MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2010) **10**, 1085–1088 doi: 10.1111/j.1755-0998.2010.02862.x

COMPUTER PROGRAM NOTE

PEAS V1.0: a package for elementary analysis of SNP data

SHUHUA XU,*† SANCHIT GUPTA* and LI JIN*†‡§

**Chinese Academy of Sciences and Max Planck Society (CAS-MPG) Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, †Key Laboratory of Computational Biology at CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China, ‡State Key Laboratory of Genetic Engineering and Ministry of Education (MOE) Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China, §China Medical City (CMC) Institute of Health Sciences, Taizhou, Jiangsu 225300, China*

Abstract

We have developed a software package named PEAS to facilitate analyses of large data sets of single nucleotide polymorphisms (SNPs) for population genetics and molecular phylogenetics studies. PEAS reads SNP data in various formats as input and is versatile in data formatting; using PEAS, it is easy to create input files for many popular packages, such as STRUCTURE, *frappe*, Arlequin, Haploview, LDhat, PLINK, EIGENSOFT, PHASE, fastPHASE, MEGA and PHYLIP. In addition, PEAS fills up several analysis gaps in currently available computer programs in population genetics and molecular phylogenetics. Notably, (i) It calculates genetic distance matrices with bootstrapping for both individuals and populations from genome-wide high-density SNP data, and the output can be streamlined to MEGA and PHYLIP programs for further processing; (ii) It calculates genetic distances from STRUCTURE output and generates MEGA file to reconstruct component trees; (iii) It provides tools to conduct haplotype sharing analysis for phylogenetic studies based on high-density SNP data. To our knowledge, these analyses are not available in any other computer program. PEAS for Windows is freely available for academic users from http://www.picb.ac.cn/~xushua/index.files/Download_PEAS.htm.

Keywords: computer program, molecular evolution, population genetics, SNP

Received 1 February 2010; revision received 15 March 2010; accepted 2 April 2010

Introduction

Recently available genome-wide data of high-density single nucleotide polymorphisms (SNPs) and the advent of next generation whole-genome sequencing data for human populations demarcated a transition from single-locus-based studies to genomics analysis of human population structure and relationship (Rosenberg *et al.* 2002, 2005; The International HapMap Consortium 2005; Friedlaender *et al.* 2008; Jakobsson *et al.* 2008; Kayser *et al.* 2008; Li *et al.* 2008). With the release of the Phase III HapMap data (http://hapmap.ncbi.nlm.nih.gov), a resource consisting of over 1.5 million SNPs genotyped in more than 1000 individuals from 11 geographically diverse populations is publicly available. The high-density genome-wide SNP data for 53 worldwide populations in Human Genome Diversity Panel (HGDP) (Cann *et al.* 2002) are also available recently (Li *et al.* 2008). In addition, a few international projects focusing on regional populations such as PanAsian SNP Project (The HUGO Pan-Asian SNP Consortium 2009) generated additional SNP data resources. The modern human genetic studies have been dramatically influenced by the development and release of these data; as a consequence, our insight and knowledge about human genome has been greatly improved because of the analysis of those SNP data.

Many software tools have been developed to extract abundant information from such large data sets. However, most of the software tools have distinctive format for input files, and it often takes much time to format the input as well as the output for additional analyses. This poses a particular challenge to biologists who are uncomfortable with programming and handling large data sets. Furthermore, some commonly used analyses are not

Correspondence: Shuhua Xu, Fax: +86 21 54920451; E-mail: xushua@picb.ac.cn or Li Jin, Fax: +86 21 65643714; E-mail: ljin007@gmail.com

included in the available software, such as calculating individual allele sharing distance, population genetic distances, haplotype sharing analyses for phylogenetic studies, forward-time simulation studies to test human migration models based on haplotype sharing analyses and so on. In addition, even for some basic data manipulations, the software available could not either handle or work very well with large data sets. We have developed a software package named PEAS to provide the users with many basic data handling and analysis tools for large SNP data set.

Features

Dynamic memory management was adopted, and all the tools in PEAS were developed to handle large SNP data set with high efficiency. All the operations of PEAS programs are file(s) to file(s), although PEAS allows the user to display results in the graphical user interface. Therefore, for the very large data set, which will take huge memory to display on the screen, the user can choose not to display data and let program run as a background process.

PEAS is versatile in manipulating data. First, it provides tools for data formatting, which facilitates the user to manipulate data prior to further analysis. These include: (i) a tool to manipulate HapMap genotype data, which formats HapMap data for various purposes; (ii) a format conversion tool to transpose data between columns and rows, and a coding translation tool to allow PEAS recognize various data formats and unify data obtained from different resources; (iii) a data split tool to allow the user to split data into multiple sets by samples or by chromosomes or by both. For example, the users may like to separate the parents from the kids of trio samples in most of the cases; (iv) a data integration tool to allow the user to integrate multiple data sets by samples or by chromosomes or by both. For example, one can integrate data sets from different population samples while performing downstream analyses; (v) a series of tools to provide the user to prepare input files for many popular softwares including STRUCTURE (Pritchard *et al.* 2000), *frappe* (Tang *et al.* 2005), Arlequin (Schneider *et al.* 2000), Haploview (Barrett *et al.* 2005), LDhat (McVean *et al.* 2004), PLINK (Purcell *et al.* 2007), EIGENSOFT (Patterson *et al.* 2006), MEGA (Kumar *et al.* 2004), PHYLIP (Felsenstein 1989), PHASE (Stephens *et al.* 2001) and fastPHASE (Stephens & Donnelly 2003); 6) a tool to allow user to format the output haplotype results of fastPHASE and PHASE as the input file of STRUCTURE, Haploview, Arlequin etc.

Second, PEAS provides tools for some basic manipulations of SNP data. These include: (i) a tool to allow the user to calculate allele and genotype frequencies and to test for deviation from the HWE (using Chi-square test); (ii) a filter tool to allow the user to filter the data by MAF, missing data proportion and HWE states; (iii) a sampling tool to allow the user to sample the subsets of data by individuals or by markers or by both; (iv) a tool to allow the users to retrieve the consensus data for multiple population samples or different resources. The tool integrates data according to the information of SNP ID, chromosome, physical position and strand (+/−).

Third, to fill up the gaps of currently available software tools, tools have been developed to focus on the population genetic analysis and phylogenetic analysis are developed. These include: (i) a tool to allow the user to calculate the allele sharing distance between each pair of individuals. This tool will generate multiple distance matrixes by bootstrapping the loci and provides the output files that can be read by MEGA (Kumar *et al.* 2004) and PHYLIP (Felsenstein 1989) programs for further processing; (ii) a tool to allow the user to calculate the distances for populations and generates multiple distance matrixes by bootstrapping the loci. The population distances supported by PEAS are Wright's F_{ST} (Weir & Hill 2002), F_{ST} distance (Latter 1972), Nei's standard distance (Nei 1972), Nei's D_A distance (Nei *et al.* 1983) and Cavalli-Sforza's D_C distance (Cavalli-Sforza & Edwards 1967). The tool also generates the output files which can be used by MEGA and PHYLIP programs for further processing; (iii) a tool to allow the user to calculate the two most commonly used LD statistics (r^2 and $|D'|$) (Lewontin 1964; Hill & Weir 1994) and to generate LD distribution report files which can be used to plot using MS Excel. This feature is especially useful for very large data set with huge number of SNP sites (Xu *et al.* 2007); (iv) a tool to carry out haplotype sharing analysis based on high-density SNP genotyping data (The HUGO Pan-Asian SNP Consortium 2009; Xu *et al.* 2009a); (v) a tool to carry out the forward-time simulations to test the evolutionary models (The HUGO Pan-Asian SNP Consortium 2009; Xu *et al.* 2009a). A summary list of the component programs and functions is shown in Table 1.

Applications

PEAS was applied in recent studies to estimate the individual and the population distances (WI *et al.* 2008; The HUGO Pan-Asian SNP Consortium 2009) for population phylogenetic analyses, calculate the LD (Xu *et al.* 2007, 2008), integrate HapMap and HGDP data (Xu & Jin 2008, 2009; The HUGO Pan-Asian SNP Consortium 2009; Xu *et al.* 2009b), conduct haplotype sharing analyses and forward-time simulation (The HUGO Pan-Asian SNP Consortium 2009; Xu *et al.* 2009a).

An executable version of PEAS along with the documentation and example data files can be freely

Table 1 The component programs and functions in PEAS package

Procedure	Program name	Function
Manipulates HapMap Data	HapMap_Data_to_Standard	Converts HapMap data to PEAS format
Transposing data between columns and rows	colTorow	Converts Linkage-like format to PEAS format
	rowTocol	Converts PEAS standard format to Linkage-like format
Re-coding genotype data	DAAtoD11	Recodes single nucleotide polymorphism (SNP) genotypes from ACGT to 11, 12, 22
	DAAtoS123	Recodes SNP genotypes from ACGT to 1, 2, 3
	DAAtoSABH	Recodes SNP genotypes from ACGT to A, B, H
	D11toS123	Recodes SNP genotypes from 11, 12, 22 to 1, 2, 3
	D11toSABH	Recodes SNP genotypes from 11, 12, 22 to A, B, H
	D11toDAA	Recodes SNP genotypes from 11, 12, 22 to ACGT
	S123toD11	Recodes SNP genotypes from 1, 2, 3 to 11, 12, 22
	S123toDAA	Recodes SNP genotypes from 1, 2, 3 to ACGT
	S123toSABH	Recodes SNP genotypes from 1, 2, 3 to A, B, H
	SABHtoD11	Recodes SNP genotypes from A, B, H to 11, 12, 22
	SABHtoS123	Recodes SNP genotypes from A, B, H to 1, 2, 3
	SABHtoDAA	Recodes SNP genotypes from A, B, H to ACGT
Data splitting	Split_by_Chr	Splits data by chromosomes
	Split_by_POP	Splits data according to population affinities
Data filtering	filter	Filters data by MAF, missing data and HWE
Consensus markers	Shared_loci	To obtain the consensus data
Data integration	Combine_sample	Combines data for multiple population samples
	Combine_snp	Combines data for different SNP markers
Basic statistics	allele_count	Calculates number of alleles for each SNP
	allele_freq	Calculates allele frequency for each SNP
	genotype_count	Calculates number of genotypes for each SNP
	genotype_freq	Calculates genotype frequency for each SNP
	hwe	To test HWE for each SNP
Data sampling	SNP_Sampler	Generates subsets of data by sampling markers
Individual distance	Ind_dis_wBootStrap	Calculates distance between individuals
Population distance	POP_dis	Calculates genetic distance, such as FST between populations
LD calculator	Pairwise_LD	Calculates the two commonly used LD statistics (r^2 and $\|D'\|$)
Haplotype-sharing analysis	HaploSharing	Calculates haplotype-sharing statistics.
Input for other programs	Arlequin_in	Generates Arlequin input file
	fastPHASE_in	Generates fastPHASE input file
	frappe_in	Generates frappe input file
	Haploview_in	Generates Haploview input file
	LDhat_in	Generates LDhat input file
	PLINK_in	Generates PLINK input file
	STRUCTURE_in	Generates STRUCTURE input file

downloaded for pure academic use from http://www.picb.ac.cn/~xushua/index.files/Download_PEAS.htm. The project is intended to remain active, and new features will be added to the software.

Acknowledgements

This work was supported by grants from the National Outstanding Youth Science Foundation of China (30625016), the National Science Foundation of China (30890034, 30971577) and the 863 Program (2007AA02Z312). LJ was also supported by the Shanghai Leading Academic Discipline Project (B111) and the Center for Evolutionary Biology. SX was also supported by the Science and Technology Commission of Shanghai Municipality (09ZR1436400) and the Knowledge Innovation Program of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences (2008KIP311). SX gratefully acknowledges the support of SA-SIBS Scholarship Program and K.C.Wong Education Foundation, Hong Kong.

References

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Cann HM, de Toma C, Cazes L *et al.* (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.

Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, **19**(Suppl 19), 233–257.

Felsenstein J (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

Friedlaender JS, Friedlaender FR, Reed FA *et al.* (2008) The genetic structure of Pacific Islanders. *PLoS Genetics*, **4**, e19.

Hill WG, Weir BS (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. *American Journal of Human Genetics*, **54**, 705–714.

Jakobsson M, Scholz SW, Scheet P *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.

Kayser M, Lao O, Saar K *et al.* (2008) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *American Journal of Human Genetics*, **82**, 194–198.

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefing in Bioinformatics*, **5**, 150–163.

Latter BD (1972) Selection in finite populations with multiple alleles. 3. Genetic divergence with centripetal selection and mutation. *Genetics*, **70**, 475–490.

Lewontin RC (1964) The interaction of selection and linkage. Ii. Optimum models. *Genetics*, **50**, 757–782.

Li JZ, Absher DM, Tang H *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.

McVean GA, Myers SR, Hunt S *et al.* (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Nei M (1972) Genetic distance between populations. *American Naturalist*, **106**, 283–292.

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J Mol Evol*, **19**, 153–170.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Rosenberg NA, Pritchard JK, Weber JL *et al.* (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Rosenberg NA, Mahajan S, Ramachandran S *et al.* (2005) Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, **1**, e70.

Schneider S, Roessli D, Excoffier L (2000) *Arlequin: A Software for Population Genetics Data Analysis. Ver 2.000*. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, Geneva.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, **73**, 1162–1169.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, **28**, 289–301.

The HUGO Pan-Asian SNP Consortium (2009) Mapping human genetic diversity in Asia. *Science*, **326**, 1541–1545.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Weir BS, Hill WG (2002) Estimating F-statistics. *Annual Review of Genetics*, **36**, 721–750.

WI WNH, AR NS, MK Z *et al.* (2008) Genetic relationship & distribution of ancestral genetic component among Peninsular Malaysia Malay Sub-Ethnic Groups. *The Malaysian Journal of Medical Sciences*, **15** (Suppl 1), 26.

Xu S, Jin L (2008) A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *American Journal of Human Genetics*, **83**, 322–336.

Xu S, Jin L (2009) Genetic landscape of Eurasia and ''admixture'' in Uyghurs Response. *American Journal of Human Genetics*, **85**, 937–939.

Xu S, Huang W, Wang H *et al.* (2007) Dissecting linkage disequilibrium in african-american genomes: roles of markers and individuals. *Molecular Biology and Evolution*, **24**, 2049–2058.

Xu S, Huang W, Qian J, Jin L (2008) Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *American Journal of Human Genetics*, **82**, 883–894.

Xu S, Jin W, Jin L (2009a) Haplotype-sharing analysis showing Uyghurs are unlikely genetic donors. *Molecular Biology and Evolution*, **26**, 2197–2206.

Xu S, Yin X, Li S *et al.* (2009b) Genomic dissection of population substructure of Han Chinese and its implication in association studies. *American Journal of Human Genetics*, **85**, 762–774.