QTL Analysis Pipeline - Complete Pipeline & Output Documentation

Table of Contents

1. Pipeline Overview
2. Complete Workflow
3. Step-by-Step Pipeline Process
4. Output Structure
5. File Formats and Interpretation
6. Results Interpretation
7. Quality Control Outputs
8. Advanced Analysis Outputs
9. Troubleshooting Outputs
10. Best Practices for Results Analysis

Pipeline Overview

What the Pipeline Does

The QTL Analysis Pipeline performs comprehensive genetic association analysis to identify variants that influence molecular traits:

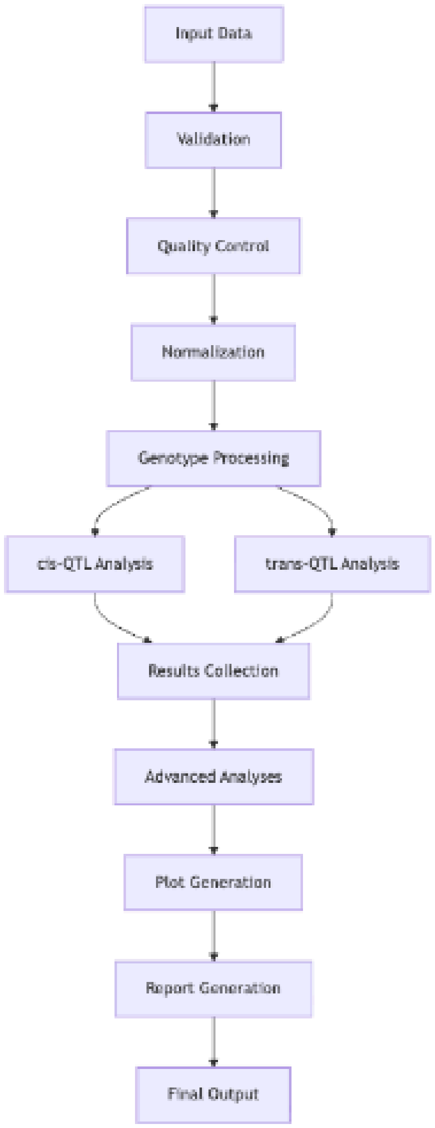
- cis-QTL Analysis: Tests variants within 1Mb of each gene
- trans-QTL Analysis: Tests variants across the entire genome
- Multi-omics Integration: Simultaneous analysis of expression, protein, and splicing
- Quality Control: Comprehensive data quality assessment
- Advanced Analyses: Interaction testing, fine-mapping, and visualization

Pipeline Architecture

```
text
```

```
Input Data → Validation → QC → Normalization → QTL Mapping → Results → Reports
     ↓           ↓         ↓         ↓            ↓           ↓        ↓
   VCF/BCF    Format    Sample    Data       Statistical  Summary  Interactive
   Phenotypes Checks    QC        Transformation  Tests    Files    Reports
   Covariates          Variant QC              Association
   Annotations         PCA                    Testing
```

Complete Workflow



High-Level Pipeline Flowchart

Runtime Expectations

Analysis Type	Sample Size	Variant Count	Expected Runtime	Memory Usage
cis-eQTL	100 samples	1M variants	1-2 hours	8-16GB
cis-eQTL	500 samples	5M variants	4-8 hours	16-32GB
trans-eQTL	500 samples	5M variants	24-48 hours	32-64GB
Multi-omics	1000 samples	10M variants	2-3 days	64-128GB

Step-by-Step Pipeline Process

Step 1: Input Validation

Purpose: Ensure all input files are correct and compatible

Process:

- Checks file existence and permissions
- Validates file formats (VCF, BED, TSV)
- Verifies sample concordance across files
- Checks chromosome naming consistency
- Validates configuration parameters

Key Outputs:

- Validation report in logs

- Sample concordance summary
- Format compatibility check

Step 2: Quality Control

Purpose: Identify and filter low-quality data

Process:

- Sample-level QC:
    - Missingness rate calculation
    - Heterozygosity analysis
    - Sex check validation
    - Relatedness detection
- Variant-level QC:
    - Missingness rate
    - Hardy-Weinberg Equilibrium
    - Minor Allele Frequency
    - Call rate thresholds
- Phenotype QC:
    - Missing value analysis
    - Outlier detection
    - Distribution assessment

Key Outputs:

- QC summary reports
- Filtered genotype data
- Sample and variant exclusion lists

Step 3: Data Normalization

Purpose: Transform data to meet statistical assumptions

Process:

- eQTL Normalization:
    - VST (Variance Stabilizing Transformation)
    - Log2 transformation

    - Quantile normalization
    - TPM normalization
- pQTL Normalization:
    - Log2 transformation with pseudocount
    - Z-score standardization
    - Quantile normalization
- sQTL Normalization:
    - Log2 transformation for PSI values
    - Arcsinh transformation
    - Z-score standardization

Key Outputs:

- Normalized phenotype files
- Normalization comparison plots
- Transformation parameters

Step 4: Genotype Processing

Purpose: Prepare genotype data for efficient analysis

Process:

- Format conversion (VCF → PLINK/BCF)
- Chromosome normalization
- Variant filtering (MAF, missingness, HWE)
- Multi-allelic site handling
- Sample matching across datasets

Key Outputs:

- Processed genotype files
- Filtering statistics
- Format-converted data

Step 5: QTL Mapping

Purpose: Perform statistical association testing

Process:

- cis-QTL Analysis:
    - Linear regression for variant-gene pairs
    - Permutation testing for FDR calculation
    - Window-based testing (default: 1Mb)
- trans-QTL Analysis:
    - Genome-wide association testing
    - Multiple testing correction
    - Conditional analysis

Statistical Model:

```
text

Phenotype ~ Genotype + Covariates + ε
```

Where:

- Phenotype: Normalized molecular trait
- Genotype: Genetic variant dosage
- Covariates: Technical and biological confounders
- ε: Error term

Key Outputs:

- Association statistics
- Nominal p-values
- FDR-corrected results
- Effect size estimates

Step 6: Advanced Analyses

Purpose: Provide deeper biological insights

Process:

- Interaction Analysis: Test for context-specific effects
- Fine-mapping: Identify causal variants
- Conditional Analysis: Independent signal detection
- Pathway Enrichment: Biological context interpretation

Key Outputs:

- Interaction results
- Credible sets
- Conditional association statistics
- Pathway analysis results

Step 7: Visualization and Reporting

Purpose: Generate interpretable results and summaries

Process:

- Manhattan plots for genome-wide results
- QQ plots for inflation assessment
- Volcano plots for effect size visualization
- Locus zoom plots for regional association
- Interactive HTML reports

Key Outputs:

- Static plots (PNG/PDF)
- Interactive plots (HTML)
- Comprehensive reports
- Summary statistics

Output Structure

Complete Directory Tree

```
text

results/
├── QTL_results/                   # Main QTL analysis results
│   ├── eqtl/                     # Expression QTL results
│   │   ├── cis/                  # cis-eQTL results
│   │   │   ├── nominals.txt     # All association results
│   │   │   ├── significant.txt  # FDR-significant results
│   │   │   ├── permutations/    # Permutation results
│   │   │   └── summary.txt      # Analysis summary
│   │   └── trans/               # trans-eQTL results
│   │       ├── nominals.txt
│   │       ├── significant.txt
│   │       └── summary.txt
│   ├── pqtl/                     # Protein QTL results
│   └── sqtl/                     # Splicing QTL results
├── GWAS_results/                  # GWAS results (if enabled)
│   ├── gwas_combined_results.txt
│   ├── individual_phenotypes/    # Per-phenotype results
│   └── qc_report.txt
├── QC_reports/                    # Quality control outputs
│   ├── genotype_qc/
│   │   ├── sample_missingness.png
│   │   ├── maf_distribution.png
│   │   ├── heterozygosity.png
│   │   └── hwe_violations.txt
│   ├── phenotype_qc/
│   │   ├── expression_qc.png
│   │   ├── protein_qc.png
│   │   └── splicing_qc.png
│   ├── sample_concordance/
│   │   ├── overlap_summary.txt
│   │   └── concordance_plot.png
│   └── comprehensive_qc_report.html
├── plots/                         # All generated visualizations
│   ├── manhattan/                # Manhattan plots
│   │   ├── eqtl_cis_manhattan.png
│   │   ├── eqtl_trans_manhattan.png
│   │   └── gwas_manhattan.png
│   ├── qq/                       # QQ plots
│   ├── volcano/                  # Volcano plots
│   ├── distribution/             # Distribution plots
│   ├── locuszoom/                # Locus zoom plots
│   ├── interactive/              # Interactive plots (HTML)
│   └── summary/                  # Summary plots
├── reports/                      # Comprehensive reports
│   ├── analysis_report.html      # Main HTML report
│   ├── pipeline_summary.txt      # Text summary
│   ├── results_metadata.json     # Results metadata
│   └── methods_section.txt       # Methods for publications
├── interaction_results/          # Interaction analysis
```

```
│   ├── age_interaction/
│   ├── sex_interaction/
│   └── summary.txt
├── fine_mapping_results/         # Fine-mapping outputs
│   ├── credible_sets/
│   ├── susie_results/
│   └── finemap_results/
├── normalization_comparison/     # Normalization assessment
│   ├── eqtl/
│   │   ├── distribution_comparison.png
│   │   ├── normalization_report.html
│   │   └── statistical_summary.txt
│   ├── pqtl/
│   └── sqtl/
├── genotype_processing/          # Processed genotype data
│   ├── filtered_genotypes.vcf.gz
│   ├── plink_format/
│   └── processing_log.txt
├── temp/                         # Temporary files (cleaned up)
└── logs/                         # Pipeline execution logs
    ├── pipeline_YYYYMMDD_HHMMSS.log
    ├── validation.log
    ├── qc.log
    └── analysis.log
```

File Formats and Interpretation

Main QTL Results Files

1. Nominal Association Results (`nominals.txt`)

Format: Tab-separated values

```text
phenotype_id  variant_id  chromosome  position  p_value  beta  se  maf
gene1         chr1_1000   1           1000      2.5e-08  0.32  0.05 0.15
gene1         chr1_2000   1           2000      1.2e-06  0.25  0.06 0.12
```

Columns:

- `phenotype_id`: Gene/protein/splicing event ID
- `variant_id`: Genetic variant identifier
- `chromosome`, `position`: Genomic coordinates
- `p_value`: Association p-value
- `beta`: Effect size (change in phenotype per additional effect allele)
- `se`: Standard error of effect size
- `maf`: Minor allele frequency

2. Significant Results (`significant.txt`)

Format: Tab-separated values with FDR information

```text
phenotype_id  variant_id  p_value  beta  p_fdr  q_value
gene1         chr1_1000   2.5e-08  0.32  0.001  0.001
gene2         chr2_5000   3.2e-07  0.28  0.015  0.015
```

Additional Columns:

- `p_fdr`: False Discovery Rate adjusted p-value
- `q_value`: Storey's q-value (similar to FDR)

3. Permutation Results (`permutations/`)

Directory containing:

- `permutation_pass_1.txt`: Results from first permutation round
- `permutation_stats.txt`: Summary of permutation distribution
- `empirical_pvalues.txt`: Empirical p-values from permutations

Quality Control Files

1. Sample QC Report

```text
Sample ID    Missing Rate  Heterozygosity  Status
```

```
sample1      0.02          0.32            PASS
sample2      0.15          0.45            FAIL_MISSING
sample3      0.01          0.29            PASS
```

2. Variant QC Report

```
text

Variant ID   Chromosome  Position  MAF  Missing Rate  HWE_P  Status
rs12345      1           1000      0.12 0.01          0.85   PASS
rs67890      1           2000      0.005 0.08         1e-08  FAIL_HWE
```

3. Sample Concordance Report

```
text

Dataset      Total Samples  Overlap Samples  Overlap Percentage
Genotypes    500           -                -
Expression   480           475              95.0%
Covariates   490           485              97.0%
```

Results Interpretation

Key Metrics to Evaluate

1. Genomic Control Lambda (λ)

What it is: Measure of test statistic inflation
Interpretation:

- λ = 1.0: Perfectly calibrated (ideal)
- 1.0 < λ < 1.05: Slight inflation (acceptable)
- λ > 1.05: Significant inflation (potential confounding)
- λ < 1.0: Deflation (rare, may indicate issues)

2. Number of Significant Associations

Expected ranges:

- cis-eQTLs: 10-80% of genes typically have cis-eQTLs
- trans-eQTLs: 1-10% of genes typically have trans-eQTLs
- pQTLs: 5-50% of proteins typically have pQTLs
- sQTLs: 5-30% of splicing events typically have sQTLs

3. Effect Size Distribution

Typical ranges:

- cis-eQTLs: |beta| = 0.1-1.0 (moderate to large effects)
- trans-eQTLs: |beta| = 0.05-0.3 (small to moderate effects)
- pQTLs: |beta| = 0.1-0.8 (moderate effects)
- sQTLs: |beta| = 0.2-1.5 (moderate to large effects)

How to Read Output Files

Example cis-eQTL Result Interpretation

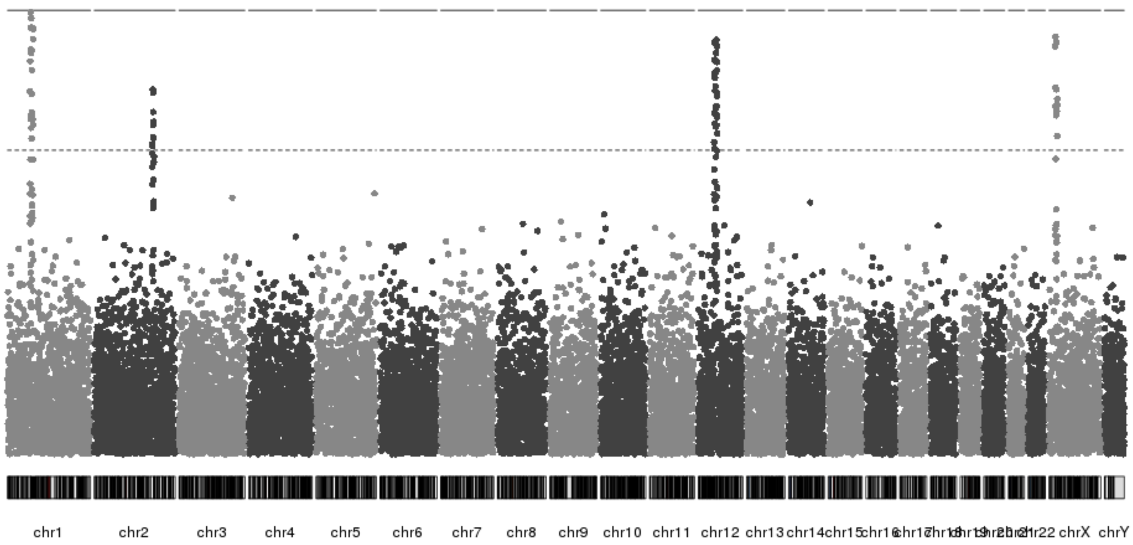
bash

```
# From significant.txt:
# gene1  chr1_1000_A_T  2.5e-08  0.32  0.001

Interpretation:
- Gene 'gene1' has a significant cis-eQTL at variant chr1:1000
- Association p-value: 2.5e-08 (highly significant)
- Effect size (beta): 0.32 → Each effect allele increases expression by 0.32 units

- FDR-adjusted p-value: 0.001 → 0.1% false discovery rate
```

Example Manhattan Plot Interpretation



What to look for:

- Peaks above red line: Genome-wide significant hits (p < 5e-8)
- Peaks above orange line: Suggestive hits (p < 1e-5)
- Chromosome patterns: Should be relatively uniform
- Inflation: Points above diagonal in QQ plot indicate inflation

Quality Control Outputs

Sample QC Assessment

1. Sample Missingness Plot

File: `QC_reports/genotype_qc/sample_missingness.png`
What to check:

- Most samples should have <5% missingness
- Remove samples with >10% missingness
- Look for bimodal distribution indicating batch effects

2. MAF Distribution Plot

File: `QC_reports/genotype_qc/maf_distribution.png`
What to check:

- Should show exponential decay (many rare variants, few common)
- Check MAF threshold is appropriate (vertical line)
- Ensure no unusual peaks or gaps

3. Heterozygosity Plot

File: `QC_reports/genotype_qc/heterozygosity.png`
What to check:

- Most samples should cluster around population mean
- Outliers may indicate sample contamination or issues
- Different clusters may indicate population stratification

Phenotype QC Assessment

1. Expression Distribution

File: `QC_reports/phenotype_qc/expression_qc.png`
What to check:

- Distribution should be smooth without extreme outliers
- Missingness pattern should be random, not systematic
- Batch effects visible as blocks in missingness heatmap

2. Sample Concordance

File: `QC_reports/sample_concordance/concordance_plot.png`
What to check:

- Overlap should be >80% between genotypes and phenotypes
- Low overlap indicates sample ID mismatches
- Investigate samples present in one dataset but not others

Advanced Analysis Outputs

Interaction Analysis Results

1. Interaction Summary

```text
phenotype_id  variant_id  p_nominal  p_interaction  beta_interaction
gene1         chr1_1000   2.5e-08    0.01          0.15
```

Interpretation:

- `p_interaction`: Significance of interaction term
- `beta_interaction`: Effect size of interaction
- Example: The genetic effect on gene1 differs by the interaction covariate

2. Stratified Results

Files: `interaction_results/age_stratified/`

- Contains results split by interaction covariate levels
- Useful for understanding direction of interaction effects

Fine-mapping Results

1. Credible Sets

```text
phenotype_id  variant_id  posterior_probability  credible_set
gene1         chr1_1000   0.45                  1
gene1         chr1_2000   0.35                  1
gene1         chr1_3000   0.15                  1
```

Interpretation:

- `posterior_probability`: Probability variant is causal
- `credible_set`: Set of variants containing causal variant with 95% probability
- Variants in same credible set should be in high LD

2. Fine-mapping Summary

- Number of credible sets per locus
- Size of credible sets (smaller = better resolution)
- Posterior probabilities of top variants

Troubleshooting Outputs

Common Issues and Diagnostic Files

1. Memory Issues

Check: `logs/pipeline_*.log` for memory warnings
Solutions:

- Increase `memory_gb` in configuration
- Enable `process_by_chromosome: true`
- Use `force_plink: true` for large datasets

2. No Significant Results

Diagnostic files to check:

- `QC_reports/comprehensive_qc_report.html` - Data quality issues
- `normalization_comparison/` - Normalization effectiveness
- `plots/qq/` - Test statistic inflation

Common causes:

- Insufficient sample size
- Poor data quality
- Inappropriate normalization
- Overly strict multiple testing correction

3. Long Runtime

Check: `logs/analysis.log` for bottleneck steps
Optimization strategies:

- Increase `num_threads`
- Use `qtl_mode: "cis"` instead of "both"
- Reduce `num_permutations`
- Use BCF instead of VCF format

Log File Interpretation

Example Log Entry Analysis

```text
2024-01-15 10:30:15 - QTLPipeline - INFO - 🔍 Running eQTL cis analysis...
2024-01-15 10:35:22 - QTLPipeline - INFO - ✅ eQTL cis: 1250 significant associations
2024-01-15 10:35:23 - QTLPipeline - WARNING - ⚠️ High genomic inflation detected: λ = 1.12
2024-01-15 10:35:24 - QTLPipeline - ERROR - ❌ trans-eQTL analysis failed: Memory allocation failed
```

Interpretation:

- cis-eQTL analysis completed successfully with 1250 hits
- Genomic inflation suggests potential confounding
- trans-eQTL failed due to memory limits

Best Practices for Results Analysis

1. Start with QC Assessment

- Always examine QC reports before interpreting results
- Check sample and variant filtering thresholds
- Verify normalization effectiveness

2. Validate Key Findings

- Check top hits in external databases (GTEx, eQTL Catalogue)
- Verify effect directions make biological sense
- Consider replication in independent datasets

3. Use Multiple Visualization Types

- Manhattan plots for genome-wide overview
- QQ plots for inflation assessment
- Locus zoom for regional context
- Volcano plots for effect size distribution

4. Consider Biological Context

- Annotate significant hits with known genes
- Check for enrichment in functional categories
- Consider tissue/cell type specificity

5. Document Analysis Decisions

- Keep configuration files for reproducibility
- Document filtering thresholds and normalization methods
- Record any manual curation steps

Example Results Workflow

bash

```
# 1. Check pipeline completed successfully
cat results/pipeline_summary.txt

# 2. Examine QC reports
open results/reports/analysis_report.html

# 3. Check significant hit counts
wc -l results/QTL_results/eqtl/cis/significant.txt

# 4. Generate custom visualizations for top hits
# (Using the provided plotting utilities)

# 5. Export results for downstream analysis
cp results/QTL_results/eqtl/cis/significant.txt my_analysis/top_qtls.txt
```