

# DETECTION AND CHARACTERIZATION OF CLONAL HEMATOPOIESIS BASED ON CIRCULATING FREE DNA DATA

Jonas Riber Jørgensen

Main Supervisor  
Prof. Jakob Skou Pedersen

Daily Supervisor  
postdoc. Gustav Alexander Poulsgaard

This thesis is submitted for the degree of  
*Master of Science in Bioinformatics*

June 2025  
Aarhus University  
Department of Bioinformatics



# ABSTRACT

---

Clonal hematopoiesis (CH) is a biological phenomenon characterized by the expansion of hematopoietic stem cell (HSC) clones originating from a single ancestor cell that acquires somatic mutations. Certain mutations can confer a competitive advantage for the affected cell, which leads to a clonal expansion. Although often asymptomatic, CH is increasingly recognized for its association with aging, hematological cancers, cardiovascular disease, and lowered mortality. Moreover, its presence can confound the interpretation of circulating free tumor DNA (ctDNA) in blood samples, posing a challenge for cancer detection using circulating free DNA (cfDNA).

This thesis aims to detect and characterize CH in a cohort of 146 individuals previously treated for stage III colorectal cancer (CRC), using whole-genome-sequenced cfDNA samples collected at multiple time points. The primary objective is to identify individuals exhibiting signs of CH, annotate and characterize them. To achieve this, a custom pipeline was developed to identify somatic variants within a curated set of genes associated with CH, and apply custom filters for accurate variant calling. The annotated patients were then analyzed to explore patient-level variation and used to perform an initial characterization via mutational spectra, based on the substitution types. From these methods, 312 CH variants were collected. And with a selection process, 3 patients exhibited a more significant signal of the CH variants.

Although the methods remain under development and results are limited by the project's time constraints, the findings provide a foundation for future refinement of CH detection in cfDNA. And eventually may lead to a broader characterization and a somatic mutation model for CH variant calling, potentially aiding in the distinction between CH and cancer variants, in a cancer detection setting.



# ACKNOWLEDGMENTS

---

I want to express my deepest gratitude to my supervisor, Prof. Jakob Skou Pedersen, and my daily supervisor, postdoc. Gustav Alexander Poulsgaard, for their guidance and expertise throughout the course of this thesis. Their insights have played a crucial role in shaping this work's direction and quality.

I am also grateful to the research department at MOMA for providing the setting and environment, as well as for providing access to the data used within this thesis.

To my fellow students and friends, thank you for the shared frustrations, whiteboard sessions, and moments of humor that made this journey more bearable and enjoyable.

Lastly, I am forever grateful to my family for their support and belief in me. Without them, this achievement would not have been possible.

*Jonas Riber Jørgensen, 2025*

# CONTENTS

---

<b>List of Figures</b>	<b>viii</b>
<b>Abbreviations</b>	<b>ix</b>
<b>Preface</b>	<b>x</b>
<b>1 Aim of the Project</b>	<b>1</b>
<b>2 Introduction</b>	<b>3</b>
2.1 Hematopoietic Stem Cells and Their Division . . . . .	3
2.2 Accumulation of Somatic Mutations . . . . .	5
2.2.1 Base Substitutions . . . . .	5
2.2.2 Replication Errors . . . . .	5
2.2.3 Double-Strand Breaks . . . . .	6
2.3 Dynamics of the Clonal Expansion . . . . .	6
2.4 Mechanisms of Clonal Hematopoiesis Driver Genes . . . . .	7
2.5 Estimated Mutational Burden in Hematopoietic Stem Cells . . . . .	8
2.6 Circulating Free DNA . . . . .	9
2.7 Challenges in Detection of Clonal Hematopoiesis . . . . .	10
<b>3 Related Works for CH Detection</b>	<b>11</b>
<b>4 Data and Methods</b>	<b>13</b>
4.1 Data . . . . .	13
4.1.1 Next-generation sequencing . . . . .	13
4.1.2 Colorectal-Cancer Data Generation . . . . .	14
4.1.3 Blacklists for Problematic Regions of the Genome . . . . .	14
4.1.4 Clonal Hematopoiesis Driver Lists . . . . .	15
4.2 Data Processing . . . . .	17
4.2.1 Data Aggregation . . . . .	17
4.2.2 Patient Selection . . . . .	17
4.2.3 Region Selection . . . . .	18
4.2.4 Quality Control for Mismatch Calling . . . . .	18
4.2.5 Mismatch Calling . . . . .	19
4.2.6 Consensus Calls . . . . .	21
4.2.7 Caller Challenges and Improvements . . . . .	22
4.2.8 Mismatch Filtering . . . . .	22
4.2.9 Germline Filtering . . . . .	23
4.2.10 Variant Annotation . . . . .	24
4.2.11 Variant Candidate Selection . . . . .	24
4.2.12 Mutation Rate . . . . .	24
4.2.13 Comparing Counts with Different Numbers of Samples . . . . .	25

4.2.14	Benjamini-Hochberg Procedure . . . . .	25
4.2.15	Fisher's Method . . . . .	25
<b>5</b>	<b>Results</b>	<b>27</b>
5.1	Mismatch Calling . . . . .	27
5.1.1	Basic Filter Application . . . . .	28
5.1.2	Germline Variant Filtering . . . . .	28
5.1.3	Upper Depth Filter . . . . .	28
5.1.4	Investigating Read Depth Bimodality . . . . .	30
5.2	Variant Analysis . . . . .	32
5.3	Comparing Patients based on the CH Variants Identified . . . . .	34
5.4	Combining p-values for Patient-Level Annotation . . . . .	37
5.4.1	Resampling . . . . .	38
5.4.2	Permutation Testing . . . . .	38
5.5	Increasing the Cohort Size . . . . .	39
5.6	Characterization of the CH Variants . . . . .	40
<b>6</b>	<b>Discussion</b>	<b>43</b>
<b>7</b>	<b>Conclusion</b>	<b>47</b>
	<b>References</b>	<b>49</b>
	<b>Supplementary</b>	<b>53</b>
	Code Access . . . . .	53
	Supplementary Figures . . . . .	53

# LIST OF FIGURES

---

1	Figure 1. Hierarchical model of HSCs' differentiation. . . . .	4
2	Figure 2. Differentiation pattern of a HSC. . . . .	5
3	Figure 3. Mechanisms of somatic mutations giving rise to CH. . . . .	6
4	Figure 4. Timeline of a clonal expansion. . . . .	7
5	Figure 5. Circulating tumor DNA in the bloodstream. . . . .	9
6	Figure 6. Next-generation sequencing workflow. . . . .	14
7	Figure 7. Timeline of the sample collection of the CRC-dataset. . . . .	15
8	Figure 8. View of a problematic region in a BAM file. . . . .	16
9	Figure 9. Schematic of caller workflow. . . . .	19
10	Figure 10. Properly-paired read orientation. . . . .	22
11	Figure 11. VAF distributions of the immediate output from the pipeline. . . . .	27
12	Figure 12. Interaction between VAF and read depth. . . . .	29
13	Figure 13. VAF distribution with the germline/somatic grouping. . . . .	29
14	Figure 14. VAF distributions of early-stage data handling. . . . .	30
15	Figure 15. Overview of the depth distributions. . . . .	31
16	Figure 16. Depth distributions including the X-chromosome. . . . .	31
17	Figure 17. Ratio of driver variants. . . . .	32
18	Figure 18. Genes of the candidate driver variants. . . . .	32
19	Figure 19. Panel showing VAF distribution of top genes. . . . .	33
20	Figure 20. VAF distribution exclusively of the SETBP1 variant. . . . .	33
21	Figure 21. Age-association with CH variants called. . . . .	34
22	Figure 22. Heatmap of the variant patient matrix. . . . .	35
23	Figure 23. Summation of significant variants in each patient. . . . .	36
24	Figure 24. Patient distribution of the significant variant counts. . . . .	36
25	Figure 25. P-value distributions of the 48 patient cohort. . . . .	37
26	Figure 26. Patients selected with resampling. . . . .	38
27	Figure 27. Distribution of permutation test outcome. . . . .	39
28	Figure 28. P-value distribution of the 65 patient cohort. . . . .	40
29	Figure 29. Mutational spectra. . . . .	40
30	Figure 30. Tri-nucleotide mutational spectrum. . . . .	42
1	Supp. Figure 1. Panel of VAF distributions for all genes. . . . .	53



# ABBREVIATIONS

---

**BH** Benjamini-Hochberg.

**cfDNA** circulating free DNA.

**CH** clonal hematopoiesis.

**CHIP** clonal hematopoiesis of intermediate potential.

**CRC** colorectal cancer.

**ctDNA** circulating free tumor DNA.

**DGD** deep generative decoder.

**FDR** false discovery rate.

**HSC** hematopoietic stem cell.

**IGV** Integrative Genomics Viewer.

**indel** insertion and deletion.

**NGS** next-generation sequencing.

**ORF** open reading frame.

**PBMC** peripheral blood mononuclear cell.

**PCR** polymerase chain reaction.

**PoN** panel of normals.

**SBS** sequencing-by-synthesis.

**VAF** variant allele frequency.

**WGS** whole-genome sequencing.