Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
ELSEVIER
GENOMICS PROTEOMICS & BIOINFORMATICS

Application Note

SNPTransformer: A Lightweight Toolkit for Genome-Wide Association Studies

Changzheng Dong*

School of Medicine, Ningbo University, Ningbo 315211, China.

Genomics Proteomics Bioinformatics 2010 Dec; 8(4): 268-273 DOI: 10.1016/S1672-0229(10)60029-0

Abstract

High-throughput genotyping chips have produced huge datasets for genome-wide association studies (GWAS) that have contributed greatly to discovering susceptibility genes for complex diseases. There are two strategies for performing data analysis for GWAS. One strategy is to use open-source or commercial packages that are designed for GWAS. The other is to take advantage of classic genetic programs with specific functions, such as linkage disequilibrium mapping, haplotype inference and transmission disequilibrium tests. However, most classic programs that are available are not suitable for analyzing chip data directly and require custom-made input, which results in the inconvenience of converting raw genotyping files into various data formats. We developed a powerful, user-friendly, lightweight program named SNPTransformer for GWAS that includes five major modules (Transformer, Operator, Previewer, Coder and Simulator). The toolkit not only works for transforming the genotyping files into ten input formats for use with classic genetics packages, but also carries out useful functions such as relational operations on IDs, previewing data files, recoding data formats and simulating marker files, among other functions. It bridges upstream raw genotyping data with downstream genetic programs, and can act as an in-hand toolkit for human geneticists, especially for non-programmers. SNPTransformer is freely available at http://snptransformer.sourceforge.net.

Key words: genome-wide association studies, tool, genotyping files, conversion

Introduction

High-throughput genotyping technologies contribute greatly to the hunt for susceptibility genes for complex diseases by constantly improving the precision of and the capacity for parallel genotyping (*1-3*). Driven by these emerging technologies, some challenging projects, such as HapMap Phase I~III (*4, 5*), ENCODE (*6, 7*) and 1000 Genomes (*8*), were proposed to explore the pattern in the human genome of common or rare variation. Broad application of the whole-genome single nucleotide polymorphism (SNP) chips into genome-wide association studies (GWAS) has also led to the discovery of more than 100 loci for nearly 40 common diseases and traits (*9*). Due to the great challenges presented by huge datasets, two strategies have been developed for performing data analysis for whole-genome genotyping data. One strategy is to use open-source or commercial packages that have been designed for GWAS. PLINK (*10*) is one of the most popular and powerful of these programs. GenABEL (*11*), GWAF (*12*), SNPAssoc (*13*) and snpMatrix (*14*)

*Corresponding author.
E-mail: dongchangzheng@nbu.edu.cn

are programs based on R language, which is a well-executed open-source framework. GWAS Analyzer (*15*), GWAS GUI (*16*) and MAVEN (*17*) provide platforms for intuitively viewing the results of GWAS. SNPTEST (*18*) also can be used for data analysis with a software suite consisting of several programs. The above software packages are easily compatible with chip data and possess more or fewer functions according to the purpose of genetic analysis. The other strategy for performing these analyses is to take advantage of classic genetic programs that implement specific functions, such as transmission disequilibrium tests (TDT) for alleles [UNPHASED (*19*)], calculating linkage disequilibrium (LD) measures and constructing LD maps [Haploview (*20*), GOLD (*21*) and JLIN (*22*)], haplotype inference [PHASE (*23*)], haplotype block partition [HapBlock (*24*)], tagSNPs selection [Tagger (*25*)] and multilocus interaction methods [MDR (*26*)]. However, most of the classic programs that are available would be not suitable for inputting the chip data directly and require custom-made input, which results in the inconvenience of converting raw genotyping files into various data formats. SNP_Tools (*27*) is an MS-Excel add-in that can convert genotyping files into several formats, such as Haploview and PHASE. Because it is an add-in program for Excel (255 columns and 65,536 rows in MS-Excel 2003), supports for the chip data are scarce. Furthermore, output formats are limited to Haploview, SNPHAP, PHASE and PedPhase. However, as classic programs implement specific algorithms for genetic analysis and provide an option for analysis of GWAS data, it is important to develop a tool to bridge these programs with raw genotyping data.

Here, we present a powerful, user-friendly, lightweight toolkit named SNPTransformer for GWAS. The major aim of SNPTransformer is to convert genotyping input (such as linkage and chip formats) into various outputs (such as packages for association, TDT, calculating LD measures, haplotype inference, haplotype block partition, tagSNPs and multilocus interaction). With this toolkit, researchers can avoid manual coding between formats and can easily construct workflows for data analysis. Additionally, accessory tools in SNPTransformer perform data previewing, relational operations on IDs, recoding data files and simulating map files that assist data conversion and GWAS analysis.

Implementation

SNPTransformer V1.0 was written using C++ Builder with a concise and user-friendly interface. It was built and tested under Windows XP. Because SNPTransformer is a lightweight toolkit, no installation or other package is required and it is compatible with other Windows platforms. All binary programs, source codes, tutorials, examples and updates are available freely under the GNU General Public License at the SNPTransformer website (http://snptransformer.sourceforge. net).

Results

The current version of SNPTransformer provides five major modules for GWAS: Transformer, Previewer, Operator, Coder and Simulator (**Table 1**).

Transformer

As the most important module of this software package, Transformer is positioned in the main window of the software and is responsible for converting file formats from genotyping data to the formats of

Table 1 Modules and functions of SNPTransformer

Module	Function	Example
Transformer	Converting genotyping files into formats of other analysis tools	Converting chip data into PLINK input
Previewer	Previewing the first N lines of large data files	Previewing annotation files of Affymetrix SNP Array 6.0
Operator	Relational operations on IDs	Retrieving annotation information for positive SNPs
Coder	Recoding genotypes between letters and numbers	Recoding AB-coding genotypes into ACGT-coding
Simulator	Simulating map or pedigree files	Simulating PLINK map files according to a SNP list

specific classic genetics analytical tools (**Figure 1** and Table 1). Input data include genotyping files, marker files and pedigree files (**Figure 2**). Formats of genotyping files can be one of the following: linkage format, which integrates genotyping data and pedigree information in a single file; chip format, which is compatible with whole-genome genotyping data of Affymetrix, Illumina and many other chip platforms (see manual in detail); or custom format, which is similar to sequencing results. A marker file stores chromosome information and the physical positions of SNPs, as well as genetic positions, which are usually set to zero. Pedigree files contain information on individual IDs, gender and disease status, or qualitative traits, and this file type is suitable not only for pedigree data such as linkage and TDT, but also for case/control data. Because a pedigree file is the same as the pedigree part (first six columns) of a linkage format, it is not required for a linkage format genotyping file.

The output formats of SNPTransformer are diverse and represent essential genetic analyses for GWAS (Figure 2). Similar to many other tools, the linkage format is considered the basic format of SNPTransformer. PLINK is one of the most popular software platforms for GWAS and can perform various genetic analyses, and the marker file of SNPTransformer references that of PLINK, leading to consistency between the input data in linkage format for SNPTransformer and those of PLINK. Haploview is a program that presents LD-based analysis tools: TDT, calculating LD measures and constructing LD maps, inferring haplotypes, partitioning haplotype blocks and selecting tagSNPs and case/control association. UNPHASED, GOLD, JLIN, PHASE, HaploBlock and Tagger each carry out one of these analysis steps. MDR excels at performing multi-locus interaction analysis and is widely used in association studies. At present, no tool related to two-locus interaction is included in SNPTransformer due to the lack of well-recognized

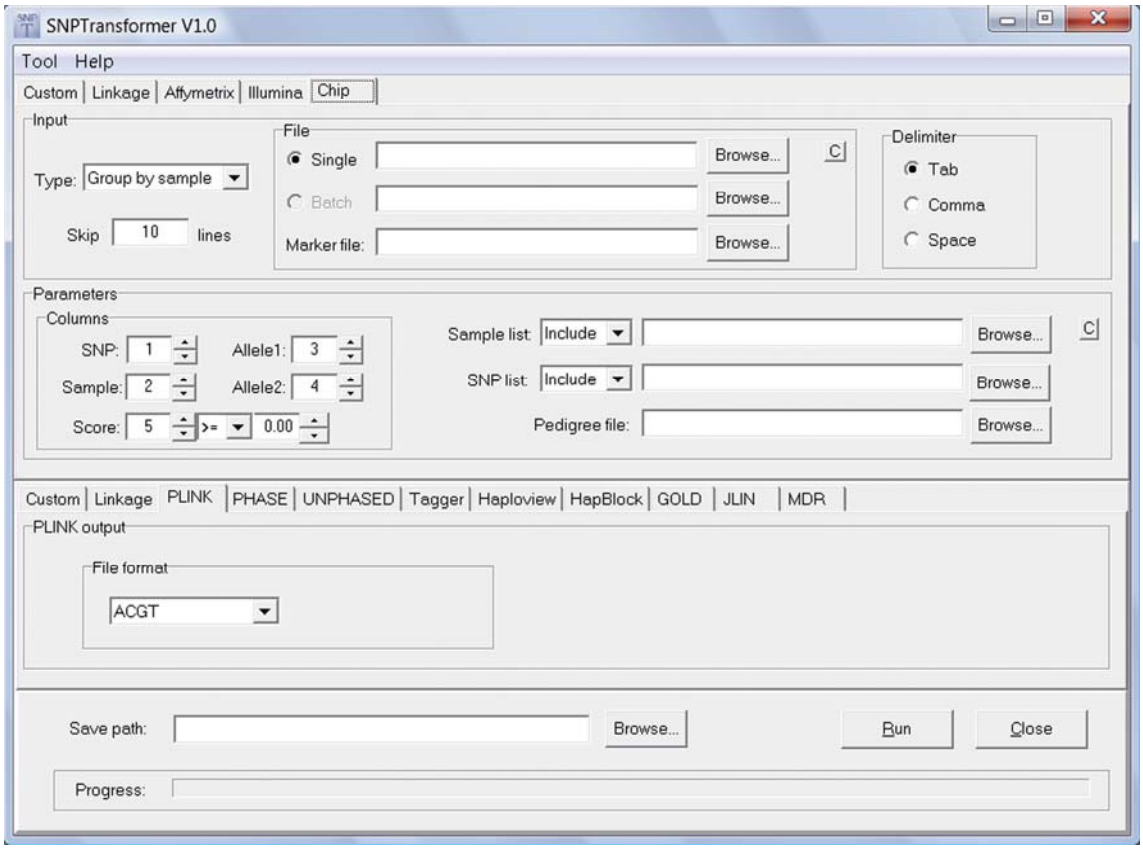

Figure 1 Screenshot of SNPTransformer. The screenshot shows the user-friendly interface of SNPTransformer. Transformer is located in the main interface of SNPTransformer and consists of two windows. The upper window is used to set input files and their relevant parameters, and the output formats are designated from the bottom.

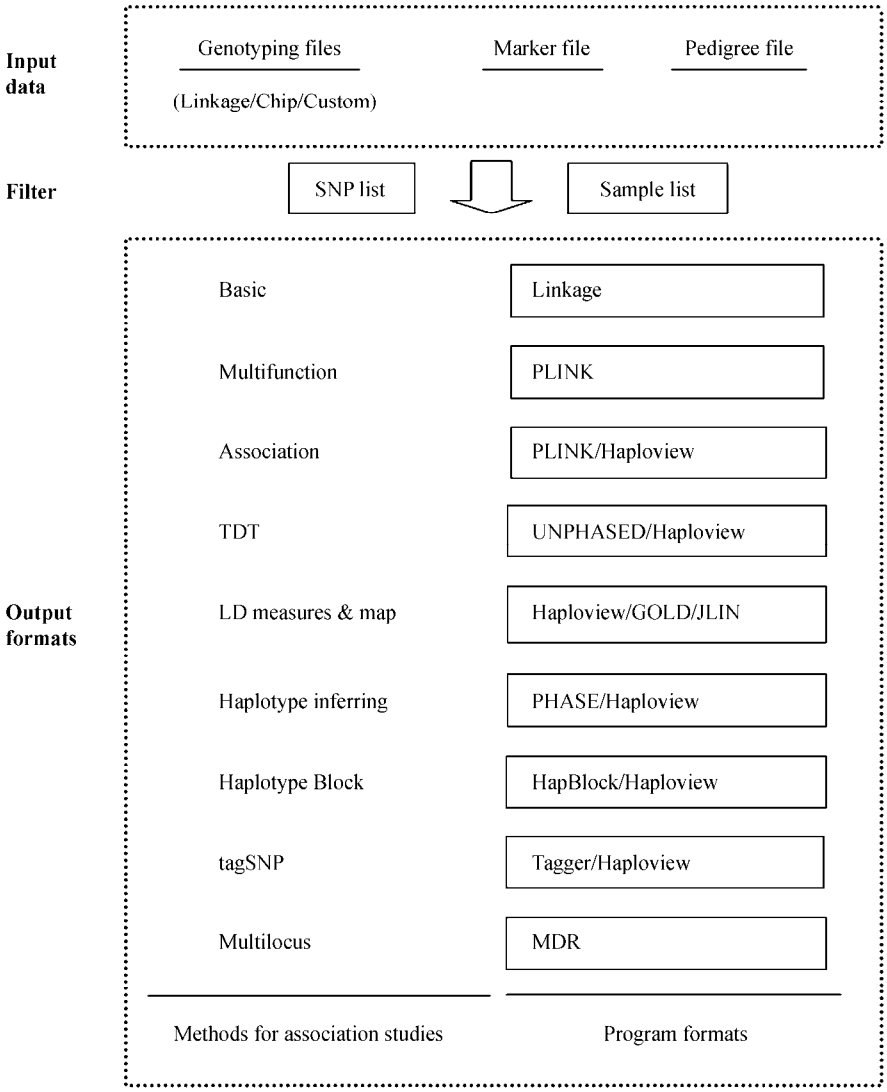

Figure 2 Workflow of Transformer. Input data of Transformer include genotyping files (linkage, chip or custom formats), marker files and pedigree files, filtered by sample and SNP lists. The output formats cover routine analyses for GWAS, and ten programs are selected as representatives.

analysis tools for this except for logistic regression, which has quite a different format from the output of SNPTransformer. Additional custom formats can be output using some options that are designed to count genotype/allele numbers and frequencies. To filter input data, SNP or sample lists can be set to improve data quality or to narrow the target scope. With these filters, Transformer can act as an extractor that searches for specific genotypes for meta-analysis.

Previewer, Operator, Coder and Simulator

The other four modules in SNPTransformer have been developed to satisfy specific demands, such as previewing data files and retrieving annotation information for positive SNPs (Table 1). Since genotyping or annotation files of whole-genome SNP chips are too large to open with generic text-editor tools, Previewer performs previewing of the top N lines of these files. With this function, users can view the format of annotation files for Affymetrix GeneChip 6.0 sets that are larger than hundreds of megabytes. Furthermore, Previewer can reorder the columns of input file to attain a new file that is arranged as required for further analysis. During the processing of genetic analyses, the relational operations can help search for specific information. For example, annotation information for positive SNPs can be retrieved by performing

the "inner join" or "left join" operation between the SNP list and the annotation file. Operations including one-item (operating on single file) and two-item operations (operating between two files) are implemented by Operator, similar to relational databases such as MySQL and Access support. Coder recodes genotypes between letters and numbers that can code one heterozygote with two alleles "A" and "T" as "AT" (ACGT-coding), "14" (1234-coding), "12" (12-coding), "AB" (AB-coding) or even "1" (012-coding). Another module, called Simulator, is used to generate pseudo map and pedigree files without the use of real information. When using Haploview to calculate LD measures, physical positions are not required to be real numbers, but rather the order of SNPs is sufficient if the pairwise distance can be ignored and if the order of SNPs is correct. In such a case, a map file can be easily simulated from a SNP list by Simulator.

Considerable work remains to be accomplished in the future to meet the needs of GWAS analysis. The first step in this process is to adopt parallel technology to further improve the speed of the analysis process. Another important aim in improving SNPTransformer is to design a personal interface for Affymetrix and Illumina SNP chips to provide more options. In the current version, these were implemented through chip interface.

Acknowledgements

This work was supported by Zhejiang Provincial Natural Science Foundation (No. Y2100240), Ningbo Natural Science Foundation (No. 2009A610142), Zhejiang Provincial Health Department Foundation (No. 2009A183), the Hulan Scholar Fund and the K.C. Wong Magna Fund at Ningbo University. The author is grateful for Yi Huang's contribution to this work, and also thanks Dr. Lingyi Lu's help with revising the manuscript.

Competing interests

The author has declared that no competing interests exist.

References

1 Margulies, M., *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.
2 Gunderson, K.L., *et al.* 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* 37: 549-554.
3 Matsuzaki, H., *et al.* 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* 1: 109-111.
4 International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437: 1299-1320.
5 International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.
6 ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* 306: 636-640.
7 Thomas, D.J., *et al.* 2007. The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res.* 35: D663-667.
8 Kaiser, J. 2008. DNA sequencing: a plan to capture human diversity in 1000 genomes. *Science* 319: 395.
9 Manolio, T.A., *et al.* 2008. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* 118: 1590-1605.
10 Purcell, S., *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81: 559-575.
11 Aulchenko, Y.S., *et al.* 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23: 1294-1296.
12 Chen, M.H. and Yang, Q. 2010. GWAF: an R package for genome-wide association analyses with family data. *Bioinformatics* 26: 580-581.
13 Gonzalez, J.R., *et al.* 2007. SNPassoc: an R package to perform whole genome association studies. *Bioinformatics* 23: 644-645.
14 David, C. and Hin-Tak, L. 2007. An R package for analysis of whole-genome association studies. *Hum. Hered.* 64: 45-51.
15 Fong, C., *et al.* 2010. GWAS analyzer: integrating genotype, phenotype and public annotation data for genome-wide association study analysis. *Bioinformatics* 26: 560-564.
16 Chen, W., *et al.* 2009. GWAS GUI: graphical browser for the results of whole-genome association studies with high-dimensional phenotypes. *Bioinformatics* 25: 284-285.
17 Narayanan, K. and Li, J. 2010. MAVEN: a tool for visualization and functional analysis of genome-wide association results. *Bioinformatics* 26: 270-272.
18 Marchini, J., *et al.* 2007. A new multipoint method for genome-wide association studies by imputation of geno-

types. *Nat. Genet.* 39: 906-913.
19 Dudbridge, F. 2008. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.* 66: 87-98.
20 Barrett, J.C., *et al.* 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263-265.
21 Abecasis, G.R. and Cookson, W.O. 2000. GOLD—graphical overview of linkage disequilibrium. *Bioinformatics* 16: 182-183.
22 Carter, K.W., *et al.* 2006. JLIN: a java based linkage disequilibrium plotter. *BMC Bioinformatics* 7: 60.
23 Stephens, M. and Donnelly, P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* 73: 1162-1169.
24 Greenspan, G. and Geiger, D. 2004. High density linkage disequilibrium mapping using models of haplotype block variation. *Bioinformatics* 20: i137-144.
25 de Bakker, P.I., *et al.* 2005. Efficiency and power in genetic association studies. *Nat. Genet.* 37: 1217-1223.
26 Hahn, L.W., *et al.* 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19: 376-382.
27 Chen, B., *et al.* 2009. SNP_tools: a compact tool package for analysis and conversion of genotype data for MS-Excel. *BMC Res. Notes* 2: 214.