Table of Contents

1. Overview
2. Single Marker Mode
    - Format of input and parameter files
    - Usage
    - Output file format
3. Multiple Marker Mode
    - Algorithm
    - Format of input and parameter files
    - Usage
    - Output file format

For questions and comments, please contact
Dr. Ke Hao (ke_hao@affymetrix.com) or
Dr. Simon Cawley (simon_cawley@affymetrix.com)

1. Overview

(1) Specific aims: Consider two collections of SNPs, denoted as P_1 and P_2, and we genotype P_1 and P_2 on the same cohort. The aim of this program (**ld_compare**) is to calculate single- and multiple-marker coverage of one SNP panel (e.g., P_1) on the other (e.g., P_2).

***(2)* Mode:** The program runs on either (1) single marker coverage mode, which accommodate diploid data or (2) multiple marker coverage mode, which accommodate phased haplotype data. The program automatic detects the run mode according to the format of command-line arguments and parameter file.

(3) System requirement: This program runs on Windows and Linux OS. The file "ld_compare.exe" is pre-compiled binary on Windows XP using Microsoft Visual Studio 2005. The file "ld_compare" is pre-compiled binary on Linux version 2.4.20-43.9.legacysmp using gcc version 3.2.

2. Single Marker Coverage Mode

- **Parameter File**

File "par.txt" serves as an example parameter file.
Line 1, the range (bp) to search upstream or downstream for predictor SNPs
Line 2~3, minor allele cutoff for SNP panels P_1 and P_2
Line 4, switch of screen output, recommended to set as 1
Line 5, switch of skipself, recommended to set as 0
Line 6, switch of freepass, recommended to set as 1
Line 7, switch of outputting all pairwise r^2
Line 8~9, path and name base for result file and empirical CDF file
Line 10, number of chromosomes each SNP panel contains
Line 11~12, name of the chromosomes
Line 13~14, path and name ped files (each line specifies one chromosome)
Line 15~16, path and name info files (each line specifies one chromosome)

- **Info File**

File "P1.ChrA.info" serves as an example info file. It's in standard linkage format.
Column1, SNP ID
Column2, chromosomal position, ascending sorted

- **Ped File**

File "P1.ChrA.ped" serves as an example info file. It's in standard linkage format.
Column1, pedigree number
Column2, individual identification number, or id
Column3, father's id number
Column4, mother's id number
Column5, sex
Column6, Proband Status
Subsequent columns, genotype data. Each SNP occupies two columns.

- **Usage**

The single-marker mode accommodates only one command-line argument which is the pass and name of the parameter file.
For example, on Windows OS, "ld_compare.exe par.txt"

- **Output File**

File "ChrB_A.chrA.lst.txt" is an example of the r^2 coverage output. For each SNP in the P_1 panel, the file lists its best coverage. The file contains five columns, (1) SnpID of P_1 SNP, (2) position of P_1 SNP, (3) SnpID of P_2 SNP, (4) position of P_2 SNP, and (5) r^2. When there is no P_2 SNP in the sliding window range, the 3rd column is assigned "0".

File "*ecdf.txt" list the combined or chromosome-specific empirical CDF (ECDF).

When Line 7 in the parameter file is set as 1, the program will output the intermediate results (all pairwise r^2) in *full.table.txt.

3. Multiple Marker Mode

- **Algorithm of Two Marker Coverage**

The multiple marker coverage calculation requires phased haplotype data. Consider three SNPs A, B and C. Each SNP carries two possible allele, denoted as A and a, B and b, and C and c, respectively. We are interested in the coverage of C by A and B.

The first step is to compute the linkage disequilibrium (LD) in term of r^2 between SNP C and SNPs A and B. SNPs A and B may form four possible haplotypes (AB, Ab, aB and ab). Therefore, A and B together can be treated as a multi-allelic marker, which carries four alleles, denoted as AB, Ab, aB and ab. Pooling {Ab, aB and ab}, we can transform this multi-allelic marker to a bi-allelic SNP, which carries alleles AB and nonAB. Easily we compute the r^2 between this new bi-allelic SNP and SNP C, and record the result as r^2_{AB}. Similarly, we can calculate r^2_{Ab} by pooling {AB, aB and ab}. Same in r^2_{aB} and r^2_{ab}. Furthermore, we compute the r^2 between SNP A and SNP C, recorded as r^2_A, as well as r^2 between SNP B and SNP C, recorded as r^2_B. Herein, we define r^2 between SNP C and SNPs A and B is simply max{ r^2_A , r^2_B, r^2_{AB} , r^2_{Ab} , r^2_{aB} and r^2_{ab} }.

The second step is to compute coverage, which requires a pre-specified threshold (r^2_{cutoff}). We defined SNP C is covered by SNP A and B if max{ r^2_A , r^2_B, r^2_{AB} , r^2_{Ab} , r^2_{aB} and r^2_{ab} } $\geq r^2_{cutoff}$. **Please note** this result is the maximum of single- and two-marker coverage. By modifying the parameter file, the program allows single marker only coverage or two marker only coverage calculation.

- **Algorithm of Three+ Marker Coverage**

There are four SNPs (A, B, C and D), and we are interested in the coverage of SNP D by SNPs A, B and C. There are 2^3=8 possible haplotypes. Again, we construct a novel bi-allelic SNP by pooling 7 haplotype together, and we obtain the r^2 after 8 iterations.

The coverage of four or more marker can be computed in the same framework, but has not been implemented at the current stage.

- **Parameter File**

File "hap_Chr24.par" serves as an example parameter file.
Line 1, the range (bp) to search upstream or downstream for predictor SNPs
Line 2, minor allele cutoff
Line 3, switch of screen output, recommended to set as 1
Line 4~6, switches of single-, two- and three- coverage, 1 = Enable
Line 7, indicator that the two SNP panels are stored jointly in hap file
Line 8, format of ped file, Broad or Oxford
Line 9~11, path and name of output file, hap file, and info file

- **Info File**

File “chr24_info.txt” serves as an example info file. It’s a tab delimited file contains four columns.
Column1, SNP ID
Column2, chromosomal position, ascending sorted
Column3, indicator that this SNP is a target SNP, whose coverage by predictor SNPs need to be computed (1=Yes)
Column4, indicator that this SNP is a predictor which is genotyped (1=Yes)

- **Haps File in “Broad” Format**

File “chr24_haps.txt” serves as an example. It’s a tab delimited file, and each line represent one chromosome. In each line, the first field stores the sample ID, the second field store the chromosome name, and the following fields are haplotype of SNPs, in the same order as in the info file. Missing data is denoted as “0”. Ambiguous haplotype is denoted as “h”, and is treated as missing data at the current stage.

- **Usage**

The multi-marker mode accommodates only one command-line argument which is the pass and name of the parameter file.
For example, on Windows OS, “ld_compare.exe hap_Chr24.par”

- **Output File**

File “chr24_TwoMarker_Rsq.ref.txt” is an output example of three marker coverage. Each line presents the results of one target SNP.