Performance Analysis of the Haplotyping Algorithm

Executive Summary

The analysis of 4,128 executions reveals that the algorithm achieves a **global efficiency of 30.4%**, solving 1,256 cases without ILP optimization and requiring optimization for 2,872 complex cases. This performance varies significantly according to input data characteristics.

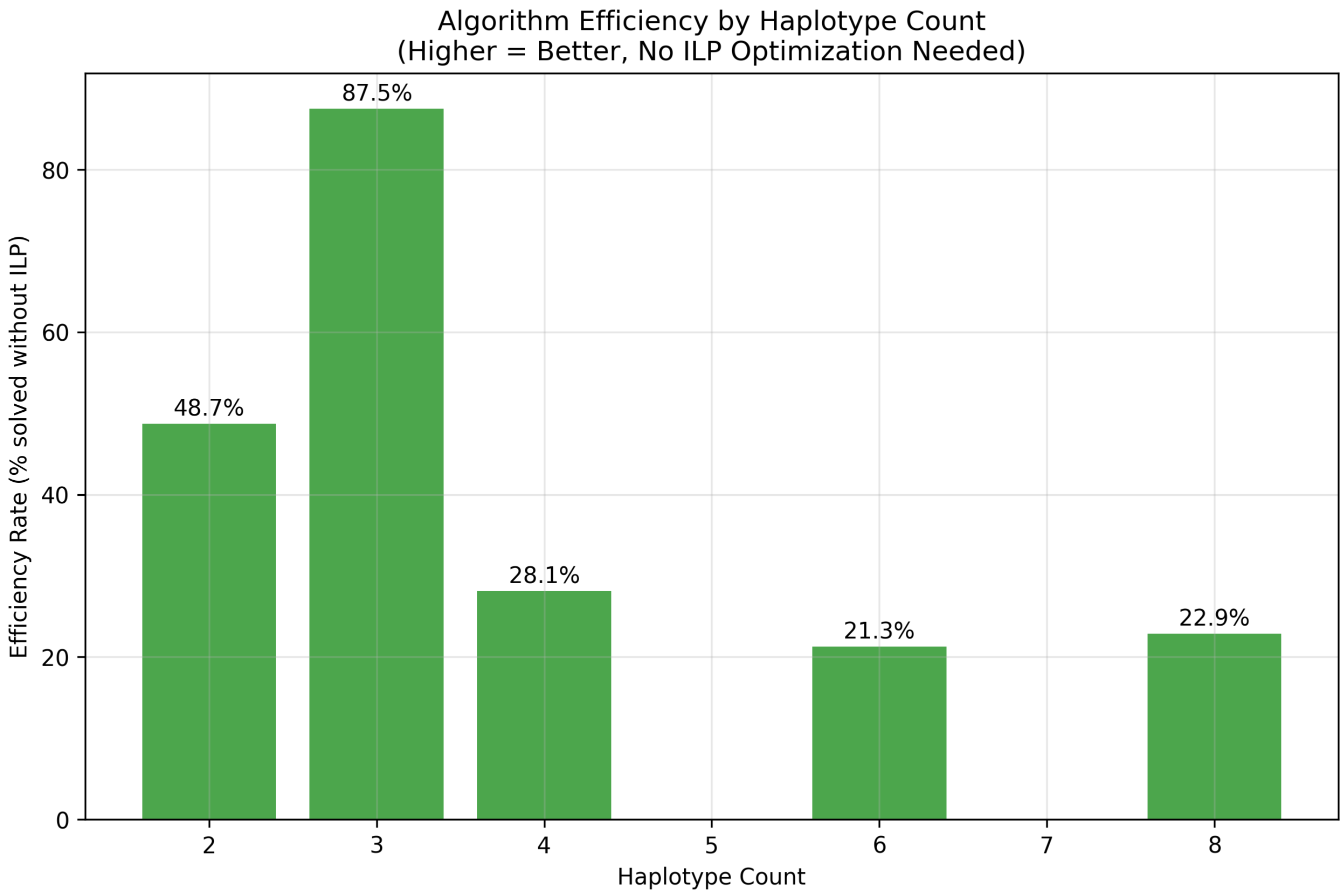
Key Results

1. Efficiency by Haplotype Count

The algorithm shows a **complex relationship** between haplotype count and efficiency:

- **3 haplotypes**: 87.5% efficiency ⚠ **LIMITED SAMPLE** (only 8 cases - not representative)
- **2 haplotypes**: 48.7% efficiency (significant sample)
- **4 haplotypes**: 28.1% efficiency
- **6-8 haplotypes**: ~22% efficiency (constant performance)

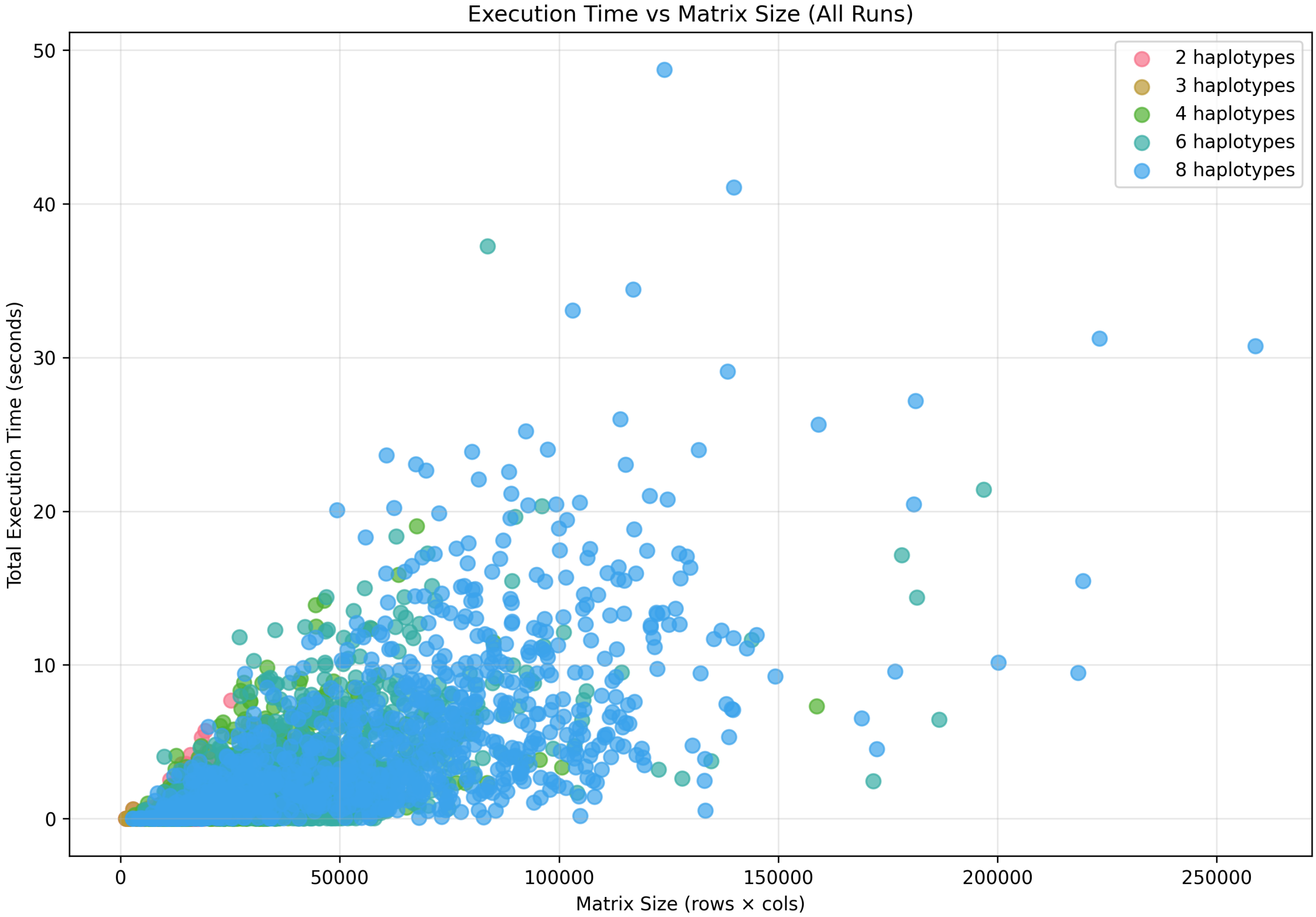
Important note: The efficiency peak at 3 haplotypes is **not statistically reliable** due to insufficient sample size (n=8). Robust conclusions are based on other categories with substantial samples.



2. Matrix Size Impact

The analysis reveals a **critical negative correlation** between matrix size and efficiency:

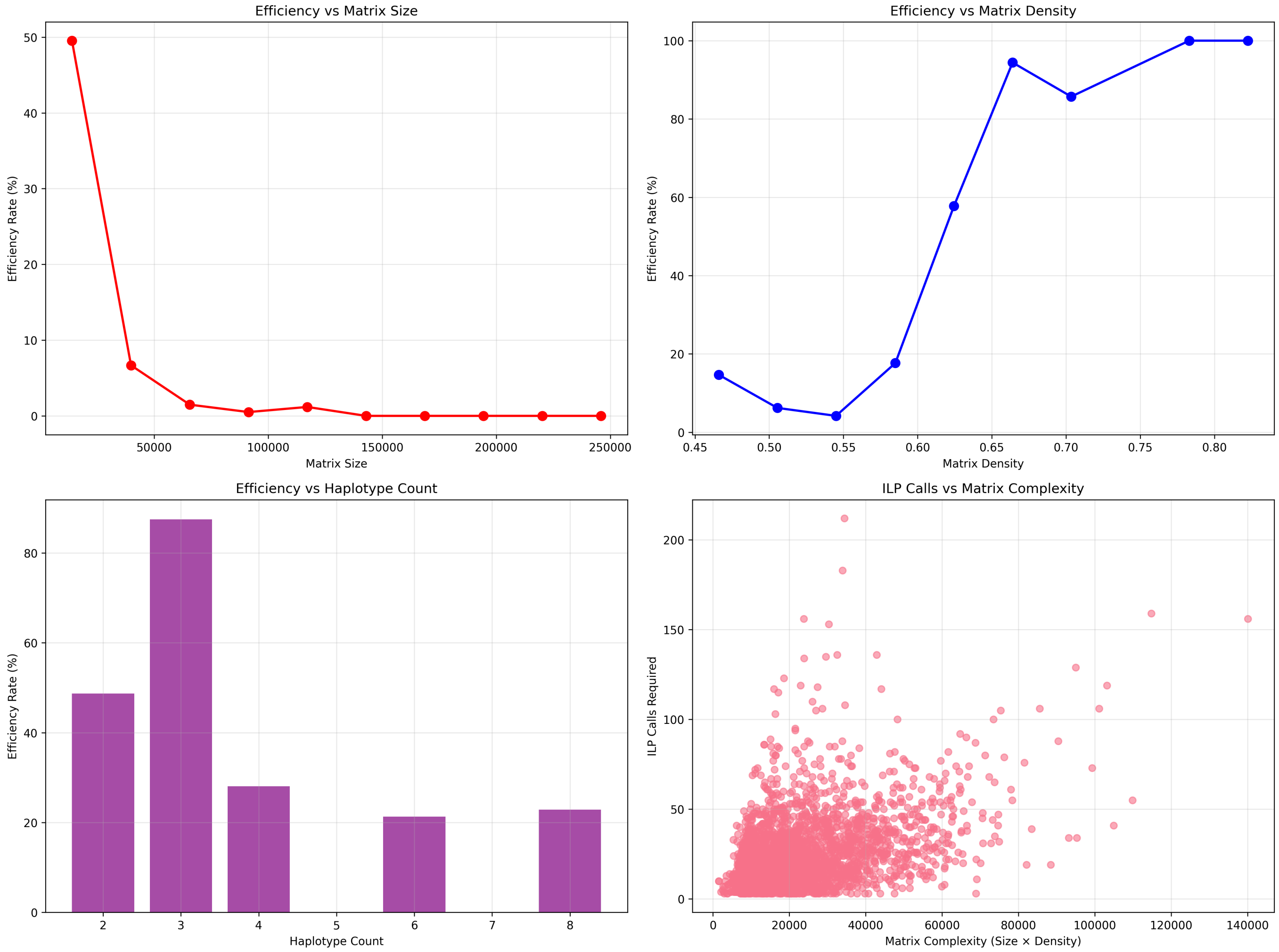
- **Matrices < 50,000 elements**: ~50% efficiency
- **Matrices 50,000-100,000**: ~7% efficiency
- **Matrices > 100,000**: <2% efficiency



3. Matrix Density Role

Matrix density presents an interesting **non-linear pattern**:

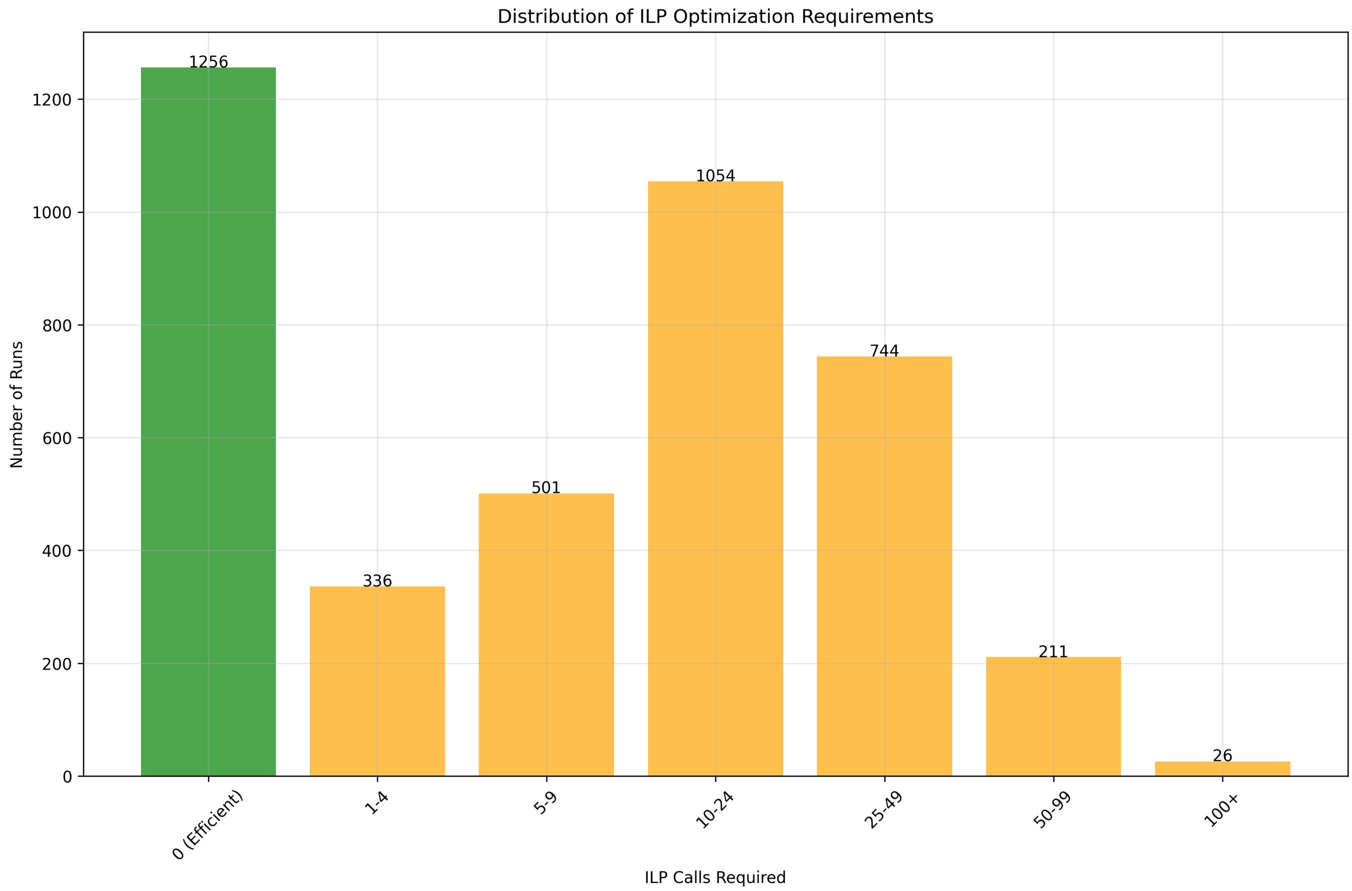
- **Density < 0.6**: ~15% efficiency
- **Density 0.6-0.65**: Peak at ~95% efficiency (**optimal zone**)
- **Density > 0.7**: Stable efficiency at ~85-100%



Distribution of Optimization Requirements

The ILP distribution analysis shows:

- **1,256 efficient cases** (0 ILP calls) - **Ideal**
- **2,872 complex cases** requiring ILP optimization:
  - 336 cases: 1-4 ILP calls (light optimization)
  - 501 cases: 5-9 ILP calls
  - 1,054 cases: 10-24 ILP calls (moderate optimization)
  - 744 cases: 25-49 ILP calls (intensive optimization)
  - 237 cases: 50+ ILP calls (very complex cases)



Performance Patterns

Efficient Cases (No ILP)

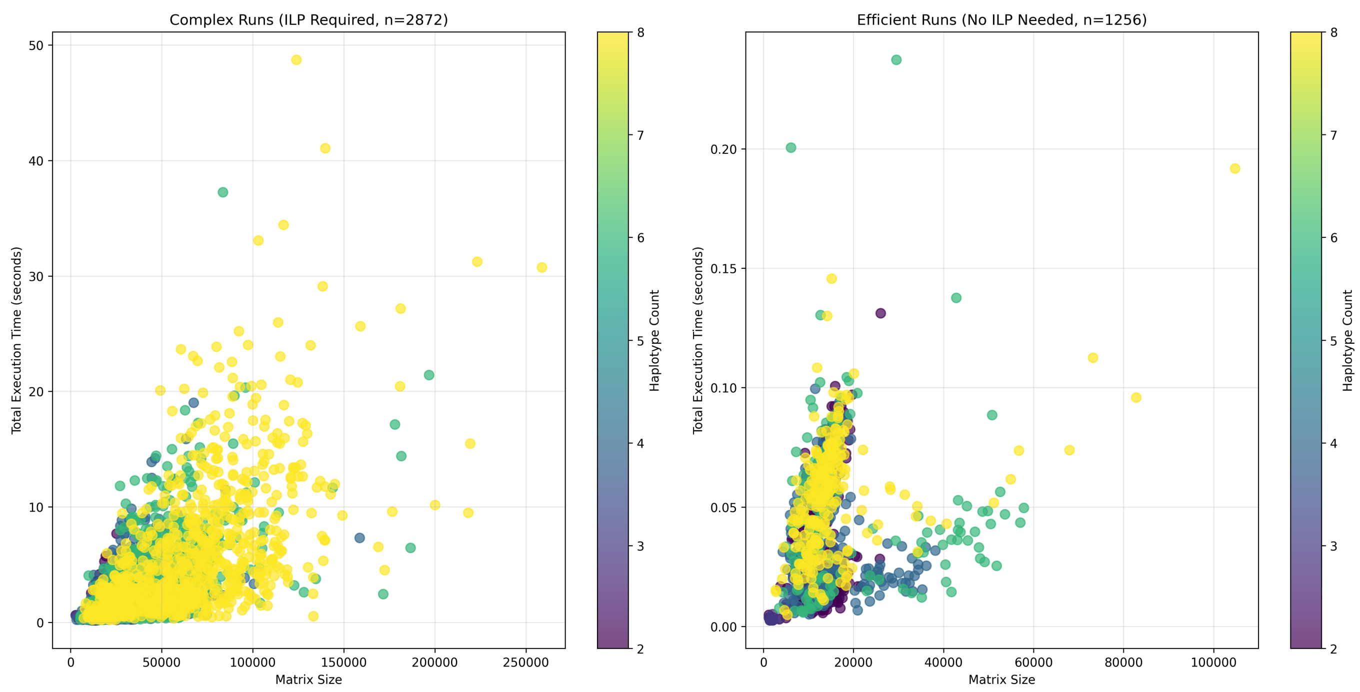
Efficient executions present distinct characteristics:

- **Smaller matrices**: optimized average size
- **High density**: generally > 0.7
- **Fast execution time**: < 0.5 seconds typically
- **Concentrated distribution**: in the low complexity zone

Complex Cases (With ILP)

Cases requiring ILP optimization show:

- **Wide range of sizes**: from small to very large matrices
- **Variable execution times**: 0.1 to 50+ seconds
- **Complexity-time correlation**: visible but with high variance
- **Extended distribution**: across the entire parameter space



Strategic Recommendations

1. Threshold Optimization

- **Target zone**: Matrices with density 0.6-0.7 and size < 50K elements
- **Preprocessing**: Preventive filtering of matrices that are too large or too sparse

2. Adaptive Strategies

- **Prediction heuristics**: Estimate complexity before execution
- **Dynamic thresholds**: Adjust parameters according to detected characteristics
- **Early optimization**: Trigger ILP more quickly for certain profiles

3. Algorithm Improvement

- **Data collection**: Increase sample for 3 haplotypes before definitive conclusions
- **Large matrices**: Develop decomposition strategies
- **Parallelization**: For cases requiring numerous ILP calls

Analysis Limitations

⚠ **Sampling bias**: The "3 haplotypes" category (n=8) requires a larger sample for statistical validation. Robust conclusions are based on categories with n>100.

⚠ **Absence of output validation**: This analysis focuses solely on performance metrics (execution time, ILP calls) without verifying the **quality or correctness of outputs** produced by the preprocessing algorithm and ILP optimization. Biological results and precision of reconstructed haplotypes are not evaluated.

Conclusion

The algorithm presents **highly input-dependent performance**. The **optimal efficiency zone** is located in matrices with density 0.6-0.7 and moderate size. However, the **rapid degradation** for complex cases indicates a need for targeted optimization to maintain performance across all usage scenarios.