BMC Bioinformatics

SOFTWARE Open Access

MI-MAAP: marker informativeness for multi-ancestry admixed populations

Siqi Chen[1,2], Sudhir Ghandikota[1,2], Yadu Gautam[1] and Tesfaye B. Mersha[1*]

* Correspondence: tesfaye.mersha@cchmc.org
[1]Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, 3333 Burnet Avenue, MLC 7037, Cincinnati, OH 45229-3026, USA
Full list of author information is available at the end of the article

Abstract

Background: Admixed populations arise when two or more previously isolated populations interbreed. A powerful approach to addressing the genetic complexity in admixed populations is to infer ancestry. Ancestry inference including the proportion of an individual's genome coming from each population and its ancestral origin along the chromosome of an admixed population requires the use of ancestry informative markers (AIMs) from reference ancestral populations. AIMs exhibit substantial differences in allele frequency between ancestral populations. Given the huge amount of human genetic variation data available from diverse populations, a computationally feasible and cost-effective approach is becoming increasingly important to extract or filter AIMs with the maximum information content for ancestry inference, admixture mapping, forensic applications, and detecting genomic regions that have been under recent selection.

Results: To address this gap, we present MI-MAAP, an easy-to-use web-based bioinformatics tool designed to prioritize informative markers for multi-ancestry admixed populations by utilizing feature selection methods and multiple genomics resources including 1000 Genomes Project and Human Genome Diversity Project. Specifically, this tool implements a novel allele frequency-based feature selection algorithm, Lancaster Estimator of Independence (LEI), as well as other genotype-based methods such as Principal Component Analysis (PCA), Support Vector Machine (SVM), and Random Forest (RF). We demonstrated that MI-MAAP is a useful tool in prioritizing informative markers and accurately classifying ancestral populations. LEI is an efficient feature selection strategy to retrieve ancestry informative variants with different allele frequency/selection pressure among (or between) ancestries without requiring computationally expensive individual-level genotype data.

Conclusions: MI-MAAP has a user-friendly interface which provides researchers an easy and fast way to filter and identify AIMs. MI-MAAP can be accessed at https://research.cchmc.org/mershalab/MI-MAAP/login/.

Keywords: Aancestry informative markers, AIMs, Lancaster estimator of Independence, LEI, MI-MAAP, Admixed population, Admixture mapping

Background

Markers with large differences in allele frequencies between ancestral populations, known as ancestry informative markers (AIMs), provide genetic information useful to infer ancestry [1]. Genetic ancestry inference is an area of considerable interest in disease genetics, population genetics, anthropology and forensics and genetic testing [2]. Identification of such a subset of genetic markers that is sufficiently informative for inferring genetic ancestries but still small enough to reduce the computational cost for resource-limited laboratories is of particular importance. With the rapid development of the field of human genetics, enormous amounts of genetic variation in large data repositories have been generated, such as 1000 Genomes Project, HapMap Project, Human Genome Diversity Project (HGDP) and Exome Aggregation Consortium (ExAC) [3–5]. Due to the growing number of the available genomic data, new computational analytics are required to effectively process massive SNP data and extract informative AIMs for multiple racial ancestries.

The task of identifying informative markers to assign individual genomic regions to correct ancestries can be difficult in admixed populations with multiple ancestral origins [6, 7]. Hence, prioritizing AIMs using various feature selection methods is of paramount significance in studies of population structure and to map risk loci via admixture mapping [8]. One of the most powerful methods for identifying such informative AIMs is to apply feature selection techniques where subsets of relevant SNPs are selected based on the classification model construction [9]. Machine learning techniques such as Support Vector Machine (SVM), Random Forest (RF) and Principal Components Analysis (PCA) can be applied to score and rank informative SNPs by modeling classifiers. Numerous AIM panels from autosomal chromosomes have been published [1–5, 8, 10]. Briefly, these studies have shown that 60–150 informative markers are useful for population clustering into major geographic regions [9, 11, 12]. Previous work to develop AIMs considered binary classification with pairs of subpopulations [12]. There are limited statistical tools to develop informative markers for three-way or multi-way admixed populations including Latino populations [11–13]. It is thus desirable to develop an efficient feature selection method, which not only identifies AIMs to estimate the admixture proportions in samples from multi-way admixed population with high accuracy, but is also computationally feasible for small laboratories.

In this study, we extend the two-way ancestry analysis into multi-way ancestry classification, and present a user-friendly web-based tool called Marker Informativeness for Multi-Ancestry Admixed Populations, MI-MAAP, to facilitate selection of informative SNPs in multi-admixed population using 1000 Genomes Project, Human Genome Diversity Project as well as user-generated data. It adopts a novel likelihood-based feature selection method, Lancaster Estimator of Independence (LEI), which has the ability to efficiently compare multiple ancestral populations using allele frequency information [14]. LEI potentially reduces the computation costs associated with individual-level genotype data and yields cost-efficient AIMs panel for ancestry inferences. The lack of easy-to-use-analytical tools is a substantial barrier for most biologists with limited computer programming background. Exploiting recent advancements in web framework technologies the MI-MAAP was developed to simplify informative marker selection for genetic studies in admixed populations. Users are able to freely access MI-MAAP and retrieve large genomic datasets and extract AIMs on-the-fly through the user-friendly interface. We also further investigated the identified SNPs or their nearby genes, with specific functional information and associated diseases that are known to be prevalent in certain ancestry. Additional information (e.g. gene expression, gene ontology, pathway information, protein information and species orthologs) is made available for users to explore their AIMs known to be prevalent in certain ancestry.

Implementation

Measure of marker Informativeness

Lancaster estimator of independence (LEI) is a maximum likelihood-based estimator of ancestry informativeness of markers using allele frequency data under Hardy-Weinberg Equilibrium (HWE). LEI can compute the quantitative score from the marker-population joint distribution and thus account for 3 or more populations simultaneously. Let m and M be the alternate and reference alleles at a marker and $X = (x_1 = mm, x_2 = mM, x_3 = MM)$ be the three genotypes. Let $Y = (y_1, y_2, \ldots, y_k)$ be the k distinct ancestral populations and f_j represents the population reference allele frequency in the j^{th} population with c_j indicating the number of individuals in the j^{th} population. Then, under the assumption of HWE in individual populations, the expected genotype counts of genotype x_1, x_2, and x_3 in the j^{th} population can be expressed as $\hat{n}_{1j} = c_j(1-f_j)^2, \hat{n}_{2j} = 2c_jf_j(1-f_j),$ and $\hat{n}_{2j} = c_jf_j^2$, respectively. If $n = c_1 + c_2 + \ldots + c_k$ is the total sample size, the joint probability distribution of X and Y, $p_{ij} = P(X = x_i, Y = y_j)$ can be estimated as $\hat{p}_{ij} = \frac{\hat{n}_{ij}}{n}$, $i = 1, 2, 3$ and $j = 1, 2, \ldots, k$. Then, the Lancaster estimator of independent (LEI) is defined as –

$$\hat{\theta}^2 = \sum\nolimits_{i,j} \frac{\hat{p}_{ij}^2}{\hat{p}_{\mathrm{i}+}\hat{p}_{+j}} - 1 \tag{1}$$

where $\hat{p}_{i+} = \frac{\sum_j n_{ij}}{n} = P(X = x_i)$ and $\hat{p}_{+j} = \frac{\sum_i n_{ij}}{n} = P(Y = y_j)$ are the marginal distributions of X and Y, respectively. $\hat{\theta}^2$ is the measure of the magnitude of independence between the two categorical variables X and Y, with $\hat{\theta}^2 = 0$ if X and Y are independent and $0 \leq \hat{\theta}^2 \leq \min(3, k) - 1$ [15] . In particular, for genotype data with biallelic markers, $0 \leq \hat{\theta}^2 \leq 1$ when the sample constitutes two populations (i.e. $k = 2$) and $0 \leq \hat{\theta}^2 \leq 2$ when the sample constitutes three or more populations (i.e. $k \geq 3$).

Programmatic architecture

MI-MAAP was built on the Django web development framework which can be run on different operating systems. PHP APIs were developed to retrieve SNP data from the local database and Python scripts were written to access them and perform feature selection calculations. Python's Scikit-Learn library was used to conduct PCA, SVM, and RF. The combination of HTML and CSS was used to construct the user interface. In addition, jQuery Javascript library as well as jQuery plugins were used to build dynamic input forms, add interaction controls to output tables, and assist in some client-side design details. The tool was deployed on Apache httpd web server.

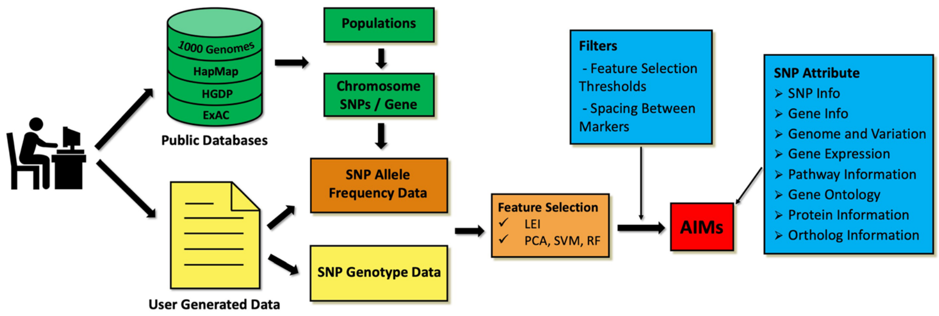


Fig. 1 Schematic representation of MI-MAAP workflow. In the flow diagram, once the user enters the information and clicks the submit button, information is mined from specific local databases, and MI-MAAP algorithm computes all comparisons for LEI to output Ancestry Informative Markers (AIMs)

Results

Workflow of MI-MAAP

MI-MAAP was designed as a web-based tool for analyzing the marker informativeness in multi-ancestry admixed populations. Its workflow is described in Fig. 1. Users have two types of input options: a) users can input a chromosome, SNP list or a single gene they are interested in from public databases (e.g., 1000 Genomes Project, International Haplotype Map (HapMap), Human Genome Diversity Project (HGDP) and Exome Aggregation Consortium (ExAC) [3–5, 10]; b) users can also upload their own SNP data files that include the population-specific allele frequency or genotype information. When the input is provided as chromosome or SNPs/gene, MI-MAAP first maps chromosomes or genes to their corresponding SNP sets and then uses a local database to retrieve and collect the population-specific SNP allele frequencies from the selected database.

Databases and populations

The local database stores allele frequency data from four different public data repositories: the 1000 Genomes Project [10], HapMap [16], Human Genome Diversity Project (HGDP) [5] and Exome Aggregation Consortium (ExAC) [17]. Databases utilized and their number of populations and sample sizes with the number of SNPs are presented in Table 1. For example, the default database, 1000 Genomes Project Phase III data, includes a total of 2504 samples from 26 populations (Fig. 2). Reference population with labels, sample sizes that are available from each public database to infer ancestry and web link are provided in Additional file 1: Table S1. The 1000 Genomes Project (n = 84.7 million SNPs) across chromosomes are provided in Table 2.

Table 1 Reference populations and sample size with the number of SNPs databases accessed by MI-MAAP

Database	Populations	Individuals	SNPs
1000 Genomes	26	2504	84.7 million
HapMap	11	1397	3.2 million
HGDP	53	1043	657,000
ExAC	6	60,706	10 million

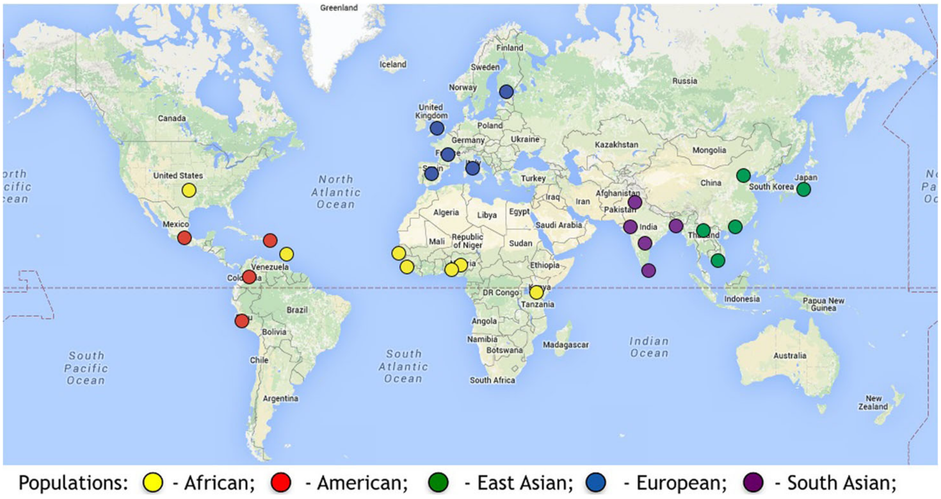


Fig. 2 Geographical distribution of 26 populations in the 1000 Genomes Project. The major continental regions are represented in 1000 Genomes Project. Source: The International Genome Sample Resource (IGSR) and 1000 Genomes Project https://www.internationalgenome.org)

Feature selection methods

After obtaining population-specific SNP data from either the public database or the user uploaded file, feature selection methods will be performed. LEI is an efficient likelihood-based approach which is computed by using the reference allele frequency data to estimate the genotype count data based on the Hardy-Weinberg Equilibrium (HWE) assumption; while PCA, SVM, and RF require the individual-level genotype data. When using summary-level allele frequency data, LEI will be applied. When using user-generated individual-level genotype data, one of the feature selection methods from LEI, PCA, SVM and RF can be performed. If LEI is chosen for genotype data, estimated allele frequencies will be first calculated from the genotype data and then used to compute LEI values. Thresholds for the computation values of the selected feature selection method can be set to filter relevant SNPs. In order to select the most informative SNPs and remove redundant and uninformative SNPs, we implemented filtering criteria including spacing between markers to account the linkage disequilibrium (LD). Removal of redundant and uninformative SNPs reduces the time and genotyping costs while increasing the power for extracting the most informative panels of SNPs. MI-MAAP also provides various ways to export the output, including Excel, CSV and PDF

Table 2 Number of SNPs on each chromosome for 1000 Genomes Phase III dataset

Chr	No. of SNPs	Chr	No. of SNPs	Chr	No. of SNPs
1	6,468,094	9	3,560,687	17	2,329,288
2	7,081,600	10	3,992,219	18	2,267,185
3	5,832,276	11	4,045,628	19	1,832,506
4	5,732,585	12	3,868,428	20	1,812,841
5	5,265,763	13	2,857,916	21	1,105,538
6	5,024,119	14	2,655,067	22	1,103,547
7	4,716,715	15	2,424,689	X	3,468,093
8	4,597,105	16	2,697,949		

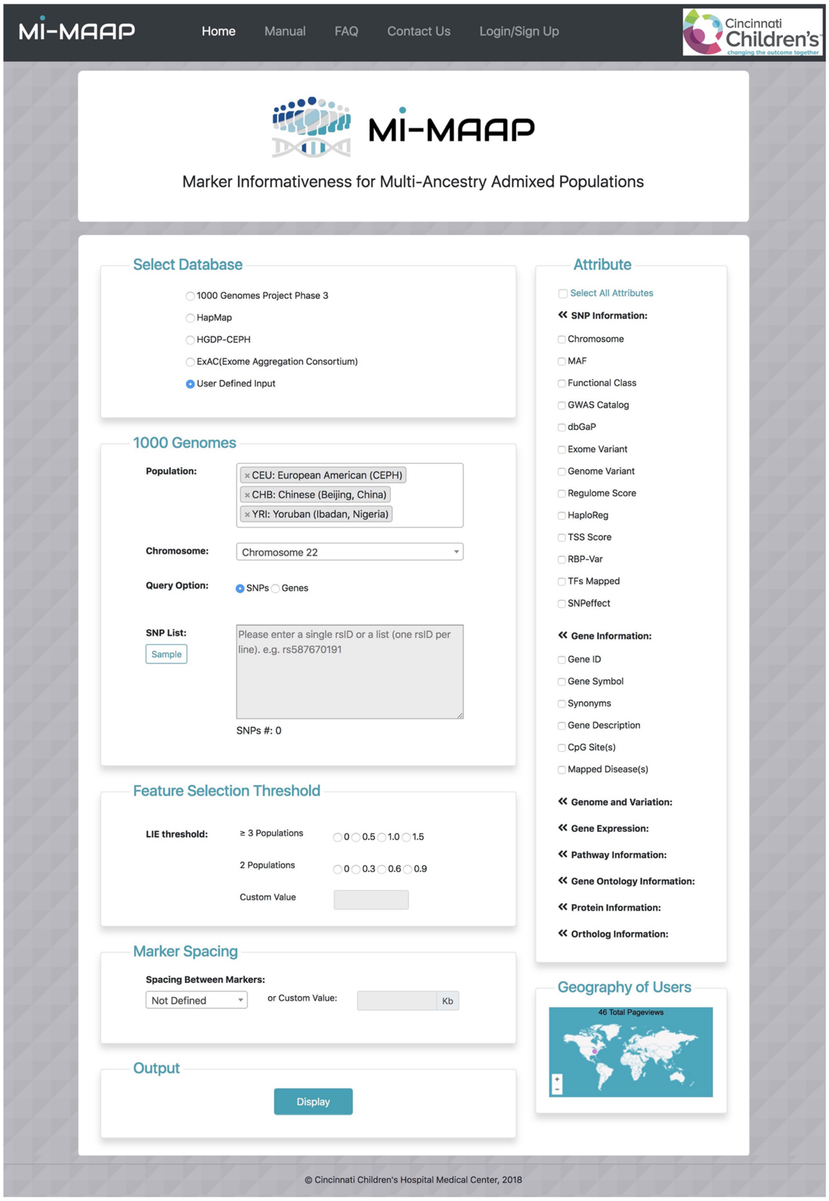
files. The ability of LEI to determine ancestry proportion in multi-ancestry admixed populations using allele frequency data without requiring the individual-level genotype data provides a more efficient alternative to the existing computationally and time-consuming genotype-based analysis for ancestry inference. When other choices are made instead of LEI, we recommend using Random Forest or SVM. In our analyses, the PCA based approach requires a larger set of informative markers to achieve similar performance as of LEI, RF, and SVM to estimate the ancestry proportion of admixed samples. We would also like to point out that the reference panels used for the construction of an informative marker set should be representative of the samples. We suggest developing AIMs from representative populations of interest.

Attributes to SNPs/genes

In addition to prioritizing AIMs, MI-MAAP allows to retrieve and explore functional annotation information associated with AIMs. These attributes are grouped into eight categories: SNP information (chromosome, alleles, MAF, functional class, Regulome Score, TSS Score, and links to GWAS Catalog, dbGap, Exome variant, Genome variant and so on), gene information (such as gene ID, gene symbol, synonyms, gene description, CpG sites and mapped diseases), genome and variation (links to ENCODE, dbVar, ClinVar and BioGPS), gene expression (such as GEO profiles, GTEx eQTL, Blood eQTL and so on), biological pathways (links to KEGG pathways, Reactome and BioCarta), gene ontology (cellular components, biological process and molecular function), protein (links to UniProt, Protein Atlas, PFAM and SMART), and species orthologs (such as Entrez IDs for chimp, rhesus, mouse, rat, zebrafish, cattle, chicken and dog) [7, 18, 19].

Components of MI-MAAP

Figure 3 shows the layout of the MI-MAAP homepage. The overall process of informative SNPs selection from public data is performed in six stages, each building on the results from the preceding stage. The first stage is to select the marker data source. One of the four public databases (1000 Genomes Project, HapMap, HGDP or ExAC) can be selected to retrieve allele frequency data of the SNPs users are interested in. In addition to the public reference datasets, users can upload their own generated data to compute the measure of marker informativeness by selecting the ‘User Defined Input’ radio button. In the second stage, based on the selected database users are able to choose the populations that are included in that database. Meanwhile, users are required to either select a chromosome in the dropdown menu or enter a list of SNPs or a gene name in the given text area. If ‘User Defined Input’ is selected, users need to upload their own SNP allele frequency data or genotype data. A single file in text or excel format of reference data can be uploaded. For the allele frequency data, the reference file must have a header line (SNPID, Chromosome, Ref_Allele, Alt_Allele, Pop1_Freq, …,Popk_Freq) and variants along the rows. Here ‘Popj_Freq’ is the reference allele frequency in population j. In addition, a dynamic formset is attached to enter the number of samples for all the populations used in the data file. For the genotype data, the file must have a header line (SampleID, Population, SNP1, SNP2, …, SNPk) and individual samples along the rows. ‘SNPj’ is the genotype of SNPj in a sample, coded as an integer



Marker Informativeness for Multi-Ancestry Admixed Populations

Select Database

○ 1000 Genomes Project Phase 3
○ HapMap
○ HGDP-CEPH
○ ExAC(Exome Aggregation Consortium)
● User Defined Input

1000 Genomes

Population: × CEU: European American (CEPH) × CHB: Chinese (Beijing, China) × YRI: Yoruban (Ibadan, Nigeria)

Chromosome: Chromosome 22

Query Option: ● SNPs ○ Genes

SNP List: Sample Please enter a single rsID or a list (one rsID per line). e.g. rs587670191

SNPs #: 0

Feature Selection Threshold

LIE threshold: ≥ 3 Populations ○ 0 ○ 0.5 ○ 1.0 ○ 1.5
2 Populations ○ 0 ○ 0.3 ○ 0.6 ○ 0.9
Custom Value

Marker Spacing

Spacing Between Markers: Not Defined or Custom Value: Kb

Output

Attribute

□ Select All Attributes

« SNP Information:
□ Chromosome
□ MAF
□ Functional Class
□ GWAS Catalog
□ dbGaP
□ Exome Variant
□ Genome Variant
□ Regulome Score
□ HaploReg
□ TSS Score
□ RBP-Var
□ TFs Mapped
□ SNPeffect

« Gene Information:
□ Gene ID
□ Gene Symbol
□ Synonyms
□ Gene Description
□ CpG Site(s)
□ Mapped Disease(s)

« Genome and Variation:
« Gene Expression:
« Pathway Information:
« Gene Ontology Information:
« Protein Information:
« Ortholog Information:

Geography of Users

© Cincinnati Children's Hospital Medical Center, 2018

Fig. 3 MI-MAAP web interface**.** The architectural design consists of six layers where a user selects populations, markers and filtering criteria to generate AIMs. The web interface allows user to either display the information on the browser or to download to a local hard drive

counting the number of the reference alleles. The third stage is the feature selection threshold section. If the allele frequency data is used, LEI will be computed. If user-generated genotype data is uploaded, users can select one of the methods from LEI, PCA, SVM and RF. LEI values are in the range of 0 to 2 for three or more -population analysis and 0 to 1 for two-population analysis. To set thresholds for LEI and other genotype-data based feature selection approaches (PCA, SVM and RF), users can click on the radio buttons or enter custom values manually in the provided space. The fourth stage is the marker spacing section where the input SNPs list can be further refined by specifying the desired physical distance between markers. Users can select a distance ranging from 50 kb to 5 Mb from the drop-down field or enter a custom value in the unit of kb (1000 base pairs). The fifth stage is the attribute section in the right column. Users can optionally select several attributes through the collapsible menu after entering the necessary input data. In the last stage, by clicking the display button at the bottom of the page, the selected feature selection method will be performed on the given input data and a computation output page will be generated.

Downloads

The output page contains a table which includes the variant ID, chromosome, reference allele, allele frequencies among the used populations, and the computation values of the selected feature selection method. The output table can be ordered by any columns by clicking the sorting arrow in the column header. The table can also be downloaded to Excel, CSV or PDF files, copied to the clipboard for further use or printed directly. At the bottom of the result page, the previously selected attributes are displayed under each category. Clicking an SNP ID out of the result table redirects to a new page on which the selected attribute information is displayed in a tabulated format. These attribute data are obtained by either directly querying a local API, or providing the hyperlinks to external resources which are shown as the corresponding database logos. For detailed information, we created a fully dedicated website with navigation bar that contains links for manual, frequently asked questions, and contact information and can be accessed at https://research.cchmc.org/mershalab/MI-MAAP/login/. In addition, a detailed documentation including instructions and how to use the tool is provided in Additional file 2: Table S2.

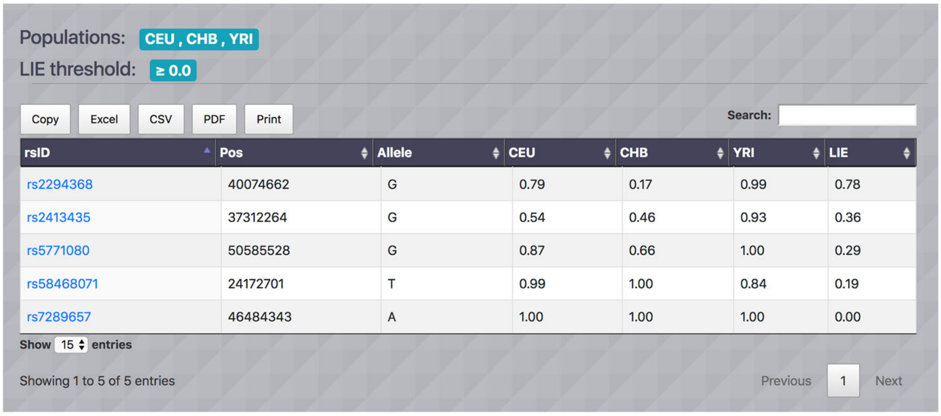
Genome wide scale and computation time and comparison with other methods

The computation time depends on the number of markers and the feature selection threshold as well as the number of populations that are selected. Mining the information for large chromosomes (e.g. chromosome 1 from 1000 Genome Project - about 6.5 million SNPs) takes longer than mining the information for smaller chromosomes (e.g. chromosome 22 from 1000 Genome Project - about 1.1 million SNPs). For example, with threshold LEI > = 0.2, analyzing LEI values of SNPs in chromosome 1 for CEU, CHB and YRI takes about 32 s and gives 233,636 AIMs, while analyzing SNPs in chromosome 22 is about 6 s and gives 43,532 AIMs. With threshold LEI > = 0.8, analyzing SNPs in chromosome 1 and chromosome 22 for the same populations give an output of 1089 AIMs and 243 AIMs, which take about 26 s and 4 s, respectively. Other genotype-based methods are not scalable to a genome-wide data and cannot be directly compared.

Illustrative examples

Ranking ancestry informative markers

Suppose we have identified a set SNPs on chromosome 22 which we want to investigate the ancestry informativeness of the SNPs among three populations CEU (Northern and Western European), CHB (Han Chinese in Beijing, China) and YRI (Yoruba in Ibadan, Nigeria). Using MI-MAAP tool, we compute LEI measure for the desired set of SNPs. Since all SNPs are investigated, we set threshold LEI > = 0. The corresponding output table with a threshold LEI > = 0 is shown in Fig. 4. SNP rs7289657 has LEI = 0



Populations: CEU , CHB , YRI

LIE threshold: ≥ 0.0

rsID	Pos	Allele	CEU	CHB	YRI	LIE
rs2294368	40074662	G	0.79	0.17	0.99	0.78
rs2413435	37312264	G	0.54	0.46	0.93	0.36
rs5771080	50585528	G	0.87	0.66	1.00	0.29
rs58468071	24172701	T	0.99	1.00	0.84	0.19
rs7289657	46484343	A	1.00	1.00	1.00	0.00

Show 15 entries

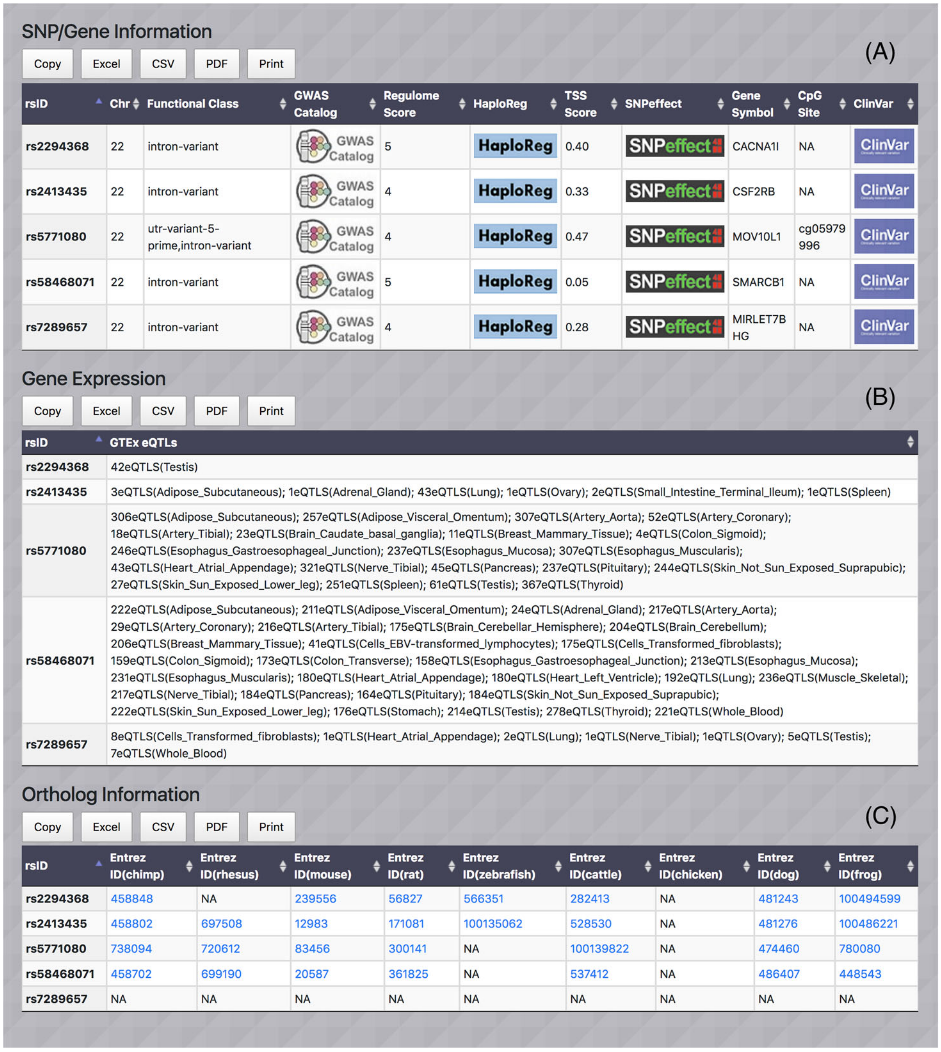
Showing 1 to 5 of 5 entries

Fig. 4 SNPs output display from chromosome 22 for population CEU, CHB and YRI. We have identified a set SNPs on chromosome 22 which we want to investigate the ancestry informativeness of the SNPs among three populations CEU, CHB and YRI. Using MI-MAAP, we compute LEI measure for the desired set of SNPs. The SNP rs2294368 has the highest LEI value (0.78) as it exhibits relatively large allele frequency differences among the three populations (CEU: 0.79, CHB: 0.17, and YRI: 0.99), which suggests that it is the most informative SNP among the SNPs under consideration to be ancestry informative

which implies the SNP is non-informative of the ancestry as expected since the allele frequencies of the three populations are the same. Meanwhile, the SNP rs2294368 has the highest LEI value (0.78) because the SNP exhibits relatively large allele frequency differences among the three populations (CEU: 0.79, CHB: 0.17, and YRI: 0.99), which suggests that it is the most informative SNP among the SNPs under consideration. We further explore the candidate SNP and gene using functional attributes are shown in Fig. 5. The functional annotation scores from RegulomeDB and GWAVA TSS show the SNP rs2294368 has Regulome score = 5 and TSS score = 0.4. By clicking the respective logos, users can get additional association study results from GWAS Catalog and dbGap, and SNP annotation information from HaploReg and SNPeffect. The SNP rs2294368 is located in the Calcium Voltage-Gated Channel Subunit Alpha1 I (CACN A1I) gene. Genome annotation and genomic variation information from ENCODE and ClinVar can be accessed through the hyperlinks provided in the Genome and Variation table. The Gene Expression table displays the number of expression quantitative locus (eQTLs) in different tissues and blood eQTLs for each gene associated with the given SNP. It shows that gene CACNA1I has 42 testis-specific eQTLs and no blood eQTLs. Pathway information is also available by visiting multiple pathway databases KEGG, Reactome and BioCarta. For gene ontology, three main categories, cellular components, biological processes and molecular functions are provided. Gene-associated protein information from four external resources UniProt, Protein Atlas, PFAM and SMART is available. Finally, the Gene Ortholog table contains Entrez Gene IDs for 9 different species (chimpanzee, rhesus, mouse, rat, zebrafish, cattle, chicken, dog and frog) and hyperlinks to the corresponding gene data are provided [18].

Ancestry classification based on 100 AIMs

To demonstrate how MI-MAAP can efficiently extract few AIMs from multiple ancestral populations, and correctly cluster different individuals into their respective



SNP/Gene Information

rsID	Chr	Functional Class	GWAS Catalog	Regulome Score	HaploReg	TSS Score	SNPeffect	Gene Symbol	CpG Site	ClinVar
rs2294368	22	intron-variant	GWAS Catalog	5	HaploReg	0.40	SNPeffect	CACNA1I	NA	ClinVar
rs2413435	22	intron-variant	GWAS Catalog	4	HaploReg	0.33	SNPeffect	CSF2RB	NA	ClinVar
rs5771080	22	utr-variant-5-prime,intron-variant	GWAS Catalog	4	HaploReg	0.47	SNPeffect	MOV10L1	cg05979996	ClinVar
rs58468071	22	intron-variant	GWAS Catalog	5	HaploReg	0.05	SNPeffect	SMARCB1	NA	ClinVar
rs7289657	22	intron-variant	GWAS Catalog	4	HaploReg	0.28	SNPeffect	MIRLET7BHG	NA	ClinVar

Gene Expression

rsID	GTEx eQTLs
rs2294368	42eQTLS(Testis)
rs2413435	3eQTLS(Adipose_Subcutaneous); 1eQTLS(Adrenal_Gland); 43eQTLS(Lung); 1eQTLS(Ovary); 2eQTLS(Small_Intestine_Terminal_Ileum); 1eQTLS(Spleen)
rs5771080	306eQTLS(Adipose_Subcutaneous); 257eQTLS(Adipose_Visceral_Omentum); 307eQTLS(Artery_Aorta); 52eQTLS(Artery_Coronary); 18eQTLS(Artery_Tibial); 23eQTLS(Brain_Caudate_basal_ganglia); 11eQTLS(Breast_Mammary_Tissue); 4eQTLS(Colon_Sigmoid); 246eQTLS(Esophagus_Gastroesophageal_Junction); 237eQTLS(Esophagus_Mucosa); 307eQTLS(Esophagus_Muscularis); 43eQTLS(Heart_Atrial_Appendage); 321eQTLS(Nerve_Tibial); 45eQTLS(Pancreas); 237eQTLS(Pituitary); 244eQTLS(Skin_Not_Sun_Exposed_Suprapubic); 27eQTLS(Skin_Sun_Exposed_Lower_leg); 251eQTLS(Spleen); 61eQTLS(Testis); 367eQTLS(Thyroid)
rs58468071	222eQTLS(Adipose_Subcutaneous); 211eQTLS(Adipose_Visceral_Omentum); 24eQTLS(Adrenal_Gland); 217eQTLS(Artery_Aorta); 29eQTLS(Artery_Coronary); 216eQTLS(Artery_Tibial); 175eQTLS(Brain_Cerebellar_Hemisphere); 204eQTLS(Brain_Cerebellum); 206eQTLS(Breast_Mammary_Tissue); 41eQTLS(Cells_EBV-transformed_lymphocytes); 175eQTLS(Cells_Transformed_fibroblasts); 159eQTLS(Colon_Sigmoid); 173eQTLS(Colon_Transverse); 158eQTLS(Esophagus_Gastroesophageal_Junction); 213eQTLS(Esophagus_Mucosa); 231eQTLS(Esophagus_Muscularis); 180eQTLS(Heart_Atrial_Appendage); 180eQTLS(Heart_Left_Ventricle); 192eQTLS(Lung); 236eQTLS(Muscle_Skeletal); 217eQTLS(Nerve_Tibial); 184eQTLS(Pancreas); 164eQTLS(Pituitary); 184eQTLS(Skin_Not_Sun_Exposed_Suprapubic); 222eQTLS(Skin_Sun_Exposed_Lower_leg); 176eQTLS(Stomach); 214eQTLS(Testis); 278eQTLS(Thyroid); 221eQTLS(Whole_Blood)
rs7289657	8eQTLS(Cells_Transformed_fibroblasts); 1eQTLS(Heart_Atrial_Appendage); 2eQTLS(Lung); 1eQTLS(Nerve_Tibial); 1eQTLS(Ovary); 5eQTLS(Testis); 7eQTLS(Whole_Blood)

Ortholog Information

rsID	Entrez ID(chimp)	Entrez ID(rhesus)	Entrez ID(mouse)	Entrez ID(rat)	Entrez ID(zebrafish)	Entrez ID(cattle)	Entrez ID(chicken)	Entrez ID(dog)	Entrez ID(frog)
rs2294368	458848	NA	239556	56827	566351	282413	NA	481243	100494599
rs2413435	458802	697508	12983	171081	100135062	528530	NA	481276	100486221
rs5771080	738094	720612	83456	300141	NA	100139822	NA	474460	780080
rs58468071	458702	699190	20587	361825	NA	537412	NA	486407	448543
rs7289657	NA	NA	NA	NA	NA	NA	NA	NA	NA

Fig. 5 SNP attributes from chromosome 22 for population CEU, CHB and YRI populations: (**a**) SNP/Gene Information, (**b**) Gene Expression Information and (**c**) Ortholog Information

geographic populations, the top 100 AIMs selected from 1000 Genome Project dataset were generated using LEI. The genotype data of the top 100 AIMs was examined using the principal components analysis (PCA) algorithm to capture the genetic structure. The PCA plots of the top 100 AIMs with the first and second principal components (PCs) for three continental ancestral populations, CEU (Northern and Western European), CHB (Han Chinese in Beijing, China) and YRI (Yoruba in Ibadan, Nigeria) is shown in Fig. 6a. The two-dimensional PCA plot reveals distinct separation of CEU, CHB and YRI racial ancestry populations. The first PC contributes 73.5% of the total variation, which clearly distinguishes between Africans from Yoruba and non-Africans samples from Han Chinese. The second PC, contributing 6.8% of the total variation, distinguishes between Europeans and Han Chinese. Figure 6b shows the PCA plot using the first and second PCAs for the admixed population ASW (African American) and its continental ancestries CEU and YRI. The first PC explains 64.5% of the total variance and the second PC explains 6.4% of the total variance. CEU and YRI samples

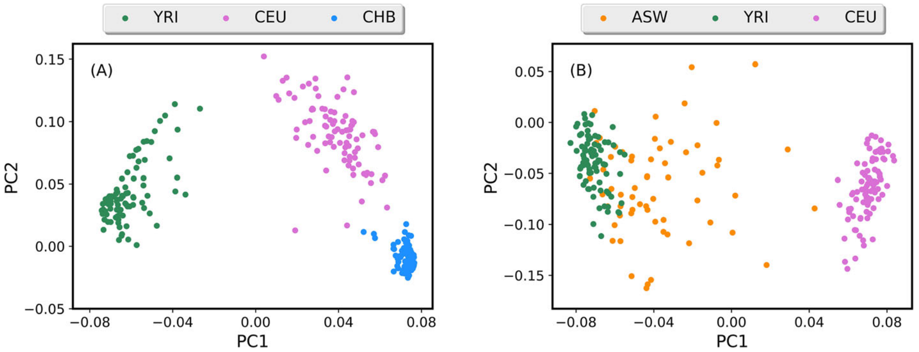


Fig. 6 Scatterplots of principal components axis of PC1 and PC2. (**a**) CEU, CHB and YRI populations: the two-dimensional PCA plot reveals distinct separation of CEU, CHB and YRI racial ancestry populations. The first PC contributes 73.5% of the total variation, which clearly distinguishes between Africans from Yoruba and non-Africans samples from Han Chinese. The second PC, contributing 6.8% of the total variation, distinguishes between Europeans and Han Chinese; (**b**) ASW, CEU and YRI populations: CEU and YRI samples form relatively dense clusters, whereas ASW has a lower density and the sample variance is large. Most of the ASW samples are much closer to YRI than CEU and CEU is separated from the other two populations along the PC1 axis. The first PC explains 64.5% of the total variance and the second PC explains 6.4% of the total variance

form relatively dense clusters, whereas ASW has a lower density and the sample variance is large. Most of the ASW samples are much closer to YRI than CEU and CEU is separated from the other two populations along the PC1 axis. These observations confirm that the African American population is an admixed population with a large contribution from YRI population, and in fact, admixture analyses of ASW samples suggested an average of 80% African ancestry and 20% European ancestry [20–22]. We showed that continental regions can be readily distinguished, while more markers are necessary to improve the classification of closely related and admixed populations. These results further validated that the markers selected from LEI were ancestry informative markers.

Selection of LEI threshold

To provide more insight into the variation of LEI values in different sets of populations, we analyzed the distribution of LEI in multiple population groups. It should be pointed out that the value of LEI is dependent on multiple factors including the number and genetic diversity among the ancestry populations, level of structures one wishes to correct and the sample size. LEI values are in the range of 0 to 1 for two populations and 0 to 2 for three or more populations. Moreover, LEI values from genetically diverse populations tend to be large while the values from similar populations tend to be smaller. When selecting the LEI threshold for AIMs, it is necessary to consider all these factors together and find the optimal LEI value can achieve the desired performance under different scenarios such as classification accuracy or admixture proportion. As a case study of LEI distribution in different sets of populations, we analyzed the distribution of LEI values and PCA clustering for two population groups, CEU, CHB and YRI, and FIN, GBR and TSI. The maximum LEI score for the diverse ancestral populations CEU-YRI-CHB is 1.21 and for the genetically similar populations FIN-GBR-TSI, the maximum LEI score is 0.52. Additional file 3: Figure S1(A) shows the distribution of LEI values from CEU-YRI-CHB populations for SNPs with LEI >0.75. In total, we identified 32,530 SNPs with LEI >0.75 from all autosomal chromosomes and over 2000 SNPs with LEI > 1.0. For the FIN-GBR-TSI populations, we identified 2313 SNPs with LEI > 0.15 and only 251 SNPs with LEI ≥0.25 (Additional file 3: Figure S1(B)). Additional file 4 shows the PCA clustering of the two sets of populations using different LEI thresholds. For CEU-YRI-CHB, SNPs with LEI > 1 separate the three populations as shown in Additional file 4: Figure S2(A). However, for FIN-GBR-TSI, SNPs with LEI ≥0.25 fail to separate three populations and further lowering the LEI threshold to 0.15 improves the PCA clustering (Additional file 4: Figure S2(B-C)). Additionally, only the top 35 LEI SNPs (LEI > 1.0) were adequate to accurately classify the CEU-YRI-CHB populations. A simulation study of the three-way admixed population with CEU, YRI, and CHB as the parental population showed that the top 1000 LEI SNPs (with LEI > 0.833) were required to achieve a root mean square error < 0.05 between the true and estimated ancestry proportion. Because of the variation of the LEI score across different sets of populations, we recommend using an iterative approach to find the optimal set of SNPs that meet the acceptable decision criteria specific to research questions.

Conclusion

Although a large number of SNPs can provide high-resolution genomic information for ancestry, a small but robust set of SNPs may be more desirable for certain applications including population structure and sample classification based on continental ancestry [23]. The small number of ancestry panel is easy to implement in cost-effective routine laboratories for forensic science and disease genetics studies in resource constraint environment. In this study, we developed MI-MAAP, a web-based tool using an allele frequency-based method to identify smaller subsets of informative markers for multi-ancestry admixed populations as well as retrieving SNP/gene related functional annotation. Building MI-MAAP based on allele frequency data without requiring the individual-level genotype data is appealing in the big data era. In summary, we used a case study to demonstrate that MI-MAAP can explore the information in the genomic data more sufficiently and efficiently from allele frequency data. We believe MI-MAAP is a useful tool for SNP panel developments in population, evolutionary, forensic and disease genetics studies.

Availability and requirements

Project name: MI-MAAP.

Project home page: https://research.cchmc.org/mershalab/MI-MAAP/login/

Project documentation page: https://research.cchmc.org/mershalab/MI-MAAP/manual/

Operating system(s): Platform independent.

Programming language: Python, JavaScript, HTML, CSS, PHP.

Other requirements: JavaScript enabled web browsers.

License: GNU General Public License.

Any restrictions to use by non-academics: GNU.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3462-5.

Additional file 1: Table S1. Reference population with labels, sample size that are available from each public database to infer ancestry and web link.

Additional file 2: Table S2. Step-by-step manual: Documentation including instructions and how to use the tool in developing ancestry SNP markers in multi-ancestry population.

Additional file 3: Figure S1. Distribution of LEI scores. (A) Histogram shows the distribution of LEI scores with threshold 0.75 computed among CEU-YRI-CHB populations. (B) Histogram shows the distribution of LEI scores with threshold 0.15 computed among FIN-GBR-TSI populations.

Additional file 4: Figure S2. PCA plots using different LEI thresholds. (A) PC1 vs PC2 produce clear separation of three ancestral populations CEU, YRI, and CHB genome-wide markers with LEI ≥1. (B) PC1 vs PC2 produce some separation of three closely related populations FIN, GBR, and TSI using markers with LEI > 0.15. We have used 2313 markers for the analysis. (C) Using markers with higher thresholds of LEI ≥0.25 failed to separate the three populations. Results were based on all 251 markers with LEI ≥0.25. PLINK 2 was used for the PCA analysis.

Abbreviations

MI-MAAP: Marker informativeness for multi-ancestry admixed populations; AIMs: Ancestry informative markers; GWAS: Genome wide association studies; GTEX: Genotype-Tissue Expression; ENCODE: The Encyclopedia of DNA Elements; LEI: Lancaster Estimator of Independence; PCA: Principal component analysis; SVM: Support vector machine; RF: Random forest; HGDP: Human genome diversity project; ExAC: Exome aggregation consortium; HWE: Hardy-Weinberg Equilibrium; HapMap: Haplotype Map; LD: Linkage disequilibrium; dbGap: Database of Genotypes and Phenotypes

Acknowledgements

Not applicable.

Author's contributions

TBM conceived this project. SC implement and constructed the web application. SC, YG, SG, and TBM wrote the manuscript. TBM supervised the research. All authors read and approved the manuscript.

Funding

This work was supported by the National Institutes of Health (NIH) grant R01HL132344. This funding source had no role in the design of the study and collection, analysis, writing and interpretation of the manuscript.

Availability of data and materials

The data supporting this work is publicly available. The 1000 Genomes Project, HapMap Project, Human Genome Diversity Project (HGDP) and Exome Aggregation Consortium (ExAC) data are available publicly from each web link as provided in Additional file 1. The MI-MAAP tool and its full documentation are freely available from https://research.cchmc.org/mershalab/MI-MAAP/login/. An accompanying Additional file 2 describes the input files with examples.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, 3333 Burnet Avenue, MLC 7037, Cincinnati, OH 45229-3026, USA. [2]Department of Electrical Engineering and Computer Science, University of Cincinnati, Cincinnati, OH 45221, USA.

Received: 22 April 2019 Accepted: 20 March 2020
Published online: 03 April 2020

References

1. Mersha TB. Mapping asthma-associated variants in admixed populations. Front Genet. 2015;6:292.
2. Baye TM, Wilke RA, Olivier M. Genomic and geographic distribution of private SNPs and pathways in human populations. Per Med. 2009;6(6):623–41.
3. International HapMap C. A haplotype map of the human genome. Nature. 2005;437(7063):1299–320.
4. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
5. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319(5866):1100–4.
6. Carithers LJ, Moore HM. The genotype-tissue expression (GTEx) project. Biopreserv Biobank. 2015;13(5):307–8.
7. GTEx. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580–5.
8. Baye TM, Wilke RA. Mapping genes that predict treatment outcome in admixed populations. Pharmacogenomics J. 2010;10(6):465–77.
9. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
10. Consortium GP. A global reference for human genetic variation. Nature. 2015;526(7571):68.
11. Baye TM, He H, Ding L, Kurowski BG, Zhang X, Martin LJ. Population structure analysis using rare and common functional variants. BMC Proc. 2011;5(Suppl 9):S8.
12. Ding L, Wiener H, Abebe T, Altaye M, Go RC, Kercsmar C, Grabowski G, Martin LJ, Khurana Hershey GK, Chakorborty R, et al. Comparison of measures of marker informativeness for ancestry and admixture mapping. BMC Genomics. 2011;12:622.

13. Amirisetty S, Hershey GK, Baye TM. AncestrySNPminer: a bioinformatics tool to retrieve and develop ancestry informative SNP panels. Genomics. 2012;100(1):57–63.
14. Wathen MJ, Gautam Y, Ghandikota S, Rao MB, Mersha TB. LEI: a novel allele frequency-based feature selection method for multi-ancestry admixed populations. Sci Rep. 2019;9(1):11103.
15. Lancaster HO. The Chi-squared Distribution. 1969. New York: Wiley; 1969.
16. International HapMap C. The international HapMap project. Nature. 2003;426(6968):789–96.
17. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45(D1):D840–5.
18. Ghandikota S, Hershey GKK, Mersha TB. GENEASE: real time bioinformatics tool for multi-omics and disease ontology exploration, analysis and visualization. Bioinformatics. 2018;34(18):3160–8.
19. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790–7.
20. Seldin MF, Pasaniuc B, Price AL. New approaches to disease mapping in admixed populations. Nat Rev Genet. 2011; 12(8):523–8.
21. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, et al. A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet. 2004;74(5): 1001–13.
22. Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, Seldin MF. A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am J Hum Genet. 2006;79(4):640–9.
23. Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, Kidd KK. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investig Genet. 2011;2(1):1.

Publisher's Note

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