Clinical Genomics – Early Stage Cancer Screening

Polina Ivanova, Catharina Hente

Table of Contents

1. Existing Studies/Approaches
   I. Sun et al.
   II. Qin et al.
   III. Tito et al.
2. Review: Genomic Analysis Pipeline
3. Working Trajectory
4. Reflection

Existing Studies

A population-scale analysis of 36 gut microbiome studies reveals universal species signatures for common diseases.

Sun, W., Zhang, Y., Guo, R. Et al. 2024

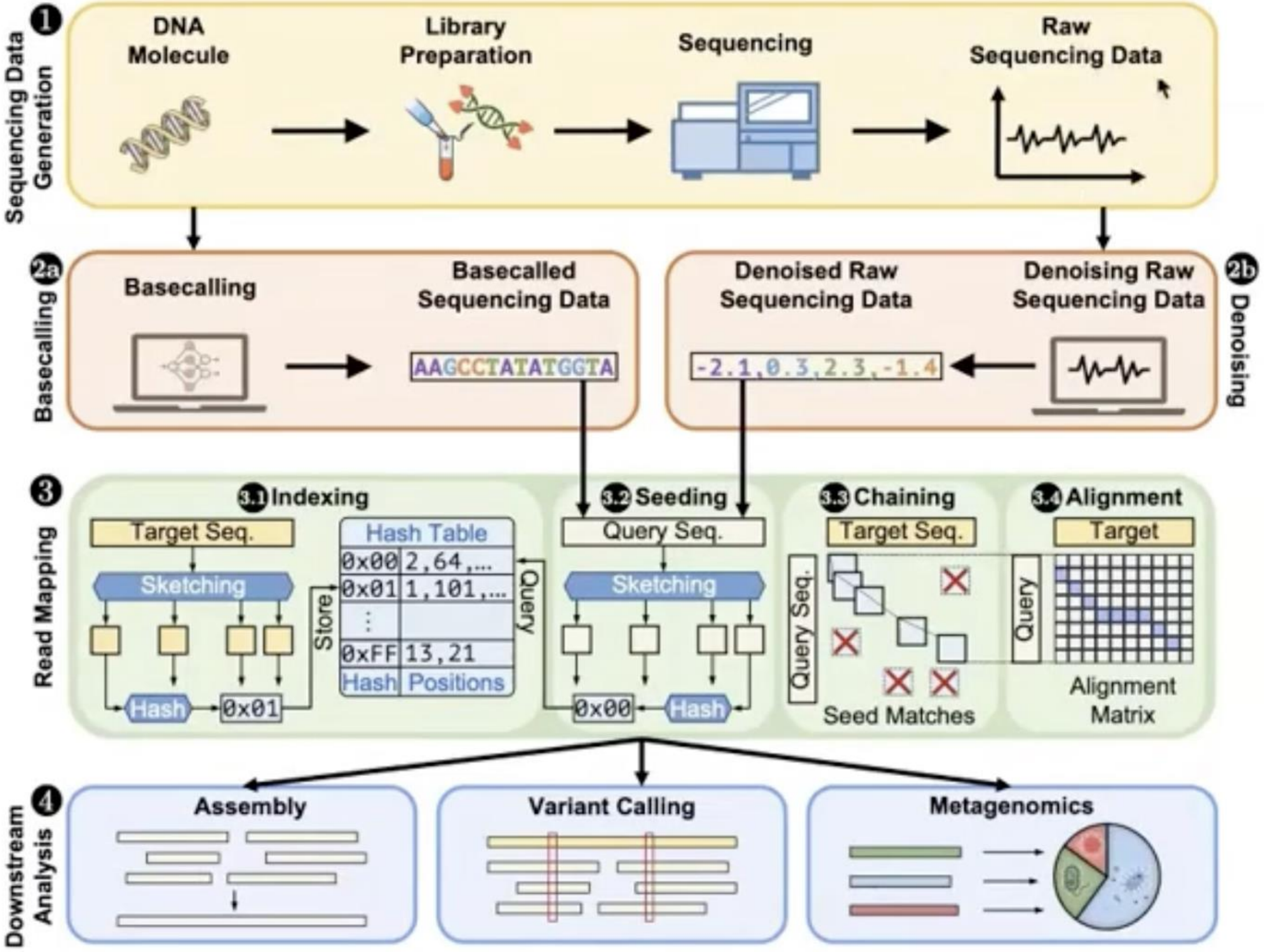
Consistent signatures in the human gut microbiome of old- and young-onset colorectal cancer.

Qin, Y., Tong, X., Mei, WJ. et al. 2024

Microbiome confounders and quantitative profiling challenge predicted microbial targets in colorectal cancer development.

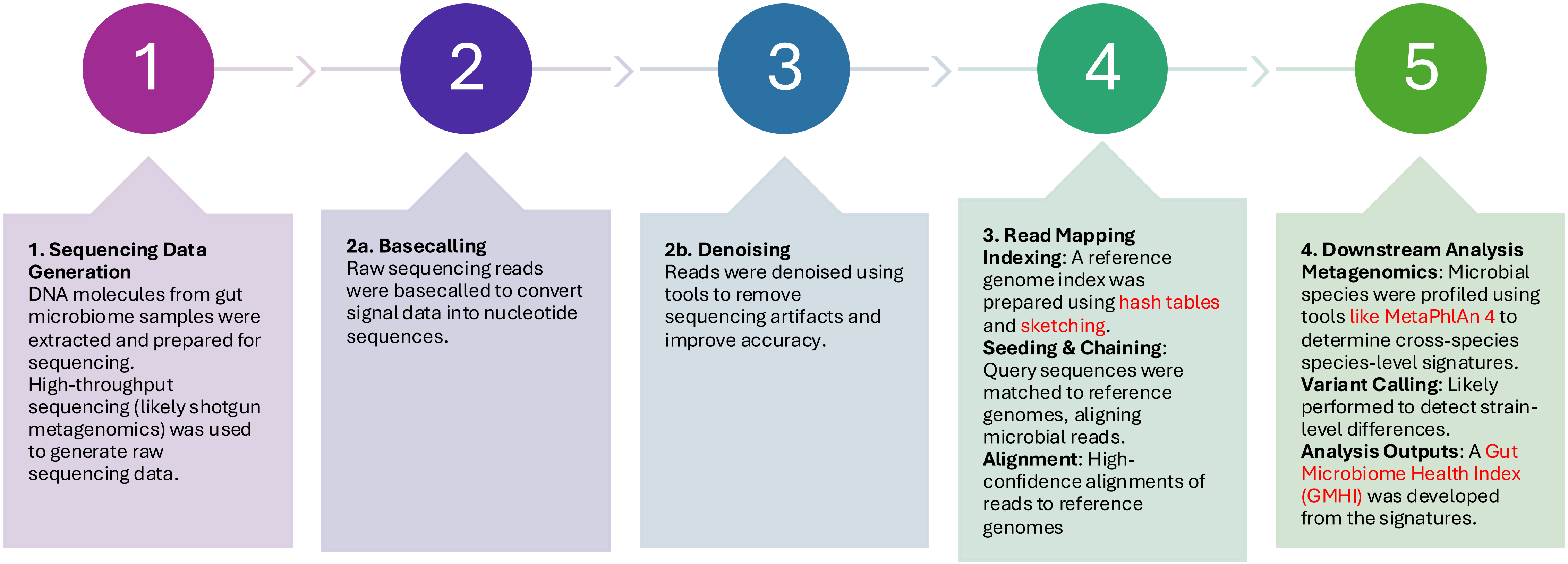
Tito, R.Y., Verbandt, S., Aguirre Vazquez, M. et al.

Genome Analysis Pipeline



Hash Table	
0x00	2,64,...
0x01	1,101,...
⋮	
0xFF	13,21
Hash	Positions

Study 1 – Gut Microbiome and Common Diseases



1. Sequencing Data Generation
DNA molecules from gut microbiome samples were extracted and prepared for sequencing.
High-throughput sequencing (likely shotgun metagenomics) was used to generate raw sequencing data.

2a. Basecalling
Raw sequencing reads were basecalled to convert signal data into nucleotide sequences.

2b. Denoising
Reads were denoised using tools to remove sequencing artifacts and improve accuracy.

3. Read Mapping
Indexing: A reference genome index was prepared using hash tables and sketching.
Seeding & Chaining: Query sequences were matched to reference genomes, aligning microbial reads.
Alignment: High-confidence alignments of reads to reference genomes

4. Downstream Analysis
Metagenomics: Microbial species were profiled using tools like MetaPhlAn 4 to determine cross-species species-level signatures.
Variant Calling: Likely performed to detect strain-level differences.
Analysis Outputs: A Gut Microbiome Health Index (GMHI) was developed from the signatures.

Study 2 – Old and Young Onset CRC

1. Sequencing Data Generation
DNA was extracted from stool samples of CRC patients and controls.
Shotgun sequencing was used to capture microbial communities.

2a. Basecalling
Raw sequencing reads were basecalled to nucleotide sequences.

2b. Denoising
Reads were denoised to filter noise and technical artifacts.

3. Read Mapping
Indexing: Reference genomes were indexed using microbial species and strain databases.
Seeding & Chaining: Query reads were matched to microbial strains.
Alignment: Reads were aligned to genomes for strain-level comparisons using tools like StrainPhlAn3.

4. Downstream Analysis
Metagenomics: Microbial strain-level profiling was conducted to detect differences between CRC onset groups.
Assembly: Possibly used to reconstruct microbial genomes for further strain analysis.
Variant Calling: Identified variations associated with disease progression.

Study 3 – CRC Microbiome Confounders and Predictions

1. Sequencing Data Generation
16S rRNA sequencing was performed on CRC patient and control samples to capture bacterial taxa.

2a. Basecalling
Basecalling converted raw signal data to nucleotide sequences for 16S regions.

2b. Denoising
Reads were denoised and processed to remove PCR and sequencing artifacts, ensuring reliable taxonomic assignment.

3. Read Mapping
Indexing: Reference databases for 16S rRNA regions were prepared.
Seeding & Chaining: Sequence reads were matched to the reference 16S sequences.
Alignment: High-confidence alignment matrices were generated to identify bacterial taxa.

4. Downstream Analysis
Metagenomics: Quantitative profiling was used to detect species- and genus-level biomarkers.
Confounder Analysis: Tools like GLMs were applied to control for confounders such as transit time and inflammation.
Variant Calling: Likely omitted due to the focus on 16S data rather than whole genomes.

Working Trajectory

Data Acquisition

Data Understanding and Preprocessing

Implementing and Understanding Preprocessing Steps (study inspired)

Initial ML Exploration using R Scripts from Original Studies

Training XGBoost in Python

Evaluation and Performance Metrics

Mitigating Bias and Expanding Scope

Final Testing and Documentation

GitHub Presentation

Reflection

- Results achieved in time we had are unfortunately inconclusive
- Potentially the issue lies in the metadata and labelling of the data from the study we used
- <u>Outlook and Improvements:</u>
  - Attempt running different algorithms on the data and compare and contrast the results
  - Try running XGboost on data from several different studies to compare/contrast/troubleshoot
  - Take data that only checks for CRC, not other diseases (at least for primary training)
  - Preferably we would have taken more time to access the data from Belgian study as it seemed to have most controls and had clearer labeling