AFFYMETRIX

The Way Ahead.™

LdCompare: rapid computation of single- and multiple-marker r^2 and genetic coverage

Ke Hao
Algorithms and Data Analysis
Affymetrix, Inc.

AFFYMETRIX
The Way Ahead.™

Overview

- Single Marker Mode
  - Format of input and parameter files
  - Output file format
- Multiple Marker Mode
  - Algorithm
  - Format of input and parameter files
  - Output file format
- Usage and Examples

AFFYMETRIX
The Way Ahead.™

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.
- LdCompare calculates single- and multiple-marker coverage of one SNP panel (e.g., P_1) on the other (e.g., P_2).
- LdCompare outputs pairwise r^2 for downstream analysis (e.g., tag SNP selection).

AFFYMETRIX
The Way Ahead.™

Running Modes

- 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.

AFFYMETRIX
The Way Ahead.™

Parameter File (Single Marker Mode)

- 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)

AFFYMETRIX

The Way Ahead.™

Parameter File Example (Single Marker Mode)

```
window=25000
maf1=0.05
maf2=0.05
verbose=1
skipself=1
freepass=1
AllPairRsq=1
ResultFileBase=../examples/Single_Marker/ChrB_A
ECDFFileBase=../examples/Single_Marker/ChrB_A
number=2
chrB
chrA
../examples/Single_Marker/P1.ChrB.ped      ../examples/Single_Marker/P2.ChrB.ped
../examples/Single_Marker/P1.ChrA.ped      ../examples/Single_Marker/P2.ChrA.ped
../examples/Single_Marker/P1.ChrB.info     ../examples/Single_Marker/P2.ChrB.info
../examples/Single_Marker/P1.ChrA.info     ../examples/Single_Marker/P2.ChrA.info
```

AFFYMETRIX
The Way Ahead.™

Linkage Format Genotype File Example

- **Info File**
  - Column1, SNP ID
  - Column2, chromosomal position, ascending sorted
- **Ped File**
  - Column1, pedigree number
  - Column2, individual identification number, or id
  - Column3, father's id number
  - Column4, mother's id number
  - Column5, gender
  - Column6, proband status
  - Subsequent columns, genotype data. Each SNP occupies two columns.

AFFYMETRIX
The Way Ahead.™

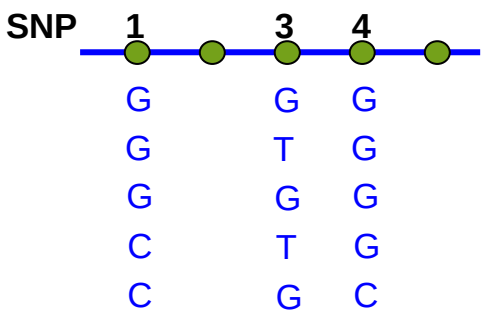
Output (Single Maker Mode)

- ***r^2 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.
- 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.

AFFYMETRIX
The Way Ahead.™

Algorithm for 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.



❖ Consider the haplotype of SNP#1 and SNP#3 as a novel multi-allelic marker, denoted as M

❖ Pooling all but one alleles, M is converted to a bi-allelic SNP. LD (r^2) between M and SNP#4 can be computed

❖Survey all possible pooling and lower order schemes , the largest r^2 achieve is defined as the multiple-marker r^2

AFFYMETRIX
The Way Ahead.™

Algorithm for Two Marker Coverage

- Compute the LD (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.
- 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} }.

AFFYMETRIX
The Way Ahead.™

Algorithm for 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.
- 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.
- max{ $r^2_{\text{three-marker}}$, $r^2_{\text{two-marker}}$, r^2_A , r^2_B , and r^2_C}
- The coverage of four or more marker can be computed in the same framework, but has not been implemented at the current stage.

AFFYMETRIX
The Way Ahead.™

Parameter File (Multi Marker Mode)

- 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 be 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

AFFYMETRIX
The Way Ahead.™

Parameter File (Multi Marker Mode)

```
window=100000
maf=0.05
verbose=1
OneMarker=1
TwoMarker=1
ThreeMarker=0
TwoPanelCoverage=1
format=Broad
ResultFileBase=../examples/Multiple_Marker/chr24_TwoMarker_Rsq.txt
../examples/Multiple_Marker/chr24_haps.txt
../examples/Multiple_Marker/chr24_info.txt
```

AFFYMETRIX
The Way Ahead.™

Haplotype File

- **Info File (4 Columns)**
  - Column#1, SNP ID
  - Column#2, chromosomal position, ascending sorted
  - Column#3, indicator that this SNP is a target SNP, whose coverage by predictor SNPs need to be computed (1=Yes)
  - Column#4, indicator that this SNP is a predictor which is genotyped (1=Yes)
- **Haps File in "Broad" Format**
  - 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.

AFFYMETRIX
The Way Ahead.™

Usage

- Makefile
  - make
  - make check
  - make clean
- Command line mode only
- The single-marker mode accommodates only one command-line argument which is the pass and name of the parameter file.
- For example, on Windows,
- `ld_compare.exe ..\..\\examples\Single_Marker\workshop.txt`
- `ld_compare.exe ..\..\\examples\Multiple_Marker\workshop1.par`
- `ld_compare.exe ..\..\\examples\Multiple_Marker\workshop2.par`
- `ld_compare.exe ..\..\\examples\Multiple_Marker\workshop3.par`

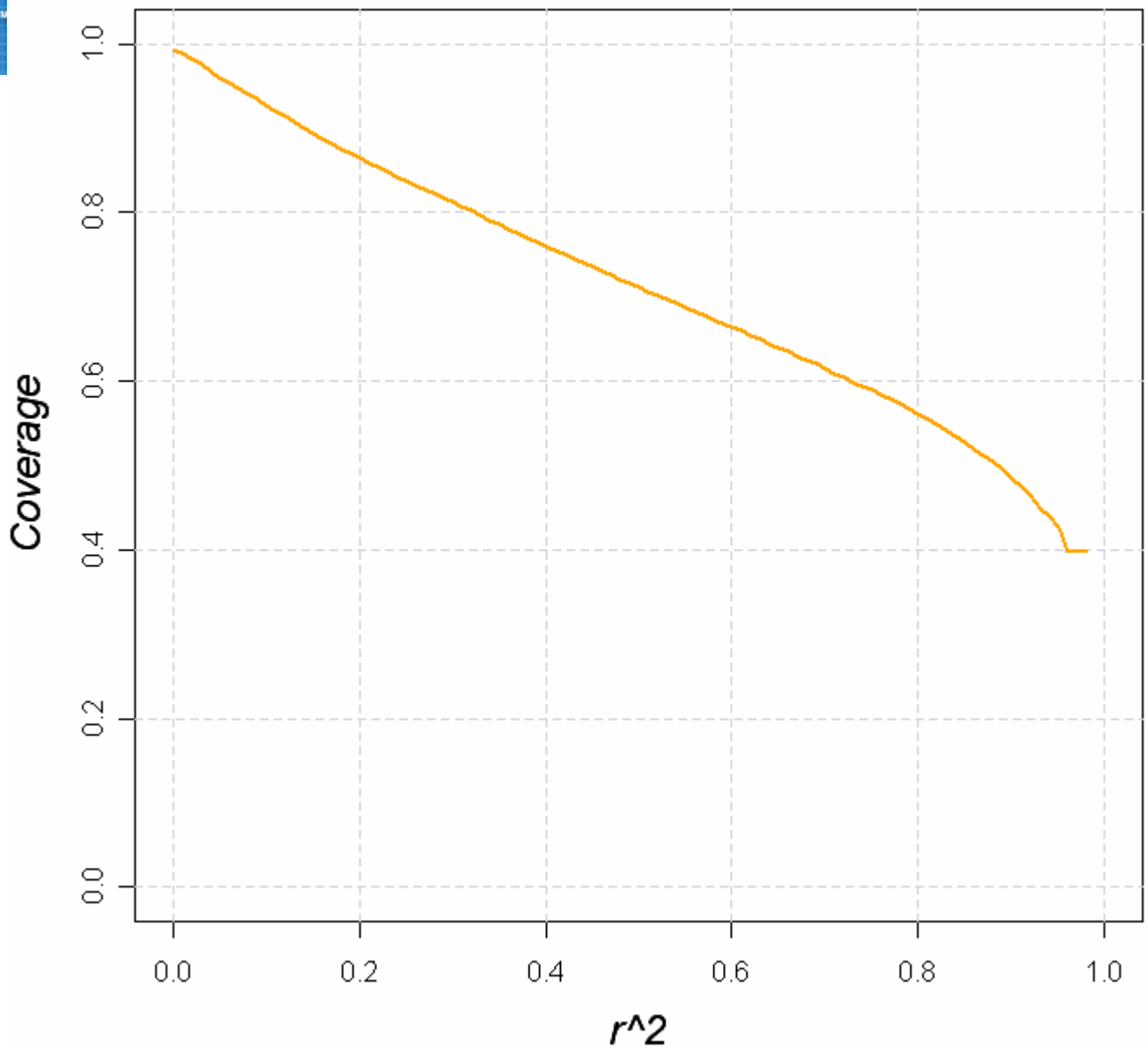
AFFYMETRIX
The Way Ahead.™

R code for plotting coverage curve

```
ylab<-"Coverage"
setPlot <- function(header="")
{
  plot(c(0,1),c(0,1),type="n",xlab="r^2",ylab=ylab,main=header,cex.lab=1.5,font.lab=3)
  xaxp <- par("xaxp")
  abline(v=seq(xaxp[1],xaxp[2],length=xaxp[3]+1),lty=2,col="lightgray")
  yaxp <- par("yaxp")
  abline(h=seq(yaxp[1],yaxp[2],length=yaxp[3]+1),lty=2,col="lightgray")
}

plotLines <- function(r2,coverage,col="black",lty=1,lwd=2)
{
  lines(r2,coverage,col=col,lwd=lwd,lty=lty)
}

myCol <- c("orange","darkgreen","purple","pink","grey")
setPlot("")
title(main = list("ChrA&B Coverage", cex=1.7,col="black", font=2))
phe<-read.table("ChrB_A.total.ecdf.txt",header=T,sep="\t")
plotLines(phe[1:99,1],phe[1:99,4],col=myCol[1])
```


ChrA&B Coverage
Coverage
0.0
0.2
0.4
0.6
0.8
1.0
r^2
0.0
0.2
0.4
0.6
0.8
1.0

AFFYMETRIX
The Way Ahead.™

R code for plotting coverage curve

```
setPlot("")
title(main = list("Multiple Marker Coverage Example", cex=1.7,col="black", font=2))
phe1<-read.table("chr24_OneMarker_Rsq.txt",header=T, sep="\t")
phe2<-read.table("chr24_TwoMarker_Rsq.txt",header=T, sep="\t")
phe3<-read.table("chr24_ThreeMarker_Rsq.txt",header=T, sep="\t")

results<-NULL
for (i in seq(0, 0.99, .01))
{
          cdf1<-mean(phe1[,7]>=i)
          cdf2<-mean(phe2[,7]>=i)
          cdf3<-mean(phe3[,9]>=i)
          results<-rbind(results,c(i,cdf1,cdf2,cdf3))
}

plotLines(results[,1],results[,2],col=myCol[1],lty=1)
plotLines(results[,1],results[,3],col=myCol[1],lty=2)
plotLines(results[,1],results[,4],col=myCol[1],lty=3)
```

Multiple Marker Coverage Example