

Validation of genotype imputation in Southeast Asian populations and the effect of single nucleotide polymorphism annotation on imputation outcome

Worachart Lert-Itthiporn, Bhoom Suktitipat, Harald Grove, Anavaj Sakuntabhai, Prida Malasit, Nattaya Tangthawornchaikul, Fumihiko Matsuda, Prapat Suriyaphol

▶ **To cite this version:**

Worachart Lert-Itthiporn, Bhoom Suktitipat, Harald Grove, Anavaj Sakuntabhai, Prida Malasit, et al.. Validation of genotype imputation in Southeast Asian populations and the effect of single nucleotide polymorphism annotation on imputation outcome. BMC Medical Genetics, 2018, 19 (1), pp.23. 10.1186/s12881-018-0534-8 . pasteur-02003873

HAL Id: pasteur-02003873

https://pasteur.hal.science/pasteur-02003873

Submitted on 25 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Lert-itthiporn et al. BMC Medical Genetics (2018) 19:23
DOI 10.1186/s12881-018-0534-8

BMC Medical Genetics

RESEARCH ARTICLE Open Access

Validation of genotype imputation in Southeast Asian populations and the effect of single nucleotide polymorphism annotation on imputation outcome

Worachart Lert-itthiporn[1,2], Bhoom Suktitipat[3,4], Harald Grove[2,4], Anavaj Sakuntabhai[5,6,7], Prida Malasit[8,9], Nattaya Tangthawornchaikul[8,9], Fumihiko Matsuda[10] and Prapat Suriyaphol[2,4*]

Abstract

Background: Imputation involves the inference of untyped single nucleotide polymorphisms (SNPs) in genome-wide association studies. The haplotypic reference of choice for imputation in Southeast Asian populations is unclear. Moreover, the influence of SNP annotation on imputation results has not been examined.

Methods: This study was divided into two parts. In the first part, we applied imputation to genotyped SNPs from Southeast Asian populations from the Pan-Asian SNP database. Five percent of the total SNPs were removed. The remaining SNPs were applied to imputation with IMPUTE2. The imputed outcomes were verified with the removed SNPs. We compared imputation references from Chinese and Japanese haplotypes from the HapMap phase II (HMII) and the complete set of haplotypes from the 1000 Genomes Project (1000G). The second part was imputation accuracy and yield in Thai patient dataset. Half of the autosomal SNPs was removed to create *Set 1*. Another dataset, *Set 2*, was then created where we switched which half of the SNPs were removed. Both *Set 1* and *Set 2* were imputed with HMII to create a complete imputed SNPs dataset. The dataset was used to validate association testing, SNPs annotation and imputation outcome.

Results: The accuracy was highest for all populations when using the HMII reference, but at the cost of a lower yield. Thai genotypes showed the highest accuracy over other populations in both HMII and 1000G panels, although accuracy and yield varied across chromosomes. Imputation was tested in a clinical dataset to compare accuracy in gene-related regions, and coding regions were found to have a higher accuracy and yield.

Conclusions: This work provides the first evidence of imputation reference selection for Southeast Asian studies and highlights the effects of SNP locations respective to genes on imputation outcome. Researchers will need to consider the trade-off between accuracy and yield in future imputation studies.

Keywords: Imputation, Reference, Genotype, Pan-Asian SNP, SNP annotation

* Correspondence: prapat.sur@mahidol.ac.th
[2]Division of Bioinformatics and Data Management for Research, Department of Research and Development, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand
[4]Center of Excellence in Bioinformatics and Clinical Data Management, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand
Full list of author information is available at the end of the article

BioMed Central

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Genome-wide association studies (GWAS) have been widely used as a reliable method for identifying genetic variants associated with a trait or complex disease. A high density of SNPs increases the chance of finding either a causal mutation for the trait or SNPs close enough to the mutation to confidently suggest a gene or another sequence feature underlying the trait. One way to overcome this problem is using imputation, a process in which samples are genotyped using a low-density SNP array and imputed with information from a reference panel genotyped on a high-density SNP array. This method will also recover genotypes that are missing because of technical issues.

Imputation has successfully helped to identify genetic susceptibilities to various diseases and phenotypes that were not recognized in a genotyped panel [1, 2]. The method relies on the number of SNPs being shared between the two panels and the amount of linkage disequilibrium (LD) between genotyped and non-genotyped SNPs [3]. A low average LD will reduce the accuracy and might require more typed SNPs. The quality of imputation also depends on the choice of reference [1]. If the reference contains genetic variants not present in the actual sample population, it will increase the noise in the data and reduce the usefulness of the imputation. One study of malaria resistance in Gambian children only identified a previously known hemoglobin S variant in the hemoglobin-β gene when a Gambian-specific reference was used [1]. Although this problem is more likely to occur in Africa, where there is a considerably lower LD compared to Europe and Asia [4], determining how to choose the best reference is relevant for any study performing imputation with publicly available reference sets.

Many studies have validated the accuracy and reliability of imputation [5–7], but most of these studies focused on populations of European descent [5, 7]. One study showed that the accuracy of using a publicly available database varied across human populations with Europeans having the highest accuracy and Africans having the lowest [6]. Because Asian populations have some unique genetic characteristics [8], it is not always possible to directly adapt information about genetics or genomics from studies in Caucasian populations [9].

Several types of software are currently available for performing genotype imputation [10–13]. Similarly, many publicly available genetics databases are accessible for public use [14, 15]. One of these is the Pan-Asian SNP genotyping database (PanSNPdb), which collects SNPs and copy number variations from 1719 samples in 71 populations from China, India, Indonesia, Japan, Malaysia, the Philippines, Singapore, South Korea, Taiwan, and Thailand [16, 17]. The genotyping process was performed using the Affymetrix GeneChip Human Mapping 50 K Xba Array.

Most of the studies on imputation have looked at the overall outcome of all SNPs [5, 18], and a few have focused on a particular region within a gene, not the whole genome [19, 20]. We proposed two objectives for the current study. The first was to identify the most preferred reference for imputation in Southeast Asian populations. Using two publicly available haplotype databases, the International HapMap Project (HMII) and the 1000 Genomes project (1000G), we compared the accuracy and yield of imputation in several Southeast Asian populations. Additionally, we looked at imputed results using genotyped samples from a study of a Thai genome cohort. The second objective was to evaluate the imputation results of different regions in the human genome using a real dataset from the Thai dengue study as a model. This is the first extensive study of imputation in Southeast Asian populations and the first illustration of imputation differences between SNPs in different regions of the genome.

Methods

This study was divided into two parts. The first part aimed at showing the difference in imputation accuracy by using different criteria for selecting a reference database. Additionally, using data from populations within the Southeast Asian region illustrated the variation in accuracy when going from one population to another. The second part used real genotype data in all autosomes to classify SNPs into different groups according to their location within genes. Imputation accuracy, GWAS significance and allele frequency were then correlated with the classification.

Sample datasets

We performed the first part of our analysis using data from PanSNPdb [16]. To illustrate the imputation accuracy in Southeast Asian populations, we selected all available samples from Indonesia (ID, $n = 288$), Malaysia (MY, $n = 217$), the Philippines (PI, $n = 125$), Singapore (SG, $n = 90$), and Thailand (TH, $n = 245$). Only SNPs that were polymorphic in all populations were used in this study ($n = 52{,}160$).

The second part was imputation accuracy and yield in a patient dataset in which we had access to phenotypes because the phenotypes allowed us to observe the effect of imputation on subsequent association tests. The subjects were 609 Thai dengue patients who were 1–15 years-old from Siriraj, Ramathibodi, and Khon Kaen hospitals. A total of 468,987 SNPs from Illumina Human Hap610 array (Illumina Inc., San Diego, CA) passed the quality control requirements (QC). The accuracy of imputation was tested for each SNP from the dengue dataset by first randomly choosing half of the SNPs from the genotyping panel to create a mutually exclusive set of SNPs: *Set 1* and *Set 2*. Then, *Set 1* SNPs were used to impute *Set 2* to create a complete SNP panel. *Set 2* was

also used to impute *Set 1* to create a complete SNP panel. Based on our results in the previous section, HMII data were used as a reference for imputing the SNPs from the dengue dataset. The total number of SNPs after imputation was 1,417,081. Post-imputation QC reduced these numbers to 858,480.

Quality control and multidimensional scaling

For all of the sample datasets, QC was performed in PLINK v1.07 [12] using standard procedures for GWAS [21]. We included all markers with a call rate > 0.95, a minor allele frequency (MAF) > 0.01, and a Hardy–Weinberg equilibrium (HWE) $> 10^{-7}$. Samples with call rates < 0.95 were excluded from the analysis along with samples that had first-degree relationship agreement, as evaluated by expected IBD sharing in PLINK v1.07. Multidimensional scaling (MDS) of Southeast Asian populations from PanSNPdb was performed in PLINK v1.07. This method allowed for visualization of principle components in the admixed population. Plotting of the MDS was conducted in R version 3.0.2 (http://www.r-project.org/).

Imputation procedure

In the first part of the study, each population from PanSNPdb was analyzed independently. Five percent of SNPs were randomly selected and removed. The same SNPs set of the removed SNPs were applied to all populations. SHAPEIT version 2 software was used to pre-phase the SNPs [22]. Imputation was accomplished with IMPUTE2 to recover the removed SNPs [23]. Each population was phased and imputed using both references in turn. According to guidelines from IMPUTE2, we imputed each chromosome separately and used windows of 5 Mb with an additional 250 kb buffer region on both sides of the analysis interval. The options used in the program were -buffer 1000, –iter 30, –burnin 10, and -k 80. The processes for random removal, phasing, and imputation were repeated five times.

The second part of our study used all the autosomal SNPs from the dengue dataset. Half of the autosomal SNPs from Thai dengue patients were removed by every second SNP (*Set 1*). Another dataset (*Set 2*) was then created in which the other half of the SNPs (*Set 1*) were removed. In this way, all SNPs were imputed once. Imputation was performed with HMII as described above. After imputation, SNPs were filtered using a QC process similar to the initial filtering of raw genotypes. Post-imputation QC excluded SNPs with MAF < 0.01, call rate < 0.95, and HWE $< 5 \times 10^{-7}$. These datasets were used in the GWAS analyses. We then selected only the imputed SNPs from the two datasets and merged them into a single dataset in which all SNPs had been imputed. This dataset was used to compare imputation accuracies of SNPs according to their location relative to known genes.

References used for the imputation were downloaded prior to the imputation process from the Impute website (http://mathgen.stats.ox.ac.uk/impute/impute.html). The references were labeled on the website as International HapMap project phase II release #22 (HMII) and 1000G phase I. A total of 1,417,081 SNPs from 90 Chinese and Japanese samples were used from HMII with an additional 39,343,900 SNPs from 1092 worldwide sample populations in the combined reference from 1000G. PanSNPdb shared 47,870 SNPs with the HapMap reference and 51,849 with the 1000G reference. The Thai dengue dataset shared 493,846 SNPs with the HapMap reference and 565,912 with the 1000G reference.

Imputation yield and accuracy

IMPUTE2 gives each imputed genotype a posterior probability score (info score) between zero and one. A higher threshold cut off for the probability score will usually result in higher accuracy but a lower yield. In this study, the posterior probability threshold was set to 0.9 to gain results with a high confidence of accuracy [24]. Genotypes with posterior probabilities < 0.9 were set to missing. Yield was reported as the percentage of non-missing genotypes within the removed SNPs and accuracy as the percentage of imputed, non-missing genotypes that matched the original genotypes.

SNP selection and annotation

All SNPs from the dengue dataset were grouped based on their location in genes. Gene annotation was collected from Illumina sample sheets (Illumina Inc., San Diego, CA) and the NCBI database of genetic variation [25]. Targeted gene regions included coding, intergenic, intronic, and untranslated regions (UTR). Some SNPs mapped to more than one location and were marked as being in a complex region. There were 53,277 SNPs that were not in any of these 5 groups and were subsequently discarded from further analysis.

The difference between imputed and genotyped data in the dengue dataset was evaluated by looking at several properties of the SNPs. Differences in accuracy and yield for each gene region were measured by varying the posterior probability threshold from 0.5 to 1.0. MAFs and *p*-values from chi-square tests for each SNP were compared between imputed and genotyped datasets. Coefficients of determination (r^2) were calculated for each comparison to estimate the concordance of imputed and genotyped SNPs. Pairwise LD, which is measured as r-squared, was calculated within a region of < 1 Mb around each SNP. The calculation was performed in PLINK v1.07 using the option –r2 with –ld-window-r2 0 and –ld-window-kb 1000. Average R-

squared values for each region were calculated and plotted in Microsoft Excel. R-squared values were also plotted against distance and averaging over 1 kb bins using R software (http://www.r-project.org/).

Results

Genotype imputation of Southeast Asian populations

The accuracy and variability of imputation for Southeast Asian samples were assessed with five populations downloaded from PanSNPdb and two publicly available references from the HapMap project and the 1000 genomes project. All populations had an average imputation accuracy of more than 92% (Fig. 1), regardless of which reference was used. Imputation with HMII as a reference gave an average accuracy of 96.57%, while for 1000G, the accuracy was 93.98% (Fig. 1a). The Thai (TH) population had the highest accuracy for both reference panels followed by Indonesia (ID), whereas the population from the Philippines (PI) had the lowest accuracy. The yield for each population was lower when imputation was performed with the HMII reference (average = 59.03%) compared to the 1000G reference (average = 68.44%) (Fig. 1b). The yields for all populations were similar when using the same reference. The only exception was the TH population imputed with the 1000G reference, which had a higher yield compared to the other populations.

Next, we looked at the results from each chromosome separately to demonstrate the variability of accuracies and yields (Fig. 2). Most results of imputation with HMII as a reference provided more than 95% accuracy (Fig. 2a). The 1000G reference provided a lower accuracy compared to HMII (Fig. 2b). The most striking result was that there was a change in standard deviation for imputation accuracy between the chromosomes. In particular, chromosomes 19 showed the lowest accuracy and higher variation in accuracy. Plotting the yield by population and chromosome showed no significant differences (Fig. 2c, d). However, chromosomes 19 also showed lower yield and chromosome 22 exhibited the highest level of variability. Because the imputation technique is, to a large extent, based on LD, we wanted to see if the higher variance could be correlated to any differences in the LD-pattern. We calculated the LD for each SNP pair with a distance between 10 kb and 1 MB. The number of these pairs that had an LD > 0.2 was recorded for each chromosome (Additional file 1: Figure S1). Chromosome 19 and 22 had the lowest values.

Trying to explain the differences in imputation accuracy, we investigated the population diversity of the five populations using MDS (Fig. 3). Whereas most samples were grouped together, all populations except TH showed large internal variation along the primary axis (C1) (Fig. 3a). The secondary axis (C2) mainly described the difference between 18 samples of the Thai Mlabri ethnic group (from Nan Province, Thailand) to the rest of the Thai samples. The third axis (C3) also mainly described the variation within Thai samples, whereas the fourth axis (C4) showed variation within PI, ID and SG (Fig. 3b).

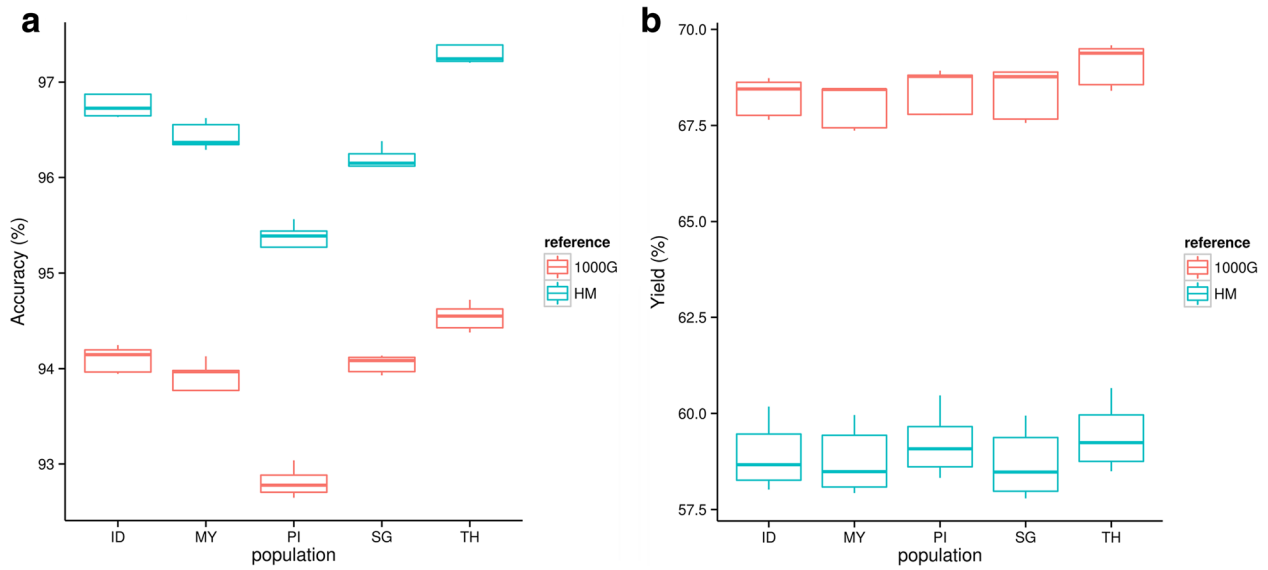


Fig. 1 Boxplot of accuracies and yields for imputation results across all populations. Five percent of randomly removed SNPs were imputed with IMPUTE2 using either the 1000 Genomes project phase I (1000G) or combined Chinese and Japanese haplotypes from the International HapMap project phase II (HMII) as a reference. The imputed SNPs were tested for accuracy with the previously removed SNPs. The same set of the removed SNPs was applied to all population dataset. The technique was repeated five times. **a** Boxplot of accuracy comparing populations and references. **b** Boxplot of yield comparing populations and references. Abbreviations: Indonesia (ID), Malaysia (MY), the Philippines (PI), Singapore (SG), and Thailand (TH)

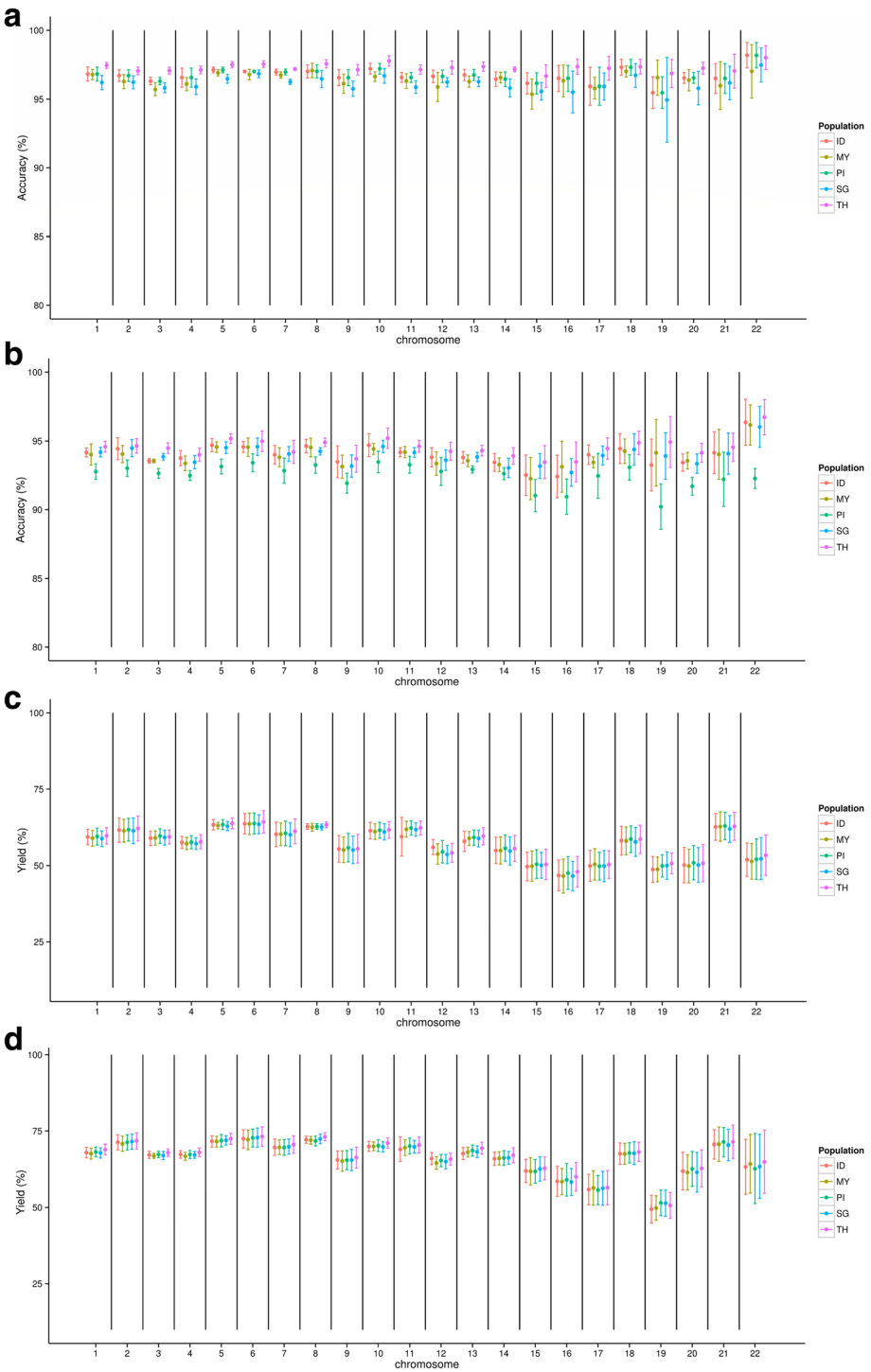


Fig. 2 Imputation accuracy and yield by chromosome. The results derived from 5% randomly removed SNPs. Imputation with IMPUTE2 was accomplished to recover the removed SNPs. The imputed SNPs were tested for accuracy with previously removed SNPs. The same set of the removed SNPs was applied to all population dataset. This process was repeated five times. **a** Imputation accuracy by chromosome using HMII as a reference. **b** Imputation accuracy by chromosome using 1000G as a reference. **c** Imputation yield by chromosome using HMII as a reference. **d** Imputation yield by chromosome using 1000G as a reference. Abbreviations: Indonesia (ID), Malaysia (MY), the Philippines (PI), Singapore (SG), and Thailand (TH)

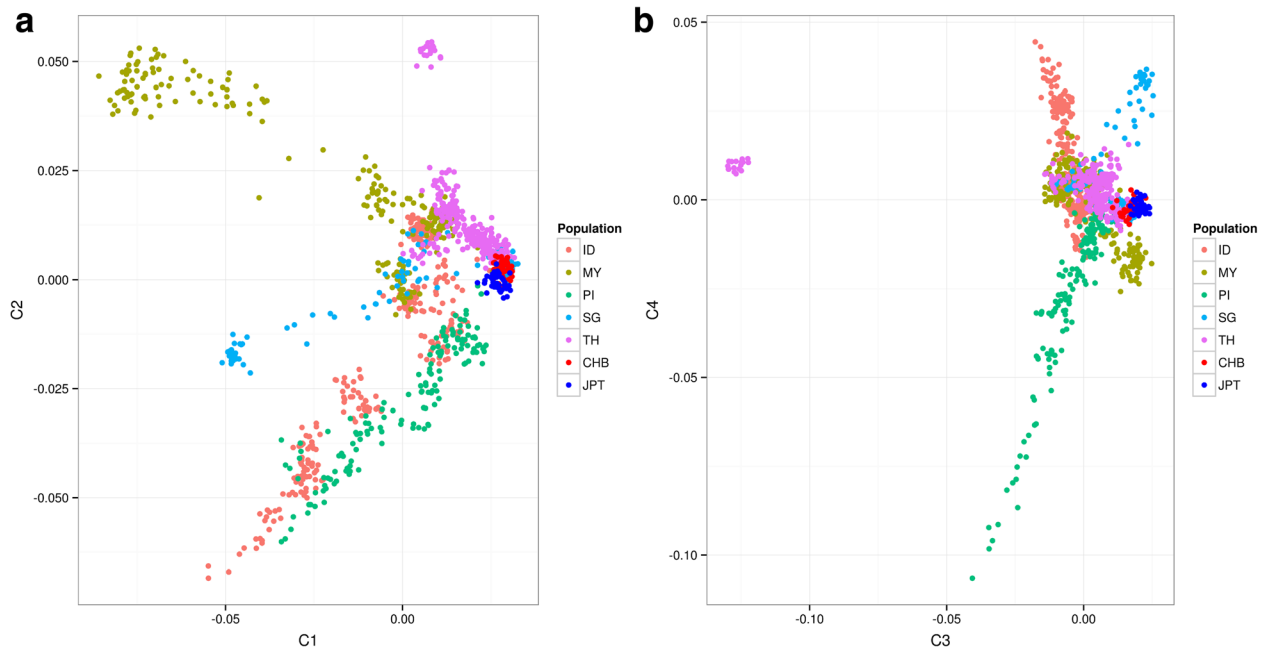


Fig. 3 Multidimensional scaling plot of Southeast Asian populations from PanSNPdb. Genotype data of samples from Southeast Asian populations were downloaded from PanSNPdb. After quality control, multidimensional scaling (MDS) was performed in PLINK v1.07. **a** Plotting of the first (C1) and the second (C2) axes. **b** Plotting of the third (C3) and the fourth (C4) axes. Abbreviations: Indonesia (ID), Malaysia (MY), the Philippines (PI), Singapore (SG), Thailand (TH), China (CHB) and Japan (JPT)

Effects of SNP location in a gene on imputation

Using the dengue GWAS dataset to evaluate the effects of SNP locations within a gene on the quality of imputation, we found 415,710 SNPs located within genes (Additional file 2: Table S1). As expected, intergenic regions contained the most SNPs, while complex regions, where SNPs have been associated with more than one gene, had the fewest. Varying the threshold settings in IMPUTE2 for accepting an imputed genotype showed that increasing the threshold led to an increase in accuracy but a decrease in yield (Additional file 1: Figure S2). Imputation results for SNPs in coding regions showed the highest yield and accuracy. Intronic and intergenic regions led to the second and third highest yield and accuracy, respectively. At a threshold of 0.5, all regions showed a similar yield. However, with an increasing threshold, the yield of coding regions increased compared to other locations. The opposite effect was observed for accuracy, and all locations approached the same level of accuracy when the threshold approached 1.

We further compared the measured MAF from the initial genotyped data to the MAF of the imputed data (Additional file 1: Figures S3-S4). Then, any imputed SNP that did not pass the quality-criteria (call rate > 0.95, MAF > 0.01, HWE > 10^{-7}) were removed (Additional file 2: Table S1). The MAF of this reduced set of SNPs were similarly compared to the same set of SNPs from the initial genotyped data (Additional file 1: Figure S4). The correlation for each region between the MAF of the genotyped data and both the imputed data and the imputed and post-imputation filtered data were calculated (Table 1). Coding regions had the fewest SNPs failing the QC, while complex regions had the lowest correlation, followed by untranslated regions (UTR), before filtering of the imputed data. After removing low quality SNPs, all regions showed a high

Table 1 Squared correlation of allele frequencies and chi-square *P*-values from SNPs in different regions

Region	Squared correlation (r^2) of minor allele frequency		Squared correlation (r^2) *p*-value from chi-square	
	Before post-imputation QC	After post-imputation QC	Before post-imputation QC	After post-imputation QC
Coding region	0.868	0.997	0.387	0.813
Complex region	0.817	0.991	0.267	0.782
Intergenic region	0.864	0.996	0.328	0.789
Intron region	0.863	0.995	0.340	0.784
UTR region	0.830	0.991	0.303	0.756

correlation in MAF between imputed and actual genotypes. Looking at the direction of change in MAF showed that SNPs with low initial MAF (MAF < 0.25) predominantly (> 80%) had even lower MAFs after imputation. SNPs with an initial MAF close to 0.5 had an equal distribution of SNPs that obtained higher and lower post imputation MAF. Imputation also appeared to systematically reduce the allele frequency of the initially low MAF for most imputed SNPs (Additional file 1: Figure S5). The same analysis was repeated using *p*-values from the binary association test instead of the allele frequency for each SNP (Additional file 1: Figures S6 and S7). All regions had a generally low correlation for the *p*-values before removing low quality SNPs ($r < 0.4$). After post-imputation QC, all regions showed an increase in correlation. Coding regions had the highest correlation both before ($r^2 = 0.39$) and after ($r^2 = 0.81$) removing low-quality SNPs. The lowest correlation for the binary association was found in complex regions before post-imputation QC ($r^2 = 0.27$) and UTR regions after the QC ($r^2 = 0.76$).

To test whether coding regions had a stronger LD compared to the other regions, PLINK v1.07 was used to calculate pairwise LD between each SNP and any other SNP within 1 Mb. LD values, presented as r-squared, are shown as averages for each region (Additional file 1: Figure S8) and as LD vs. distance plot (Additional file 1: Figure S9). SNPs within coding regions had the highest average LD to neighboring SNPs ($r^2 = 0.253$), followed by SNPs in intronic regions ($r^2 = 0.232$).

Discussion

Imputation of genotyped datasets is a common practice when performing genome-wide association studies. This technique is used to fill in missing genotypes and to increase the density by adding information from SNPs that are not present in the original dataset [5, 26]. Imputation of SNPs that are not available in the dataset serves several purposes. First, if available SNP arrays are designed based on a specific population, such as Europeans, the SNPs may not cover the areas of interest for another population. SNPs important to populations from Southeast Asia might therefore be underrepresented or missing from these arrays. This situation has been reported for populations from Africa [27] and Mexico [28]. Second, if data are collected independently between groups of case and control populations, the datasets might have been genotyped on different SNP sets. Third, GWAS usually requires a high number of SNPs to increase chance to detect association signals. Although current genotyping arrays could contain more than a million markers, imputation still adds more SNPs for denser full genome coverage.

The choice of reference panels can affect the accuracy of imputation through the genetic variation of the samples and the genetic relationship between the samples in the reference panel and the imputed references [6, 27]. We studied these effects in five populations from Southeast Asia. The 1000G reference provided the highest yield, while the HMII reference had the highest accuracy. These results were the same for all Southeast Asian populations. IMPUTE2 software provides the posterior probability score for each imputed genotype. There was increasing accuracy with a decreasing yield when the probability threshold increased (Additional file 1: Figure S2). At the threshold 1.0, the result showed a large drop in yield but only a limited increase in accuracy. At the threshold 0.9, the slope of the yield and accuracy were significantly changed. This threshold might be a good starting point for using IMPUTE2 for Southeast Asian populations.

Genotype imputation of the Thai population had the highest accuracy in the current study. Previous work has shown that in the PanSNPdb database, the Thai population had the highest relationship to the Chinese and Japanese populations out of the other four study populations [16]. The increased accuracy can therefore be explained by this closer relationship. To determine if population diversity influenced the average imputation accuracy, classical MDS was used to display the variation within Southeast Asian populations. The main variation in the MDS plot (C1) was related to the sub-populations in 4 of the 5 main populations, and the only exception was the Thai population (Fig. 3). The Thai samples form a more homogenous group compared to other populations, and this outcome can help explain why they had the most accurate results. Eighteen individuals from the Thai Mlabri group, which is a hunter-gatherer group in Northern Thailand, clustered away from other Thai samples, which is consistent with the findings of previous studies [17, 29]. Nevertheless, further investigation is necessary to understand the effects of population stratification on imputation results.

Plotting the imputation accuracy and yield for each chromosome revealed that the variation within populations was the highest for the smaller chromosomes, especially chromosomes 19 and 22. The same results were also observed for yield. In particular, the population from the Philippines showed a large increase in variability for these chromosomes when using the 1000G reference. This result might indicate a specific issue with the SNP selection for these chromosomes, such as the LD structure being different or the relationship to the reference being lower in these regions. One possible reason for the increased variability was found by looking at the long distance LD in each chromosome. Chromosome 19 and 22 was shown to have the fewest SNPs connected by LD above 0.2 when comparing inter-SNP distances above 10 kb (Additional file 1: Figure S1). The accuracy of resulting imputation for these two chromosomes will therefore be more dependent on which SNPs are removed and will show more variance

between replications with randomly selected SNPs. The results also clearly show that only taking the average numbers for accuracy or yield into account will result in overlooking potentially important information.

Coding sequences have been the favored area to search for functional mutations because these sequences are more informative and easier to interpret due to the direct link to a protein and the possibility of functional changes [30–32]. This approach made it useful to compare the imputation results between coding and other SNP regions. Our results show that a higher percentage of SNPs in coding regions passed the post-imputation QC than in other regions. SNPs in coding regions also had the highest accuracy. This outcome correlates well with the results showing coding regions to have the highest LD with the surrounding SNPs (Additional file 1: Figure S8). This is an important factor to consider when discussing significant results because imputed SNPs in coding regions will have a higher accuracy compared to SNPs in less conserved areas. This trend does not mean that we can omit non-coding SNPs because they have been shown to be associated with phenotypes in more than one-third of GWAS [33–35].

Our study also compared GWAS results from a dataset from Thai dengue fever patients to see how imputation affected the reported results. Imputation appeared to systematically reduce the allele frequency of the initially minor allele for most imputed SNPs (Additional file 1: Figure S5). This effect was more pronounced for SNPs with an initial low MAF and will make imputing low frequency alleles difficult. Our results demonstrated that imputation tended to increase the common allele. This outcome was especially problematic if the genotypes are very rare variants [26]. Even if the most significant SNPs had higher significance in the imputed dataset, this trend was not seen when comparing *p*-values from GWAS before and after imputation.

The imputed genotypes were also subjected to QC-filtering, which is similar to the QC being performed on the raw genotype data. It was previously shown that post-imputation QC did not influence the imputation outcome [36]. However, we observed an improvement in the correlation between measured MAF and *p*-values from imputed data and genotyped data. Without the SNPs failing the QC, the correlation between imputed data and genotyped data was close to one for allele frequency and had improved *p*-values. Even if over half the SNPs were removed in this step, the remaining data had higher quality and were more trustworthy. This difference in post-QC improvement could be due to the initial imputation accuracy. Post-imputation QC might be more important if the initial imputation results are less accurate. This result is also supported by a previous experiment that similarly demonstrated Hardy–Weinberg disequilibrium is a crucial step for post-imputation filtering [37].

This study demonstrates that the expected accuracy and yield of imputation in various Southeast Asian populations varies between populations. Our reference comparison of HMII and 1000G in imputation in Thai GWAS showed that using a larger reference provided a higher yield but caused a reduction in accuracy compared to a smaller but more related reference. We also extensively showed the imputation results with respect to SNP localization near genes using Thai genome-wide genotypes as a model. This study provides crucial information for investigators undertaking imputation, especially in Southeast Asian populations.

Conclusions

This work provides the first evidence of imputation reference selection for Southeast Asian studies and highlights the effects of SNP locations respective to genes on imputation outcome. Researchers will need to consider the trade-off between accuracy and yield in future imputation studies.

Additional files

Additional file 1: Figure S1. Long distance LD pr. Mb separated by the chromosome for each population LD for each SNP pair with a distance between 10 kb and 1 Mb taken into account. **Figure S2.** Accuracy and yield of imputation between each location of SNPs. **Figure S3.** Comparing minor allele frequencies of SNPs between imputed and actual genotypes before quality control of imputed results separately plot by each SNP location. **Figure S4.** Comparing allele frequencies of SNPs between imputed and actual genotypes after quality control of imputed results separately plotted by each SNP location. **Figure S5.** Proportion of SNPs with lower AF after imputation vs initial AF. **Figure S6.** Comparing *p*-values between cases and controls of SNPs between imputed and actual genotypes before quality control of imputed results separately plotted by each SNP location. **Figure S7.** Comparing *p*-values between cases and controls of SNPs between imputed and actual genotypes after quality control of imputed results separately plotted according to each SNP location. **Figure S8.** Average linkage disequilibrium as r-squared for each region. SNPs were assigned to gene locations. **Figure S9.** LD, measured as r^2, plotted against physical distance. (PDF 1286 kb)

Additional file 2: Table S1. Summary of SNPs used in study of SNP annotation to imputation outcome. (XLS 26 kb)

Abbreviations

1000G: the 1000 Genomes project; HMII: the International HapMap Project; ID: Indonesia; LD: linkage disequilibrium; MY: Malaysia; PanSNPdb: Pan-Asian SNP genotyping database; PI: The Philippines; QC: Quality control; SG: Singapore; TH: Thailand

Acknowledgements

We thank all team members of the Pan-Asian SNP Genotyping Database especially Dr. Sissades Tongsima and Dr. Chumpol Ngamphiw for collecting and curating PanSNPdb data.

Funding

This research was supported by National Research Council of Thailand (NRCT) (grant number 92/2550). Additionally, the research was partial funded by the Office of the Higher Education Commission and Mahidol University under the National Research Universities Initiative. W. Lert-itthiporn was supported by a scholarship from the Medical Scholars Program, Mahidol University. P. Malasit was supported by NSTDA Chair Professor Grant. P. Suriyaphol was supported

by a National Research University (NRU) Grant through Mahidol University and TRF-office for R&D.

Availability of data and materials
The data supporting the conclusions of this article are included within the article and its additional files. The datasets from Thai dengue patients generated and/or analysed during the current study are not publicly available due to ethical restrictions enforced by the Siriraj Institutional Review Board, but are available from the corresponding author on reasonable request.

Authors' contributions
WL participated in the imputation experiment and analyzed the association study as well as the SNP regional effects on imputation and contributed to the article. HG analyzed pre- and post-imputation results and wrote the article. FM and AS designed the study and participated in genotyping the DNA samples. PM and NT conducted the dengue patient's recruitment and DNA. BS and PS participated in the study design and wrote the article. All authors read and approved the final manuscript.

Ethics approval and consent to participate
The ethics of this study were approved by the Siriraj Institutional Review Board Certificate of Approval. This program is completed to certify that the Siriraj Institutional Review Board is in full compliance with international guidelines for human research protection such as the Declaration of Helsinki, the Belmont Report, CIOMS Guidelines and the International Conference on Harmonization in Good Clinical Practice (ICH-GCP). Written informed consent was obtained from both the participants and legal guardians.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Molecular Medicine Graduate Program, Faculty of Science, Mahidol University, Bangkok, Thailand. [2]Division of Bioinformatics and Data Management for Research, Department of Research and Development, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand. [3]Integrative Computational BioScience Center, Department of Biochemistry, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand. [4]Center of Excellence in Bioinformatics and Clinical Data Management, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. [5]Unité de Génétique Fonctionnelle des Maladies Infectieuses, Department Genome and Genetics, Institut Pasteur, Paris, France. [6]Centre National de la Recherche Scientifique, URA3012, Paris, France. [7]Systems Biology of Diseases Research Unit, Faculty of Science, Mahidol University, Bangkok, Thailand. [8]Medical Biotechnology Research Unit, National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency, Bangkok, Thailand. [9]Division of Dengue Hemorrhagic Fever Research, Department of Research and Development, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand. [10]Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan.

Received: 7 December 2016 Accepted: 24 January 2018
Published online: 13 February 2018

References
1. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, Kivinen K, Bojang KA, Conway DJ, Pinder M, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet. 2009;41(6):657–65.
2. Li J, Guo YF, Pei Y, Deng HW. The impact of imputation on meta-analysis of genome-wide association studies. PLoS One. 2012;7(4):e34486.
3. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 2009;5(5):e1000477.
4. Teo YY, Small KS, Kwiatkowski DP. Methodological challenges of genome-wide association analysis in Africa. Nat Rev Genet. 2010;11(2):149–60.
5. Zhao Z, Timofeev N, Hartley SW, Chui DH, Fucharoen S, Perls TT, Steinberg MH, Baldwin CT, Sebastiani P. Imputation of missing genotypes: an empirical evaluation of IMPUTE. BMC Genet. 2008;9:85.
6. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet. 2009;84(2):235–50.
7. Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A. A comprehensive evaluation of SNP genotype imputation. Hum Genet. 2009;125(2):163–71.
8. Barbujani G, Colonna V. Human genome diversity: frequently asked questions. Trends Genet. 2010;26(7):285–95.
9. Pillai NE, Okada Y, Saw WY, Ong RT, Wang X, Tantoso E, Xu W, Peterson TA, Bielawny T, Ali M, et al. Predicting HLA alleles from high-resolution SNP data in three southeast Asian populations. Hum Mol Genet. 2014;23(16):4443–51.
10. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.
11. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39(7):906–13.
12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
13. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34(8):816–34.
14. Wong KM, Langlais K, Tobias GS, Fletcher-Hoppe C, Krasnewich D, Leeds HS, Rodriguez LL, Godynskiy G, Schneider VA, Ramos EM, et al. The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. Nucleic Acids Res. 2017;45(D1):D819–26.
15. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). Nucleic Acids Res. 2017;45(D1):D896–901.
16. Ngamphiw C, Assawamakin A, Xu S, Shaw PJ, Yang JO, Ghang H, Bhak J, Liu E, Tongsima S, Consortium HP-AS. PanSNPdb: the Pan-Asian SNP genotyping database. PLoS One. 2011;6(6):e21451.
17. Yang X, Xu S, Hugo Pan-Asian SNP Consortium. Identification of close relatives in the HUGO Pan-Asian SNP database. PLoS One. 2011;6(12):e29502.
18. Finkel TH, Li J, Wei Z, Wang W, Zhang H, Behrens EM, Reuschel EL, Limou S, Wise C, Punaro M, et al. Variants in CXCR4 associate with juvenile idiopathic arthritis susceptibility. BMC Med Genet. 2016;17:24.
19. Han S, Kim-Howard X, Deshmukh H, Kamatani Y, Viswanathan P, Guthridge JM, Thomas K, Kaufman KM, Ojwang J, Rojas-Villarraga A, et al. Evaluation of imputation-based association in and around the integrin-alpha-M (ITGAM) gene and replication of robust association between a non-synonymous functional variant within ITGAM and systemic lupus erythematosus (SLE). Hum Mol Genet. 2009;18(6):1171–80.
20. Li L, Li Y, Browning SR, Browning BL, Slater AJ, Kong X, Aponte JL, Mooser VE, Chissoe SL, Whittaker JC, et al. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS One. 2011;6(9):e24945.
21. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, et al. Quality control procedures for genome-wide association studies. Curr Prot Hum Genet. 2011;Chapter 1:Unit1. 19.
22. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013;10(1):5–6.
23. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529.
24. Lert-itthiporn W, Suriyaphol P. Genotype imputation in Thai population. In: Poster presented at: the Human Genome Meeting 2015. Kuala Lumpur: Human Genome Organisation; 2015.
25. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.

26. Zheng HF, Rong JJ, Liu M, Han F, Zhang XW, Richards JB, Wang L. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. PLoS One. 2015;10(1):e0116487.
27. Sung YJ, Gu CC, Tiwari HK, Arnett DK, Broeckel U, Rao DC. Genotype imputation for African Americans using data from HapMap phase II versus 1000 genomes projects. Genet Epidemiol. 2012;36(5):508–16.
28. Krithika S, Valladares-Salgado A, Peralta J, Escobedo-de La Pena J, Kumate-Rodriguez J, Cruz M, Parra EJ. Evaluation of the imputation performance of the program IMPUTE in an admixed sample from Mexico City using several model designs. BMC Med Genet. 2012;5:12.
29. Consortium HP-AS, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen CH, Chen J, et al. Mapping human genetic diversity in Asia. Science. 2009;326(5959):1541–5.
30. Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, Barker JN, Barrett JC, Blackburn H, Brand O, Burren O, et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature. 2013;498(7453):232–5.
31. Lee EK, Gorospe M. Coding region: the neglected post-transcriptional code. RNA Biol. 2011;8(1):44–8.
32. Bayegan AH, Garcia-Martin JA, Clote P. New tools to analyze overlapping coding regions. BMC Bioinf. 2016;17(1):530.
33. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009;461(7261):199–205.
34. Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be? Nucleic Acids Res. 2016;44(13):6046–54.
35. Cowie P, Hay EA, MacKenzie A. The noncoding human genome and the future of personalised medicine. Expert Rev Mol Med. 2015;17:e4.
36. Southam L, Panoutsopoulou K, Rayner NW, Chapman K, Durrant C, Ferreira T, Arden N, Carr A, Deloukas P, Doherty M, et al. The effect of genome-wide association scan quality control on imputation outcome for common variants. Eur J Hum Genet. 2011;19(5):610–4.
37. Shriner D. Impact of Hardy-Weinberg disequilibrium on post-imputation quality control. Hum Genet. 2013;132(9):1073–5.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

