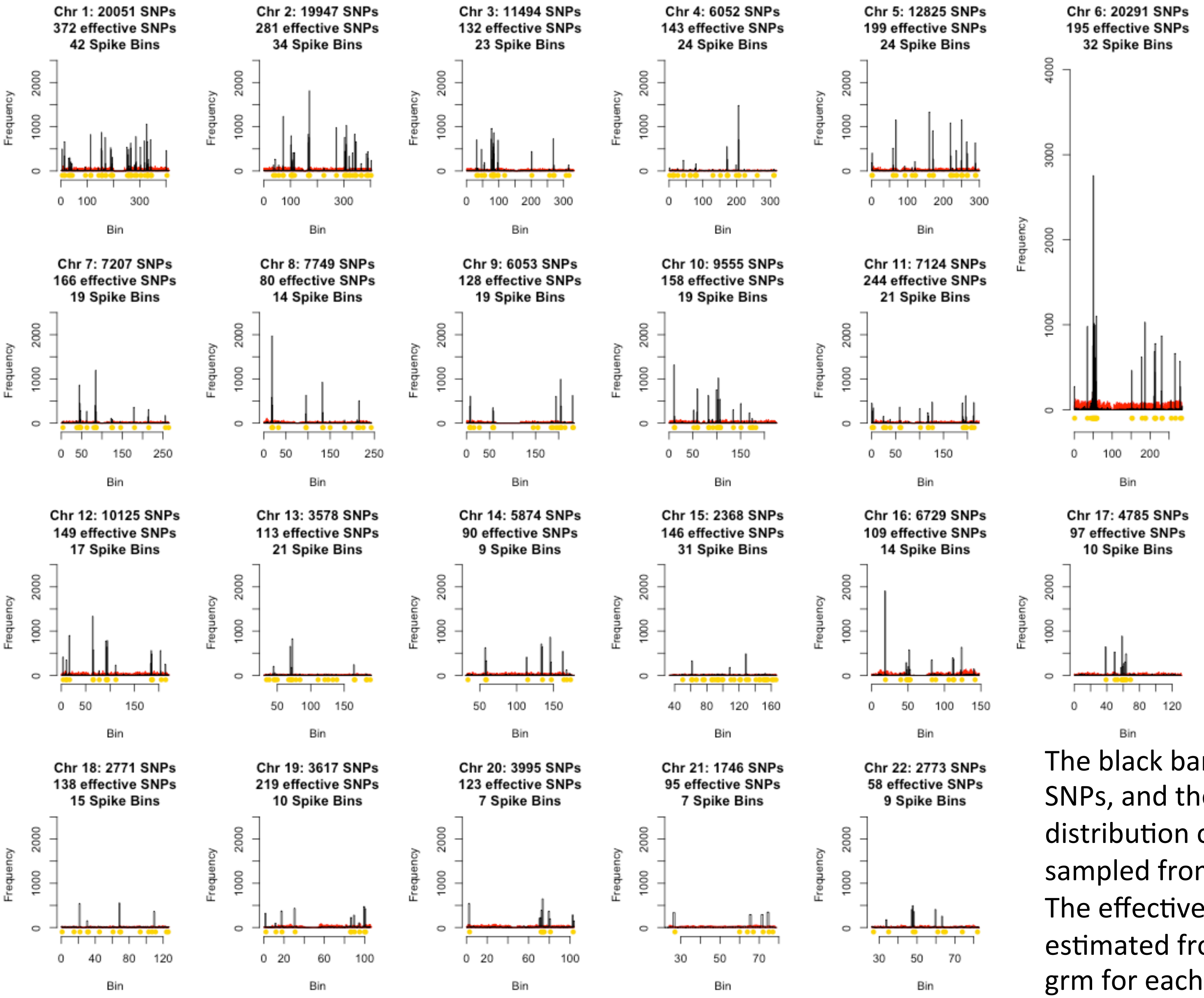
Inflammatory Bowel Disease (IBD) Immunochip QC steps

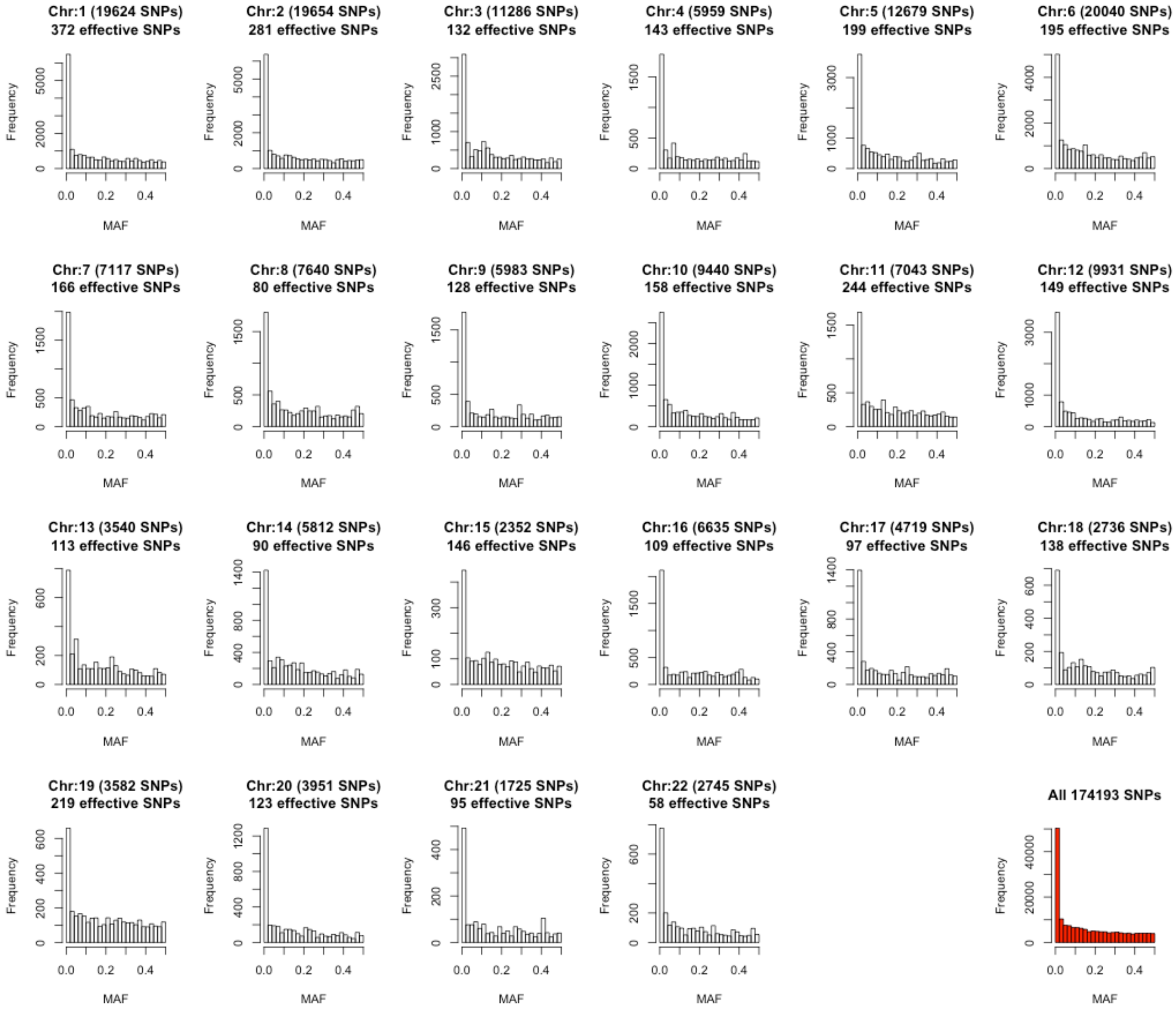
Publication: Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. Chen, et al, ***Hum Mol Genet***, 2014, 17:4710-4720

Bird's eye view for IBD immunochip



Each chromosome was split into a number of bins, the size of which is 600Kb, about 0.2cM.
Each yellow dot indicates a bin that is selected as "spike".

The black bars represent clustered iChip SNPs, and the red bars represent the distribution of the same number of SNPs sampled from HapMap imputed data. The effective number of markers were estimated from the distribution of the grm for each chromosome.



MAF distribution

The MAF spectrums resemble each other over chromosomes.

The effective number of markers were estimated from the empirical distribution of each chromosome's GRM, N(-1/Ne, 1/sqrt(Me)).

Due to highly LD between the markers, Me is small on iChip.

QC Stage 1 (by IBD consortium)

- Stage 1 QC has already been conducted by the IBD consortium

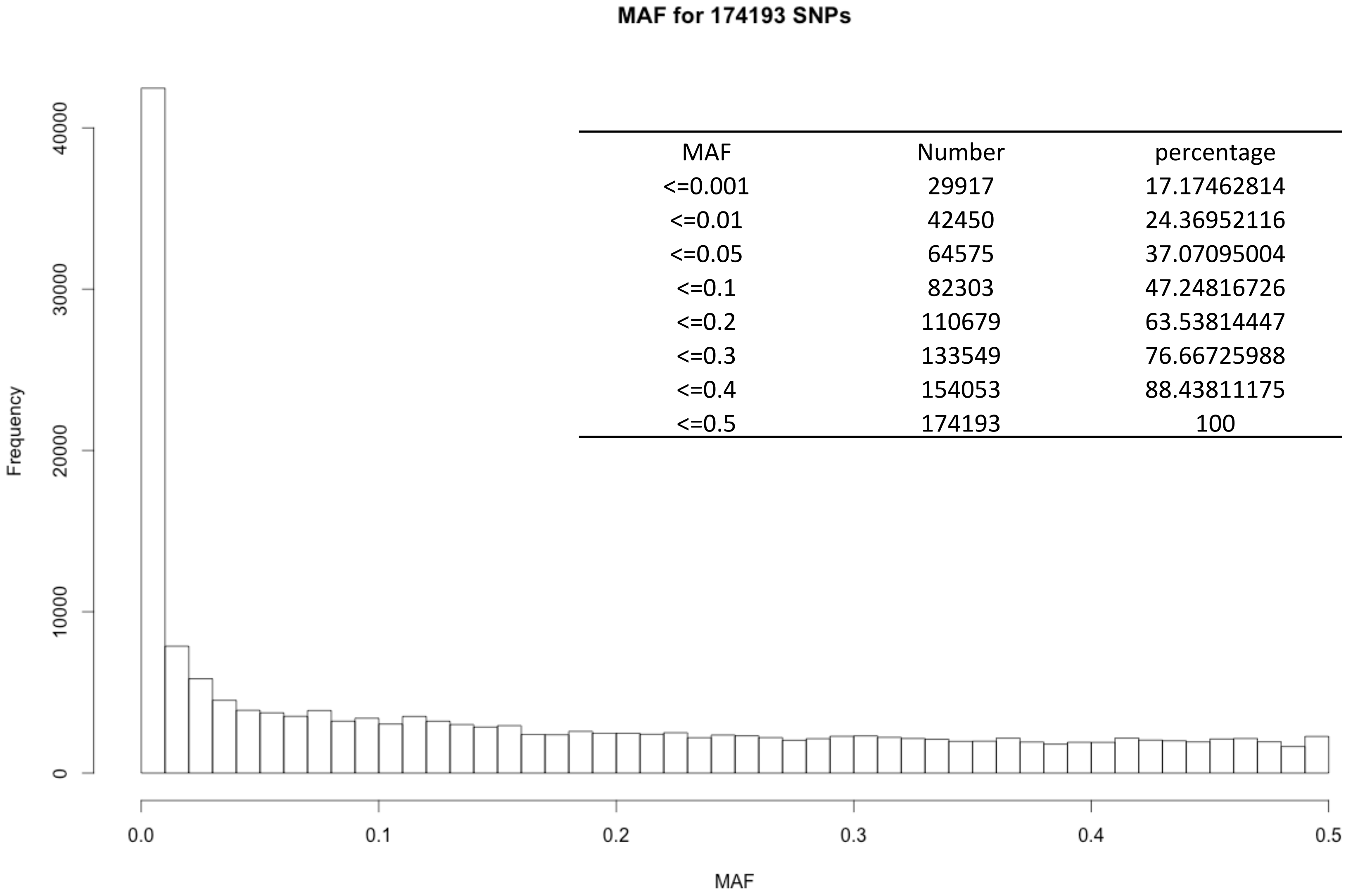
- Removed 2471 related individuals (IBD>0.4, including duplicated ones. We will revisit, and see GRM section for details)
- Removed 12433 variants failed heterogeneous frequency test in 2+ batches
- Removed 16886 variants that are not in 1KG phase 2 (We put them back)
- Blanked out variants failed HWE test in only one batch
- Blanked out variants failed heterogeneous frequency test in only one batch
- We further removed individuals whose affection statues are unknown (1178 individuals)

- After stage 1 QC, it yields 33,306 cases, 28,248 controls, and 174193 SNPs.

QC Stage 2: in-house QC

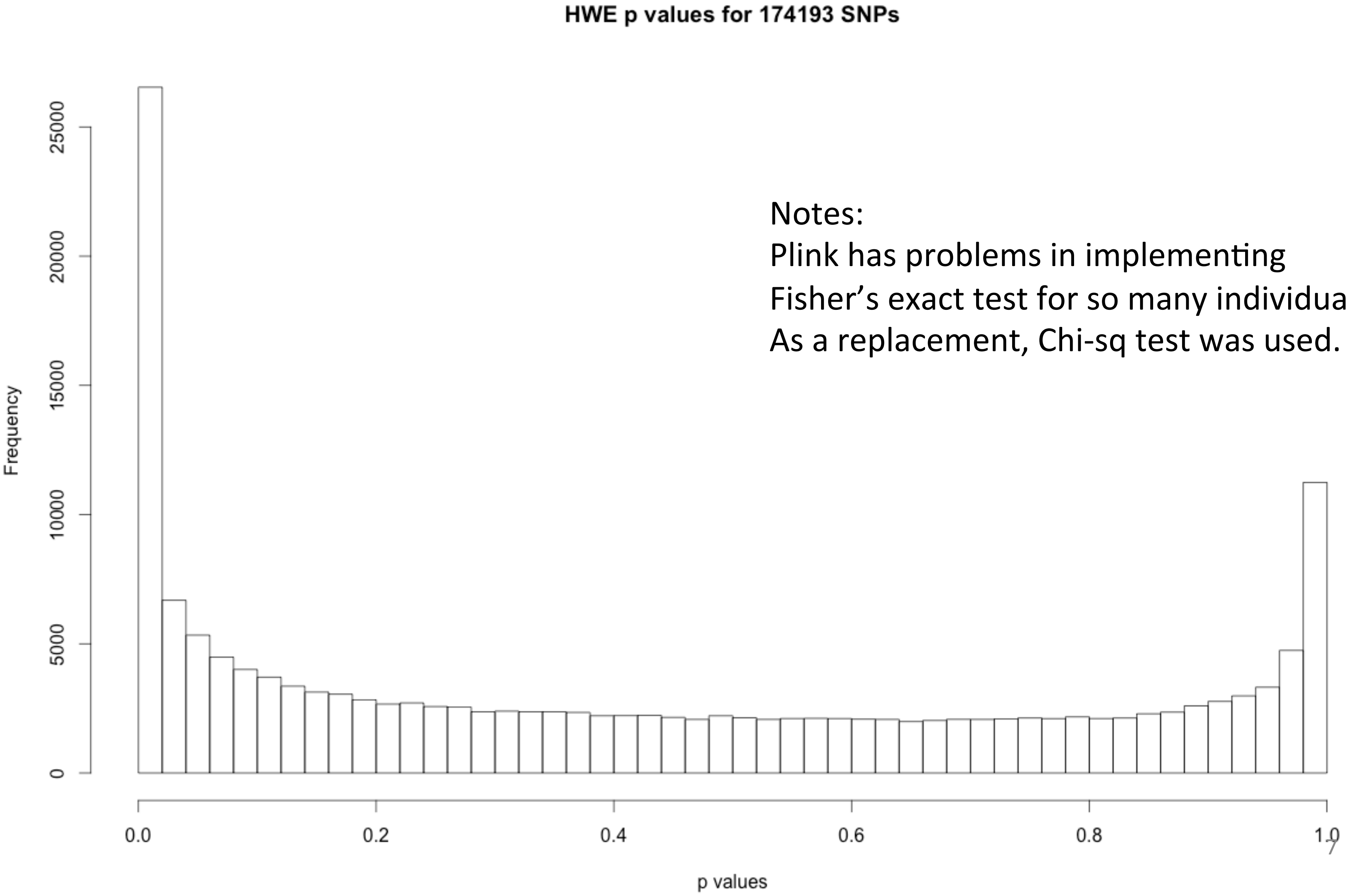
- MAF
- HWE test with Chi-sq test (plink failed to implement Fisher’s exact test for this data)
- Inbreeding
- Missing rate

MAF



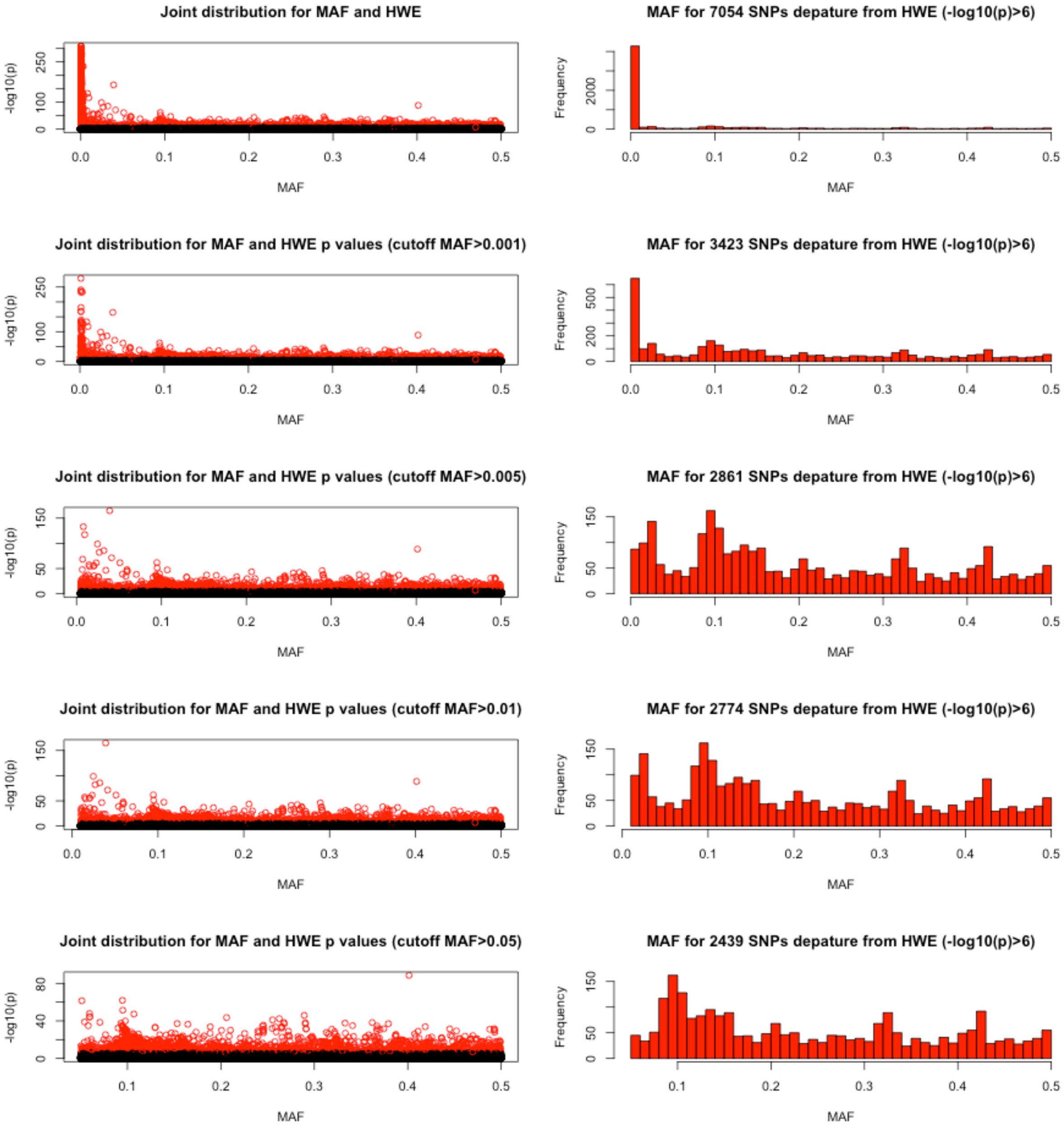
MAF	Number	percentage
<=0.001	29917	17.17462814
<=0.01	42450	24.36952116
<=0.05	64575	37.07095004
<=0.1	82303	47.24816726
<=0.2	110679	63.53814447
<=0.3	133549	76.66725988
<=0.4	154053	88.43811175
<=0.5	174193	100

HWE



Notes:
Plink has problems in implementing Fisher’s exact test for so many individuals. As a replacement, Chi-sq test was used.

Joint distribution for MAF and HWE p-values



Joint distribution for MAF and HWE

MAF for 7054 SNPs depature from HWE (-log10(p)>6)

Joint distribution for MAF and HWE p values (cutoff MAF>0.001)

MAF for 3423 SNPs depature from HWE (-log10(p)>6)

Joint distribution for MAF and HWE p values (cutoff MAF>0.005)

MAF for 2861 SNPs depature from HWE (-log10(p)>6)

Joint distribution for MAF and HWE p values (cutoff MAF>0.01)

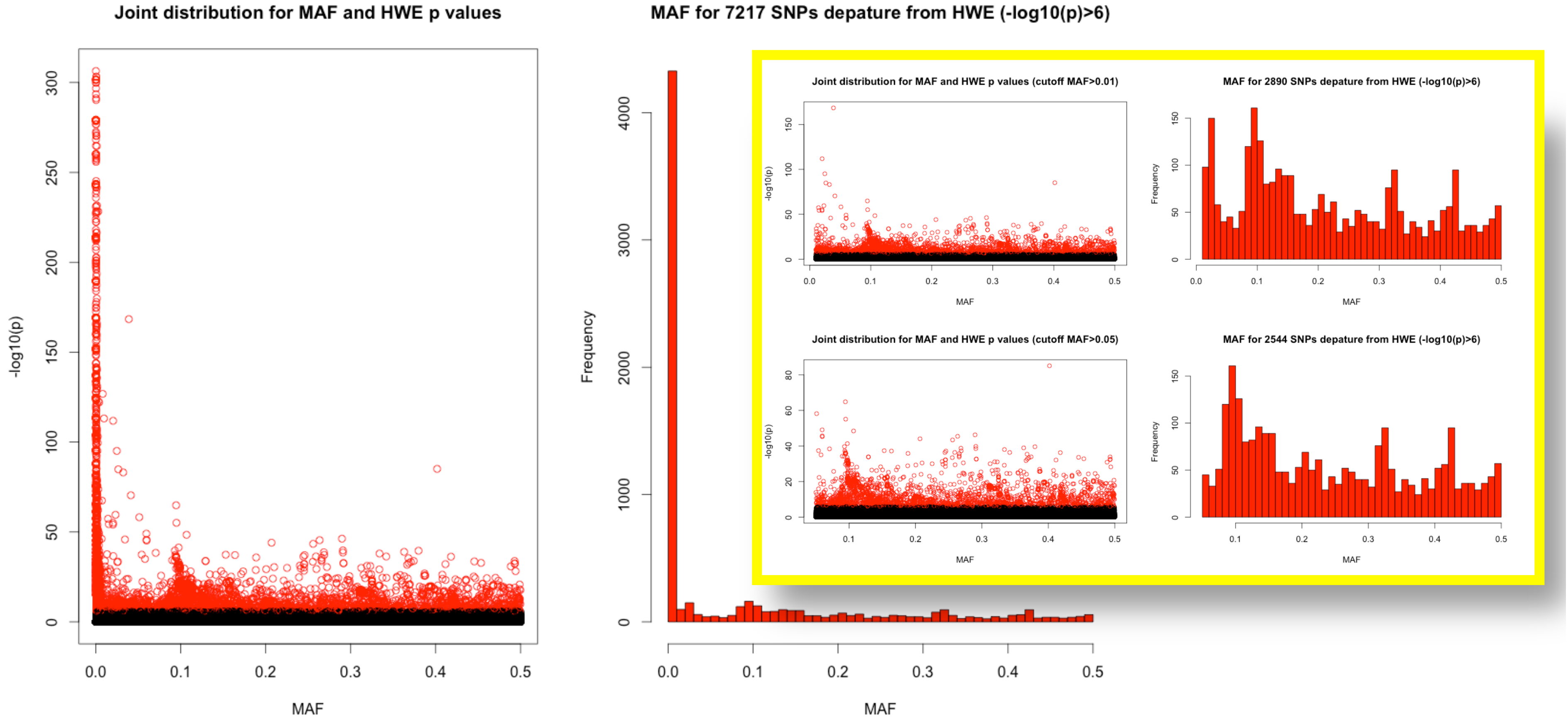
MAF for 2774 SNPs depature from HWE (-log10(p)>6)

Joint distribution for MAF and HWE p values (cutoff MAF>0.05)

MAF for 2439 SNPs depature from HWE (-log10(p)>6)

Joint distribution for MAF and HWE p-values, the red spots are the ones that had p-values greater than 6 after taking –log10(p). Eliminating low MAF, <=0.005, made the HWE p-values evenly distributed. Unexpected high homozygosity caused the low p-values, especially for rare variants.

MAF and HWE pvalues



Joint distribution for MAF and HWE p values

MAF for 7217 SNPs depature from HWE (-log10(p)>6)

Joint distribution for MAF and HWE p values (cutoff MAF>0.01)

MAF for 2890 SNPs depature from HWE (-log10(p)>6)

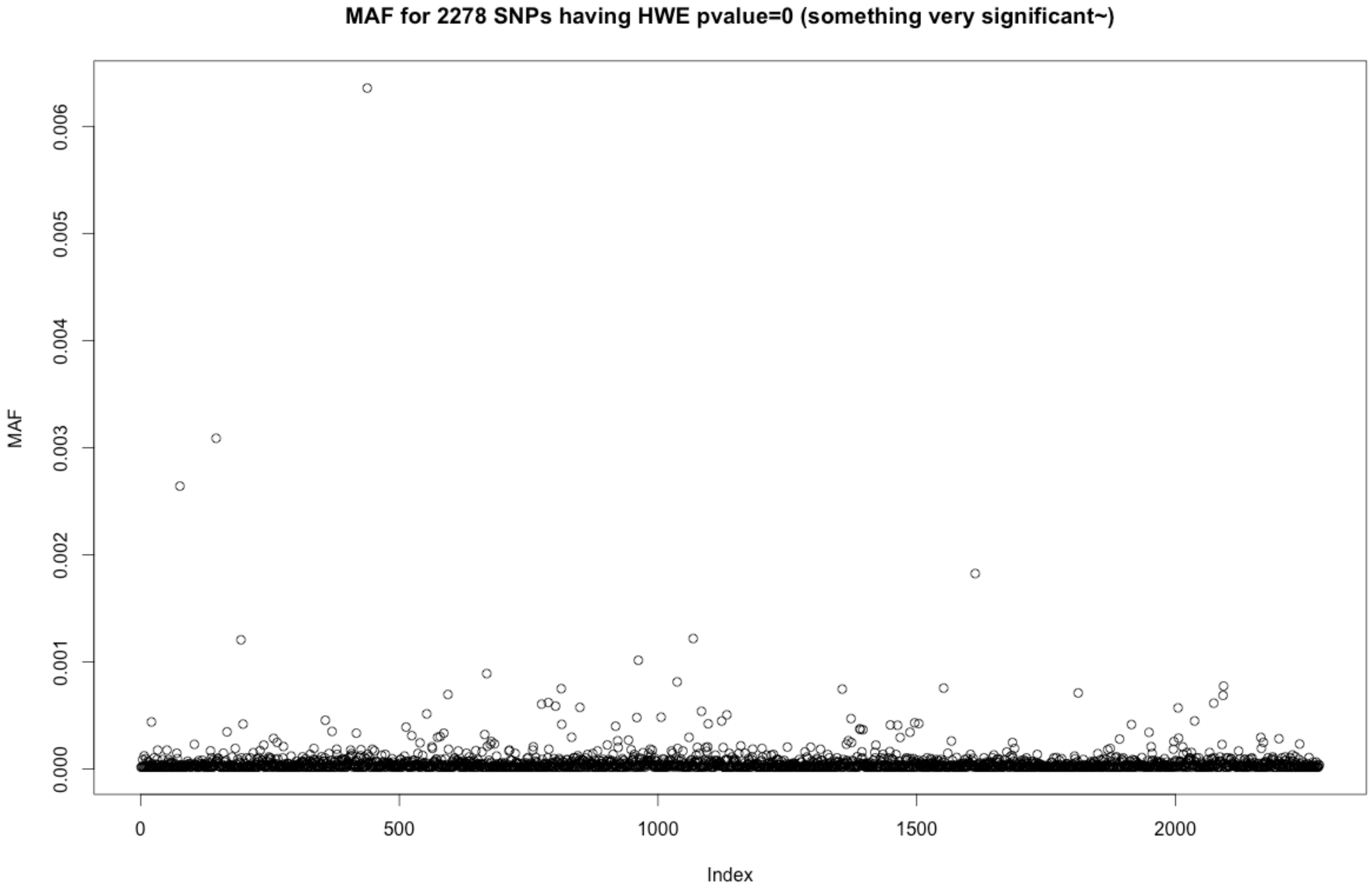
Joint distribution for MAF and HWE p values (cutoff MAF>0.05)

MAF for 2544 SNPs depature from HWE (-log10(p)>6)

In IBD iChip data, a lot of extremely low p-value for HWE were ascribed to low frequent SNPs.

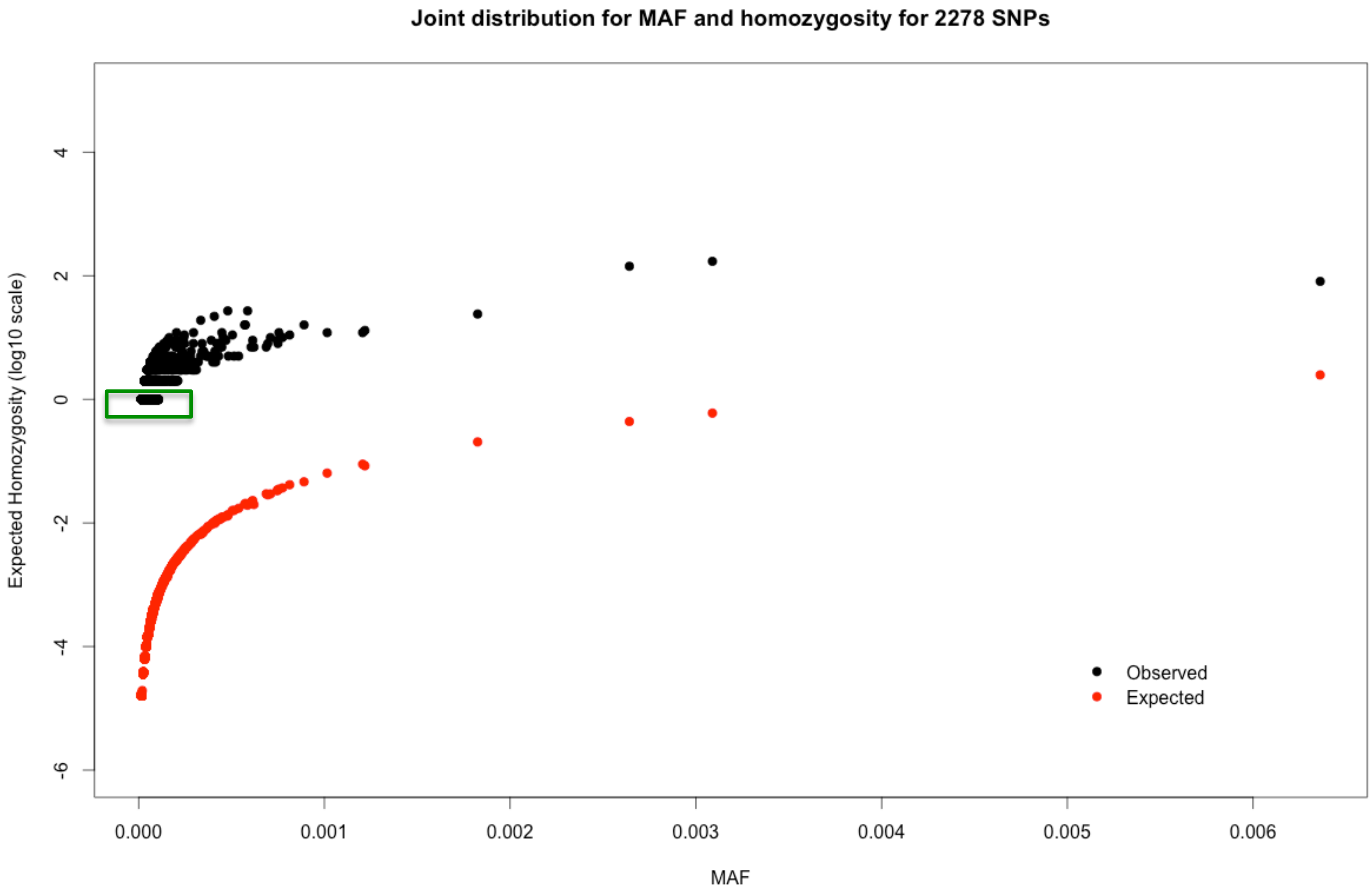
The next two slides investigate why so many low frequent SNPs showed such low p-values.

MAF for SNPs that had pvalue=0 for HWE test



Extracted SNPs had extremely significant p-values, which were represented as “0” in the plink HWE output. For these 2278 found SNPs, their MAF distribution were illustrated.

Expected and observed homozygosity

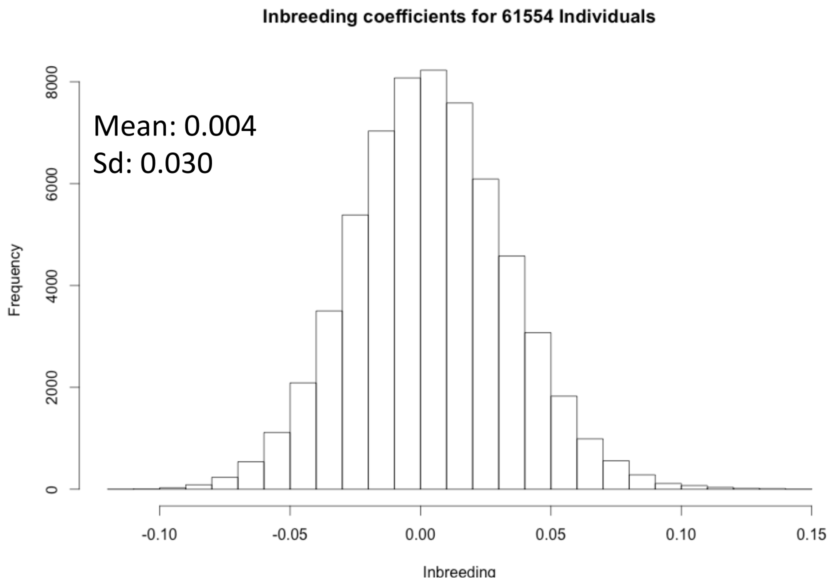


For these extracted 2278 SNPs, their observed (black) and expected (red) homozygosity were compared. Expected homozygosity was calculated: MAF^2*(non-missing sample size).

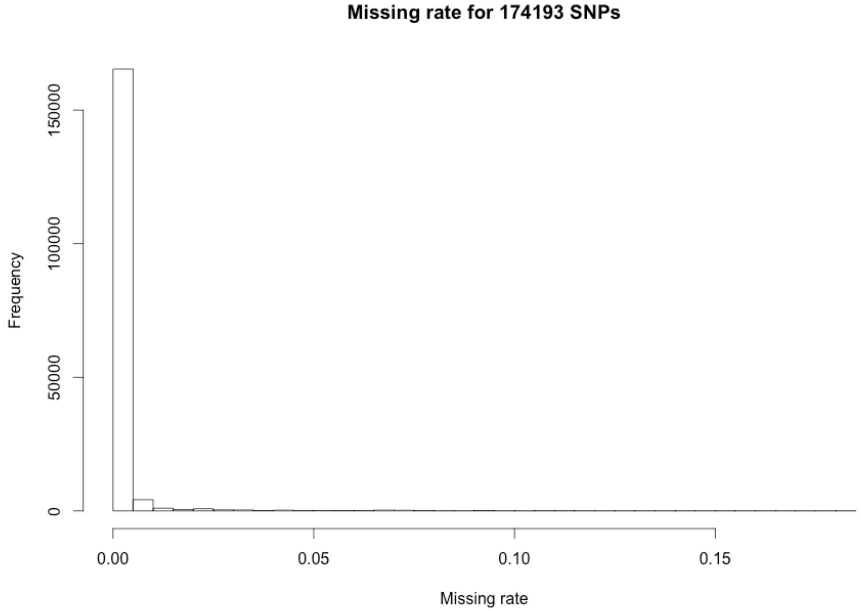
As a matter of fact, at least one homozygosity (in green box) were detected for all these SNPs. To scale up the difference, log10 transformation was taken for homozygosity. Notes: for MAF=0, log10 goes negative infinite.

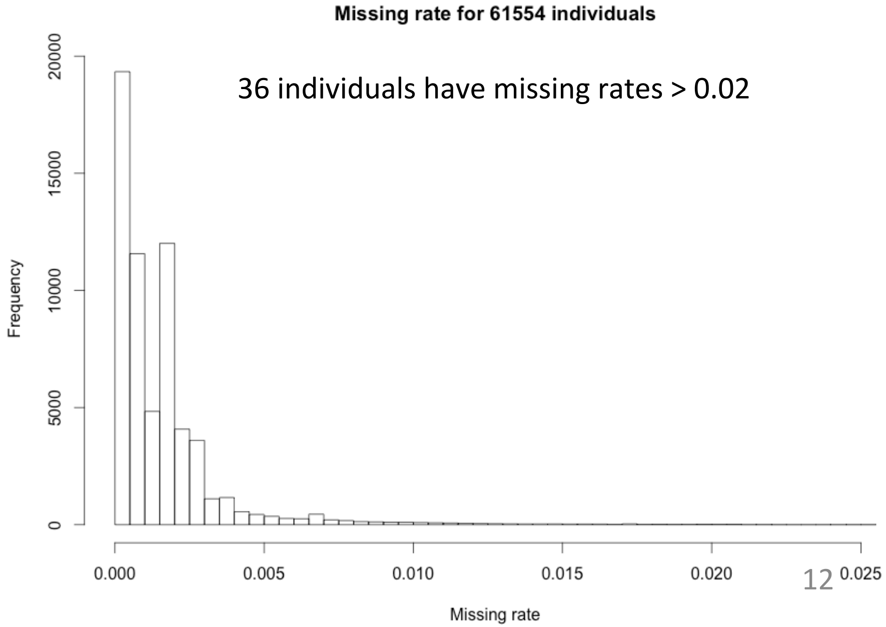
Too many – one seems too many for low frequency ones – homozygosity brought about absurd high p-values for HWE.

Inbreeding, missing rates



Mean: 0.004
Sd: 0.030





36 individuals have missing rates > 0.02

QC2 criteria

- MAF>0.001 (excluded 29971 SNPs)
- HWE>10E-6 (excluded 14787 SNPs)
- Missing rate per person<0.02 (excluded 36 Individuals)
- It ends up with 61518 individuals, 140853 SNPs.

- MAF>0.01 (excluded 42447 SNPs)
- HWE>10E-6 (excluded 14787 SNPs)
- Missing rate per person<0.02 (excluded 36 Individuals)
- It ends up with 61518 individuals, 128972 SNPs.

Two methods for calculating effective number of markers

- Method I: simulation methods
  After QC2, we conducted simulation as below
  sample 5000 individuals
  1 randomly assign 0 or 1 to each individual
  2 conduct GWAS and calculate Chisq statistic for each marker
  3 sum Chisq statistic, denoted as C_S
  Repeat 1-3 for 1000 rounds of simulation, and calculate var(C_S)

- M_e=M*[2M/var(C_S)].
- M_e=2822, about 2% of the total markers.

- Method II: GRM Methods
  1 Calculate sampling variance for lower off-diagonal elements.
  2 the reciprocal of that is Me.
- **See Supplementary notes in Chen G-B, Front Genet 5:107 for details.**

GRM stats

1,892,262,921 relatedness scores were generated over 61518 individuals (Thank Hong). The correlation between the two GRMs, estimated under MAF cutoff on 0.01 and 0.001, is 0.994.

- Given MAF cutoff on 0.01 (denoted as GRM_01).
- GRM was generated on 128972 SNPs.
- 90109 pairs have relatedness greater than 0.1
- Assuming the theoretical distribution for lower GRM triangle is N(1/(n-1), 1/sqrt(m)), the estimated distribution for GRM_001 is N(1.628e-5, 0.021), which gives n_hat=61425 and m_hat=2222.

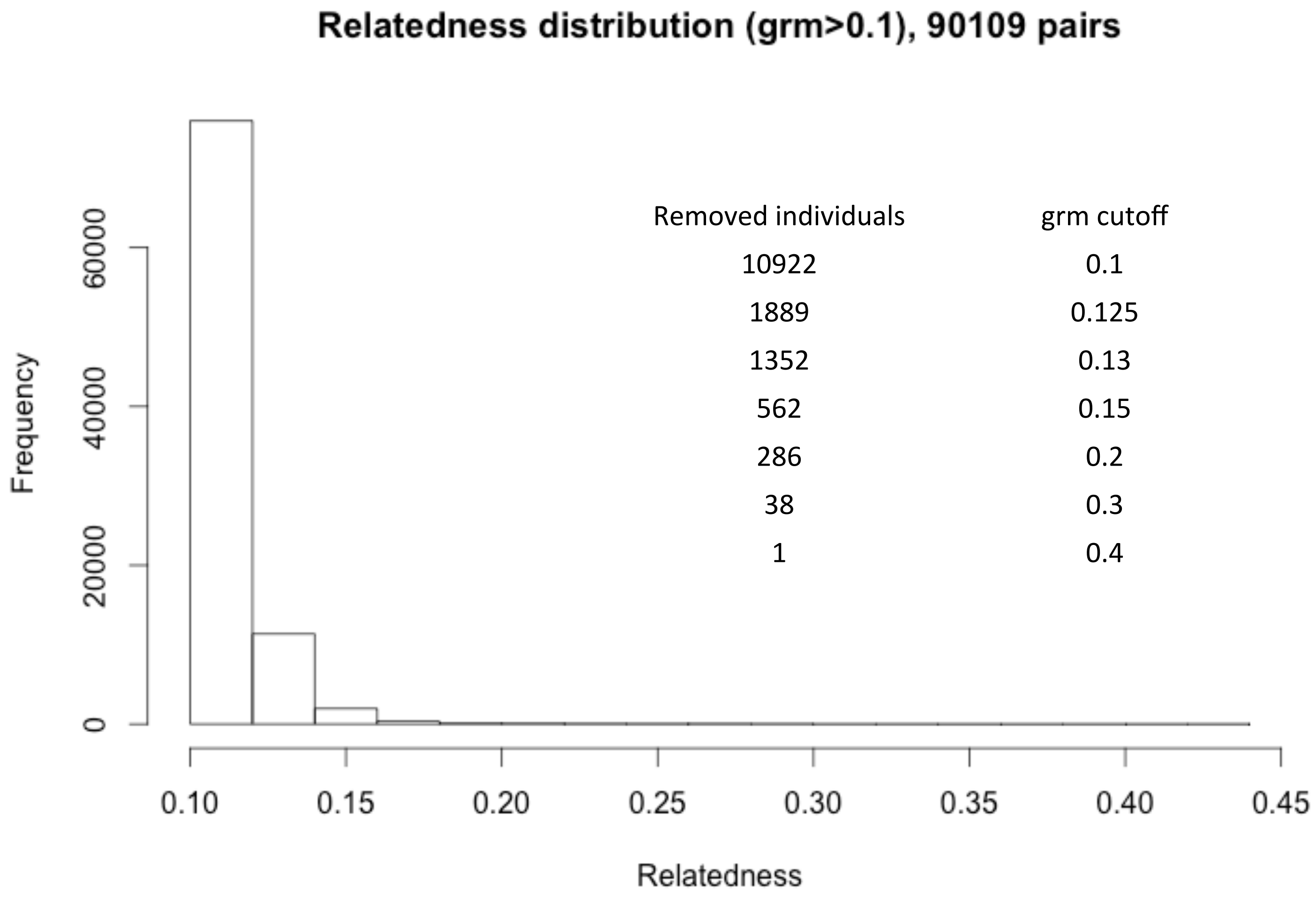
  "n" is sample size, and "m" is the number of markers if assume sampling variance only.

- Given MAF cutoff on 0.001 (denoted as GRM_001).
- GRM was on 140853 SNPs.
- 176773 pairs have relatedness greater than 0.1
- The estimated distribution for GRM_01 is N(1.627e-5, 0.0198), which gives n_hat=61462 and m_hat=2551.

Given MAF cutoff on 0.05 (denoted as GRM_05).
GRM was estimated over 107185 SNPs
N(1.627e-05, 0.00239)
N_hat=61462 and m_hat=1748

MAF cutoff on 0.01

Distribution of the GRM_001 at its right tail (grm > 0.1)



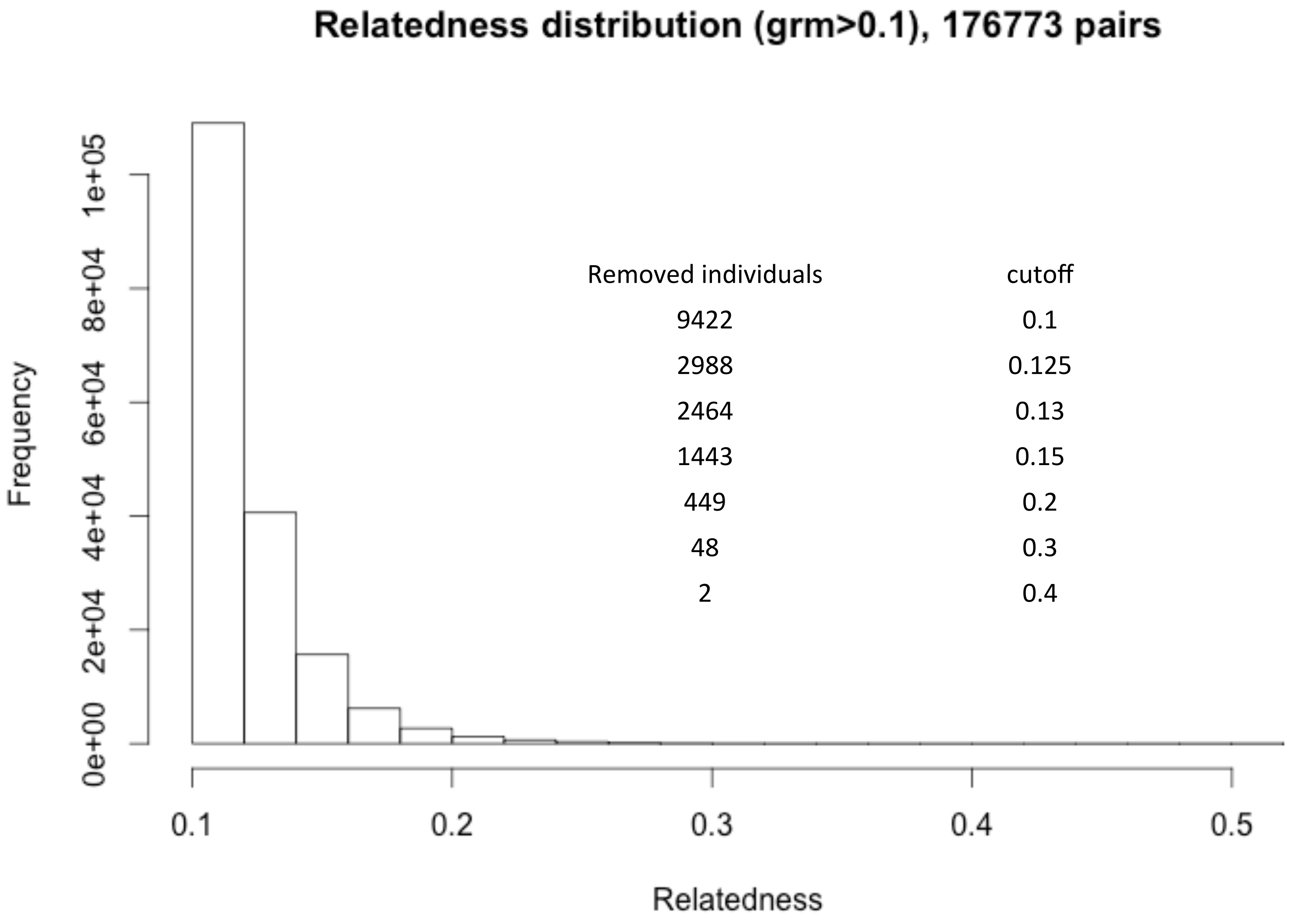
Relatedness distribution (grm>0.1), 90109 pairs

Removed individuals	grm cutoff
10922	0.1
1889	0.125
1352	0.13
562	0.15
286	0.2
38	0.3
1	0.4

The biggest relatedness was 0.504.

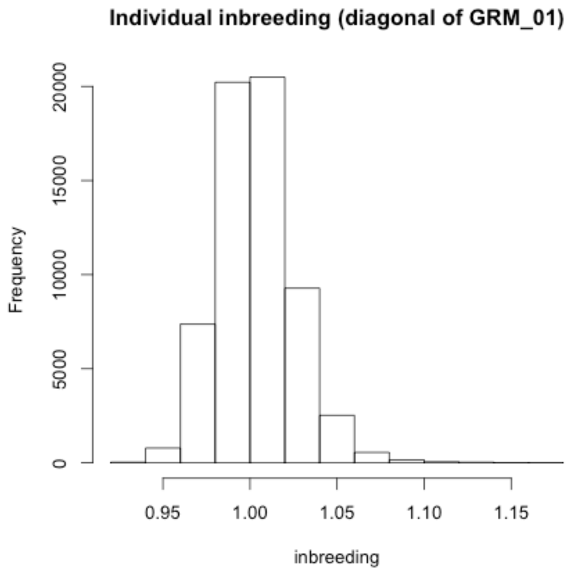
MAF cutoff on 0.001

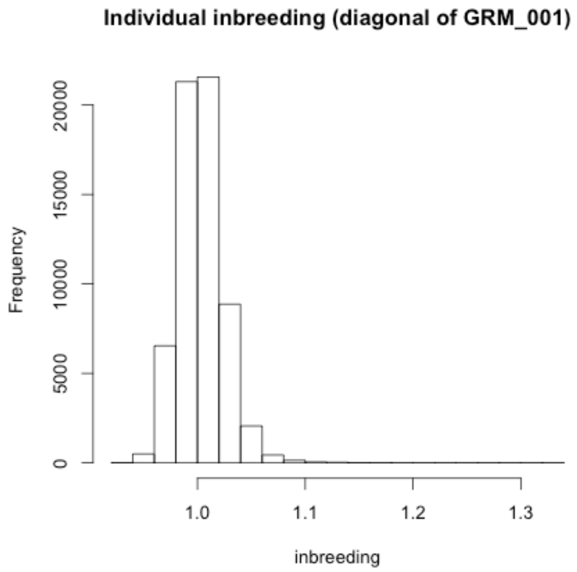
Distribution of the GRM_01 at its right tail (grm > 0.1)



Removed individuals	cutoff
9422	0.1
2988	0.125
2464	0.13
1443	0.15
449	0.2
48	0.3
2	0.4

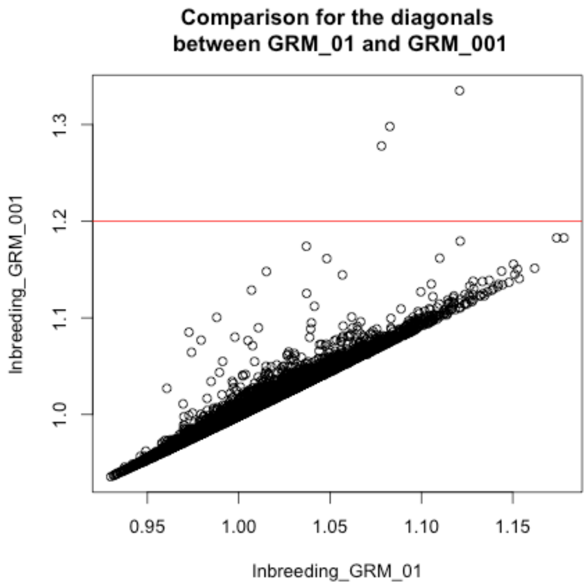
Inbreeding (diagonal in both GRMs)


Individual inbreeding (diagonal of GRM_01)

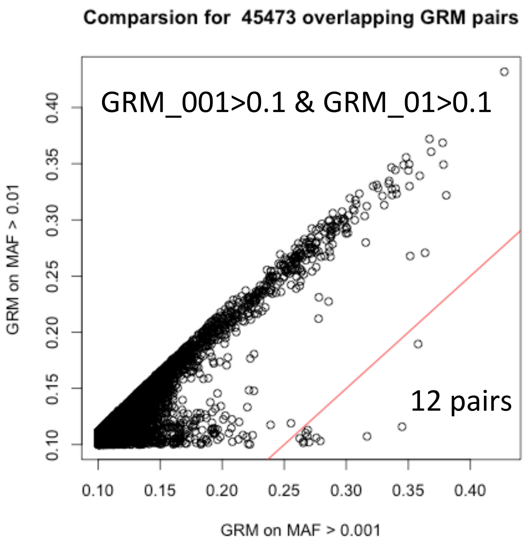

Individual inbreeding (diagonal of GRM_001)

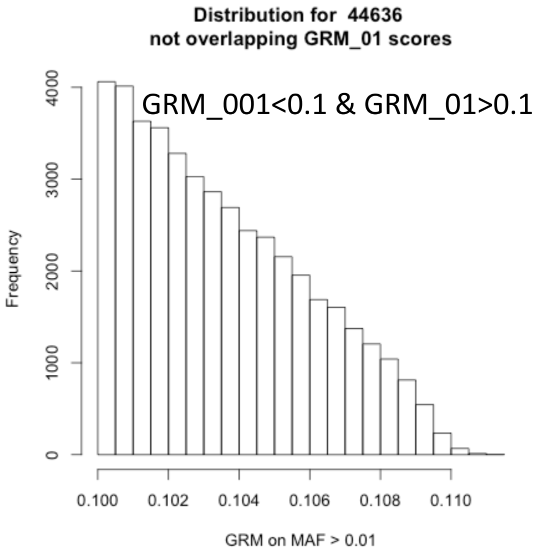
Inbreeding here refers to the diagonal elements extracted from two GRMs. The top left and right illustrate the diagonals in GRM_01 and GRM_001, respectively. In GRM_001 high inbreeding, as high as 1.3 or so, individuals show up.

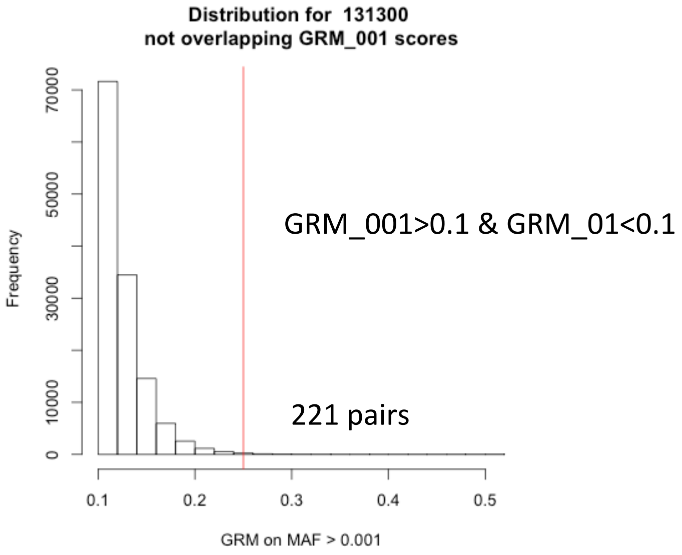
Among the three individuals that have GRM_001 greater than 1.2 two are Swedish samples (IG3199901 and IG3186478), whose relatedness is 0.5, and another one is Iranian (IS3072563). All three individuals have nearly consecutive homozygotes at very low MAF loci.


Comparison for the diagonals between GRM_01 and GRM_001

Further examination for discrepancy between two GRM sets







- For these two GRM sets, I split them into three schemes:
  1) GRM_001>0.1 & GRM_01>0.1 (top left)
  2) GRM_001>0.1 & GRM_01<0.1 (down left)
  3) GRM_001<0.1 & GRM_01>0.1 (top right)
- Scheme 1: It yielded 45473 pairs, and of them 12 were found having **GRM_001-GRM_01>0.15** (top left).
- Scheme 2: we further picked up GRM_001>0.25 so that also abs(GRM_001-GRM_01)>0.15, and it picked up another 221 pairs (bottom left). Those pairs might reflect that the weight were drawn by the low allele frequency spectrum ranged from 0.001 to 0.01.

- These 233 pairs were ascribed to 187 individuals. And among them, the most suspected individuals were IG3199901 and IG3186478, both of who were genotyped in **Unaffected_tbalschun_icbatch3**. Between them the relatedness was 0.50, and both independently showed many high relatedness, greater than 0.2, with other picked individuals.

- Scheme 3: there were no relatedness scores too big to be worried (top right).

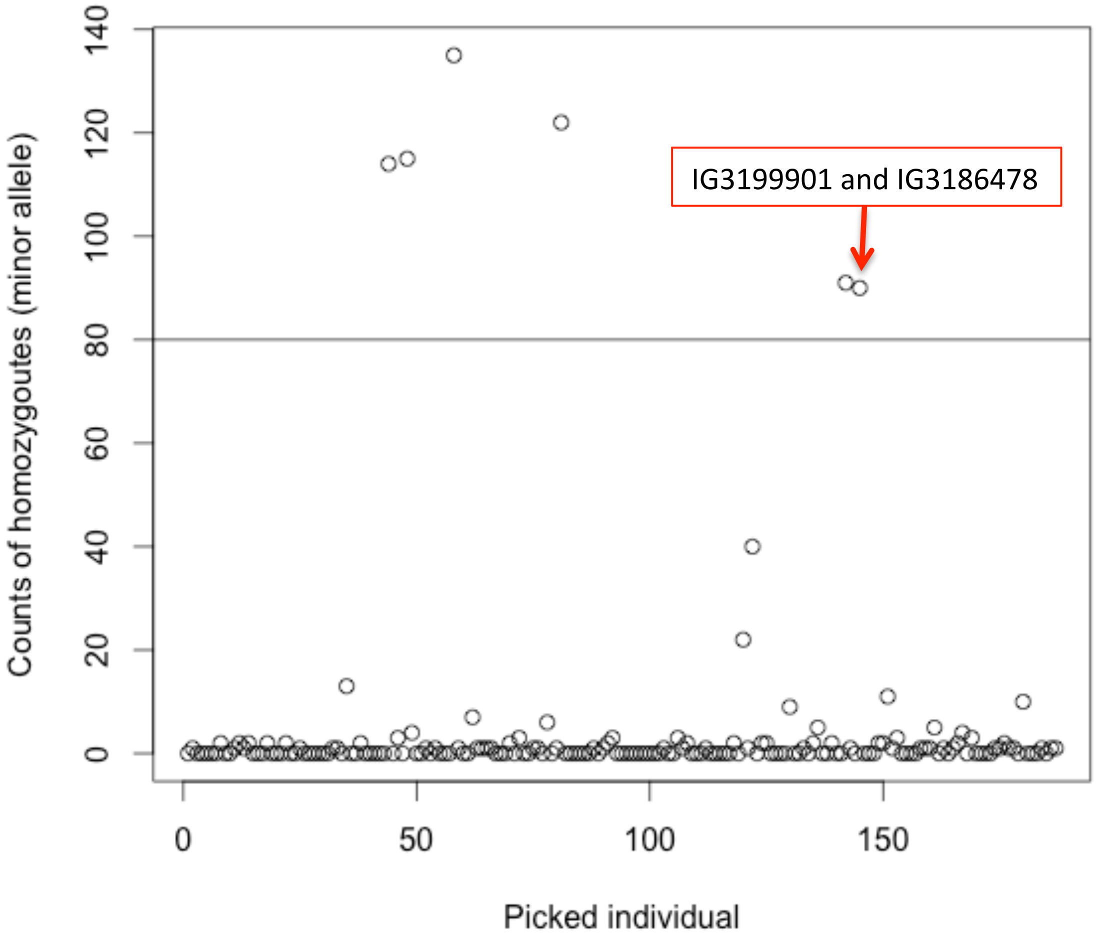
- Notes: GRM_001 refers to GRM constructed with alleles whose MAF cutoff was 0.001, and GRM_01 0.01.

Find GRM cut-off for iChip

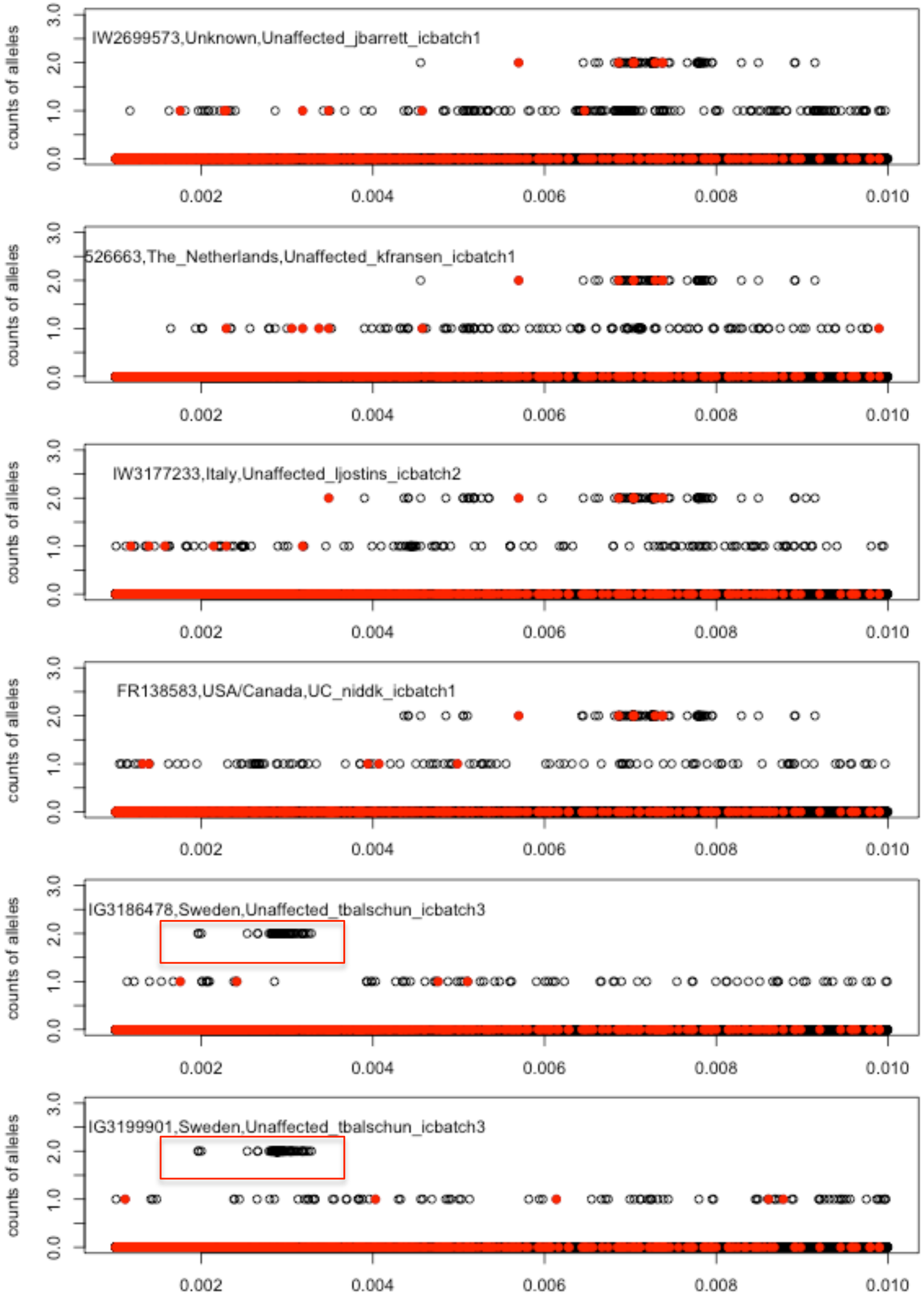
- Due to small number of effective number of markers, grm cut-off is different from what we use for generic chips.
  Me for generic chip is about 30,000
  Me for iChip is about 3,000.
- The equation to find grm cut-off is qnorm(1-1/(n*(n-1)/2))*sqrt(1/Me)

  n is sample size, Me is effective number.
  Given 60,000 samples, Me=3000, the cut-off is about 0.12 for iChip data.

Enriched homozygotes in 11881 rare variants, ranging from 0.001 to 0.01, for these 187 extracted individuals



- The counts of homozygotes for the 187 picked individuals were illustrated for their homozygotes in the allelic spectrum between 0.001 and 0.01.
- Six individuals had more than 80 homozygous were further examined in the next slides.
- IG3199901 and IG3186478 had homozygotes in rare MAF loci (see the next slide).



X axis represents allele frequency, and the Y axis represents counts of the reference alleles at each loci.

For the last two individuals, IG3199901 and IG3186478, whose relatedness was less than 0.1 in GRM_01, had relatedness in GRM_001 was 0.5, which was contributed by the about these 90 homozygotes (red squared), as roughly calculated below.

Given in total ~140000 SNPs, the relatedness ascribed to these 90 homozygotes, say the average MAF of them was about 0.003, can be calculated as: (2-2*0.003)^2/(2*0.003) *90/140000=0.42. It concluded that their relatedness, as high as 0.50, was largely contributed by the low frequency homozygotes.

Once another individual has genotype "1" along these loci, the relatedness with either of these two individuals will be about (2-2*0.003)*(1-2*0.003)/(2*0.003) *90/140000=0.21.

These high homozygotes were really cryptic, if they were not brought about by genotyping. In fact, these homozygotes are quite clustered to each other.

Notes:

1) for these 11881 low MAF loci, 837, highlighted in red, were not include in 1KG.

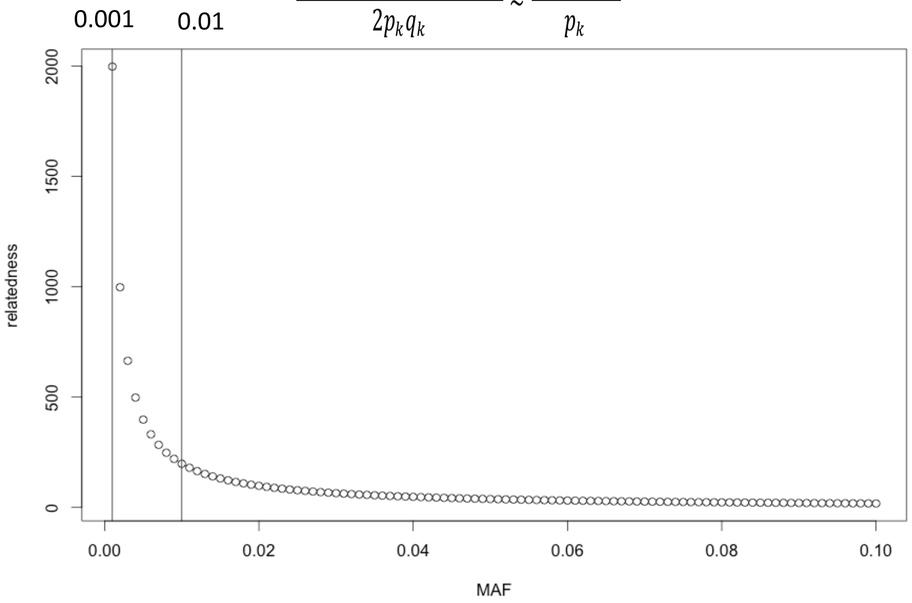
2) These 90 homozygotes loci, red squared in the last two panels, are located next to each other along **Chromosome 12** . Very likely, a rare haplotype, covering these 90 loci, resides at the region.

Rare variants can generate pseudo relationship

- The weight drawn by rare variants can be substantial especially when both individual have homozygote for the rare allele.
  The figure shows the relatedness score ascribes to a pair of homozygotes (x=2).

- It may generate pseudo relationships, such as full sib/parent-offspring, or half sib, because of a couple of homozygotes.

$$\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_kq_k} \approx \frac{2(1-p_k)}{p_k}$$

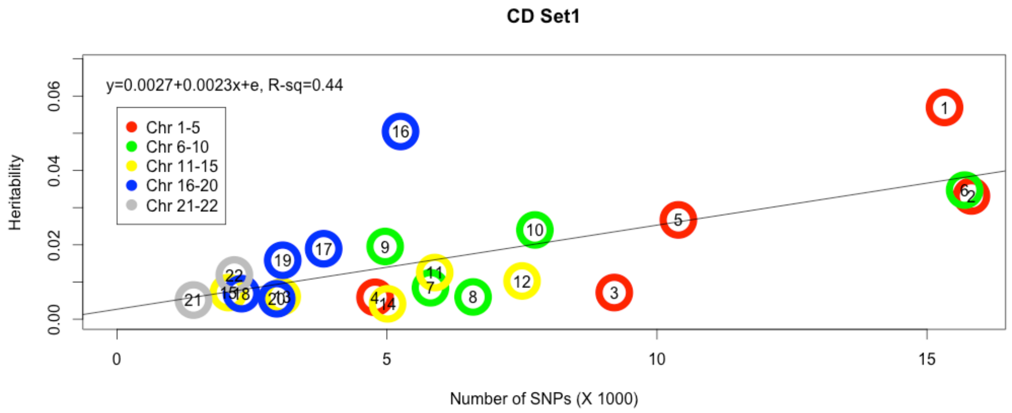


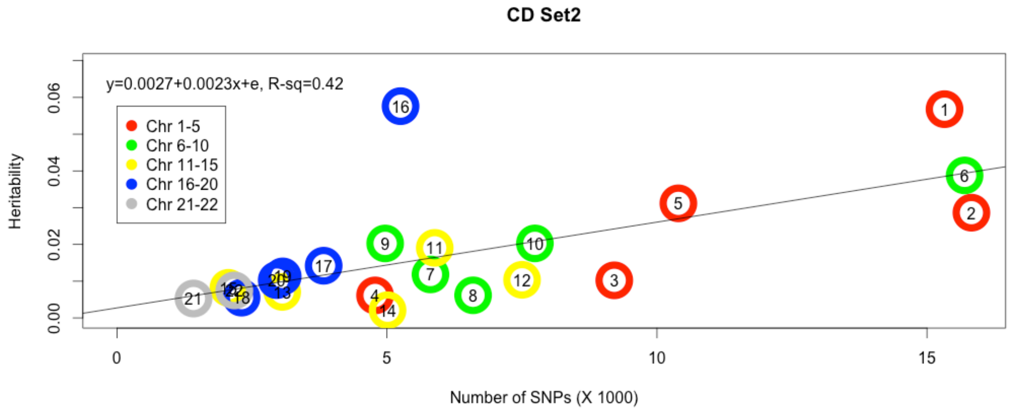
Individual 1	Individual 2	GRM_001	GRM_01	Pseudo relationship
IG3199901 (90 homozygotes with MAF 0.003)	IG3186478 (90 homozygotes with MAF 0.003)	0.504	<0.1	full sib
IG3199901 (90 homozygotes with MAF 0.003)	FP229832 (90 heterozygotes with MAF 0.003)	0.252	<0.1	half sib

Phenotypes

- The top 10 PCs were investigated the variance explained for CD and UC, respectively.
- Although 22699 CD cases, 18101 UC cases, and 36099 controls in total, only 19761 CD cases, 14833 UC cases, and 28999 controls had their top 10 PCs recorded.

Joint analysis for CD





The top 10 PCs were fitted as the covariates. 22 chromosomes were fitted altogether.

23404 individuals in set 1. The sum of the heritability was 0.377.

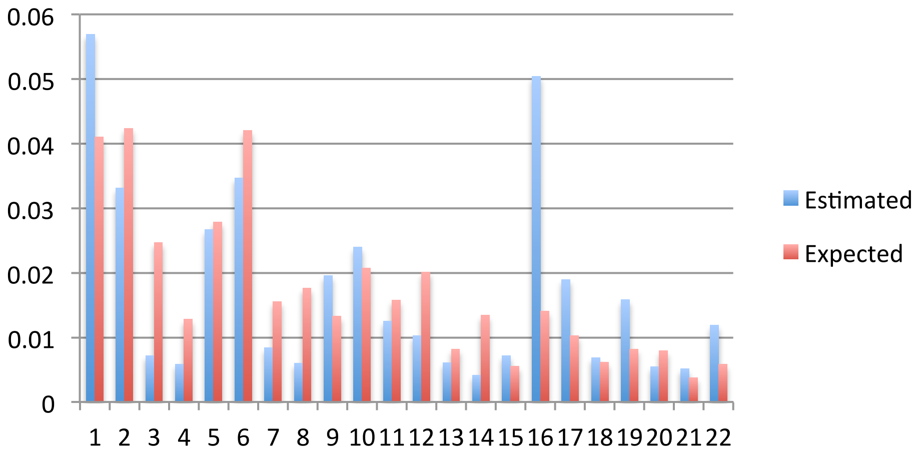
23408 individuals in set 2. The sum of the heritability was 0.389.

Chr16 stood atop in CD (but not UC) probably because of NOD2.

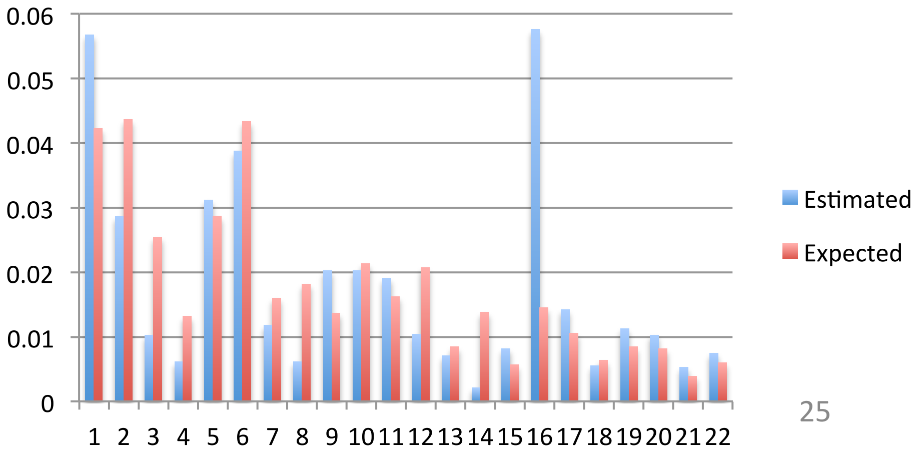
The difference between chromosomal heritability was not significant.

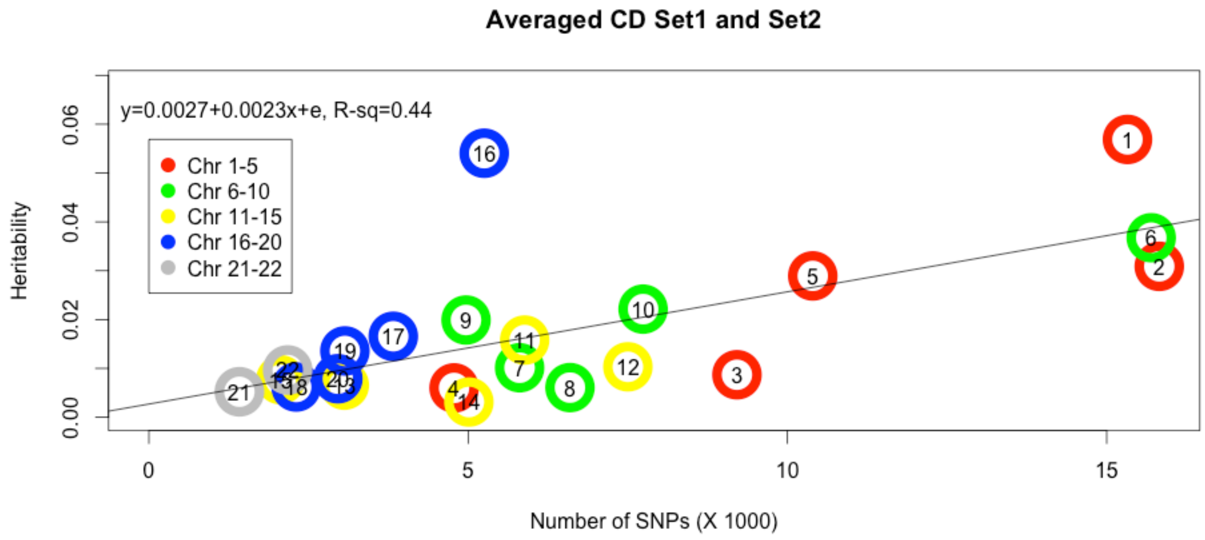
The whole genome gave 0.42.

CD Set 1



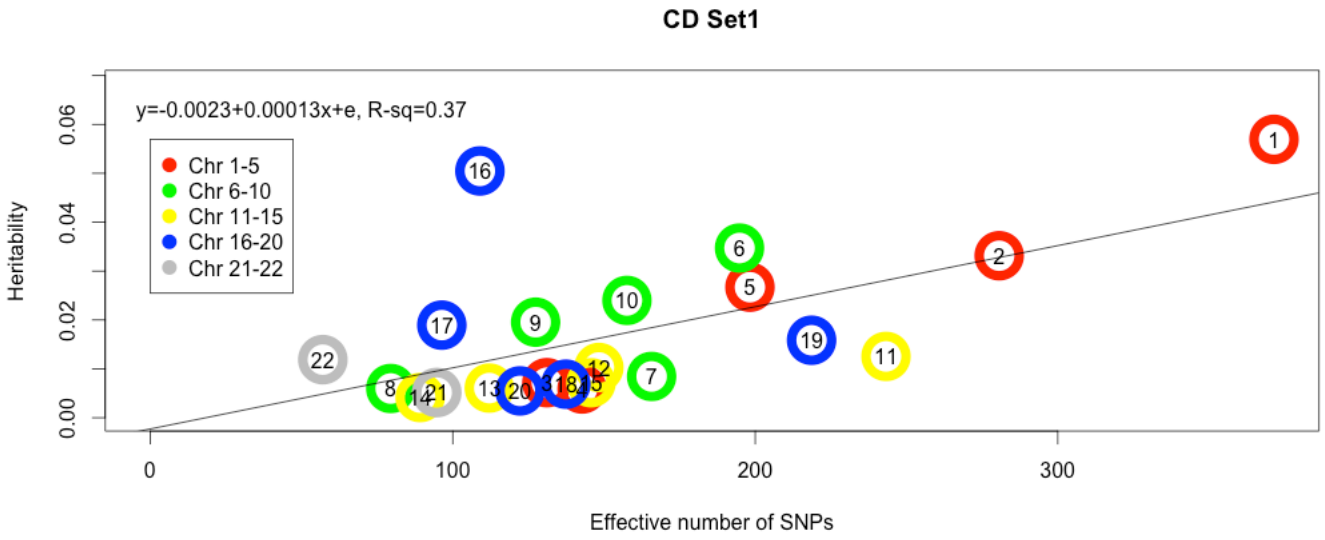
CD Set 2

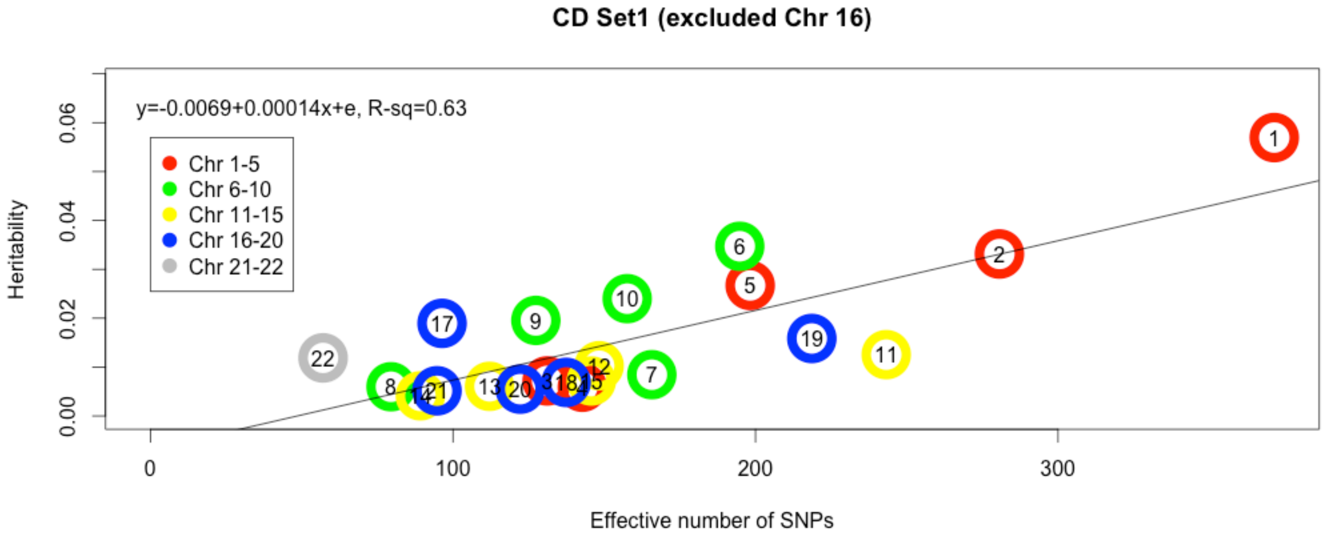




Average Set 1 and Set 2 for CD

As the difference between he estimated chromosomal heritability was not significant, we took the average of the heritability for each chromosome.

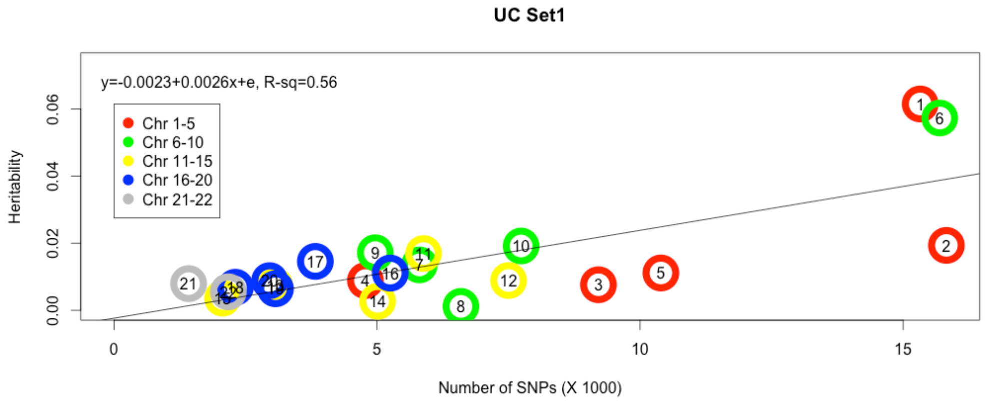


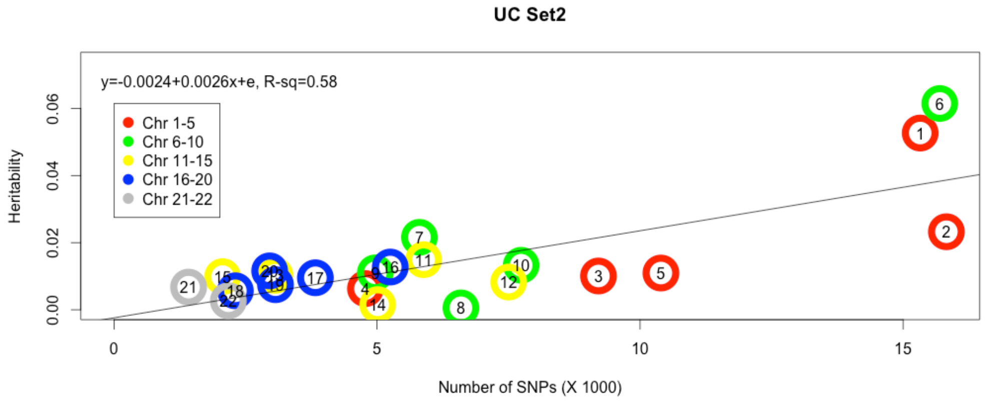


Joint Analysis for CD, regression on effective number of markers

Regression with or without Chr 16 (NOD2)

Joint Analysis for UC





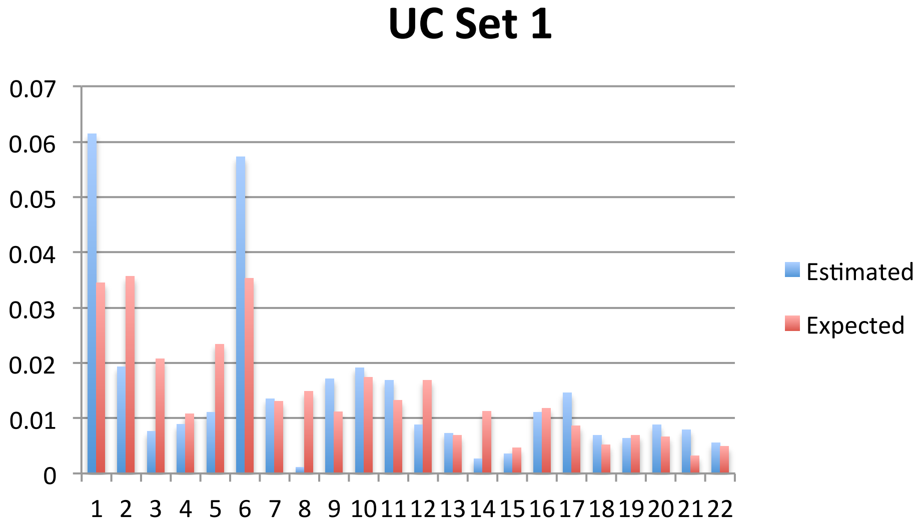
The top 10 PCs were fitted as the covariates. 22 chromosomes were fitted altogether.

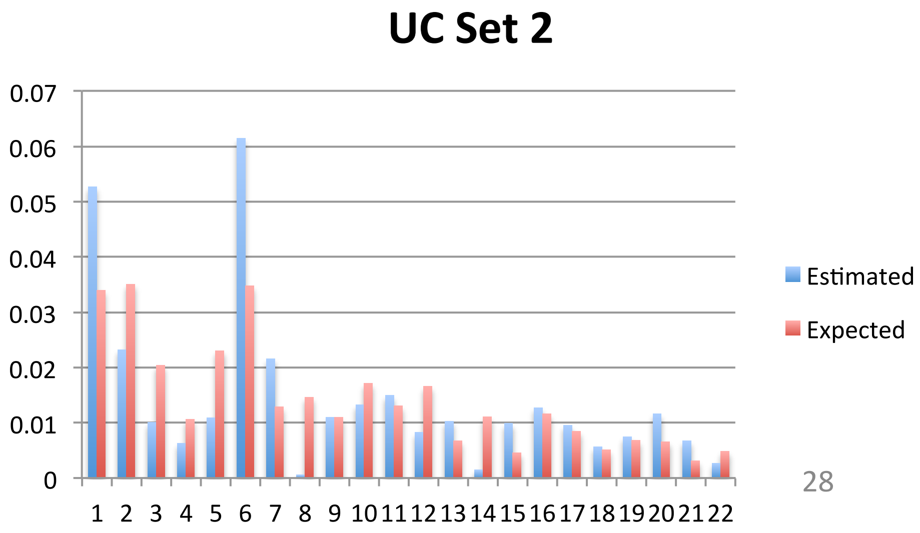
21215 individuals in set 1. The sum of the heritability was 0.317.

21307 individuals in set 2. The sum of the heritability was 0.312.

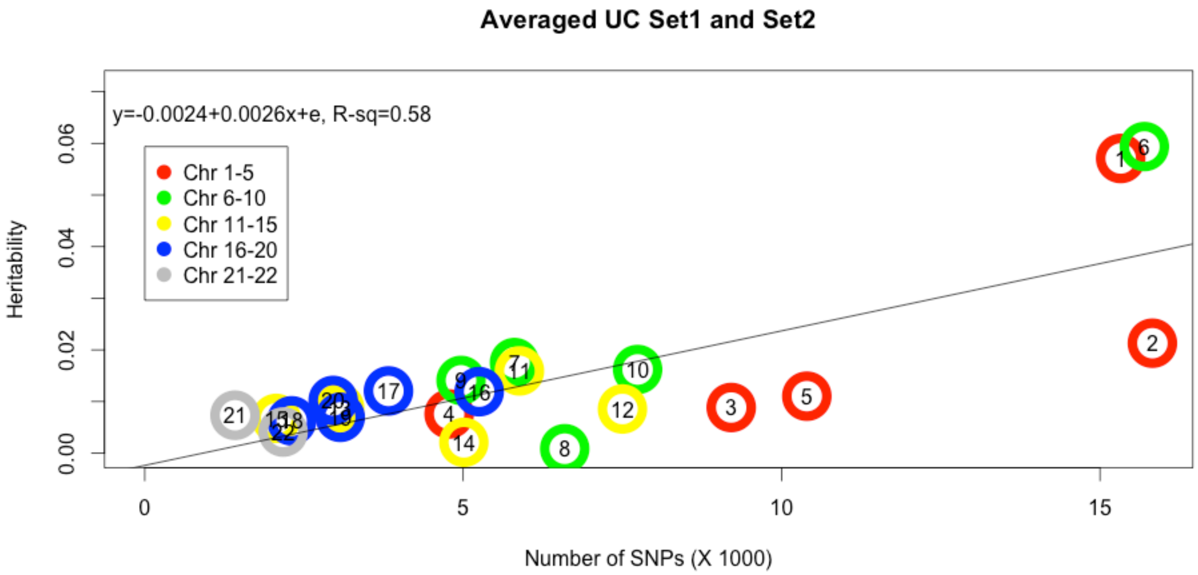
The difference between chromosomal heritability was not significant.

The whole genome grm gave 0.34.

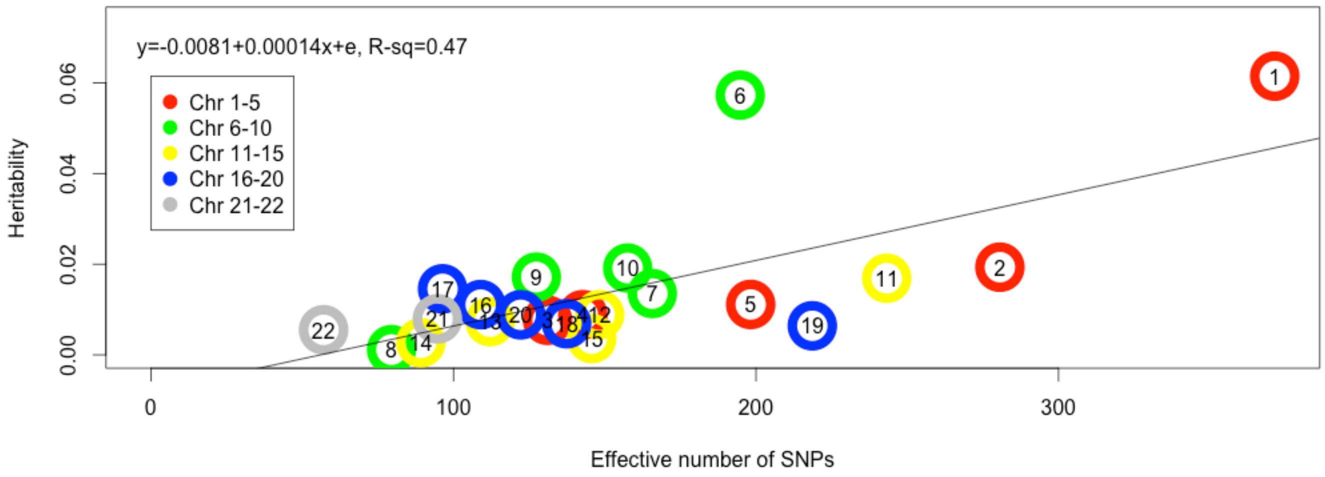




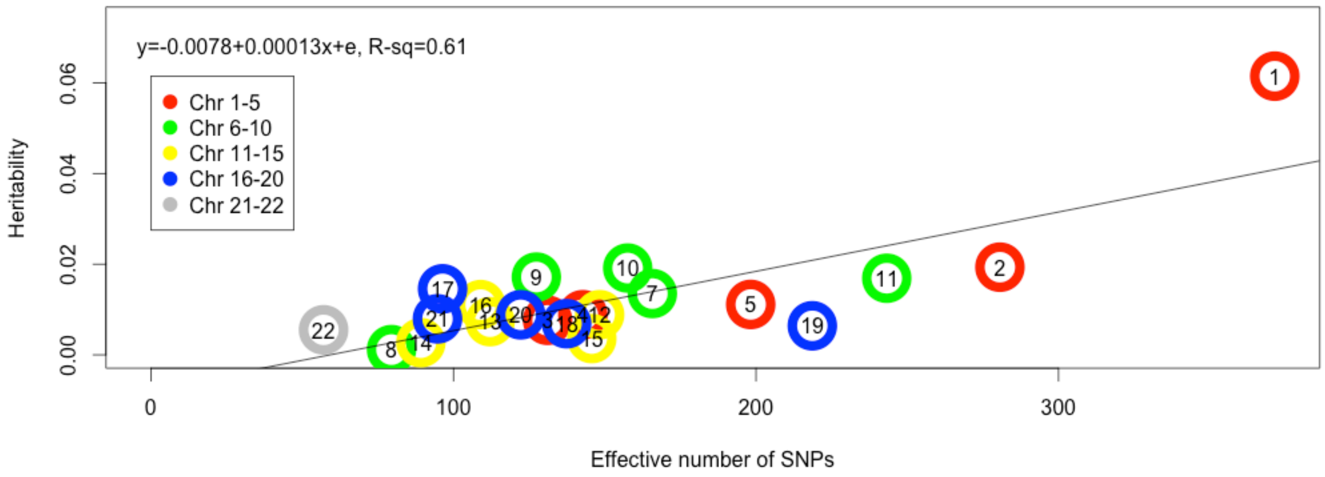
Joint Analysis for CD, regression on effective number of markers



UC Set1



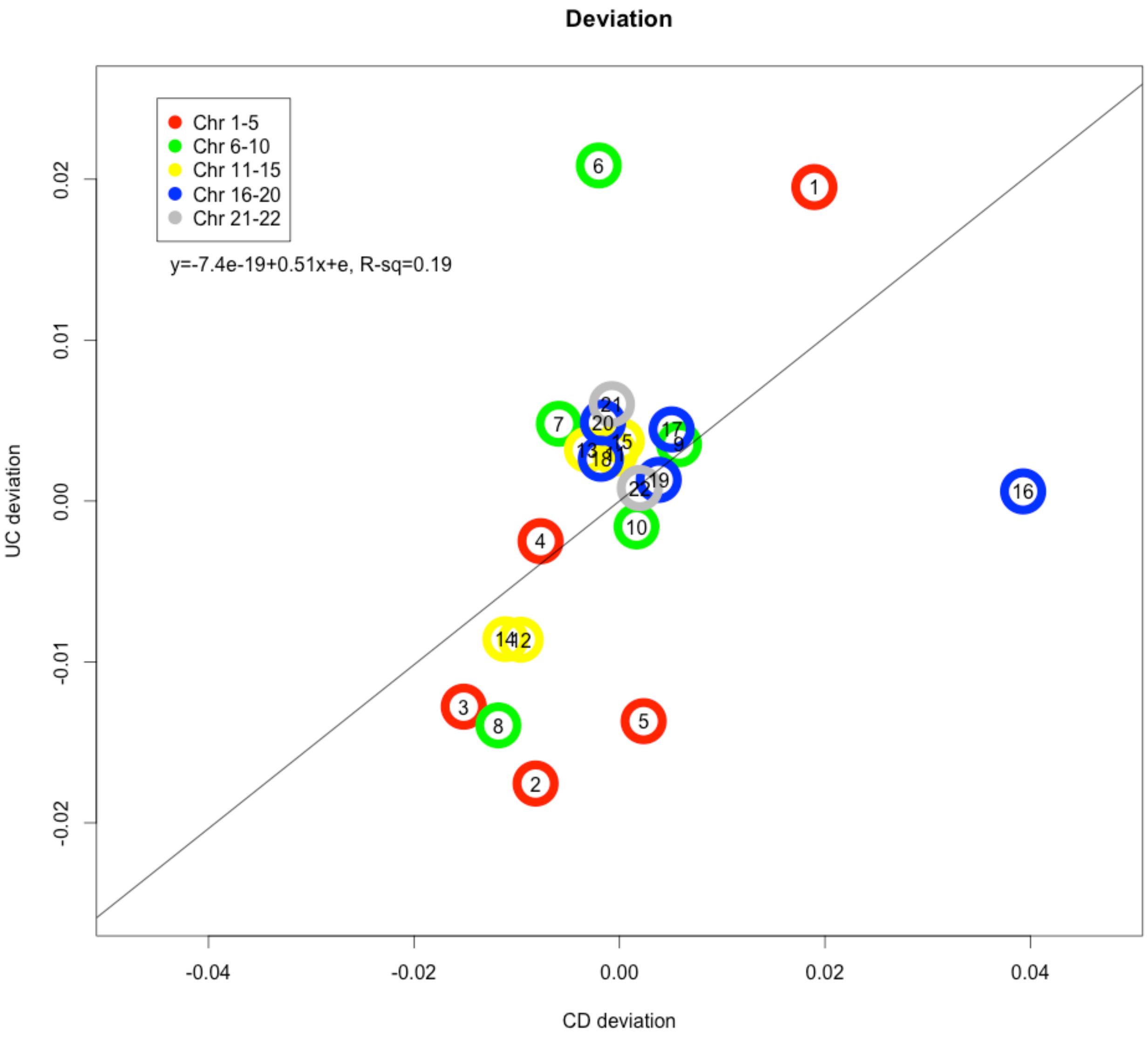
UC Set1 (excluded Chr 16)



Joint Analysis for UC, regression on effective number of markers

With or without Chr 6.

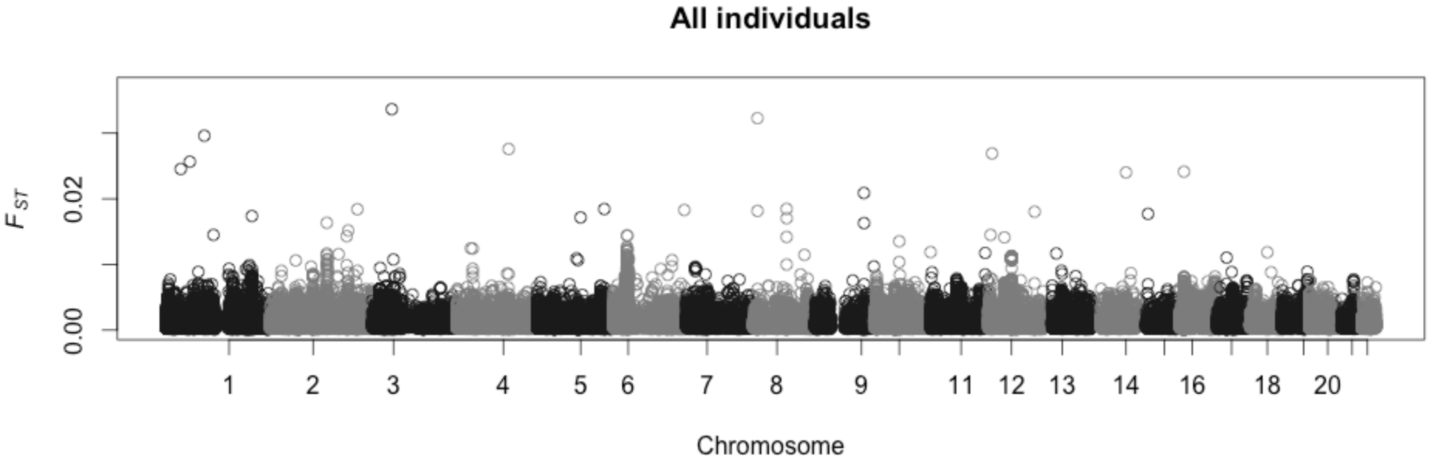
Correlation between deviations



Regressing the mean, over two subsets, of the heritability of each chromosome for UC on that for CD.

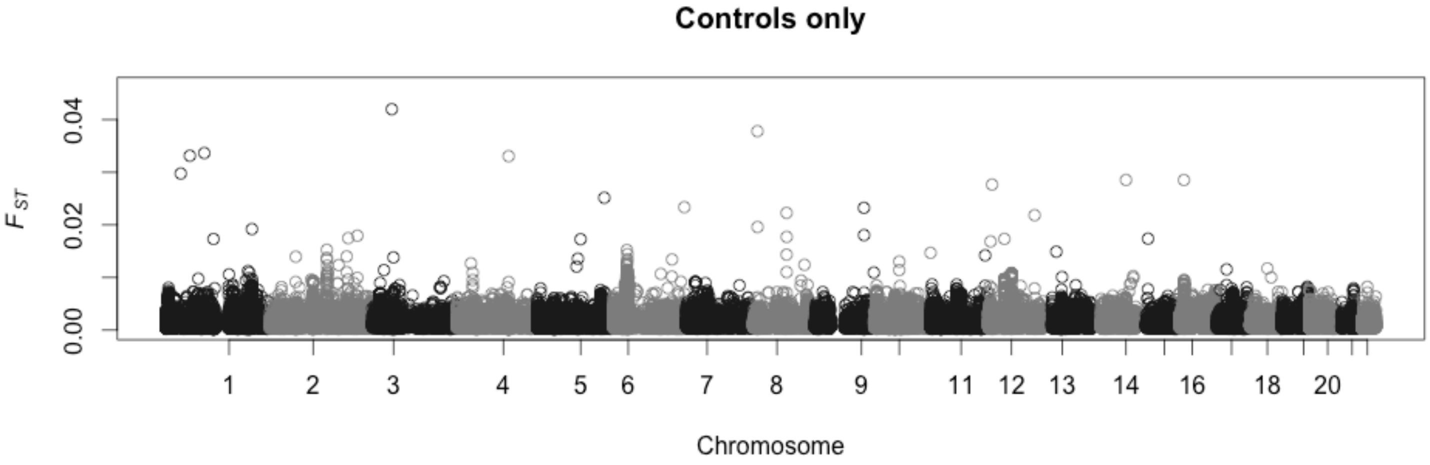
The correlation was about 0.43, whereas Hong got 0.4 from the GWAS data.

F_{ST}



The Fst statistics were calculated for each locus given the data after our in-house QC, MAF cutoff on 0.001.
140853 SNPs.
61518 individual.

mean of Fst = 0.00139
sd of Fst = 0.00118



28154 controls.
mean of Fst=0.00158
sd of Fst=0.00123