A binary format for genetic data designed for large whole genome studies that enable both marker and strand based analyses

Athos Antoniades, Loizos Loizou, Aristos Aristodimou,
Constantinos S. Pattichis, *Senior Member, IEEE*

***Abstract*— Recent advances in genotyping technology have enabled large studies with data from thousands of subjects to contain half a million or more of single nucleotide polymorphisms (SNPs) marker per subject. This rapid increase in the size of data has generated the need to compress the data in order to reduce the storage capacity requirements and the memory required at run time to perform analysis on the data. The availability of so many markers across the whole genome has created opportunities for new methodologies to be implemented that take advantage of the relatively high density of the markers to perform analyses that take into account the Linkage Disequilibrium (LD), an effect where some combinations of genetic markers are non-randomly associated. Classical techniques for transforming genotypic data into a binary format are already in use by several applications however we demonstrate in this paper that the traditional transformations are not adequate for certain types of analyses as some information key to new methodologies of analyses is lost. We propose a new protocol for formatting binary genotypic data that can be used in all types of analyses while still offering a very high compression rate.**

I. Introduction

Until recently in genetic studies, due to limitations in genotyping technology, only a small subset of the genome could be analyzed. The introduction of affordable high throughput genotyping technologies allowed the assay of more than half a million loci variants (SNPs) per subject across the whole genome. Genetic association studies applying such technology have now become common as they allow investigation of the majority of common loci variants in the genome; such studies are typically called genome wide association scans (GWAS). In this paper we present a non-lossy format of encoding the data in the datasets generated from GWAS to a binary form.

The substance that encodes the genetic instructions of living organisms is Deoxyribonucleic acid (DNA). DNA consists of two long units called strands having a shape of a double helix [1], [2]. The genetic code is specified by the four nucleotide "letters" A (adenine), C (cytosine), T (thymine), and G (guanine). There are multiple different types of variations that can occur on the DNA strands, however traditionally the genetic analyses seem to focus on analyzing Single Nucleotide Polymorphisms (SNPs). SNP variation occurs when a single nucleotide, such as an A, replaces one of the other three nucleotide letters—C, G, or T [3]. For a variation to be considered a SNP it also needs to be present in at least 1% of the population. Due to the Linkage Disequilibrium effect (LD), SNPs serve as biological markers for pinpointing a region on the genome associated with a phenotypic trait even thought the SNP itself is not necessarily the variation responsible for the association. Rather it's one or more of the variations that are in LD with the SNP that is causing the association.

The need to reduce the size of the data is owed to the 100 fold increase in the genotyping capacity available today combined with the massive reduction in cost. Today's technology has the ability to analyze datasets up to 550,000 or even 1 million SNPs per subject with >99% accuracy, at a rate of >100 K genotypes per day and at a cost of around 20–30 cents per genotype [4], [5], [6].

Traditionally the genetic data in these studies is stored in the QTDT format introduced in the program Merlin [7]. Input files describe relationships between individuals in a dataset, store marker genotypes, disease status and quantitative traits and provide information on marker locations and allele frequencies.

There is already a technique for compressing this data by utilizing a binary encoding [8]. However, that technique was focused at encoding SNPs as markers to be used in analyses rather than strand based information. Today as more GWAS datasets became available researchers are developing innovative new methodologies to analyze them. Some of these methodologies are not relying so much on the SNPs as markers; rather they look at the sequence of genotyped alleles on each strand of DNA separately. The methodology used in plink to encode the data looses the information of which strand holds each allele's genotype for heterozygote SNPs, therefore making it impossible to run analyses that use strand information using the binary input format for GWAS. Also, recent technological advances in genotyping are enabling the detection of deletion regions. This may result in future datasets to incorporate markers in them that

Manuscript received August 8, 2008.

Athos Antoniades is with the Department of Computer Science, University of Cyprus, Nicosia, CY (corresponding author: +357-22-892685; e-mail: athos@ cs.ucy.ac.cy).

Loizos Loizou is with the Department of Computer Science, University of Cyprus, Nicosia, CY. (e-mail: cs06ll1@cs.ucy.ac.cy).

Aristos Aristodimou is with the Department of Computer Science, University of Cyprus, Nicosia, CY (e-mail: cs06aa2@cs.ucy.ac.cy).

Constantinos S. Pattichis is with the Department of Computer Science, University of Cyprus, Nicosia, CY (e-mail: pattichi@cs.ucy.ac.cy)

may have an allele missing not due to a genotyping error but due to a deletion on one of the two strands. The encoding format proposed in this paper will offer an efficient lossles compression that is also capable of encoding deletions separately from errors in genotyping or quality control removed markers.

The structure of the paper is as follows: Section II presents the traditional methodology of formatting GWAS data as well as the current alternative to compressing the data and the format proposed in this paper. Section III offers a comparison of the three formats identifying advantages and disadvantages of each. Section IV concludes describing the situations under which each format is optimal.

II. Methods and Material

A. The Commonly Used Format QTDT Merlin [7]

The QTDT file format described in Merlin is the one traditionally used for this type of data. It's split into three files; Pedigree files contain phenotypes for discrete and quantitative traits and marker genotypes for a specific number of subjects. They are usually white-space delimited files. The first (usually 6) columns contain information about the subject (Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype). The combination of the information of each subject must be unique. The next columns contain biallelic markers; typically SNPs. Marker genotypes are encoded as two consecutive integers, one for each allele, or using the letters "A", "C", "T" and "G". To denote missing alleles a sentinel value is used, typically "0".

Map files contain information for each single nucleotide polymorphism. They are used to analyze genetic markers into the equivalent pedigree file. Each line per marker usually contains 3, 4 or 5 columns (chromosome, SNP identifier, morgans or centimorgans and base-pair position). Each column is separated by white space.

Dat files describe the pedigree file. They include one row per data item in the pedigree file, indicating the data type providing a one-word label for each item.

B. Plink's Method for Binary Ped Files [8]

Plink is an excellent, open source program offering a comprehensive range of basic large-scale whole genome association analysis methodologies. It has been widely adopted since high dimensionality GWAS have become available as it enables researchers to efficiently analyze these large datasets in a computationally efficient manner.

In plink there is an encoding format for transforming QTDT Merlin data into binary formatted files. The approach used in the plink method, uses 2 bits for encoding biallelic markers with 4 possible states. Plink uses the encoding for each genotype given in Table I [8].

Testing on plink binary format showed that the exported binary file was 15 times smaller than the original file. The drawback however is that encoding of the heterozygote allele is the same regardless of what strand it's actually derived from. Therefore any analyses that rely on the sequence of the alleles on the strand will be missing this information.

TABLE I
PLINK BINARY PED FILE ENCODING [8]

Allele On Strand +	Allele On Strand -	Marker Encoding
A	A	00
A	a	01
a	A	
A	a	11
Missing Data	Missing Data	10

One analyses technique that needs the lost information is imputation. Imputation analysis is the practice of 'filling in' missing data with plausible values. It is a method for uncovering the genetic basis of human disease and it is used for inferring genotypes at observed or unobserved SNPs that can detect causal variants that have not been directly genotyped [9]. It is in essence an in-silico approach to discover the probability of the existence of a specific genotype for loci that haven't being directly genotyped but are known to be in LD with genotyped markers.

C. Proposed Method

Since imputation analyses require knowledge of which strand the alleles of heterozygote SNPs exist on, we need to encode each allele on each strand separately for all cases. There are a minimum of three states each allele can be in, these include the two possible nucleotides (commonly denoted as A and a) as well as the possibility of missing data at that location. The smallest number of bits that can encode the 3 states of an allele is 2, however with 2 bits we can actually encode a fourth state. In many existing studies this may not be used, even though it will have no impact on the capacity requirements the databases will have for storage. However, we propose that the fourth state is set to denote markers that are in deleted regions. This utilizes the extra available coding identifying a deleted allele from a missing allele due to quality control concerns, or genotyping error.

An analytical technique that enables the detection of these deletions that has started being applied is copy number variants (CNV) analyses. It refers to the genetic trait of differences in the number of copies of a particular region (for example a gene) present in the genome of an individual [10], [11]. To perform CNV analyses most algorithms rely on raw data from the genotyping platform. CNV algorithms are able to detect deletion regions as well as regions that are duplicated that may exist in some individual's strands.

However it should be clearly noted that CNV incorporates more information than just deleted regions. It can also detect regions that exist in more than one copy per strand. That information is not reflected in the proposed protocol; therefore methodologies that use that information would still rely on an external file with CNV regions.

In the proposed format the data are structured as a two dimensional vector of alleles. The first dimension's size is equal to the number of strands the subject has. Typically in humans all chromosomes have two strands with the exception of X and Y chromosomes in males that each have 1 strand. Even in these cases a second strand can exist listing the alleles of the second strand on males as missing.

Each element in the vector of each strand will encode an allele. The allele will be 2 bits long enabling encoding of a total of 4 states per allele, missing data, nucleotide 1, nucleotide 2 or deleted.

Table II presents the 4 different states that can be coded per allele. The term "Unknown" is used rather than the more typical "missing" to denote alleles that it's unclear what their genotypes are or if they are deleted so as not to confuse it with the deleted state that defines alleles that do not exist on that strand.

Table III shows how two alleles are encoded and create a biallelic allele such as SNPs, while still enabling the identification of deleted alleles from missing values if that information is available. By comparing two alleles together and using the encoding as proposed above the resulted encoding is shown in Table III.

The two strands in each Subject's vector need to be perfectly aligned, that is, the *i* th element of each vector will point to the same Marker's alleles one for each strand. This makes it easy to access information in the way it was traditionally accessed. To access the *i*'th marker's alleles the two bits at position *i* in each strand will carry a total of 4 bits, using the encoding column of Table III the genotype of Marker *i* can then be identified.

TABLE II
PROPOSED ALLELE ENCODING

Allele	Encoding
Unknown	00
A	01
a	10
Deleted	11

TABLE III
PROPOSED ENCODING FOR BI-ALLELIC MARKERS

Allele 1	Allele 2	Encoding
Unknown	Unknown	0000
Unknown	A	0001
Unknown	a	0010
Unknown	Deleted	0011
A	Unknown	0100
A	A	0101
A	a	0110
A	Deleted	0111
a	Unknown	1000
a	A	1001
a	a	1010
a	Deleted	1011
Deleted	Unknown	1100
Deleted	A	1101
Deleted	a	1110
Deleted	Deleted	1111

III. EXPERIMENTAL RESULTS

To compare the different formats we will consider the size of the resulting files in relation to the original QTDT Merlin format, the amount of information lost through the encoding of the original QTDT Merlin format and the ability of each format to retain different types of information available today. Table IV provides a summary of the comparison.

A. Information Loss

On one hand loosing information due to encoding is undesired; however, if the encoding is used to simply speed up analyses and reduce scratch space then it's not an issue as long as the original raw file is kept for future analyses. However if an algorithm is developed that needs the information of which strand each heterozygote marker's alleles are on , or if there are deletion regions overlapping the markers in either strand, then utilizing the encoding methodology proposed in this paper will in a single table or file, encode all of this information efficiently. Another issue is the actual storage of the data for long term use or for transferring over the internet. Utilizing a non-lossy approach to compressing the data that incorporates all genetic information into a single file can reduce the resources

TABLE IV
COMPARISON OF FILE FORMATS

QTDR Merlin	Proposed Protocol	Plink 's protocol
Information Loss for bi-allelic markers		
Used as Reference	None	strand location of heterozygote alleles Aa,aA Missing one of the two alleles A0, 0A, a0,0a
Added Information capability		
tri or quad allelic markers	Alleles deleted from a specific strand	None
Size*		
Pedigree File 3.6 GB Map File 12.6 MB **Total:** 3.612 GB	One binary file 496 MB	.bed file : 229.6 MB .fam file : 30 KB .bim file : 14.1 MB **Total** : 243.7 MB

*Using a dataset containing **1804** subjects with **532578** SNPs per subject.

required.

B. Added Information Capability

In this paper we focused on bi-allelic markers as they are the ones that current high throughput genotyping technologies are able to genotype. However it should be noted that the format of QTDT Markers enables encoding of tri or quad allelic markers as well since each allele is encoded as an ASCII character. Neither plink's binary ped file nor the format proposed could handle tri or quad allelic markers. Large datasets with tri-allelic markers are not existing today (to the best of our knowledge) while information on deleted markers is available through CNV analyses and other methodologies available to the various genotyping platforms. Therefore tri-allelic and quad allelic markers were ignored in this encoding until high throughput genotyping technologies make these data available. A new protocol could then be easily generated for supporting this analysis.

C. Size of Encoded Data

The compression rate can easily be estimated since both encodings are deterministic, however we also provide as an example an actual test we performed using an average dataset size of 1806 subjects and 532,579 markers. The plink binary ped file was able to compress the file to $1/15^{th}$ of it's original size while the proposed method compressed the file to exactly double the size of that achieving just $1/7.5^{th}$ of the original size. However, both compressions produce enough of a compression to overcome the issue of the large data since the data can now fit on the average computer's physical memory (RAM) as today's typical computer has at least 1 GB.

IV. Conclusion

The proposed format encodes genotypic data into a binary form in order to compress it and at the same time preserve all the information relating to the bi-allelic markers and the strand location of each allele. However it does double in size the resulting datasets from existing encoding methodologies that are lossy, but loose information only necessary to certain type of analytical approaches. The analytical approaches that would benefit from the proposed format of encoding are primarily the ones that take into account the strand on which heterozygote alleles are based on, the existence of the marker on just one of two possible strands, identifying if the second wasn't available due to genotyping errors or a deletion over the marker on that strand. Due to the need to use 2 bits per allele for encoding while only needing 3 states for each allele we were left with an available fourth state. We propose that the fourth state it is used to denote markers that are deleted as this information is becoming commonly available from genotyping platforms available already; however, future researchers may choose to use the fourth state to code a different state an allele can be in. Also in cases where compression of the data in a non-lossy way for storage, backup or data transfer is used, the methodology proposed would be ideal.

References

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walters, "Molecular Biology of the Cell 4th ed." New York and London: Garland Science, ch. 1, 2002.

[2] J. M. Butler, Forensic DNA Typing. Elsevier. ISBN 978-0-12-147951-0, pp. 14–15, 2001.

[3] M. P. Weiner, T. J. Hudson, Introduction to SNPs: Discovery of markers for disease, Biotechniques Suppl:4-7, 2002.

[4] S. Jenkins and N. Gibson, "High-Throughput SNP Genotyping," Comparative and Functional Genomics, vol. 3, no. 1, pp. 57-66, 2002.

[5] P. Y. Kwok, "High-throughput genotyping assay approaches," *Pharmacogenomics*, vol. 1, pp. 95–100, Feb 2000.

[6] J. N. Hirschhorn, M. J. Daly, "Genome-wide association studies for common diseases and complex traits, *Nature Review Genetics*, vol. 6 pp. 95–108, 2005.

[7] G. R. Abecasis, S. S. Cherny, W. O. Cookson and L. R. Cardon, "Merlin-rapid analysis of dense genetic maps using sparse gene flow trees," *Nature Genetics*, vol. 30, pp. 97-101, 2002.

[8] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W De Bakker, Daly MJ and Sham PC, "PLINK: a toolset for whole-genome association and population-based linkage analysis," *American Journal of Human Genetics*, vol. 81, pp. 559-75, Sep. 2007.

[9] J. Redon, "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444-454, Nov. 2006.

[10] R. E. Fay, "Valid inference from imputed survey data," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 41-48, 1993.

[11] J. L. Freeman, et al., "Copy number variation: New insights into genome diversity," *Genome Research*, vol. 16, pp. 949–61, 2006.