Binary Phased Table (bedp)

Ryan Lewis
08/2020

What is bedp?

- **bedp is a file format whose design is based on the framework of the PLINK file format, .bed**
  - This allows bedp to utilize the PLINK tool set
- **It is designed to hold genetic data in a binary format and allows users to access said data in a computationally efficient manner**

Reference: https://www.cog-genomics.org/plink2/formats

Why do we need bedp?

- **The PLINK .bed format stores data in 2 bits, giving 4 options**
  - 00 homozygous for first allele
  - 01 missing genotype
  - 10 heterozygous
  - 11 homozygous for second allele
- **The 4 options provided do not allow for storing Phased Haplotype Data**

bedp Structure

- **bedp is comprised of 3 files**
  - .bim
  - .fam
  - .bed - *must use extension name '.bed', not '.bedp', in order to use the PLINK tool set*
- **The file name for the 3 files must be identical in order to identify correspondence**
  - ex. file1.bed, file1.bim, file1.fam

.bim File

- **The .bim holds data on the SNPs present in the data set. Each line in the file represents 1 SNP and each line has 6 columns (tab delimiter)**

22	rs587755077	0	16050115	A	G
22	rs587654921	0	16050213	T	C
22	rs587712275	0	16050319	T	C
22	rs587769434	0	16050527	A	C
22	rs587638893	0	16050568	A	C
22	rs587720402	0	16050607	A	G

Chromosome Identification

.bim File

- **The .bim holds data on the SNPs present in the data set. Each line in the file represents 1 SNP and each line has 6 columns (tab delimiter)**

22	rs587755077	0	16050115	A	G
22	rs587654921	0	16050213	T	C
22	rs587712275	0	16050319	T	C
22	rs587769434	0	16050527	A	C
22	rs587638893	0	16050568	A	C
22	rs587720402	0	16050607	A	G

RSID or SNP Identifier

.bim File

- **The .bim holds data on the SNPs present in the data set. Each line in the file represents 1 SNP and each line has 6 columns (tab delimiter)**

22	rs587755077	0	16050115	A	G
22	rs587654921	0	16050213	T	C
22	rs587712275	0	16050319	T	C
22	rs587769434	0	16050527	A	C
22	rs587638893	0	16050568	A	C
22	rs587720402	0	16050607	A	G

Position in cM, default = 0

.bim File

- **The .bim holds data on the SNPs present in the data set. Each line in the file represents 1 SNP and each line has 6 columns (tab delimiter)**

22	rs587755077	0	16050115	A	G
22	rs587654921	0	16050213	T	C
22	rs587712275	0	16050319	T	C
22	rs587769434	0	16050527	A	C
22	rs587638893	0	16050568	A	C
22	rs587720402	0	16050607	A	G

Base-pair coordinate (position)

.bim File

- **The .bim holds data on the SNPs present in the data set. Each line in the file represents 1 SNP and each line has 6 columns (tab delimiter)**

22	rs587755077	0	16050115	A	G
22	rs587654921	0	16050213	T	C
22	rs587712275	0	16050319	T	C
22	rs587769434	0	16050527	A	C
22	rs587638893	0	16050568	A	C
22	rs587720402	0	16050607	A	G

1st Allele: corresponds to the clear bits, 0 (minor)

.bim File

- **The .bim holds data on the SNPs present in the data set. Each line in the file represents 1 SNP and each line has 6 columns (tab delimiter)**

22	rs587755077	0	16050115	A	G
22	rs587654921	0	16050213	T	C
22	rs587712275	0	16050319	T	C
22	rs587769434	0	16050527	A	C
22	rs587638893	0	16050568	A	C
22	rs587720402	0	16050607	A	G

2nd Allele: corresponds to the set bits, 1 (major)

.fam File

- **The .fam file holds data on the individuals contained in the data set. Each line represents 1 individual and has six columns(tab delimiter)**

NA21089	NA21089	0	0	0	-9
NA21090	NA21090	0	0	0	-9
NA21091	NA21091	0	0	0	-9
NA21092	NA21092	0	0	0	-9
NA21093	NA21093	0	0	0	-9
NA21094	NA21094	0	0	0	-9

Family ID, default = Individual ID

.fam File

- **The .fam file holds data on the individuals contained in the data set. Each line represents 1 individual and has six columns(tab delimiter)**

NA21089	NA21089	0	0	0	-9
NA21090	NA21090	0	0	0	-9
NA21091	NA21091	0	0	0	-9
NA21092	NA21092	0	0	0	-9
NA21093	NA21093	0	0	0	-9
NA21094	NA21094	0	0	0	-9

Individual ID

.fam File

- **The .fam file holds data on the individuals contained in the data set. Each line represents 1 individual and has six columns(tab delimiter)**

NA21089	NA21089	0	0	0	-9
NA21090	NA21090	0	0	0	-9
NA21091	NA21091	0	0	0	-9
NA21092	NA21092	0	0	0	-9
NA21093	NA21093	0	0	0	-9
NA21094	NA21094	0	0	0	-9

Father ID, default = 0

.fam File

- **The .fam file holds data on the individuals contained in the data set. Each line represents 1 individual and has six columns(tab delimiter)**

NA21089	NA21089	0	0	0	-9
NA21090	NA21090	0	0	0	-9
NA21091	NA21091	0	0	0	-9
NA21092	NA21092	0	0	0	-9
NA21093	NA21093	0	0	0	-9
NA21094	NA21094	0	0	0	-9

Mother ID, default = 0

.fam File

- **The .fam file holds data on the individuals contained in the data set. Each line represents 1 individual and has six columns(tab delimiter)**

NA21089	NA21089	0	0	0	-9
NA21090	NA21090	0	0	0	-9
NA21091	NA21091	0	0	0	-9
NA21092	NA21092	0	0	0	-9
NA21093	NA21093	0	0	0	-9
NA21094	NA21094	0	0	0	-9

Sex, male = 1, female = 2, default = 0

.fam File

- **The .fam file holds data on the individuals contained in the data set. Each line represents 1 individual and has six columns(tab delimiter)**

NA21089	NA21089	0	0	0	-9
NA21090	NA21090	0	0	0	-9
NA21091	NA21091	0	0	0	-9
NA21092	NA21092	0	0	0	-9
NA21093	NA21093	0	0	0	-9
NA21094	NA21094	0	0	0	-9

Phenotype, control = 1, case = 2, default = -9

.bedp File

- **Binary with little-endian format, meaning the bits are read right to left within the byte.**
- **The first 3 bytes contain the header:** *1101100 00011011 00000001*
- **The remaining bytes in the file hold the phased haplotype data**

.bedp File

- **Phased Haplotype data held in 2 bits:**
  - 00 = homozygous for allele 1 (minor)
    - can represent null value as well
  - 01 = heterozygous, allele 1 is in the 1st position
  - 10 = heterozygous, allele 2 is in the 1st position
  - 11 = homozygous for allele 2 (major)

.bedp File

- **The file is a sequence of X blocks of N/4(rounded up) bytes, where X is the number of SNPs and N is the number of Individuals.**
  - ex. A data set of 14,207 SNPs and 300,013 individuals
    - 14,207(300,013/4) = 14,207(75,004) = 1,065,581,828
    - File size = 1,065,581,828 + 3 *(header)* = ~1.07GB

.bedp File

- The low-order(1st) bits of a block’s first byte stores the first individual’s haplotype data.
- The next two bits store the second individual's haplotype data, and so on for the 3rd and 4th individual.
- The second byte stores haplotype data for the 5th-8th samples, the third byte stores codes for the 9th-12th, etc.
- If N is not divisible by four, the extra high-order bits in the last byte of each block are always zero.

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

The 1st individual's 1st haplotype is 11, homozygous, allele 2 (major)

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

The 2nd individual's 1st haplotype is "10", heterozygous, allele 2 in 1st position

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

The 5th individual's 1st haplotype is "10", heterozygous, allele 2 in 1st position

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

The 6th individual's 1st haplotype is "11", homozygous, allele 2 (major)

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

Null Value

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

The 1st individual's 4th haplotype is "10", heterozygous, allele 2 in 1st position

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

The 2nd individual's 4th haplotype is 11, homozygous, allele 2 (major)

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

The 5th individual's 4th haplotype is 11, homozygous, allele 2 (major)

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

The 6th individual's 4th haplotype is 10, heterozygous, allele 2 in 1st position

.bedp File

Example: 4 SNPs and 6 individuals *(4(6/4) = 4(2) = 8 bytes)*

(WITHOUT HEADER)

1st byte	2nd	3rd	4th	5th	6th	7th	8th
10101011	00001110	11111110	00001111	01010111	00000101	11111110	00001011

Null Value

Converting to bedp

- **Using C++, a command line executable was created to convert files in the vcf.gz format into bedp**
- **vcf.gz requirements:**
  - Must be phased (haplotype separator = “|”, not “/”)
  - No missing haplotype values (there is no option in bedp)
  - Only the “GT” SNP data is converted

GitHub Repository: https://github.com/Ryan-J-Lewis/VCFtoBEDP

Using bedp with PLINK

- **PLINK is an open source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner. http://zzz.bwh.harvard.edu/plink/**
- **Currently only the --snp, --keep, and --remove, options have been tested and confirmed to work.**
- **To maintain data integrity the option,**

  "--keep-allele-order" must be used

Using bedp with PLINK

- **Examples:**

--snp

```
ryan@laptop:~/plink$ ./plink --bfile input --snp rs587697622 --keep-allele-order --make-bed --out output
```

This command will pull only the snp “rs587697622” for all individuals and store it in the files “output.bed”, “output.bim”, and “output.fam”

--keep

```
ryan@laptop:~/plink$ ./plink --bfile input --keep HG00096 --keep-allele-order --make-bed --out output
```

This command will pull all of the SNPs for the Individual “HG00096” and store it in the files “output.bed”, “output.bim”, and “output.fam”

--remove

```
ryan@laptop:~/plink$ ./plink --bfile input --remove HG00096 --keep-allele-order --make-bed --out output
```

This command will pull all individuals except “HG00096” and store it in the files “output.bed”, “output.bim”, and “output.fam”

Question?

ryan.j.lewis@uth.tmc.edu