Ryo Yamada [1] Hiroto Kawakami [2] Masao Yamaguchi [3]
Eri Tatsu [1] Akihiro Sekine [4] Kazuhiko Yamamoto [1]
Yusuke Nakamura [4] Tatsuhiko Tsunoda [3]

[1] Laboratory for Rheumatic Diseases, SNP Research Center, RIKEN, Tokyo, JAPAN

[2] Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, JAPAN

[3] Laboratory for Medical Informatics, SNP Research Center, RIKEN, Tokyo, JAPAN

[4] Laboratory for Genotyping, SNP Research Center, RIKEN, Tokyo, JAPAN

Correspondence to Ryo Yamada, ryamada-tky@umin.ac.jp

Acknowlegments

We Thank Hisanori Haga [2] for discussing on presentation materials in detail.

Key Words

- Population Structure
- Multiple Genetic Markers
- SNP
- Spurious Association
- Genomewide Association Study
- Markov Chain Monte Carlo Simulation
- Japanese

Abstract

Linkage disequilibrium (LD) mapping, using a dense map of single nucleotide polymorphisms (SNPs), has been advocated as the method of choice to locate loci of complex genetic diseases. However it has also been suggested that genetic population structure could lead to many spurious associations between genetic markers and a disease phenotype when LD mapping was performed, especially in case of studies using samples from ethnically structured populations.

Japanese population has been considered to be less structured, and, therefore, more appropriate for LD mapping than many other populations in the world, because of its ethnic isolation from both historical and geographical standpoints. However no analysis on their genetic homogeneity based on assayed genotype data has been reported. We evaluated homogeneity of two sub-populations, each of which were from one of the two largest Japanese metropolises by analyzing genotype data of 188 individuals for 303 unlinked SNPs on autosomal chromosomes. We adopted two analytical methods to evaluate structure. One method was to assess existence of difference in population structure between two sub-populations, by analyzing chi square statistics computed for multiple contingency tables constructed for allele distribution of each SNP between sampling sub-populations. The other was to infer population structure by model-based clustering method using Markov Chain Monte Carlo algorithm. We found that, if any, only subtle difference in population structure between two sub-populations existed. The possible allele difference between two sub-populations seemed small enough not to severely interfere case-control association studies for complex genetic traits.

Organization of This Presentation

- Six sets of simulated genotype data of two sub-populations with various degree of difference in allele frequencies of SNPs were analyzed with two methods for genetic structure.
- Real genotype data of two sub-populations in Japan were analyzed in the same way.
- The results of real data were compared with the results of simulated data.

Methods

Construction of genotype data for the structural analysis of simulated data

- Suppositions:
  - Two sub-populations were supposed, each of which had randomly mated and were homogeneous.
  - 180 SNPs were supposed, all of which were:
    - Biallelic.
    - In Hardy-Weinberg equilibrium.
    - Unlinked each other.
  - Genotypes of individuals, consisting of genotype data sets, were assigned as if individuals were randomly sampled from each sub-population.
  - Fifty individuals were sampled for each sub-population.
- Six data sets were constructed with variations in allele frequencies of a part of SNPs between two sub-populations.

Methods

Allele frequencies of SNPs in two sub-populations for 6 sets of simulated data (1)

Table 1 Number of SNPs and their allele frequencies in 2 sub-populations.

	No. SNPs	Sub-population 1	Sub-population 2	Difference of allele frequencies between sub-populations	Average allele frequency
	20	0.2	0.2	0	0.2
	20	0.3	0.3	0	0.3
	20	0.4	0.4	0	0.4
	10	0.2－Δa×0.5*	0.2＋Δa×0.5	-Δa	0.2
	10	0.2＋Δa×0.5	0.2－Δa×0.5	Δa	0.2
	10	0.3－Δa×0.5	0.3＋Δa×0.5	-Δa	0.3
	10	0.3＋Δa×0.5	0.3－Δa×0.5	Δa	0.3
	10	0.4－Δa×0.5	0.4＋Δa×0.5	-Δa	0.4
	10	0.4＋Δa×0.5	0.4－Δa×0.5	Δa	0.4
	10	0.2－Δb×0.5&	0.2＋Δb×0.5	-Δb	0.2
	10	0.2＋Δb×0.5	0.2－Δb×0.5	Δb	0.2
	10	0.3－Δb×0.5	0.3＋Δb×0.5	-Δb	0.3
	10	0.3＋Δb×0.5	0.3－Δb×0.5	Δb	0.3
	10	0.4－Δb×0.5	0.4＋Δb×0.5	-Δb	0.4
	10	0.4＋Δb×0.5	0.4－Δb×0.5	Δb	0.4
total	180	-	-		

Δa* and Δb & are parameters to make difference of allele frequency between sub-populations 1 and 2.
Δa and Δb are specified for each set of simulation on Table2

Allele frequencies of SNPs in two sub-populations for 6 sets of simulated data (2)

Table 2 Parameters to give allele frequency difference between sub-populations 1 and 2

Name of simulation set	Δa	Δb
1	0.025	0.025
2	0.025	0.05
3	0.05	0.1
4	0.1	0.1
5	0.05	0.15
6	0.1	0.2

Methods

Analytical Methods of Structure (1)

- Analytical Method 1

  Evaluation of sum of multiple chi statistics calculated for individual SNP[1]

- Analytical Method 2

  Inference of structure by Markov Chain Monte Carlo simulation method[2]

  – [1] Jonathan K. Prichard and Noah A. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 65: 220-228,1999

  – [2] Jonathan K. Prichard, Matthew Stephens and Peter Donnelly. Inference of population structure using multilocus genotype data. Genetics. 155: 945-959, 2000

Methods

Analytical Method 1

- Principles of sum of multiple chi statistics calculated for individual SNP
  a. Null hypothesis and alternative hypothesis
     - Null Hypothesis: All the samples are from an identical population.
     - Alternative hypothesis: Samples are from two distinct sub-populations.
  b. Chi square value for each SNP genotype data
     - χ_i^2 (i=1,2,…N; N:Number of SNPs analysed) represents chi square value calculated for 2×2 contingency table of the observed number of alleles from two sub-populations.
  c. Sum of multiple chi square value(S) and degree of freedom(df)
     - $S = \sum_{i=1}^{N} \chi_i^2$
     - $df = N$
     - S is statistically evaluated as chi square value with degree of freedom being df.

Methods

Analytical Method 2

- Inference of structure by Markov Chain Monte Carlo (MCMC) simulation method*
  a. Parameter matrices were set as below:
     i. X: Genotypes of the sampled individuals
     ii. Z: Belonging sub-populations of the individuals
        – Number of sub-populations : Two
     iii. P: Allele frequencies of SNPs in all sub-populations
  b. Belonging sub-population was assigned to each individual at random at the beginning.
  c. MCMC algorithm was applied and converged result was obtained.
     i. Sample $P^{(m)}$ from $Pr(P|X,Z^{(m-1)})$#.
        • Beta distribution was used for allele frequency distribution of biallelic markers.
        • Gibbs sampler with adaptive rejection sampling was adopted.
     ii. Sample $Z^{(m)}$ from $Pr(Z|X,P^{(m)})$.
  d. Results of multiple runs were summed, considering phenomenon to converge into symmetrical modes.

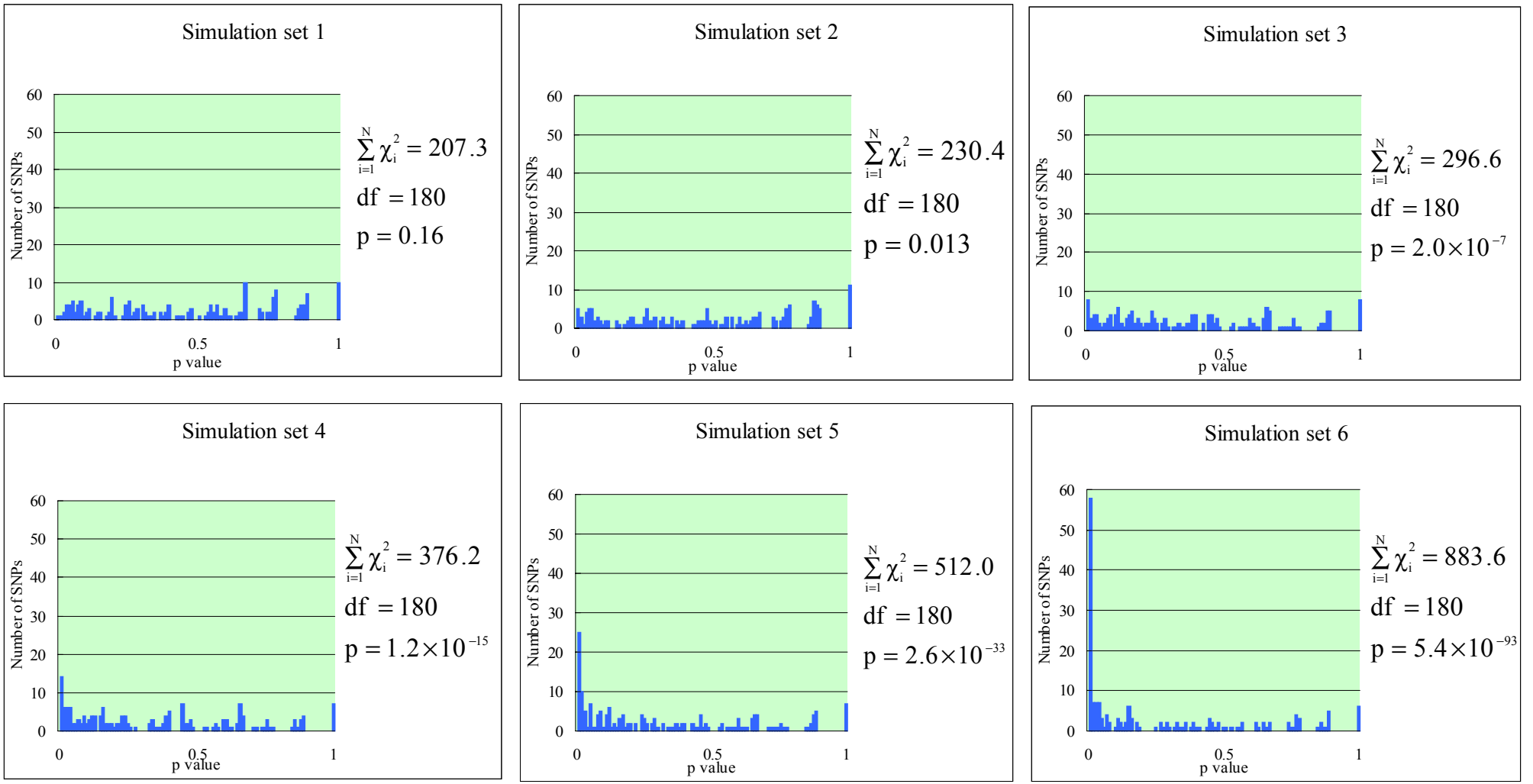
* Program source will be available in near future on request by the author.

m denotes number of iterations and Pr(Y|W) denotes conditional probability distribution of Y when W.

Results

Results of analysis 1(Sum of Chi Square Statistics) of simulated data

Fig1 p value distribution of individual SNPs and sum of chi square values and their corresponding p value



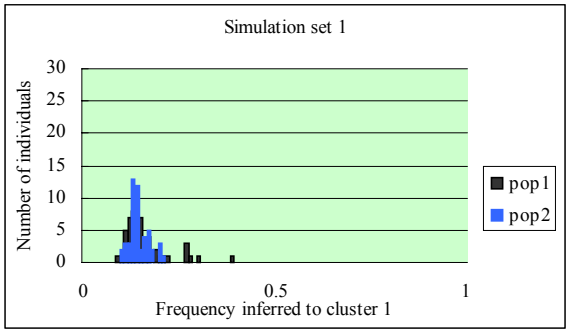
6 histograms represents result of 6 simulated data,respectively.

Null hypothesis in set 1 was not rejected. For sets 3, 4, 5 and 6 null hypothesis was strongly rejected.

Results

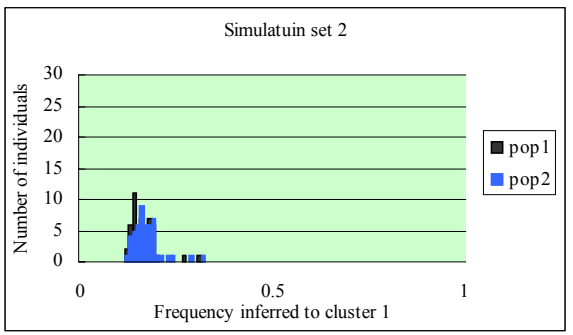
Results of analysis 2 (MCMC) of simulated data

Fig2 Histograms showing how often each individual was inferred to cluster 1



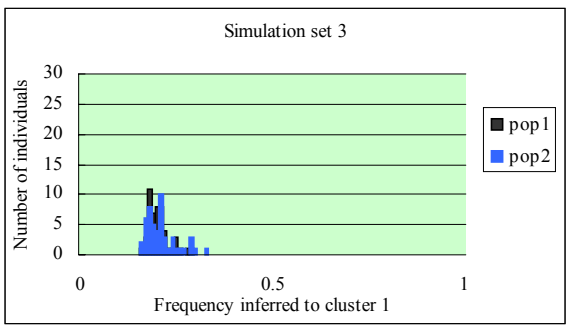
pop1
average 0.15
variance 0.011

pop2
average 0.14
variance 0.001



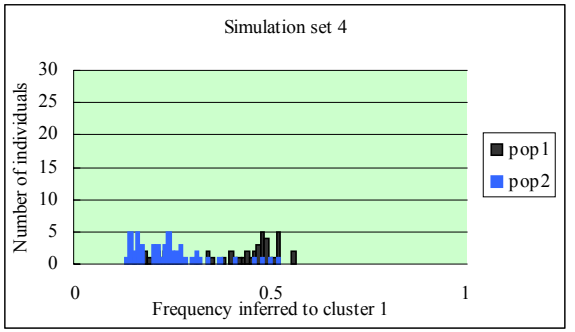
pop1
average 0.16
variance 0.001

pop2
average 0.17
variance 0.001



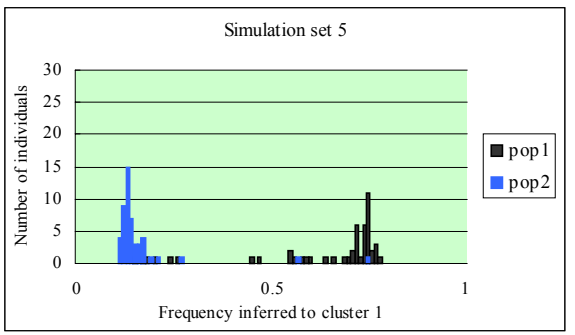
pop1
average 0.20
variance 0.001

pop2
average 0.20
variance 0.002



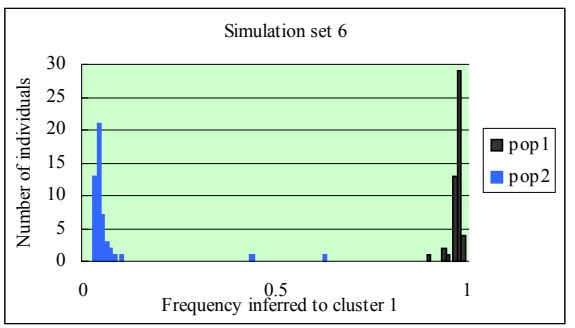
pop1
average 0.39
variance 0.014

pop2
average 0.24
variance 0.009



pop1
average 0.63
variance 0.033

pop2
average 0.16
variance 0.012



pop1
average 0.97
variance 0.000

pop2
average 0.06
variance 0.010

Distributions of sub-populations 1 and 2 of simulation sets 1, 2 and 3 appeared to be similar. Simulation sets 4, 5 and 6 segregated two sub-populations with appropriate progression of clarity of discrimination.

Methods

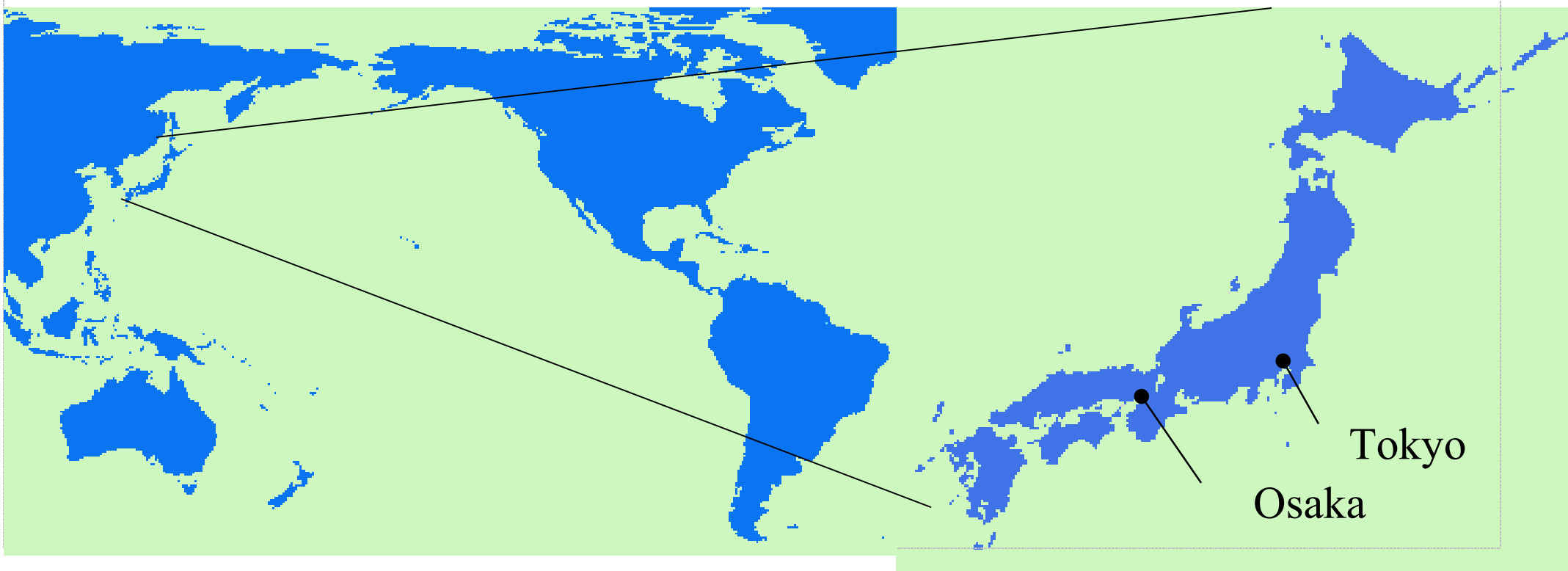
Samples(1)

Construction of genotype data for the structural analysis of real data

Tokyo: Sampling Metropolis 1

Osaka: Sampling Metropolis 2

Fig3 Location of Tokyo and Osaka



Methods

Samples(2)

Construction of genotype data for the structural analysis of real data

- Historical, geographical and ethnical information of sampling sub-populations

  The land of Japan consists of 4 major islands and many small islands, locating at the rim of northwestern Pacific ocean.

  Japanese ancestry is considered to be formed by migration of at least a few ethnic groups from the Eurasian continent before the end of the last glacial epoch, admixed by the Pacific islanders.

  No major migration or admixture happened for the last few thousands years, forming a strongly culturally distinct population with a limited degree of ethnic diversity.

  Tokyo and Osaka metropolitan areas are characterized by high-degree of influxes of people from diverse areas of Japan, followed by random mating, which is believed to have canceled genetic structure present in the past.

  Major differences between Tokyo and Osaka areas:

  - Population of Tokyo area has been formed by influx of people from almost all the areas of Japan for the last 4 hundreds years without a core ancestral population.
  - Population of Osaka area has been formed by admixture of many people from many areas of Japan, especially of south-western area, to the core ancestral population, which had been formed before early middle ages.

Methods

SNPs(1)

Construction of genotype data for the structural analysis of real data

- Source of SNPs:
  - SNPs on autosomal chromosomes from IMS-JST SNPs database, discovered by direct-sequencing method of multiple Japanese individuals'genomic DNAs targeting mainly on and around known and expected genes throughout the human genome. (http://snp.ims.u-tokyo.ac.jp/)
- Number of SNPs:
  - 4373 SNPs were genotyped initially.
  - 303 of 4373 SNPs were adopted for structure analyses by criteria described on the following sheet of paper.

Methods

SNPs(2)

Construction of genotype data for the structural analysis of real data

- SNPs were selected for analyses of population structure by the criteria as below:
  - Each SNP should:
    - Be assayed with appropriate call for more than 170 individuals.
    - Be biallelic.
    - Have minor allele frequency more than 0.1.
    - Be unlinked each other:
      - » Two SNPs on an identical chromosome were considered to be unlinked each other when linkage disequilibrium (LD) index D'* were less than 0.3

        *D': Index of LD was calculated as below:

        1. Haplotype frequencies were estimated by EM-algorithm
        2. $D' = |(P_{AB}^{**} \times P_{ab} - P_{ab} \times P_{aB})| / \text{Minimum}((P_{AB}+P_{aB}) \times (P_{aB}+P_{ab}), ((P_{AB}+P_{Ab}) \times (P_{Ab}+P_{ab}))$,

           AB, Ab, aB, and ab represent 4 haplotypes created by 2 SNPs.

           ** P_{cd} represents frequency of haplotype.
    - Be successfully assayed. Our internal criteria # was adopted for judgment of success of assays:

      # It based on the data in SRC, RIKEN, TOKYO, JAPAN (unpublished):

      - » Intensity of fluorescence signals should be more than a threshold.
      - » More than 90 % of samples should be called.
      - » Chi square statistic for Hardy-Weinberg Equilibrium should give p-value of larger than 1.0×10^{-3}.

Methods

Genotyping Assay

Construction of genotype data for the structural analysis of real data

- Invader assay with multiplex polymerase chain reaction (PCR), characterized as below:
  - Multiplex PCR was performed against 100 genomic regions simultaneously in a single reaction tube.
  - Each amplified fragment contained at least one SNP.
  - A specialized 384-well card system was adopted for the ultra high-throughput genotyping system.
  - Very small amount of genomic DNA was required for genotyping:
    - 40 ng of genomic DNA was used as a template for each multiplex PCR, meaning only 0.4 ng of genomic DNA per genotyping of one SNP.

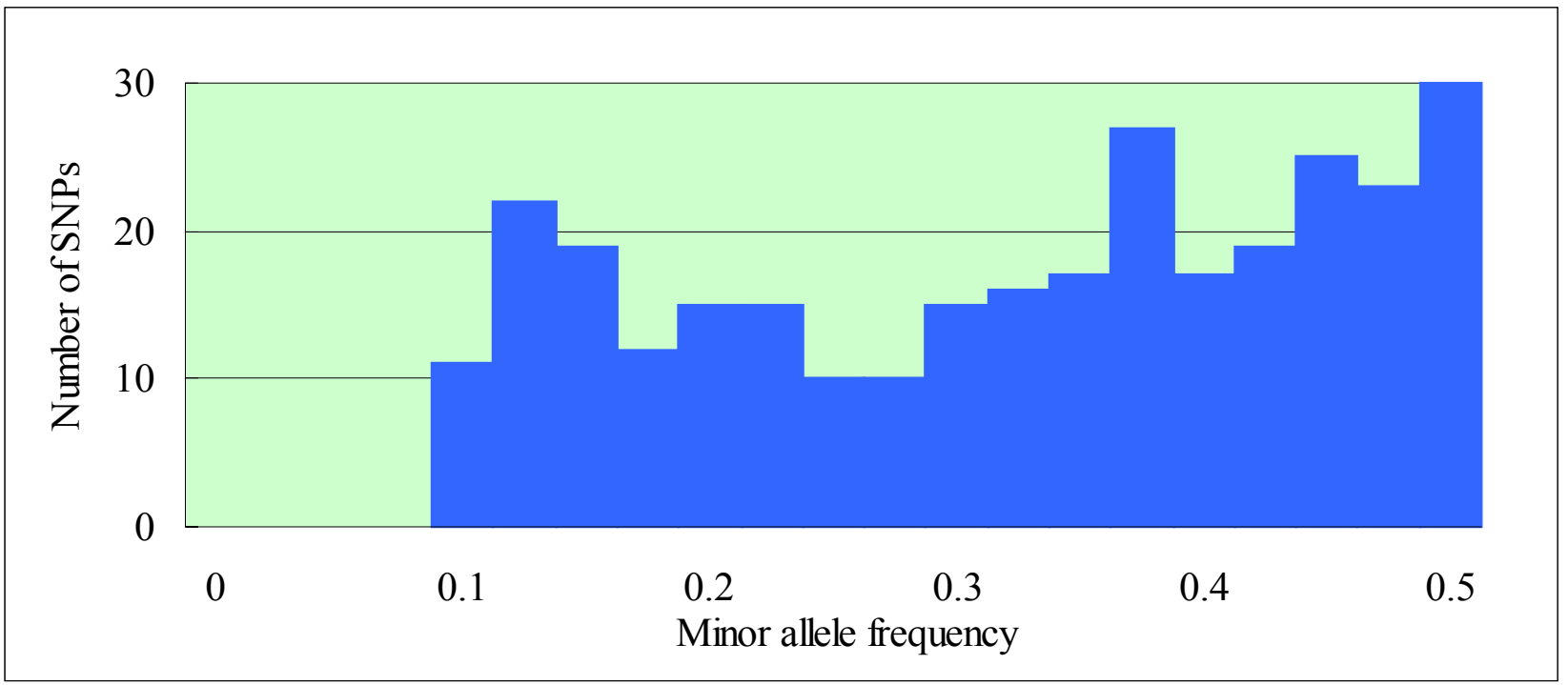
A high-throughput SNP typing system for genome-wide association studies.
J Hum Genet. 2001;46(8):471-7.
Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y.

Results

Characterization of selected SNPs of real data

- Characterization of SNPs included for analyses of population structure
  - 303 SNPs out of 4373 were selected.
  - Average rate of successful call of genotyping of 303 SNPs was 0.99
  - Minor allele frequency of each SNP was distributed as below:

Fig4 Distribution of minor allele frequency of 303 SNPs

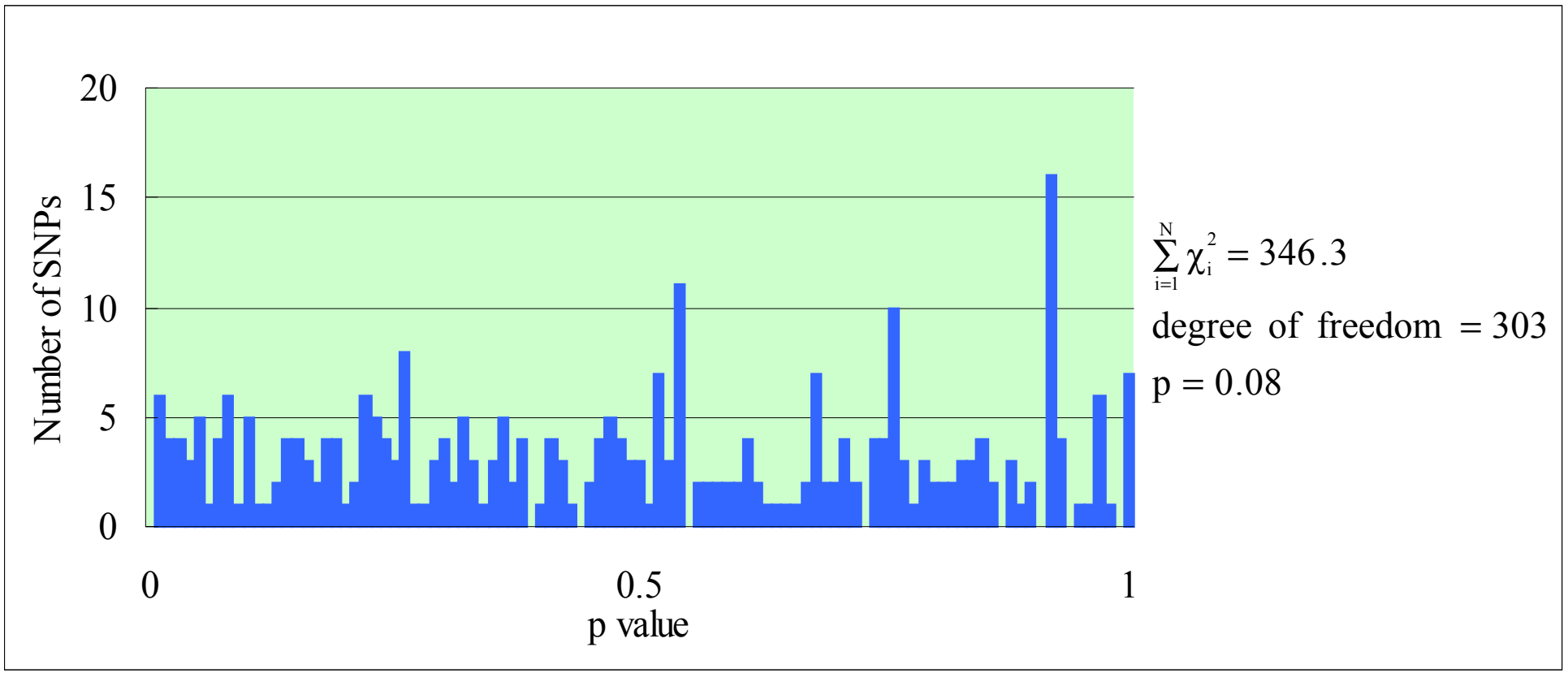


SNPs with higher minor allele frequency seemed to be included more frequently than the distribution in reality.

Results

Result of analysis 1(Sum of Chi Square Statistics) of real data

Fig5 p value distribution of 303 chi square tests and sum of chi square values and its corresponding p value with degree of freedom of 303

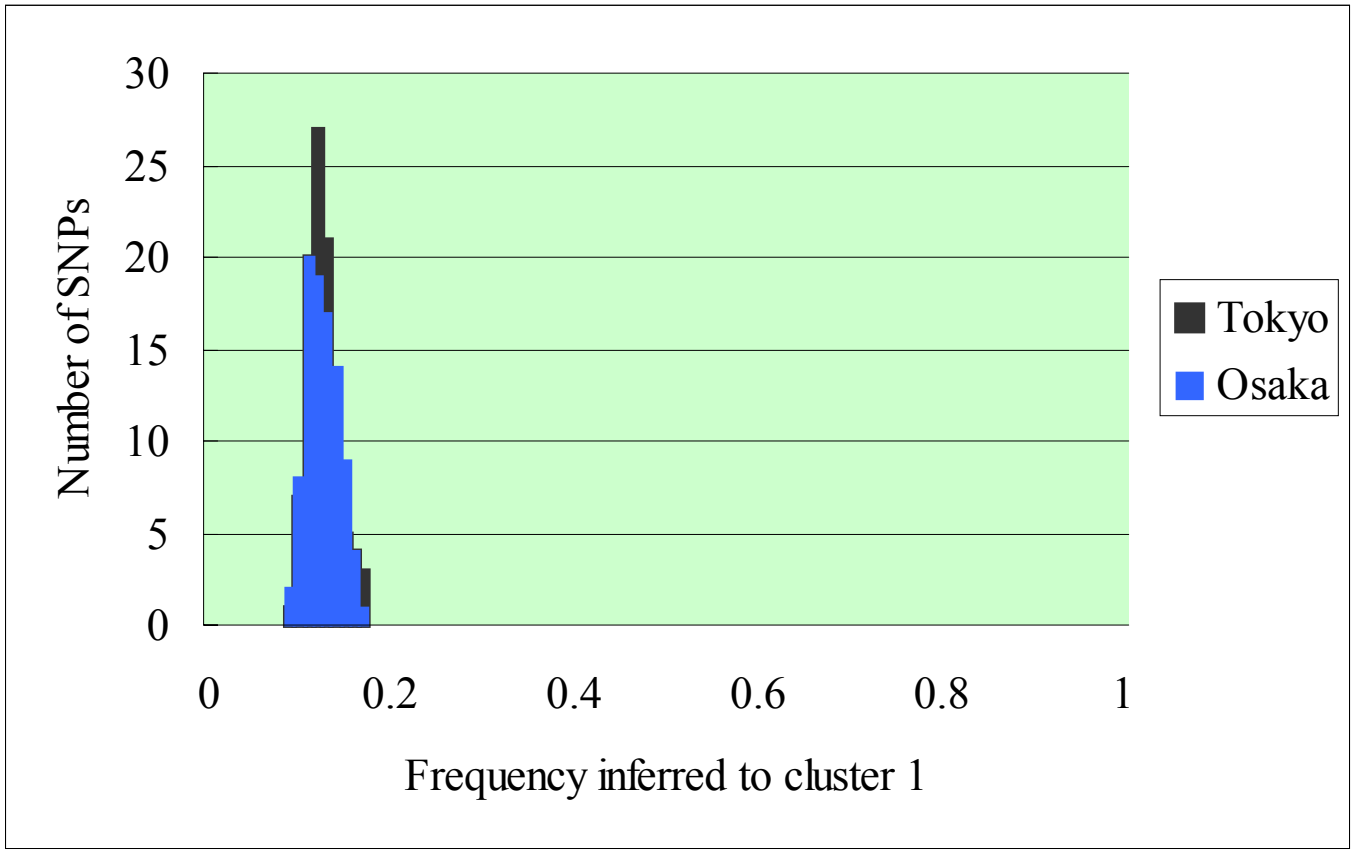


Null hypothesis that Tokyo and Osaka sub-populations were identical was not rejected with cut off p-value of 0.05.

Results

Result of analysis 2(MCMC) of real data

Fig6 Histogram of frequency that each individual was inferred to cluster 1



Tokyo		
	avarage	0.12
	variance	0.0003
Osaka		
	avarage	0.12
	variance	0.0003

Almost identical distributions were obtained for 2 sampling areas.

These distributions were similar to one of simulation sets 1, 2 and 3.

It suggested the difference between Tokyo and Osaka was as small as difference between two sub-populations in simulation sets 1, 2 or 3.

Discussion

The two analytical methods were applied to simulated genotype data and real data, both of which represented subtly heterogeneous two sub-populations. Result of analyses of simulated data revealed both analytical methods could dissect two populations when SNPs with allele frequency difference between populations more than 0.1 existed. Analysis 1 (sum-of-chi square values) showed statistical significance, even when SNPs with allele frequency difference of 0.05 dominated, if such SNPs existed frequently, although analysis 2 (MCMC) did not show visible discrimination in the distributions. In a situation, where population-based sampling association studies were carried out, difference of allele frequencies among subtly stratified populations would be negligible, as far as the difference between ethnic sub-populations was smaller than the difference between phenotypic sub-groups. The comparison between results of simulated data and real data gave an idea that the genetic structure between Tokyo and Osaka was somewhere between a simulated data with maximum allele frequency difference of 0.025 and another set with maximum allele frequency difference of 0.05. When complex genetic traits are analyzed with LD mapping, allele frequency difference between phenotypically distinct groups distributes most likely more than 0.05. From this standpoint, samples from Japanese metropolises should be treated as homogeneous.