AARHUS UNIVERSITY

Master's Thesis

Characterization of ctDNA and Classification Using Machine Learning Methods

Students
Victoria Hagelskjær Vilstrøm
Student number 202007866
Exam number 186463

Eftychia Daskalaki
Student number 202402146
Exam number 227492

Supervisors
Jakob Skou Pedersen
Mathilde Hartvig Diekema

June 2026

Section for Bioinformatics and
Computational Biology
Aarhus University

Contents

Abstract

Liquid biopsies provide a minimally invasive approach for analyzing tumor derived material in blood plasma. These samples contain circulating tumor DNA (ctDNA), which reflects tumor biology and has potential applications in cancer detection and disease monitoring when distinguished from healthy cell free DNA (cfDNA). However, ctDNA is typically present at low abundance, which limits reliable detection and poses significant analytical challenges. Fragmentomic features have emerged as informative signals of fragment origin that may improve discrimination between ctDNA and healthy cfDNA. These features capture patterns related to DNA fragmentation, providing complementary information to conventional mutation based approaches. In this thesis, we investigated multiple fragmentomic features, including fragment length, end motifs and mismatch rates, to characterize differences between ctDNA and healthy cfDNA fragments and determine predictive power. The results demonstrated consistent differences across several feature types. The ctDNA fragments exhibited shorter fragment lengths across nucleosome associated peaks, altered end motif frequencies and elevated overall mismatch rates. Notably, mismatch related features emerged as a previously underexplored source of fragmentomic signal, highlighting their potential relevance for future liquid biopsy research. A predictive model further confirmed that these features contain information relevant for fragment origin classification, with fragment length and mismatch rates identified as the most important contributors. However, the overall predictive performance indicated that these features alone are insufficient for robust classification.

Acknowledgements

Eftychia

I would like to thank my supervisor Jakob Skou Pedersen for giving me the opportunity to work on this thesis and for his valuable guidance during the semester. His ideas and proposals were really inspiring and motivating. I would also like to thank my daily supervisor, Mathilde Hartvig Diekema for her willingness to provide support to any question during the process. Furthermore, I would like to express my gratitude to my colleague, Victoria Hagelskjær Vilstrøm, for considering me as a thesis collaborator and for our excellent cooperation. Working with her was a fruitful experience, from which I evolved not only scientifically as a bioinformatician but also personally. Lastly, I want to thank my friends and family, who have always been by my side during my journey.

Victoria

I would like to thank my supervisor Jakob Skou Pedersen, for his guidance, constructive feedback and insightful ideas throughout this thesis, which were inspiring and shaped the direction of this work. My appreciation also extends to my daily supervisor Mathilde Hartvig Diekema, for her academic guidance, support and valuable feedback on the report. Furthermore, I am deeply grateful to my collaborator Eftychia Daskalaki. Without her great support, understanding and partnership through difficult times, finishing this thesis would not have been possible. I truly enjoyed working with her. Lastly, a special thanks goes to family for the amazing encouragement and patience they have shown me. Having them by my side has been a constant source of strength. Above all, I want to thank my mother for her kindness, support and encouragement, which has deeply influenced my life. She continues to motivate me every day and this thesis is a tribute to her.