# Near-independent signals from cojo and clumping

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Run** | **Option** | **cis** | **trans** | **total** | **Comments/location+** |
| **GCTA** |  |
| 1 | LD blocks | 210 | 147 | 357 | only SNPs, cojo/aild-snp/INF1.jma.\*, also doc/INF1.paper.xlsx |
| + indels | LD blocks | 220 | 155 | 375 | SNPs+indels, cojo/aild-indel/INF1.jma.\* |
| 2 | default | 234 | 173 | 407 | –cojo-collinear 0.9 –cojo-wind 10000, doc/SCALLOP\_INF1-260419.xlsx |
| 3 | small R2 & window | 189 | 186 | 375 | –cojo-collinear 0.1 –cojo-wind 500, doc/SCALLOP\_INF1-260419.xlsx |
| **PLINK** |  |
| 4 | LD blocks | 594 | 252 | 846 | only SNPs, clumping/aild-snp/INF1.jma.\*, also doc/INF1.paper.xlsx |
| + indels | LD blocks | 621 | 258 | 879 | SNPs+indels, clumping/aild-indel/INF1.jma.\* |
| 5 | INTERVAL LD panel | 657 | 275 | 932 | –clump-r2 0.1 –clump-kb 500, doc/SCALLOP\_INF1-120419.xlsx |
| 6 | 1000G LD panel | 405 | 229 | 634 | –clump-r2 0.1 –clump-kb 500, clumping/INF1.1KG.r2-0.1.clumped.\* |
| 7 | INTERVAL data | 424 | 188 | 612 | –clump-r2 0.1 –clump-kb 500, doc/SCALLOP\_INF1-120419.xlsx |
| 8 | 1000G LD panel | 402 | 226 | 628 | –clump-r2 0.1 –clump-kb 1000, on tryggve |

+The directories are relative to /scratch/jhz22/INF, i.e., doc/, cojo/ and clumping/, Results in 2 and 3 include regions in high LD excluded in other analyses.

A few observations can be made,

* indels lead to more signals in cojo (1) and clumping (4) analyses.
* **default GCTA –cojo-collinear and –cojo-wind parameters did quite well in numbers, esp. taking ~30 away regions in LD** (1, 2).
* Although it looks close, GCTA –cojo-collinear produces considerably less signals compared to –cojo-collinear 0.9 (3).
* the number of signals increase with the values of GCTA parameters (2, 3), yet moderate changes in LD window have less impact than the reference panel (5, 8).
* PLINK –clump gives more signals than GCTA –cojo (1, 4 and 2, 5).
* **Specification of sliding LD windows disregarding AILD patterns in clumping gives 53 additional signals** (4, 5).
* Thanks to the larger sample size and perhaps greater variant number, INTERVAL as LD reference leads to more signals than 1000Genomes (5, 6).
* Summary statistics from larger sample size gives more signals (5, 7).
* Unpruned results are likely to give more cis signals but this is subject to scrutiny perhaps on individual cases.

It can be concluded that it is desirable to employ approximately independent LD blocks for both GCTA (1) and PLINK (4), and also that reference such as UK10K+1KG would be desirable with respect to both sample size and variant number.