Technical Report

Vol. 00 no. 0 2005, pages 1–5
doi:

Genetic and population analysis

Rapid analysis of large SNP datasets

Matthias Wjst[1,*]

[1]GSF National Research Center of Environment and Health, München-Neuherberg, Germany

Received on Februar 27, 2006; revised on xxx, 2006; accepted on xxx, 2006

Advance Access publication . . .

ABSTRACT

Motivation: Current genotyping technologies allow testing of more than 500,000 SNPs in thousands of individuals. Data storage and analysis of such large datasets is far from being trivial and generates multiple bottlenecks.

Here I show that by the use of a small C library in conjunction with fast R statistical routines also a single trait analysis of 500,000 SNPs in 270 individuals can be performed in less than 1 hour (online supplement box 1 and 2). Additional speed gain can be achieved by using parallel computing as most tasks involve independent, serial data processing. SNP allele frequencies in the current Affymetrix 500K array turn out to be rather equally distributed with a peak at low allele frequencies. Approximately 1% of all SNPs are in multicopy regions while 2-6% show low call rates; taken together they explain only a quarter of all SNPs that deviate from Hardy-Weinberg equilibrium (online supplement tables 1-4, figures 1-3). Less than half of all SNPs are situated on high LD blocks.

In summary, also large SNP datasets can be analyzed on a desktop computer. The overall data quality will be high if a few caveats are taken into account.

1 INTRODUCTION

New genotyping technologies like the Affymetrix GeneChip Mapping 500K Array Set or llumina's HumanHap300 BeadChip with 317,000 SNPs allow for the first time dense whole-genome association studies. The Affymetrix assay uses two arrays, each capable of genotyping on average 250,000 SNPs (approximately 262,000 for Nsp1 arrays and 238,000 for Sty1 arrays). Nsp1 and Sty1 denote the initial restriction enzyme cutting of the native DNA, before being ligated, amplified, fragmented, end labeled, and hybridized to a chip that is scanned for fluorescence signals. The analysis of this new generation of SNP genotyping chips creates a challenge of storing and managing large datasets.

Although there is considerable experience with large-scale gene expression arrays for more than one decade, the number of expression experiments is usually limited. After normalization expressed RNAs are often tested only for clustering and even then only top-ranked expressed RNAs are further evaluated. Most of these tasks can still be done with conventional desktop computers. The problem is somewhat different with SNP genotyping chips where each of the 500,000 SNPs need to be tested in up to 5,000 individuals. Association studies do not only require testing of several traits and subgroup analysis of billions of genotypes, there are also more complex tasks like binning into high LD regions and haplotype construction.

The initial attempt to load the genotypes of such a SNP chip study in an industry standard database was not successful. Although having used this database for some time, the import of 540 chip sets ended several times with error messages. Accessing the database over the intranet by the ODBC interface turned out to be unacceptable slow, given the large amount of data transferred with each query. Running the analysis directly on the database host computer was also not an option as the additionally CPU load significantly downgraded database performance, ending up with analysis times of up to 1 week for a single trait.

As there is currently also no information available about genomic sequence coverage, allele frequency, Hardy-Weinberg distribution, and LD block binning of these genotyping sets, the goal for this study was twofold: (1) to optimize data storage for a 135 million SNP project to work on a conventional desktop computer and (2) to provide a first view on allele frequencies as well as other background data. All analysis should be platform independent and available with software in the public domain.

2 METHODS

Clinical sample. Individuals contributing DNA samples to the hapmap project have come from a total of 270 people and 4 ethnical diverse groups. The Yoruba people of Ibadan, Nigeria, provided 30 sets of samples from two parents and an adult child (YRI). Another 30 U.S. trios collected in 1980 from U.S. residents with Northern and Western European ancestry come from the Centre d'Etude du Polymorphisme Humain registry (CEU). In Japan, 45 unrelated individuals from the Tokyo area provided samples (JPT) as well 45 unrelated individuals from Beijing/China (CHB) [Altshuler].

Genotype data. SNPs for the 500K Array Set have been selected for inclusion for their suitability to the laboratory protocol and minimum allele frequencies. Following sample interrogation genotypes were called by the proprietary Affymetrix GeneChip Genotyping Analysis Software (GTYPE, see GeneChip Operating Software User's Guide for instructions on .dat and .cel file generation). GTYPE 4.0 provides a convenient interface to group file sets and export into various formats where here the tab delimited format was selected. For this project, data were downloaded from the public Affymetrix website (as a single zip file (500K_HapMap270.zip) that contains in two large ASCII text files SNP genotypes in 500,568 rows and 270 individuals arranged in columns.

Data storage and analysis. For data storage the SQLite 3.3.1 library was used. All genotype data were imported into single tables by the n-col and the 3-col method. A small routine in R software was then used to count alleles for benchmarking. All benchmarks were performed on a standard Windows XP system with a Intel® Pentium® M processor with 2.0 GHz, using 1 GB main memory, and a 20 GB harddisk partition freshly format-

*To whom correspondence should be addressed.

ted with NTFS using 64 MB clusters with deactivated virus scanner. For parallel computing R version 2.2.1 [R Core Team]) were translated from source into 64-bit code to run on a linux cluster IBM e1350 under Suse Linux Enterprise Server 9. Hardy-Weinberg equilibrium was determined also by ethnical group using an exact test [Wigginton]. All SNP sequence (plus/minus 16 bases were aligned to the finished human genome assembly (hg17, May 2004), all known human repeat sequences and to the mitochondrial genome (Cambridge reference sequence 02/07/2005) using blat software. LD blocks and haplotype information was generated using haploview [Barrett] and results transformed into BED files that can be merged into genome browser data [Kent] (http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&position=chr12&hgt.customText=http://cooke.gsf.de/Affx500Kbl.gz).

3 RESULTS

I implemented first a data preprocessing step by a short Perl script (online supplement box 1). This script stripped headers, recoded missing values and prepared another script that will import all data into SQLite format (www.sqlite.org). SQLite is a small C library, that forms a self-contained, embeddable, zero-configuration, SQL-92 compliant database engine supporting up to 2 terabytes in size. This procedure was by factor 5 to 100 faster compared to bulk data upload to an industry standard client-server database (online supplement table 1). Reason for the excellent performance of SQLITE may be seen by the simple API, the small and efficient code without any superfluous database features.

Initial row/column structure was either unchanged (n-col method) or reverted to each genotype in a single line with SNP identifier, person identifier and genotype (3-col method). The 3-col method was tested as most databases show maximum column limits. Depending on release version and operating system these limits range from 250 (PostgreSQL), 255 (MS SQL Server), 1024 (mySQL), 1000 (Oracle) or 8000 (IBM DB2). The n-col and 3-col approach differed markedly where n-col was by factor 40 faster than 3-col.

None of both methods showed memory problems while the 3-col approach is probably suited to billions of genotypes. The n-col approach has single peak memory usage due to a single array transpose step that could lead to memory problems by testing more than 1,000 probands. Due to the modular approach, larger proband sizes may be easily compensated by using smaller batch sizes. 3-col had even more disadvantages when building the final analysis dataset by its large overhead (the "XML" effect, leading to 10-fold file size) and much higher indexing time as this has to be done on 135 million and not only 500,000 datasets.

All further operations were then optimized in the R programming environment (www.r-project.org) during tedious steps to avoid unnecessary data read/write disk access, sorting, indexing or merging. Loop numbers and array sizes of processed genotypes were varied in a wide range. A single outer loop (running 50 times including 10,000 SNPs) with one nested loop (running 10 times including 1,000 SNPs) produced the best benchmarks (online supplement box 2 and table 1). Too small arrays needed to much I/O operations slowing down the overall performance, while too large arrays need too much time to address certain array positions. During project start I even used a third inner loop (running 1,000 times on each single SNPs) which was then replaced by a short function that is being applied to all columns. This had even more advantages when computation was being parallelized between multiple processors. Each loop level could be used to divide computing load to different processors. This choice depends heavily on the overall architecture of a network, e.g. number of nodes, memory and message passing time between nodes. Parallelizing of outer loops would need a major rewrite of the program (and making it specific for our local cluster architecture) while parallelizing the most inner loop is less demanding. The R package "snow" already offers a load balancing version of the "apply" function where it should be possible in theory to downsize the current processing time by dividing processing time by the number of nodes. Although extra overhead is created by additional master - slave communication this may be compensated by the higher CPU speed of the single nodes. If this prediction is correct, is the subject of current test runs. So far, processing time was only been optimized until the most inner loop. All functions that are being applied at this stage (e.g. allele counting, HWE estimation, LD binning, haplotype construction) need to be extra benchmarked for large datasets.

In addition I established some quality checks for the Affymetrix 500K array set. HWE was calculated by an implementation optimized for large samples while any deviation from HWE in an outbred population without major stratification might indicate genotyping errors. Unexpectedly 5,817 SNPs showed multiple alignments in the human genome (list available upon request). Although some of these may be false positives, most of them are leading to a distortion of HWE (online tables 2 and 3). More than the double amount of SNPs are not in HWE compared to "single hit SNPs", a reason why these SNPs should be masked from further analysis. By testing all SNP sequences in a collection of repeats in the human genome, I also found multiple sequence matches there (and even two on the mitochondrial genome). Missing genotypes might also indicate unequal amplification and indeed HWE decreased with increasing number of missing genotypes (online supplement table 3). "Multiple hit SNPs" together with those showing more than 10% missing values, however, still do not explain more than 20,8% (CEU) or 25,5% (YRI) of those SNPs not in HWE making some more unidentified errors possible mainly where HWE is severly violated. On a genome-wide level SNPs with deviation from HWE seem to be randomly distributed making the current selection a valuable screening set. In a last step, I have estimated LD block sizes which expand on average between 4,2 and 5,7 SNPs. As shown earlier [Altshuler] probands of African origin as less as 40% of all SNPs inside high LD regions.

In summary, the use of a simple C database library linked to fast R routines makes the analysis of large SNP datasets possible even on inexpensive desktop workstation. 1% of the current Affymetrix 500K SNP set may be masked for unreliable results in multi-copy regions as well as all SNPs with less than 90% call rate.

REFERENCES

Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J. and Donnelly, P. (2005) A haplotype map of the human genome, Nature, 437, 1299-1320.

Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps, Bioinformatics, 21, 263-265.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC, Genome Res, 12, 996-1006.

R Core Development Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Wigginton, J.E., Cutler, D.J. and Abecasis, G.R. (2005) A note on exact tests of Hardy-Weinberg equilibrium, Am J Hum Genet, 76, 887-893.