N. (2006) PRISM/PRDM6, a transcriptional repressor that promotes the proliferative gene program in smooth muscle cells. *Mol. Cell. Biol.*, **26**, 2626–2636.
24. Gaal, E.I., Salo, P., Kristiansson, K., Rehnstrom, K., Kettunen, J., Sarin, A.P., Niemela, M., Jula, A., Raitakari, O.T., Lehtimaki, T. *et al.* (2012) Intracranial aneurysm risk locus 5q23.2 is associated with elevated systolic blood pressure. *PLoS Genet.*, **8**, e1002563.
25. Gewies, A., Castineiras-Vilarino, M., Ferch, U., Jahrling, N., Heinrich, K., Hoeckendorf, U., Przemeck, G.K., Munding, M., Gross, O., Schroeder, T. *et al.* (2013) Prdm6 is essential for cardiovascular development in vivo. *PLoS One*, **8**, e81833.
26. Bers, D.M. and Despa, S. (2009) Na+ transport in cardiac myocytes; implications for excitation-contraction coupling. *IUBMB Life*, **61**, 215–221.
27. Wakimoto, K., Kobayashi, K., Kuro, O.M., Yao, A., Iwamoto, T., Yanaka, N., Kita, S., Nishida, A., Azuma, S., Toyoda, Y. *et al.* (2000) Targeted disruption of Na+/Ca2+ exchanger gene leads to cardiomyocyte apoptosis and defects in heartbeat. *J. Biol. Chem.*, **275**, 36991–36998.
28. 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. *et al.* (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.*, **41**, 527–534.
29. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.
30. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
31. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
32. Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.
33. Ioannidis, J.P., Patsopoulos, N.A. and Evangelou, E. (2007) Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One*, **2**, e841.
34. Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J. and de Bakker, P.I. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, **24**, 2938–2939.
35. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, **22**, 1790–1797.

Downloaded from http://hmg.oxfordjournals.org/ at Kyoto University on March 21, 2015

ARTHRITIS & RHEUMATOLOGY
Vol. 66, No. 12, December 2014, pp 3395–3403
DOI 10.1002/art.38867
© 2014, American College of Rheumatology

# Association Between Antinuclear Antibodies and the HLA Class II Locus and Heterogeneous Characteristics of Staining Patterns

## The Nagahama Study

Chikashi Terao, Koichiro Ohmura, Ryo Yamada, Takahisa Kawaguchi, Masakazu Shimizu, Yasuharu Tabara, Meiko Takahashi, Kazuya Setoh, Takeo Nakayama, Shinji Kosugi, Akihiro Sekine, Fumihiko Matsuda, and Tsuneyo Mimori
on behalf of the Nagahama Study Group

*Objective.* **While antinuclear antibodies (ANAs) are observed in healthy populations as well as in patients with autoimmune diseases such as systemic lupus erythematosus (SLE), the detailed genetic background of ANAs has remained unclear. We undertook this study to identify the genetic determinants of ANAs in the general population in order to elucidate the underlying mechanisms of ANA production and to distinguish disease susceptibility genes from ANA production genes.**

*Methods.* **A total of 9,575 Japanese volunteers were registered, and their ANA levels were quantified using indirect immunofluorescence to analyze correlates of ANA positivity. Genetic studies were performed using 7,148 of the 9,575 subjects. We performed a genome-wide association study using 3,185 subjects genotyped for 303,506 single-nucleotide polymorphisms (SNPs), followed by a replication study of 3,963 subjects. HLA–DRB1 and HLA–DQB1 alleles were imputed, and associations between ANA positivity and the SNPs or the HLA alleles associated with SLE were analyzed.**

*Results.* **Female sex and old age were associated with ANA positivity, except for the nucleolar pattern. The T allele of rs2395185 in the HLA locus, which was in moderate linkage disequilibrium with HLA–DRB1*0405, was significantly associated with ANA positivity ($P = 1.3 \times 10^{-11}$). The T allele of rs2395185 displayed increasing effects on the frequency of speckled and homogeneous patterns ($P = 7.5 \times 10^{-12}$ and $P = 2.2 \times 10^{-11}$, respectively) but decreasing effects on the frequency of the nucleolar pattern ($P = 0.0045$). The 7 SNPs and 4 HLA–DRB1 alleles associated with SLE did not display strong associations with ANA positivity.**

*Conclusion.* **SNP rs2395185 linked with HLA–DRB1*0405 is a genetic determinant of ANA production in the Japanese population. Overlapping of loci for susceptibility to SLE and to ANA positivity was limited. The nucleolar pattern showed different associations from other staining patterns, both with correlates of ANA positivity and with the HLA locus.**

Antinuclear antibodies (ANAs) are autoantibodies that recognize various nuclear and cytoplasmic proteins, and they are frequently observed in patients with a broad range of diseases including systemic lupus erythematosus (SLE), hepatic disease, malignant disease, lung disease, and a variety of infections (1–6). The distribution patterns of fluorescent types of ANAs (such as speckled, homogeneous, nucleolar, or discrete speck-

Supported by the Ministry of Education, Culture, Sports, Science, and Technology of Japan (Grant-in-Aid for Scientific Research), Kyoto University (University Grant), the Japan Society for the Promotion of Science (Program for Enhancing Systematic Education in Graduate School Grant), and the Takeda Science Foundation (research grant).

Chikashi Terao, MD, PhD, Koichiro Ohmura, MD, PhD, Ryo Yamada, MD, PhD, Takahisa Kawaguchi, MSc, Masakazu Shimizu, PhD, Yasuharu Tabara, PhD, Meiko Takahashi, PhD, Kazuya Setoh, MSc, Takeo Nakayama, MD, PhD, Shinji Kosugi, MD, PhD, Akihiro Sekine, PhD, Fumihiko Matsuda, PhD, Tsuneyo Mimori, MD, PhD: Kyoto University, Kyoto, Japan.

Address correspondence to Chikashi Terao, MD, PhD, Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Shogoin-Kawahara-cho 54, Sakyo-ku, Kyoto 606-8507, Japan. E-mail: a0001101@kuhp.kyoto-u.ac.jp.

Submitted for publication March 1, 2014; accepted in revised form August 28, 2014.

led patterns) also provide useful information for differential diagnosis (7–9). Previous studies have suggested that it is not unusual to find healthy individuals who are positive for ANAs (10). Since ANAs are included in the classification criteria for SLE as well as those for autoimmune hepatitis (11,12), analyzing the kinds of variables that affect the levels of ANAs would be helpful for avoiding excessive or deficient classification of these diseases as well as for gaining insight into their etiologies.

Although previous studies showed that ANA positivity was associated with female sex, old age, and being overweight (13,14), genetic components affecting ANA positivity in healthy individuals have never been addressed. Genome-wide association studies (GWAS) have detected many genes that confer susceptibility to connective tissue diseases, including SLE (15–18), and have elucidated the genetic background of biomarkers in general populations (19). Because almost all patients with SLE are positive for ANAs, it is important to confirm that SLE-related genes in the previous GWAS were not merely derived from their associations with ANA positivity.

At present, the number of large-scale studies addressing ANA levels in healthy subjects is quite limited. Detailed analyses of the correlates and genetic components of ANAs in healthy individuals would provide clues to the mechanisms responsible for the production of autoantibodies and the development of autoantibody-mediated autoimmune diseases (20,21). In the present study, we quantified circulating levels of ANAs in 9,575 Japanese volunteers for detailed analyses of the distributions and effects of correlates on ANA production. We also performed a GWAS in 7,148 of the 9,575 subjects to detect susceptibility loci that affect ANA production.

## SUBJECTS AND METHODS

This study was approved by the Ethics Committee of Kyoto University Graduate School and Faculty of Medicine.

**Study population.** This study was performed as a part of the Nagahama Prospective Genome Cohort for Comprehensive Human Bioscience (the Nagahama Study), a community-based prospective multiomics cohort study conducted by the Center for Genomic Medicine at Kyoto University (22). A total of 9,809 volunteers ages 30–75 years in Nagahama City, Shiga Prefecture, Japan were recruited for this study. Written informed consent was obtained from each participant, and all were asked to complete a detailed questionnaire including present and past illnesses and lifestyle.

**Exclusion criteria.** We excluded volunteers from the association studies if they lacked necessary information or had ever been told that they have or had an autoimmune disease. We also excluded individuals whose answers to the questionnaire suggested that they might have an autoimmune disease. As a result, a total of 9,575 subjects remained for this study. A detailed flow chart of sample exclusion is shown in Supplementary Figure 1 (available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38867/abstract).

**Quantification of ANAs and C-reactive protein (CRP).** ANAs and CRP in serum samples from volunteers were quantified (23) at SRL, one of the largest clinical laboratory testing companies in Japan. ANAs were quantified by serum dilution using indirect immunofluorescence with HEp-2 cells (TFB). Titers of ANAs with detailed staining patterns (speckled, homogeneous, nucleolar, cytoplasmic, and discrete speckled patterns) were also reported for these subjects. A cutoff level of 1:40 for positivity was applied according to the manufacturer's instructions.

**Selection of potential correlates.** Age, sex, body weight, smoking, alcohol use, and serum CRP level were selected as potential correlates based on a previous US study (14). CRP was quantified by highly sensitive methods using nephelometry, with a detection limit of 0.051 mg/liter, as previously reported (23).

**Statistical analysis of nongenetic studies.** The subjects were divided into 2 subgroups based on sex, 9 subgroups based on age (5-year intervals), and 18 subgroups based on sex and age. Associations between ANAs and age and/or sex were assessed by standardized logistic regression analysis. Odds ratios were also calculated with 95% confidence intervals. The associations between ANAs and potential correlates were analyzed by logistic regression analysis, with sex and age as covariates. Statistical analyses were performed using R statistical software (http://www.r-project.org) or SPSS version 18. We set significance levels in a conservative manner using Bonferroni correction for multiple testing.

**GWAS.** DNA samples from 3,710 of the 9,809 participants in the Nagahama Study were genome-scanned using Illumina HumanHap610, HumanHapOmni2.5-4, or Human HapOmni2.5-8 arrays. A total of 392,801 single-nucleotide polymorphisms (SNPs) that were common between the arrays were selected for the GWAS. We selected 3,185 subjects with call rates of >0.95 who did not show a high degree of kinship (PI_HAT <0.35) and who did not have connective tissue diseases. SNPs that showed $P$ values less than $5 \times 10^{-7}$ and in Hardy-Weinberg equilibrium ($P > 1 \times 10^{-7}$) with a success rate of >0.95 and a minor allele frequency of >0.05 were selected for a replication study using a TaqMan Assay (Applied Biosystems) with 3,963 of the participants. Population stratification was assessed with genomic control (24). Logistic regression analysis was performed to analyze the genetic influence on the production of ANAs for each SNP, corrected by age and sex. Logistic regression analysis was also used for the conditioning analysis. The associations of the 2 studies were combined using the inverse-variance method. The Jonckheere-Terpstra test was used to assess increasing effects of SNPs on ANA levels in subjects positive for ANAs.

**HLA imputation.** The HLA–DRB1 locus (the established HLA locus associated with SLE in previous reports) and the HLA–DQB1 locus were imputed using the GWAS data with HLA*IMP:02 (25). The imputation accuracy was evaluated by kappa coefficient with the use of imputation and genotyping data for 589 patients with rheumatoid arthritis and 932 healthy subjects for HLA–DRB1, as previously described

(23), and for 114 patients with thyroid diseases for HLA–DQB1 (Terao: unpublished observations). We analyzed whether each allele of HLA–DRB1 and HLA–DQB1 with imputation accuracy >70% was associated with ANA positivity by logistic regression analysis with additive or dominant models.

**Evaluation of linkage disequilibrium (LD).** LD between SNPs and HLA–DRB1 alleles was obtained from previous studies (17,26,27). For LD calculation between HLA–DRB1 and HLA–DQB1 alleles, we used genotyping data of 1,000 unrelated healthy Japanese subjects (Terao: unpublished observations).

**Evaluation of effects of SLE-related SNPs.** A total of 7 SNPs that displayed associations with SLE beyond levels significant in GWAS in a Japanese population (15) and the 5 SNPs in the HLA locus that displayed independent associations with SLE in Europeans (28) were selected to assess their effects on ANA positivity. The associations between these SNPs and ANA positivity were analyzed based on imputation by MaCH (29), using 192 samples in the Nagahama Study genotyped by HumanHapOmni2.5-8, HumanHapOmni2.5s, and HumanExome arrays or using East Asian panels in the 1000 Genomes Project as a reference when they were not directly genotyped.

**Statistical analysis of genetic studies.** Statistical calculations were performed using Plink software version 1.07 (30) and R statistical software. For all genetic analyses including the GWAS, we set significance levels using the Bonferroni correction for multiple testing.

## RESULTS

A total of 9,575 subjects were analyzed for their ANA levels in the current study (Table 1). ANA titers in 45.2%, 12.5%, and 2.8% of the volunteers were ≥1:40, ≥1:80, and ≥1:160, respectively (see Supplementary Table 1, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38867/abstract). When we analyzed potential correlates of ANA positivity, female sex and old age had higher correlations with ANA positivity, as shown in previous studies (13,14) (corrected $P$ [$P_{corr}$] $< 1.0 \times 10^{-10}$) (see Supplementary Figure 2 and Supplementary Table 2, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38867/abstract).

When we focused on each staining pattern, 43.7%, 25.3%, 4.7%, 0.9%, and 2.0% of subjects had ANAs with speckled, homogeneous, nucleolar, discrete speckled, and cytoplasmic patterns, respectively, at titers of ≥1:40 (Table 1). The multiple logistic regression analyses revealed that the nucleolar pattern was not associated with age or sex (see Supplementary Table 2 and Supplementary Figure 3, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38867/abstract). Considering the higher frequency of the nucleolar pattern compared with that of the discrete speckled pattern, these results indicated that age and sex do not influence the positivity for each staining pattern in the same manner. Positivity for the speckled pattern was strongly correlated with positivity for all ANAs (see Supplementary Figure 4, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38867/abstract). Associations between other potential correlates and ANAs are shown in Supplementary Table 3 (available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38867/abstract). High CRP levels showed an association with ANA positivity ($P_{corr}$ = 0.0029). We did not find a significant association between obesity and ANA positivity.

Next, we performed a GWAS for ANA positivity. A total of 3,185 participants and 303,506 markers that had passed criteria of inclusion and quality control were used for logistic regression analysis, with age and sex as covariates. As a result, the Q–Q plot indicated an inflation factor of 1.02, suggesting that the current study was free from population stratification (Figure 1). A significant association of rs9405108 in the HLA locus was observed at a $P$ value of $8.9 \times 10^{-8}$. Conditioning rs9405108 to detect further associated markers in this region did not result in any markers showing significant associations ($P > 1.0 \times 10^{-4}$) (data not shown). No SNPs in non-HLA regions displayed suggestive associations ($P > 1.0 \times 10^{-5}$). We performed a replication study for rs9405108 using 3,963 participants (Table 1). For technical reasons, SNP rs2395185, which is almost in complete LD with rs9405108 (D′ = 1 and $r^2$ = 0.999), was genotyped instead of rs9405108. As a result, the

**Table 1.** Characteristics of the subjects in the current study*

| | All subjects (n = 9,575) | GWAS (n = 3,185)† | Replication study (n = 3,963)† |
|---|---|---|---|
| Women | 66.9 | 66.0 | 67.0 |
| Age, mean ± SD years | 53.3 ± 13.4 | 52.0 ± 14.1 | 53.7 ± 13.5 |
| ANA titer ≥1:40 | | | |
| All | 45.2 | 48.4 | 42.5 |
| Speckled | 43.7 | 46.8 | 41.1 |
| Homogeneous | 25.3 | 29.0 | 21.3 |
| Nucleolar | 4.7 | 5.1 | 4.2 |
| Discrete speckled | 0.9 | 0.8 | 0.9 |
| Cytoplasmic | 2.0 | 1.6 | 2.3 |

* Except where indicated otherwise, values are the percent. ANA = antinuclear antibody.
† In the genome-wide association study (GWAS), DNA samples were genome-scanned using Illumina HumanHap610, HumanHapOmni2.5-4, or HumanHapOmni2.5-8 arrays. Genotyping in the replication study was performed using a TaqMan Assay.