

## Cheeta User Manual: SNP-SNP Interaction Analysis

Cheeta is a GPU-accelerated toolkit for exhaustive genome-wide SNP-SNP interaction analysis. It provides three complementary models to accommodate diverse biological hypotheses: the nine-genotype model, the multifactor dimensionality reduction (MDR) model, and the dominant-recessive model. All methods run on a single consumer-grade GPU and are capable of processing biobank-scale datasets.

### ✧ General Options

The following options are common to all subcommands and are optional.

Option	Description	Default
-threads	Number of threads per GPU block. Usually left at default.	256
-set_gpu	GPU device ID to use (useful for multi-GPU systems).	0

### 1. Nine-Genotype Model (`genotype_interaction`)

This model exhaustively evaluates all nine possible joint genotype combinations of two SNPs. For each combination, a 2×2 contingency table is constructed and tested for association with case/control status.

#### (1) Command

```
cheeta genotype_interaction -file0 <case_file> -file1 <control_file> -o <output_file> [-threads <num>] [-alpha_cut <p_value>] [-or_cut <or_threshold>] [-set_gpu <id>]
```

#### (2) Parameters

Parameter	Required	Description
-file0	Yes	Path to the case genotype file. Rows: SNPs, columns: individuals. Values: 0 (AA), 1 (Aa), 2 (aa), 9 (missing). Tab-separated.
-file1	Yes	Path to the control genotype file. Same format as -file0.
-o	Yes	Output file path.
-alpha_cut	No	Significance threshold for the chi-square p-value (default: 1e-5)
-or_cut	No	Odds ratio confidence interval filter. Only pairs whose OR confidence interval lies entirely outside $[1/\theta, \theta]$ are reported (default: 2.0).

#### (3) Output Columns

```
Genotype_Label SNP0 SNP1 a b c d chi_square chi_pvalue OR  
OR_lower OR_upper
```

Title	Description
Genotype_Label	One of the nine combinations, e.g., AA*BB_vs_other.
SNP0, SNP1	Zero-based indices of the two SNPs in the input file.
a	Case count with the target genotype combination.
b	Case count with any other combination.
c	Control count with the target combination.
d	Control count with any other combination.
chi_square	Yates-corrected chi-square statistic.
chi_pvalue	P-value from the chi-square test.
OR	Point estimate of the odds ratio.
OR_lower, OR_upper	95% confidence interval of the OR.

#### (4) Example

```
cheeta genotype_interaction -file0 cases.txt -file1 controls.txt -o nine_genotype_result.txt
-alpha_cut 1e-6 -or_cut 2.0
```

## 2. Multifactor Dimensionality Reduction (MDR) Model (mdr\_interaction)

This model collapses the nine genotype combinations into two risk categories (high-risk vs. low-risk) using a sample-size correction factor, then performs a single statistical test per SNP pair.

### (1) Command

```
cheeta mdr_interaction -file0 <case_file> -file1 <control_file> -o <output_file> [-threads
<num>] [-alpha_cut <p_value>] [-or_cut <or_threshold>] [-set_gpu <id>]
```

### (2) Parameters

Parameter	Required	Description
-file0	Yes	Path to the case genotype file. Rows: SNPs, columns: individuals. Values: 0 (AA), 1 (Aa), 2 (aa), 9 (missing). Tab-separated.
-file1	Yes	Path to the control genotype file. Same format as -file0.
-o	Yes	Output file path.
-alpha_cut	No	Significance threshold for the chi-square p-value (default: 1e-5)
-or_cut	No	Odds ratio confidence interval filter. Only pairs whose OR confidence interval lies entirely outside $[1/\theta, \theta]$ are reported (default: 2.0).

### (3) Output Columns

```
Model SNP0 SNP1 a b c d chi_square chi_pvalue OR OR_lower
OR_upper
```

- Model: Always high\_risk\_vs\_low\_risk.
- Other columns have the same meaning as in the nine-genotype model, but a and c are the pooled counts of high-risk combinations, while b and d are the pooled counts of low-risk combinations.

### (4) Example

```
cheeta mdr_interaction -file0 cases.txt -file1 controls.txt -o mdr_result.txt -alpha_cut 1e-5
-or_cut 1.5
```

## 3. Dominant-Recessive Model (domrec\_interaction)

This model implements four classical Mendelian inheritance patterns based on the reference alleles: Dominant-Dominant (DD), Dominant-Recessive (DR), Recessive-Dominant (RD), and Recessive-Recessive (RR). For each pattern, a single 2×2 table is tested per SNP pair.

### (1) Command

```
cheeta domrec_interaction -file0 <case_file> -file1 <control_file> -o <output_file> [-threads
<num>] [-alpha_cut <p_value>] [-or_cut <or_threshold>] [-set_gpu <id>]
```

### (2) Parameters

---

Parameter	Required	Description
-file0	Yes	Path to the case genotype file. Rows: SNPs, columns: individuals. Values: 0 (AA), 1 (Aa), 2 (aa), 9 (missing). Tab-separated.
-file1	Yes	Path to the control genotype file. Same format as -file0.
-o	Yes	Output file path.
-alpha_cut	No	Significance threshold for the chi-square p-value (default: 1e-5)
-or_cut	No	Odds ratio confidence interval filter. Only pairs whose OR confidence interval lies entirely outside $[1/\theta, \theta]$ are reported (default: 2.0).

---

### (3) Output Columns

```
Model SNP0 SNP1 a b c d chi_square chi_pvalue OR OR_lower
OR_upper
```

- Model: One of the four patterns, e.g., (AA+Aa)\*(BB+Bb)\_vs\_other (DD).
- Other columns as defined above.

### (4) Example

```
cheeta domrec_interaction -file0 cases.txt -file1 controls.txt -o domrec_result.txt -alpha_cut
```

1e-5 -or\_cut 2.0

## ❖ **Important Notes**

### **1. Input file format**

- Rows = SNPs, columns = individuals.
- Values: 0 (homozygous reference, AA), 1 (heterozygous, Aa), 2 (homozygous alternative, aa), 9 (missing).
- Tab-separated, no header row.
- The number of SNPs (rows) must be identical in both -file0 and -file1. The number of samples (columns) may differ; the algorithm automatically handles imbalanced case/control sizes.

### **2. Missing data handling**

- Individuals with missing genotype (9) are excluded from the analysis for that specific SNP pair. No imputation is performed.

### **3. Output files**

- Results are sorted by SNP indices and then by model/combination.
- Only results that pass both the chi-square and OR-confidence-interval filters are written.
- A log file (same base name as the output file with .log extension) is automatically generated, containing runtime statistics, memory usage, and group partitioning details.

### **4. Performance tips**

- Use the default -threads unless you have specific hardware tuning needs.
- For very large datasets (e.g., >500k SNPs), consider using a GPU with larger memory to reduce the number of groups.
- The first run may include CUDA kernel compilation overhead; subsequent runs are faster.

### **5. Getting Help**

- To display a brief help message listing all available commands:

`cheeta -h`