**Table S1.** Options for LAMPLINK. The bold letters indicate the required options.

(a) Options to detect statistically significant SNP combinations.

| Option            | Description                                                                 |
|-------------------|------------------------------------------------------------------------------|
| --lamp            | Find statistically significant combinations of SNPs                        |
| --file (or --bfile) <in_filename> | Input filename without the extension                                      |
| --model-dom (or --model-rec) | dom: dominant exclusive model, rec: recessive exclusive model         |
| --out <lamp_out_filename> | Output filename (default is “lamplink”)                                    |
| --fisher          | Use Fisher’s exact test as a statistical test (default is chi-squared test) |
| --ci <value>      | Output (1 - value) confidence intervals for odds ratios                    |
| --sglev <value>   | Set a statistical significance level (default is 0.05)                    |
| --upper <value>   | Set a maximum MAF value (default is 0.1)                                    |

(b) Options to eliminate redundant SNP combinations.

| Option             | Description                                                                 |
|-------------------|------------------------------------------------------------------------------|
| --lamp-ld-remove  | Eliminate redundant SNP combinations                                         |
| --file (or --bfile) <in_filename> | Input filename without the extension                                      |
| --comb <lamp_out_filename> | Filenames generated by the --lamp option without the extension         |
| --out <out_filename> | Output filename (default is “lamplink”)                                    |
| --lamp-r2 <value>  | Set the threshold for the $r^2$ (default is 0.8)                          |

**Table S2.** List of columns in output files of the --lamp option.

(a) Columns in <lamp_out_filename>.lamp

| Column name | Description                                                                 |
|-------------|------------------------------------------------------------------------------|
| COMBID      | Find combinatorial SNPs                                                      |
| Raw_P       | Input filename without the extension                                         |
| Adjusted_P  | dom: dominant exclusive model, rec: recessive exclusive model               |
| COMB        | Output filename (default is “lamplink”)                                     |

(b) Columns in <lamp_out_filename>.lamplink

| Column name | Description                                                                 |
|-------------|------------------------------------------------------------------------------|
|CHR          | Chromosome number                                                           |
|SNP          | SNP name                                                                     |
|A1           | Minor allele name                                                           |
|A2           | Major allele name                                                           |
|Test         | Genetic model selected (DOM: dominant exclusive model, REC: recessive exclusive model) |
|AFF          | The numbers of case individuals having A1 and A2 alleles, respectively       |
|UNAFF        | The numbers of control individuals having A1 and A2 alleles, respectively    |
|P            | Raw p-value                                                                  |
|OR           | Odds ratio                                                                   |
|COMBIDs      | A member of the combination COMBID (presence: 1, absence: 0)                 |
|CHSQ         | Chi-squared score (if chi-squared test is used)                              |
|DF           | Degree of freedom (if chi-squared test is used)                              |
|Lx           | Lower bound of x% confidence interval for odds ratios (if --ci is used)      |
|Ux           | Upper bound of x% confidence interval for odds ratios (if --ci is used)      |
Table S3. Combinations containing at least three SNPs accumulated in Japanese individuals with statistical significance detected by LAMPLINK. The numbers in the Bonferroni correction column indicate the maximum size of SNP combination investigated. NA means that the combination is not investigated. The combinations with * were not eliminated after Procedure 2.

| ID  | SNP            | Chr | Position (bp) | Gene      | LAMPLINK          | Adjusted p-value | Bonferroni correction |
|-----|----------------|-----|---------------|-----------|--------------------|------------------|----------------------|
| 1   | rs34902660     | 6   | 25,850,874    | SLC17A3   | 7.7695E-05         | 0.012638         | NA                   |
|     | rs2298091      | 6   | 26,158,211    | HIST1H2BD |                    |                  | 1                    |
|     | rs1150723      | 6   | 28,283,939    | PGBD1     |                    |                  | 1                    |
| 2   | rs2303080      | 5   | 7,878,311     | MTRR      | 0.012638           | 0.019122         | 1                    |
|     | rs2287779      | 5   | 7,889,103     | MTRR      |                    |                  | 1                    |
|     | rs2287780      | 5   | 7,889,191     | MTRR      |                    |                  | 1                    |
|     | rs16879334     | 5   | 7,891,393     | MTRR      |                    |                  | 1                    |
|     | rs3815990      | 12  | 121,253,285   | CAMKK2    |                    |                  | 1                    |
| 3*  | rs2472647      | 5   | 141,331,138   | PCDHGA1   |                    | 0.019122         | 1                    |
|     | rs36012859     | 6   | 132,734,332   | VNN3      |                    |                  | 1                    |
|     | rs17238245     | 15  | 61,951,918    | VPS13C    |                    |                  | 1                    |
| 4*  | rs79825658     | 3   | 57,508,536    | DNAH12    | 0.019122           | 0.019122         | 1                    |
|     | rs2472647      | 5   | 141,331,138   | PCDHGA1   |                    |                  | 1                    |
|     | rs1710011      | 7   | 34,827,570    | NPSR1     |                    |                  | 1                    |
| 5*  | rs79825658     | 3   | 57,508,536    | DNAH12    | 0.019122           | 0.019122         | 1                    |
|     | rs2242244      | 19  | 53,900,671    | PRKCG     |                    |                  | 1                    |
|     | rs114591455    | 19  | 55,796,187    | NLRP11    |                    |                  | 1                    |
| 6   | rs3772534      | 3   | 105,702,190   | CBLB      |                    | 0.019122         | 1                    |
|     | rs2298091      | 6   | 26,158,211    | HIST1H2BD |                    |                  | 1                    |
|     | rs34902660     | 6   | 25,850,874    | SLC17A3   |                    |                  | 1                    |
| 7   | rs34902660     | 6   | 25,850,874    | SLC17A3   |                    | 0.019122         | 1                    |
|     | rs2298091      | 6   | 26,158,211    | HIST1H2BD |                    |                  | 1                    |
|     | rs115152597    | 6   | 31,637,065    | PRRC2A    |                    |                  | 1                    |
| 8   | rs2298548      | 18  | 9,221,885     | ANKRDI2   | 0.023541           | 0.02354          | NA                   |
|     | rs17498752     | 18  | 9,254,787     | ANKRDI2   |                    |                  | NA                   |
|     | rs59633812     | 18  | 9,275,544     | ANKRDI2   |                    |                  | 1                    |
|     | rs17499116     | 18  | 9,275,565     | ANKRDI2   |                    |                  | 1                    |
|     | rs114591455    | 19  | 55,796,187    | NLRP11    |                    |                  | 1                    |
| 9   | rs2303080      | 5   | 7,878,311     | MTRR      | 0.023541           | 0.02354          | NA                   |
|     | rs2287780      | 5   | 7,889,191     | MTRR      |                    |                  | 1                    |
|     | rs16879334     | 5   | 7,891,393     | MTRR      |                    |                  | 1                    |
|     | rs115837224    | 6   | 30,625,708    | ATAT1     |                    |                  | 1                    |
| 10  | rs2303080      | 5   | 7,878,311     | MTRR      | 0.041435           | 0.041435         | NA                   |
|     | rs2287779      | 5   | 7,889,103     | MTRR      |                    |                  | 1                    |
|     | rs228780       | 5   | 7,889,191     | MTRR      |                    |                  | 1                    |
|     | rs16879334     | 5   | 7,891,393     | MTRR      |                    |                  | 1                    |
|     | rs41292755     | 13  | 38,689,594    | FREM2     |                    |                  | 1                    |
SUPPLEMENTARY TEXT

S1 DETAILS OF LAMPLINK

LAMPLINK performs a case-control analysis for GWAS data using Fisher’s exact test or chi-squared test, and enumerates statistically significant combinations associated with a given phenotype. LAMPLINK is implemented as a set of options added to PLINK (version 1.07), as shown in Supplementary Table S1. Because LAMPLINK can receive PLINK format files, it can be used by replacing the command name “plink” with “lamplink”. The result files are identical format to PLINK except for the LAMPLINK specific results about combinatorial effects shown in Supplementary Table S2. LAMPLINK runs with C++ and Python 2.7 on Linux.

Fig. 1(a) shows a typical analytical procedure with LAMPLINK to detect SNP combinations. The details are described in the following subsections.

Detection of statistically significant SNP combinations: The --lamp option with --model-dom (or --model-rec) can be used for enumerating statistically significant SNP combinations (Procedure 1 in Fig. 1(a)). The input and output filenames are specified with the --file (or --bfile for binary format) and --out options, respectively. When you set --model-dom, LAMPLINK detects statistically significant combinations of SNPs according to a dominant exclusive model, while --model-rec focuses on a recessive exclusive model. These two genetic models are defined in Section S3 in Supplementary Text. The options used with --lamp are listed in Supplementary Table S1(a).

LAMP can be slow if SNPs with high minor allele frequency (MAF) are included. The --lamp option firstly eliminates SNPs whose MAF is higher than the threshold, which is specified with the --upper option.

Then, the --lamp option performs two steps. In the first step, the GWAS dataset is converted into the format used in LAMP according to the user-specified genetic model. The genotype for each SNP has three patterns: dominant homozygote, heterozygote and recessive homozygote. To handle the data in LAMP, these genotypes are categorized into two classes: risk and non-risk classes. If the dominant exclusive model is selected, heterozygote and recessive homozygote correspond to a risk class. Otherwise, recessive homozygote is regarded as a risk class.

In the second step, statistically significant SNP combinations are enumerated with LAMP. LAMP first computes the adjusted significance level $\delta$ so as to control FWER below $\alpha$, and then enumerates SNP combinations whose p-values are below $\delta$. LAMP performs statistical tests for SNP combinations according to the user-specified genetic model. When we focus on effect of a single SNP, the statistical test between two groups of individuals having a genotype in the risk class and others is performed on each SNP.

When considering combinations of SNPs, statistical tests comparing individuals who have risk class genotypes for all SNPs in the combination and individuals who have at least one non-risk class genotype in the combination are performed. Further details are provided in Section S2. The significance level $\alpha$ can be changed using the --sglev option. For statistical assessment, chi-squared test and Fisher’s exact test can be used.

The input files are in the PED and MAP formats, which are identical to PLINK. Binary files including BED, BIM, and FAM formats are also acceptable by using the --bfile option.

LAMPLINK results including statistically significant combinations are exported to files: “<lamp_out_filename>.lamp” and “<lamp_out_filename>.lamp”. The former file reports all SNP combinations statistically significantly associated with the phenotype. The latter file reports detailed information about each SNP in a format similar to the result generated by PLINK for association analysis. All columns of the result files are listed in Supplementary Table S2.

LAMPLINK can compute confidence intervals with the --ci option, and the results are provided in “<lamp_out_filename>.lamp”. For a SNP combination with an odds ratio OR, $(1 - \beta)$ confidence intervals are defined as

$$\exp\{\log OR \pm \Phi^{-1}(1 - \beta/2)\sigma\},$$

where $\Phi$ and $\sigma$ represent the standard normal distribution and the standard deviation of the logarithm of the odds ratio, respectively, computed by

$$\sigma = \sqrt{\frac{1}{N_{case}^2} + \frac{1}{N_{ctrl}^2} + \frac{1}{N_{case} - n_{case}} + \frac{1}{N_{ctrl} - n_{ctrl}}}$$

where $N_{case}$, $N_{ctrl}$, $n_{case}$, and $n_{ctrl}$ denote the numbers of case and control individuals, and the numbers of case and control individuals having the risk class genotypes, respectively.

Elimination of redundant SNP combinations: While epistatic combinations should consist of SNPs across different linkage disequilibrium (LD) regions, combinations of SNPs in the same LD region may be frequently listed in the previous step because LAMPLINK does not use the LD information, which may prevent our understanding of SNP-phenotype associations. The --lamp-ld-remove option is useful to select informative combinations (Procedure 2 in Fig. 1(a)).

Using the --lamp-ld-remove option eliminates SNP combinations whose members have $r^2$ higher than the user-specified threshold, assuming that they are located in the same LD region. It calculates the $r^2$ for all pairs of SNPs whose chromosomes are identical in each combination. If all of the pairs have $r^2$ higher than the given threshold, the combination is removed from the results. The default threshold value of $r^2$ is 0.8, which can be changed by using the --lamp-r2 option.

All options with --lamp-ld-remove are listed in Supplementary Table S1(b). The file containing the results produced using the --lamp option should be specified with the --comb option.

Similar to Procedure 1, --file (or --bfile) can be used to set input files, and the results are exported to two files “<out_filename>.lamp” and “<out_filename>.lamplink”.

S2 LAMP

LAMP (Terada et al., 2013) is a generic multiple testing procedure to detect combinatorial features statistically significantly associated with a target value. When it is applied to case-control GWAS analysis, SNP combinations statistically significantly associated with a phenotype can be found. LAMP employs Tarone’s procedure (Tarone, 1990) to overcome the low sensitivity of Bonferroni correction and counts the exact number of testable combinations that have the possibility of causing a false positive result with an itemset mining algorithm (Uno et al., 2003).

Supplementary Table S4 shows an example dataset used in LAMP. Each row represents an individual. Each column indicates...
is calculated using Fisher’s exact test or chi-squared test. Four individuals (\(t_1, \ldots, t_8\)) and four genotypes (\(s_1, \ldots, s_4\)).

| Individual | SNP/marker | Case-control |
|------------|------------|--------------|
|            | \(s_1\) | \(s_2\) | \(s_3\) | \(s_4\) | \(c\) |
| \(t_1\)   | 0  | 1  | 1  | 0  | Case |
| \(t_2\)   | 1  | 1  | 1  | 0  | Case |
| \(t_3\)   | 0  | 1  | 1  | 1  | Control |
| \(t_4\)   | 1  | 0  | 0  | 0  | Control |
| \(t_5\)   | 0  | 1  | 1  | 1  | Control |
| \(t_6\)   | 1  | 1  | 0  | 0  | Control |
| \(t_7\)   | 1  | 0  | 0  | 1  | Control |
| \(t_8\)   | 0  | 1  | 0  | 0  | Control |

Table S5. Contingency table to test combination \(S = \{s_1, s_2\}\). The p-value is calculated using Fisher’s exact test or chi-squared test.

\[ c = \text{Control} \]
\[ \begin{array}{c|c|c|c}
   & have S & not have S \\
\hline
   c = \text{Case} & 1 & 4 & 5 \\
   & 1 & 2 & 3 \\
\hline
   & 2 & 6 & 8 \\
\end{array} \]

Because this statistical test is performed for each SNP combination, a multiple testing procedure must be applied. LAMPLINK uses LAMP as the multiple testing procedure, which uses the Bonferroni inequality to calculate the upper bound of FWER. LAMP adjusts the significance level in a similar way to Bonferroni correction, apart from the division of given tests into two categories: testable and untestable. Testable ones have the possibility of causing a false positive result, while untestable ones have no possibility of doing so. By using the fact that the existence of untestable ones does not increase FWER, LAMP only counts the number of testable ones in the Bonferroni factor, and safely reduces the correction factor compared with Bonferroni correction. Furthermore, an extremely large number of combinations should be considered in GWAS data, which requires an intractable amount of computational time. LAMP solves this problem by accelerating the calculation using an itemset mining technique.

### S3 GENETIC MODEL

While effects of single loci and interactions between two loci have been studied, few studies have considered interactions of multiple loci. When such high-order interactions are considered, various genetic models are generated. From these models, we introduced two models for LAMPLINK: the dominant exclusive model (`--model-dom`) and the recessive exclusive model (`--model-rec`).

Given a single SNP, we have dominant and recessive models. A genotype is categorized into a risk or non-risk class according to the selected model as follows. A biallelic SNP can be described by three genotypes: \(AA\), \(Aa\), and \(aa\), in which ‘\(A\)’ and ‘\(a\)’ represent major and minor alleles, respectively. We suppose that risk and non-risk classes are described as 1 and 0, respectively. In the dominant model, \(Aa\) and \(aa\) are assigned to 1 (risk class), and \(AA\) is 0 (non-risk class). In the recessive model, \(aa\) is considered as 1, and \(AA\) and \(Aa\) are 0.

Given a SNP combination, various genotype patterns can be considered. For example, when we have two SNPs, \(A\) and \(B\), since each has three genotypes, there are nine genotypes in total. Among

(a) Dominant exclusive model

| Genotype of B | Genotype of A |
|---------------|---------------|
| \(AA(0)\)     | 0 0 0         |
| \(Aa(1)\)     | 0 1 1         |
| \(aa(1)\)     | 0 1 1         |

(b) Recessive exclusive model

| Genotype of B | Genotype of A |
|---------------|---------------|
| \(AA(0)\)     | 0 0 0         |
| \(Aa(0)\)     | 0 0 0         |
| \(aa(1)\)     | 0 1 1         |
Table S7. Risk and non-risk classes of combinations of three SNPs A, B and C in the dominant and recessive exclusive models. A, B and C indicate major alleles, and a, b and c are minor alleles. 1 and 0 represent risk and non-risk classes, respectively. The usage is similar to the DNA codon table.

(a) Dominant exclusive model

| Genotype of B | Genotype of A | Genotype of C |
|---------------|---------------|---------------|
| BB (0)        | AA (0)        | CC (0)        |
| BB (0)        | Aa (1)        | Cc (1)        |
| BB (0)        | aa (1)        | cc (1)        |
| Bb (1)        | AA (0)        | CC (0)        |
| Bb (1)        | Aa (1)        | Cc (1)        |
| Bb (1)        | aa (1)        | cc (1)        |
| bb (1)        | AA (0)        | CC (0)        |
| bb (1)        | Aa (1)        | Cc (1)        |
| bb (1)        | aa (1)        | cc (1)        |

(b) Recessive exclusive model

| Genotype of B | Genotype of A | Genotype of C |
|---------------|---------------|---------------|
| BB (0)        | AA (0)        | CC (0)        |
| BB (0)        | Aa (1)        | Cc (1)        |
| BB (0)        | aa (1)        | cc (1)        |
| Bb (1)        | AA (0)        | CC (0)        |
| Bb (1)        | Aa (1)        | Cc (1)        |
| Bb (1)        | aa (1)        | cc (1)        |
| bb (1)        | AA (0)        | CC (0)        |
| bb (1)        | Aa (1)        | Cc (1)        |
| bb (1)        | aa (1)        | cc (1)        |

S4 APPLICATION TO 1000 GENOMES PROJECT

We applied LAMPLINK to human exome data provided by 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010), including 12,758 SNPs and 697 individuals from seven populations. We investigated time performance when running with the --lamp option using datasets of various sizes which were generated by randomly selecting subsets of human exome data provided by the 1000 Genomes Project. LAMPLINK was performed with the following settings: the dominant exclusive model and two-sided significance test. The statistically significant SNP combinations were enumerated under the following setting: Two-sided Fisher’s exact test was used, and the significance level was set to $\alpha = 0.05$. The dominant exclusive model was selected, and the maximum MAF threshold was set to 0.1. The experiments were run on a machine with an Intel Xeon E5-2680v2 processor at 2.6 GHz running Red Hat Enterprise Linux 6.4. The calculation time was 21.281 seconds.

S4.1 Statistically significant SNP combinations

Procedure 1 in Fig. 1(a) detected 106 statistically significant SNP combinations, including 49 SNP combinations that consisted of more than five SNPs. 10 SNP combinations consisted of three or more SNPs, all of which are listed in Supplementary Table S3. These combinations could not be detected when Bonferroni correction was used instead of LAMP. Our findings also show that no combinations with more than five SNPs were significant because all combinations were investigated in LAMPLINK.

Figs. S1(a) and (b) illustrate two statistically significant combinations associated with IDs 3 and 10 in Supplementary Table S3, respectively. ID 3 consisted of three SNPs: rs4276247, rs17238245, and rs36012859, which are located within the genes PCDHGA1, VPS13C, and VNN3, respectively. The fact that these SNPs are located within different genes on different chromosomes (Fig. S1(c)) shows the ability of LAMPLINK to detect novel interactions of SNPs. The p-value of this combination was smaller than the p-values of any single SNP, suggesting the existence of an epistatic effect among the three SNPs. Furthermore, none of the pairs of the three SNPs was statistically significant, showing the importance of investigating higher-order interactions.

ID 10 consisted of five SNPs. 15 individuals have a genotype in the risk class for all SNPs in this combination, and 11 of them are Japanese. Four SNPs (rs2287779, rs2287780, rs2303080, rs16879334) are located within the same gene MTRR, and the remaining one (rs41292755) is within the gene FREM2 on a different chromosome from MTRR (Fig. S1(c)). Because none of the single SNPs or pairs showed a statistically significant association with Japanese, as shown in the right diagram of Fig. S1(b), investigation of higher-order combinations is necessary to detect this combination. Since most of the SNPs are located within the same gene and may be in the same LD region, it might be considered that the combination is redundant. Procedure 2 with --lamp-ld-remove can detect the SNP combinations with high LD ($r^2$ value between rs2287779 and rs16879334 is almost 1) and eliminate them. In contrast to this, the SNP combination in Fig. S1(a) was kept even after Procedure 2.

These two results show that LAMPLINK has the ability to detect statistically significant SNP combinations from genome-wide case-control data. By replacing the phenotype with a disease, it might be possible to identify causal mutations of complex diseases.

S4.2 Time performance

We investigated time performance when running with the --lamp option using datasets of various sizes which were generated by randomly selecting subsets of human exome data provided by the 1000 Genomes Project. LAMPLINK was performed with the following settings: the dominant exclusive model and two-sided significance test. The statistically significant SNP combinations were enumerated under the following setting: Two-sided Fisher’s exact test was used, and the significance level was set to $\alpha = 0.05$. The dominant exclusive model was selected, and the maximum MAF threshold was set to 0.1. The experiments were run on a machine with an Intel Xeon E5-2680v2 processor at 2.6 GHz running Red Hat Enterprise Linux 6.4. The calculation time was 21.281 seconds.

The statistically significant SNP combinations were enumerated under the following setting: Two-sided Fisher’s exact test was used, and the significance level was set to $\alpha = 0.05$. The dominant exclusive model was selected, and the maximum MAF threshold was set to 0.1. The experiments were run on a machine with an Intel Xeon E5-2680v2 processor at 2.6 GHz running Red Hat Enterprise Linux 6.4. The calculation time was 21.281 seconds.
LAMPLINK has the ability to identify combinatorial effects of pairs of SNPs with a phenotype. The calculation of high-order interactions requires a lot of time when the dataset contains a large number of individuals. This increment is caused by LAMPLINK taking more time for a statistical test without multiple testing correction, whereas the --lamp option applies it to the results by computing the adjusted significance level to control FWER below $\alpha$ theoretically. Furthermore, LAMPLINK can detect the combinations of three or more SNPs, showing higher-order epistatic interactions.

LAMPLINK has higher sensitivity than Bonferroni correction in analyzing 1000 Genomes Project data. Supplementary Table S3 reports adjusted $p$-values by Bonferroni correction when we apply it to exhaustively identify combinatorial effects of up to three, four, and five SNPs. All of the results are 1, indicating that Bonferroni correction overlooks the combinations. The Bonferroni factor (the number of combinations tested) is $2.37 \times 10^{11}$ when the maximum combination size is limited to three, while the correction factor in LAMP is $2.61 \times 10^6$.

LAMPLINK is the first implementation that can detect statistically sound high-order interactions from tens of thousands of markers. Hence, LAMPLINK may contribute to the identification of combinatorial effects from multiple markers by re-analysis of existing GWAS datasets.

**S5 DISCUSSION**

We developed LAMPLINK, in which functions to discover the statistically sound high-order combinatorial effects of SNPs were added to widely used GWAS software, PLINK. Because it was implemented as a set of additional functions to PLINK, all existing functions in PLINK and existing files for it can be used. All of the additional results detected by LAMPLINK are exported to independent files with a format similar to that of PLINK files, and hence the results can be easily integrated with existing analysis procedure with PLINK.

LAMPLINK can be utilized to add statistical validity to the --epistatic option in PLINK, which analyzes the relationships of pairs of SNPs with a phenotype. This option returns $p$-values without multiple testing correction, whereas the --lamp option applies it to the results by computing the adjusted significance level to control FWER below $\alpha$ theoretically. Furthermore, LAMPLINK can detect the combinations of three or more SNPs, showing higher-order epistatic interactions.

LAMPLINK has higher sensitivity than Bonferroni correction in analyzing 1000 Genomes Project data. Supplementary Table S3 reports adjusted $p$-values by Bonferroni correction when we apply it to exhaustively identify combinatorial effects of up to three, four, and five SNPs. All of the results are 1, indicating that Bonferroni correction overlooks the combinations. The Bonferroni factor (the number of combinations tested) is $2.37 \times 10^{11}$ when the maximum combination size is limited to three, while the correction factor in LAMP is $2.61 \times 10^6$.

LAMPLINK is the first implementation that can detect statistically sound high-order interactions from tens of thousands of markers. Hence, LAMPLINK may contribute to the identification of combinatorial effects from multiple markers by re-analysis of existing GWAS datasets.

**S6 COMMAND LINES**

We show the command lines for identifying SNP combinations statistically significantly accumulated in the Japanese. The script used in the following commands is available at [http://a-terada.github.io/lamplink/](http://a-terada.github.io/lamplink/).
At first, the following eight files were downloaded from the 1000 Genomes Project FTP site.

1. CEU.genotypes.vcf
2. CHB.genotypes.vcf
3. CHD.genotypes.vcf
4. JPT.genotypes.vcf
5. LWK.genotypes.vcf
6. TSI.genotypes.vcf
7. YRI.genotypes.vcf
8. integrated_call_samples.20130502.ALL.ped

The first seven files were obtained from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/exon/snps/, and the last file was obtained from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

Because 1000 Genomes Project data were provided as VCF formatted files, they were converted to PLINK format using VCFtools (v0.1.12b) (Danecek et al., 2011), independently, with the following command.

```bash
$ for pop in CEU CHB CHD JPT LWK TSI YRI; do
  vcftools --vcf ${pop}.exon.201009.genotypes.vcf \
  --plink --out ${pop} done
```

This command generated the 14 files:

- {CEU, CHB, CHD, JPT, LWK, TSI, YRI}.ped
- {CEU, CHB, CHD, JPT, LWK, TSI, YRI}.map

The data were provided for each population. Therefore, we merged them using the --merge-list option of LAMPLINK, which is identical to the option in PLINK.

```bash
$ ./lamplink --file CEU --merge-list allfiles.txt \
  --recode --out exome1000genomes allpop
```

The two files were generated:

- exome1000genomes_allpop.ped
- exome1000genomes_allpop.map

With the following commands, PLINK formatted files that regard Japanese individuals as cases were generated.

```bash
$ python population2pheno.py \ exome1000genomes_allpop.ped JPT > JPTallpop.ped
$ mv exome1000genomes_allpop.map JPTallpop.map
```

The two files were generated:

- JPTallpop.ped
- JPTallpop.map

We run LAMPLINK to identify SNP combinations statistically significantly accumulated in Japanese.

```bash
$ ./lamplink --file JPTallpop --lamp --model-dom \  --sglev 0.05 --upper 0.1 --fisher \  --out JPTallpop_result
```

Two files were generated:

- JPTallpop_result.lamp
- JPTallpop_result.lamplink

The redundant combinations were eliminated with the following command, which requires JPTallpop_result. (lamp, lamplink).

```bash
$ ./lamplink --file JPTallpop --lamp-ld-remove --comb JPTallpop_result --out JPTallpop_result.ldremove
```

The two result files were generated:

- JPTallpop_result.ldremove.lamp
- JPTallpop_result.ldremove.lamplink

REFERENCES

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158.

Tarone, R. E. (1990). A modified Bonferroni method for discrete data. *Biometrics*, 46(2), 515–522.

Terada, A., Okada-Hatakeyama, M., Tsuda, K., and Sese, J. (2013). Statistical significance of combinatorial regulations. *Proc Natl Acad Sci U S A*, 110(32), 12996–13001.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061–1073.

Uno, T., Asai, T., Uchida, Y., and Arimura, H. (2003). LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets. In Proc of PAMI’03.