Parallel Computing with DNA Forensics Data

Students: Luka Camus - Lucas Hauszler

Authors: Adam Michaleas, Philip Fremont-Smith, Chelsea Lennartz, Darrell O. Ricke

May 2025

Contents

Abstract

The analysis of complex mixture of DNA and other advanced forensics capabilities have been made possible by the forensic calculation of SNPs. The paper describes a pipeline using multiples tools that runs in parallel, and enable performance and scalability of multiple samples. Illumina and Ion Torrent generate sequencing datasets that are fed to GrigoraSNPs to allele call the HTS sequences. The results of this operation are automatically loaded in the IdPrism for the end-users to visualize the DNA mixture analysis. The whole pipeline can run in about 7 min for 100 millions sequences.

Introduction

Today, forensic scientists can identify subjects from DNA (deoxyribonucleic acid) samples that have DNA from multiple subjects by comparing DNA samples against a reference. The current criminal forensic solutions compare length polymorphism of STR[1] to detect a selected set of loci[2]. Forensic scientists can also analyse DNA to determine the genetic relation of peoples.

From 2017, the USA's CODI[3] has 20 loci[1] that are valuable for the analysis of DNA mixture which include typically less than 4 contributors, and the identification of first degree relationship. To identify up to 10 contributors, MIT developped custom SNP[4] panels. They are a collection of DNA spots attached to a surface that are used to detect polymorphism in a population. With SNP panels using the TranslucentID[5] method, more than 10 contributors can be identified in DNA mixtures.

The probability of random man not excluded, or P(RMNE), is the forensic calculation of a DNA profile match for someone not in the DNA mixture. Other forensic statical calculations donc scale with large number of loci and allele mismatches.

This paper[3] finds a way to avoid performance scaling issues that come from sizing 20 STRs to processing hundred of millions of DNA sequences. For this purpose, it uses five tools:

- GrigoraSNPs can allele call HTS sequences
- FastID[4] enables the reference and mixture of SNP profile comparaison against large datasets
- Fast P(RMNE) enable the rapid probability calculation of RMNE
- TranslucentID identifies the contributors in a DNA mixture with a large number of contributors
- IdPrism is a visualization tool for end users that show the results of FastID

Those tools are put together in a fully automated pipeline, and are tested and benchmarked with Ion Torrent and Illumina DNA sequencing platforms.

Related Work

A first work is `Short Tandem Repeats Forensics`. It details tools about the DNA sequencing of STR profiles that criminal DNA forensics relies heavily on. Those tools are STRait Razor (analysis of massively parallel sequencing (MPS) data from different marker systems, including STR), Illumina ForenSeq Software (software designed to support forensic genomic applications), and EuroForMix (can analyse complex DNA profiles with artefacts).

The second work is `Single Nucleotide Polymorphism DNA forensics`. Identifying multiple contributors of complex DNA mixtures can prove to be challenging, and can be facilitated

[1] Short Tandem Repeat
[2] a locus: fixed position on a chromosome where a particular gene or genetic marker is located
[3] Combined DNA Index System
[4] single nucleotide polymorphism

by identifying less frequently occurring DNA bases[5] in panels of selected SNPs loci. Taking a panel of multiple SNP alleles, the more contributors of a complex DNA mixture can be identified the larger the panel of SNPs.

Materials and Methods

4.1 HTS Sequencing

Ion Torrent and Illumina sequencing libraries, as well as their library quantitation are both prepared with different kits, each according to the manufacturers' instructions. Those libraries include AmpliSeq library kit, IonXpress Barcode Adapter and Ion Library TaqMan Quantation Kit for IonTorrent; Ion AmpliSeq Library PLUS, AmpliSeq UD Indexes, and NEBNext Library Quant Kit are used for Illumina. Ion Torrent necessite template preparation, and both perform sequencing using different kits according to the manufacturers' instructions.

The DNA output information of Illumina and ThermoFisher are generated in BAM or FASTQ file format, and both plateforms were used for benchmarking by generating files with 20, 50, 100, and 250 millions reads.

4.2 Forensic Panel

An in silico[6] SNP panel was created for 10298 SNPs, whose reference profiles were extracted from the 1000 Genomes project[2], then combined to create mixtures from known contributors.

4.3 Batch Layer

GrigoraSNP Tool uses SCALA language, and makes use of the Akka framework to enable the parallel computation of SNP allele calling of barcoded multiplexed HTS data. IT's designed to be scalable.

4.4 Serving Layer

IdPrism was developed to be an end-user plateform using Ruby on Rails, and a relational database. It provides them with an easy access of the analysis, and uses either FastID or TranslucentID depending of the number of contributors in a mixture.

4.5 Hardware Platforms

Both AMD and Intel systems plateform were used for the performance benchmarking plots. The AMD model has 64 CPU cores with 512GB RAM, and the Intel 40 CPU cores and 80 threads with 192 GB RAM.

4.6 High Performance Computing System Implementation

The HPC system for the pipeline developement is the MIT Laboratory Tx-Green's system; its implementation of the GrigoraSNP pipeline uses LLMapReduce to generate results from

[5] minor alleles

[6] performed on a computer, or via simulation software

FASTQ datasets in parallel. It uses the map-reduce programming model to allow users to process a large amount of data in the same program.

4.7 Standalone System Implementation

The pipelines running in parallel use standalone Linux systems to generate the analysis' results. They were tested and validated on both hardware platforms, and GNU Parallel was used for benchmarking.

4.8 Illumina MiSeq Pipeline

The detection of new DNA sequence runs in the Illumina MiSeq pipeline is automated. They ingest only on success and run every 15 min by default. A run marked as successfull will locate the appropriate FASTQ files for processing. The datasets are copied on the data analysis server to be decompressed in parallel using pigz[7]. The FASTQ files for each sequence run are then combined to be put as argument for GrigoraSNP that executes SNP allele calling. The results are then uploaded to IdPrism.

4.9 ThermoFisher Ion S5 XL Pipeline

The detection of new sequence runs is also automated, and data is only ingested on success run. The pipeline runs every 15 min by default. A success run of the Ion S5 Sequencer is marked as so in the `drmaa stdout.txt`. The pipeline will then ingest BAM output files in parallel. SAMtools is then invoked to convert BAM files to FASTQ. GrigoraSNP is then called on the generated file, and the results uploaded to IdPrism.

Serving Layer

IdPrism has references and mixture samples loaded for analysis and visualization. It's a DNA forensic analysis system that is fully automated, visualize DNA results for end users upon demand, and automatically searches new mixtures against existing references in the back-end database as the sample profiles are updated. A pie chart is generated with the DNA contributor concentration after an estimation for better readability.

Results

This pipeline is able to identify all twelve contributors using FastID with very low (RMNE) values (8.3e-130). GNUParallel and LLMapReduce enabled both scalability and parallel computing for the GrigoraSNPs batch layer for both environments (standalone and HPC). No matter the number of datasets runs in parallel (from 1 to 250), the analysis time increases only slightly for bigger numbers with the use of GNUParallel and LLMapReduce, showing a near linear performance of the pipeline.

[7] parallel implementation of gzip

Discussion

The innovation described in the paper enable the adoption of SNP forensic panel without requiring the expansion of current data centers. They also enable DNA forensic analysis on laptops as they can run on HPC systems or a laptop.

Future Work

The future goal of this pipeline would be to support hybrid STR and SNP panels.

Conclusion

This paper shows an advancement for forensic scientists to identify people, should it be in the criminal forensic, or to more broadly identify kinship between people. And it is possible in complex DNA mixture containing more contributors. Using the pipeline and developed tools, the 12 contributors have been able to be identified. The pipeline is able to be as scalable and as performant as it is thanks to parallel systems, both on the software and the hardware sides. The benchmark solutions and framework used for the GrigoraSNP tool allow parallel computing, which makes scalability nearly linear. Moreover, the two hardware components used have been designed to handle parallel computation thanks to number of CPU and threads.

Added to an automatic loading of results in the end-user platform (IdPrism), the SNP data analysis typically completes in about 7 minutes for 100 millions sequences, and can even run on a laptop.

References

[1] URL: https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet.

[2] URL: https://www.internationalgenome.org/.

[3] A. Michaleas, P. Fremont-Smith, C. Lennartz, and D. O. Ricke. "Parallel Computing with DNA Forensics Data". In: *IEEE High Performance Extreme Computing Conference (HPEC)*. 2022. URL: https://ieeexplore.ieee.org/document/9926352.

[4] Ricke D. O. "FastID: Extremely fast forensic DNA comparisons". In: *2017 IEEE High Performance Extreme Computing Conference (HPEC)*. 2017. URL: https://ieeexplore.ieee.org/document/8091056.

[5] D. O. Ricke, J. Watkins, P. Fremont-Smith, M. S. Petrovick, T. Boettcher, and E. Schwoebel. "TranslucentID: analysis of complex DNA SNP mixtures with large numbers of donors". In: *Australian Journal of Forensic Sciences* (2021). URL: https://doi.org/10.1080/00450618.2019.1699958.