SOFTWARE

Open Access

MI-MAAP: marker informativeness for multiancestry admixed populations



Siqi Chen^{1,2}, Sudhir Ghandikota^{1,2}, Yadu Gautam¹ and Tesfaye B. Mersha^{1*}

* Correspondence: tesfaye.mersha@ cchmc.org ¹Department 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. AlMs 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



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. 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 in a credit line to the data.

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 $Y = (y_1, y_2, ..., y_k)$ 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 $n = c_1 + c_2 + ... + c_k$ 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, ..., k. Then, the Lancaster estimator of independent (LEI) is defined as –

$$\hat{\theta}^{2} = \sum_{i,j} \frac{\hat{p}_{ij}^{2}}{\hat{p}_{i+}\hat{p}_{+j}} - 1$$
(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 $0 \le \hat{\theta}^2 \le \min(3, k) - 1$ [15]. In particular, for genotype data with biallelic markers, $0 \le \hat{\theta}^2 \le 1$ when the sample constitutes two populations (i.e. k = 2) and $0 \le \hat{\theta}^2 \le 2$ when the sample constitutes three or more populations (i.e. $k \ge 3$).

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.



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.

 $\label{eq:stable} \textbf{Table 1} \text{ 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
НарМар	11	1397	3.2 million
HGDP	53	1043	657,000
ExAC	6	60,706	10 million



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 d	lataset
--	---------

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	Х	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 timeconsuming 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

1	IVIALK		drained Deputations
7 F		s mondaveneed for Mala 7 meessay 7	amixed Populations
s	Select Databa	ase	Attribute
1	_ 1000	Conomec Drainet Diago 2	Select All Attributes
	HanM	an	SNP Information:
		-CEPH	Chromosome
	O ExACI	Exome Aggregation Consortium)	MAF
	O User I	Defined Input	Functional Class
1 1			GWAS Catalog
7			dbGaP
1	000 Genom	es	Exome Variant
	Population:	× CEU: European American (CEPH)	Genome Variant
		× CHB: Chinese (Beijing, China)	Regulome Score
		× YRI: Yoruban (Ibadan, Nigeria)	HaploReg
	Chromosome:	Chromosome 22	TSS Score
			RBP-Var
1 11	Query Option:	SNPs ○ Genes	TFs Mapped
11/11			SNPerrect
	SNP List:	Please enter a single rsID or a list (one rsID per	K Gene Information:
1	Sample	line), e.g. rs587670191	Gene ID
11/1			Gene Symbol
1 7			Synonyms
		SNPs #: 0	Gene Description
			CpG Site(s)
/ /	acture Color	tion Throphold	Mapped Disease(s)
	eature Selec	cion mesnoid	≪ Genome and Variation:
	LIE threshold:	≥ 3 Populations	K Gene Expression:
			Pathway Information:
1		2 Populations 0 0.3 0.6 0.9	Gene Ontology Information:
		Custom Value	S Protein Information:
-N	Aarker Spaci	ng	S Ortholog Information:
1 1	Spacing Between	Markere.	
1	Not Defined	or Custom Value: Kb	Geography of Users
			46 Total Pageviews
	Dutput		
		Display	

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 usergenerated 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

lD	A Pos	\$ Allele	¢ CEU	\$ СНВ	\$ YRI	¢ LIE
2294368	40074662	G	0.79	0.17	0.99	0.78
2413435	37312264	G	0.54	0.46	0.93	0.36
5771080	50585528	G	0.87	0.66	1.00	0.29
58468071	24172701	т	0.99	1.00	0.84	0.19
7289657	46484343	А	1.00	1.00	1.00	0.00
7289657 ow 15 ¢ entries owing 1 to 5 of 5 entrie:	46484343 s	A	1.00	1.00	1.00 Previ	0.00 ous 1

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 A11) 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 CACNA11 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

sID	Chr 🖨	Function	al Class	\$ ⁰	GWAS Catalog	Regulome Score	\$н	laploReg	TSS Score	SNPeffec		Gene Symbol ^{\$}	CpG Site [‡]	ClinVar
s2294368	22	intron-va	riant		GWAS	5		HaploReg	0.40	SNPe	fect	CACNA1I	NA	ClinVa
s2413435	22	intron-va	riant		GWAS	4		HaploReg	0.33	SNPe	fect	CSF2RB	NA	ClinVa
s5771080	22	utr-variar prime,int	nt-5- ron-variant		GWAS Catalog	4		HaploReg	0.47	SNPe	fect	MOV10L1	cg05979 996	ClinVa
s58468071	22	intron-va	riant		GWAS Catalog	5		HaploReg	0.05	SNPe	fect	SMARCB1	NA	ClinVa
s7289657	22	intron-va	riant		GWAS	4		HaploReg	0.28	SNPe	fect	MIRLET7B HG	NA	ClinVa
ene Exp	ressi	ion												4
Copy Ex	cel	CSV	PDF	Print										(B)
sID ^	GTEx	eQTLs												
-0004000														
52294368	42eQ	LS(Testis)											
s2294368 s2413435 s5771080	42eQ 3eQTI 306e0 18eQT 246e0	TLS(Testis _S(Adipose QTLS(Adip TLS(Artery QTLS(Esop) e_Subcutan ose_Subcut _Tibial); 23e hagus_Gas	eous); 1e aneous); 2QTLS(Bi troesoph	eQTLS(Adrenal ; 257eQTLS(Ac rain_Caudate_b nageal_Junction	_Gland); 43eQ lipose_Viscera lasal_ganglia); h); 237eQTLS()TLS(Lu al_Omer ; 11eQT (Esopha	ung); 1eQTLS(O ntum); 307eQTI [LS(Breast_Man agus_Mucosa);	ovary); 2eQ LS(Artery_A nmary_Tissi 307eQTLS	TLS(Small_Ir orta); 52eQ ue); 4eQTLS (Esophagus_	testine_T TLS(Arter Colon_Si Muscular	erminal_lleur y_Coronary); gmoid); is);	n); 1eQTLS	(Spleen)
s2294368 s2413435 s5771080	42eQ 3eQTI 306eQ 18eQT 246eQ 43eQ 27eQT 222eQ	LLS(Testis _S(Adipose TLS(Adip TLS(Artery TLS(Esop TLS(Heart TLS(Skin_S QTLS(Adip	e_Subcutan ose_Subcut Tibial); 23e hagus_Gas Atrial_Appe un_Expose ose_Subcut	eous); 1e aneous); QTLS(Bi troesoph endage); d_Lower_ aneous);	eQTLS(Adrenal ; 257eQTLS(Ac rain_Caudate_t nageal_Junction 321eQTLS(Ne _leg); 251eQTL ; 211eQTLS(Ad	_Gland); 43eQ lipose_Viscera asal_ganglia); n); 237eQTLS(rve_Tibial); 45 S(Spleen); 61e pose_Visceral)TLS(Lu al_Omer ; 11eQT (Esopha ieQTLS eQTLS(I_Omen	ung); 1eQTLS(O ntum); 307eQTI [LS(Breast_Man agus_Mucosa); 5(Pancreas); 233 (Testis); 367eQ ntum); 24eQTLS	Vary); 2eQ LS(Artery_A nmary_Tiss 307eQTLS 7eQTLS(Pit TLS(Thyroi 6(Adrenal_G	TLS(Small_Ir orta); 52eQ ue); 4eQTLS (Esophagus_ uitary); 244d d) iland); 217eQ	testine_T Colon_Si Muscular QTLS(Sk	erminal_lleur y_Coronary); gmoid); is); in_Not_Sun_l ry_Aorta);	n); 1eQTLS Exposed_Si	(Spleen) uprapubic
s2294368 s2413435 s5771080 s58468071	42eQ 3eQTI 306eQ 18eQT 246eQ 43eQ 27eQT 22eQ 29eQ 206eQ 159eQ 231eQ 217eQ 222eQ	ILS(Testis _S(Adiposi _TLS(Adip _TLS(Artery, _TLS(Esop TLS(Heart, _LS(Skin_S _TLS(Adipin TLS(Adipin TLS(Artery _TLS(Artery _TLS(Brea _TLS(Sion) _TLS(Colory _TLS(Sion) _TLS) e_Subcutan pse_Subcut Tibial); 23e hagus_Gas Atrial_Appo un_Expose pse_Subcut _Coronary); st_Mammar a_Sigmoid); hagus_Mus a_Tibial); 18 Sun_Expos	eous); 1e aneous); 2QTLS(Be troesoph endage); d_Lower_ aneous); 216eQT y_Tissue 173eQTI cularis); 4eQTLS(ed_Lowe	eQTLS(Adrenal ; 257eQTLS(Adrenal ; 257eQTLS(Adrenal ; 210eQTLS(Ne]eg); 251eQTL ; 211eQTLS(Ad LS(Artery_Tibi ; 211eQTLS(Ce LS(Colon_Tran 180eQTLS(He [Pancreas); 164 r_leg); 176eQT	_Gland); 43eQ lipose_Viscera asal_ganglia);); 237eQTLS(rve_Tibial); 45 S(Spleen); 61e pose_Visceral al); 175eQTLS(lls_EBV-transf sverse); 158eQ rt_Atrial_Appe leQTLS(Pitults LS(Stomach);	TLS(Lu al_Omer ; 11eQT Esopha ieQTLS eQTLS(LOmen (Brain_(formed QTLS(E endage ary); 18 214eQ	ung); 1eQTLS(O ntum); 307eQTI TLS(Breast_Man agus_Mucosa); (Pancreas); 23; (Testis); 367eQ ntum); 24eQTLS Cerebellar_Hem Lymphocytes); 5sophagus_Gast s); 180eQTLS(H 84eQTLS(Skin, P TLS(Testis); 27	ivary); 2eQ LS(Artery_A mary_Tissi 307eQTLS(Pit TLS(Thyroio G(Adrenal_G iisphere); 2 175eQTLS troesophag eart_Left_V Vot_Sun_Eft_ Vot_Sun_S(T)	TLS(Small_Ir orta); 52eQ Je); 4eQTLS Esophagus, uitary); 2444 d) land); 217eC 04eQTLS(B (Cells_Trans eal_Junctior entricle); 19 posed_Supr yroid); 221e	testine_T LS(Arter Colon_Si Muscular QTLS(Sk TLS(Arter ain_Cerel formed_fi); 213eQ 2eQTLS(L apubic); QTLS(WH	erminal_lleur y_Coronary); gmoid); is); in_Not_Sun_ ry_Aorta); pellum); broblasts); fLS(Esophag ung); 236eC iole_Blood)	n); 1eQTLS Exposed_Si jus_Mucosa JTLS(Musci	(Spleen) uprapubic a); le_Skeleta
sz294368 s2413435 s5771080 s58468071 s7289657	42eQ 3eQTI 306eQ 18eQT 246eQ 43eQ 27eQT 222eQ 29eQ 206eQ 159eQ 231eQ 217eQ 217eQ 222eQ 8eQTI 7eQTL	ILS(Testis _S(Adipose TLS(Adip TLS(Artery, TLS(Esop TLS(Heart, TLS(Skin_S TLS(Adipe TLS(Artery) TLS(Kolo TLS(Colo TLS(Colo TLS(Colo TLS(Nerve TLS(Skin_ _S(Cells_T _S(Whole_ _S(Whole_)) a_Subcutan pse_Subcut Tibial); 23e hagus_Gas Atrial_Appq un_Expose pse_Subcut Coronary); st_Mammar _Sigmoid); hagus_Mus _Tibial); 18 Sun_Expos ransformed Blood)	eous); 1e aneous); aQTLS(Bi troesoph andage); d_Lower aneous); 216eQT y_Tissue 173eQTI cularis); 4eQTLS(ed_Lowe _fibrobla	eQTLS(Adrenal ; 257eQTLS(Ac rain_Caudate_t aageal_Junction 321eQTLS(Ne Jeg); 251eQTL ; 211eQTLS(Ad LS(Arter_Tibi); ; 41eQTLS(Ce LS(Colon_Tran 180eQTLS(Hez (Pancreas); 164 rr_Jeg); 176eQT sts); 1eQTLS(H	Gland); 43eQ ipose_Viscera asal_ganglia);); 237eQTLS(ve_Tibal); 45 S(Spleen); 61e pose_Visceral b); 175eQTLS(IIs_EBV-transf sverse); 158e verse); 158e verse); 158e verse); 158e tal_Appp Lag(Stomach); leart_Atrial_App	2TLS(Lu al_Omer ; 11eQT (Esopha ieQTLS) eQTLS(l_Omen (Brain_(Gormed QTLS(E endage ary); 18 214eQ ppenda	ung); 1eQTLS(Q ntum); 307eQTI LS(Breast_Man agus_Muccosa); 3(Pancreas); 23; (Pancreas); 23; (Testis); 367eQ thum); 24eQTLS Cerebellar_Her Lymphocytes); 5sophagus_Gast Cerebellar_Her Lymphocytes); 5sophagus_Gast 2, 180eQTLS(H MacQTLS(Skin_M 2TLS(Testis); 27 age); 2eQTLS(L	ivary); 2eQ LS(Artery_f nmary_Tiss 307eQTLS(Pit TLS(Thyroio 6(Adrenal_G nisphere); 2 troesophag eart_Left_V Not_Sun_Ex 8eQTLS(T) ung); 1eQTI	TLS(Small_Ir norta); 52eQ je); 4eQTLS [Esophagus_ uitary); 244d d) land); 217eC OdeQTLS(B (Cells_Trans eal_Junctior entricle); 19 posed_Supr nyroid); 221 s.S(Nerve_Til	testine_T [Colon_Si Muscular QTLS(Sk TLS(Arte ain_Cerei ormed_fi); 213eQ 2eQTLS(L apubic); QTLS(Wh ial); 1eQT	erminal_lleur y_Coronary); gmoid); is); in_Not_Sun_l ry_Aorta); pellum); broblasts); fLS(Esophag ung); 236eC iole_Blood) LS(Ovary); {	n); 1eQTLS Exposed_Si jus_Mucosa JTLS(Musci 5eQTLS(Te:	(Spleen) uprapubic; a); le_Skeleta stis);
s_2294368 s2413435 s5771080 s58468071 s7289657 Drtholog	42eq 3eqTI 306e0 18eqTI 246e0 43eq 27eqTI 222e0 29eq 206e0 159e0 231e0 231e0 222e0 8eqTI 7eqTI	ILS(Testis .S(Adipose 2TLS(Adip ILS(Artery, 2TLS(Esop TLS(Heart, ILS(Skin_S) 2TLS(Skin_S) 2TLS(Color 2TLS(Color 2TLS(Color 2TLS(Color 2TLS(Color 2TLS(Color 2TLS(Color 2TLS(Skin_S) 2TLS(Skin_S) .S(Colls_TLS(Skin_S) .S(Whole_) matior) a_Subcutan ose_Subcut Tibial); 23e hagus_Gas Atrial_Appu un_Expose ose_Subcut Coronary); st_Mammar _Sigmoid); hagus_Mus _Tibial); 18 Sun_Expos ransformed 3lood)	eous); 1e aneous); eQTLS(B troesoph andage); d_Lower_ aneous); 216eQT y_Tissue 173eQT cularis); 4eQTLS(ed_Lower_ fibrobla	aQTLS(Adrenal ; 257eQTLS(Ac rain, Caudate, junction 321eQTLS(Ne, jeg); 251eQTLS(Ne jeg); 251eQTLS(Ac LS(Artery, Tibi); 41eQTLS(Ac LS(Artery, Tibi); 41eQTLS(Her Pancreas); 164 T-Jeg); 176eQT sts); 1eQTLS(H	Gland); 43eQ ijpose_Viscerar asal_ganglia);); 237eQTLS S(Spleen); 61 pose_Visceral); 175eQTLS Ills_EBV-transf sverse); 158eG tra_tAtrial_App teQTLS(Phituik LS(Stomach); leart_Atrial_Ap	NTLS(Lu II_Omer 11eQT IEsophic eQTLS eQTLS(I_Omen (Brain_1 Gormed, QTLS(E endage ary); 18 214eQ popenda	ung); 1eQTLS(Q Intum); 307eQT ILS(Breast, Man agus, Mucosa); (Pancreas); 232(G Crestis); 367eQ (Testis); 367eQ (Testis); 367eQ Lymphocytes); 5:09hagus, Gas 5:); 180eQTLS(H 34eQTLS(Skin, J TLS(Testis); 27 age); 2eQTLS(L	tvary); 2eQ LS(Artery_A nmary_Tiss 307eQTLS)? ZeQTLS(Pit TLS(Thyroi 6(Adrenal_C isphere); 2 175eQTLS 175eQTLS troesophag eart_Left\ Vot_Sun_Ex '8eQTLS(TI ung); 1eQTI	TLS(Small_Ir orta); 52eQ Je); 4eQTLS (Esophagus, uitary); 244d d) land); 217eC 04eQTLS(B (Cells_Trans eal_Junction entricle); 19 posed_Supr nyroid); 221e _S(Nerve_Til	testine_T LS(Arter Colon_Si Muscular QTLS(Sk TLS(Arte ain_Cerel formed_fi); 213eQ 2eQTLS(L apubic); QTLS(WH ial); 1eQT	erminal_lleur y_Coronary); gmoid); is); in_Not_Sun_1 ry_Aorta); pellum); broblasts); fLS(Esophag ung); 236eC wole_Blood) 'LS(Ovary); {	n); 1eQTLS Exposed_Si jus_Mucoss RTLS(Muscl 5eQTLS(Ter	(Spleen) uprapubic a); le_Skeletai stis);
s2413435 s2413435 s5771080 s58468071 s7289667 Ortholog Copy ED	42eQ 3eQTI 306ec 18eQT 246ec 43eQ 27eQT 222ec 231eC 231eC 231eC 231eC 231eC 231eC 217eQ 222ec 231eC 217eQT 159ec 231eC 159ec 231eC 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec 159ec	ILS(Testis .S(Adipose 2TLS(Adip TLS(Adip TLS(Adip TLS(Esop TLS(Esop TLS(Esop TLS(Skin_S 2TLS(Adip TLS(Adip TLS(Adip TLS(Color 2TLS(Skin_S 2TLS(Kin S(Kin S(Color 2TLS(Skin LS(Whole_ CSV) a_Subcutan pase_Subcut Tibial); 23c hagus_Gas Atrial_Appo un_Expose pase_Subcut _Coronary); st.Mammar _Sigmoid); st.Mammar _Tibial); 18 Sun_Expos ransformed 3lood) PDF	eous); 1e aneous); aQTLS(Bi troesoph andage); d_Lower_ aneous); 216eQT 216eQT 216eQT cularis); 4eQLLS(ed_Lower_ _fibrobla	22TLS(Adrenal 225AeQTLS(Adrenal 225AeQTLS(Adrenal 2321eQTLS(Ne 1962), 251eQTLS(Ne 1962), 251eQTLS(Ad LS(Atreny_Tbit) 2, 141eQTLS(CLS(Colon_Tran 180eQTLS(Hei Pancreas); TdFancreas); TdFanc	Gland); 43eQ lipose_Viscera asal_ganglia);); 23PeQTLS(vrc_Tibial); 45 S(Spleen); 61c pose_Visceral b); 175eQTLS(lib_ERV-trans); 158e drrt_Atrial_Appe deQTLS(Phiula) LS(Stomach); leart_Atrial_App	RTLS(LU i_Omer 11eQT Esophai eQTLS(cormed QTLS(E endage ary); 18 214eQ oppenda	ung); 1eQTLS(C Intum); 307eQT ILS(Breast, Mana agus, Mucosa); (Pancreas); 233 (Feisis); 387eQ attum); 24eQTLS Cerebellar, Hen Lymphozytes); soophagus, Gast Saophagus, Gast AleQTLS(Kith, Markon); 71LS(Testis); 27 age); 2eQTLS(L	Ivary); 2eQ LS(Artery_A mmary_Tiss 307eQTLS 7eQTLS(Pit TLS(Thyroi (Adrenal_C nisphere); 2 175eQTLS troesophag eart_Left_V Voct_Sun_Ex 18eQTLS(TI ung); 1eQTI	TLS(Small_Ir corta); 52eQ ue); 4eQTLS (Esophagus, titary); 244d d) land); 217eQ 04eQTLS(B (Cells, Trans eal_Junction entricle); 19 posed_Supr nyroid); 221e .S(Nerve_Tii	testine_T "LS(Arter Colon_Si Muscular QTLS(Sk TLS(Arter ain_Cerei ormed_fi); 213eQ 2eQTLS(L apubic); QTLS(WH ial); 1eQT	erminal_lleur y_Coronary); js); in_Not_Sun_l ry_Aorta); broblastophag ung); 2366 hole_Blood) 'LS(Ovary); f	n); 1eQTLS Exposed_Si jus_Mucosa JTLS(Musci	(Spleen) uprapubic; a); le_Skeletai stis); (C)
s2413435 s2413435 s5771080 s58468071 s7289657 Ortholog Copy E s1D 4	42eQ 3eQTI 306e(18eQTI 226e(22eQ 22eQ 22eQ 22eQ 22eQ 22eQ 22eQ 23eQ 22eQ 22	TLS(Testis .S(Adipos: 2TLS(Adip TLS(Adip TLS(Feart TLS(Heart TLS(Adip) TLS(Kin_S) 2TLS(Adip) TLS(Kin_S) 2TLS(Kin_S) 2TLS(Kin_S) TLS(Skin_S) TLS(Skin_S) Colour TLS(Skin_S) Colour TLS(Skin_S) Colour CSV 2 4 mp)	Subcutan Se_Subcut Tibial); 23c hagus_Gas Atrial_App un_Expose Subcut Coronary); st_Mammara Sigmoid); hagus_Mus Tibial); 18 Sun_Expose ansformed Blood) PDF Entrez ID(rhesus	eous); 1¢ aneous); QTLS(B, Vroesoph d_Lower, aneous); 216eQT y_Tissue aneous); 216eQT y_Tissue d_Lower, fibrobla Print print print print	aQTLS(Adrenal ; 257eQTLS(Ac rain, Caudate, j aageal_Junction 321eQTLS(Ne leg); 251eQTLS(Ne leg); 251eQTLS(Ne LS(Atrey_Tbib); 41eQTLS(Ac LS(Atrey_Tbib); 41eQTLS(He Ignarceas); 16 Pancreas); 16 Pancreas; 16 Pancreas	Gland); 43eQ Jacobia Context Jacobia C	RTLS(LL I_Omer ; 11eQT Esophi eQTLS C_Omen (Brain_t QTLS(Omen dgr 214eQ 214eQ ppenda 214eQ ppenda	ung); 1eQTLS(Q ntum); 307eQT LS(Breast, Lma gaus, Mucosa); (Pancreas); 232 (Testis); 367eQ (Testis); 367eQ Lymphocytes); 5:09hagus, Gas ;) 180eQTLS(H 34eQTLS(Skin, <u>1</u> TLS(Testis); <u>27</u> age); 2eQTLS(L	LS(Artery_A nmary_Tiss 307eQTLS 7eQTLS(Pit TLS(Thyroio 6(Adrenal_G iisphere); 2 175eQTLS 175eQTLS 175eQTLS 175eQTLS 175eQTLS 175eQTLS 10; 10; 10; 10; 10; 10; 10; 10; 10; 10;	TLS(Small_Ir TLS(Small_Ir S2eQ 19; AeQTLS Esophagus, uitary); 2444 3) land); 217eC 04eQTLS(B (Cells_Trans ea_Junction entricle); 19 posed_Supr yroid); 221E .S(Nerve_Til)))))))))))))	testine_T "LS(Arter Colon_Si 2 Muscular Mu	erminal_lleur y_Coronary); gmoid); is); in_Not_Sun_l ry_Aorta); pellum); broblasts); TLS(Esopha usle_Blood) 'LS(Ovary); f Covary); f	n); 1eQTLS Exposed_Si jus_Mucoss jTLS(Muscl 5eQTLS(Ter 5eQTLS(Ter 5eQTLS(Ter 5eQTLS(Ter 5eQTLS(Ter 5eQTLS(Ter 5eQTLS(Ter 5eQTLS(Ter 5eQTLS(Ter 5eQTLS))	(Spleen) uprapubic; a); le_Skeleta stis); (C) ntrez)(frog)
s22134368 s2413435 s5771080 s58468071 s7289657 Ortholog Copy E3 s1D 4 s2294368	42eQ 3eQTI 306e(18eQT 226eQ 22eQ 206e(159eQ 23eQ 23eQ 23eQ 23eQ 23eQ 23eQ 23eQ 23	ILS(Testis .S(Adipos: 2TLS(Adipos: 2TLS(Adipos: 2TLS(Adipos: 2TLS(Adipos: 2S(Adipos: 2S(Skin_S) 2TLS(Adipos: 2TLS(Adipos: 2TLS(Skin_S) 2TLS(Skin_S) 2TLS(Skin_S) 2TLS(Skin_S) 2S(Cells_TI .S(Whole_ CSV z mp) \$	e_Subcutan pse_Subcutan pse_Subcut hagus_Gas Atrial_Appu um_Expose sse_Subcut Coronary); st_Mammar _Sigmoid); hagus_Mus _Tibialy; 18 Sun_Expos ansformed Blood) PDF Entrez ID(rhesus NA	eous); 1c aneous); qCTLS(Rey troesoph nndage); 216eqT y_Tissue 173eqT 173eqT 173eqT Lower 4eqTLS(cularis); fibrobla Print \$	2QTLS(Adrenal ; 257eQTLS(Ac rain, Caudate, Junction 321eQTLS(Ne jeg); 251eQTLS(Ne jeg); 251eQTLS(Ne LS(Artery_Tibi); 41eQTLS(Ad LS(Artery_Tibi); 41eQTLS(Ad LS(Artery_Tibi); 14eQTLS(He Figuress); 15 Figuress); 15 Figuress); 15 Figuress); 16 Figuress);	Gland); 43eQ Gland); 43eQ asal_ganqlia); h); 237eQTLS(tve_Tibial); 45 S(Splent); 61 pose_Visceral h); 175eQTLS(lls_EBV-transfills, EBV-transfills, EBV-transfills, EBV-transfills, LS(Stomach); leart_Atrial_App eCTLS(Pittial_App leart_Atrial_App Entrez D(rat) 56827	RTLS(LL I_Omerr 11eQTLS eQTLS(L_Omen (Brain_1_ QTLS(E formed 2TLS(E comendage ary); 18 214eQ oppenda	ung); 1eQTLS(C ntum); 307eQT LS(Breast, Man agus, Mucosa); (Pancreas); 233(Trastis); 367eQ (Trastis); 367eQ (Trastis); 367eQ Lymphocytes); 5:09hagus, Gas); 180eQTLS(Ksin, J 130eQTLS(Ksin, J 130eQTLS(LS); 3:09hagus, Gas); 2eQTLS(L 130eQTLS(LS); 3:09hagus, Gas); 2eQTLS(L); 3:09hagus, Gas); 3:09hagus, Gas); 3:09hagu	Vary); 2eQ S(Artery_F mary_Tisssi 307eQTLS 307eQTLS 7eQTLS(Pit TLS(Thyroio (Adrenal_C inisphere); 2 175eQTLS 175eQTLS 175eQTLS 175eQTLS 175eQTLS 175eQTLS 175eQTLS 10(cattld 282413	TLS(Small_Ir TLS(Small_Ir S2eQ Weix 40 TLS (Esophagus, Uitary); 2444 3) Iand); 217eC 04eQTLS(B (Cells_Trans ea_Junction entricle); 19 posed_Supr syroid); 221e .S(Nerve_Tit)) (NA NA	testine_T "LS(Arter Colon_5") Muscular Muscular Muscular TLS(Arter armed_fi,); 213eQ 213eQ 123eQ	erminal_lleur y_Coronary); gmoid); is); is); ry_Aorta); pellum); broblasts); TLS(Esophagung); 236eG uole_Blood) TLS(Ovary); { Entresz ID(dog 48124	n); 1eQTLS Exposed_SI jus_Mucoss TTLS(Muscl 5eQTLS(Te: 5eQTLS(Te: 5eQTLS(Te: 5eQTLS(Te: 5eQTLS(Te: 5eQTLS(Te: 5eQTLS(Te: 5eQTLS(Te: 5eQTLS(Te: 5eQTLS) 5eQTLS(Te: 5eQTLS) 5eQTLS(Te: 5eQTLS) 5eQTLS(Te: 5eQTLS) 5eQTLS 5eQTLS(Te: 5eQTLS) 5eQTLS) 5eQTLS(Te: 5eQTLS) 5eQTLS) 5eQTLS(Te: 5eQTLS) 5eQTLS) 5eQTLS(Te: 5eQTLS) 5eQTLS) 5eQTLS) 5eQTLS(Te: 5eQTLS) 5e	(Spleen) uprapubic; a); le_Skeletal stis); (C) ntrez p(frog) 00494599
sz2494368 s2413435 s5771080 s558468071 s7289657 Ortholog Copy E s1D E s2294368 s2413435	42eQ 3eQTI 306ect 18eQT2 246ect 43eQ 27eQ 22eeQ 231eQ 231eQ 231eQ 231eQ 231eQ 231eQ 231eQ 232ect 8eQT1 Infor Entre: ID(chi 4588/ 4588/	TLS(Testis S(Adipose 2TLS(Adip) TLS(Atery) TLS(Kesop TLS(Kesop TLS(Kesop TLS(Kesop TLS(Kesop TLS(Color TLS(Sesop TLS(Color TLS(Sesop TLS(Kesop TLS	e_Subcutan pse_Subcutan pse_Subcut hagus_Gas Atrial_Appu um_Expose sse_Subcut Coronary); st.Mammar _Sigmoid); hagus_Mus _Tibialy; 18 Sun_Expos ansformed Blood) PDF PDF Entrez ID(rhesus NA 697508	eous); 1e aneous); vroesoph andage); 216eQTI 73eQT 173eQT	2QTLS(Adrenal ; 257eQTLS(Ac rain, Caudate, J aageal_Junction 321eQTLS(Ne leg); 251eQTLS(Ne leg); 251eQTLS(Ne LS(Artery_Tbib); 41eQTLS(Ac LS(Artery_Tbib); 41eQTLS(Ac LS(Colon_Train 810eQTLS(Het Prancreas); 16 Prancreas); 16 Prancreas; 16	Gland); 43eQ lipose_Viscera asal_ganglia); h); 237eQTLS(X); pose_Visceral h); 757eQTLS(I); pose_Visceral h); 757eQTLS(I); lipose_Visceral h); 757eQTLS(I); h);	PTLS(LL al_Omera Esophate garLS) garLS(E Brain_1 formed 27LS(E Brain_1 formed 27LS(E Brain_1 formed 27LS(E Brain_1 formed 27LS(E Brain_1 formed 27LS(E Brain_1 formed 27LS(E formed D) formed 27LS(E formed D) formed 27LS(E formed D) formed 27LS(E formed D) formed 27LS(E formed D) formed for	ung); 1eQTLS(C ntum); 307eQT LS(Breast, Man gaus, Mucosa); (Pancreas); 232(Testis); 367eQ (Testis); 367eQ Lymphocytes); 5:09hagus, Gas ;): 180eQTLS(Ksin, J TLS(Testis); 27 age); 2eQTLS(L trez zabrafish) (8351 0135062	Entrez Bolzesta 282413 282413 282413 282413 282413 282530	TLS(Small_Ir TLS(Small_Ir S2eQ be); &eQTLS (Esophagus, uitary); 2444 d) aland); 217eC 04eQTLS(B (Cells_Trans ea_Junction entricle); 19 posed_Supry yroid); 2216 S(Nerve_Til D) (NA NA NA	testine_T "LS(Arter Colon_5" Muscular QTLS(Sk TLS(Arter and Color (Colo	erminal_lleur y_Coronary); gmoid); is); ry_Aorta); pellum); broblasts); ILS(Esophagung); 236eC inole_Blood) 'LS(Ovary); 5 Entress ID(dog 48124 48127	n); 1eQTLS Exposed_SI jus_Mucoss TTLS(Muscl 5eQTLS(Te:) ¢ Ef) 10 6 10	(Spleen) uprapubic, a); le_Skeletaa stis); ((C) 00494599 00494599 00494599
sz2494368 s2413435 s5771080 s558468071 s7289657 Copy E s2294368 s2413435 s5771080	42eQ 3eQTI 306ec 18eQT 246ec 246ec 22ec 22ec 22ec 231eC 231eC 231eC 231eC 231eC 231eC 231eC 231eC 231eC 231eC 231eC 22ec 206ec 4589 4589 4589 73805 73805	TLS(Testis S(Adipos) DTLS(Adip) DTLS(Adip) TLS(S(Hart, LS(Atrey) TLS(S(Hart, LS(Atrey) DTLS(Adip) TLS(Adip) TLS(S(Hart, LS(Atrey) TLS(Atrey) TL	Subcutan bas_Subcutan base_Subcut Tibial); 23& hagus_Gas Atrial_Appuu_ Expose sse_Subcut Coronary); st_Mammar _Sigmoid); hagus_Mus _Tibial); 18 Sun_Expos ansformed Blood) PDF Entrez ID(rhesus NA 697508 720612	eous); 1e aneous); yqTLS(B aneous); 216eQTI 73eQT 1	2QTLS(Adrenal ; 257eQTLS(Ac rain, Caudate, J aageal_Junction 321eQTLS(Ne leg); 251eQTLS(Ne leg); 251eQTLS(Ne LS(Atrey_Tbib); 41eQTLS(Ac LS(Atrey_Tbib); 41eQTLS(He ISC(Join, Train 80eQTLS(He ISC(Join, Train 80eQTLS(He ISC)); 1eQTLS(He ISC)); 1eQTLS(HE));	Gland); 43eQ lipose_Viscera asal_ganglia);); 237eQTLS(vv; Tibial); 45 S(Spleen); 61c pose_Visceral a)); 175eQTLS(lis_E8V-trans); 158c tr.Atrial_Apg leart_Atrial_Ap; leart_Atrial_Ap; leart_Atrial_Ap; Entrez ID(rat) 556827 171081 300141	PTLS(LU al_Omera 11eQT Esophic eQTLS) eQTLS(comend QTLS(C	ung); 1eQTLS(Q ntum); 307eQT [LS(Breast, Man agus, Mucosa); (Pancreas); 232(Testis); 367eQ (Testis); 367eQ (Te	Entrez ID(cattle) 282413 1001998 1001998 1001998 1001998 1001998 1001998 1001998 1001998 1001998	TLS(Small_Ir TLS(Small_Ir S2eQ be); &eQTLS (Esophagus, uitary); 2444 d) aland); 217eC 04eQTLS(B (Cells_Trans ea_Junction entricle); 19 posed_Supr yrroid); 2216 .S(Nerve_Til D) k k NA NA 22 NA	testine_T LS(Arter Wascular ain_Cere formed_file apubic); 213eQTLS(W apubic); 213eQ apubic); 213eQ apubic); 213eQ apubic); 213eQ apubic); 213eQTLS(W apubic); 213eQTLS	erminal_lleur y_Coronary); gmoid); is); ry_Aorta); pellum); broblasts); ILS(Esophaz, ung); 236eC inole_Blood) 'LS(Ovary); 5 E blocd inole_Blood) 'LS(Ovary); 5 Clock inole_Blood) 'LS(Ovary); 5 Clock inole_Blood)	n); 1eQTLS Exposed_SI jus_Mucoss TTLS(Muscl 5eQTLS(Te:) ¢ Ef (1) 3 10 6 10 74	(Spleen) uprapubic le_Skeletal stis); (C) 00494599 00494599 00494599 00494599 00494599 00494599 00494599 00494599 00494599 0049595 00495959 004959 004959 00495959 00495959 004959
s22194368 s2413435 s5771080 s5771080 s57289657 Ortholog Copy E s2294368 s2413435 s5771080 s58468071	42eQ 3eQTI 306et 18eQTI 18eQTI 246et 43eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 159Q 22eQ 23eQ 159Q 22eQ 23eQ 22eQ 23eQ 159Q 22eQ 23eQ 159Q 22eQ 23eQ 22eQ 23eQ 159Q 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 23eQ 22eQ 22	ILS(Testis ILS(Atleps) ITLS(Atlep) ITLS(Est) ITLS(Est) ItLS(Heart ILS(Heart ILS(Heart ILS(Heart) IL	Subcutan Se_Subcut Tibial); 23c hagus_Gas Atrial_Appuu_Expose Se_Subcut Coronary); st_Mammar _Sigmoid); hagus_Mus _Tibial); 18 Sun_Expos _Tibial); 18 Sun_Expos _ansformed ansfo	eous); 1e aneous); VQTLS(B); VQTLS(B); VQTLS(B); VQTLS(B); 216eQT 726eQT 73eQT 73eQT 73eQT 73eQT 74eQTLS; 4eQLLS; 4eQLLS; 4eQLLS; 4eQLLS; 6 Fibrobla 9 Fib	20TLS(Adrenal ; 257eQTLS(Adrenal ; 257eQTLS(Adrenal ; 257eQTLS(Adrenal ; 257eQTLS(Adrenal ; 251eQTL ; 211eQTLS(Adrenal ; 211eQT	Gland); 43eQ lipose_Viscera asal_ganglia);); 23PeQTLS(vv; Tibial); 45 S(Spleen); 61c pose_Visceral a)); 175eQTLS(lis_E8V-trans); 158c tr.: Atrial_Ape teQTLS(Pituita LS(Stomach); leart_Atrial_Ape teQTLS(Pituita LS(Stomach); leart_Atrial_Ape teQTLS(Pituita LS(Stomach); leart_Atrial_Ape teQTLS(Pituita LS(Stomach); leart_Atrial_Ape 10(rat) 56827 171081 300141 361825	Content of the second sec	ung); 1eQTLS(Q ntum); 307eQTI (LS(Breast_Man agus_Muco3); (Pancreas); 233((Pancreas); 233(C Creste); 367eQ (Trestis); 367eQ Lymphocytes); 5:0800TLS(H 34eQTLS(Ksin_J TLS(Testis); 27 age); 2eQTLS(L trez zobrafish) 63551 0135062	LS(Artery); 2eQ LS(Artery),4 Immary,Tissi 307eQTLS(Prit TLS(Thyroio) 6(Adrenal_G isphere); 2 175eQTLS Troesophag eart_Left_V Not_Sun_Ex 8eQTLS(T) Ung); 1eQTI ID(cattle 282413 528530 1001398 537412	LLS(Small_Ir orta); 52eQ Je); 4eQTLS (Esophagus, uitary); 244a J) land); 217eQ OdeQTLS(B (Cells_Trans eal_Jiol2 (Cells_Trans eal_Jiol2 S(Nerve_Til D(NA NA 22 NA NA	testine_T LS(Arter Wascular ain_Cere orrmed_file apubic); 213eQ QTLS(Wi ial); 1eQ T S V apubic); 1eQ T S V apubic); 213eQ V apubic); 212eQ V apubic); 212eQ V apubic); 212eQ V apubic); 212eQ V apubic); 212eQ V apubic); 212eQ V apubic); 212eQ V apubic); 212eQ V apubic); 212eQ V apubic); 212eQ V Apubic); 212eQ V V Apubic); 212eQ V V V V V V V V V V V V V V V V V V V	erminal_leur y_Coronary); gmoid); is); in_Not_Sun_l ry_Aorta); pellum); broblasts); TLS(Esophac ung); 236eC isole_Blood) 'LS(Ovary); f 48127 48127 48127 48127	m); 1eQTLS Exposed_Si pus_Mucoss DTLS(Musci 5eQTLS(Tes 5)	(Spleen) uprapubic le_Skeleta stis); (C) 00494596 0049596 0049566 0049566 0049566 0049566 0049566 0049566 0049566 0049566

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



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 64.4% 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 \geq 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 \geq 0.25 failed to separate the three populations. Results were based on all 251 markers with LEI \geq 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

¹Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, 3333 Burnet Avenue, MLC 7037, Cincinnati, OH 45229-3026, USA. ²Department 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.
- 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.
- 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.

- 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.
- 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.
- 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.
- 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.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

