Journal of Human Genetics (2009) 54, 615–623

www.nature.com/jhg

REVIEW

Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse

Hirofumi Nakaoka[1,2] and Ituro Inoue[1]

Meta-analysis is a useful tool to increase the statistical power to detect gene–disease associations by combining results from the original and subsequent replication studies. Recently, consortium-based meta-analyses of several genome-wide association (GWA) data sets have discovered new susceptibility genes of common diseases. We reviewed the process and the methods of meta-analysis of genetic association studies. To conduct and report a transparent meta-analysis, the search strategy, the inclusion or exclusion criteria of studies and the statistical procedures should be fully described. Assessing consistency or heterogeneity of the associations across studies is an important aim of meta-analysis. Random effects model (REM) meta-analysis can incorporate between-study heterogeneity. We illustrated properties of test for and measures of between-study heterogeneity and the effect of between-study heterogeneity on conclusions of meta-analyses through simulations. Our simulation shows that the power of REM meta-analysis of GWA data sets (total case–control sample size: 5000–20 000) to detect a small genetic effect (odds ratio (OR)=1.4 under dominant model) decreases as between-study heterogeneity increases and then the mean of OR of the simulated meta-analyses passing the genome-wide significance threshold would be upwardly biased (*winner's curse* phenomenon). Addressing observed between-study heterogeneity may be challenging but give a new insight into the gene–disease association.
Journal of Human Genetics (2009) **54,** 615–623; doi:10.1038/jhg.2009.95; published online 23 October 2009

Keywords: genome-wide association study; heterogeneity; meta-analysis; winner's curse

INTRODUCTION

Population-based association studies provide a powerful approach to the identification of susceptibility genes underlying common diseases.[1,2] A very large amount of information about genetic variants in the human genome has been accumulated through the International Human Genome Sequencing Project and the International HapMap Project.[3–6] Combined with the establishment of high-throughput single-nucleotide polymorphism (SNP) typing systems, genome-wide association (GWA) studies have been widely applied.[7] Accordingly, gene–disease associations have been reported.

Replication studies were extensively implemented to establish the credibility of the initial positive findings. However, comprehensive reviews of the published literatures in the era of the candidate gene approach show that most of the initial positive associations were not reproduced in the subsequent replication studies.[8–13] These findings suggest that a large number of original findings were false-positive reports and another possibility is that most of the studies were underpowered to detect small genetic effect.[8,9] Furthermore, inconsistency or between-study heterogeneity of results of genetic associations can be observed regardless of whether the associations are true or not,[10,14] and it may be attributed to population stratification, genotyping errors, differences in the pattern of linkage disequilibrium (LD) structure and other factors.[15,16] In the era of GWA studies, this problem remains one of the most difficult issues of genetic association studies.[10,15,16] For example, the large-scale international study of Parkinson's disease failed to replicate 13 SNPs identified by the previous GWA study.[17]

In these circumstances, meta-analysis can be a useful tool to combine both statistically significant and nonsignificant results from individual studies on the same research question. In case–control study, the odds ratios (ORs) for individual studies are combined to calculate a summary OR. Meta-analysis improves the estimation of a summary OR and 95% confidence interval (CI) and increases the statistical power to detect gene–disease associations.[18] Therefore, conclusions from a meta-analysis are more robust than those from a single small study. In addition, meta-analysis is useful to investigate the consistency or heterogeneity of the associations across studies. Testing for and quantifying between-study heterogeneity is an

[1]Division of Molecular Life Science, School of Medicine, Tokai University, Isehara, Kanagawa, Japan and [2]The Japan Health Sciences Foundation, Chuo-ku, Tokyo, Japan
Correspondence: Professor I Inoue, Division of Molecular Life Science, Tokai University, School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan.
E-mail: ituro@is.icc.u-tokai.ac.jp
Received 6 August 2009; revised 4 September 2009; accepted 15 September 2009; published online 23 October 2009

important aim of meta-analyses to determine whether there are differences underlying the results of the study.[19,20] Addressing the observed between-study heterogeneity could generate a new insight into the gene–disease association.[20]

In this review, we begin with describing the process of meta-analysis of genetic association studies. The statistical backgrounds, methodological issues and sources of between-study heterogeneity of meta-analysis of genetic association studies are briefly reviewed. Finally, we present the results of our simulation study to illustrate the effect of between-study heterogeneity on conclusions of meta-analyses.

LITERATURE-BASED META-ANALYSIS

In a basic meta-analysis, data are retrospectively collected from published literatures to assess whether a gene–disease association of interest is true or not.[18] When planning a meta-analysis, it is important to define precise search strategy beforehand.[21] If relevant studies are excluded or inadequate studies are included, conclusions of the meta-analysis may be biased.[22] The literature search is conducted in databases such as PubMed and EMBASE. The HuGe Published Literature database (http://www.cdc.gov/genomics/hugenet/) is also useful, as it includes published literatures on genetic associations and other human genome epidemiology.[23] It is important to collect the largest possible number of studies; therefore, we should use appropriate key words. Once the search has been completed, bibliographies of retrieved articles should be examined for further relevant publications.

These processes make up the essential part of the methods section of a meta-analysis, because literature-based meta-analysis is subjected to bias caused by difficulty to identify and include all conducted and relevant studies,[13,24] and small difference in selected literatures may alter conclusions of meta-analyses on the same genetic association.[25] However, the essential features of the search strategy have not fully reported in most meta-analyses of genetic association studies.[26] In order to avoid such biases, it may be recommended to have two or more different researchers conducting the same search.[21] When conducting and reporting a literature-based meta-analysis, flowchart detailing the exclusion and inclusion criteria and the number of studies excluded and included at each step of the literature search is useful (Figure 1).

Meta-analysis of genetic association studies may be subjected to publication bias.[18,26] Publication bias tends to occur when small studies showing negative or nonsignificant results remain unpublished and may result in the overestimation of the genetic effect. If the presence of publication bias is suspected by statistical tests,[27,28] conclusions from the meta-analysis should be cautiously reported and the potential impact of the publication bias should be mentioned.[18]

The results obtained from the meta-analysis would be assessed by the following: (i) the size of the summary OR; (ii) the extent and possible cause of between-study heterogeneity; and (iii) the sufficiency and stability of the meta-analysis by using the cumulative and recursive cumulative meta-analysis approaches.[29–31] In the cumulative meta-analysis, studies are sorted chronologically and a summary OR is calculated when a new study is added.[29] As a result, we can present how the summary OR has shifted over time. The recursive cumulative meta-analysis is an extension of the cumulative meta-analysis, where the relative change in the summary OR by adding a new study is evaluated.[30,31]

CONSORTIUM-BASED META-ANALYSIS

Consortium-based meta-analysis is the meta-analysis of individual patient data through the collaboration of consortium of investigators.

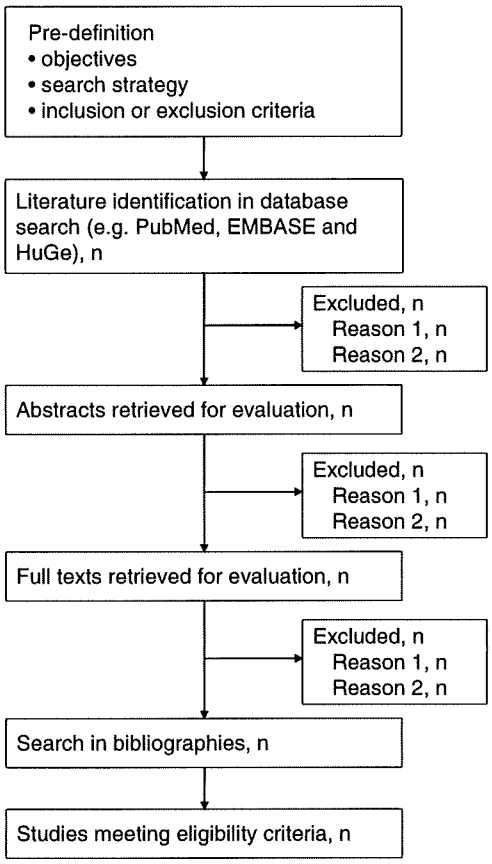


Figure 1 Flowchart detailing the exclusion and inclusion criteria and the number of studies excluded and included at each step of the literature search.

Consortium-based meta-analysis attains increased attention,[32–34] because integration of several GWA data sets has been designed and new susceptibility genes have been discovered.[35–39] Although meta-analysis of GWA studies can be implemented using reported ORs and 95% CIs or *P*-values from different GWA studies, it is preferable to reanalyze several GWA data sets with individual patient data.[35] In the latter case, one can use imputation techniques for missing data when SNPs have been genotyped in some platforms but not in others.[40] Barrett *et al.*[39] conducted a meta-analysis of three GWA data sets for Crohn's disease that used different genotyping platforms using imputation methods. The combined GWA data sets included 635 547 SNPs in 3230 cases and 4829 controls. They used the GWA data sets at the screening stage. The power of the meta-analysis was reported to be 0.74 to detect associations with per allele OR of 1.2 and with risk allele frequency of 0.2 at the significance level of $P=1.0\times10^{-5}$. The meta-analysis of the GWA data sets and additional replication data sets confirmed 11 previously reported loci and identified genome-wide significant signals for novel 21 loci.

GENETIC ASSOCIATION STUDY-SPECIFIC METHODOLOGICAL ISSUES

There are methodological issues relevant to meta-analysis of genetic association studies: (i) assessment of Hardy–Weinberg equilibrium (HWE) and (ii) definition of genetic models.

Deviation from HWE in control samples is the most commonly used test for genotyping error.[41] However, the test for HWE has relatively low statistical power to detect genotyping error.[42]

Furthermore, SNPs that are not in HWE can be used for inference about genetic model of disease susceptibility at the locus.[43] Although there is no consensus how meta-analyses should handle the studies that are not in HWE, three strategies have been applied: including all studies regardless of departure from HWE,[44] performing sensitivity analyses in order to evaluate whether the genetic effects are different between subgroups of studies classified according to test for HWE[26,45–47] and excluding studies showing statistically significant departure from HWE.[18] Reporting the extent of departure from HWE measured by such as α,[48] the inbreeding coefficient,[49] and the disequilibrium parameter[50] is also useful.[44]

In a genetic association study, subjects are classified into three exposure groups (*AA*, *Aa* and *aa*). Let *A* be the susceptibility allele, there are several methods of dichotomizing these exposure groups for conducting a meta-analysis:[26] by comparing allele frequency, by assuming a specific mode of inheritance (recessive, dominance, complete overdominant or codominant) and by performing multiple pairwise comparisons. All these methods, with exception of the method performing multiple pairwise comparisons, assume a particular genetic model. When performing multiple pairwise comparisons or testing multiple genetic models, results of all analyses undertaken should be reported. In order to choose most likely genetic model describing the genetic architecture underlying a disease of interest, Minelli *et al.*[51] presented a 'genetic model free' approach. Their procedure is based on the estimation of the ratio (λ) of the log OR of *Aa* versus *aa* compared with the log OR of *AA* versus *aa*. λ will be 0 under a recessive model, 0.5 under a codominant model and 1 under a dominant model.

ESTIMATION OF A SUMMARY OR AND TEST FOR AND MEASURE OF BETWEEN-STUDY HETEROGENEITY

The statistical methods of combining the results of different studies are described. We consider a meta-analysis of k separate genetic association studies to estimate the genetic effect (θ) for dichotomous disease outcome quantified by log OR. Let θ_i and $\hat{\theta}_i$ be the true and observed log OR for ith case–control study, respectively ($i=1, \ldots, k$). Let v_i denote the variance of $\hat{\theta}_i$, the weight for ith study is given by $w_i=1/v_i$ (that is, the inverse of the variance). OR for each study is given by $\mathrm{OR}_i=a_id_i/b_ic_i$. $\hat{\theta}_i = \ln(\mathrm{OR}_i)$. v_i is defined as $v_i=1/a_i+1/b_i+1/c_i+1/d_i$, where a_i and b_i correspond to numbers of affected individuals with and without the susceptible genotype, respectively, and c_i and d_i correspond to numbers of unaffected individuals with and without the susceptible genotype, respectively.

There are two commonly used procedures for combining $\hat{\theta}_i$s: 'fixed effects model' (FEM) and 'random effects model' (REM). FEM assumes that θ_is are homogeneous across studies (that is, $\theta_1=\theta_2=\ldots=\theta_k$) and all differences are due to chance. Inverse-variance, Mantel-Haenszel[52] and Peto's[53] methods are commonly used for FEM meta-analysis. Using the inverse-variance method for combining the results across studies, a summary log OR under FEM is calculated as a weighted average of the study estimates: $\hat{\theta}_{\mathrm{FEM}} = (\sum_{i=1}^{k} w_i\hat{\theta}_i)/(\sum_{i=1}^{k} w_i)$. The variance of $\hat{\theta}_{\mathrm{FEM}}$ is given by $v_{\mathrm{FEM}} = 1/\sum_{i=1}^{k} w_i$.

The assumption underlying FEM should be examined with the test for heterogeneity, Cochran's Q test.[54] Test statistics of Cochran's Q test is

$$Q = \sum_{i=1}^{k} w_i\left(\hat{\theta}_i - \hat{\theta}_{\mathrm{FEM}}\right)^2$$

Under the null hypothesis of homogeneity (that is, $\theta_1=\theta_2=\ldots=\theta_k$), this statistics approximately follows a χ^2 distribution with $k-1$ degrees of freedom. Cochran's Q test has relatively low statistical power to detect between-study heterogeneity, especially when the number of studies is small;[55] therefore, the test is usually preformed at the significance level of 0.1.[56]

REM assumes that the genetic effects may vary across studies because of genuine difference and/or differential biases. The estimate of the between-study variance (τ^2) is included into the weight as $w_i' = 1/(w_i^{-1}+\hat{\tau}^2)$. A summary log OR under REM are estimated as follows: $\hat{\theta}_{\mathrm{REM}} = (\sum_{i=1}^{k} w_i'\hat{\theta}_i)/(\sum_{i=1}^{k} w_i')$. The variance of $\hat{\theta}_{\mathrm{REM}}$ is approximated as $v_{\mathrm{REM}} = 1/\sum_{i=1}^{k} w_i'$.

In DerSimonian and Laird[57] REM meta-analysis, the τ^2 is estimated as follows:

$$\hat{\tau}^2_{\mathrm{DL}} = \frac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \left(\sum_{i=1}^{k} w_i^2/\sum_{i=1}^{k} w_i\right)}$$

When $Q<k-1$, $\hat{\tau}^2_{DL}$ takes negative value. In practice, $\max\{0, \hat{\tau}^2_{DL}\}$ is used. Therefore, the precision of a summary log OR with REM ($1/v_{\mathrm{REM}}$) can never exceed that with FEM ($1/v_{\mathrm{FEM}}$).

The 95% CI for $\hat{\theta}$ is given by $\hat{\theta} \pm 1.96\times\sqrt{v}$. Test statistic of test for the genetic effect is given by $Z = \hat{\theta}/\sqrt{v}$. Under the null hypothesis, Z follows a standard normal distribution.

Higgins and Thompson[58] proposed three criteria (H, R and I^2) for measure of heterogeneity, which have following desired characteristics: (i) dependence on the extent of heterogeneity, (ii) scale invariance (that is, comparison can be made across meta-analyses with different scales and different outcomes) and (iii) size invariance (that is, independence on the number of studies included). $H = \sqrt{Q/(k-1)}$ is the relative excess of Q to its degrees of freedom. Mittlbock and Heinzl[59] proposed $H^2_{\mathrm{M}} = \frac{Q-(k-1)}{k-1}$ as a modification of H. H^2_{M} is the proportion of between-study variance to within-study variance. In practice, $\max\{0, H^2_{\mathrm{M}}\}$ is used. H^2_{M} values over 1.0 indicate considerable heterogeneity.[59] $R = \sqrt{v_{\mathrm{REM}}/v_{\mathrm{FEM}}}$ is the ratio of the standard error of a summary effect with REM to the standard error with FEM. R represents the inflation of the CI for REM compared with FEM. H and R coincide when all studies have equal weight.[58] $I^2 = 100\times\frac{Q-(k-1)}{Q}$. I^2 can take negative value, but $\max\{0, I^2\}$ is used in practice. I^2 represents the proportion of between-study variance to the total variation in study estimates and ranges from 0 to 100%. I^2 is most widely used for measure of heterogeneity. I^2 values over 50% indicate large heterogeneity.[58,60] Potential drawback of I^2 is that CIs are very large, especially when the number of studies is small.[61]

If heterogeneity is present or suspected by the statistical test or measures, there are several commonly used approaches: (i) performing sensitivity analysis by excluding one or more studies showing outlier effect size, (ii) stratifying the studies into homogeneous subgroups such as racial groups and applying FEM for each subgroup and (iii) implementing REM when observed heterogeneity could not be addressed. Some researchers recommend that the use of REM is preferable compared with FEM, because both models give similar summary effects when there is no between-study heterogeneity, FEM gives narrower CI for summary effect compared with REM when between-study heterogeneity exists and a negative result of test for heterogeneity does not always indicate homogeneity when the number of studies is small.[25]

SOURCE OF HETEROGENEITY

A number of reasons have been advanced for heterogeneity in the genetic effects across the results of various studies.[8,13,14,47] False-positive results in the initial studies and false-negative results in small replication studies are implicated as the most likely reasons for non-replications.[8–10,13,14] Inconsistency and between-study heterogeneity may be caused because of biases or genuine differences in the genetic effects across populations. We review briefly in this article.

Biases

Differential biases due to population stratification, misclassification of clinical outcome, genotyping error and overestimation of genetic effect in the first study can be sources of between-study heterogeneity.

The presence of population stratification tends to spurious associations. It can be caused when there are undetected genetically different subgroups within a study population and disease prevalence differs among these subgroups.[11,62] The effect of population stratification on the results of genetic association studies is debatable.[62–66] According to systematic reviews of meta-analyses of genetic association studies, it is not so much frequent that difference in racial or ethnic groups could explain heterogeneity.[9,67]

Inadequate assignment of cases and controls may cause misclassification bias. Although there is a possibility that misclassification of cases and controls would weaken the gene–disease association, the results of misclassification bias may be modest unless the trait is common.[13,32]

Ioannidis *et al.*[10] conducted a systematic review of 36 meta-analyses including a total of 370 genetic association studies. Statistically significant between-study heterogeneity was observed in 14 meta-analyses. Restricting to meta-analyses with at least 15 studies, 7 of 9 meta-analyses showed significant heterogeneity. In 25 or 26 meta-analyses, the first study showed more predisposing or protective OR than subsequent replication studies. Using cumulative meta-analysis plots, the authors depicted the process that strong associations claimed in the first study were regressed toward null associations, as subsequent replication studies were accumulated over time. Similar findings were reported in Lohmueller *et al.*[9] Associations passing predetermined thresholds of statistical significance tend to overestimate the size of the genetic effect, especially when the sample size of the study is small and the threshold is stringent in multiple testing situations.[68–74] Such an upward bias is called as *winner's curse* phenomenon.[9,69]

Genuine differences

Differences in the pattern of LD structure over chromosomal regions of interest across populations are implicated as a cause of between-study heterogeneity in the genetic effects. Zondervan and Cardon[75] show that marker allelic OR can vary according to the extent of LD between marker and true disease allele in terms of D' and according to mismatch between disease allele frequency and marker allele frequency. This issue may be especially pronounced in the GWA settings because the SNPs that most efficiently surrogate the other SNPs in a genomic region with high LD (that is, tag SNPs) rather than putative functional SNPs have been used to increase genome coverage. When the extent of LD between tag SNP and true disease allele varies across studied populations, the observed ORs could vary across studies.

Many common diseases are implicated to have a complex etiology involving multiple genetic and environmental factors including their interactions. Gene–disease associations can be modified when the gene–gene or gene–environment interaction exists. If these interactions are not identified and controlled for, the gene–disease associations would be heterogeneous across populations according to distribution of a genetic variant or prevalence of a particular environmental exposure. It is needed to conduct a consortium-based meta-analysis of individual patient data in large scale to account for gene–gene or gene–environment interactions.[47]

SIMULATION STUDY

We conducted a simulation study to illustrate (i) the power of Cochran's Q test, (ii) the properties of measures of between-study heterogeneity (I^2 and H_M^2) and (iii) the type I error rate and the power of meta-analysis for detecting the gene–disease association in the presence of between-study heterogeneity.

We consider meta-analysis of k case–control association studies to estimate the overall genetic effect (θ; log OR) of disease outcome. The exposure status (*AA*, *Aa* and *aa*) of subjects included in each case–control study are ascertained in the sampling manner outlined below.[70] The values $y \in \{1, 0\}$ are labels encoding case (1) or control (0). Let A denote the susceptibility allele, we assume the dominant model and then the SNP genotype predictor value x was designed as 1=*AA* or *Aa*, 0=*aa*. Under the assumption of HWE, the frequency of x written as f_x is calculated based on the disease allele frequency f_A: $f_1 = 1-(1-f_A)^2$. The logistic regression model for ith study ($i = 1, 2, \ldots, k$) is produced as follows:

$$\log(\Pr(Y=1|x)/(1-\Pr(Y=1|x))) = \alpha_i + \theta_i x$$

where α_i is the intercept and θ_i is the log OR for ith study. θ_i is drawn from $N(\theta, \tau^2)$. τ^2 is the between-study variance. α_i can be calculated by using the equation for the prevalence of the disease $\pi = \sum_x \frac{\exp(\alpha_i+\theta_i x)}{1+\exp(\alpha_i+\theta_i x)} \times f_x$. The genotypes of case and control subjects are generated based on the conditional probabilities of x given by y as follows:

$$\Pr(X=x|Y=1) = \frac{f_x}{\pi} \times \frac{\exp(\alpha_i+\theta_i x)}{1+\exp(\alpha_i+\theta_i x)},$$

$$\Pr(X=x|Y=0) = \frac{f_x}{1-\pi} \times \frac{1}{1+\exp(\alpha_i+\theta_i x)}$$

For each study, the genotypes of case–control samples were generated and then the OR and its variance were calculated. Then, the ORs for k studies were combined by FEM and REM meta-analyses. Cochran's Q test was conducted and the I^2 and H_M^2 were measured.

We considered simple five simulation scenarios of meta-analyses. The description of simulation scenarios is shown in Table 1. The scenarios I, II and III were designed to be same in sample size within each study but different in the number of included studies. In scenarios III, IV and V, numbers of studies were different but total number of case–control samples included in meta-analysis was fixed at 20 000. The pairs of scenarios I and V or II and IV were designed to have the same number of studies but differ in sample size within each study.

We examined 126 parameter combinations for each scenario. The between-study variance (τ^2) varied from 0.0 to 0.02 with increments of 0.001. The true summary OR ($\exp(\theta)$) was set to be 1.0, 1.4 or 2.0. The disease allele frequency f_A was assigned to be 0.1 or 0.3. The disease prevalence π was fixed at 0.01. The values of τ^2 were based on the literature values reported by Moonesinghe *et al.*[76] for the confirmed 10 loci in a meta-analysis of three GWA studies of type 2 diabetes.[77] Therefore, our simulation would reflect the possible range of between-study variance. For each scenario and parameter combination, 100 000 simulations were carried out.

Table 1 Description of five simulation scenarios of meta-analysis

Scenario	k	$n_{case}/n_{control}$
I	5	500/500
II	10	500/500
III	20	500/500
IV	10	1000/1000
V	5	2000/2000

k denotes the number of included studies and n_{case} and $n_{control}$ are the number of cases and controls within each study, respectively.

The empirical power of Cochran's Q test was evaluated by the proportion of the simulation runs crossing the significance level of 0.1 when $\tau^2 > 0.0$. The top row of Figure 2 shows the powers of Cochran's Q test obtained with five scenarios as the function of τ^2 when the overall OR=1.0 and f_A=0.1 or 0.3. For each scenario, the power increased as τ^2 increased. Comparing among scenarios I, II and III, the power increased as the number of studies increased. When total number of case–control samples was fixed (that is, comparing among scenarios III, IV and V), the powers were similar but scenarios with smaller number of studies showed higher power

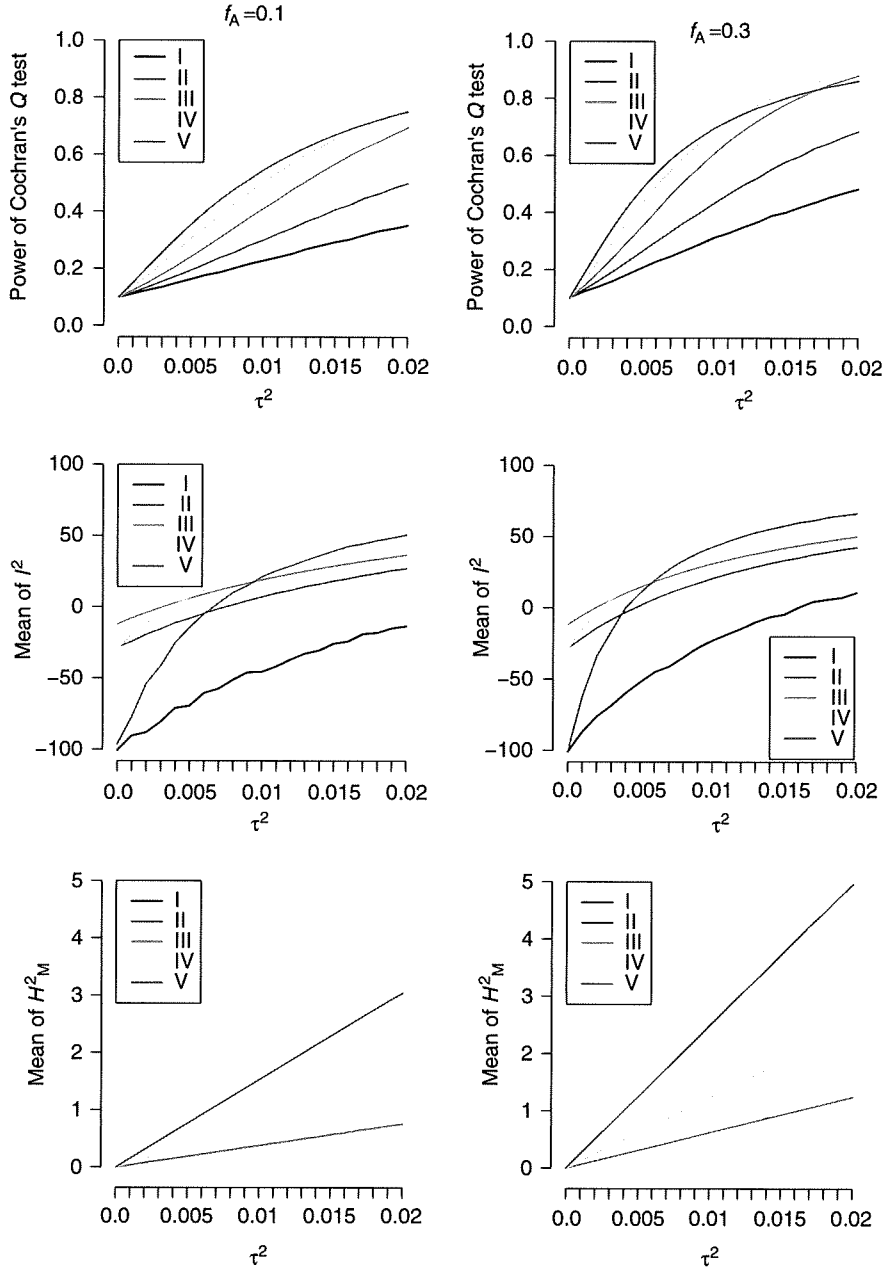


Figure 2 Behaviors of test for and measures of between-study heterogeneity for five simulation scenarios as the function of τ^2, the disease allele frequency f_A=0.1 or 0.3, and the overall odds ratio (OR)=1.0. The top row shows the power of the Cochran's Q test at the significance level of 0.1. The middle and bottom rows show the means of I^2 and H^2_M, respectively. The lines of H^2_M for scenarios I, II and III are overlapping. The description of each simulation scenario is in Table 1.

when τ^2 was small. When numbers of studies were identical (that is, two pairwise comparisons of scenarios I versus V or II versus IV), meta-analyses with larger sample size showed higher power for the same τ^2. The powers obtained with f_A=0.3 were higher than those with f_A=0.1. For most of our parameter settings, the powers of Cochran's Q test did not reach at 0.8, although the significance level was set to be 0.10.

The means of 100 000 simulated values for the measures of heterogeneity (I^2 and H_M^2) are shown as the function of τ^2 when the overall OR=1.0 and f_A=0.1 or 0.3 (the middle and bottom rows of Figure 2). In practice, max{0, I^2} and max{0, H_M^2} are used to restrict the ranges of these measures as positive. As the simulation study of Mittlbock and Heinzl,[59] unrestricted values of I^2 and H_M^2 were used to obtain unbiased distributions for these measures in this study. These two measures presented monotonic increases as τ^2 increased. I^2 and H_M^2 increased as the sample size per study increased (scenarios I versus V or II versus IV). The two measures obtained with f_A=0.3 were higher than those with f_A=0.1. These results indicate that I^2 and H_M^2 increased as within-study variance, $k/(\sum_{i=1}^{k} w_i)$, decreased. Comparing scenarios I, II and III shows the important difference between I^2 and H_M^2: whereas I^2 increased as the number of studies increased, H_M^2 did not change (the lines of H_M^2 for scenarios I, II and III are overlapping in the bottom rows of Figure 2). This suggests that H_M^2 may be a good indicator of comparing the extent of between-study heterogeneity across meta-analyses. Similar results and further discussion are provided by Mittlbock and Heinzl.[59] The 95% intervals of simulated I^2 and H_M^2 were large, especially when the number of studies is small (Supplementary Figure S1).

The type I error rate in meta-analysis was assessed as the proportion of the simulation runs showing significant summary OR at the significance level of 0.05 when the null hypothesis was true (that is, the true overall OR=1.0). Figure 3 shows the type I error rates of five scenarios when f_A=0.1 or 0.3. When there was no between-study variance (τ^2=0.0), the type I error rates under FEM were well controlled at 0.05, but REM showed slightly conservative results (the type I error rate$\approx$0.04). As τ^2 increased, the type I error rates under FEM rapidly inflated, but those under REM slightly increased. The type I error rates under both models for the same τ^2 increased when sample size per study was large or f_A=0.3. We should note that the use of FEM could increase the type I error rate even to the extent that the between-study heterogeneity could not be fully identified by Cochran's Q test and two measures I^2 and H_M^2. For example, in case of τ^2=0.005 and f_A=0.3, the type I error rate under FEM for five scenarios were 8.5–19.2% (Figure 3). For the parameter setting, the powers of Cochran's Q-test were 20.6–48.3%, the means of I^2 were −51.9 to 20.8% and the means of H_M^2 were 0.31–1.25 (Figure 2).

The power of detecting a gene–disease association was evaluated as the proportion of simulation runs reaching the significance level of 5.7×10^{-7}, assuming the consortium-based meta-analysis of GWA data sets. As shown in Figure 3, applying FEM meta-analysis to heterogeneous genetic associations could lead to false-positive findings; therefore, we considered only REM when assessing the power of

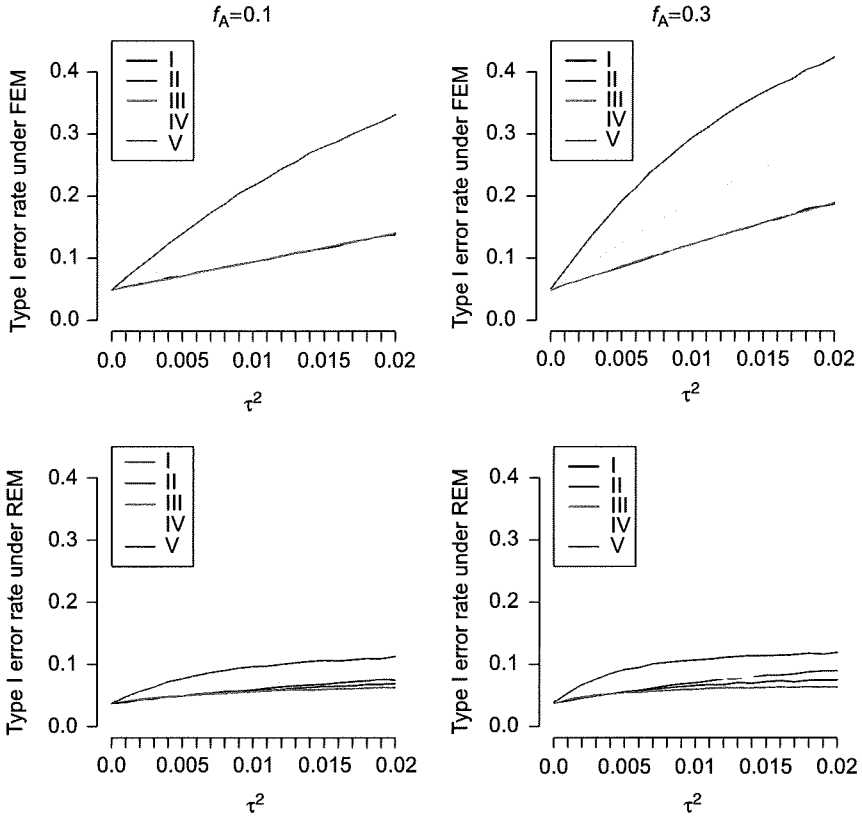


Figure 3 The type I error rate in fixed effects model (FEM) and random effects model (REM) meta-analyses at the significance level of 0.05 for five scenarios as the function of τ^2 and the disease allele frequency f_A=0.1 or 0.3. The top and bottom rows show the type I error rates when applying FEM and REM, respectively. The lines of the type I error rate under FEM for scenarios I, II and III are overlapping. The description of each simulation scenario is in Table 1.

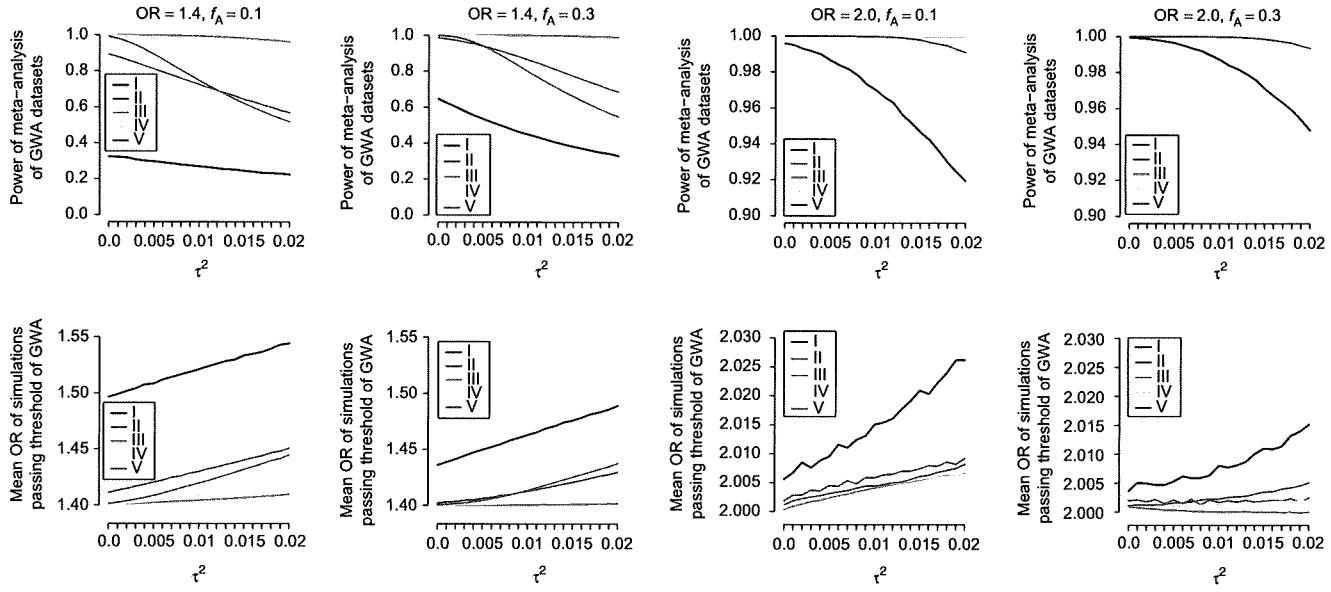


Figure 4 Simulations for the powers in random effects model (REM) meta-analyses of detecting a gene–disease association at the significance level of 5.7×10^{-7} (the top row) and the mean odds ratio (OR) of the simulations passing the threshold (the bottom row) as the function of τ^2, the disease allele frequency f_A=0.1 or 0.3, and the overall OR=1.4 or 2.0. When the overall OR=2.0, the lines of the powers for scenarios II, III and IV are overlapping. The description of each simulation scenario is in Table 1.

meta-analysis. The top row of Figure 4 shows the result, assuming the dominant model and f_A=0.1 or 0.3. When the true overall OR=1.4, the power for each scenario gradually decreased as τ^2 increased. Comparing scenarios III, IV and V, the decreases in the power for the same τ^2 were larger in the scenarios with large sample size per study. While the values of v_{FEM} for scenarios III, IV and V were not different, the values of v_{REM} for scenarios III, IV and V varied when between-study heterogeneity was present. For the same τ^2 (>0), the following inequality was true: v_{REM} for scenario V $>v_{REM}$ for scenario IV $>v_{REM}$ for scenario III. When $\theta\neq0$, the mean of the distribution of the Z-test under REM is $\lambda=\theta/\sqrt{v_{REM}}$. The power of detecting gene-disease association of effect size of θ is[78]

$$\text{Power}=1-\Phi(C_{\alpha/2}-\lambda)+\Phi(-C_{\alpha/2}-\lambda)$$

where Φ is the cumulative distribution function of the standard normal and $C_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution. Along with the inequality described above, the decrease in the power for the same τ^2 is larger in the scenarios with large sample size per study when the total sample sizes are equal across scenarios. When the overall OR was set to be 2.0, the powers did not so much decrease in the simulated range of τ^2. Furthermore, we calculated the mean OR of the simulations passing the genome-wide significance threshold (P-value$<5.7\times10^{-7}$). The estimates of mean OR were upwardly biased, especially in scenarios whose powers of detecting gene–disease associations were low (the bottom row of Figure 4). On the other hand, if the meta-analyses were sufficiently powered (for example, the true overall OR=2.0), upward biases were not so pronounced in the simulated range of τ^2.

Our simulation suggests that the power of meta-analysis of GWA data sets to detect small genetic effect would decrease due to between-study heterogeneity ($\tau^2\sim0.02$). As a result, the discovered gene–disease association could have inflated effect (*winner's curse* phenomenon). Such a *winner's curse* phenomenon can be seen even to the extent that the between-study heterogeneity could not be fully identified. Similar results were obtained when different genetic models (that is, recessive and additive in log-odds scale models) were examined (data not shown).

CONCLUSION

We reviewed the process and the methods of meta-analysis of genetic association studies. To conduct and report a transparent meta-analysis, the search strategy, the inclusion or exclusion criteria of studies and the statistical procedures should be fully described. Assessment of HWE and determination of genetic model are methodological issues relevant to meta-analysis of genetic association studies.

In genetic association studies of common disease, effect size of consistently replicated gene–disease associations were found to be small (OR=1.2–1.5);[15] therefore, meta-analysis of GWA data sets is the most important approach to increase the power to detect such gene–disease associations.[35]

Our simulation shows that the power of REM meta-analysis of GWA data sets to detect a small genetic effect could decrease due to between-study heterogeneity and then the mean OR of the simulated meta-analyses that passing the genome-wide significance threshold would be upwardly biased. Recently, Moonesinghe *et al.*[76] show that the required sample size in meta-analysis to detect an overall association with adequate power at a significant level increases as between-study heterogeneity increases and when the between-study heterogeneity exceeds a threshold, meta-analysis cannot reach the power regardless of how large included studies are. At the same time, empirical evaluation of published meta-analyses[61] and our simulation study show the uncertainty of estimated between-study heterogeneity is large unless many studies are combined.

These findings suggest that when a meta-analysis of GWA data sets shows association signals reaching genome-wide significance with small between-study heterogeneity, the result should be cautiously reported and further replication studies by institutions other than GWA teams are required.[35] Moreover, when a large number of data sets are available, challenges to explain and reduce the observed

between-study heterogeneity may become important.[74,76] The knowledge about the potential causes of between-study heterogeneity may help. Such post-GWA research will enable us to map the causative variant finely[79] or to detect polymorphisms associated with clinically important subtypes of diseases.[80]

1 Lander, E. S. The new genomics: global views of biology. *Science* **274,** 536–539 (1996).
2 Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).
3 The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
4 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
5 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
6 Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
7 Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–1605 (2008).
8 Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4,** 45–61 (2002).
9 Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33,** 177–182 (2003).
10 Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nat. Genet.* **29,** 306–309 (2001).
11 Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nat. Rev. Genet.* **2,** 91–99 (2001).
12 Freely associating. *Nat. Genet.* **22,** 1–2 (1999).
13 Colhoun, H. M., McKeigue, P. M. & Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet* **361,** 865–872 (2003).
14 Ioannidis, J. P. Non-replication and inconsistency in the genome-wide association setting. *Hum. Hered.* **64,** 203–213 (2007).
15 Khoury, M. J., Little, J., Gwinn, M. & Ioannidis, J. P. On the synthesis and interpretation of consistent but weak gene–disease associations in the era of genome-wide association studies. *Int. J. Epidemiol.* **36,** 439–445 (2007).
16 NCI-NHGRI Working Group on Replication in Association StudiesChanock, S. J. Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J. *et al.* Replicating genotype–phenotype associations. *Nature* **447,** 655–660 (2007).
17 Elbaz, A., Nelson, L. M., Payami, H., Ioannidis, J. P., Fiske, B. K., Annesi, G. *et al.* Lack of replication of thirteen single-nucleotide polymorphisms implicated in Parkinson's disease: a large-scale international study. *Lancet Neurol.* **5,** 917–923 (2006).
18 Munafo, M. R. & Flint, J. Meta-analysis of genetic association studies. *Trends Genet.* **20,** 439–444 (2004).
19 Lau, J., Ioannidis, J. P. & Schmid, C. H. Summing up evidence: one answer is not always enough. *Lancet* **351,** 123–127 (1998).
20 Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE* **2,** e841 (2007).
21 Sagoo, G. S., Little, J. & Higgins, J. P. Systematic reviews of genetic association studies. Human Genome Epidemiology Network. *PLoS Med.* **6,** e28 (2009).
22 Egger, M. & Smith, G. D. Bias in location and selection of studies. *BMJ* **316,** 61–66 (1998).
23 Lin, B. K., Clyne, M., Walsh, M., Gomez, O., Yu, W., Gwinn, M. *et al.* Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database. *Am. J. Epidemiol.* **164,** 1–4 (2006).
24 Tang, J. L. Selection bias in meta-analyses of gene–disease associations. *PLoS Med.* **2,** e409 (2005).
25 Kavvoura, F. K. & Ioannidis, J. P. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum. Genet.* **123,** 1–14 (2008).
26 Attia, J., Thakkinstian, A. & D'Este, C. Meta-analyses of molecular association studies: methodologic lessons for genetic epidemiology. *J. Clin. Epidemiol.* **56,** 297–303 (2003).
27 Begg, C. B. & Mazumdar, M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50,** 1088–1101 (1994).
28 Egger, M., Davey Smith, G., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315,** 629–634 (1997).
29 Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F. & Chalmers, T. C. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N. Engl. J. Med.* **327,** 248–254 (1992).
30 Ioannidis, J. P., Contopoulos-Ioannidis, D. G. & Lau, J. Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. *J. Clin. Epidemiol.* **52,** 281–291 (1999).
31 Ioannidis, J. & Lau, J. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. *Proc. Natl Acad. Sci. USA* **98,** 831–836 (2001).
32 McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9,** 356–369 (2008).
33 Seminara, D., Khoury, M. J., O'Brien, T. R., Manolio, T., Gwinn, M. L., Little, J. *et al.* The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology* **18,** 1–8 (2007).
34 Ioannidis, J. P., Bernstein, J., Boffetta, P., Danesh, J., Dolan, S., Hartge, P. *et al.* A network of investigator networks in human genome epidemiology. *Am. J. Epidemiol.* **162,** 302–304 (2005).
35 Zeggini, E. & Ioannidis, J. P. Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10,** 191–201 (2009).
36 Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40,** 638–645 (2008).
37 Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* **447,** 661–678 (2007).
38 Evangelou, E., Maraganore, D. M. & Ioannidis, J. P. Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease. *PLoS ONE* **2,** e196 (2007).
39 Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40,** 955–962 (2008).
40 Browning, S. R. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum. Genet.* **124,** 439–450 (2008).
41 Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A. *et al.* Detection of genotyping errors by Hardy–Weinberg equilibrium testing. *Eur. J. Hum. Genet.* **12,** 395–399 (2004).
42 Cox, D. G. & Kraft, P. Quantification of the power of Hardy–Weinberg equilibrium testing to detect genotyping error. *Hum. Hered.* **61,** 10–14 (2006).
43 Wittke-Thompson, J. K., Pluzhnikov, A. & Cox, N. J. Rational inferences about departures from Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76,** 967–986 (2005).
44 Minelli, C., Thompson, J. R., Abrams, K. R., Thakkinstian, A. & Attia, J. How should we use information about HWE in the meta-analyses of genetic association studies? *Int. J. Epidemiol.* **37,** 136–146 (2008).
45 Zintzaras, E. & Lau, J. Synthesis of genetic association studies for pertinent gene–disease associations requires appropriate methodological and statistical approaches. *J. Clin. Epidemiol.* **61,** 634–645 (2008).
46 Thakkinstian, A., McElduff, P., D'Este, C., Duffy, D. & Attia, J. A method for meta-analysis of molecular association studies. *Stat. Med.* **24,** 1291–1306 (2005).
47 Salanti, G., Sanderson, S. & Higgins, J. P. Obstacles and opportunities in meta-analysis of genetic association studies. *Genet. Med.* **7,** 13–20 (2005).
48 Lindley, D. Statistical inference concerning Hardy–Weinberg equilibrium. *Bayesian Stat.* **3,** 307–326 (1988).
49 Weir, B. S. in *Genetic Data Analysis II: Methods for Discrete Population Genetic Data* (Sinauer Associates, Sunderland, 1996).
50 Hernandez, J. L. & Weir, B. S. A disequilibrium coefficient approach to Hardy–Weinberg testing. *Biometrics* **45,** 53–70 (1989).
51 Minelli, C., Thompson, J. R., Abrams, K. R., Thakkinstian, A. & Attia, J. The choice of a genetic model in the meta-analysis of molecular association studies. *Int. J. Epidemiol.* **34,** 1319–1328 (2005).
52 Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22,** 719–748 (1959).
53 Yusuf, S., Peto, R., Lewis, J., Collins, R. & Sleight, P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog. Cardiovasc. Dis.* **27,** 335–371 (1985).
54 Cochran, W. G. The combination of estimates from different experiments. *Biometrics* **10,** 101–129 (1954).
55 Hardy, R. J. & Thompson, S. G. Detecting and describing heterogeneity in meta-analysis. *Stat. Med.* **17,** 841–856 (1998).
56 Petitti, D. B. Approaches to heterogeneity in meta-analysis. *Stat. Med.* **20,** 3625–3633 (2001).
57 DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control Clin. Trials* **7,** 177–188 (1986).
58 Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21,** 1539–1558 (2002).
59 Mittlbock, M. & Heinzl, H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat. Med.* **25,** 4321–4333 (2006).
60 Higgins, J. P., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327,** 557–560 (2003).
61 Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ* **335,** 914–916 (2007).
62 Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361,** 598–604 (2003).
63 Wacholder, S., Rothman, N. & Caporaso, N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J. Natl. Cancer Inst.* **92,** 1151–1158 (2000).
64 Wacholder, S., Rothman, N. & Caporaso, N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol. Biomarkers Prev.* **11,** 513–520 (2002).
65 Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36,** 512–517 (2004).

66 Thomas, D. C. & Witte, J. S. Point: population stratification: a problem for case–control studies of candidate-gene associations? *Cancer Epidemiol. Biomarkers Prev.* **11,** 505–512 (2002).
67 Ioannidis, J. P., Ntzani, E. E. & Trikalinos, T. A. 'Racial' differences in genetic effects for complex diseases. *Nat. Genet.* **36,** 1312–1318 (2004).
68 Garner, C. Upward bias in odds ratio estimates from genome-wide association studies. *Genet. Epidemiol.* **31,** 288–295 (2007).
69 Zollner, S. & Pritchard, J. K. Overcoming the winner's curse: estimating penetrance parameters from case–control data. *Am. J. Hum. Genet.* **80,** 605–615 (2007).
70 Ghosh, A., Zou, F. & Wright, F. A. Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *Am. J. Hum. Genet.* **82,** 1064–1074 (2008).
71 Ioannidis, J. P. Why most discovered true associations are inflated. *Epidemiology* **19,** 640–648 (2008).
72 Kraft, P. Curses—winner's and otherwise—in genetic epidemiology. *Epidemiology* **19,** 649–651 (2008); discussion 657–658.
73 Yu, K., Chatterjee, N., Wheeler, W., Li, Q., Wang, S., Rothman, N. *et al.* Flexible design for following up positive findings. *Am. J. Hum. Genet.* **81,** 540–551 (2007).
74 Ioannidis, J. P., Thomas, G. & Daly, M. J. Validating, augmenting and refining genome-wide association signals. *Nat. Rev. Genet.* **10,** 318–329 (2009).
75 Zondervan, K. T. & Cardon, L. R. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5,** 89–100 (2004).
76 Moonesinghe, R., Khoury, M. J., Liu, T. & Ioannidis, J. P. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc. Natl Acad. Sci. USA* **105,** 617–622 (2008).
77 Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316,** 1341–1345 (2007).
78 Hedges, L. V. & Pigott, T. D. The power of statistical tests in meta-analysis. *Psychol. Methods* **6,** 203–217 (2001).
79 Helgason, A., Palsson, S., Thorleifsson, G., Grant, S. F., Emilsson, V., Gunnarsdottir, S. *et al.* Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat. Genet.* **39,** 218–225 (2007).
80 Garcia-Closas, M., Hall, P., Nevanlinna, H., Pooley, K., Morrison, J., Richesson, D. A. *et al.* Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet.* **4,** e1000054 (2008).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

hybrid sterility involves both the unusual abundance and retention of OdsHmau protein in the *D. simulans* testis, as well as an unusual localization and possibly decondensation of the *D. simulans* Y chromosome. We conclude on the basis of these data that hybrid male sterility is caused by a gain-of-function interaction between OdsHmau and some component of the *D. simulans* Y chromosome heterochromatin, with this protein-DNA interaction representing the Dobzhansky-Muller incompatibility.

OdsH shares similarities with the hybrid sterility genes *Prdm9* (or *Meisetz*) in mouse (*23*) and *Overdrive* (*Ovd*) in *Drosophila* (*24*), all of which encode proteins with putative DNA-binding domains. Satellite DNAs have also been implicated in hybrid inviability, including a pericentric satellite locus (*Zhr*) (*25*, *26*) and a gene encoding a heterochromatin-binding protein (*Lhr*) (*27*). Thus, rapidly evolving repetitive DNA elements driven by genetic conflict may represent a major evolutionary force driving sequence divergence of speciation genes that would ultimately result in hybrid incompatibilities (*13*, *14*, *28*).

References and Notes

1. E. Mayr, *Systematics and the Origin of Species from the Viewpoint of a Zoologist* (Columbia Univ. Press, New York, 1942).
2. J. A. Coyne, H. A. Orr, *Speciation* (Sinauer Associates, Sunderland, MA, 2004).
3. C. C. Laurie, *Genetics* **147**, 937 (1997).
4. R. M. Kliman *et al.*, *Genetics* **156**, 1913 (2000).
5. C. T. Ting, S. C. Tsaur, M. L. Wu, C. I. Wu, *Science* **282**, 1501 (1998).
6. S. Sun, C. T. Ting, C. I. Wu, *Science* **305**, 81 (2004).
7. D. E. Perez, C. I. Wu, *Genetics* **140**, 201 (1995).
8. D. E. Perez, C. I. Wu, N. A. Johnson, M. L. Wu, *Genetics* **134**, 261 (1993).
9. S. D. Hueber, I. Lohmann, *Bioessays* **30**, 965 (2008).
10. C. T. Ting *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 12232 (2004).
11. K. Tabuchi, S. Yoshikawa, Y. Yuasa, K. Sawamoto, H. Okano, *Neurosci. Lett.* **257**, 49 (1998).
12. M. Nei, J. Zhang, *Science* **282**, 1428 (1998).
13. S. Henikoff, K. Ahmad, H. S. Malik, *Science* **293**, 1098 (2001).
14. S. Henikoff, H. S. Malik, *Nature* **417**, 227 (2002).
15. L. Fishman, A. Saunders, *Science* **322**, 1559 (2008).
16. A. Daniel, *Am. J. Med. Genet.* **111**, 450 (2002).
17. N. Aulner *et al.*, *Mol. Cell. Biol.* **22**, 1218 (2002).
18. M. Ashburner, K. G. Golic, R. S. Hawley, *Drosophila: A Laboratory Handbook* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 2005).
19. G. Cenci, S. Bonaccorsi, C. Pisano, F. Verni, M. Gatti, *J. Cell Sci.* **107**, 3521 (1994).
20. B. D. McKee, *Curr. Top. Dev. Biol.* **37**, 77 (1998).
21. J. E. Tomkiel, *Genetica* **109**, 95 (2000).
22. J. Forejt, *Trends Genet.* **12**, 412 (1996).
23. O. Mihola, Z. Trachtulec, C. Vlcek, J. C. Schimenti, J. Forejt, *Science* **323**, 373 (2009).
24. N. Phadnis, H. A. Orr, *Science* **323**, 376 (2009).
25. K. Sawamura, M. T. Yamamoto, T. K. Watanabe, *Genetics* **133**, 307 (1993).
26. P. M. Ferree, D. A. Barbash, *PLoS Biol.* **7**, e1000234 (2009).
27. N. J. Brideau *et al.*, *Science* **314**, 1292 (2006).
28. H. S. Malik, S. Henikoff, *Cell* **138**, 1067 (2009).

29. We thank C-I. Wu for the *D. simulans* fertile and sterile introgression lines; C. Ting for scientific discussions and sharing data; G. Findlay for initial observations on OdsH cytology; and K. Ahmad, S. Biggins, N. Elde, S. Henikoff, N. Phadnis, T. Tsukiyama, and D. Vermaak for comments on the manuscript. Supported by NIH training grant PHS NRSA T32 GM07270 (J.J.B.), and grants from the Mathers foundation and NIH R01-GM74108 (H.S.M.). H.S.M. is an Early-Career Scientist of the Howard Hughes Medical Institute.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1181756/DC1
Materials and Methods
Figs. S1 to S8
References

10 September 2009; accepted 13 October 2009
Published online 22 October 2009;
10.1126/science.1181756
Include this information when citing this paper.

Mapping Human Genetic Diversity in Asia

The HUGO Pan-Asian SNP Consortium*†

Asia harbors substantial cultural and linguistic diversity, but the geographic structure of genetic variation across the continent remains enigmatic. Here we report a large-scale survey of autosomal variation from a broad geographic sample of Asian human populations. Our results show that genetic ancestry is strongly correlated with linguistic affiliations as well as geography. Most populations show relatedness within ethnic/linguistic groups, despite prevalent gene flow among populations. More than 90% of East Asian (EA) haplotypes could be found in either Southeast Asian (SEA) or Central-South Asian (CSA) populations and show clinal structure with haplotype diversity decreasing from south to north. Furthermore, 50% of EA haplotypes were found in SEA only and 5% were found in CSA only, indicating that SEA was a major geographic source of EA populations.

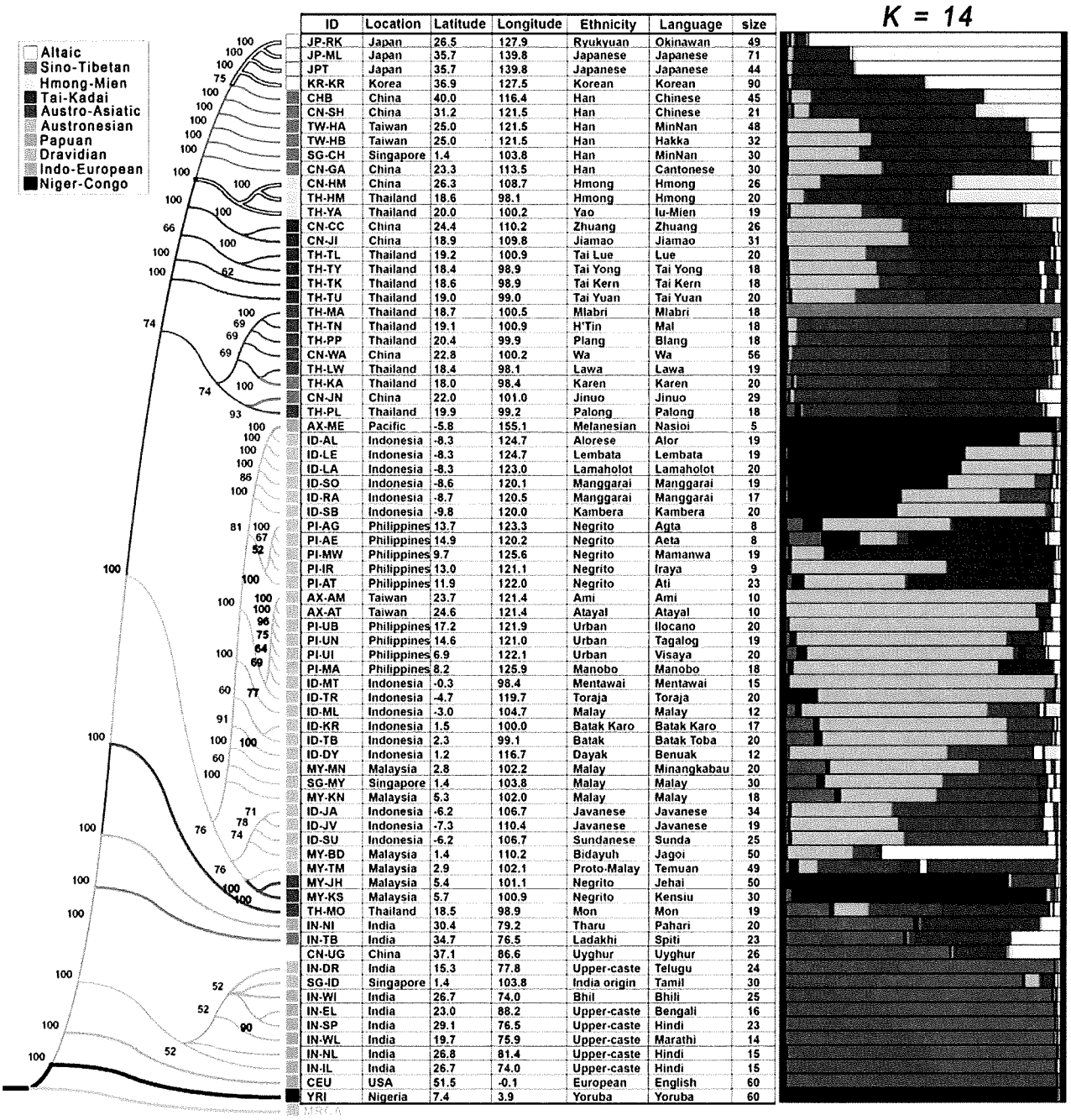
Several genome-wide studies of human genetic diversity focusing primarily on broad continental relationships, or fine-scale structure in Europe, have been published recently (*1*–*8*). We have extended this approach to Southeast Asian (SEA) and East Asian (EA) populations by using the Affymetrix GeneChip Human Mapping 50K Xba Array. Stringently quality-controlled genotypes were obtained at 54,794 autosomal single-nucleotide polymorphisms (SNPs) in 1928 individuals representing 73 Asian and two non-Asian HapMap populations (*9*). Apart from developing a general description of Asian population structure and its relation to geography, language, and demographic history, we concentrated on uncovering the geographic source(s) of EA and SEA populations.

We first performed a Bayesian clustering procedure using the STRUCTURE algorithm (*10*) to examine the ancestry of each individual. Each person is posited to derive from an arbitrary number of ancestral populations, denoted by K. We ran STRUCTURE from $K = 2$ to $K = 14$ using both the complete data set and SNP subsets to exclude those in strong linkage disequilibrium (Fig. 1 and figs. S1 to S13). At $K = 2$ and $K = 3$, all SEA and EA samples are united by predominant membership in a common cluster, with the other cluster(s) corresponding largely to Indo-European (IE) and African (AF) ancestries. At $K = 4$, a component most frequently found in Negrito populations that is also shared by all SEA populations emerges, suggesting a common SEA ancestry. Each value of K beyond 4 introduces a new component that tends to be associated with a group of populations united by membership in a linguistic family, by geographic proximity, by a known history of admixture, or, especially at higher Ks, by membership in a small population isolate. The results obtained using *frappe* (*11*), a maximum-likelihood–based clustering analysis, showed a general concordance with those of STRUCTURE (figs. S14 to S26). These analyses show that most individuals within a population share very similar ancestry estimates at all Ks, an observation that is consistent also with a phylogeny relating individuals (fig. S27) based on an allele-sharing distance (*12*). Therefore, we proceeded to evaluate the relationships among populations. A maximum-likelihood tree of populations, based on 42,793 SNPs whose ancestral states were known (Fig. 1), showed that all the SEA and EA populations make up a monophyletic clade that is supported by 100% of bootstrap replicates. This pattern remained even after data from 51 additional populations and 19,934 commonly typed SNPs from a recent study were integrated into the tree (fig. S28). These observations suggest that SEA and EA populations share a common origin.

STRUCTURE/*frappe* and principal components analyses (PCA) (*13*) (Figs. 1 and 2 and figs. S1 to S26) identify as many as 10 main population components. Each component corresponds largely to one of the five major linguistic groups (Altaic, Sino-Tibetan/Tai-Kadai, Hmong-Mien, Austro-Asiatic, and Austronesian), three ethnic categories (Philippine Negritos, Malaysian Negritos, and East Indonesians/Melanesians) and two small population isolates (the Bidayuh of Borneo and the hunter-gatherer Mlabri population of central and northern Thailand). The STRUCTURE results

*All authors with their affiliations appear at the end of this paper.
†To whom correspondence should be addressed. E-mail: ljin007@gmail.com (L.J.); liue@gis.a-star.edu.sg (E.T.L.); seielstadm@gis.a-star.edu.sg (M.S.); xushua@picb.ac.cn (S.X.)

(Fig. 1 and figs. S1 to S13), population phylogenies (Fig. 1 and figs. S27 and S28), and PCA results (Fig. 2) all show that populations from the same linguistic group tend to cluster together. A Mantel test confirms the correlation between linguistic and genetic affinities ($R^2 = 0.253$; $P < 0.0001$ with 10,000 permutations), even after controlling for geography (partial correlation = 0.136; $P <$ 0.005 with 10,000 permutations). Nevertheless, we identified eight population outliers whose linguistic and genetic affinities are inconsistent [Affymetrix-Melanesian (AX-ME), Malaysia-Jehai (MY-JH)



ID	Location	Latitude	Longitude	Ethnicity	Language	size
JP-RK	Japan	26.5	127.9	Ryukyuan	Okinawan	49
JP-ML	Japan	35.7	139.8	Japanese	Japanese	71
JPT	Japan	35.7	139.8	Japanese	Japanese	44
KR-KR	Korea	36.9	127.5	Korean	Korean	90
CHB	China	40.0	116.4	Han	Chinese	45
CN-SH	China	31.2	121.5	Han	Chinese	21
TW-HA	Taiwan	25.0	121.5	Han	MinNan	48
TW-HB	Taiwan	25.0	121.5	Han	Hakka	32
SG-CH	Singapore	1.4	103.8	Han	MinNan	30
CN-GA	China	23.3	113.5	Han	Cantonese	30
CN-HM	China	26.3	108.7	Hmong	Hmong	26
TH-HM	Thailand	18.6	98.1	Hmong	Hmong	20
TH-YA	Thailand	20.0	100.2	Yao	Iu-Mien	19
CN-CC	China	24.4	110.2	Zhuang	Zhuang	26
CN-JI	China	18.9	109.8	Jiamao	Jiamao	31
TH-TL	Thailand	19.2	100.9	Tai Lue	Lue	20
TH-TY	Thailand	18.4	98.9	Tai Yong	Tai Yong	18
TH-TK	Thailand	18.6	98.9	Tai Kern	Tai Kern	18
TH-TU	Thailand	19.0	99.0	Tai Yuan	Tai Yuan	20
TH-MA	Thailand	18.7	100.5	Mlabri	Mlabri	18
TH-TN	Thailand	19.1	100.9	H'Tin	Mal	18
TH-PP	Thailand	20.4	99.9	Plang	Blang	18
CN-WA	China	22.8	100.2	Wa	Wa	56
TH-LW	Thailand	18.4	98.1	Lawa	Lawa	19
TH-KA	Thailand	18.0	98.4	Karen	Karen	20
CN-JN	China	22.0	101.0	Jinuo	Jinuo	29
TH-PL	Thailand	19.9	99.2	Palong	Palong	18
AX-ME	Pacific	-5.8	155.1	Melanesian	Nasioi	5
ID-AL	Indonesia	-8.3	124.7	Alorese	Alor	19
ID-LE	Indonesia	-8.3	124.7	Lembata	Lembata	19
ID-LA	Indonesia	-8.3	123.0	Lamaholot	Lamaholot	20
ID-SO	Indonesia	-8.6	120.1	Manggarai	Manggarai	19
ID-RA	Indonesia	-8.7	120.5	Manggarai	Manggarai	17
ID-SB	Indonesia	-9.8	120.0	Kambera	Kambera	20
PI-AG	Philippines	13.7	123.3	Negrito	Agta	8
PI-AE	Philippines	14.9	120.2	Negrito	Aeta	8
PI-MW	Philippines	9.7	125.6	Negrito	Mamanwa	19
PI-IR	Philippines	13.0	121.1	Negrito	Iraya	9
PI-AT	Philippines	11.9	122.0	Negrito	Ati	23
AX-AM	Taiwan	23.7	121.4	Ami	Ami	10
AX-AT	Taiwan	24.6	121.4	Atayal	Atayal	10
PI-UB	Philippines	17.2	121.9	Urban	Ilocano	20
PI-UN	Philippines	14.6	121.0	Urban	Tagalog	19
PI-UI	Philippines	6.9	122.1	Urban	Visaya	20
PI-MA	Philippines	8.2	125.9	Manobo	Manobo	18
ID-MT	Indonesia	-0.3	98.4	Mentawai	Mentawai	15
ID-TR	Indonesia	-4.7	119.7	Toraja	Toraja	20
ID-ML	Indonesia	-3.0	104.7	Malay	Malay	12
ID-KR	Indonesia	1.5	100.0	Batak Karo	Batak Karo	17
ID-TB	Indonesia	2.3	99.1	Batak	Batak Toba	20
ID-DY	Indonesia	1.2	116.7	Dayak	Benuak	12
MY-MN	Malaysia	2.8	102.2	Malay	Minangkabau	20
SG-MY	Singapore	1.4	103.8	Malay	Malay	30
MY-KN	Malaysia	5.3	102.0	Malay	Malay	18
ID-JA	Indonesia	-6.2	106.7	Javanese	Javanese	34
ID-JV	Indonesia	-7.3	110.4	Javanese	Javanese	19
ID-SU	Indonesia	-6.2	106.7	Sundanese	Sunda	25
MY-BD	Malaysia	1.4	110.2	Bidayuh	Jagoi	50
MY-TM	Malaysia	2.9	102.1	Proto-Malay	Temuan	49
MY-JH	Malaysia	5.4	101.1	Negrito	Jehai	50
MY-KS	Malaysia	5.7	100.9	Negrito	Kensiu	30
TH-MO	Thailand	18.5	98.9	Mon	Mon	19
IN-NI	India	30.4	79.2	Tharu	Pahari	20
IN-TB	India	34.7	76.5	Ladakhi	Spiti	23
CN-UG	China	37.1	86.6	Uyghur	Uyghur	26
IN-DR	India	15.3	77.8	Upper-caste	Telugu	24
SG-ID	Singapore	1.4	103.8	India origin	Tamil	30
IN-WI	India	26.7	74.0	Bhil	Bhili	25
IN-EL	India	23.0	88.2	Upper-caste	Bengali	16
IN-SP	India	29.1	76.5	Upper-caste	Hindi	23
IN-WL	India	19.7	75.9	Upper-caste	Marathi	14
IN-NL	India	26.8	81.4	Upper-caste	Hindi	15
IN-IL	India	26.7	74.0	Upper-caste	Hindi	15
CEU	USA	51.5	-0.1	European	English	60
YRI	Nigeria	7.4	3.9	Yoruba	Yoruba	60

Fig. 1. Maximum-likelihood tree of 75 populations. A hypothetical most-recent common ancestor (MRCA) composed of ancestral alleles as inferred from the genotypes of one gorilla and 21 chimpanzees was used to root the tree. Branches with bootstrap values less than 50% were condensed. Population identification numbers (IDs), sample collection locations with latitudes and longitudes, ethnicities, language spoken, and size of population samples are shown in the table adjacent to each branch in the tree. Linguistic groups are indicated with colors as shown in the legend. All population IDs except the four HapMap samples are denoted by four characters. The first two letters indicate the country where the samples were collected or (in the case of Affymetrix) genotyped, according to the following convention: AX, Affymetrix; CN, China; ID, Indonesia; IN, India; JP, Japan; KR, Korea; MY, Malaysia; PI, the Philippines; SG, Singapore; TH, Thailand; and TW, Taiwan. The last two letters are unique IDs for the population. To the right of the table, an averaged graph of results from STRUCTURE is shown for $K = 14$.

(Negrito), Malaysia-Kensiu (MY-KS) (Negrito), Thailand-Mon (TH-MO), Thailand-Karen (TH-KA), China-Jinuo (CN-JN), India-Spiti (IN-TB), and China-Uyghur (CN-UG); see table S3]. These linguistic outliers tend to cluster with their geographic neighbors or [especially evident in the principal component (PC) plots of Fig. 2] occupy an intermediate position between their geographic neighbors and the more-distant members of their linguistic group. These patterns are consistent either with substantial recent admixture among the populations (*14*–*16*), a history of language replacement (*17*), or uncertainties in the linguistic classifications themselves (for example, the controversial Altaic family, which groups Korean and Japanese with Uyghur).

Considerable gene flow among Asian populations was observed among subpopulations in these clusters, including those groups believed to practice endogamy based on linguistic, cultural, and ethnic information. In fact, most populations studied, even at lower *K*s, show evidence of admixture in the STRUCTURE analyses. For example, the Han Chinese have grown to become the largest ethnic group today in a demographic expansion that has occurred mostly within historical times. STRUCTURE reveals that the six Han Chinese population samples in our study show varying degrees of admixture (Fig. 1 and figs. S1 to S26) between a northern Altaic cluster and a Sino-Tibetan/Tai-Kadai cluster, which most frequently appears in the ethnic groups sampled from southern China and northern Thailand. Finally, most of the Indian populations showed evidence of shared ancestry with European populations, which is consistent with the recent observations (*18*) and our understanding of the expansion of Indo-European–speaking populations (Fig. 1 and figs. S1 to S26).

The geographic source(s) contributing to EA populations have long been debated. One hypothesis suggests that all SEA and EA populations derive primarily from a single initial migration, which entered the continent along a southern, largely coastal route (*19*, *20*). Another hypothesis argues for at least two independent migrations into East Asia, first along a southern route, followed later by a series of migrations along a more northern route that served to bridge European and EA populations, but with little contribution to populations in Southeast Asia (*20*). The topology of a maximum-likelihood tree (Fig. 1 and fig. S28) displays a largely south-to-north ordering of the populations, and a plot of the first two PCs (Fig. 2) similarly orients most populations according to their geographic coordinates. The average

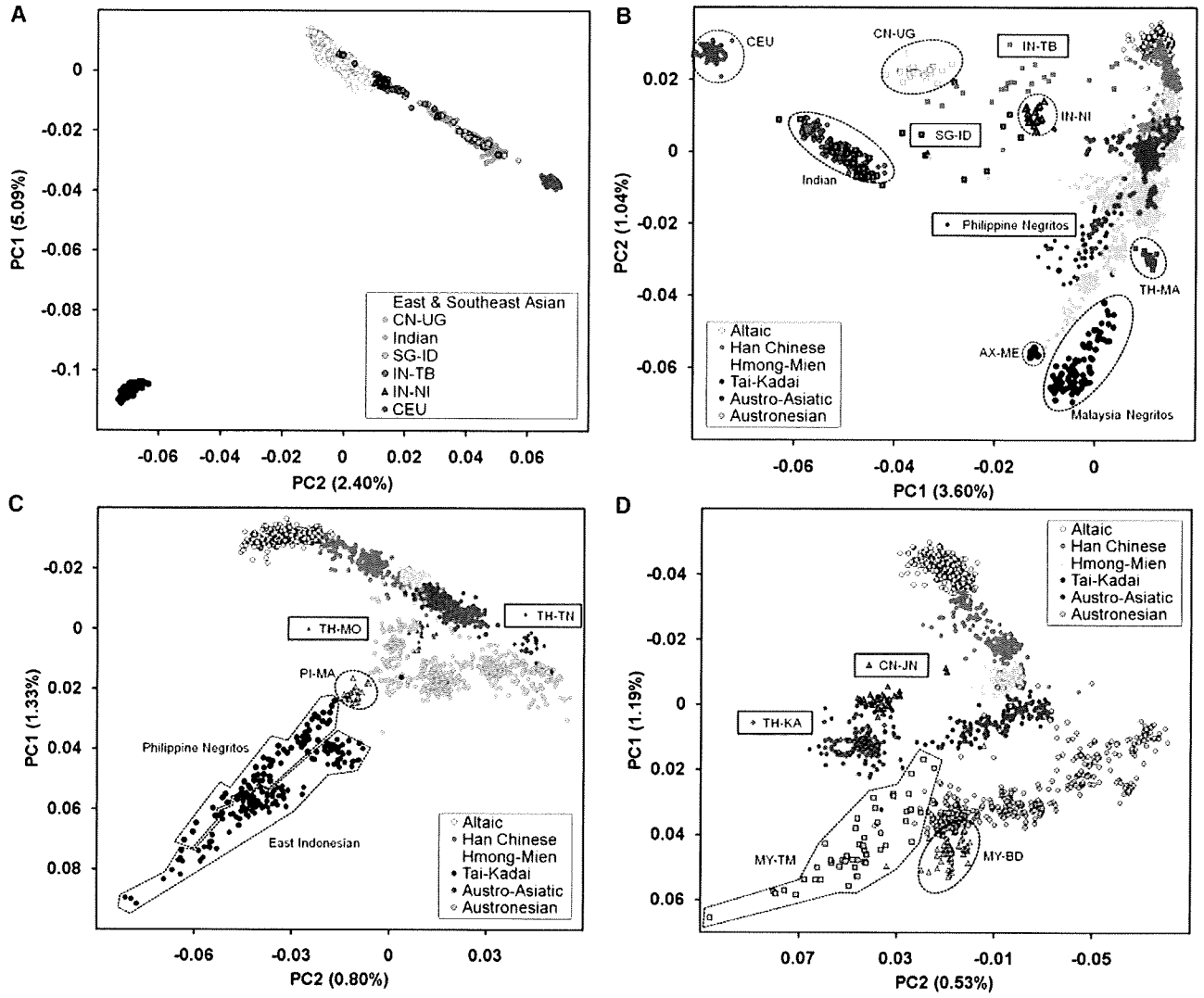


Fig. 2. Analysis of the first two PCs. **(A)** 1928 individuals representing all 75 populations. **(B)** 1868 individuals representing 74 populations (excluding YRI). **(C)** 1471 individuals representing 58 populations (excluding all Indians, CN-UG, TH-MA, AX-ME, and Negritos from Malaysia). **(D)** 1235 individuals representing 44 populations (excluding Philippine Negritos, PI-MA, and East Indonesians).

value of the first PC is highly correlated with the latitude at which the populations were sampled (R^2 = 0.79, P < 0.0001). Such a pattern could result simply from isolation-by-distance (IBD), as suggested by Ding *et al.* (*21*), although a recent study failed to detect IBD in East Asia with data from the Human Genome Diversity Project (*22*).

In an effort to distinguish between long-term historical divergence and the effects of IBD, we applied partial and multiple Mantel tests to the data (*23*) [see supporting online material (SOM) text for details]. The primary approach was to ascertain the differential correlation between genetic distance, geographical distance, and a group indicator matrix as an indication of prehistoric population divergence. The partial correlation coefficient of genetic and geographic distances was 0.228 (P < 0.0006), after controlling for the group indicator matrix (inferred from STRUCTURE/*frappe* analyses), whereas the partial correlation of the genetic and group indicator matrices was 0.403 (P < 0.0001) after controlling for geography. The superior association between genetic distance and the group indicator matrix as measured by the correlation coefficients suggests that prehistorical population divergence is the favored model over IBD in explaining the data (*24*). This conclusion is supported by simulation studies that also suggest that the observed patterns cannot be explained by simple IBD effects alone (see SOM text for details).

To further refine the analysis, we looked to haplotype organization to limit the effect of fluctuations in single-nucleotide determinations and to increase the resolution around genetic diversity. The IBD model predicts a correlation of genetic distance with geographical distance but not genetic diversity and geographic distance (*24*). By contrast, we found (Fig. 3A) that haplotype diversity is strongly correlated with latitude (R^2 = 0.91, P < 0.0001), with diversity decreasing from south to north, which is consistent with a loss of diversity as populations moved to higher latitudes. In estimating the contribution of SEA and Central-South Asian (CSA) haplotypes to the EA gene pool by haplotype sharing analyses (*16*), we found that more than 90% of haplotypes in EA populations could be found in SEA and CSA populations, of which about 50% were found in SEA and EA only and 5% found in CSA only (Fig. 3B, see also SOM text). Phylogenetic analysis of private haplotypes indicates greater similarity between EA and SEA populations relative to EA and CSA populations (Fig. 3C). These observations suggest that the geographic source(s) contributing to EA populations were mainly from SEA populations, with rather minor contributions from CSA,

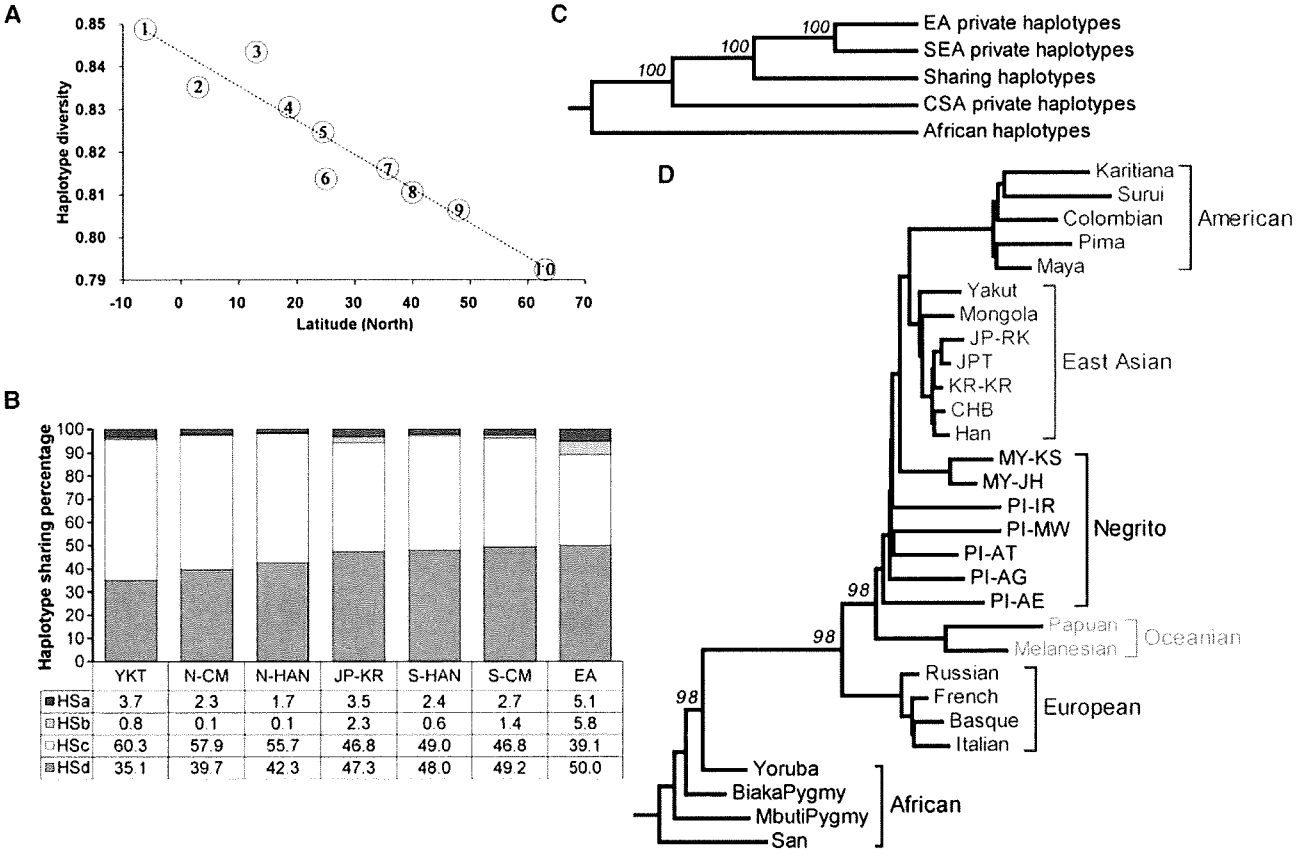


Fig. 3. Analysis of haplotype diversity, haplotype sharing, and population phylogeny. **(A)** Haplotype diversity versus latitudes. Haplotypes were estimated from combined data, and diversity was measured by heterozygosity of haplotypes. HSa, b, c, and d and the corresponding colors show the percentages of EA group haplotypes in each class: HSa, found in CSA only; HSb, found in neither CSA nor SEA; HSc, found in both CSA and SEA; HSd, found in SEA only. Latitudes (*y* axis) for groups were obtained from the center of sample collection locations. Circled numbers are as follows: 1, Indonesian; 2, Malay; 3, Philippine; 4, Thai; 5, Southern Chinese minorities; 6, Southern Han Chinese; 7, Japanese and Korean; 8, Northern Han Chinese; 9, Northern Chinese minorities; and 10, Yakut. Haplotype heterozygosity of each group was estimated from 100-kb bins and taking together all haplotypes within each group. R^2 for the regression line is 0.91 (P < 0.0001). **(B)** Haplotype sharing analysis for EA populations and groups. YKT, Yakut; N-CM, Northern Chinese minorities; N-HAN, Northern Han Chinese; JP-KR, Japanese and Korean; S-HAN, Southern Han Chinese; S-CM, Southern Chinese minorities; EA, East Asian. **(C)** Phylogeny of group private haplotypes. EA private haplotypes: haplotypes found only in EA samples; SEA private haplotypes: haplotypes found only in SEA samples; CSA private haplotypes: haplotypes found only in CSA samples; Shared haplotypes: haplotypes found in all EA, SEA, and CSA samples; African haplotypes were used as outgroup. **(D)** Maximum-likelihood tree of 29 populations. The tree is based on data from 19,934 SNPs. Bootstrap values were based on 100 replicates. Only values on splitting of African and non-African, European and Oceanian and Asian, and Oceanian and Asian are shown.

and that this clinal structure of EA populations arose from prehistoric population divergence rather than IBD or gene flow from CSA populations.

On the basis of increased cultural, linguistic, and genetic diversity, the origins of SEA populations are thought to be more complex than the origins of those to their north. Notably, the Negritos of the Philippines and Malaysia differ from neighboring populations in aspects of their physical appearance, prompting intense speculation about models of human settlement in Southeast Asia. The two-wave hypothesis, which suggests that ancestral Negrito populations settled in Southeast Asia, Australia, and Oceania before a more northerly migration originating in or near the Middle East, and spreading both toward Europe and Northeast Asia via Central Asia (*25*), has been supported by phylogenetic trees constructed from data on a limited number of protein markers (*24*, *25*). The topology of our population trees, both with and without the data from additional European and Asian populations discussed in (*1*), is inconsistent with regard to this genetic similarity of European and EA populations (Figs. 1 and 3D). Instead, on the basis of variation at a large number of independent SNPs, we observed that there is substantial genetic proximity of SEA and EA populations (fig. S28). An identical pattern is seen in the population tree of Li *et al.* (*1*) based on all of their 642,690 SNPs. Our forward-time simulation results under extreme ascertainment scenarios (SOM text) show that the observed phylogeny is not the result of ascertainment bias. Simulation studies also suggest that substantial levels of migration between populations after their initial separation are unlikely to distort the topology of the phylogeny (SOM text).

To unambiguously infer population histories represents a considerable challenge (*26*). Although this study does not disprove a two-wave model of migration, the evidence from our autosomal data and the accompanying simulation studies (figs. S29 and S30) point toward a history that unites the Negrito and non-Negrito populations of Southeast and East Asia via a single primary wave of entry of humans into the continent.

References and Notes

1. J. Z. Li *et al.*, *Science* **319**, 1100 (2008).
2. M. Kayser *et al.*, *Am. J. Hum. Genet.* **82**, 194 (2008).
3. N. A. Rosenberg *et al.*, *PLoS Genet.* **1**, e70 (2005).
4. N. A. Rosenberg *et al.*, *Science* **298**, 2381 (2002).
5. J. Novembre *et al.*, *Nature* **456**, 98 (2008).
6. M. Nelis *et al.*, *PLoS One* **4**, e5472 (2009).
7. C. Tian *et al.*, *PLoS Genet.* **4**, e4 (2008).
8. O. Lao *et al.*, *Curr. Biol.* **18**, 1241 (2008).
9. The International HapMap Consortium, *Nature* **426**, 789 (2003).
10. J. K. Pritchard, M. Stephens, P. Donnelly, *Genetics* **155**, 945 (2000).
11. H. Tang, J. Peng, P. Wang, N. J. Risch, *Genet. Epidemiol.* **28**, 289 (2005).
12. J. L. Mountain, L. L. Cavalli-Sforza, *Am. J. Hum. Genet.* **61**, 705 (1997).
13. N. Patterson, A. L. Price, D. Reich, *PLoS Genet.* **2**, e190 (2006).
14. S. Xu, L. Jin, *Am. J. Hum. Genet.* **83**, 322 (2008).
15. S. Xu, W. Huang, J. Qian, L. Jin, *Am. J. Hum. Genet.* **82**, 883 (2008).
16. S. Xu, W. Jin, L. Jin, *Mol. Biol. Evol.* **26**, 2197 (2009).
17. L. Reid, in *Language Contact and Change in the Austronesian World.* T. Dutton, T. Tryon, Eds. (Mouton de Gruyter, Berlin, 1994) pp. 443–475.
18. Indian Genome Variation Consortium, *J. Genet.* **87**, 3 (2008).
19. J. Y. Chu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11763 (1998).
20. B. Su *et al.*, *Am. J. Hum. Genet.* **65**, 1718 (1999).
21. Y. C. Ding *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 14003 (2000).
22. A. Manica, F. Prugnolle, F. Balloux, *Hum. Genet.* **118**, 366 (2005).
23. M. P. Telles, J. A. Diniz-Filho, *Genet. Mol. Res.* **4**, 742 (2005).
24. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ, 1993).
25. L. L. Cavalli-Sforza, M. W. Feldman, *Nat. Genet.* **33**, 266 (2003).
26. G. Hellenthal, A. Auton, D. Falush, *PLoS Genet.* **4**, e1000078 (2008).

27. The entire consortium thanks all individuals who volunteered their DNA for this project. It is this collaboration between scientists and the public that is essential to progress in our field. All SNP data have been submitted to dbSNP with the submission handle PASNPI and will become accessible in dbSNP Build 131. See SOM text for a complete listing of all acknowledgments.

The HUGO Pan-Asian SNP Consortium

Mahmood Ameen Abdulla,[1] Ikhlak Ahmed,[2] Anunchai Assawamakin,[3,4] Jong Bhak,[5] Samir K. Brahmachari,[2] Gayvelline C. Calacal,[6] Amit Chaurasia,[2] Chien-Hsiun Chen,[7] Jieming Chen,[8] Yuan-Tsong Chen,[7] Jiayou Chu,[9] Eva Maria C. Cutiongco-de la Paz,[10] Maria Corazon A. De Ungria,[6] Frederick C. Delfin,[6] Juli Edo,[1] Suthat Fuchareon,[3] Ho Ghang,[5] Takashi Gojobori,[11,12] Junsong Han,[13] Sheng-Feng Ho,[7] Boon Peng Hoh,[14] Wei Huang,[15] Hidetoshi Inoko,[16] Pankaj Jha,[2] Timothy A. Jinam,[1] Li Jin,[17,38]† Jongsun Jung,[18] Daoroong Kangwanpong,[19] Jatupol Kampuansai,[19] Giulia C. Kennedy,[20,21] Preeti Khurana,[22] Hyung-Lae Kim,[18] Kwangjoong Kim,[18] Sangsoo Kim,[23] Woo-Yeon Kim,[5] Kuchan Kimm,[24] Ryosuke Kimura,[25] Tomohiro Koike,[11] Supasak Kulawonganunchai,[4] Vikrant Kumar,[8] Poh San Lai,[26,27] Jong-Young Lee,[18] Sunghoon Lee,[5] Edison T. Liu,[8]† Partha P. Majumder,[28] Kiran Kumar Mandapati,[22] Sangkot Marzuki,[29] Wayne Mitchell,[30,31] Mitali Mukerji,[2] Kenji Naritomi,[32] Chumpol Ngamphiw,[4] Norio Niikawa,[40] Nao Nishida,[25] Bermseok Oh,[18] Sangho Oh,[5] Jun Ohashi,[25] Akira Oka,[16] Rick Ong,[8] Carmencita D. Padilla,[10] Prasit Palittapongarnpim,[33] Henry B. Perdigon,[6] Maude Elvira Phipps,[1,34] Eileen Png,[8] Yoshiyuki Sakaki,[35] Jazelyn M. Salvador,[6] Yuliana Sandraling,[29] Vinod Scaria,[2] Mark Seielstad,[8]† Mohd Ros Sidek,[14] Amit Sinha,[2] Metawee Srikummool,[19] Herawati Sudoyo,[29] Sumio Sugano,[37] Helena Suryadi,[29] Yoshiyuki Suzuki,[11] Kristina A. Tabbada,[6] Adrian Tan,[8] Katsushi Tokunaga,[25] Sissades Tongsima,[4] Lilian P. Villamor,[6] Eric Wang,[20,21] Ying Wang,[15] Haifeng Wang,[15] Jer-Yuarn Wu,[7] Huasheng Xiao,[13] Shuhua Xu,[38]† Jin Ok Yang,[5] Yin Yao Shugart,[39] Hyang-Sook Yoo,[5] Wentao Yuan,[15] Guoping Zhao,[15] Bin Alwi Zilfalil,[14] Indian Genome Variation Consortium[2]

[1]Department of Molecular Medicine, Faculty of Medicine, and the Department of Anthropology, Faculty of Arts and Social Sciences, University of Malaya, Kuala Lumpur, 50603, Malaysia. [2]Institute of Genomics and Integrative Biology, Council for Scientific and Industrial Research, Mall Road, Delhi 110007, India. [3]Mahidol University, Salaya Campus, 25/25 M. 3, Puttamonthon 4 Road, Puttamonthon, Nakornpathom 73170, Thailand. [4]Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, Thailand Science Park, Pathumtani 12120, Thailand. [5]Korean BioInformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), 111 Gwahangno, Yuseong-gu, Deajeon 305-806, Korea. [6]DNA Analysis Laboratory, Natural Sciences Research Institute, University of the Philippines, Diliman, Quezon City 1101, Philippines. [7]Institute of Biomedical Sciences, Academia Sinica, 128 Sec 2 Academia Road Nangang, Taipei City 115, Taiwan. [8]Genome Institute of Singapore, 60 Biopolis Street 02-01, 138672, Singapore. [9]Institute of Medical Biology, Chinese Academy of Medical Science, Kunming, China. [10]Institute of Human Genetics, National Institutes of Health, University of the Philippines Manila, 625 Pedro Gil Street, Ermita Manila 1000, Philippines. [11]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411-8540, Japan. [12]Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan. [13]National Engineering Center for Biochip at Shanghai, 151 Li Bing Road, Shanghai 201203, China. [14]Human Genome Center, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia. [15]MOST-Shanghai Laboratory of Disease and Health Genomics, Chinese National Human Genome Center Shanghai, 250 Bi Bo Road, Shanghai 201203, China. [16]Department of Molecular Life Science Division of Molecular Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara-A Kanagawa-Pref A259-1193, Japan. [17]State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, 220 Handan Road, Shanghai 200433, China. [18]Korea National Institute of Health, 194, Tongil-Lo, Eunpyung-Gu, Seoul, 122-701, Korea. [19]Department of Biology, Faculty of Science, Chiang Mai University, 239 Huay Kaew Road, Chiang Mai 50202, Thailand. [20]Genomics Collaborations, Affymetrix, 3420 Central Expressway, Santa Clara, CA 95051, USA. [21]Veracyte, 7000 Shoreline Court, Suite 250, South San Francisco, CA 94080, USA. [22]The Centre for Genomic Applications (an IGIB-IMM Collaboration), 254 Ground Floor, Phase III Okhla Industrial Estate, New Delhi 110020, India. [23]Soongsil University, Sangdo-5-dong 1-1, Dongjak-gu, Seoul 156-743, Korea. [24]Eulji University College of Medicine, 143-5 Yong-dudong Jung-gu, Dae-jeon City 301-832, Korea. [25]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. [26]Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, National University Hospital, 5 Lower Kent Ridge Road, 119074, Singapore. [27]Population Genetics Lab, Defence Medical and Environmental Research Institute, DSO National Laboratories, 27 Medical Drive, 117510, Singapore. [28]Indian Statistical Institute (Kolkata) 203 Barrackpore Trunk Road, Kolkata 700108, India. [29]Eijkman Institute for Molecular Biology, Jl. Diponegoro 69, Jakarta 10430, Indonesia. [30]Informatics Experimental Therapeutic Centre, 31 Biopolis Way, 03-01 Nanos, 138669, Singapore. [31]Division of Information Sciences, School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore. [32]Department of Medical Genetics, University of the Ryukyus Faculty of Medicine, Nishihara, 207 Uehara, Okinawa 903-0215, Japan. [33]National Science and Technology Development Agency, 111 Thailand Science Park, Pathumtani 12120, Thailand. [34]Monash University (Sunway Campus), Jalan Lagoon Selatan, 46150 Bandar Sunway, Selangor, Malaysia. [35]RIKEN Genomic Sciences Center, W502, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan. [36]Department of Biochemistry, University of Hong Kong, 3/F Laboratory Block, Faculty of Medicine Building, 21 Sasson Road, Pokfulam, Hong Kong. [37]Laboratory of Functional Genomics, Department of Medical Genome Sciences Graduate School of Frontier Sciences, University of Tokyo (Shirokanedai Laboratory), 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. [38]Chinese Academy of Sciences-Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Rd., Shanghai 200031, China. [39]Genomic Research Branch, National Institute of Mental Health, National Institutes of Health, 6001 Executive Boulevard, Bethesda, MD 20892 USA. [40]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Tobetsu 061-0293, Japan.

Supporting Online Material

www.sciencemag.org/cgi/content/full/326/5959/1541/DC1
Materials and Methods
SOM Text
Figs. S1 to S38
Tables S1 to S4

1 June 2009; accepted 13 October 2009
10.1126/science.1177074

Immunogenetics (2009) 61:257–270
DOI 10.1007/s00251-009-0364-0

ORIGINAL PAPER

HLA-A allele associations with viral *MER9-LTR* nucleotide sequences at two distinct loci within the *MHC alpha* block

Jerzy K. Kulski · Atsuko Shigenari · Takashi Shiina · Kazuyoshi Hosomichi · Makoto Yawata · Hidetoshi Inoko

Received: 25 November 2008 / Accepted: 25 February 2009 / Published online: 18 March 2009

Abstract The study of the association of the *Human Leukocyte Antigen* (*HLA*) alleles and polymorphic retrotransposons such as *Alu*, *HERV*, and *LTR* at various loci within the *Major Histocompatibility Complex* allows for a better identification and stratification of disease associations and the origins of *HLA* haplotypes in different populations. This paper provides sequence and association data on two structurally polymorphic *MER9-LTR* retrotransposons that are located 54 kb apart and in close proximity to the multiallelic *HLA-A* gene involved in the regulation of the human immune system. Direct *DNA* sequencing and analysis of the PCR products identified *DNA* nucleotide variations between the *MER9-LTR* sequences at the two loci and their associations with *HLA-A* alleles as potential haplotype and evolutionary markers. All *MER9-LTR* sequences were haplotypic when associated with common *HLA-A* alleles. The number of SNP loci was 2.5 times greater for the solo *LTR* at the *AK* locus, which is located closer to the *HLA-A* gene than the solo or 3′ *LTR* at the *HG* locus. Our study shows that the nucleotide variations of the *MER9-LTR DNA* sequences are additional informative markers in fine mapping *HLA-A* genomic haplotypes for future population, evolutionary, and disease studies.

Keywords *HERVK9 · LTR · MER9 · HLA-A · MHC ·* Polymorphism · Haplotype · Duplication

J. K. Kulski (✉)
Centre for Forensic Science, The University of Western Australia,
M420, 35 Stirling Highway,
Crawley, Western Australia, Australia
e-mail: kulski@mac.com

J. K. Kulski · A. Shigenari · T. Shiina · K. Hosomichi · H. Inoko
Department of Molecular Life Science,
Division of Basic Medical Science and Molecular Medicine,
Tokai University School of Medicine,
Isehara, Kanagawa, Japan

A. Shigenari
e-mail: ashige@is.icc.u-tokai.ac.jp

T. Shiina
e-mail: tshiina@is.icc.u-tokai.ac.jp

K. Hosomichi
e-mail: hoso@is.icc.u-tokai.ac.jp

H. Inoko
e-mail: hinoko@is.icc.u-tokai.ac.jp

M. Yawata
Department of Structural Biology, School of Medicine,
Stanford University,
Stanford, CA, USA
e-mail: GM5M-YWT@j.asahi-net.or.jp

Introduction

Retrotransposons are a class of mobile genetic elements that have been transposed replicatively within their host genomes via RNA intermediates and contribute to at least 44% of the human genomic content as interspersed repeat elements (Lander et al. 2001). The retrotransposons are basically classified into two groups, those with long terminal repeats (*LTR*s) and those without (Kapitonov et al. 2004). The *LTR* group consists mostly of solitary *LTR* structures and human endogenous retroviruses (*HERV*s) with the flanking 5′ and 3′ *LTR*s as part of their *DNA* sequence (Mager and Medstrand 2003). The non-*LTR* group consists mostly of various long and short interspersed repeat families such as *Alu*, *L1*, *L2*, and *SVA* (Belancio et al. 2008). Most retrotransposons are replicatively inactive and are permanently fixed within the human genome, but a small percentage remains active and is structurally poly-

morphic (absent or present) within the genome of populations (Bennett et al. 2004). After insertion, all retrotransposons would have been structurally polymorphic in populations for a time prior to their fixation. In this regard, structurally polymorphic retrotransposons are useful evolutionary markers and molecular clocks in population studies because their presence or absence is related by descent after the insertion event (Batzer and Deininger 2002; Bennett et al. 2004; Kulski and Dunn 2005; Terreros et al. 2005).

Some polymorphic retrotransposons within the *Major Histocompatibility Complex* (*MHC*) genomic region (chromosome position *6p21.3*) have been associated with *Human Leukocyte Antigen* (*HLA*) alleles potentially allowing for a better identification and stratification of disease associations and the origins of *HLA* haplotypes in different populations. Although studies during the past decade have focused mainly on polymorphic *Alu* retrotransposons within the *MHC* region of normal and diseased populations (Dunn et al. 2006, 2007; Kulski et al. 2001; Kulski and Dunn 2005), recent attention has turned to polymorphic *HERV* elements (Kulski et al. 2008). One example of a polymorphic *HERV* within the *MHC* class I region is the *human endogenous retrovirus K9* (*HERVK9* alias *HERV-K HML-3*) family, which has approximately 150 copies distributed in the human genome (Mager and Medstrand 2003; Mayer and Meese 2005) and is transcribed in various normal and abnormal tissues (Medstrand and Blomberg 1993; Seifarth et al. 2005). The provirus includes 5′ and 3′ *LTR* flanking sequences called the *5′* and *3′ MER9-LTR* (Kapitonov et al. 2004). Single copies of the *MER9-LTR* sequence (solo or *sMER9* or s*LTR*) have been generated by homologous recombination between the two flanking *LTR* that have resulted in the deletion of the internal *HERVK9* sequence (Kulski et al. 2005; Kulski et al. 2008). The *sMER9* sequences are found more frequently in the genome than the internal proviral sequences and the flanking *5′* and *3′ MER9* sequences (Mager and Medstrand 2003). The *HERVK9* sequences were first fixed in the genome about 35 Myr ago, before the emergence of the Old World monkeys (OWM; Mayer and Meese 2005).

There are at least two copies of *HERVK9* and up to three copies of *sMER9-LTR* sequences within the *MHC* class I genomic region. One of the *HERVK9* sequences is telomeric of the *HLA-C* gene (Kulski et al. 1999), and the other is located between the *HLA-H* pseudogene and the *HLA-G* coding gene at the *HG* locus ~63 kb telomeric of the *HLA-A* gene (Stewart et al. 2004; Kulski et al. 2008). There are at least three *sMER9-LTR* sequence loci: the *HG* locus (Stewart et al. 2004; Kulski et al. 2008), the *AK* locus immediately telomeric of *HLA-A* (Kulski et al. 2005), and the *BC* locus, which is telomeric of the *HLA-B* gene and centromeric of the *HERVI* sequence (Kulski et al. 1999). The *MER9-LTR*s at the *HG* and *AK* loci appear to have evolved within the *MHC* genomic region as products of a multigenic duplication at some time after the emergence of the rhesus macaque, but probably before the emergence of the great apes (Kulski et al. 2004, 2005). The *sMER9-LTR. HG* sequence, representing a solitary *MER9* sequence and a *HERVK9* deletion at the *HG* locus, occurs at an average frequency of 41% in Japanese and 66% in African Americans or Caucasians and has a strong hitchhiking association with *HLA-A* alleles (Kulski et al. 2008). Because of the strong association between some *HLA-A* alleles and the structural *MER9* polymorphism at the *HG* locus, the loss or absence of the expected association or linkage can be informative in assessing the frequency of recombination between the two loci. This assessment might be strengthened further with a better understanding of the associations between the *MER9* nucleotide sequences and *HLA-A* alleles. However, there is little or no published analysis on the nucleotide variation of the *MER9-LTR* at either the *HG* or the *AK* loci and their association with *HLA-A* alleles.

The aim of the present study was to determine whether there are nucleotide differences between the *5′*, *3′* and *sMER9* sequences at the *HG* and *AK* loci and whether these differences might be associated with *HLA-A* alleles as potential haplotypic and stratification markers.

Materials and methods

DNA samples

Fifty-five *HLA DNA* samples, extracted from B-lymphoblastoid cell lines of different ethnic origins and genotyped and/or serotyped for *HLA* alleles at the *HLA-A*, *-B*, and *-DRB1* loci (Table 1) were purchased from the European Collection of Cell Cultures (http://www.ecacc.org.uk/), now the Health Protection Agency Culture Collection (http://www.hpacultures.org.uk/products/celllines/hlatyped/search.jsp). These samples were chosen for *MER9* sequencing and analysis because they were mostly homozygous (48 of 55 samples) for *HLA-A*. Some Australian Aborigine (two of five samples) and Japanese (four of nine samples) cell lines were heterozygous for *HLA-A*.

PCR and sequencing of *MER9-LTR*

The *MER9-LTR* sequences were PCR-amplified using the *DNA* samples extracted from the cell lines shown in Table 1 and sequenced for *MER9* at the *AK* and *HG* locus. The *sMER9* or *3′ MER9* sequences at the *HG* locus (Fig. 1) were amplified by PCR using the primer pairs PCR-A del (3Si2/3ASe2) and PCR-B ins (1Se1/3ASe2), respectively, as previously described (Kulski et al. 2008). The PCR-A del

Table 1 Homozygous and heterozygous *HLA DNA* cell lines sequenced for *MER9* at the *HG* and *AK* loci

No.	Cell line name	IHW number[a]	Ethnic origin	*HLA* alleles			*HERVK9*		
				A	*B*	*DRB1*	*3′ MER9. HG*	*sMER9. HG*	*sMER9.AK*
1	WBD001816	9154		01	17	07		+	+
2	APD	9291		01	60	0402		+	+
3	HAM 013	9178	South African Caucasoid	01	08	1201 0301		+	+
4	E4181324	9011	Australian Caucasoid	0101	52011	15021		+	+
5	J0528239	9041	Italian	0101	3502	1104		+	+
6	VAVY	9023	French	0101	0801	0301		+	+
7	LO541265	9086	Australian Caucasoid	0101	0801	0301		+	+
8	PF04015	9088	French	0101	0801	0301		+	+
9	COX	9022	South African Caucasoid	0101	0801	0301		+	+
10	WATANABE	9126	Oriental	02	46	8032		+	+
11	HAY, KJ	9196	Australian Aborigine	02	15	1301 14		+	+
12	BSM	9032	Dutch	0201	1501	04		+	+
13	BOLETH BO	9031	Swedish	0201	1501	0401		+	+
14	WT9, 31227ABO	9061	Italian	0201	1801	1401		+	+
15	KOSE	9056	German	0201	3503	1302 1401		+	+
16	BER	9093	German	0201	1302	0701		+	+
17	DBB	9052	Amish	0201	5701	0701		+	+
18	HID	9074	Japanese	0201	4001 4006	09		+	+
19	SPO010 SPO	9036	Italian	0201	4402	1101		+	+
20	AWELLS_WEL	9090	Australian Caucasoid	0201	4402	0401		+	+
21	EK	9054	Scandinavian	0201	4402	1401		+	+
22	BM16	9038	Italian	0201	1801	1201		+	+
23	TAB089	9066	Japanese	0207	4601	8031		+	+
24	WAL, FD	9129	Caucasoid	03	07	1501	+		+
25	HO104	9082	French	03	07	?	+		+
26	DRI, SM	9128		03	07 35	0101 1501	+		+
27	PLH	9047	Scandinavian	0301	4701	0701	+		+
28	PGF	9318	English	0301	0702	1501	+		+
29	SCHU	9013	French	0301	0702	1501	+		+
30	EA	9081	Scandinavian	0301	0702	1501	+		+
31	WT100BIS	9006	Italian	1101	3501	0101	+		+
32	HOSONUM	9130	Oriental	24	07	0101	+		DELETION
33	KUROIWA	9131	Oriental	24	07	0101	+		DELETION
34	TISI variant			01	57	07		+	+
35	SA	9001	Japanese	2402	0702	01	+		DELETION
36	AKIBA	9286	Japanese	2402	5201	1502	+		DELETION
37	LKT3	9107	Japanese	2402	5401	0405	+		DELETION
38	QBL	9020	Dutch	2601	1801	0301		+	+
39	MOU MANN M	9050	Danish	2902	44031	0701		+	+
40	LBF, LBUF	9048	Caucasoid	3001	1302	0701	+		+

Table 1 (continued)

No.	Cell line name	IHW number[a]	Ethnic origin	*HLA* alleles			*HERVK9*		
				A	*B*	*DRB1*	*3′ MER9. HG*	*sMER9. HG*	*sMER9.AK*
41	SPL SPACH	9101	Amish	31	1501	0802	+		+
42	SSTO variant			31	15	08	+		+
43	WON, PY	9156	Oriental	33	58	0301	+		+
44	HAU, ML	9157	Oriental	33	58	0301	+		+
45	LWAGS	9079	Ashkenasi Jewish	3301	1402	0102	+		+
46	IBW9	9049	Sardinian	3301	1402	0701	+		+
47	WON I	9194	Australian Aborigine	34	40 56	08 14		+	+
48	WON C	9195	Australian Aborigine	34	40 15	08 12		+	+
49	NON L	9192	Australian Aborigine	01 10	51 55	04 8		+	+
50	EK-TOK	9354	Japanese	11 2602	35 46	0405 1101		+	+
51	KOZ	9310	Japanese	24 26	40 54	09	+	+	+
52	LKT14	9103	Japanese	24 26	40 51	09	+	+	+
53	BEA, PL	9138		02 29	44 62	0401 07		+	+
54	LKT12	9073	Japanese	2402 3101	3501 5201		+		+
55	IHL, AD031	9117	Australian Aborigine	02 31	27 40	04 08	+	+	+

[a] The IHW number is the International Immunohistocompatibility Workshop number at http://www.hpacultures.org.uk/products/celllines/hlatyped/search.jsp

primer pair 5′-GTCACCCCCTAGAAGGAGACC-3′ and 5′-CAGAAGACTCAGGATGGAGTCTCC-3′ produced an amplified product size of 556 bp. The PCR-B ins primer pair 5′-AGATGCAGATCCCGATTCCTGC-3′ and 5′-CAGAAGACTCAGGATGGAGTCTCC-3′ produced an amplified product size of 625 bp. The *sMER9* at the *AK* locus (PCR-C solo) was amplified as a 547 bp product with the primer pair M9AK.S (5′ to 3′ GTCATCCTCCAGAAG GAGACT) and M9AK.AS (3′ to 5′ CACAAGACTCAG CATGGAGTCTTC). The genomic coordinates for the PCR primer pair products on the human chromosome 6 reference sequence NW_001838980.1 (NCBI) are 2921215 to 2921771 for the *HG* locus (PCR-A del) and 2974721 to 2975264 for the *AK* locus (PCR-C solo).

The conditions for PCR for all primer pairs were the same. Each PCR assay was performed in 10 µl aliquots using 2 pmol of each primer (200 nmol/l), 1 ng of genomic *DNA*, 0.25 units of TaKaRa LA *Taq* polymerase, 0.08 µl of dNTP mixture (2.5 mM each), and 5 µl of 2×GC reaction buffer 1 with 5 mM $MgCl_2$ purchased from TaKaRa, Shiga, Japan. The PCR was performed in eight strips of 0.2-ml thin-walled PCR tubes (QSP) using a GenAmp 9700 thermal cycler (Applied Biosystems) programmed for 35 cycles with a denaturation (96°C—30 s) and annealing (62°C—3 min) step at each cycle. The reaction products were stained with ethidium bromide, and sizes were compared with molecular size markers by horizontal gel electrophoresis in 2% agarose using tris–borate–EDTA running buffer. Control samples (without *DNA* template) were run to ensure there was no amplification of contaminating *DNA*. Reference control *DNA* from the COX and PGF cell lines (Table 1) were used for each PCR run to verify the structural polymorphisms.

Homozygous PCR products of the *sMER9* and *3′ MER9* sequences amplified from the 55 *DNA* samples (Table 1) were sequenced directly with BigDye terminator Cycle Sequencing FS Ready Reaction Kit Ver. 3.1 (Applied Biosystems, Foster City, CA, USA) according to the instructions provided by the manufacturer using the sense and antisense PCR primers as sequencing primers. The sequences were analyzed using an automated *DNA* sequencer (ABI PRISMTM 3130 *DNA* Sequencer; Applied Biosystems). The *MER9* nucleotide sequences were amplified from the *DNA* samples in Table 1, sequenced directly

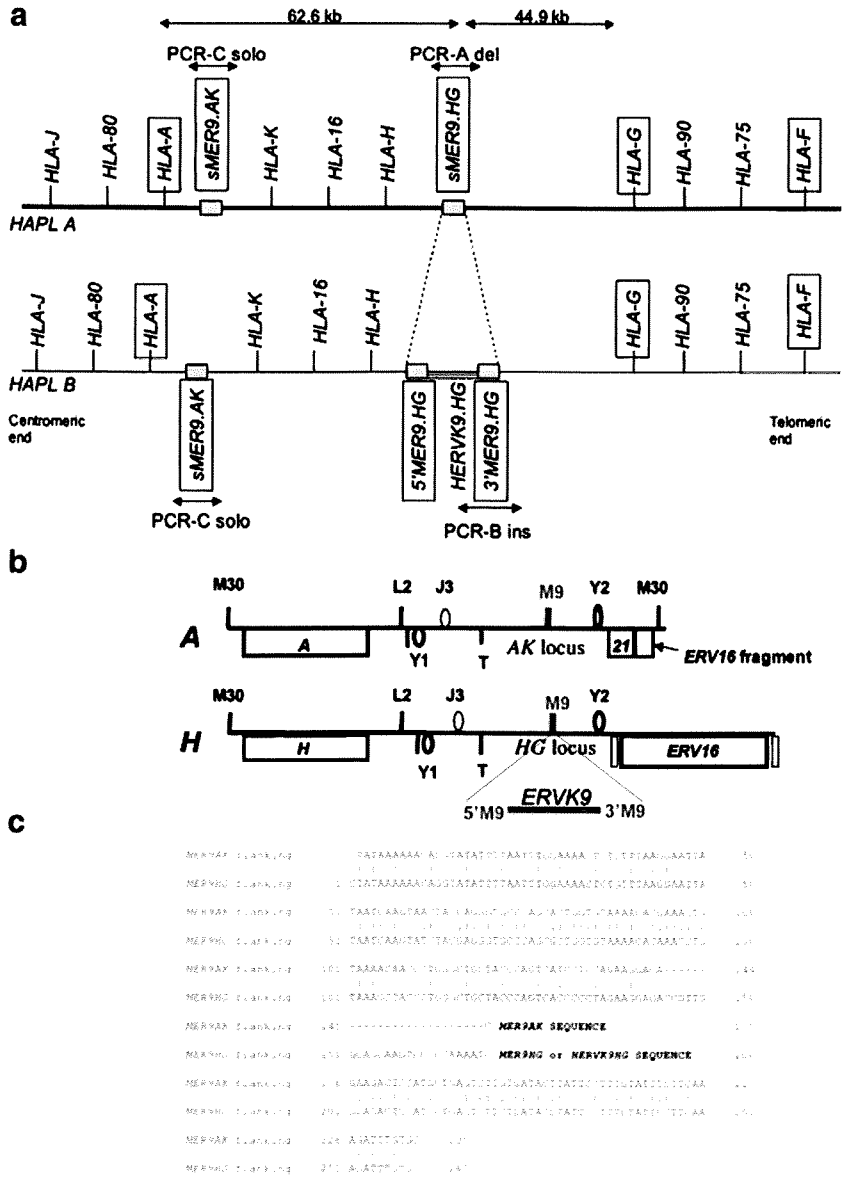


Fig. 1 Location of *MER9-LTR* duplication products at the paralogous loci *AK* and *HG*. **a** Genomic map of the *solo* (*s*), *5′* and *3′ MER9* sequences within the *HLA* class I gene clusters (*HLA-J* to *HLA-F*) of the *alpha* block of two different *MHC* haplotypes *HAPL A* and *HAPL B*. The *double horizontal arrows* labeled *PCR-A del*, *PCR-B ins*, and *PCR-C solo* indicate the *MER9 DNA* regions that were amplified by PCR for sequencing. The *HLA* gene positions are not drawn to scale. The *double horizontal arrow at the top of the figure* indicates the approximate distance between the *HERVK9.HG* deletion at the *HG* locus and the *HLA-A* locus. The distance between the *HLA-G* locus and *MER9.HG* locus is approximately 44.9 kb. **b** Genomic map shows the location of the *MER9-LTR* and *HERVK9* within the *MHC alpha* block duplicons *A* (*top*) and *H* (*bottom*). The *ERVK9*, *MER9* (*M9*), and retroelements are shown relative to the locations of the *HLA* class I genes labeled A, H, and 21 (*shaded boxes*) and *ERV16* sequences (*open boxes*). *Alu* elements are represented by the oval symbols where the letter *J* is *AluJ*, *S* is *AluS*, and *Y* is *Alu*Y, and the number following the *Alu* subfamily designation indicates its paralogy. The lines labeled *M9* and *M30* along the *horizontal line* represent the retroelements *MER9* and *MER30*, respectively. *T* and *L2* represent a fragmented *THE* sequence and *LINE2* sequence, respectively. The *symbols above the horizontal line* represent sequences in the 5′ to 3′ orientation, whereas those below the line are 3′ to 5′. A variety of retroelements were omitted from the map for convenience of presentation. The presence and absence of the internal *ERVK9* sequence within the duplicon *H* is indicated by the *labeled triangle* below the M9 of duplicon *H*. The figure is modified from Kulski et al. (2005). **c** *DNA* sequence alignment of nucleotides flanking the 5′ and 3′ ends of the *MER9-LTRs* or *HERVK9* sequences at the *AK* and *HG* loci. The 15 (7.3%) nucleotide differences in 205 of the aligned nucleotides, excluding the gaps, represent 20.4 Myrs of sequence divergence between the two duplicons

in both the 5′ and 3′ directions and submitted online via Sakura to DDBJ (http://www.ddbj.nig.ac.jp/). The sequences were assigned the accession numbers AB443932 to AB443937 and AB447373 to AB447385 in the DDBJ/EMBL/GenBank nucleotide sequence databases. The accession number given for the orangutan (*Popy*) *MER9/HERVK9* sequence at the *HG* locus of the CHORI-253 BAC library was AB453920.

MER9-LTR sequence and phylogenetic analysis

Additional *MER9 DNA* sequences for multiple alignments and for the reconstruction of a phylogenetic tree were obtained by extracting the *MER9-LTR* from genomic sequences available within the public *DNA* databases at DDBJ, EMBL, and GenBank. The NCBI (http://www.ncbi.nlm.nih.gov/) nucleotide accession numbers (cell line, *MHC* class I allele) of previously sequenced *MER9-LTR* within genomic *MHC* sequences of humans (Hampe et al. 1999; Stewart et al. 2004; Horton et al. 2008), chimpanzee (Anzai et al. 2003), and rhesus macaque (Kulski et al. 2004; Shiina et al. 2006) downloaded for analysis were CR388220 (DBB, *HLA-A2*), AL671277 and AL645929 (PGF, *HLA-A3*), AL645935 and AL671561 (COX, *HLA-A1*), BX005091 and BX284699 (SSTO, *HLA-A32*), CR382333 and BX927141 (MANN, *HLA-A29*), AL845454 (QBL, *HLA-A26*), CT009517 (MCF, *HLA-A2*), AF055066 (Hampe, *HLA-A2*), CR847781 (MANN, *HLA-C* locus), AC192848 (*PatrA*, chimpanzee), CU104658 (*GogoA*, gorilla), and AB128049 (*Mamu*, rhesus macaque). The nucleotide positions of the *MER9-LTR* were first located within the genomic sequences of different accession numbers by using RepeatMasker v3.1.6 (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker), and the annotated *MER9-LTR* sequences were then manually extracted from the genomic sequences using the BLAST extraction tool at NCBI (http://www.ncbi.nlm.nih.gov/).

The *MER9 DNA* sequences were aligned by CLUSTALW, and a phylogenetic tree was reconstructed by the maximum parsimony method using the MEGA4 software (Tamura et al. 2007) with the close-neighbor interchange search set to one replication and random addition trees set to ten replications.

Multiple alignments of *MER9 DNA* sequences were also examined using the multiple alignment programs provided by the CLC Free Workbench (http://www.clcbio.com/) and GeneDoc (http://www.nrbsc.org/gfx/genedoc/index.html). Needle, a Needleman–Wunsch algorithm and part of the EMBOSS Pairwise Alignment Algorithms at EMBL-EBI (http://www.ebi.ac.uk/emboss/align/index.html), was used to calculate the percent similarity and to identify the SNP and gap positions between the two *MER9 DNA* sequences as required.

Results

Location and PCR analysis of the *MER9-LTR* structural polymorphisms

The locations of the *MER9-LTR* sequences at the duplicated *AK* and *HG* loci within the *alpha* block of the *MHC* class I region are shown in Fig. 1. The genome coordinates for the *AK* and *HG* loci on the human genome view, build 36.3, are 30,011.9 to 30,012.45 and 29,951.3 to 29,957.8 kbp, respectively, at NCBI (http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9606). The *AK* locus appears to only have the *sMER9-LTR* and not the *5′* or *3′ MER9-LTR* indicative of the *HERVK9* insertion, whereas the *HG* locus has either the 5′ and 3′ LTRs flanking the internal *HERVK9*. *HG* insertion or the *sMER9-LTR* due to the deletion of the *HERVK9* internal sequence at the *HG* locus (Kulski et al. 2005). Thus, the *MER9-LTR* at the *HG* locus is a structural polymorphism in that the *3′ MER9-LTR*, or the *solo MER9-LTR* are structural alleles that are absent and/or present within an individual and vary in frequency between different populations (Kulski et al. 2008). The *MER9-LTR* at the *AK* and *HG* loci are probably duplicated or paralogous products because they share the same insertion site with the same or similar nucleotide sequences flanking their insertion site (Figs. 1b and c). Apart from a few nucleotide differences between the two paralogous sites, there is a 25-nucleotide indel that is located immediately upstream of the 5′ end of the *MER9-LTR* insertion.

The PCR results for the absence or presence of the *MER9-LTR* at the *AK* and *HG* loci in 55 *DNA* samples that are homozygous (48 samples) or heterozygous (seven samples) for different *HLA-A* alleles are shown in Table 1. The PCR primer sets amplified all the *DNA* samples except those at the *AK* locus for five *DNA* samples with the homozygous HLA-A24 alleles. The failure of the PCR amplification at the *AK* locus however, is consistent with a 50-kb genomic deletion near the *HLA-A* gene of the *HLA-A24* haplotypes (Watanabe et al. 1997), which appears to include the *MER9-LTR AK* locus, but not the *MER9-LTR HG* locus. PCR detected the 20 homozygous *3′ MER9.HG* sequences that were linked with the homozygous *HLA-A3* (seven samples), *-A11* (one sample), *-A24* (five samples), *-A30* (one sample), *-A31* (two samples), and *-A33* (four samples) and the 28 homozygous *sMER9.HG* sequences that were linked to *HLA-A1* (ten samples), *-A2* (14 samples), *-A26* (one sample), *-A29* (one sample), and *-A34* (two samples).

MER9-LTR SNPs and *HLA-A* haplotypes

MER9 PCR products obtained from the *DNA* of the cell lines listed in Table 1 were sequenced directly two or three