Quality Control

# Parameters

|  |  |
| --- | --- |
| **Parameter** | **Plink2** |
| Autosomes only | Chr1-22 |
| SNPs only (remove multiallelic) | --snps-only |
| ~~Remove duplicates (force-first)~~ | ~~--rm-dup force-first~~ |
| ukbb info score > 0.8 | --extract [IDs with info score > 0.8] |
| call rate > 0.95 (missingness) | --geno 0.05 --mind 0.05 |
| alternate frequency (MAF) > 0.001 && < 0.999 | --maf 0.001 --max-maf 0.999 |
| HWE > 1e-10 | --hwe 1e-10 |
| Remove indels | --exclude [indels marked by “:” in rsid] |
|  |  |
|  |  |

Info score obtained from ukb [resource 1967](https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=1967).



# Log

### 6/1/2021

* Additional QC to remove duplicates and make it SNPs only

### 6/10/2021

* Rm-dup removes all duplicate rsids even if they have different alternative alleles
* Snps-only removes all multiallelic snps
* Remove indels by



* Number of multiallelic SNPs (based off logs) and Number of Indels:

1: 747315 out of 806159 variants; --exclude: 747013 variants remaining.

2: 822332 out of 886385 variants; --exclude: 821943 variants remaining.

3: 696782 out of 751003 variants; --exclude: 696482 variants remaining.

4: 698477 out of 754656 variants; --exclude: 698167 variants remaining.

5: 641525 out of 692257 variants; --exclude: 641247 variants remaining.

6: 646613 out of 700392 variants; --exclude: 646164 variants remaining

7: 568983 out of 613394 variants; --exclude: 568706 variants remaining.

8: 545560 out of 584587 variants; --exclude: 545337 variants remaining.

9: 424097 out of 455496 variants; --exclude: 419599 variants remaining

10: 488070 out of 525362 variants; --exclude: 487815 variants remaining

11: 485763 out of 522796 variants; -exclude: 485554 variants remaining

12: 452383 out of 489413 variants; --exclude: 452196 variants remaining

13: 349568 out of 379125 variants; --exclude: 349414 variants remaining

14: 306790 out of 332046 variants; --exclude: 306639 variants remaining.

15: 263328 out of 283775 variants; --exclude: 263199 variants remaining

16: 297454 out of 317177 variants; --exclude: 297287 variants remaining

17: 249060 out of 268622 variants; --exclude: 248831 variants remaining

18: 272277 out of 293978 variants; --exclude: 272179 variants remaining

19: 209118 out of 225955 variants; --exclude: 208984 variants remaining

20: 213110 out of 228953 variants; --exclude: 213011 variants remaining

21: 127450 out of 137409 variants; --exclude: 127398 variants remaining

22: 125453 out of 134895 variants; --exclude: 125383 variants remaining

Summary:

--snps-only: 9,631,508 out of 10,383,835 variants

Multiallelic SNPs removed: 752,327

--exclude indels: 9,622,548 variants remaining

Indels removed: 8,960

### 9/15/2021

* Ukbb 3rd degree relatives evidence: datafield 22020

<https://www.biorxiv.org/content/10.1101/166298v1.full>

* Ukbb white British – datafield 22006