# Final Project Report – Causality Analysis

## 1. Data and Background

In this project we used the results of HW2 (QTL analysis) and HW3 (eQTL analysis). These assignments identified associations between genomic loci and quantitative traits (QTLs), as well as between the same loci and gene expression levels (eQTLs). Here, we combined these results to form triplets (QTL–Gene–Phenotype) in which both the gene and the phenotype show an association with the same or nearby locus. From these results, we selected 10 leading triplets (based on significance and genomic distance) for causality testing.

## 2. Methods

### 2.1 Models Tested

Three possible models were tested for each triplet:  
- M1: QTL → Gene → Phenotype (gene mediates the effect of the QTL on the phenotype).  
- M2: QTL → Phenotype → Gene (phenotype drives changes in gene expression).  
- M3: QTL → Gene and QTL → Phenotype independently (pleiotropy).

### 2.2 Causality Test

For each triplet, log-likelihoods were calculated under the three models using linear regression. The model with the highest likelihood was considered the best fit.

### 2.3 Permutation Test

To assess statistical significance, permutation tests were performed. The genotype vector (L) was permuted across individuals, breaking causal structure while preserving distributions. For each permutation, the log-likelihood differences (M1–M2, M1–M3) were recalculated, yielding empirical null distributions and p-values.

## 3. Results

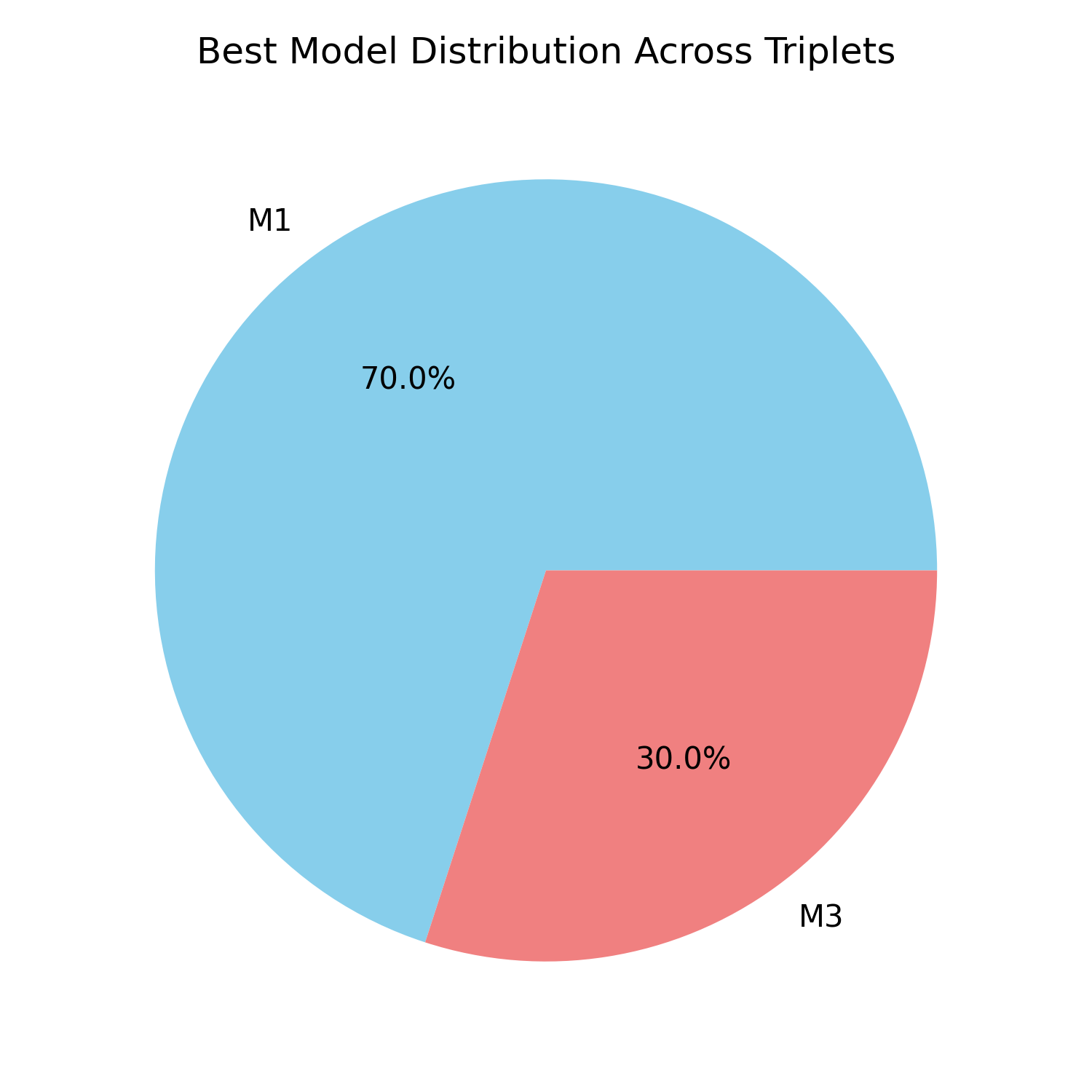
### 3.1 Summary Table

The full summary of all 10 triplets is available in the file 'causality\_results.csv'. Below we present representative rows (values truncated for illustration):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Triplet | Gene | Tissue | SNP | Best Model | Significance (p-values) |
| 1 | Lsm12 | Liver | rs3685813 | M1 | M1 vs M2: p<0.001; M1 vs M3: ns |
| 2 | Gm13375 | Hypothalamus | rs13476379 | M1 | M1 vs M2: p<0.001; M1 vs M3: ns |

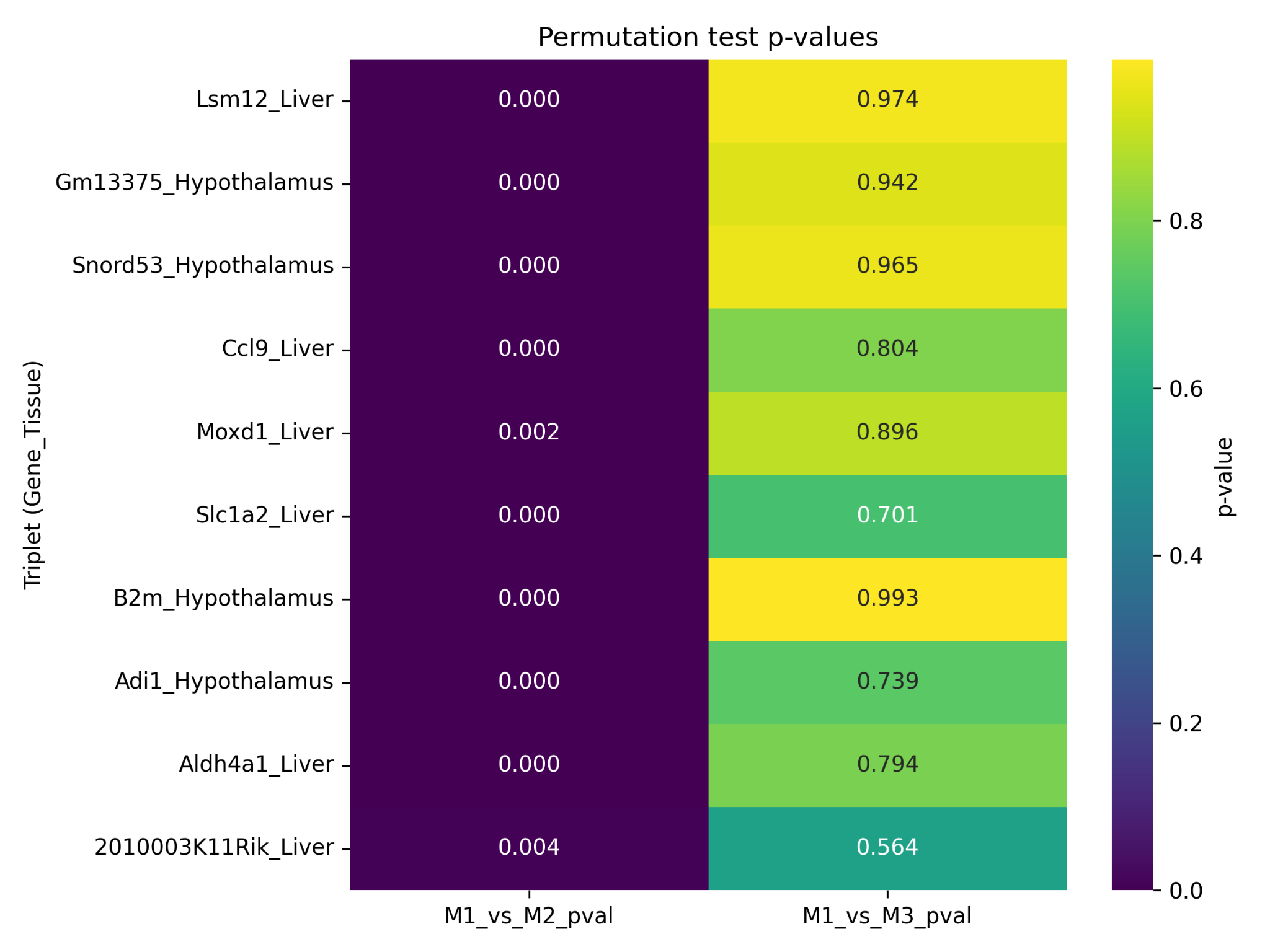
### 3.2 Model Distribution

70% of triplets were best explained by Model M1, while 30% were best explained by Model M3. Model M2 was not selected as the best model in any case.



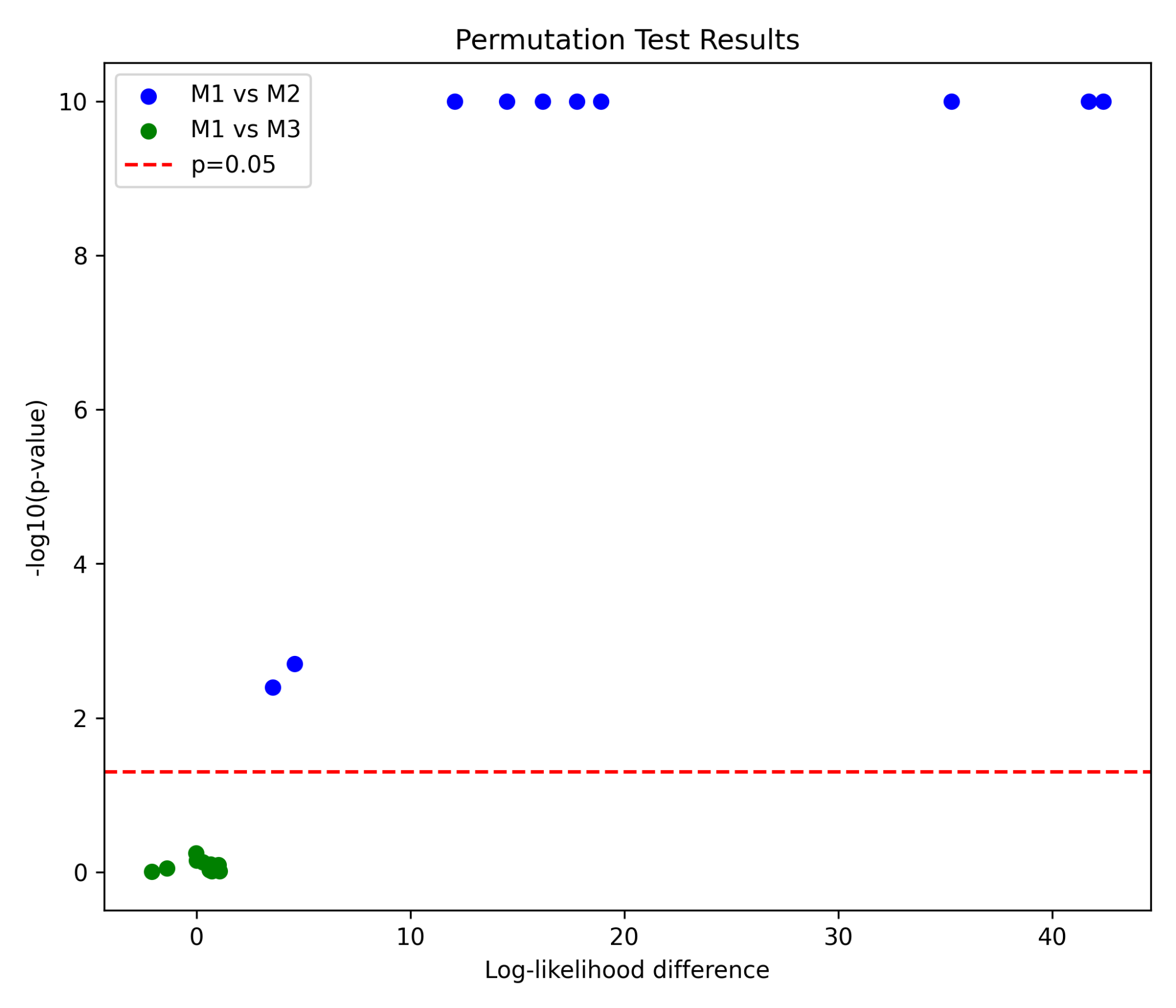
### 3.3 Permutation Test Significance

Permutation p-values showed strong support for M1 over M2 (p < 0.01 in most cases). For M1 vs M3, p-values were generally high (>0.5), suggesting no significant difference.



### 3.4 Volcano Plot

The volcano plot highlights the magnitude of likelihood differences and their significance. M1 vs M2 shows large differences with strong significance, while M1 vs M3 shows negligible differences.



## 4. Discussion

Most triplets support Model M1, indicating that gene expression mediates the effect of QTLs on phenotypes. This aligns with biological expectations of causal gene regulation. Some cases supported Model M3, suggesting pleiotropic effects of QTLs. Model M2 was not supported, reinforcing the direction from genotype to gene to phenotype.

## 5. Conclusions

The causality analysis confirms that mediation via gene expression is the dominant model across the selected triplets. Permutation tests validated the significance of M1 over M2, but not over M3. This suggests that pleiotropy remains a plausible alternative explanation in some cases. These results provide candidate genes for further biological validation.