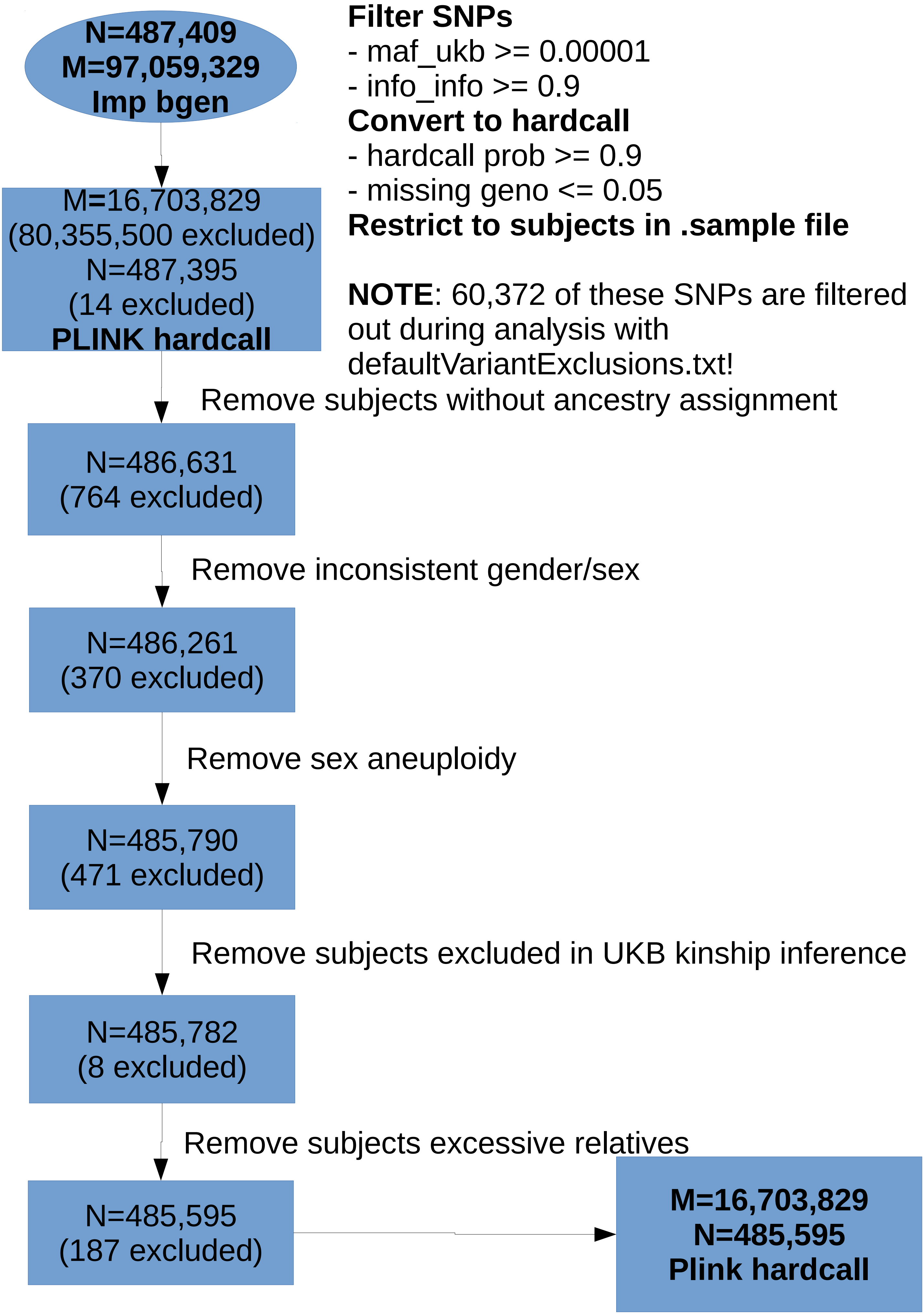


N=487,409
M=97,059,329
Imp bgen

Filter SNPs
- maf_ukb >= 0.00001
- info_info >= 0.9

Convert to hardcall
- hardcall prob >= 0.9
- missing geno <= 0.05

Restrict to subjects in .sample file

NOTE: 60,372 of these SNPs are filtered out during analysis with defaultVariantExclusions.txt!

M=16,703,829
(80,355,500 excluded)
N=487,395
(14 excluded)
PLINK hardcall

Remove subjects without ancestry assignment

N=486,631
(764 excluded)

Remove inconsistent gender/sex

N=486,261
(370 excluded)

Remove sex aneuploidy

N=485,790
(471 excluded)

Remove subjects excluded in UKB kinship inference

N=485,782
(8 excluded)

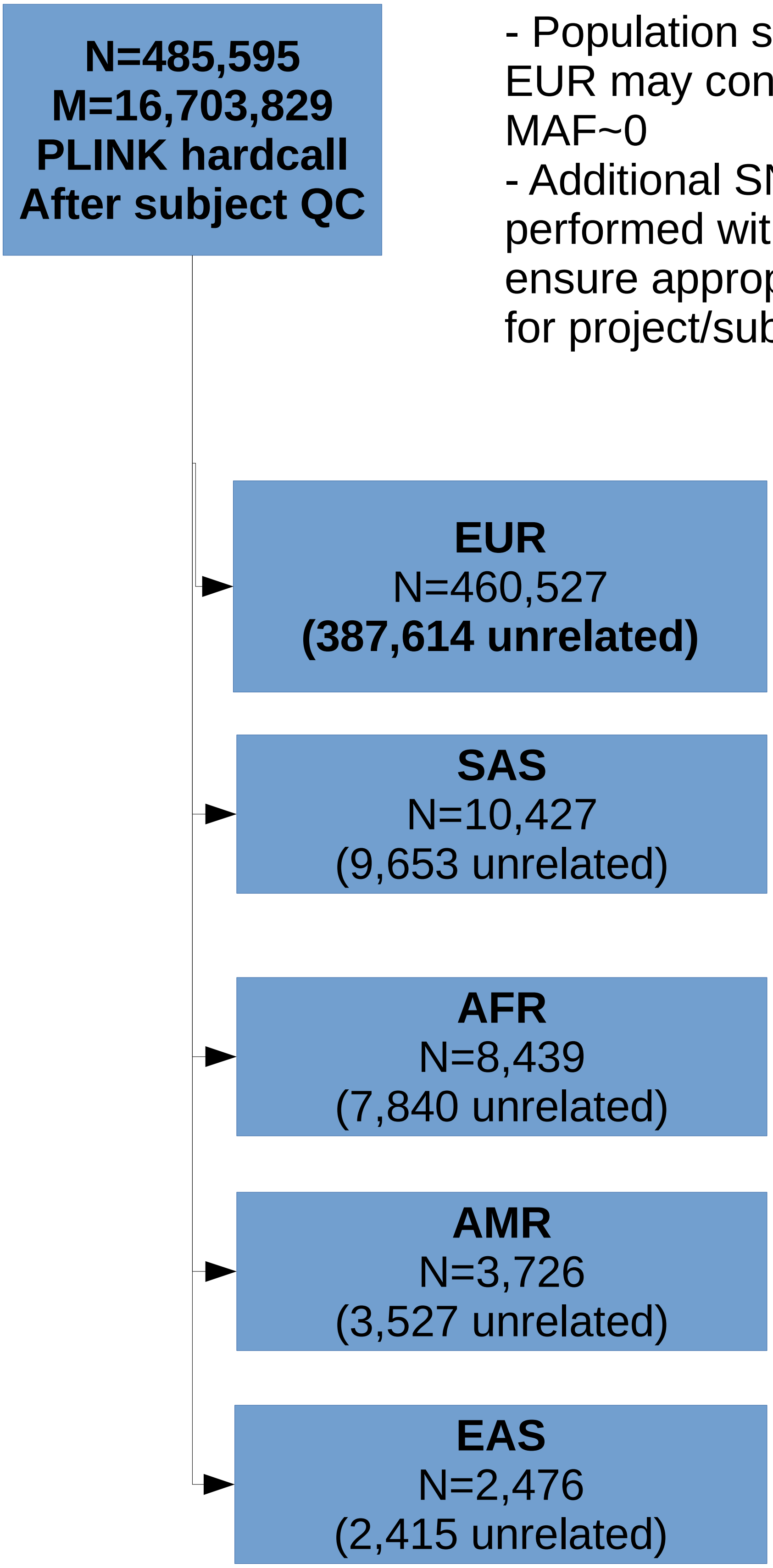
Remove subjects excessive relatives

N=485,595
(187 excluded)

M=16,703,829
N=485,595
Plink hardcall

Split populations (after subject QC)

- M same for all population samples
- Population samples other than EUR may contain SNPs with MAF~0
- Additional SNP filtering is typically performed within each project (to ensure appropriate MAF threshold for project/subsample)



N=485,595
M=16,703,829
PLINK hardcall
After subject QC

EUR
N=460,527
(387,614 unrelated)

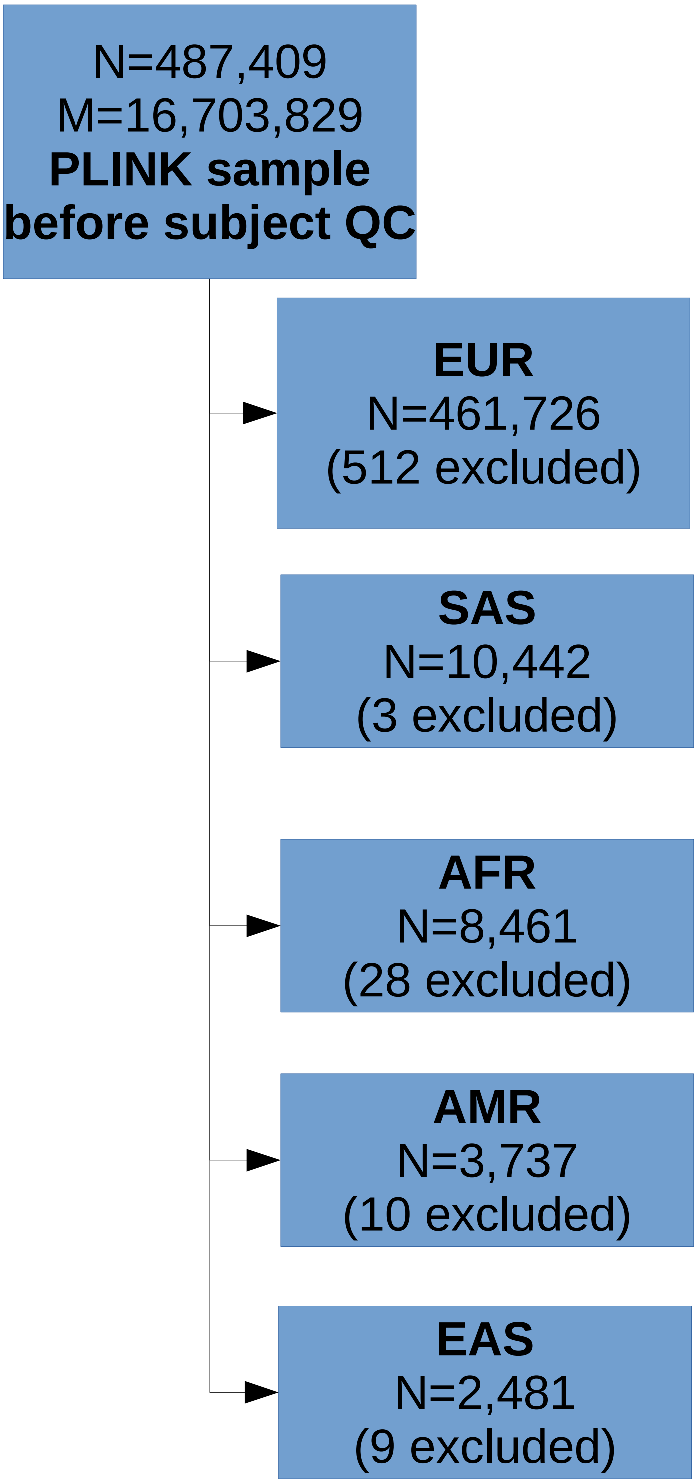
SAS
N=10,427
(9,653 unrelated)

AFR
N=8,439
(7,840 unrelated)

AMR
N=3,726
(3,527 unrelated)

EAS
N=2,476
(2,415 unrelated)

Ancestry assignment procedure



N=487,409
M=16,703,829
PLINK sample before subject QC

EUR
N=461,726
(512 excluded)

SAS
N=10,442
(3 excluded)

AFR
N=8,461
(28 excluded)

AMR
N=3,737
(10 excluded)

EAS
N=2,481
(9 excluded)

Assign 1kg ancestry

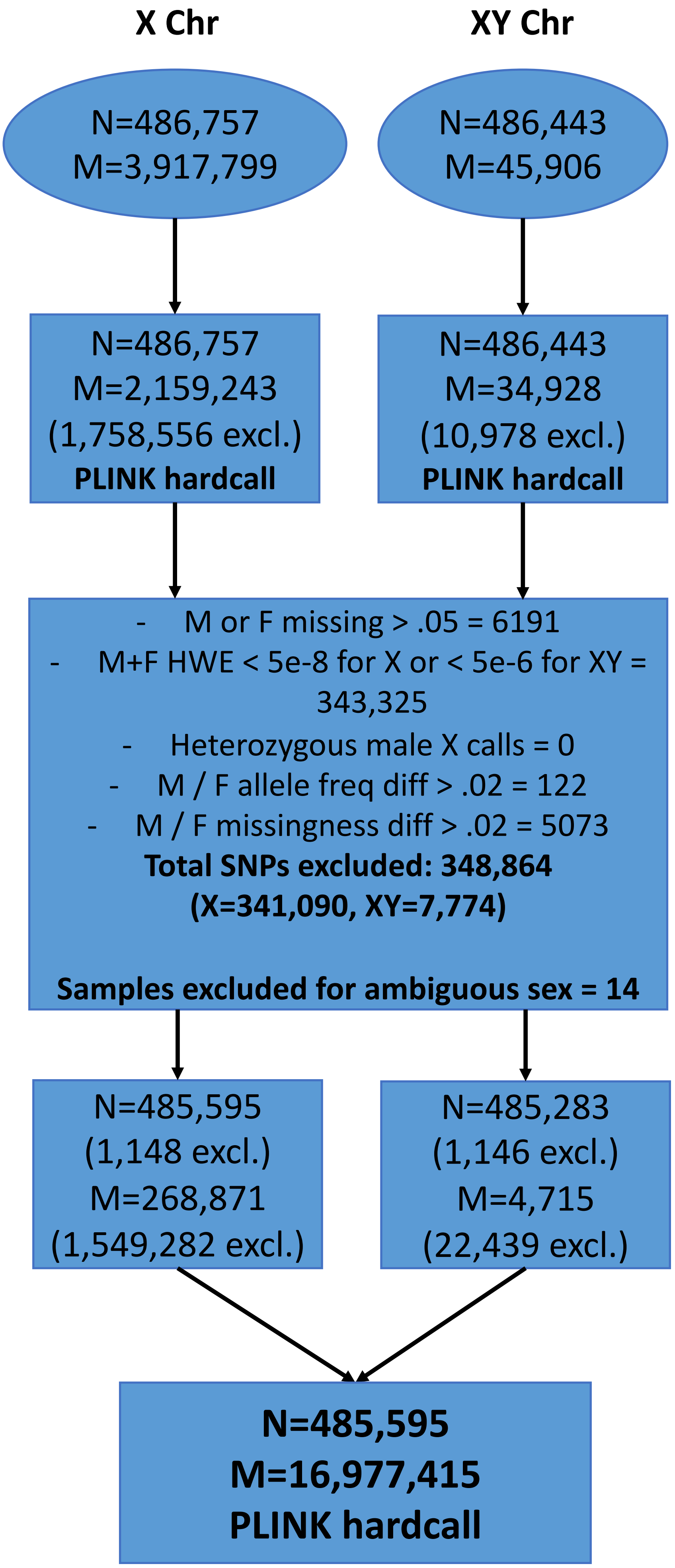
- restrict to ukb-1kgph3 overlapping SNPs (650,232 SNPs)
- apply filters (593,693 SNPs)
  - maf_ukb < 0.001
  - hwe_ukb < 0.000001
  - remove longrange LD regions (Price et al., 2008)
  - remove non-HRC SNPs
  - merge ukb and 1kg and restrict to nonmissing SNPs

-prune SNPs pairwise

- window = 1500kb
- step = 150snps
- r2 = 0.1

(145,692 SNPs)

- compute 30 PCs in 1kg
- project ukb subjects on these PCs
- assign ancestry to closest 1kg subpopulation
- exclude subjects with Mahalanobis distance > 6 S.D. from subpop average



X Chr

XY Chr

N=486,757
M=3,917,799

N=486,443
M=45,906

Filter SNPs

- maf_ukb > .0001

Convert to hardcall

- hardcall prob >= 0.9
- Missing geno < .05

N=486,757
M=2,159,243
(1,758,556 excl.)
PLINK hardcall

N=486,443
M=34,928
(10,978 excl.)
PLINK hardcall

Sex-specific QC

- M or F missing > .05 = 6191
- M+F HWE < 5e-8 for X or < 5e-6 for XY = 343,325
- Heterozygous male X calls = 0
- M / F allele freq diff > .02 = 122
- M / F missingness diff > .02 = 5073

Total SNPs excluded: 348,864
(X=341,090, XY=7,774)

Samples excluded for ambiguous sex = 14

Filter SNPs

- INFO > 0.9
- unique position

Filter samples

- passed autosomal QC

N=485,595
(1,148 excl.)
M=268,871
(1,549,282 excl.)

N=485,283
(1,146 excl.)
M=4,715
(22,439 excl.)

Merge with autosomal genotype files

N=485,595
M=16,977,415
PLINK hardcall