Reframing Population Genetic Structure as a Quantum Optimization Problem

Andrew A. Davinack

Department of Biological, Chemical, and Environmental Sciences, Wheaton College Massachusetts, Norton, MA 02766

Corresponding author: davinack_drew@wheatoncollege.edu

ABSTRACT

Population genetic structure is commonly inferred using statistical and ordination-based methods that emphasize variance partitioning or likelihood-based clustering. While powerful, these approaches may undersample the full space of possible population partitions, particularly in systems characterized by weak genetic differentiation and high connectivity. Here, I present a proof-of-concept framework that reframes population genetic distance data as a combinatorial optimization problem, enabling structure to be interrogated through a distinct computational lens. Pairwise genetic distances derived from mitochondrial COI sequences of the shell-boring polychaete *Polydora websteri* are represented as a weighted graph and optimized using a quantum-inspired implementation of the Max-Cut problem via the Quantum Approximate Optimization Algorithm (QAOA). Using small, tractable datasets, I demonstrate that this approach recovers partitions consistent with classical analyses without claiming improved inference or computational advantage. Rather, the contribution of this work lies in establishing a transparent and reproducible mapping between population genetic distance structure and quantum-ready optimization frameworks, providing methodological groundwork for future studies using high-dimensional genomic SNP data.

I. INTRODUCTION

Population genetic inference underpins a wide range of biological applications, from reconstructing evolutionary history and identifying barriers to gene flow, to informing conservation planning and managing biological invasions [1 – 4]. Classical population genetic analyses rely on a well-established toolkit which includes summary statistics such as F_{ST} (Wright's fixation index), ordination methods such as principal component analysis (PCA), clustering approaches and phylogeographic reconstruction, to identify genetic structure and connectivity among populations [5 – 8]. For decades, these methods have proven powerful and interpretable, particularly for small to moderate datasets and simple demographic scenarios. However, as genomic datasets continue to grow in size and complexity, there is increasing interest in alternative computational frameworks that can interrogate genetic structure in fundamentally different ways [9 – 11].

Recent advances in quantum information science have opened the possibility of applying quantum algorithms to hard optimization problems that arise across scientific domains [12 – 15]. In particular, combinatorial optimization problems, many of which are NP-hard, are central to tasks such as graph partitioning, clustering, and network analysis [15]. Once such problem is Max-Cut, which seeks an optimal bipartition of a weighted graph that maximizes the sum of edge weights crossing the partitioning [12, 13, 17]. Max-Cut is a canonical benchmark problem in quantum computing and has motivated the development of variational quantum algorithms such as the Quantum Approximate Optimization Algorithm (QAOA) [16 – 18].

Here, I explore the use of Max-Cut as a conceptual bridge between population genetic distance data and quantum optimization. Genetic distance matrices, derived from DNA sequence data, can be naturally represented as weighted graphs, where nodes correspond to individuals or

populations and edge weights reflect genetic divergence. Casting genetic structure inference as a graph partitioning problem allows the application of quantum optimization techniques that differ fundamentally from classical population genetic workflows. Importantly, this approach does not seek to replace existing statistical models of population history, but rather to evaluate whether quantum algorithms can uncover structure through a distinct computational lens.

In this study, I present a proof-of-concept application of QAOA to genetic distance data derived from mitochondrial COI gene sequences of the shell-boring polychaete worm *Polydora websteri*. *Polydora* is a marine parasite that infects commercially reared shellfish such as oysters and scallops, and has been shown to exhibit low genetic differentiation and high haplotype sharing across global geographic regions [19]. Using small tractable sample sizes, I demonstrate how genetic distances can be encoded into a Max-Cut formulation and optimized using quantum circuits executed on quantum simulators. I explicitly compare quantum-derived partitions with exact classical solutions, emphasizing algorithmic behavior, convergence properties, and circuit depth (p-level) rather than biological novelty.

Importantly, the goal of this study was not to supplant established population genetic methods, nor to claim superior biological inference at present. Rather, the novelty lies in the explicit mapping of population genetic distance structure to a quantum-ready optimization framework and the demonstration that such mappings are feasible, interpretable, and reproducible using existing quantum software infrastructure. This work provides a foundation for future investigations using larger multilocus datasets—such as RAD-seq–derived SNP matrices—where the combinatorial complexity of population partitions may exceed the practical limits of classical approaches.

II. METHODS

Study System and Genetic Distance Data

Mitochondrial cytochrome c oxidase subunit I (COI) sequence data for the shell-boring polychaete *Polydora websteri* were retrieved from the NCBI GenBank database. A total of twelve individuals were selected, representing four geographically distinct regions: Cape Cod, Massachusetts, USA; South Africa; China; and Australia (three individuals per population). Previous population genetic studies on *P. websteri* have reported extensive haplotype sharing and global homogenization among these regions along with low levels of genetic differentiation, consistent with widespread anthropogenic-mediated dispersal and high connectivity [19]. Sequences were aligned using the MAFFT algorithm, and pairwise genetic distances were calculated under the kimura-2-parameter (K2P) substitution model using the Biopython library [20]. The resulting genetic distance matrix represents pairwise evolutionary divergence among sampled individuals and serves as the sole biological input to the quantum optimization framework described below. No demographic parameters or population genetic models were assumed beyond the computation of pairwise distances.

Graph Representation of Genetic Distances

Pairwise genetic distances were represented as a weighted, undirected graph $G = (V, E)$, where each node $v_i \in V$ corresponds to an individual/sample and each edge $e_{ij} \in E$ connects individuals *i and j.* Edge weights were defined as the normalized genetic distance between the corresponding pair of individuals, such that larger weights reflect greater genetic divergence. To ensure numerical stability and facilitate comparison across datasets, the distance matrix was normalized by its maximum value such that all edge weights lie in the interval [0, 1]. Missing or

undefined distances (e.g., due to alignment gaps) were conservatively set to zero, effectively removing their contribution to the optimization objective. This graph-based representation allows genetic structure inference to be reframed as a combinatorial optimization problem, independent of explicit population genetic assumptions.

Max-Cut Formulation

Genetic structure inference was formulated as a Max-Cut problem, a canonical NP-hard optimization task. Given the weighted graph described in the previous section, the objective is to partition the nodes into two disjoint sets such that the sum of edge weights crossing the partition is maximized.

Formally, for a binary assignment vector $x \in \{0,1\}^n$, the cut value is defined as:

$$C(x) = \sum_{i<j} w_{ij} \cdot I\left(x_i \neq x_j\right),$$

where w_{ij} denotes the normalized genetic distance between nodes i and j.

This formulation does not assume that the biological populations are bifurcating or that two clusters represent true biological evolutionary units; rather, it provides a mathematically well-defined optimization landscape for evaluating how quantum algorithms partition population genetic distance graphs.

Exact Classic Optimization

For small datasets (n ≤ 12), the Max-Cut problem was solved exactly used brute-force enumeration of all possible binary partitions, fixing one node to remove symmetry between equivalent cuts. This provided the true global optimum against which quantum-derived solutions

could be evaluated. Exact solutions were used solely for validation purposes and were not feasible for large datasets due to the exponential growth of the search space.

Quantum Optimization via QAOA

Quantum optimization was performed using the Quantum Approximate Optimization Algorithm (QAOA). QAOA is a variational, hybrid quantum-classical algorithm designed to approximate solutions to combinatorial optimization such as Max-Cut. The Max-Cut objective was encoded into a problem Hamiltonian acting on n qubit, where each qubit represents the binary assignment of a node in the graph. The algorithm alternates between applying the problem Hamiltonian and a mixing Hamiltonian for a specific number of layers, p, with each layer parameterized by angles Υ and β.

In this study, QAOA circuits were constructed manually to ensure compatibility with current quantum simulation backends. Circuits were executed using Qiskit's Aer simulator, and parameter optimization was performed using COBYLA classical optimizer. For all reported results, we used $p = 1$, corresponding to a single alternation layer, to maintain circuit simplicity and interpretability.

Sampling and Solution Extraction

After parameter optimization, the final QAOA circuit was sampled to generate a distribution over bitstrings corresponding to candidate cuts. The expected cut value was computed from the full distribution, while the best sampled bitstring was identified and evaluated using the classical cut objective. Quantum-derived solutions were compared to exact classical optima in terms of cut

value and node assignments, without post hoc adjustment or biological interpretation. A summary of the analytical pipeline is shown in figure 1.

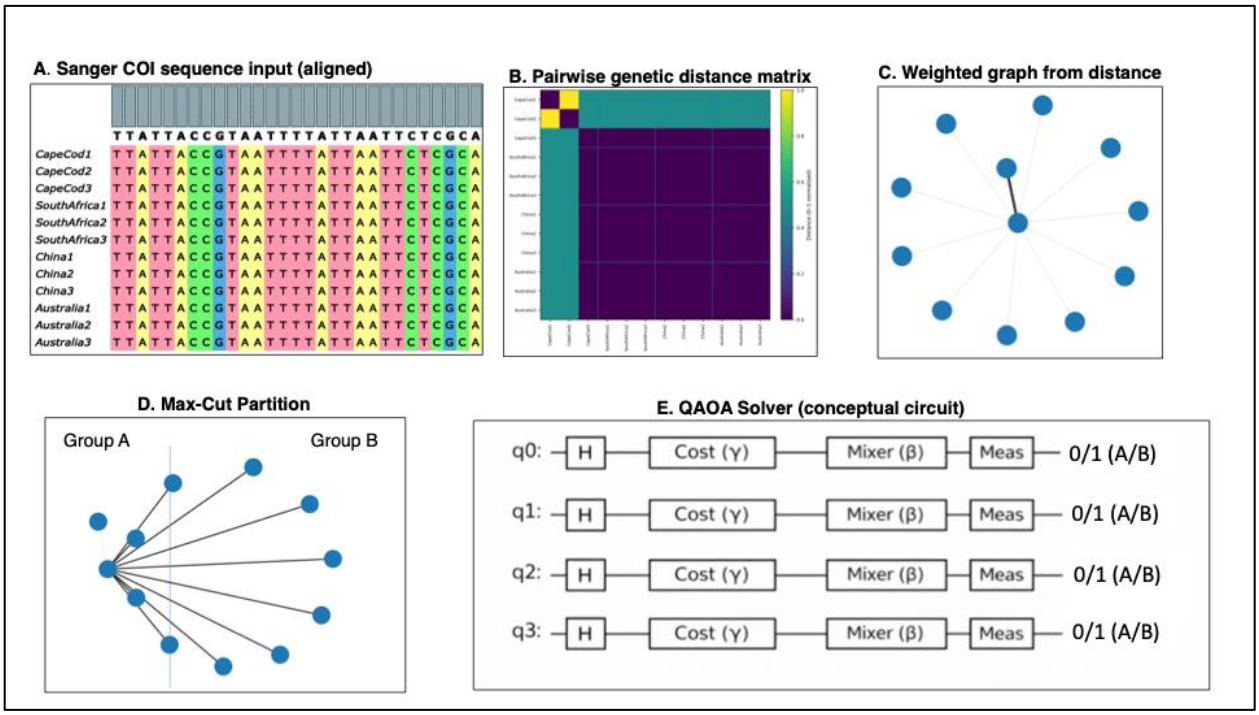

Figure 1. Conceptual pipeline for applying quantum optimization to population genetic distance data. (A) Multiple sequence alignment of mitochondrial cytochrome c oxidase I (COI) sequences from geographically distinct populations of *Polydora websteri*. (B) Pairwise genetic distance matrix computed under the Kimura-2-parameter (K2P) model and normalized to the interval [0,1]. (C) Weighted, undirected graph representation of genetic distances, where nodes correspond to individual sequences and edge weights reflect normalized pairwise genetic distances. (D) Formulation of genetic structure inference as a Max-Cut optimization problem, in which nodes are partitioned into two groups such that the sum of edge weights crossing the partition is maximized. Node layout is schematic and does not imply spatial or evolutionary proximity. (E) Conceptual depiction of the Quantum Approximate Optimization Algorithm (QAOA) applied to the Max-Cut problem; qubits are initialized in superposition, evolved under cost (γ) and mixer (β) operators, and measured to produce bitstrings encoding candidate partitions with the highest-scoring partition retained (partition A or B).

Software and Reproducibility.

All analyses were implemented in Python using NumPy and Qiskit [19, 20]. Scripts for loading genetic distance matrices, constructing optimization problems, executing QAOA circuits, and evaluating results were written to be fully reproducible and platform independent. All computations were performed on a classical macintosh workstation using quantum simulators;

no access to physical quantum hardware was required. All datasets including the aligned DNA sequence file, distance matrix and python scripts are available via Github (https://github.com/parasiteguy/population-genetic-max-cut.git)

III. Results

Pairwise K2P genetic distances computed from COI sequences of *Polydora websteri* exhibited low overall divergence, consistent with previous population genetic analyses reporting extensive haplotype sharing and weak geographic structure. Normalization of the distance matrix yielded a dense, weighted graph in which most edges carried small but nonzero weights, reflecting shallow differentiation among individuals across sampling regimes. This graph representation provided a suitable test case for evaluating whether quantum optimization recovers structure consistent with classical expectations under conditions of genetic homogeneity.

For the dataset consisting of 12 individuals sampled across four geographic regions (Cape Cod, South Africa, China, and Australia), the Max-Cut problem was solved exactly using brute-force enumeration. The optimal partition maximized the sum of normalized genetic distances across the cut and resulted in a highly asymmetric split, separating two Cape Cod individuals from the remaining 10 sequences. The exact solution achieved a maximum cut value of 9.99 (normalized units). Importantly, this portion did not correspond to an ecologically interpretable population split and instead reflects the optimization objective acting on subtle distance variation within an otherwise homogenous dataset.

The same Max-Cut problem was solved using the Quantum Approximate Optimization Algorithm (QAOA) with a single alternation layer ($p = 1$) executed on a classical quantum simulator. Parameter optimization converged to a set of angles Υ and β that maximized the

expected cut value under the QAOA ansatz. The optimized quantum stated yielde an expected cut value of 6.57, lower than the exact optimum, as expected for shallow-depth QAOA. However, sampling from the optimized quantum circuit produced bitstrings corresponding to the exact optimal cut identified by the brute-force classic solver. Thus, despite the reduced expected value at $p = 1$, QAOA successfully recovered the globally optimal solution through probabilistic sampling.

The best quantum-derived partition was identical to the exact classical Max-Cut solution in both cut value and cluster membership. This result demonstrates that, even at minimal circuit depth and without quantum advantage, *the QAOA framework can faithfully encode and recover optimization structure present in genetic distance graphs.* Notably, the quantum optimization did not reveal population structure inconsistent with classical analysis for this dataset. Instead, the quantum solution reproduced the same weakly structured partition implied by low genetic divergence and minimal classic population differentiation. These results confirm that quantum optimization can be applied to population genetic distance matrices in a mathematically consistent and computationally tractable manner using existing quantum software infrastructure.

IV. Discussion

This study demonstrates that population genetic structure inference can be reformulated as a combinatorial optimization problem and addressed using quantum-inspired methods. By representing genetic distance matrices as weighted graphs and applying Max-Cut optimization, a statistical inference task was translated into a problem class that is native to quantum algorithms. This reframing is conceptually distinct from classical population genetic analyses such as F_{ST}, Analysis of Molecular Variance (AMOVA), PCA or clustering-based methods (e.g.

STRUCTURE), which rely on variance partitioning, eigen-decomposition, or likelihood-based inference [7, 23]. In contrast, the Max-Cut formulation seeks global partitions that maximize pairwise genetic dissimilarity across groups, providing an alternative lens through which genetic structure can be interrogated.

Interpretation of Quantum Results in a Biologically Homogenous Dataset

Application of both exact classical Max-Cut and shallow-depth QAOA to COI data from *Polydora websteri* recovered equivalent partitions characterized by weak and biologically uninformative splits. This outcome is consistent with prior analyses of the species that reported extensive haplotype sharing and low geographic differentiation across this species' range [19]. Perhaps more importantly, the quantum approach did not produce spurious or contradictory structure relative to classical methods. Instead, it reproduced the same absence of meaningful population subdivision, indicating that quantum optimization behaves sensibly when applied to biologically homogenous datasets. This result reinforces that the framework does not artificially impose structure where none exist. At the same, the ability of QAOA to recover the exact optimal cut through probabilistic sampling – despite lower expected objective values at $p = 1$ – demonstrates that quantum variational algorithms can faithfully encode and explore genetic distance landscapes.

No Claim of Quantum Advantage

I should explicitly note here that this work does not claim quantum advantage, improved biological inference, or computational speedup. All quantum computations were executed using classical simulators and the problem size (12 individuals representing four populations – three

per population) remains well within the tractable regime of brute-force classical solvers. Rather, the contribution of this study is methodological: it establishes a clear, reproducible mapping between population genetic data structures and quantum optimization frameworks, and it validates that this mapping behaves consistently with biological expectations under controlled conditions. Such groundwork is essential before more ambitious claims regarding scaling behavior, advantage, or discovery can be credibly assessed.

Why Classical Methods May Miss Structure in High-Dimensional Regimes

Classical population genetic tools are highly effective for many applications, but they often rely on heuristics or dimensionality reductions that may obscure complex or weakly expressed structure in large, multilocus datasets [24]. For example, PCA and related methods emphasize variance along dominant axes, potentially overlooking subtle but globally optimal partitions. Similarly, likelihood-based clustering methods can be sensitive to model assumptions and initialization.

In contrast, combinatorial optimization formulations such as Max-Cut operate directly on pairwise relationships without requiring assumptions about population number, Hardy-Weinberg equilibrium, or linkage equilibrium. Quantum optimization algorithms are designed to explore rugged, high-dimensional solution landscapes that become increasingly difficult for classical heuristics as problem size grows. This distinction suggests that quantum approaches may prove most informative not for small mitochondrial datasets that were used here, but for multilocus SNP datasets generated via RAD-seq or whole-genome sequencing, where the dimensionality and complexity of genetic distance graphs increase dramatically.

Implications for Future Quantum-Enabled Population Genetics and Genomics

The results presented here motivate several directions for future work. Increasing the QAOA depth parameter (p) will allow exploration of more expressive quantum circuits and may improve solution quality for larger or more complex datasets. Systematic scaling analyses across increasing numbers of individuals and loci will be critical for identifying regimes where quantum optimization diverges from classical heuristics. Beyond Max-Cut, other graph-based optimization problems, such as community detection, minimum bisection, or modularity maximization, may offer alternatives formulations better aligned with specific biological data questions. Integrating multilocus distance measures or SNP-based similarity graphs represents a particularly promising avenue, as these datasets more fully capture population structure than single-locus mitochondrial markers.

More broadly, this work represents the first attempt to bridge evolutionary biology and quantum information science. In terms of contextualizing the role of quantum computing in biological data analysis, early-stage applications should focus on conceptual clarity, algorithmic transparency, and biological interpretability rather than premature claims of advantage. By providing a concrete, reproducible example of quantum optimization applied to population genetic data, this study establishes a foundation for interdisciplinary collaboration and positions population genetics, and evolutionary biology as a whole as a domain where quantum methods may eventually offer novel insights.

297 **REFERENCES**

298 1. Avise JC (2000) Phylogeography: the history and formation of species. Harvard University
299 Press.

300 2. Funk WC, McKay JK, Hohenlohe PA, Allendorf AW (2012) Harnessing genomics for
301 delineating conservation units. *Trends in Ecology & Evolution* 27: P489 – P496.

302 3. Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why,
303 how and so what? *Molecular Ecology* 19: 4113 – 4130.

304 4. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population
305 structure. *Evolution* 38: 1358 – 1370.

306 5. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from
307 metric distances among DNA haplotypes: application to human mitochondrial DNA restriction
308 data. *Genetics* 131: 479 – 491.

309 6. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a
310 new method for the analysis of genetically structured populations. *BMC Genetics* 11: 94.

311 7. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*
312 2: e190.

313 8. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial
314 population genetic variation. *Nature Genetics* 40: 646 – 649.

315 9. Ellegren H (2014) Genome sequencing and population genomics in non-model organisms.
316 *Trends in Ecology & Evolution* 29: P51 – P63.

317 10. Schrider DR, Kern AD (2018) Supervised machine learning for population genetics: a new
318 paradigm. *Trends in Genetics* 34: P301 – P312.

11. Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics* 9: 1003905.

12. Preskill J (2018) Quantum computing in the NISQ era and beyond. *Quantum* 2: 79.

13. Farhi E, Goldstone J, Gutmann S (2014) A quantum approximate optimization algorithm. *arXiv: 1411.4028.*

14. Cerezo M, Arrasmith A, Babbush R, Benjamin SC, Endo S, Fujii K, McClean JR, Mitarai K, Yuan X, Cincio L, Coles PJ (2021) Variational quantum algorithms. *Nature Reviews Physics* 3: 625 – 644.

15. Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N, Lloyd S (2017) Quantum machine learning. *Nature* 549: 195 – 202.

16. Karp RM (2009) Reducibility among combinatorial problems. In *50 Years of Integer Programming 1958 – 2008: from the Early Years to the State-of-the-Art* (pp. 219 – 241). Berlin, Heidelberg: Springer Berlin Heidelberg.

17. Lucas A (2014) Ising formulations of many NP problems. *Frontiers in Physics* 2: 5.

18. Farhi E, Harrow AW (2016) Quantum supremacy through the quantum approximate optimization algorithm. *arXiv 1602.07674.*

19. Davinack AA, Strong M, Brennessel B (2024) Worms on the Cape: An integrative survey of polydorids infestation in wild and cultivated oysters (*Crassostrea virginica*) from Massachusetts, USA. *Aquaculture* 581: 740366.

20. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Wilczynski B, de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25: 1422 – 1423.

21. Harris CR, Millman J, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Rio JF, Wiebe M, Peterson P, Gerard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. *Nature* 585: 357 – 362.

22. Javadi-Abhari A, Treinish M, Krsulich K, Wood CJ, Lishman J, Gacon J, Martiel S, Nation PD, Bishop LS, Cross AW, Johnson BR, Gambetta JM (2024) Quantum computing with Qiskit. *arXiv*: 2405.08810

23. Lawson DJ, van Dorp L, Falush D (2018) A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications* 9: 3258.

24. Davinack AA (2025) The promise of quantum computing for population genetics and molecular ecology. *Frontiers in Quantum Science and Technology* 4: 1657832.