Journal of Advanced Engineering Trends
ISSN : 2682 - 2091

Vol. 40, No.2. July 2021

http://jaet.journals.ekb.eg

Pre-processing Steps for Genome-wide High-density NARAC Dataset Facilitates its Haplotype Block Partitioning

Fatma S. Ibrahim[1,*], Mohamed N. Saad[1], Ashraf M. Said[1], Hesham F. A. Hamed[2]

[1] *Biomedical Engineering Department, Faculty of Engineering, Minia University, Minia, Egypt*

[2] *Electrical Engineering Department, Faculty of Engineering, Minia University, Minia, Egypt, Faculty of Engineering, Egyptian Russian University, Cairo, Egypt*

* corresponding author, E-mail: fatmasayed93@mu.edu.eg

ABSTRACT

The pre-processing phase is a crucial step to prepare any data for deep considerable analysis. Genome-wide data is considered big data; dealing with such data is not an easy task and still poses a significant challenge. The genome-wide association study (GWAS) is based on enormous high-density data with high throughput. This paper has illustrated the main pre-processing steps on data from North American Rheumatoid Arthritis Consortium (NARAC) for preparing it for haplotype block partitioning using different methods and with different platforms. This paper's main objective is to summarize the steps of pre-processing the raw genotyped dataset to prepare it for haplotype block partitioning and further analyses. Besides, we present each practical step by clear tables for better visualizing, elucidation, and workflow interpretation. Besides, we aimed to overcome the missing data and normalize the output in a standardized format. Eventually, this will improve the understanding of such data formats and build the foundation stone of critical genome-wide experiments and studies. Thus, this work could a guide for other researchers who use similar data. The pre-processed data will be applied to imputation, BigLD block partitioning under R and Haploview methods. Our sequence of pre-processing steps includes preparing the characters to be in a form that is suitable for imputation. The next step is recording data in 0,1,2 format to be proper for the BigLD. We were finally preparing data for Haploview to provide clear haplotype block partitioning, association analysis, and furthermore.

Keywords: *Minor allele frequency, Genome-wide studies, Single nucleotide polymorphism. Haplotype block. Linkage disequilibrium.*

1. INTRODUCTION

Genetic variations, especially single nucleotide polymorphisms (SNPs), hold an enormous promise for scientists in diagnosis, treatment, prevention, cure many diseases, and the causes of certain phenotypes [1]–[4]. SNPs are the most common genetic variation type [5], [6], and they compose about 90% of all human genetic variation [7]. Each SNP stands for a substitution in a single DNA nucleotide [8], [9]. SNPs are characterized by having at least a frequency of 1% of an allele in a population [2]. Commonly, SNPs do not occur within genes and are found in non-coding regions [10], [11]; yet how they contribute to disorders and phenotypes is unknown. Several

Revised: 18, November 2020, Accepted: 28, November 2020

research disciplines are based on SNPs including; large-scale population-based genetic association studies [12], family-based disease inheritance tracking, gene mapping [13], crop genetics [12] pharmacogenetic applications [16] and, etc. [15]. SNPs act as a biomarker, i.e., identifier for traits and diseases, therefore work very well in studies that are non-hypothesis driven [17], [18]. In the last years, SNPs serve as a powerful tool prominently in complex disease association studies and uncovered the risk loci [12], [19]–[22]. Since SNPs are inherited together through generations, SNPs at different loci are linked to each other, and this property

of the association is called linkage disequilibrium (LD) [23]–[25]. As a consequence of the recombination events, the association is cut and form block-like structures through the genome sequence called haplotype blocks [26], [27]. Likewise, haplotype blocks are treated as biomarkers and are often preferred over a single SNPs approach for several rationales in many cases [21], [28], [29]. Haplotype blocks reduce markers density by gathering several correlated SNPs as one block, thus reducing the required analyses and tests [30]. Besides, haplotype blocks allow capturing information about the evolutionary history of the population and species. Furthermore, the haplotype blocks approach considers the interactions and interrelationship between SNPs [31], [32]. The advent of new DNA technology facilitates SNP discovery and significantly accelerate high-throughput sequencing and genotyping on a genome-wide scale [33]–[36]. In humans, millions of SNPs are identified. Dealing with such big data is a considerable challenge and a time-consuming process [5], [22]. Data pre-processing is an essential step toward further analysis. This paper aims to address the genotype dataset to prepare it for haplotype block partitioning using different methods. We have conducted this experimental analysis on genomic-scale patient-control data from the North American Rheumatoid Arthritis Consortium (NARAC) [37]. NARAC data consist of 531,689 SNPs for 2,062 participants (868 rheumatoid arthritis (RA) patients and 1,194 controls). Missing data were estimated by imputation [38]. We have handled the data format to fit the following haplotype block partitioning methods: Gabriel Confidence Interval test (CIT) [39], four-gamete test [40], solid spine (SSLD) [41], and BigLD [42]–[44].

A sample of SNPs data is presented, and all steps of pre-processing are illustrated in the next sections.

2. SYSTEM DESCRIPTION

Figure 1 is a scheme that summarizes all the pre-processing steps from the raw data into an output that would be fitting for further analysis.

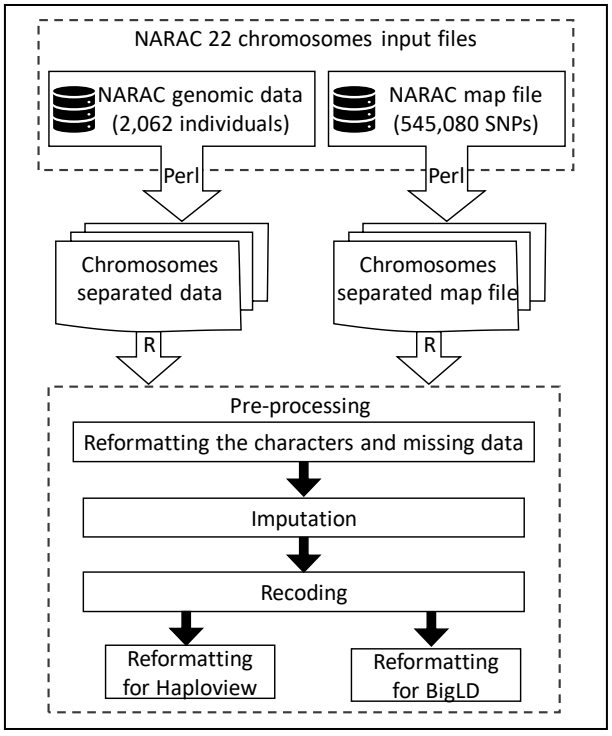

Fig.1. Summary of system and experimental design.

NARAC dataset comes in two files, the SNPs array file and the map file [37]. We have split the raw dataset into disjointed 23 matrices

for the 23 chromosomes. Our study and analysis are based only on the autosomal SNPs because allosome pairs have different structures, two X chromosomes in females and one X and one Y chromosome in males [45]. We neglected the allosome SNPs and worked only on the 22 chromosomes. The number of SNPs in the sex chromosomes are 13,395. After excluding the sex chromosomes' SNPs, we end up with 531,689 SNPs. The size of the SNPs matrices is 1,096,342,718 (531,689 SNPs ×2,062 participants). Each cell in the matrix is five-length character vectors, which are: whitespace, the letter of the first allele, then underscore, the letter of the second allele, and whitespace. The missing data is represented by "? _?". The data dividing has been executed using Perl, and the rest of the steps have been implemented by the R programming language.

Table 1. A sample of the row dataset after trucking the information columns. The genotypes appear in the format where there is an underscore between the characters. "?_?" is for missing data.

rs10439884	rs2260810	rs1296971	rs2257224
G_G	A_A	A_A	G_G
G_G	A_A	A_A	G_G
G_G	A_A	A_A	G_G
G_G	A_A	A_A	A_A
G_G	A_A	A_A	G_G
G_G	A_A	A_A	G_G
G_G	A_A	A_A	G_G
G_G	G_G	C_C	A_G
A_G	A_G	A_C	A_G
A_G	A_G	A_C	G_G
?_?	A_G	A_C	A_G
G_G	A_G	A_C	G_G

3. REFORMATTING THE DATA

We have excluded the sex chromosomes; only autosomal SNPs were picked up and pre-processed. Table 1 shows a portion of the chromosome 21 genotype data. We have truncated the first nine descriptive columns since the SNPs' data are targeted in the pre-processing step. The row data format is not suitable for many genetic analysis tools.

For instance, to pre-process the data for imputation, first, we have removed the underscore between alleles and recode missing data from "? _?" to "NA." Table 2 displays the reformatting for imputation. Data translation is done through the R package "Stringr" [46], [47].

Table 2. The reformatted sample for preparing the data for imputation. The missing value is spotlighted by bold font.

rs10439884	rs2260810	rs1296971	rs2257224
GG	AA	AA	GG
GG	AA	AA	GG
GG	AA	AA	GG
GG	AA	AA	AA
GG	AA	AA	GG
GG	AA	AA	GG
GG	AA	AA	GG
GG	GG	CC	AG
AG	AG	AC	AG
AG	AG	AC	GG
NA	AG	AC	AG
GG	AA	AA	GG

4. IMPUTATION AND RECODING

Imputation is a feasible process to overcome missing data by inferring untyped genotypes based on neighbouring SNPs [48], [49]. Not all tools are practically executable with missing data [42]. Several studies have recommended a genotype pre-processing before genetic analysis [29], [50]. Childers et al. have carried out genotype imputation for untyped SNPs to the same dataset using three different packages; BIMBAM, IMPUTE, and TUNA. Childers et al.'s results show that some untyped SNP demonstrated a significant association with RA [51]. We have proposed the probability distribution method of genotype imputation. We recode the data and impute the missing

genotypes using the R package "Synbreed" based on marginal allele distribution [52], [53].

Table 3. The sampled data after imputation in biallelic format.

rs10439884	rs2260810	rs1296971	rs2257224
GG	AA	AA	GG
GG	AA	AA	GG
GG	AA	AA	GG
GG	AA	AA	AA
GG	AA	AA	GG
GG	AA	AA	GG
GG	AA	AA	GG
GG	GG	CC	AG
AG	AG	AC	AG
AG	AG	AC	GG
GG	AG	AC	AG
GG	AA	AA	GG

The same imputation has been made for the NARAC dataset at a chromosomal level by Ibrahim et al. [44]. The imputation process took about 150 hours for the genome-wide data. Table 3 is the sampled data after imputation in the biallelic format. Data imputation completes the SNPs' coverage and facilitates the genome-wide analysis by combining the results genotypes that depend on different genotyping platforms. Table 4 is the sampled data after recoding, where 0 is for the major homozygous genotype, 1 for the heterozygous genotype, and 2 for the minor homozygous genotype.

Table 4. Sampled data after imputation and recoding in 0,1 and 2 format.

rs10439884	rs2260810	rs1296971	rs2257224
0	0	2	2
0	0	2	2
0	0	2	2
0	0	2	0
0	0	2	2
0	0	2	2
0	0	2	2
0	2	0	1
1	1	1	1
1	1	1	2
1	1	1	1
0	0	2	2

The pre-processed data is saved as a genotype object that combines the SNPs array and the corresponding map file. The new map matrix that defines the SNP genotypes is presented in Table 5. The filter SNPs' arrays thus become valid for haplotype block partitioning. The data format in Table 3 is convenient to undergo the BigLD haplotype block partitioning and several genetic analyses. The sampled data after imputation in biallelic markers are shown in Table 4. The pre-processed data is saved as a genotype object that combines the SNPs array and the corresponding map file. The filter

Table 5. The sample map file with defined genotypes in each locus. "Refer." is the reference genotype, also called minor homozygous. "Heter." For the heterozygous genotype and "Alter." for the alternative genotype, which also called the minor homozygous

rsID	Chr	Position	Refer.	Heter.	Alter.
rs10439884	21	9993822	GG	AG	AA
rs2260810	21	13562271	AA	AG	GG
rs1296971	21	13609442	CC	AC	AA
rs2257224	21	13690214	AA	AG	GG

SNPs' arrays thus become valid for haplotype block partitioning. Ultimately, we prepare the data into the final format to fit the Haploview input format. The Haploview input takes up two columns for every single genotype, one for each allele. Table 12 shows the sampled in input Haploview format. The data is non-family-based, so the same ID has been appointed for the family ID, and the individual ID in the first two columns, and paternal and maternal IDs have been set as "0" in the third and fourth columns. The Individual's gender was assigned 1 for male and 2 for female in the fifth column. The sixth column is for the affection status to be used for association tests; the controls = one and cases =2. The rest of the columns are for marker genotypes. Each marker is represented by two columns, one for each allele. and coded as the following: 1=A, 2=C, 3=G, and 4=T [41].

5. Conclusion

This empirical studyhas shown the detailed steps for preparing high-density data of 1,096,342,718 genotyped SNPs. Three main steps are adopted; reformatting, imputation, and recoding. The problem of missing data has been solved in a reasonable computing time by imputation. The proposed approach explains how to deal with such a big dataset, making it suitable for haplotype block partitioning using Haploview or BigLD. Our pre-processing steps pose a practical guide for the researchers and geneticists who want to make similar analyses. We have gathered all codes and build them under the R programming language, which simplified the process and the workflow. This paper evokes a vivid image of how to pre-process genome-wide data under a personal computer environment in a clearly defined manner. The entire time computing was relatively long; about 750 hours using Intel Core i7-8550U CPU 1.99 GHz with 16 GB RAM. Thus, in future work, parallel computing should be considered. However, the most time-consuming process was the reformatting, which has taken about 350 hours, then preparing data to Haploview, which has taken about 250 hours. The imputation process has

Table 6. The sample of the input format for the Haploview

Family name	Individual ID	Father' s ID	Mother' s ID	Gender	Affection	rs10439884	rs10439884	rs2260810	rs2260810	rs1296971	rs1296971	rs2257224	rs2257224
D0024949	D0024949	0	0	2	1	3	3	1	1	1	1	3	3
D0024302	D0024302	0	0	2	1	3	3	1	1	1	1	3	3
D0023151	D0023151	0	0	2	1	3	3	1	1	1	1	3	3
D0022042	D0022042	0	0	2	1	3	3	1	1	1	1	1	1
D0021275	D0021275	0	0	2	1	3	3	1	1	1	1	3	3
D0021163	D0021163	0	0	2	1	3	3	1	1	1	1	3	3
D0020795	D0020795	0	0	2	1	3	3	1	1	1	1	3	3
D0020691	D0020691	0	0	2	1	3	3	3	3	2	2	3	1
D0019121	D0019121	0	0	2	1	1	3	3	1	1	2	3	1
D0018942	D0018942	0	0	2	1	1	3	3	1	1	2	3	3
D0016405	D0016405	0	0	2	1	3	3	3	1	1	2	3	1
D0016076	D0016076	0	0	2	1	3	3	3	1	1	2	3	3

taken 150 hours. The previously mentioned imputation methods have had a wide range of processing time, i.e., IMPUTE has more than 400 hours processing time while TUNA took more than 12 hours. We cannot make an accurate comparison between our proposed method and other methods since both studies' hardware configuration should be the same.

Nowadays, the high throughput genome-scale data are available more than any time before. It is essential to address such data to exploit them and focus on understanding and preparing them in different forms, which eventually help to utilize them in various platforms.

4. ACKNOWLEDGMENT

The authors acknowledge the Genetic Analysis Workshop grant [R01 GM031575] for providing the NARAC dataset. This work is based on data gathered with the support of grants from the National Institutes of Health [NO1-AR-2-2263, RO1-AR-44422], and the National Arthritis Foundation (NAF).

REFERENCES

[1] S. Alicia, A review of HLA allele and SNP associations with highly prevalent infectious diseases in human populations. *Swiss Med. Weekly,*150(1516), 1–14, 2020

[2] A. J. Brookes, The essence of SNPs. *Gene*, 234(2), 177–186, 1999.

[3] Bethesda, Understanding Human Genetic Variation: NIH Curriculum Supplement Series. 2007.

[4] N. Vankadari, Overwhelming Mutations or SNPs of SARS-CoV-2: A Point of Caution. *Gene*, (just-accepted) 144792, 2020.

[5] 1000 Genomes Project Consortium, A global reference for human genetic variation.*Nature*, 526 (7571), 68–74, 2015.

[6] A. Syvänen, Accessing Genetic Variation : Genotyping Single Nucleotide Polymorphisms. *Nat. Rev. Genet.*, 2(12), 930–942, 2001.

[7] E. Lander and L. Kruglyak, Genetic dissection of complex traits : reporting linkage results. *Nat. Genet.*, 11(3), 241–247, 1995.

[8] R. Sachidanandam, D. Weissman, Steven C. Schmidt, J. M. Kakol, L. D. Stein, and G. Marth, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822), 928–934, 2001.

[9] 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061, 2010.

[10] K. A. Frazer, S. S. Murray, N. J. Schork, and E. J. Topol, Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, 10(4), 241–251, 2009.

[11] F. Crea, P. L. Clermont, A. Parolia, Y. Wang, and C. D. Helgason, The non-coding transcriptome as a dynamic regulator of cancer metastasis. *Cancer metastasis Rev.*, 33(1), 1–16, 2014.

[12] T. A. Manolio *et al.*, Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753, 2009.

[13] D. G. Wang *et al.*, Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280 (5366),1077–82, 1998.

[14] A. Rafalski, Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.*, 5(2), 94–100, 2002.

[15] C. Ding, Other ' applications of single nucleotide polymorphisms. *Trends Biotechnol.*,25(7), 279–283, 2007.

[16] K. M. Giacomini *et al.*, The Pharmacogenetics Research Network : From SNP Discovery to Clinical Drug Response. *Clin. Pharmacol. Ther.*, 81(3),328–345, 2007.

[17] K. E. Fowler, R. Pong-wong, J. Bauer, E. J. Clemente, C. P. Reitter, and N. A. Affara, Genome-wide analysis reveals single nucleotide polymorphisms associated with fatness and putative novel copy number variants in three pig breeds. *BMC Genomics*, 14(1),1, 2013.

[18] M. N. Saad, M. S. Mabrouk, A. M. Eldeib, and O. G. Shaker, Effect of MTHFR,

TGFβ1, and TNFB polymorphisms on osteoporosis in rheumatoid arthritis patients, *Gene*, 568(2),124–128, 2015.

[19] J. N. Hirschhorn and M. J. Daly, Genome-wide Association Studies for Common Diseases and Complex Traits. *Nat. Rev. Genet.*, 6(2), 95–108, 2005.

[20] G. Glusman, H. C. Cox, and J. C. Roach, Whole-genome haplotyping approaches and genomic medicine. *Genome Med.*, 6(9), 1–16, 2014.

[21] J. Fu, E. A. M. Festen, and C. Wijmenga, Multi-ethnic studies in complex traits. *Hum. Mol. Genet.*, 20(2), 206–213, 2011.

[22] C. Bycroft *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726),203–209, 2018.

[23] A. Blomhoff *et al.*, Linkage disequilibrium and haplotype blocks in the MHC vary in an HLA haplotype-specific manner assessed mainly by DRB1 * 03 and DRB1 * 04 haplotypes. *Genes Immun.*,7(2) 130–140, 2006.

[24] N. Li and M. Stephens, Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data.*Genetics*, 165(4), 2213–2233, 2003.

[25] B.S. Weir, Linkage Disequilibrium, and Association Mapping. *Annu. Rev. Genomics Hum. Genet.*, 9,129-142., 2008.

[26] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander, High-resolution haplotype structure in the human genome. *Nat. Genet.*, 29(2), 229–232, 2001.

[27] H. Yoshikawa *et al.*, Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. *Science*, 294(5547), 1719–1723, 2001.

[28] M. N. Saad, Identification of rheumatoid arthritis biomarkers based on single nucleotide polymorphisms and haplotype blocks : A systematic review and meta-analysis. *J. Adv. Res.*, 7(1), 1–16, 2016.

[29] M. N. Saad, M. S. Mabrouk, A. M. Eldeib, and O. G. Shaker, Studying the effects of haplotype partitioning methods on the RA-associated genomic results from the North American Rheumatoid Arthritis Consortium (NARAC) dataset. *J. Adv. Res.*, 18, 113–126, 2019.

[30] P. Ballesta, C. Maldonado, P. Pérez-Rodríguez, and F. Mora, SNP and Haplotype-Based Genomic Selection of Quantitative Traits in Eucalyptus globulus. *Plants*, 8(9), 331, 2019.

[31] A. J. Lorenz, M. T. Hamblin, and J. Jannink, Performance of Single Nucleotide Polymorphisms versus Haplotypes for Genome-Wide Association Analysis in Barley.*PLoS One*, 5(11), 14079, 2010..

[32] E. Capriotti, R. Calabrese, and R. Casadio, Sequence analysis Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information.*Bioinformatics*, 22(22), 2729–2734, 2006.

[33] D. Grattapaglia, O. B. Silva-junior, M. Kirst, B. M. De Lima, D. A. Faria, and G. J. P. Jr, High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus : assay success, polymorphism, and transferability across species. *BMC Plant Biol.*, 11(65), 1–18, 2011.

[34] M. M. L. Wong, C. H. Cannon, and R. Wickneswari, Development of high-throughput SNP-based genotyping in Acacia auriculiformis x A . mangium hybrids using short-read transcriptome data. *BMC Genomics.*, 13(1), 726, 2012.

[35] J. W. Davey, P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter, Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.*, 12(7), 499–510, 2011.

[36] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song, Genotype and SNP calling from next-generation sequencing data, *Nat. Rev. Genet.*, 12(6), 443–451, 2011.

[37] C. I. Amos *et al.*, Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. *BMC Proc.*, 3(S7), S2, 2009.

[38] K. K. Id, S. T. Id, F. K. Id, and G. T. Id, A genotype imputation method for identified haplotype reference information using recurrent neural network. *PLoS Comput. Biol.*, 16(10), 1–21, 2020

[39] S. B. Gabriel *et al.*, The structure of haplotype blocks in the human genome. *Science*, 296(5576), 2225–2229, 2002.

[40] N. Wang, J. M. Akey, K. Zhang, R. Chakraborty, and L. Jin, Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. *Am. J. Hum. Genet.*, 71(5), 1227–1234, 2002.

[41] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, Haploview : analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2), 263–265, 2005.

[42] S. A. Kim, C. Cho, S. Kim, S. B. Bull, and Y. J. Yoo, A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics*, 34(3), 388–397, 2018.

[43] S. A. Kim, C. Cho, S. Kim, S. B. Bull, and Y. J. Yoo, Supplementary Information for ' A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics*, 34(3), 1–40, 2018.

[44] F. S. Ibrahim, Haplotype Block Partitioning for NARAC Dataset Using Interval Graph Modeling of Clusters Algorithm. 2018 9th Cairo International Biomedical Engineering Conference (CIBEC). IEEE, 134-137, 2018.

[45] A. J. Griffiths, D. T. S. Jeffrey H Miller, Richard C Lewontin, and William M Gelbart, *An Introduction to Genetic Analysis*. W. H. Freeman, 2000.

[46] H. Wickham, stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. 2019, [Online]. Available: https://cran.r-project.org/web/packages/stringr/stringr.pdf

[47] H. Wickham, Stringr: modern, consistent string processing. *R J.*, 2(2), 38–40, 2010.

[48] Y. Li, C. Willer, and S. Sanna, Genotype Imputation. *Annu. Rev. Genomics Hum. Genet.*,10, 387–406, 2009.

[49] H. Density, U. L. Phasing, H. D. Daetwyler, G. R. Wiggans, B. J. Hayes, and J. A. Woolliams, Imputation of Missing Genotypes From Sparse to High Density Using Long-Range Phasing. *Genetics*, 189(1), 317–327, 2011.

[50] K. Damkliang, P. Tandayya, U. Sangket, S. Mahasirimongkol, and E. Pasomsab, An Efficient Process for Enhancing Genotype Imputation in Genome-wide Association Studies Using High-Performance Computing. *Int. Comput. Sci. Eng. Conf. (ICSEC)*. 1–6, 2015.

[51] D. K. Childers, G. Kang, N. Liu, G. Gao, and K. Zhang, Application of imputation methods to the analysis of rheumatoid arthritis data in genome-wide association studies. *BMC Proc.*, 3(7), S24, 2009.

[52] V. Wimmer, T. Albrecht, H.-J. Auinger, and Chris-Carolin Schön, synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*, 28(15), 2086–2087, 2012.

[53] R. Nielsen, T. Korneliussen, A. Albrechtsen, Y. Li, J. Wang, SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One*, 7(7), 37558, 2012.

اعتماد تسلسل من المعالجة المسبقة لمجموعة بيانات NARAC عالية الكثافة على مستوى الجينوم للاستخدام الأمثل لتقسيم كتل النمط الفرداني

الملخص:

تعتبر عملية المعالجة المسبقة خطوة هامة لإعداد أي بيانات لتحليل دقيق وعميق. وتعد البيانات على مستوى الجينوم –الذي يشمل كل المحتوى الوراثي– بيانات هائلة ولذلك فإن التعامل معها ليس بالمهمة السهلة ولا يزال يمثل تحديًا كبيرًا للباحثين. وتعتمد دراسة الارتباط على مستوى الجينوم GWAS على تلك البيانات عالية الكثافة وذات الإنتاجية الوافرة. في هذه الورقة البحثية، قمنا بتوضيح خطوات المعالجة المسبقة الرئيسية لبيانات من اتحاد أمريكا الشمالية لالتهاب المفاصل الروماتويدي (NARAC) لإعدادها وتقسيمها الى كتل النمط الفرداني Haplotype blocks باستخدام طرق ومنصات مختلفة وقد تم تطبيق البيانات المعالجة مسبقًا على عملية احتساب البيانات المفقودة Imputation وطريقة BigLD لتقسيم كتلة النمط الفرداني باستخدام لغة البرمجة R وكذلك على الطرق المستخدمة في برنامج Haploview. يتضمن تسلسل المعالجة المسبقة لدينا إعداد الحروف لتكون في شكل مناسب لعملية احتساب البيانات المفقودة. الخطوة التالية هي تسجيل البيانات بتنسيق 0،1،2 لتكون مناسبة لطريقة BigLD. وفي النهاية قمنا بتجهز البيانات لطرق Haploview. وعليه فان هذا العمل يسهم في تسهيل تقسيم بيانات الاختلافات الوراثية من نوع تعددات أشكال النوكليوتيدات المفردة الى كتل النمط الفرداني Haplotype blocks بشكل واضح، ويساعد علاوة على ذلك في تحليل الارتباطات.