

GERBIL 1.0 Instruction Manual

BY GAD KIMMEL AND RON SHAMIR

1 Introduction

GERBIL is a software for simultaneously phasing genotypes into haplotypes and block partitioning. The software is written in C++, and can be used under both WINDOWS and LINUX platforms. The software is based on an algorithm described in [1].

2 Input Files

There is one input file: genotypes file. Each line represents one genotype. Each genotype is represented by pairs of letters, which correspond to SNPs: an unordered pair of allele readings, one from each copy of the chromosome. Missing data entries are represented by '?'. Pairs of SNPs are separated by a space or a tab. The first two lines are the number of rows (genotypes) and columns (SNPs). Here is an example of a genotypes file:

```
rows: 3
columns: 4
AG      TT      GG      CG
AG      ??      GC      CC
AA      CC      CC      CC
```

The above genotypes file contains 3 genotypes with 4 SNPs. The first SNP in the first genotype is $\{A, G\}$ (a heterozygote SNP), the second SNP is $\{T, T\}$ (a homozygote SNP), etc. In each column (which corresponds to a site) only two possible nucleotides are allowed.

3 Output Files

GERBIL has five output files:

1. 012 matrix file: a translation of the genotypes file into a 012 matrix format. Each component in the matrix corresponds to a genotype SNP (i.e., a pair of nucleotides). '0' stands for a homozygote SNP in a specific column, '1' stands for the other homozygote SNP in this column, '2' stands for a heterozygote SNP, and '?' stands for a missing entry allele. For example, the above genotype file is translated into the following 012 matrix:

```
2002
2?20
0110
```

2. Blocks file: Each line represents one block. The line is composed of two integer numbers, separated by a tab. The numbers correspond to starting and ending SNP of each block. Here is an example of a blocks file:

```
1      14
15     21
22     26
27     32
```

The above blocks file corresponds to a genotypes file with 32 SNPs. The file describes 4 blocks. The first block starts at SNP 1 and ends at SNP 14, the second block starts at SNP 15 and ends at SNP 21, etc.

3. Information file: contains the optimized parameters of the model, given the data. A detailed description of the model and parameters appears in [1]. The file contains for each block
 - The common haplotypes.
 - Their frequencies.
 - The probability of getting 1 in each SNP in the common haplotype (see [1]).
4. Resolved haplotypes file: The two consecutive lines $2i - 1$ and $2i$ represent the two resolved haplotypes of genotype i . For example, the following resolved haplotypes file is one possible solution for the above genotype file:

```
ATGG
GTGC
A?CC
G?GC
ACCC
ACCC
```

5. Resolved haplotype composition file: describes the common haplotype composition of each resolved haplotype. The two lines after the number i : represent the resolved haplotypes of genotype i . The j -th number in each line corresponds to the index of the common haplotype in block j . The translation of these indices into common haplotypes is as described in the parameters file. For example, the following file:

```
1:
1 4 3
2 1 4
2:
1 4 2
5 1 3
3:
5 2 2
1 2 4
```

describes 3 pairs of haplotypes. In the first haplotype of genotype 1, the common haplotype in the first block is 1, the common haplotype in the second block is 4, and in the third block it is 3.

4 Running GERBIL

In both WINDOWS and LINUX the usage is the same:

```
gerbil <genotypes filename>
```

or:

```
gerbil <genotypes filename> -o <output filename>
```

For example:

```
gerbil genotypes.txt
```

5 Questions and Bugs Reports

For questions and bugs reports e-mail: kgad@tau.ac.il.

References

- [1] G. Kimmel and R. Shamir. GERBIL: Genotype Resolution and Block Identification using Likelihood. Manuscript, School of Computer Science, Tel Aviv University, May 2004.