eda_t2d-credible-kyle-metabo.Rmd

Jason Torres

February 2, 2017

```r
"%&%" <- function(a,b) paste0(a,b)
library("data.table")
library("dplyr")
```

```
## -------------------------------------------------------------------------

## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!

## -------------------------------------------------------------------------

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```r
serv.dir <- "/Users/jtorres/FUSE/"
cred.dir <- serv.dir %&% "reference/credible_sets/from_kyle/"
cred.file <- cred.dir %&% "metabochip.chr_added.vcf"

df <- fread(cred.file)
```

Evaluate and reformat data frame

```r
str(df)
```

```
## Classes 'data.table' and 'data.frame':   19266 obs. of  8 variables:
##  $ V1: chr  "chr1" "chr1" "chr1" "chr1" ...
##  $ V2: int  120437718 120437884 120438577 120439109 120440029 120441998 120442257 120443424 12044354
##  $ V3: chr  "rs2793823" "rs2641348" "rs147294252" "rs6668119" ...
##  $ V4: chr  "G" "A" "G" "G" ...
##  $ V5: chr  "A" "G" "A" "C" ...
##  $ V6: int  100 100 100 100 100 100 100 100 100 100 ...
##  $ V7: chr  "PASS" "PASS" "PASS" "PASS" ...
##  $ V8: chr  "LOCUS=NOTCH2;PROB=0.00871;" "LOCUS=NOTCH2;PROB=0.01154;" "LOCUS=NOTCH2;PROB=0.00048;" "
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
locus <- as.character(sapply(df$V8,function(string){
  gsub("LOCUS=","",strsplit(string,split=";")[[1]][1])
}))
prob <- as.character(sapply(df$V8,function(string){
  gsub("PROB=","",strsplit(string,split=";")[[1]][2])
}))
df <- select(df,one_of("V1","V2","V3","V4","V5"))
names(df) <- c("chr","pos","rsid","A1","A2")
df <- cbind(df,locus,prob)
df<- as.data.frame(df)
df$prob <- as.numeric(df$prob)
```

There are **49** loci in this file

Build locus summary data frame

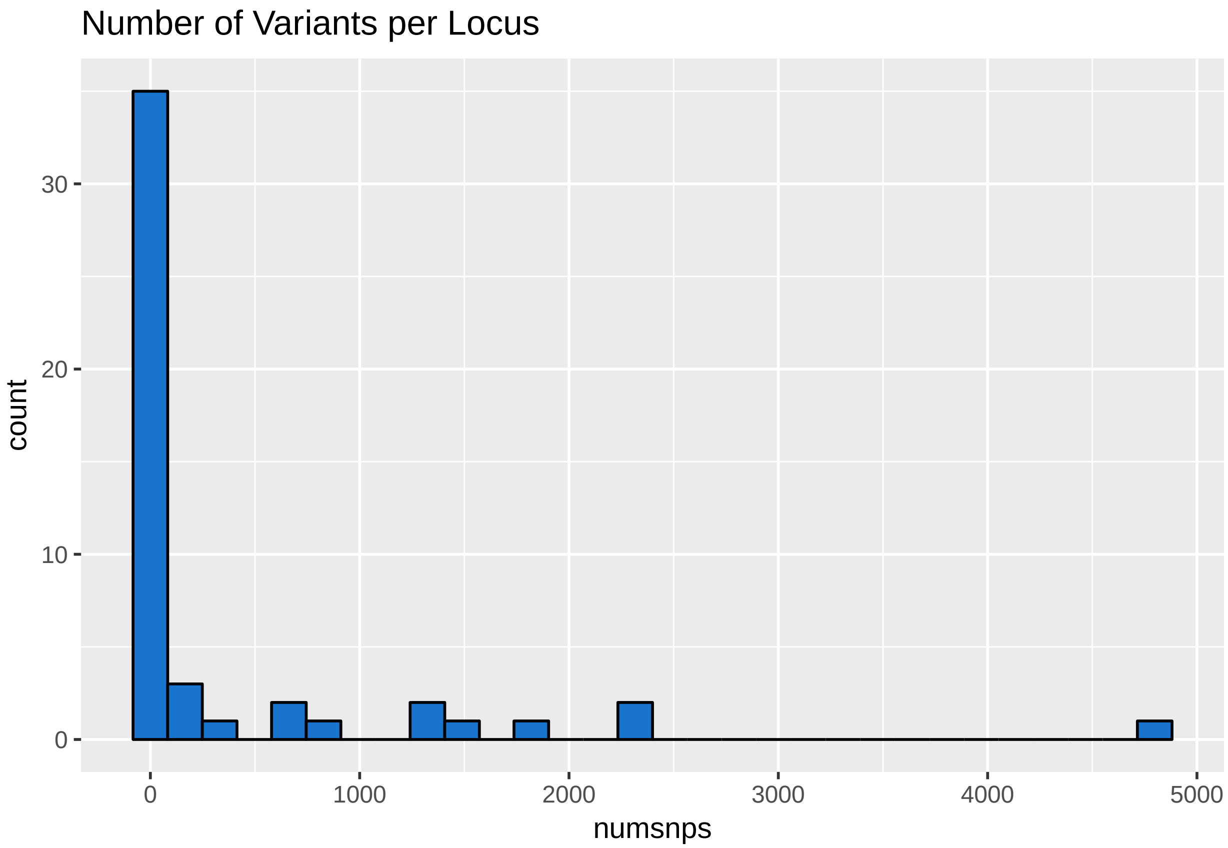
```r
loci <- unique(df$locus)
loc <- loci[1]
numsnps <- as.integer(sapply(loci, function(loc){
  length(filter(df,locus==loc)$prob)
}))
prop01 <- as.numeric(sapply(loci, function(loc){
  sum(filter(df,locus==loc)$prob > 0.01)/length(filter(df,locus==loc)$prob)
}))
prop05 <- as.numeric(sapply(loci, function(loc){
  sum(filter(df,locus==loc)$prob > 0.05)/length(filter(df,locus==loc)$prob)
}))
prop10 <- as.numeric(sapply(loci, function(loc){
  sum(filter(df,locus==loc)$prob > 0.10)/length(filter(df,locus==loc)$prob)
}))
prop20 <- as.numeric(sapply(loci, function(loc){
  sum(filter(df,locus==loc)$prob > 0.20)/length(filter(df,locus==loc)$prob)
}))

loc.df <- data.frame(loci,numsnps,prop01,prop05,prop10,prop20,
                     stringsAsFactors = FALSE)
```

Histograms summarizing loci distributions

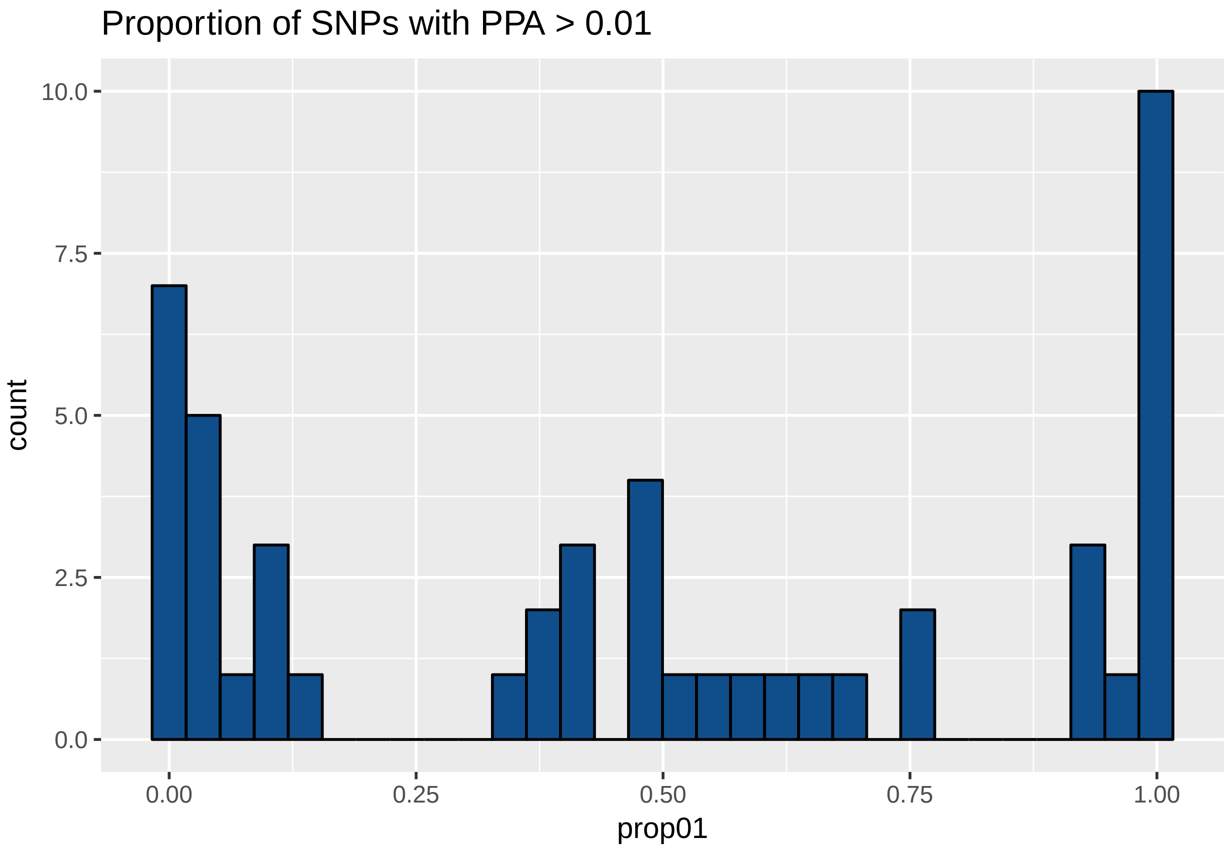
```r
plt1 <- ggplot(data=loc.df) +
  geom_histogram(aes(x=numsnps),color="black",
  fill="dodgerblue3") + ggtitle("Number of Variants per Locus");plt1
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



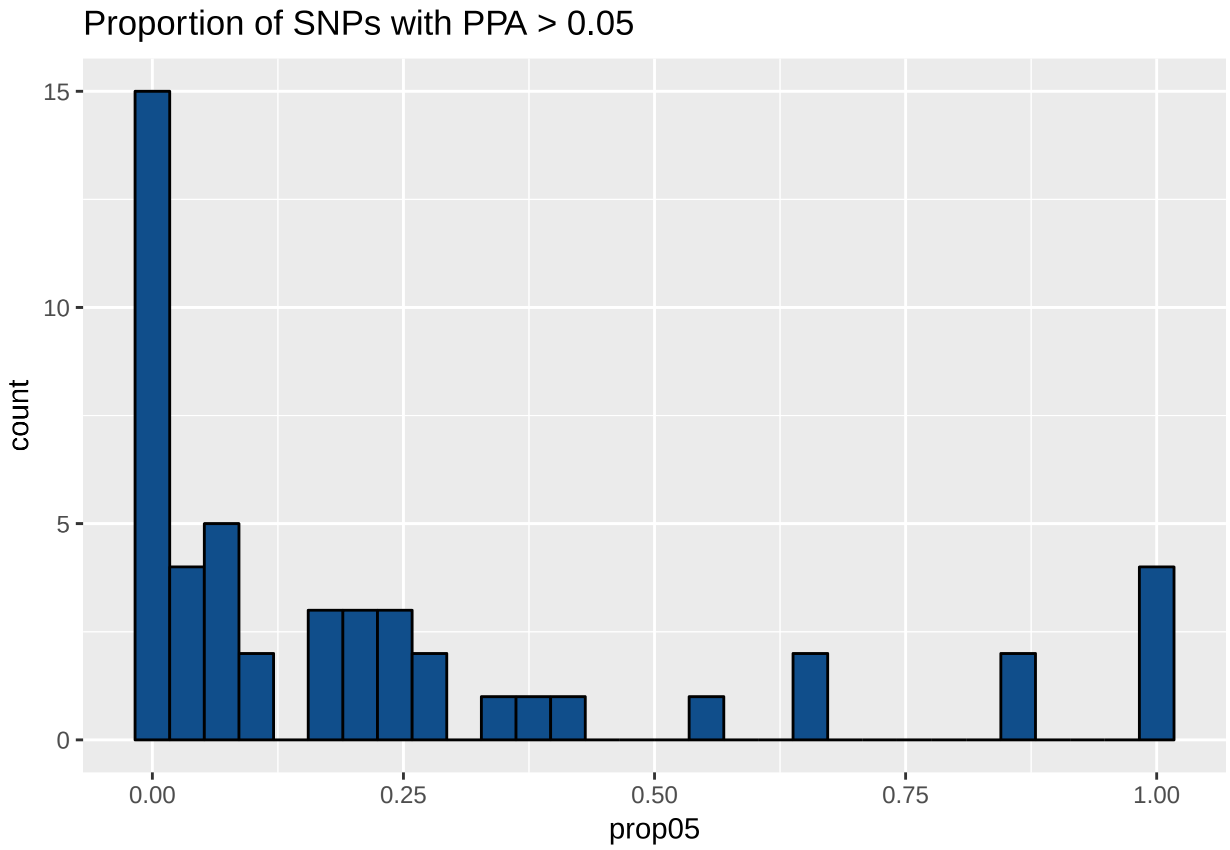
```r
plt2 <- ggplot(data=loc.df) +
  geom_histogram(aes(x=prop01),color="black",
  fill="dodgerblue4") + ggtitle("Proportion of SNPs with PPA > 0.01");plt2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



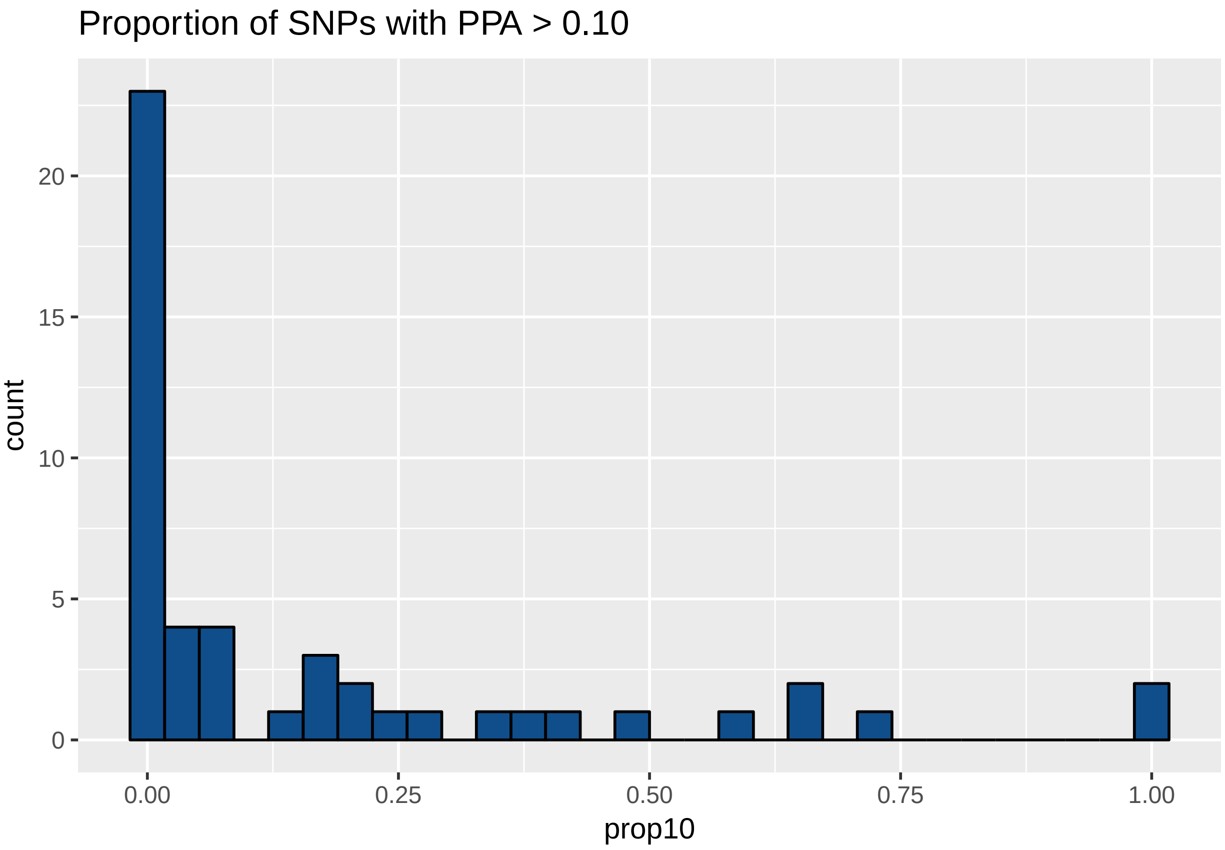
```
plt3 <- ggplot(data=loc.df) +
  geom_histogram(aes(x=prop05),color="black",
  fill="dodgerblue4") + ggtitle("Proportion of SNPs with PPA > 0.05");plt3
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



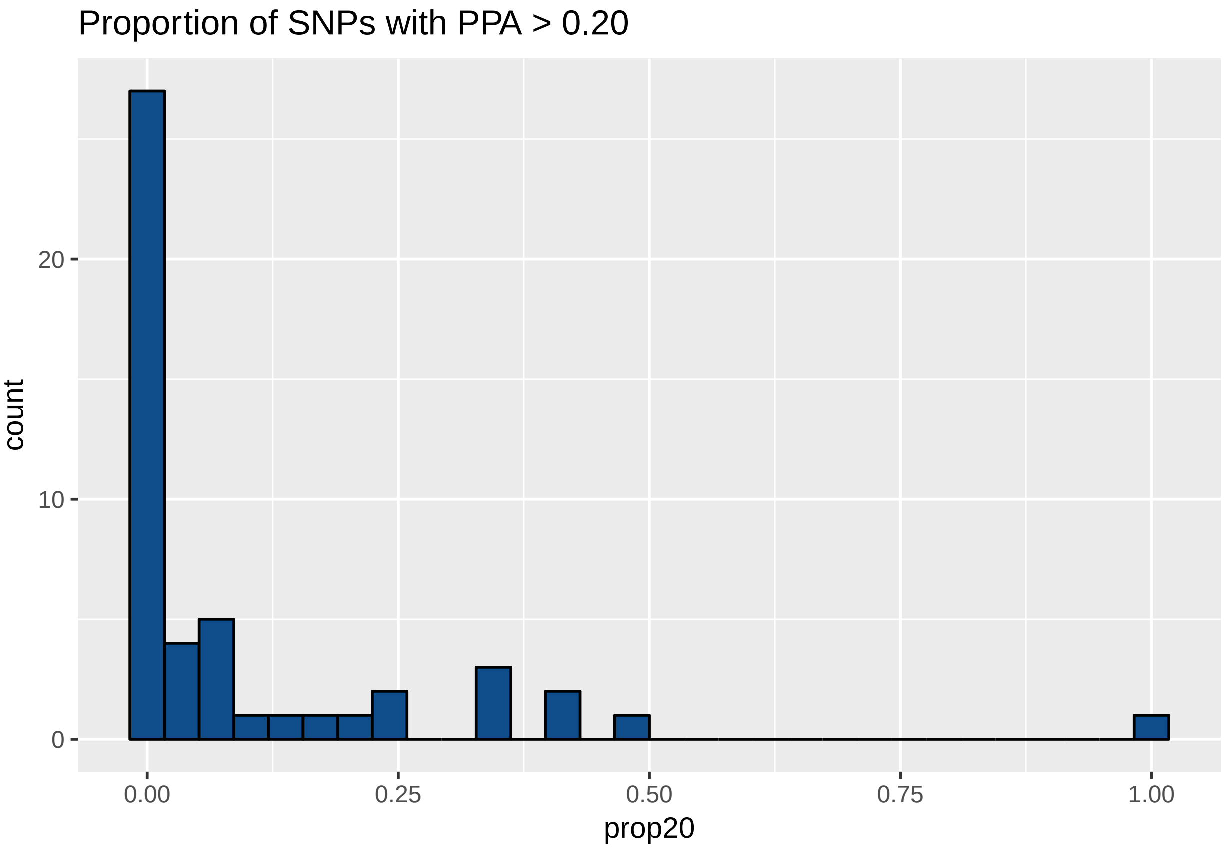
```
plt4 <- ggplot(data=loc.df) +
  geom_histogram(aes(x=prop10),color="black",
  fill="dodgerblue4") + ggtitle("Proportion of SNPs with PPA > 0.10");plt4
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
plt5 <- ggplot(data=loc.df) +
  geom_histogram(aes(x=prop20),color="black",
  fill="dodgerblue4") + ggtitle("Proportion of SNPs with PPA > 0.20");plt5
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
loc.df <- arrange(loc.df,desc(prop01))
summary(loc.df)
```

```
##      loci              numsnps           prop01             prop05
##  Length:49          Min.   :   1.0   Min.   :0.001042   Min.   :0.000000
##  Class :character   1st Qu.:  13.0   1st Qu.:0.058824   1st Qu.:0.004905
##  Mode  :character   Median :  27.0   Median :0.476191   Median :0.100000
##                     Mean   : 393.2   Mean   :0.488629   Mean   :0.241847
##                     3rd Qu.: 171.0   3rd Qu.:0.923077   3rd Qu.:0.259259
##                     Max.   :4799.0   Max.   :1.000000   Max.   :1.000000
##      prop10            prop20
##  Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.03704   Median :0.002941
##  Mean   :0.16971   Mean   :0.100857
##  3rd Qu.:0.21053   3rd Qu.:0.105263
##  Max.   :1.00000   Max.   :1.000000
```

```
loc.df
```

```
##                      loci numsnps      prop01       prop05       prop10
## 1                  TCF7L2       3 1.000000000 1.0000000000 1.0000000000
## 2        KCNQ1.rs74046911       3 1.000000000 1.0000000000 0.6666666667
## 3                  MTNR1B       1 1.000000000 1.0000000000 1.0000000000
## 4                   HNF1B       7 1.000000000 0.8571428571 0.7142857143
## 5                   PPARG      27 1.000000000 0.2222222222 0.0000000000
## 6                   ZBED3       5 1.000000000 0.4000000000 0.4000000000
## 7                  CDKAL1       8 1.000000000 0.8750000000 0.5000000000
## 8                SLC30A8       6 1.000000000 0.6666666667 0.6666666667
## 9      CDKN2B.rs10811660       6 1.000000000 0.6666666667 0.3333333333
## 10     CDKN2B.rs10757283       5 1.000000000 1.0000000000 0.6000000000
## 11               ADAMTS9      27 0.962962963 0.2592592593 0.0000000000
## 12                 ADCY5      17 0.941176471 0.3529411765 0.1764705882
## 13                  GCKR      13 0.923076923 0.5384615385 0.3846153846
## 14                CDC123      12 0.916666667 0.2500000000 0.2500000000
## 15                  HHEX      40 0.775000000 0.1000000000 0.0000000000
## 16               IGF2BP2      50 0.760000000 0.0400000000 0.0000000000
## 17                 GLIS3      10 0.700000000 0.2000000000 0.2000000000
## 18                 GRB14      24 0.666666667 0.2500000000 0.0833333333
## 19                CENTD2      27 0.629629630 0.2592592593 0.1481481481
## 20   KCNQ1.chr11_2692322      12 0.583333333 0.1666666667 0.1666666667
## 21                 JAZF1      16 0.562500000 0.1875000000 0.1875000000
## 22                  WFS1      82 0.524390244 0.0365853659 0.0000000000
## 23                  IRS1      65 0.492307692 0.0153846154 0.0000000000
## 24                 CILP2      29 0.482758621 0.0689655172 0.0689655172
## 25                KCNJ11      21 0.476190476 0.3809523810 0.2857142857
## 26                  PRC1      51 0.470588235 0.1568627451 0.0000000000
## 27                 PROX1      19 0.421052632 0.2105263158 0.2105263158
## 28                 HMGA2      72 0.416666667 0.0555555556 0.0000000000
## 29                BCL11A      35 0.400000000 0.2571428571 0.0285714286
## 30                NOTCH2     108 0.388888889 0.0000000000 0.0000000000
## 31       HNF1A.rs1169288      27 0.370370370 0.0740740741 0.0370370370
## 32                   FTO      72 0.361111111 0.0000000000 0.0000000000
## 33        DGKB.rs1974620     266 0.124060150 0.0000000000 0.0000000000
## 34                TSPAN8      76 0.118421053 0.1052631579 0.0394736842
## 35                   GCK      18 0.111111111 0.0555555556 0.0555555556
## 36                 KLF14     171 0.111111111 0.0058479532 0.0000000000
## 37                 HNF4A      17 0.058823529 0.0588235294 0.0588235294
## 38        GIPR.rs2238689      26 0.038461538 0.0384615385 0.0384615385
## 39                 THADA     247 0.036437247 0.0242914980 0.0080971660
## 40        GIPR.rs4399645     704 0.031250000 0.0028409091 0.0014204545
## 41       HNF1A.rs1800574     899 0.027808676 0.0022246941 0.0022246941
## 42       MC4R.rs17066842    1275 0.025098039 0.0000000000 0.0000000000
## 43 HNF1A.chr12_121440833    1427 0.011212334 0.0049053959 0.0000000000
## 44                C2CD4B    1851 0.008103728 0.0000000000 0.0000000000
## 45       KCNQ1.rs2237895     680 0.007352941 0.0029411765 0.0029411765
## 46       KCNQ1.rs2283220    2258 0.003985828 0.0004428698 0.0000000000
## 47        KCNQ1.rs458069    2309 0.001732352 0.0004330879 0.0000000000
## 48   MC4R.chr18_57739289    1343 0.001489203 0.0000000000 0.0000000000
## 49       DGKB.rs10276674    4799 0.001041884 0.0006251302 0.0004167535
##         prop20
## 1  0.333333333
## 2  0.333333333
## 3  1.000000000
## 4  0.142857143
## 5  0.000000000
## 6  0.400000000
## 7  0.250000000
## 8  0.500000000
## 9  0.333333333
## 10 0.400000000
## 11 0.000000000
## 12 0.058823529
## 13 0.076923077
## 14 0.250000000
## 15 0.000000000
## 16 0.000000000
## 17 0.200000000
## 18 0.041666667
## 19 0.000000000
## 20 0.083333333
## 21 0.187500000
## 22 0.000000000
## 23 0.000000000
## 24 0.034482759
## 25 0.000000000
## 26 0.000000000
## 27 0.105263158
## 28 0.000000000
## 29 0.000000000
## 30 0.000000000
## 31 0.037037037
## 32 0.000000000
## 33 0.000000000
## 34 0.013157895
## 35 0.055555556
## 36 0.000000000
## 37 0.058823529
## 38 0.038461538
## 39 0.004048583
## 40 0.000000000
## 41 0.001112347
## 42 0.000000000
## 43 0.000000000
## 44 0.000000000
## 45 0.002941176
## 46 0.000000000
## 47 0.000000000
## 48 0.000000000
## 49 0.000000000
```

```
write.table(df,cred.dir%&%"metabochip.chr_added.txt",row.names=FALSE,
            sep="\t",quote=FALSE)
write.table(loc.df,cred.dir%&%"metabochip.chr_added.locusSummary.txt",row.names=FALSE,
            sep="\t",quote=FALSE)
```