IGG3: a tool to rapidly integrate large genotype datasets for whole-genome imputation and individual-level meta-analysis

Miao-Xin Li1,2, Lin Jiang2, Patrick Yu-Ping Kao1, Pak-C. Sham3,4, You-Qiang Song1,4,∗

1Department of Biochemistry, The University of Hong Kong, Pokfulam, Hong Kong, 2The Center for Experiment, Hunan University of Commerce, Changsha, Hunan 410205, China, 3Department of Psychiatry and 4The Centre for Reproduction, Development and Growth, The University of Hong Kong, Pokfulam, Hong Kong

Received on February 9, 2009; revised on March 25, 2009; accepted on March 30, 2009

Advance Access publication April 3, 2009

ABSTRACT
Summary: There is an urgent and increasing demand for integrating large genotype datasets across genome-wide association studies and HapMap project for whole-genome imputation and individual-level meta-analysis. A new algorithm was developed to efficiently merge raw genotypes across large datasets and implemented in the latest version of IGG, IGG3. In addition, IGG3 can integrate the latest phased and unphased HapMap genotypes and can flexibly generate complete sets of input files for six popular genotype imputation tools. We demonstrated the efficiency of IGG3 by simulation tests, which could rapidly merge genotypes in tens of thousands of large genotype chips (e.g. Affymetrix Genome-Wide Human SNP Array 6.0 and Illumina Human/r-m-duo) and in HapMap III project on an ordinary desktop computer.

Contacts: songy@hkucc.hku.hk; limx54@yahoo.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
Genotype integration across various platforms is becoming popular and in demand in current genome-wide association studies (GWAS) as whole-genome genotype imputation (Marchini et al., 2007; Nicolae, 2006) gets feasible and almost routine. Over 3 million genotyped single nucleotide polymorphisms (SNPs) can be freely downloaded as imputation reference in HapMap II (Prazer et al., 2007). A larger reference panel is also ready in HapMap III comprised of an encouraging size of 1301 individuals. Besides, by combining raw genotypes of multiple cohorts from different genotyping platforms, an individual-level meta-analysis across GWAS is usually ideal because many tests, which are impossible for study-level meta-analysis, are allowed to be conducted including epistasis, gene-based tests and tests for other phenotypes not originally considered (de Bakker et al., 2008).

However, consistent genotype integration is usually laborious in practice. Genotypes of the same SNP in different datasets may have different allele names according to either forward/+ or reverse−strand orientation. Imputation tools may smartly flip the alleles when encountering inconsistent allele names or observing large allele frequencies differences (de Bakker et al., 2008). But this strategy does not work for A/T or C/G SNPs since they are complementary bases (an A/T SNP cannot be distinguished from a T/A SNP). Convenient tools which facilitate genotype integration are in need (de Bakker et al., 2008). Although IGG (Li et al., 2007) worked well in consistently integrating genotypes in hundreds of small or medium chips for conventional genome-wide linkage and association scans, it is inefficient for the ongoing GWAS, in which thousands of large chips and the HapMap III dataset may be involved. Besides, it cannot integrate phased HapMap genotypes and has no specific modules to facilitate genotype imputation.

Therefore, we developed a new algorithm to efficiently merge large amount of genotype data and devised functions to facilitate whole-genome imputation. The algorithm and functions have been implemented in the latest version of IGG, IGG3. It is an open source and platform-independent Java package.

2 ALGORITHMS AND FEATURES

2.1 Algorithm to speed up integration of large datasets
This new algorithm includes three critical characteristics. First, genotypes of SNP are encoded into binary codes, which will save much storage space in both Random Access Memory (RAM) and hard disk. A 4-bit variable ([X1, X2], [X3, X4]) is designed to denote a genotype of a biallelic SNP. Here, X1 and X2 denote the existence status of two alleles X3 and X4 of a genotype, respectively. If X1 is 0, the allele X3 is missing and the value of X3 is ignored. When X1 is 1, X3 can be either 0 or 1 to denote two different alleles. Similarly, X2 and X4 indicate together the other allele of the genotype. Conventionally, a 16-bit space is required to denote a genotype by two characters in the computer. It is also the manner in which genotypes are presented in a text file. But our variable now just needs 4 bits ([X1−X4]) to denote a genotype, which can save 75% space compared with these characters. In IGG3’s program, the 4-bit genotype variable is implemented by 2 bit- vectors. One bit-vector is for the existence status and the other bit-vector for the genotypes. Figure 1 shows an example to illustrate the programming implementation.

Second, a hash function is used to quickly access given SNPs in a SNP annotation list. The SNP/ probe set IDs in the chips are employed to generate unique hash codes based on the hash function. Given the probe set ID of a SNP, the SNP location in the list can be calculated by the hash function. In this way the SNPs information can be directly retrieved without iterating the whole list. Third, to further relieve the burden in RAM, genotype chip files loaded at a time are automatically divided into a number of buffers and are dealt with at multiple times. The whole integration ends up with a final fusion of integrated binary genotypes between buffers (detailed in Supplementary Material 1). This divide-and-conquer strategy is particularly effective when
2.2 Integration of phased and unphased HapMap genotypes (Phase-I+II-III)

IGG3 uses annotation-guided method to integrate HapMap genotypes into local projects. These annotation data were extracted and compiled from a number of files downloaded from the NCBI ftp site, ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/, which includes dbSNPs reference ID, physical positions on the human genome assembly, allele names, flanking sequences and allele frequencies of all HapMap SNPs. The allele names and flanking sequences were converted for forward strands according to the human genome assembly. The strand information of HapMap genotypes is known. IGG3 asks users to load HapMap genotypes on the forward strand for integration. Thus, the flanking sequences in annotation data are consistent with the loaded HapMap genotypes. Based on these flanking sequences, the HapMap genotypes can be consistently merged into integrated chip genotypes (Li et al., 2007). Two other features in the integration of HapMap genotypes are described in the Supplementary Material 2. Similar method has been used to integrate general genotype datasets with strand information (Detailed in Supplementary Material 3).

2.3 Facilitate genotype imputation

Aside from the integration of HapMap genotypes (including the phased ones), in which the troublesome discordance of allele-strand problem has been completely solved, IGG3 has two other features with regard to facilitation of genotype imputation. First, it can now flexibly generate complete sets of input files for six popular genotype imputation tools, Plink (http://pngu.mgh.harvard.edu/~purcell/plink), Merlin (http://www.sph.umich.edu/csg/abecasis/Merlin), IMPUTE (http://www.stats.ox.ac.uk/~marchini/software/gwas/impute.html), MACH (http://www.sph.umich.edu/csg/abecasis/MaCH), BEAGLE (http://www.stat.auckland.ac.nz/~browning/beagle/beagle.html) and fastPhase (http://depts.washington.edu/ventures/UW_Technology/Express_Licenses/fastPHASE.php). The connection to multiple tools can help on comparisons of imputation results by different methods, which may be an important way to produce convincing conclusions. Second, IGG3 gives an example command to suggest the usage of the generated files for imputation by the tool being chosen. Here is a suggested command for IMPUTE: ‘An example command to run IMPUTE for genotype imputation on chromosome #; # geno -t test.sh haplo -l test.txt -r legend’.

3 TESTING AND DISCUSSION

Several genotype datasets were generated to test the performance of IGG3. The datasets were made up of a number of virtual chip genotype files simulated for four large whole-genome chips: Affymetrix Genome-Wide Human SNP Array 6.0 and 5.0, Illumina Human1m-duo and Human650y-quad BeadChips. Under the continuous uniform distribution U(0, 1), genotypes of SNPs were randomly assigned according to their available frequencies in the HapMap CHB+JPT population. Allele frequency 0.5 was set for SNPs without HapMap frequencies. The simulation function can also be launched by users on IGG3’s menu, Tools \rightarrow Generate Sample. As a practical reference, the test was conducted on an ordinary desktop computer, Intel Core™2 Duo CPU 3.00GHz, RAM 2.0GB and 32-bit Windows Vista™ Home Edition.

We compared the running time and required RAM between IGG3 and IGG (Li et al., 2007) was originally designed for relatively modest chips like the Affymetrix Mapping 250K Nsp Array. It simply used characters to present genotypes and binary search method to locate an SNP in the annotation list, and had no adoption of the divide-and-conquer design. The comparison results are shown in Supplementary Table 1. IGG was found to work poorly in handling these large chips. Given the maximum RAM 1488 MB, IGG could, at most, only integrate 120 subjects’ genotype chips (30 for each chip type) at a time. But IGG3 merely required 706 MB RAM to do this. More importantly, IGG3 ran about 10 times faster than IGG to integrate the whole-genome genotypes. Furthermore, a huge dataset was used to test the performance of IGG3. 20 000 generated subjects’ genotype chips (5000 for each chip type) and the HapMap (II-III) genotypes. The total size of the dataset was ∼316.2 GB in the hard disk. This huge dataset could not be processed by IGG at all due to limited RAM on this ordinary computer. However, IGG3 still worked well in this situation. These testing results indicate a proof of concept for the efficiency of the new algorithm to integrate large amount of genotypes cross different genotyping platforms.

ACKNOWLEDGEMENTS

We gratefully acknowledge the BIOSUPPORT project (http://www.bioinfo.hku.hk), the Computer Centre. Funding: Research Grant Council of Hong Kong (HKU7688/05M & HKU7752/08M, YQS); Research Fund for the Control of Infectious Diseases (No. 08070652, to Y.-Q.S.).

Conflict of Interest: none declared.

REFERENCES


