Online Suppplement

Table 1: Benchmarks of large scale SNP data storage and analysis: Processing times for a typical run on a 2 GHz Pentium CPU.

Table 2: Probability for departure from Hardy-Weinberg equilibrium and multiple copy sequence occurrence of 490,032 SNPs (X-chromosomal-SNPs were excluded) tested in 210 individuals (60 children of of CEU and YRI trio families were excluded).

Table 3: Probability for departure from Hardy-Weinberg equilibrium and missing genotypes in 490,032 SNPs (X-chromosomal-SNPs were excluded) tested in 210 individuals (60 children of CEU and YRI trio families were excluded).

Table 4: LD block number and size in 270 individuals tested for 500,568 SNPs by ethnic origin.

Figure 1: Genomewide distribution of allele frequencies in 270 individuals tested for 500,568 SNPs.

Figure 2: Deviation from Hardy-Weinberg equilibrium in 60 European hapmap parents (CEU) tested for 490,032 SNPs (X-chromosomal-SNPs were excluded). The threshold for large circles is $p<1e-5$.

Figure 3: Distribution of Hardy-Weinberg equilibrium of 490,032 SNPs (X-chromosomal-SNPs were excluded) by ethnic origin.

Supplementary methods

Table 1: Benchmarks of large scale SNP data storage and analysis: Processing times for a typical run on a 2 GHz Pentium CPU.

Software	Task	N-column method	3-column method
Perl	StyHapMap and NspHapMap data conversion	0:01:04	0:09:49
SQLite	SQLite import, creating index on ProbeSetID	0:01:50	1:21:07
R	500,000 SNP allele count: SQLite format	1:52:40	18:30:58
R	as above; using R data files	0:51:00	na*

*not available

Table 2: Probability for departure from Hardy-Weinberg equilibrium and multi copy sequence occurrence of 490,032 SNPs (X-chromosomal-SNPs were excluded) tested in 210 individuals (60 children of of CEU and YRI trio families excluded).

Population	grouped p-value for departure from HWE	unique sequence, N SNPs		multiple sequence occurrence, N SNP	
CEU	[0, 1e-05]	247	0,1%	67	1,2%
	(1e-05, 0.05]	17171	3,5%	564	10,3%
	(0.05, 1]	467158	96,4%	4825	88,4%
YRI	[0, 1e-05]	187	0,0%	44	0,8%
	(1e-05, 0.05]	20427	4,2%	577	10,6%
	(0.05, 1]	463962	95,7%	4835	88,6%
CHB	[0, 1e-05]	54	0,0%	5	0,1%
	(1e-05, 0.05]	15726	3,2%	438	8,0%
	(0.05, 1]	468796	96,7%	5013	91,9%
JPT	[0, 1e-05]	43	0,0%	9	0,2%
	(1e-05, 0.05]	14579	3,0%	441	8,1%
	(0.05, 1]	469954	97,0%	5006	91,8%

Table 3: Probability for departure from Hardy-Weinberg equilibrium and missing genotypes in 490,032 SNPs (X-chromosomal-SNPs were excluded) tested in 210 individuals (60 children of CEU and YRI trio families were excluded).

Population	p-value group, departure from HWE	missing genotypes [0,5%]		(5%,10%]		(10%,15%]		(15%,20%]	
CEU	[0, 1e-05]	196	0,0%	49	0,1%	27	0,2%	42	0,5%
	(1e-05, 0.05]	11192	2,7%	3409	6,6%	1589	10,6%	1545	17,5%
	(0.05, 1]	403325	97,3%	47984	93,3%	13408	89,2%	7266	82,1%
YRI	[0, 1e-05]	115	0,0%	40	0,1%	17	0,1%	59	0,5%
	(1e-05, 0.05]	11941	3,0%	4242	7,1%	2156	11,8%	2665	20,8%
	(0.05, 1]	387316	97,0%	55328	92,8%	16057	88,1%	10096	78,8%
CHB	[0, 1e-05]	18	0,0%	6	0,0%	8	0,0%	27	0,2%
	(1e-05, 0.05]	9930	2,4%	2381	5,7%	1645	8,5%	2208	14,4%
	(0.05, 1]	403637	97,6%	39279	94,3%	17812	91,5%	13081	85,4%
JPT	[0, 1e-05]	22	0,0%	12	0,0%	5	0,0%	13	0,1%
	(1e-05, 0.05]	10375	2,4%	1894	5,9%	1186	8,5%	1565	14,9%
	(0.05, 1]	422918	97,6%	30334	94,1%	12792	91,5%	8916	85,0%

Table 4: LD block number and size in 270 individuals tested for 500,568 SNPs by ethnic origin (for abbreviations see methods).

	CEU	YRI	CHB	JPT
LD blocks	47.355	47.631	40.685	41.075
SNP on LD block	269.006 (53,7%)	200.773 (40,1%)	218.860 (43,7%)	224.790 (44,9%)
mean SNP number per block	5,68	4,21	5,38	5,47

Figure 1: Genomewide distribution of allele frequencies in 270 individuals tested for 500,568 SNPs.

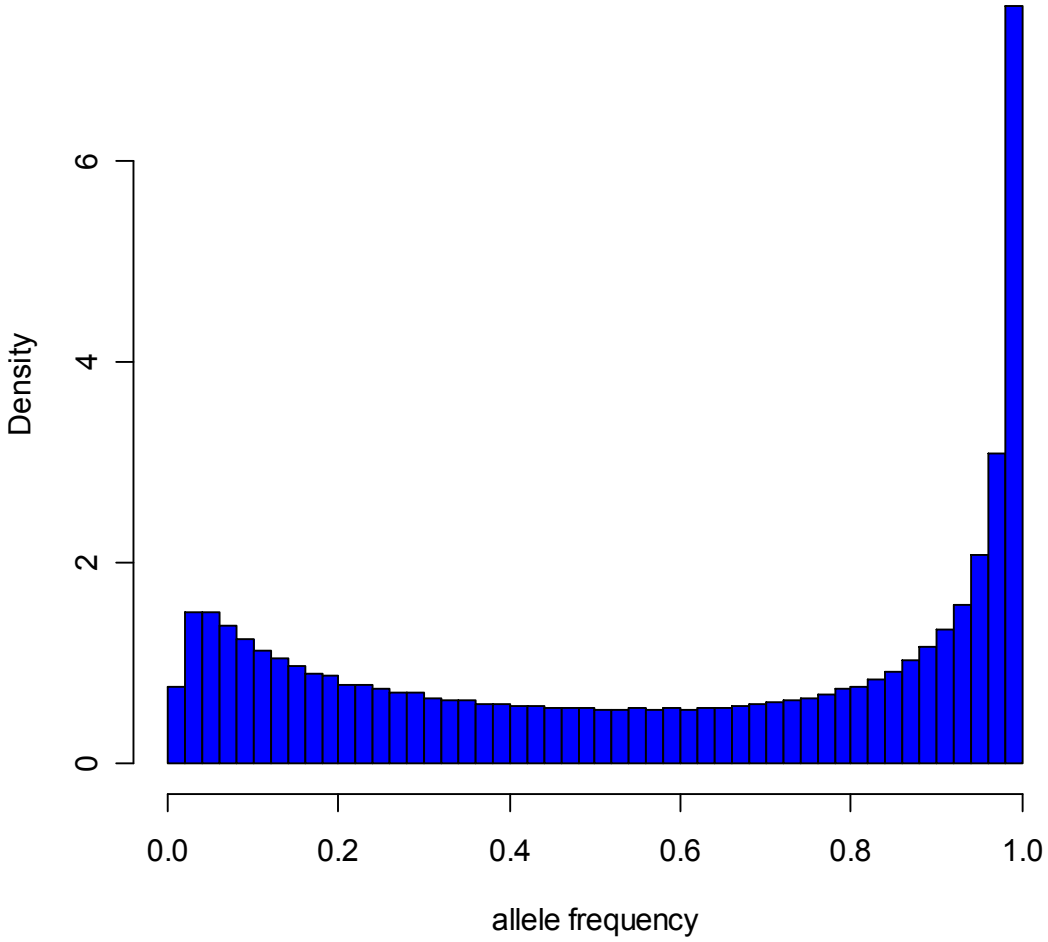


Figure 2: Deviation from Hardy-Weinberg equilibrium in 60 European hapmap parents (CEU) tested for 490,032 SNPs (X-chromosomal-SNPs were excluded). The threshold for large circles is p<1e-5.

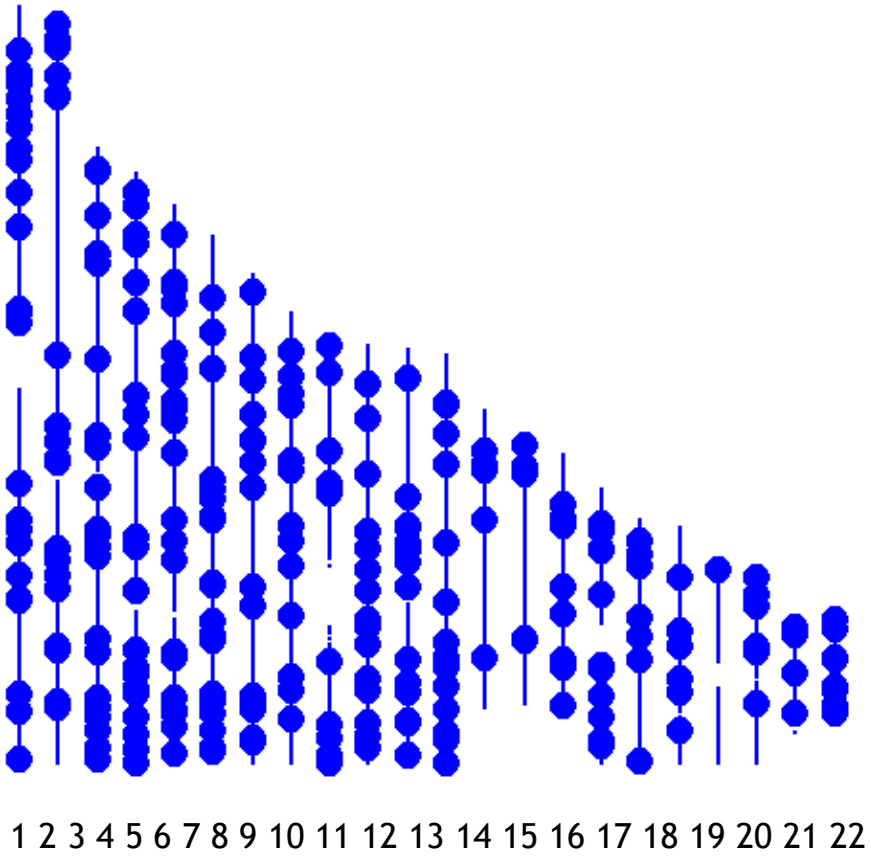
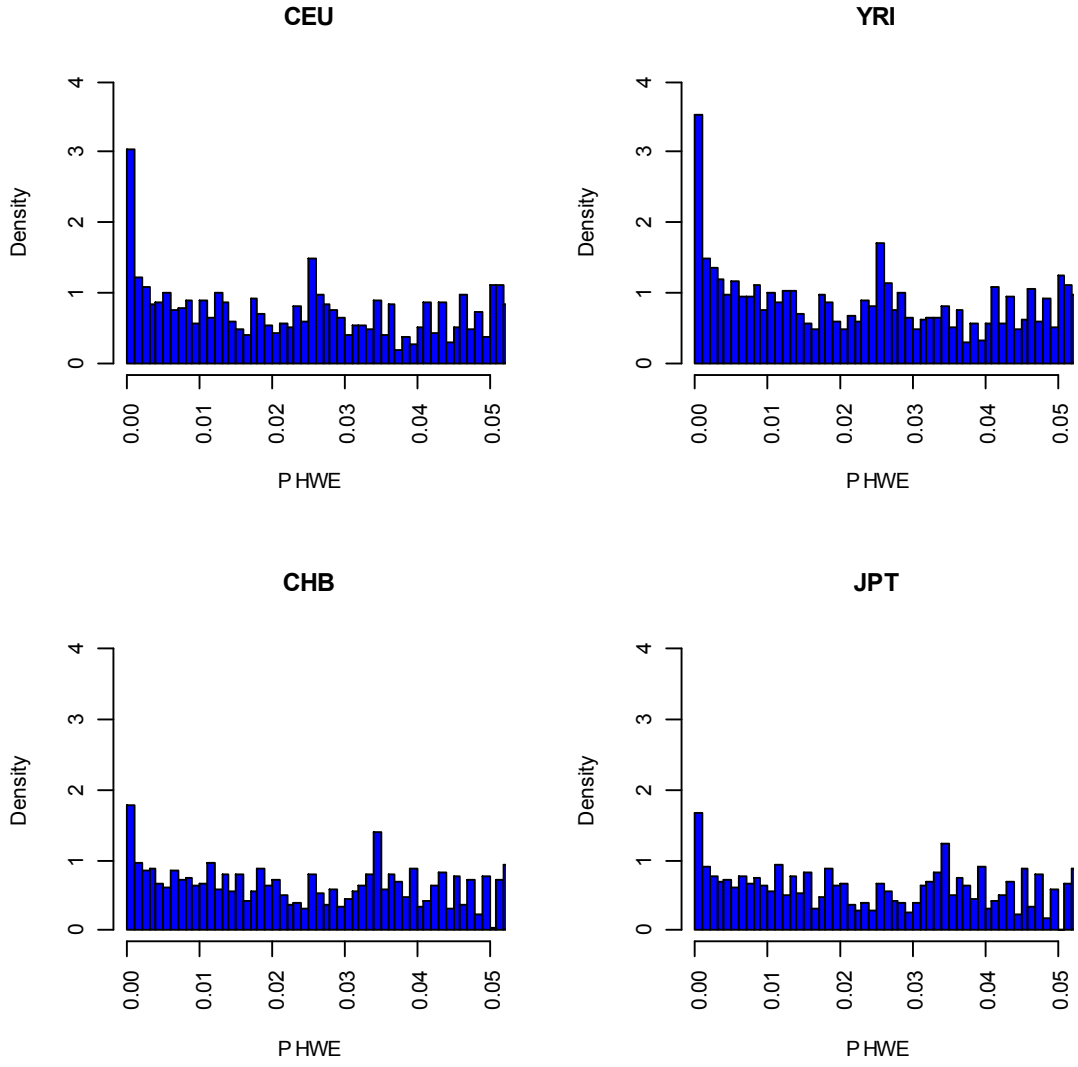


Figure 3: Distribution of Hardy-Weinberg equilibrium of 490,032 SNPs (X-chromosomal-SNPs were excluded) by ethnic origin (for abbreviations see methods).



Box 1: Perl script to convert genotyping data and create SQL file

```
#!/usr/local/bin/perl

open(OUTFILE, ">>E:\\affx.dat");
&convert("E:\\Nsp_HapMap270.txt");
&convert("E:\\Sty_HapMap270.txt");
close OUTFILE;
exit(0);

sub convert {
        $i=0;
        my ($file) = @_;
        open(INFILE,$file);
        while (<INFILE>) {
                if ($i == 0 & $j != 1) {
                        @header = split(/\t/,"id$_");
                        $h="";
                        for $ii ( 1 .. $#header ) {
                                $h="$h,$header[$ii] nvarchar(2)";
                        }
                        open(OUTSQL, ">E:\\affx.sql");
                        print OUTSQL <<"EOFMRK";
/*call sqlite3 affx.db < affx.sql*/
CREATE TABLE haplo270geno (ProbeSetID nvarchar20$h);
.separator "\\t"
.import "E:\\\\affx.dat" haplo270geno
CREATE INDEX idx ON haplo270geno (ProbeSetID);
.q
EOFMRK
                        close OUTSQL;
                        $j=1;
                }
                else {
                        s/\r//gi;
                        s/NN//gi;
                        print OUTFILE "$_";
                }
                $i++;
        }
        print "$file ...\n";
        close INFILE;
}
```

Box 2: R script to analyse data (non parallelized version)

```
library(RSQLite)
source(paste(dir,"hwe.R",sep=""))
mytable <- function(col2) { table(factor(col2,levels=c("AA","AB","BB"))) }

SQLite(max.con = 16, fetch.default.rec = 1000000, force.reload = TRUE)
m <- dbDriver("SQLite")
con <- dbConnect(m, dbname = "e:/affx_n-col.db")

S<- c("SELECT haplo270indiv.sampleLSID as indivID, haplo270indiv.*")
F<- c("FROM haplo270indiv")
rs <- dbSendQuery(con, paste(S,F,sep=" "))
db0 <- fetch(rs,n=-1)
db0[,"indivID"]<-as.factor(db0[,"indivID"])

S<- c("SELECT haplo270SNP.ProbeSetID")
F<- c("FROM haplo270SNP")
rs <- dbSendQuery(con, paste(S,F,sep=" "))
db1 <- fetch(rs,n=-1)

S<- c("SELECT *")
F<- c("FROM haplo270geno")

for (i in 0:50) {

      range <- (i*10000+1):((i+1)*10000)
      if (i==50) {range <- (500001):(500568)}
      W<- paste("WHERE haplo270geno.ProbeSetID IN
      ('",paste(db1[range,"ProbeSetID"],collapse="\',\'"),"');",sep="")
      rs <- dbSendQuery(con, paste(S,F,W,sep=" "))
      db2 <- fetch(rs,n=-1)
      db2[,271] <- gsub("\r","",db2[,271])
      db2 <- as.data.frame(t(db2))
      colnames(db2) <- as.character(as.matrix(db2[1,]))
      db2 <- db2[-1,c(db1[range,"ProbeSetID"])]
      db2[,"indivID"] <- rownames(db2)
      db2 <- merge(db0,db2,by="indivID")
      save(db2,file=paste("e:\\affx",formatC(i,width=2,flag="0"),".Rdata",sep=""))

      if (i!=50) {
            freqtable <- as.data.frame(matrix(nrow=10000,ncol=8,NA))
            for (ii in seq(0,9000,1000)) {
               selcol <- c(1:9,(10+ii):(10+ii+1000-1))
               tmpdb2 <- db2[,selcol]
               tmp <- t(apply(tmpdb2[,10:(10+1000-1)],2,mytable))
               freqtable[(ii+1):(ii+1000),1:4] <-
               cbind(names(tmpdb2)[10:(10+1000-1)],tmp)
            }
      }
      if (i==50) {
            freqtable <- as.data.frame(matrix(nrow=568,ncol=8,NA))
            tmpdb2 <- db2[,c(1:9,10:577)]
            tmp <- t(apply(tmpdb2[,10:577],2,mytable))
            freqtable[1:568,1:4] <- cbind(names(tmpdb2)[10:577],tmp)
      }
      write.table(freqtable,file="e:/freqtable.txt",sep="\t",quote=FALSE,na="",
      row.names=FALSE,col.names=FALSE,append=TRUE,dec=",")
}
```