19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

An Efficient Process for Enhancing Genotype Imputation in Genome-wide Association Studies Using High Performance Computing

Kasikrit Damkliang*, Pichaya Tandayya*, Unitsa Sangket†, Surakameth Mahasirimongkol‡ and Ekawat Pasomsab§
*Department of Computer Engeering, Faculty of Engineering,
†Center for Genomics and Bioinformatics Research, Faculty of Science,
Prince of Songkla University, Hat Yai, Songkhla, Thailand 90112
‡Medical Genetic Center, Medical Life Sciences Institute, Department of Medical Sciences,
Ministry of Public Health, Bangkok, Thailand 11000
§Department of Pathology, Faculty of Medicine, Ramathibodi Hospital,
Mahidol University, Bangkok, Thailand 10400
Email: kasikrit.d@psu.ac.th, pichaya@coe.psu.ac.th, unitsa.s@psu.ac.th, surakameth.m@dmsc.mail.go.th, ekawat.pas@mahidol.ac.th

***Abstract*—Genotype imputation based analysis usually consumes computational and data intensive. This paper presents a practical and efficient process for enhancing the genotype imputation based analysis on Single Nucleotide Polymorphism (SNP) using High Performance Computing (HPC). Our process is split into data quality control, haplotype estimation, and imputation. We validate and measure the process on a standard workstation and a server for pilot dataset of chromosome 22 from Genetic Analysis Workshop 16 (GAW16) provided by the North American Rheumatoid Arthritis Consortium (NARAC). The NARAC dataset consists of 2,062 individuals and 545,080 SNP variants. We use 1000 Genomes database as reference panels. Our process correctly and rapidly produces results more than ordinary steps of the genotype imputation based analysis.**

I. Introduction

Many research topics in Genome-Wide Association studies (GWAS) of Bioinformatics have not been revealed and resolved yet. The GWAS research usually tends to intensively consume both computational and data resources, especially Single Nucleotide Polymorphism (SNP) analysis. There are at least 1% or 30 million bases of SNPs and it shares $\geq$10 million common genetic variants with Minor Allele Frequencies (MAF) $\geq$5% in human genome. The GWAS objective is to discover which SNPs influence gene expression. However, geneticists has found that variations of DNA base sequences occur on some positions of the SNPs. If a lot of missing variations are un-genotyped, then these missing variations may induce ambiguous SNP analysis. As a result, the gene expression analysis also is ambiguous. The geneticists use genotype imputation for evaluating evidences of genetic markers or SNPs association of datasets that they are not directly genotypes. Therefore, the genotype imputation is useful for estimating un-genotyped gene positions using existing GWAS data. The most recent technology only provides about one million variants. Different genotyping platforms may provide different information [1] [2] [3].

In general, the analysis of genotype imputation usually requires High Performance Computing (HPC) as it is computational and data intensive. From the point of Computer Science and Engineering, we have found that bioinformaticians usually lack of customized sufficient tools in the genotype imputation including hardware, software, and process. As a result, they always face inconsistency in running analysis environment. They may also not comprehend what is behind available software, especially the issues concerning both of computer and software architecture.

This paper presents a practical and efficient process for enhancing genotype imputation based analysis on SNPs using HPC. We employ a dataset of Genetic Analysis Workshop 16 (GAW16) provided by the North American Rheumatoid Arthritis Consortium (NARAC) [4].

In the next Section, we describe attributes of the NARAC dataset and preliminary data format conversion. In Section III, we review related tools for the genotype imputation based analyzed such as genotype imputation tools and data quality control tools. Then, we compare these tools and select interested tools for our proposed process. In Section IV, we preliminary measure the performance of the related legacy tools with samples dataset from the NARAC. In Section V, we propose the efficient process for genotype imputation based analysis. We split the process into three steps: data quality control, haplotypes estimation, and imputation. Then, we validate and measure the performance of our process on a standard workstation and a server using chromosome 22. Discussion and Conclusions are presented in Section VI and VII respectively.

II. The NARAC Dataset

The GAW16 is a public dataset consisting of 545,080 SNP-genotype fields from the Illumina 550K chip. There are 2,062 individuals all around the USA, which come from 860 cases and 1,194 controls. The NARAC dataset is contained in 2 files with the comma-delimited format. The file narac.csv contains a header line and 2,062 records of individuals and the file narac.map contains a header line and 545,080 records of the

978-1-4673-7825-3/15/$31.00

TABLE I. EXAMPLE RECORDS OF THE RAW NARAC DATASET

ID	Sex	rs5747620	rs5747968	rs2236639
D0024949	F	A_G	A_A	G_G
D0024302	F	G_G	A_A	G_G
D0023151	F	?_?	A_A	G_G
D0022042	M	?_?	C_C	G_G
D0021275	M	A_A	A_A	G_G

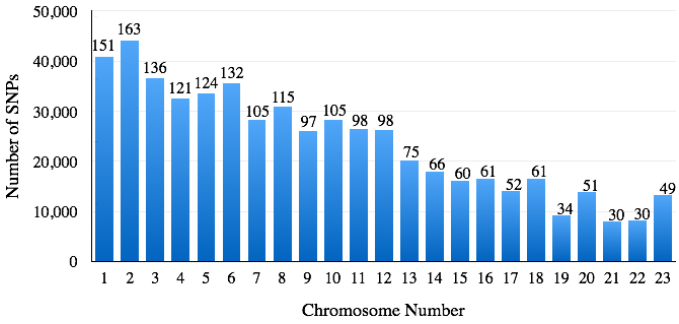

Fig. 1. The number of SNPs and execution time for format conversion

TABLE II. EXAMPLE RECORDS OF THE PEDIGREE FILE

FamilyID	IndividualID	FatherID	MotherID	Sex	rs5747620	rs5747968	rs2236639
0001	D0024949	0	0	2	1/4	1/1	4/4
0002	D0024302	0	0	2	4/4	1/1	4/4
0003	D0023151	0	0	2	0/0	1/1	4/4
0004	D0022042	0	0	1	0/0	3/3	4/4
0005	D0021275	0	0	1	1/1	1/1	4/4

SNPs-genotype fields (one line per an variant) [4]. Each record consists of fixed nine fields and a various number of SNP-genotype fields depending on each chromosome. For example, Chromosome 1 contains around 40,000 SNPs-genotype fields, whereas Chromosomes 21 and 22 contain around 8,000 SNPs-genotype fields each. The total of SNPs-genotype fields of the dataset is 545,080 records. For example, In Table I, a record is an individual data representative of Chromosome 22. The SNPs-genotype fields are in the format "X_X" where X is a base (A, T, C, and G). Missing SNP genotypes are coded as "?_?" and, these SNPs do not have A_T and C_G alleles.

In our preliminary study, we have composed simple shell scripts for converting the raw NARAC data into the Merlin format (data and pedigree files), which will be estimate inputs to MACH [5] for haplotypes estimation. We can identify each number of SNPs-genotype fields in each chromosome (1 - 23) using a simple shell script and results are shown in Figure 1. Figure 1 also shows execution times used in hours converting from the raw NARAC into the Merlin format. For example, converting Chromosome 22 is about 30 hours, running on a standard Linux workstation. We compare the result to all chromosomes and can presume that it may about or more three months to complete the conversion. This is a large obstacle in pre-processing data stage. However, there is a related research concerning splitting data from the NARAC by numbers of SNPs [5]. We approach to utilize a distributed database system applying data partitioning such as by chromosome, by SNP and/or by individual. Table II shows example records of the pedigree file of Chromosome 22. We transform family IDs using running numbers, father IDs and mother IDs to zero, for marking unrelated individuals.

TABLE III. GENOTYPE IMPUTATION TOOLS

Tool	Developer
MACH/Minimac/Minimac3	University of Michigan
Beagle	University of Auckland
IMPUTE2/SHAPEIT	University of Oxford
PLINK	Massachusetts General Hospital

III. RELATED TOOLS

In this section, we describe genotype imputation based analysis related tools, and data quality control tools. Then, we compare the tools using their attributes.

A. Genotype Imputation Tools

There are many genotype imputation tools which run as command-line programs on Unix and Linux-based as shown in Table III [2] [6]. The main algorithm of the tools is Markov-Chain Monte Carlo (MCMC), numerical approximation algorithms. The genotype imputation tools generally utilize it for phasing unobserved or hidden states of data by iterating steps for a hidden markov model [7] [8] [9]. Moreover, outputs of the programs can be inputs into other tools such as GenABEL [5], to determine further gene expression and quantitative traits analysis.

1) MACH & Minimac/Minimac3: MACH 1.0, Markov Chain based haplotype, is the most popular genotype imputation tool developed by University of Michigan [2] [10] [11]. The current version is a pre-release. It is a sequential program. MACH supports both SNPs genotype in missing alleles and SNPs inference in samples of unrelated individuals. Minimac is a high throughput/multi-threading version of MACH that implemented using OpenMP [12] also developed by the University of Michigan [6]. It only supports inferring untyped markers (Method 2 of MACH). Therefore, input of Minimac must already has been phased before using the program. In version 3, it accepts only the Variant Call Format (VCF) input format [13]. There are many programs for converting data formats to the VCF such as PLINK [14], MACH2VCF and, SHAPEIT [15]. MACH and Minimac recommend two steps of genotype imputation. The first step is to phase the samples into a series of estimated haplotypes. The second step is to carry out direct impute with these phased haplotypes. In case that a reference panel is renewed, it is not necessary to phase the haplotypes again, only the second step has to do re-imputation.

2) SHAPEIT: SHAPEIT stands for Segmented HAPlotype Estimation and Imputation Tool [14]. SHAPEIT estimates haplotypes or does phasing from genotypes or sequencing data. It is developed by University of Oxford. SHAPEIT recommends to do pre-phasing imputation together with IMPUTE2 which is to be described in the next section. SHAPEIT is free for academic use only.

3) IMPUTE2: IMPUTE2 is a haplotype-phasing and imputation tool. It works based on ideas of Howie et al. 2009 [8]. It supports many modes of genotype imputation. However, we are interested in only the imputation with one phased reference panel (pre-phasing) in this paper.

4) ParaHaplo: ParaHaplo 3.0 is a parallel version of genotype imputation software packages running on a supercomputer, developed by the RIKEN research laboratory, Japan [16]. The paralleled version of haplotype estimation is 20 times

TABLE IV. GENOTYPE IMPUTATION TOOLS COMPARISON

Tool	Objective	Programming Model	Input Data Format	Open Source
MACH	Phasing/Imputation	Sequential	Merlin	Yes
Minimac	Imputation	OpenMP	MACH	Yes
Minimac3	Imputation	OpenMP	VCF	Yes
ParaHaplo	Phasing/Imputation	MPI	Merlin	Yes
SHAPEIT	Phasing	OpenMP	Many	No
IMPUTE	Phasing/Imputation	Sequential	IMPUTE2	No
GenABEL	Data QC	Sequential	N/A	Yes
PLINK	Data Conversion/QC	Sequential	PLINK	Yes

faster than the non-parallel version. We have considered our datasets and concluded that the haplotype blocking method is another step to be added in the process of our approach. It is not necessary to break down the datasets into haplotype blocks due to datasets attributes. However, we have tested ParaHaplo 3.0 with phased haplotypes data of the NARAC-Chromosome 22 on 4-cores 3.0 GHz CPU, 8 GB RAM, Linux and a cluster of National e-Science Infrastructure Consortium, Thailand, it has 12 nodes Intel Xeon 2.66 GHz, total memory of 576 GB and PBS job scheduler. Both of tests failed due to memory segmentation faults.

B. Data Quality Control Tools

Bioinformaticians highly recommend doing data quality control (QC) [11]. The basic data QC is to filter un-frequent alleles from variants. The most popular data QC is GenABEL. Main objectives of GenABEL are to analyze quantitative traits and also support data QC. By the way, there is a parallel version of the R-GenABEL which is ParallABEL. ParallABEL splits the NARAC dataset into a number of subsets, which depends on available processors. The paper [5] reported that partitioning the chromosome dataset by the number of SNPs shows the highest parallel performance. Another QC program for genotype imputation based analysis is the PLINK software package [14] [17], a tool for handling the SNP data, developed by Massachusetts General Hospital. PLINK also is a tool for data format conversion.

C. Genotype Imputation Tools Comparing

We summarize related genotype imputation based analysis tools and identify their attributes as shown in Table IV. Most programs have their own formats for input and output data, and also provide data format conversion. For example, SHAPEIT supports the PLINK PED/MAP file format, PLINK BED/BIM/FAM file format, and Oxford GEN/SAMPLE file format. SHAPEIT also supports format conversion of its output to the VCF and Oxford formats. There are many format conversion programs. The most is PLINK. In Table V, Autumn, L. reported genotype imputation tools performance comparison by including all steps required such as Beagle, IMPUTE2 and, Minimac [18]. The performance showed that pre-phasing an original dataset before imputation can improve significantly computation time.

IV. PRELIMINARY PERFORMANCE MEASUREMENT

We have measured preliminary performance of some related tools for genotype imputation steps as shown in Figure 2 with a small sample dataset and the reference panel consists

TABLE V. GENOTYPE IMPUTATION TOOLS PERFORMANCE COMPARISON

Tool	Total Number of SNPs	Computation Time (Hrs.)
IMPUTE2	668,180	23
BEAGLE	484,023	213
IMPUTE2 with Pre-phasing	668,180	8
BEAGLE with Pre-phasing	484,023	34
Minimac	667,870	18

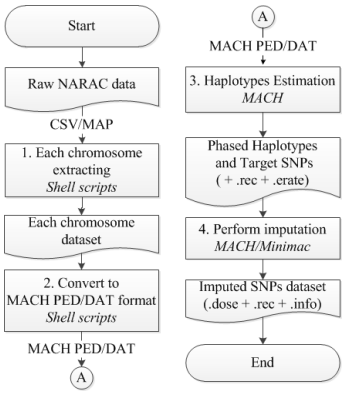

Fig. 2. The ordinary workflow of genotype imputation

TABLE VI. EXECUTION TIME OF IMPUTATION STEP APPLYING CHROMOSOME 22 ON A WORKSTATION

Tool	Execution Time (Min.)
MACH	Fatal Error
Minimac	13
Minimac-omp	Segmentation Fault

of Chromosome 22 (2,062 individuals, 8205 SNPs) of the NARAC dataset. The reference panel is the March 2010 release of Europe (CEU) phased data population from the 1000 Genomes Project (http://www.1000genomes.org). The reference panel contains 120 haplotypes. Our workstation operating on Linux Kernel 2.6.32, CentOS 6.6 has a 4-cores single CPU 3.0 GHz and 8 GB RAM.

Figure 3 shows the haplotypes estimation or phasing times on the Linux workstation by varying number of individuals. We configured the number of states to 400 and the number of iterations to 50 for MACH 1.0. The program took about one and a half hours to complete phasing 95 individuals. It took about 59 hours or two and a half days to complete phasing all individuals of Chromosome 22. After that, we performed the imputation step using MACH 1.0 with phased data from the previous step. It took about five minutes to complete imputation for 80 individuals. In case of 2,062 individuals, MACH returned a fatal error as shown in Table VI, whereas Minimac took 13 minutes, and the multi-threaded version of Minimac returned segmentation faults. We collected the execution times for only successful running steps as shown in Table VII. It took about 60 hours for the genotype imputation of the NARAC-Chromosome 22 and 108 hours or four and a half days for completing all steps. Furthermore, it probably needs at least three and a half months for the genotype imputation analysis of all 23 chromosomes.

V. RESULTS

SNP analysis is workflow-oriented. Workflow is a representative of automatic executable instructions that generate results.

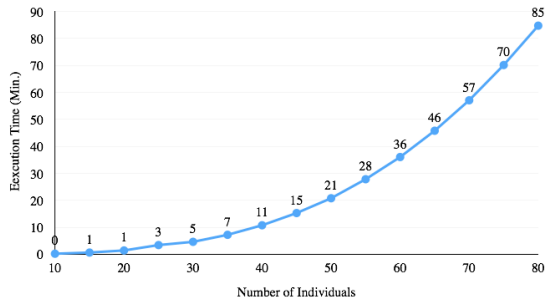

Fig. 3. Haplotypes estimation time using MACH

TABLE VII. EXECUTION TIME FOR RUNNING ALL STEPS OF CHROMOSOME 22 ON A WORKSTATION

Step	Tool	Execution Time (Hrs.)
Data Conversion	Simple shell scripts	48.88
Haplotypes Estimation	MACH	58.44
Perform Imputation	Minimac	0.21

Some processes' outputs can be some inputs of any process into many downstream processes in the workflow [19]. In this section, we present our proposed efficient process for genotype imputation and it's performance measurement on a workstation and a server.

A. Proposed Efficient Process

We propose a new efficient process for the genotype imputation based analysis in the flowchart using the NARAC dataset as a case study shown in Figure 4. The initial step is format conversion from the raw dataset of NARAC to compatible formats for haplotypes estimation. We simply use shell scripts for raw format conversion and select the PLINK PED/MAP file format for the output. The second step is data quality control using PLINK. It returns outputs in the PLINK BED/BIM/FAM file format. The preprocessed data quality control is piped to estimate haplotypes using SHAPEIT which returns outputs in the IMPUTE file formats such as HAPS and SAMPLE. The final step is imputation analysis using Minimac3, but it does not support the previously file formats. Minimac3 supports only the VCF format. Therefore, we have to convert the outputs of phased haplotypes into the VCF format which achieved by SHAPEIT itself after the haplotypes estimation step. In data QC steps, we configure parameter using basic usage recommended by PLINK within 0.05 for the MAF threshold. We validate the proposed process by testing with a large dataset on the 4-cores single 3.0 GHz CPU, 8 GB RAM, Linux workstation and a server described in the next section.

B. Performance Measurement on a standard workstation

Figure 5 shows haplotypes estimation comparison between MACH and SHAPEIT on the Linux workstation by varying the number of individuals. We configure the number of CPUs to four according to the number of CPU cores and 400 states of haplotypes are used. SHAPEIT dramatically saves more running times than MACH, from many hours to $<$10 minutes. Then, we perform imputation step of 100 samples with the Phase 1 Version 3 of Chromosome 22 from the 1000 Genomes reference panel. The reference panel consists of

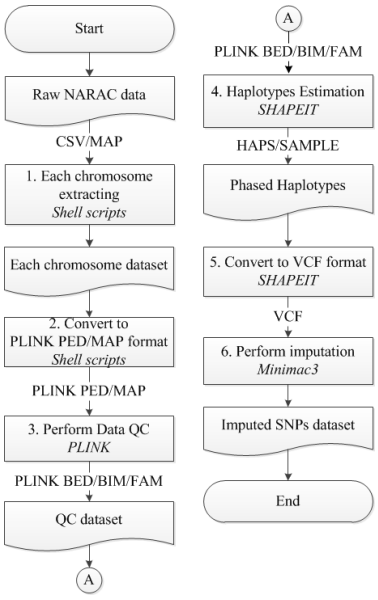

Fig. 4. The proposed workflow of genotype imputation

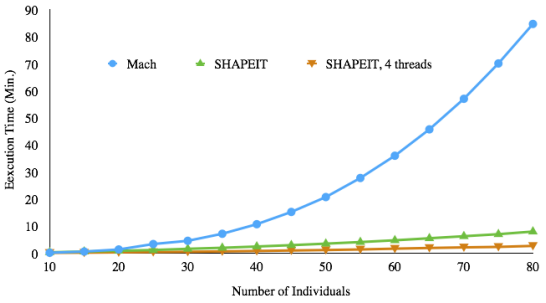

Fig. 5. Haplotypes estimation comparison between MACH and SHAPEIT on a workstation

365,644 markers and 2,184 haplotypes. The result were slightly different. The single Minimac3 took 8 hours and 9 minutes, whereas it took 8 hours and 40 minutes using the number of four CPU-cores parameter.

C. Performance Measurement on a server

The testing server has 32 Intel Xeon 2.3 GHz CPU, total memory of 132 GB, and PBS job scheduler. Table VIII shows execution time for all steps in applying the NARAC-Chromosome 22 genotype imputation on the server. We perform the phasing step for all individuals (2,062) for four times using 400 states and 35 MCMC iterations configuration using SHAPEIT and 32 CPUs parameter. It took almost one hour. In the next step, we converted outputs' format of SHAPEIT (haps, sample) into the VCF format of Minimac3's input within SHAPEIT itself. The execution time was 11 seconds. The last step was the imputation analysis step using Minimac3 with the same configuration. The reference panel is Chromosome 22 Phase 3 Version 5; there are 652,195 markers and 5,008 haplotypes. The NARAC-Chromosome 22 has 47 markers and 4,124 haplotypes. The imputation time was 39.25 hours and all steps was 72.25 hours.

Figure 6 shows execution times used only the genotype

TABLE VIII. THE ALL STEPS EXECUTION TIME FOR GENOTYPE IMPUTATION OF CHROMOSOME 22 ON THE 32-CPUS SERVER

Step	Execution Time(Hrs.)
1. Extract raw data (Shell scripts)	<1 min
2. Convert to PLINK format (Shell scripts)	32
3. Data QC *(PLINK)*	<1 min
4. Estimate haplotypes *(SHAPEIT, 32 CPUs)*	1
5. Convert to VCF format *(SHAPEIT)*	<1 min
6. Perform imputation *(Minimac3, 32 CPUs)*	39.25

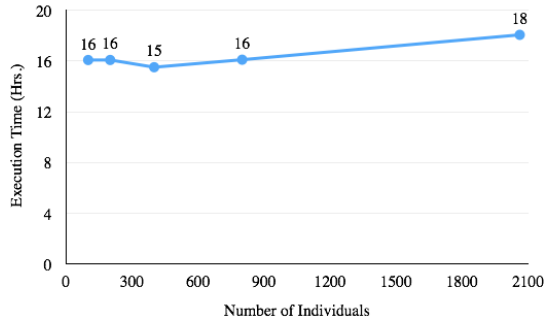

Fig. 6. Genotype imputation execution time for Chromosome 22 on the 24-CPUs server and 200-states of haplotypes estimation

imputation steps (Step 3 to Step 6 in Figure 4) by varying the number of individuals of Chromosome 22. We configured parameter configuration for 200 states, 24 CPUs and 35 MCMC iterations on the server. It took 18 hours for all individuals and the execution times of the genotype imputation were slightly different.

D. Performance Comparison

We also conducted another test case by varying the number of states on both the server and workstation but fixed the number of MCMC iterations to 35 and states to 400. In the server, we configured using 32 CPUs parameter whereas there were four CPUs on the workstation. Figure 7 depicts execution time comparison for only the genotype imputation steps (Step 3 to Step 6 in Figure 4). It is quite different amongst the number of states than the number of individuals. It took about 41 hours to complete the analysis of Chromosome 22 and it would take about one month for all chromosomes with the large reference panel on the 32-CPUs server. We identify the execution times of haplotypes estimation by SHAPEIT as shown in Figure 8. The execution time was about one hour for the 400-states estimation. In addition, we classify execution times of parameter estimation and imputation steps by Minimac3 running on the server as shown in Figure 9. The execution time of the parameter estimation steps was about one day for the 400-states analysis whereas the execution time of the imputation steps was about three hours. The parameter estimation steps are more computational intensive than the imputation steps.

VI. DISCUSSION

Our bioinformatician has phased and imputed a dataset at the RIKEN research laboratory in Japan with a reference panel using manual distribution and execution using shell scripts

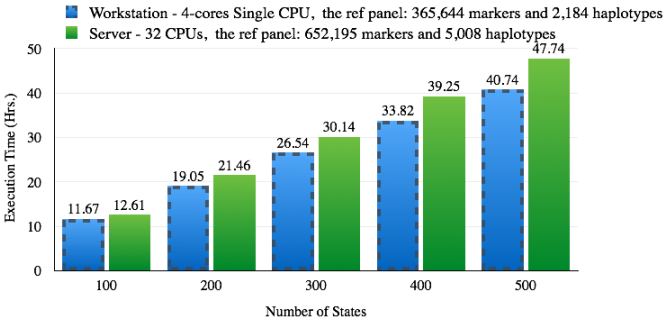

Fig. 7. Genotype imputation execution time comparison for Chromosome 22 on the 32-CPUs server

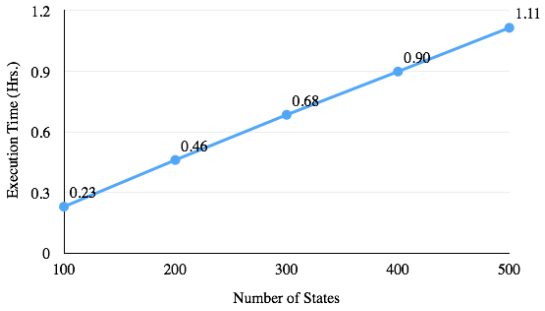

Fig. 8. The execution time of haplotypes estimation by SHAPEIT on the 32-CPUs server

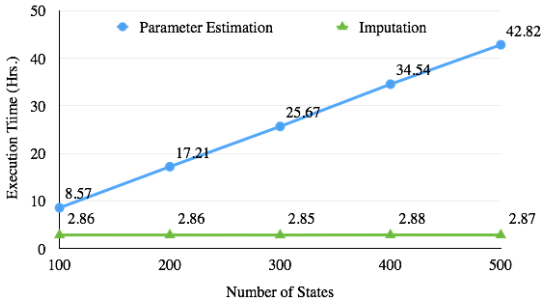

Fig. 9. Execution times of parameter estimation and imputation using Minimac3 on the 32-CPUs server

on a large cluster. The cluster has 80K nodes, each node is installed SPARC64 XIfx 1.975 GHz, 32 cores, and 32 GB RAM [20]. The bioinformatician has managed two tracks of genotype imputation consisting of the phasing and imputing step using MACH, the phasing step using MACH, and then the imputation step using Minimac. It took a month to do so. Both tracks do not require data format conversion for their data flows. As we described in the previous section, Minimac and Minimac3 are different. Minimac is the previous version that supports MACH outputs for performing imputation. Minimac3 is the latest version and supports only the VCF format which is a standard format of reference panels provided by the 1000 Genomes project.

In this work, we propose a new efficient process that performs haplotypes estimation using the SHAPEIT-based method and phased-haplotypes imputation using the Minimac3-based method. Even though the formats of data flows between haplotypes estimation and imputation analysis are different, the

format conversion in our proposed process is not computational intensive. It takes less than one minute. We classify that the Minimac3-parameter estimation step is the most computational intensive amongst the relevant steps, followed by the SHAPEIT-haplotypes estimation and the Minimac3-imputation step respectively. It may take about one month for analyzing all chromosomes using our proposed process. From our performance comparison experiments, we are confident that it can be faster for automatic process or workflow orientation if utilizing a proper parallelized running paradigm.

In GWAS, there usually are a large number of samples (>200). Our testing dataset also is some sort of GWAS samples. It has 2,062 samples and 545,080 markers. In our performance measurement on the Linux workstation and the server, we configured the states-parameter up to 500 which means conditioning haplotypes were used in the estimations. The recommended number of states used in the MCMC is closely related to the individuals being estimated and significantly influences intensive computation. However, the number of 400 states is widely used and acceptable. Another important factor influences the performance is the size of reference panels [2]. For example, the haplotypes estimation and imputation for Chromosome 22 consume about 41 hours running time with the 5,008 haplotypes reference panel by SHAPEIT and Minimac3 respectively. There are freely access of available reference panel providers such as the 1000 Genomes project and HapMap project [21].

VII. Conclusions

We propose the practical and efficient process for enhancing genotype imputation based analysis on SNPs using HPC. We split the analysis into three main steps consisting of the data quality control using PLINK, haplotypes estimation using SHAPEIT, and imputation using Minimac3 for speeding up the SNPs analysis process. Our practical process is a novel idea that no anyone has ever proposed before. We present the result of performance measurement of the NARAC-Chromosome 22 with the large reference panels on the 4-cores single CPU standard Linux workstation and the 32-CPUs server. The proposed process runs all steps successfully and returns the outputs for further gene expression analysis. The execution time is acceptable without any related tool improvement.

This paper proposes a new practical process that will be beneficial to bioinformaticians at the beginner or intermediate level. We are on progress for enhancing genotype imputation based analysis tools to support parallelization running on a large cluster.

Acknowledgment

The NARAC data was supported by the GAW grant (R01GM031575) and the NIH grant that supports a collection of RA data (AR44422). We also thank the National e-Science Infrastructure Consortium, Thailand, for providing computing resources.

References

[1] Y. Li, C. Willer, S. Sanna, and G. Abecasis, "Genotype imputation," *Annu Rev Genomics Hum Genet*, vol. 10, pp. 387-406, Sep. 2009.

[2] J. Marchini, and B. Howie, "Genotype imputation for genome-wide association studies," *Nature Reviews Genetics*, vol. 11, pp. 449-511, Jul. 2010.

[3] K. Hao, E. Chudin, J. McElwee, and E.E. Schadt, "Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies," *BMC Genetics*, vol. 10, no. 27, Jun. 2009.

[4] G.M. Peloso, N. Timofeev, and K.L. Lunetta, "Principal-component-based population structure adjustment in the North American Rheumatoid Arthritis Consortium data: impact of single-nucleotide polymorphism set and analysis method," *BMC Proceedings*, vol. 3, no. Suppl 7, pp. S108, Dec. 2009.

[5] U. Sangket, S. Mahasirimongkol, W. Chantratita, P. Tandayya, and Y.S. Aulchenko, "ParallABEL: an R library for generalized parallelization of genome-wide association studies," *BMC Bioinformatics*, vol. 11, no. 217, Apr. 2009.

[6] B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G.R. Abecasis, "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing," *Nat Genet*, vol. 44, no. 8, pp. 955-959, Jul. 2012.

[7] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE 77*, 1989, pp.257286.

[8] B.N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genetics*, vol. 5, no. 6, pp. e1000529. Jun. 2009.

[9] C.Y. Cheung, E.A. Thompson, and E.M. Wijsman, "GIGI: an approach to effective imputation of dense genotypes on large pedigrees," *Am J Hum Genet*, vol. 92, no. 4, pp. 504-516, Apr. 2013.

[10] Y. Li, C.J. Willer, J. Ding, P. Scheet, and G.R. Abecasis, "MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes," *Genetic epidemiology*, vol. 34, no. 8, pp. 816-834, Dec. 2010.

[11] Willer CJ, Sanna S, Jackson AU el al. C.J. Willer, S. Sanna, and A.U. Jackson el al, "Newly identified loci that influence lipid concentrations and risk of coronary artery disease, " *Nat Genet*, vol. 40, no. 2, pp. 161-169, Feb. 2008.

[12] M. Hudik, and M. Hodon, "Performance optimization of parallel algorithms," *Communications and Networks*, vol. 14, no. 4, pp. 436-446, Aug. 2014.

[13] 1000 Genomes. (2015, May.). VCF (Variant Call Format) version 4.1. [Online]. Available: http://www.1000genomes.org/wiki/analysis/variant call format/vcf-variant-call-format-version-41.

[14] S. Purcell, B. Neale, and K. Todd-Brown et al, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, no. 3, pp. 559-575, Sep. 2007.

[15] O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, et al. J. O'Connell, D. Gurdasani, and O. Delaneau et al, "A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness, " *PLoS Genet*, vol. 10, no. 4, pp. e1004234, Apr. 2014.

[16] K. Misawa, and N. Kamatani, "ParaHaplo 3.0: A program package for imputation and a haplotype-based whole-genome association study using hybrid parallel computing," *Source Code Biol Med*, vol. 6, no. 1, pp. 10, May. 2011.

[17] C.A. Anderson, F.H. Petterssonm, G.M. Clarke, L.R. Cardon, A.P. Morris, and K.T. Zondervan, "Data Quality Control in Genetic Case-Control Association Studies," *Nature Protocols*, vol. 5, no. 9, pp. 15641573, Aug. 2010.

[18] Autumn L. (2013, Jun.). Comparing BEAGLE, IMPUTE2, and Minimac Imputation Methods for Accuracy, Computation Time, and Memory Usage. [Online]. Available: http://blog.goldenhelix.com/?p=1911.

[19] K. Wolstencroft, R. Haines, and D. Fellows et al, "The Taverna Workflow Suite: Designing and Executing Workflows of Web Services on the Desktop, Web or in the Cloud," *Nucleic Acids Research*, vol. 41, no. Web Server issue, pp. W557W561, May. 2013.

[20] RIKEN. (2015, May). Supercomputer System Summary of Operations. [Online]. Available: http://accc.riken.jp/secure/5082/ operation_summary_en050401.pdf.

[21] R.A. Gibbs, J.W. Belmont, P. Hardenbol, T.D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, and Y. Shen, "The International HapMap Project," *Nature*, vol. 426, no. 6968, pp. 789-796, 2003.