BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 3 2004, pages 439–443
DOI: 10.1093/bioinformatics/btg446

Frequency Finder: a multi-source web application for collection of public allele frequencies of SNP markers

Tu H. Nguyen[1,], Chunyu Liu[1,2], Elliot S. Gershon[1] and Francis J. McMahon[3]*

[1]Department of Psychiatry, University of Chicago, IL 60616, USA [2]National Laboratory of Medical Genetics of China, Central South University, Changsha Hunan, 410078, People's Republic of China and [3]Mood and Anxiety Disorder Program, National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20892, USA

Received on August 19, 2003; accepted on September 17, 2003
Advance Access Publication January 22, 2004

ABSTRACT

Summary: Publicly available single nucleotide polymorphism (SNP) allele frequencies are an important resource for the selection of genetic markers that may be most useful for gene mapping and association studies. Data mining these allele frequencies through disparate public databases and Websites is time consuming and can result in inconsistent findings. We have developed a web-based software tool, Frequency Finder, to acquire SNP allele frequencies from multiple public data sources and return a summarized result to the user. Our software optimizes and automates the search of candidate markers, decreasing the amount of time it would take to extract pertinent data manually. We have included several methods to output the data, including on-screen and as a compressed text file. We show that Frequency Finder accurately retrieves available frequency data from the available sources. Using this tool, we detect significant differences between Asian, African and Caucasian populations in the allele frequency spectra of 246 097 SNPs. While limited to public databases that provide web-based access to allele frequencies, Frequency Finder provides a single, user-friendly interface for retrieving allele frequencies for large batches of SNPs from multiple data sources.

Availability: Frequency Finder is available for public use at http://mapgenetics.nimh.nih.gov/frequencyfinder and at http://bluegenes.bsd.uchicago.edu/frequencyfinder

Contact: tnguyen@yoda.bsd.uchicago.edu

INTRODUCTION

Single nucleotide polymorphisms (SNPs) have become the marker of choice for association studies with disease genes because they are abundant, easier than microsatellites to genotype automatically, and informative for linkage disequilibrium mapping when selected for appropriate allele frequencies. The amount of publicly available SNPs and allele frequency data is rapidly increasing, making the efficient selection of optimally informative SNPs increasingly important (Brookes, 1999; Marth *et al*., 2001). Limiting genotyping studies to a single data source can result in incomplete information including false positives and high degrees of homology in gene families (Marsh *et al*., 2002).

Resources for allele frequency data include The SNP Consortium (Thorisson and Stein, 2003), dbSNP (Wheeler *et al*., 2003), ALFRED (Cheung *et al*., 2000), HGVBase (Fredman *et al*., 2002) and Celera Genomics. However, the efficient selection of polymorphic SNPs from multiple public databases is not intuitive. Each of these databases differs in data structure, update frequency and sources of frequency data. For example, TSC often has the most up-to-date data, but dbSNP contains an accumulation of frequencies from submitters other than TSC, and Celera Genomics provide population-specific allele frequencies from their own genotyping assays. ALFRED and HGVBase do not use the same dbSNP identifiers as the other databases, further complicating data mining efforts.

Recent work has attempted to centralize and automate the process of SNP extraction from disparate public databases (Riva and Kohane, 2002). For example, SNPper, a web application developed by The Children's Hospital Informatics Program (http://snpper.chip.org), allows the extraction of SNPs within a genomic region, along with allele frequency information gathered from TSC. Other existing tools are valuable, but focus on obtaining detailed information about individual SNPs. Often, data mining these SNPs can be inefficient because data such as allele frequencies are not available

*To whom correspondence should be addressed at Knapp Research Center, 924E 57th Street, Chicago, IL 60637, USA.

at one central location. Frequency Finder offers investigators an efficient method to select and screen large lists of publicly available SNPs for subsequent studies. We describe here the development and implementation of Frequency Finder, provide data supporting the validity of Frequency Finder data retrievals, demonstrate the use of Frequency Finder to perform a large-scale allele frequency analysis and discuss the strengths and limitations of this tool in practical use.

DEVELOPMENT AND IMPLEMENTATION

Frequency Finder is a web application that uses a browser-based interface to input data into the Java framework to perform search and analysis. The application runs under an open source web server in a UNIX environment. Java Server Pages (JSP) allows interaction between the users and the application, including data and file processing. Java servlets are used to parse the form data, along with uploaded files. We use the O'Reilly public Java Application Program Interface (API) to perform the multipart processing of the uploaded file.

We import reference data into the application through dynamic and local resources.

(1) *Local database queries* UNIX shell scripts are integrated into the application to perform automated downloads of the raw frequency data from the SNP Consortium (http://snp.cshl.org) and from NCBI dbSNP (http://www.ncbi.nlm.nih.gov/SNP). This data is converted and stored in a MySQL 4.0 database server (http://www.mysql.com). Indexing key identifiers optimizes batch queries that are made from the uploaded SNP list.

(2) *Real-time data request from external resources* Additional SNP allele frequencies are obtained through a real-time programmatic request to the Applied Biosystems' (http://abassays.celera.com) genomic assays public data, which is an interface to the Celera database. Our software parses the HTML data and validates the presence and ordering of expected data by examining the column headings. If the ordering of the expected data has changed, the program will try to rearrange the data to the correct positions. The program will abort seeking data from ABI, continue with the local search, and notify the user and the administrator of the failed attempt if the validation techniques fail to be resolved. Otherwise, the relevant matches are appended into the result set. This algorithm can be used to integrate with additional resources as these become available.

The user inputs a list of SNPs into the system either by uploading a line-delimited file, by entering the SNP identifiers into the HTML text box, or by specifying what chromosomal region boundaries to query positionally. Chromosomal regions map to the latest available builds from the NCBI's sequence map. Our system currently queries against dbSNP accession (rs) numbers, dbSNP individual submission accession numbers (ss) and local TSC (tsc) identifiers.

After processing the requested SNPs, the program outputs an HTML-generated table of dbSNP accession (rs) numbers, allele frequency data (both major and minor alleles, if available), populations typed and data sources (Fig. 1). We used the rs accession numbers as the preferred output because (1) these numbers are usually stable after they have been validated and (2) most gene mapping studies use this accession number as the standard SNP nomenclature. SNP allele frequencies from multiple sources are displayed in adjacent columns. Genomic map positions are also included with the results, facilitating SNP selection for further study. The user has the option to save the results for future reference, print them, or export them in a compressed, tab-delimited ASCII file for further analysis.

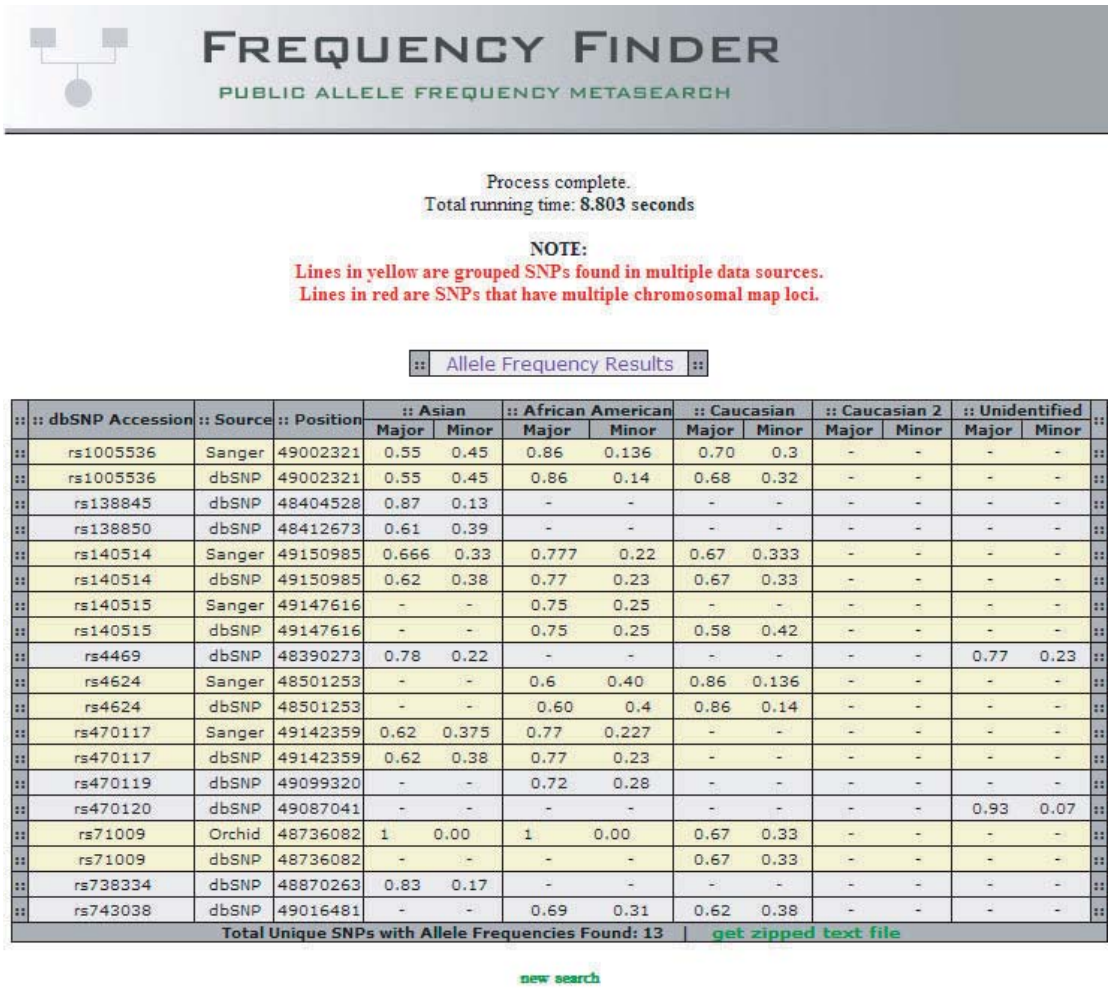
RESULTS

Data sources. Frequency Finder was used to search 11 449 SNPs randomly selected from a Chromosome 15 by use of the UCSC Genome Browser (http://genome.ucsc.edu). The search, performed in April, 2003, took 3 min and retrieved 582 SNPs with associated allele frequencies in one or more ethnic groups. Of these, 318 frequency results (55%) were unique to dbSNP, 109 (19%) were unique to ABI and 56 (9%) were unique to TSC. The remaining 99 (18%) were found in multiple sources.

Validation of negatives. No frequency data were found for 10 867 SNPs. In order to verify that no public frequency data actually existed for these SNPs in any of the databases searched, a random sample of 100 SNPs that produced negative results were searched manually against ABI, TSC and dbSNP, and also automatically with SNPper. No additional frequency data was found for any of these SNPs in any of the databases we searched. Given the sample size of 100, the actual false negative rate has a 95% probability of lying between 0.0 and 3.62%.

Large-scale searching. To demonstrate the value of Frequency Finder for large-scale searching of SNP allele frequencies, we searched all 4 145 589 SNPs available in the NCBI dbSNP Build 114 using Frequency Finder. Allele frequencies were returned for 246 097 SNPs (5.9%). As demonstrated above with the chromosome 15 SNPs, a substantial fraction of SNPs (179 926 or 73%) had allele frequency data available unique to one public database (Fig. 2).

The retrieved data was used to compare SNP allele frequency spectra among three populations represented in the dataset, Asians (162 706 frequencies), African Americans (136 289 frequencies) and Caucasians (97 210 frequencies) (Fig. 3). The spectra differed significantly between populations ($\chi^2 = 12\,191$, df $= 6$, $P < 0.0001$). Asian and African American allele frequency spectra were similar, with over 76% of SNPs returning common minor allele frequencies of



FREQUENCY FINDER

PUBLIC ALLELE FREQUENCY METASEARCH

Process complete.
Total running time: **8.803 seconds**

NOTE:
Lines in yellow are grouped SNPs found in multiple data sources.
Lines in red are SNPs that have multiple chromosomal map loci.

Allele Frequency Results

:: dbSNP Accession	:: Source	:: Position	:: Asian		:: African American		:: Caucasian		:: Caucasian 2		:: Unidentified	
			Major	Minor	Major	Minor	Major	Minor	Major	Minor	Major	Minor
rs1005536	Sanger	49002321	0.55	0.45	0.86	0.136	0.70	0.3	-	-	-	-
rs1005536	dbSNP	49002321	0.55	0.45	0.86	0.14	0.68	0.32	-	-	-	-
rs138845	dbSNP	48404528	0.87	0.13	-	-	-	-	-	-	-	-
rs138850	dbSNP	48412673	0.61	0.39	-	-	-	-	-	-	-	-
rs140514	Sanger	49150985	0.666	0.33	0.777	0.22	0.67	0.333	-	-	-	-
rs140514	dbSNP	49150985	0.62	0.38	0.77	0.23	0.67	0.33	-	-	-	-
rs140515	Sanger	49147616	-	-	0.75	0.25	-	-	-	-	-	-
rs140515	dbSNP	49147616	-	-	0.75	0.25	0.58	0.42	-	-	-	-
rs4469	dbSNP	48390273	0.78	0.22	-	-	-	-	-	-	0.77	0.23
rs4624	Sanger	48501253	-	-	0.6	0.40	0.86	0.136	-	-	-	-
rs4624	dbSNP	48501253	-	-	0.60	0.4	0.86	0.14	-	-	-	-
rs470117	Sanger	49142359	0.62	0.375	0.77	0.227	-	-	-	-	-	-
rs470117	dbSNP	49142359	0.62	0.38	0.77	0.23	-	-	-	-	-	-
rs470119	dbSNP	49099320	-	-	0.72	0.28	-	-	-	-	-	-
rs470120	dbSNP	49087041	-	-	-	-	-	-	-	-	0.93	0.07
rs71009	Orchid	48736082	1	0.00	1	0.00	0.67	0.33	-	-	-	-
rs71009	dbSNP	48736082	-	-	-	-	0.67	0.33	-	-	-	-
rs738334	dbSNP	48870263	0.83	0.17	-	-	-	-	-	-	-	-
rs743038	dbSNP	49016481	-	-	0.69	0.31	0.62	0.38	-	-	-	-

Total Unique SNPs with Allele Frequencies Found: 13 | get zipped text file

new search

Fig. 1. The result set of a sample request from the Frequency Finder. A batch upload of 53 SNPs was processed to produce these results.

0.10 or greater. By contrast, in the Caucasian population, only 52% of SNPs returned minor allele frequencies >0.10.

CONCLUSION AND DISCUSSION

Frequency Finder provides a single, user-friendly interface for retrieving allele frequencies for large batches of SNPs from multiple data sources. Additional features retrieve mapping data and real-time data from the ABI database. Using Frequency Finder, we have shown that the majority of SNP frequencies publicly available can be found in only one of the three main databases (dbSNP, TSC and ABI), with little overlap between them. We have also shown that Frequency Finder effectively locates at least 96% of frequency data available in these databases at the time the search is performed. Although only 6% of all SNPs currently have public frequency data, this fraction is expected to rise dramatically in the coming years with the advent of large, publicly funded studies of human genetic variation.

Frequency Finder has some important limitations. It currently mines data from only three established resources. However, the tool can communicate with any resource that provides a web-based interface to access allele frequencies. This is because we have built a flexible data-mining algorithm to integrate additional data sources though the HTTP protocol. Since Frequency Finder searches the HTML-content of the targeted databases, changes in the data layout at a database can pose a challenge. To address this, we have implemented a technique to validate real-time data requests. Most HTML pages for which our application parses data are contained in a table structure. Exploiting this structure, we utilize the column headings of the data table to perform intermediary validation on data position, additions and deletions. In the case of positional changes to the expected data, Frequency Finder

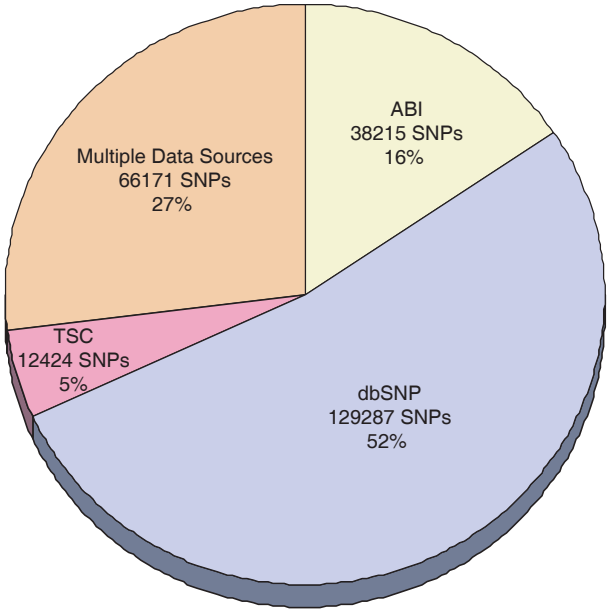


Fig. 2. Results of a whole genome query of dbSNP Build 114. Frequency Finder successfully retrieved allele frequencies for 246 097 SNPs.

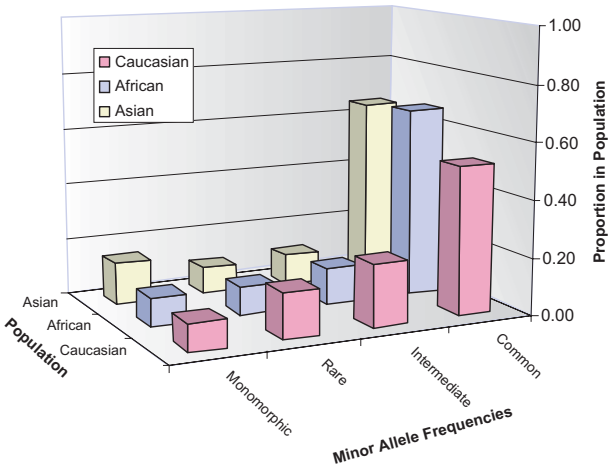


Fig. 3. Allele frequency spectra of 246 097 SNPs retrieved from public databases. The spectra differ significantly between populations ($\chi^2 = 12\,191$, df $= 6$, $P < 0.0001$). Prior to analysis, minor allele frequencies were divided into bins called 'monomorphic' (0.00), 'rare' (0.01–0.05), 'intermediate' (0.05–0.10) and 'common' (0.10–0.50).

attempts to alter its data structure to encompass the changes in the target database. If the change cannot be resolved or if the expected data column is not present, the application notifies the user and system administrator while proceeding with the data mining of other available resources. This ensures maximal data return, while eliminating the chance of false data retrieval.

Our application currently searches for frequency data that is available in the general population or in any of five groups, including Asians, Africans, Caucasians, CEPH and Unidentified. This provides a basis for determining which SNPs would be sufficiently polymorphic to proceed with genotypic studies in a particular population. For more extensive population genetic studies, databases such as ALFRED or HGVBase could be used. We have not yet included ALFRED and HGVBase into our application because these data sources query SNPs by methods other than a dbSNP cluster accession number.

In addition to our ability to filter the results by limiting the report to user-specified frequency ranges, we report marker positions in the result set through a cross-reference to NCBI's sequence map. This feature guarantees that the reported marker positions reflect the current build of the human sequence assembly. This additional feature should streamline the process of building a map of potential SNPs for association experiments.

Further development will improve the performance of Frequency Finder by storing data from real-time resources, such as allele frequencies retrieved from ABI, into a local database. This should alleviate performance discrepancies inherent in dynamic queries of real-time resources, where response time will be affected by both network latency and the remote system response. Of course, local data storage is subject to hardware limitations at the storage site, and introduces the possibility that user-requested data may not contain the latest available frequencies posted on the remote sites. This latter problem is minimized by nightly updates taken from the external data sources. We are also in the process of developing a web service integration tool, allowing external web-based software to retrieve the results from the Frequency Finder for real-time requests of the frequency data. Our application would be triggered through a HTTP request and the results would be sent in the formats of XML, HTML or tab-delimited text files. We would use existing technologies such as XML Remote Procedure Calls (XML-RPC) or Simple Object Access Protocol (SOAP) to implement this service.

In this paper, we demonstrated the use of Frequency Finder for large-scale SNP allele frequency searches by searching over 4 million SNPs and retrieving close to 250 000 allele frequencies obtained in three major population groups. Comparison between the groups indicated that a smaller proportion of SNPs return common allele frequencies in Caucasian populations compared to Asian or African American populations. This is consistent with previous suggestions in the literature that Caucasians are less genetically diverse than African Americans (Rosenberg *et al*., 2002; Reich *et al*., 2001).

The explosion of human genetic information from the Human Genome Project places increasing demands on scientists who wish to exploit fully the available information.

Frequency Finder provides a rapid, accurate and automated approach to mining SNP frequency data. We believe this program will be a valuable addition to the informatics toolbox.

ACKNOWLEDGEMENTS

We would like to thank Nirmala Akula, Sridhar Prathikanti and Yu-Sheng Chen for additional ideas and feedback. We would also like to thank Applied Biosystems (ABI) for granting us permission to use their public allele frequency data, obtained from the Assays-on-Demand™ SNP Genotyping Products Website. This work was supported by the NIH R01 MH61613 and NIH R01 MH065560, and Young Investigator Grant to C.L. from NARSAD (National Alliance for Research in Schizophrenia and Affective Disorders).

NOTE ADDED IN PROOF

We have recently updated our application to retrieve minor allele frequencies from the HapMap Data Coordination Center (http://www.hapmap.org).

REFERENCES

Brookes,A.J. (1999) The essence of SNPs. *Gene*, **234**, 177–186.

Cheung,K.H., Osier,M.V., Kidd,J.R., Pakstis,A.J., Miller,P.L. and Kidd,K.K. (2000) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. *Nucleic Acids Res.*, **28**, 361–363.

Fredman,D., Siegfried,M., Yuan,Y.P., Bork,P., Lehvaslaiho,H. and Brookes,A.J. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.*, **30**, 387–391.

Marsh,S., Kwok,P. and McLeod,H.L. (2002) SNP databases and pharmacogenetics: great start, but a long way to go. *Hum. Mutat.*, **20**, 174–179.

Marth,G., Yeh,R., Minton,M., Donaldson,R., Li,Q., Duan,S., Davenport,R., Miller,R.D. and Kwok,P.Y. (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat. Genet.*, **27**, 371–372.

Reich,D.E., Cargill,M., Bolk,S., Ireland,J., Sabeti,P.C., Richter,D.J., Lavery,T., Kouyoumjian,R., Farhadian,S.F., Ward,R. and Lander,E.S. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.

Riva,A. and Kohane,I.S. (2002) SNPper: retrieval and analysis of human SNPs. *Bioinformatics*, **18**, 1681–1685.

Rosenberg,N.A., Pritchard,J.K., Weber,J.L., Cann,H.M., Kidd,K.K., Zhivotovsky,L.A. and Feldman,M.W. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Thorisson,G.A. and Stein,L.D. (2003) The SNP Consortium website: past, present and future. *Nucleic Acids Res.*, **31**, 124–127.

Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.