**WTCCC Data Description**

1. **Chiamo SNP**

**SNP marker:** 500568 deleted: 30956 final: 469612

Column for individual; Missing: N; case: 1 control: 0; genotype: 2 characters, separated by tab.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | CHIAMO | deleted | Final sample | 1:2  (male/female) | 1:0(case/control) |
| 58C | 1504 | 24 | 1480 | 741/739 |  |
| NBC | 1500 | 42 | 1458 | 705/753 |  |
| RA | 1999 | 139 | 1860 |  | 1860/2938 |
| CAD | 1988 | 62 | 1926 |  | 1926/2938 |
| T2D | 1999 | 75 | 1924 |  | 1924/2938 |
| Total | 8990 | 342 | 8648 |  |  |

Samples are completely the same as WTCCC paper; 31011 SNP markers deleted in paper, 55 more than ours.

Chiamo\_data format:

1. Original data:

Each file is presented in tab-delimited format and contains one genotype per line. Regardless of the way the SNPs are organized, all assays are sorted according to SNPs so that the file can be readily separated into sample blocks. The following is a brief example of the genotype data format:

SNP SAMPLE GENOTYPE SCORE

rs1234567 WTCCC12345 CC 0.9262

rs1234567 WTCCC12346 TC 0.8650

rs1234567 WTCCC12347 CC 0.9117

1. Reorganize the original data to matrix format:

Output each chromosome in a file. For the Chiamo data, the recommended probabiliy threshold for inclusion is 0.9 and above. If the score<0.9, output missing NN

SNP SAMPLE1 SAMPLE2 … SAMPLE n

rs1234567 CC TC … TC

rs1234568 AA AG … GG

rs1234569 CC CG … NN

1. Combine 23 Chromosome data together

SNP SAMPLE1 SAMPLE2 … SAMPLE n

rs1234567 CC TC … TC

rs1234568 AA AG … GG

rs1234569 CC CG … NN

…

.

.

.

rs234567 CC TC … TC(Chr2)

rs234568 AA AG … GG

rs234569 CC CG … NN

.

.

1. Remove bad samples
2. Combine case-control data and remove bad SNPs at the same time.

Program: A->B: proc\_wtccc\_data(); { proc\_wtccc\_by\_chr\_2(strFile,outFile);}

B ->C: combine\_chr\_CHIAMO();

C ->D: chiamo\_sample\_exclution();

D->E: combine\_CHIAMO\_case\_control();

1. **BRLMM96 SNP**

**SNP marker:** 500568

Row for individual; Missing: N; case: 1 control: 0; genotype: 2 characters

|  |  |  |  |
| --- | --- | --- | --- |
|  | BRLMM96 | Male/female(1:2) | Case/control |
| RA | 1972 |  | 1972/3004 |
| CAD | 1944 |  | 1944/3004 |
| T2D |  |  | 1991/3004 |
| 58C | 1504 | 752/752 |  |
| NBC | 1500 | 720/780 |  |
| Total |  |  |  |

BRLMM96\_data format:

1. Original data:

Each file is presented in tab-delimited format and contains one genotype per line. Regardless of the way the SNPs are organized, all assays are sorted according to sample so that the file can be readily separated into sample blocks. The following is a brief example of the genotype data format:

SNP SAMPLE GENOTYPE SCORE

rs1234567 WTCCC12345 CC 0.9262

rs1234568 WTCCC12345 TC 0.8650

rs1234569 WTCCC12345 CC 0.9117

1. Reorganize the original data to matrix format:

Output each chromosome in a file. For the Chiamo data, the recommended probabiliy threshold for inclusion is 0.5 and below. If the score>0.5, output missing NN

SAMPLE rs1 rs2 … rs n

WTCCC12345 CC AG … CT

WTCCC12346 CG AG … TT

WTCCC12347 GG AG … CT

1. Combine 23 Chromosome data together

SAMPLE rs1 rs2 … rs n … rsk+1 rsk+2 … rs …

WTCCC12345 CC AG … CT …CC AG … AA

WTCCC12346 CG AG … TT … CG AG … TT

WTCCC12347 GG AG … CT … GG AG … AT

…

CHR1…CHR2… CHRX

1. Combine case-control data.

Program: A->B: proc\_wtccc\_data(); { proc\_wtccc\_by\_chr\_2(strFile,outFile); }

B ->C: combine\_chr();{ combine\_2\_chr(strFile1,strFile2,outFile); }

C ->D: combine\_BRLLM96\_case\_control();

1. **Chiamo control SNP :**

Combine 58C and NBC 2938 control, remove all bad samples and bad SNPs; test the sex differential.

Program: wtccc\_control\_sex();