Forensic Science International: Genetics 31 (2017) 40–47

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Secure and robust cloud computing for high-throughput forensic microsatellite sequence analysis and databasing

Sarah F. Bailey[a,b], Melissa K. Scheible[a,b], Christopher Williams[a,b], Deborah S.B.S. Silva[a,b], Marina Hoggan[a], Christopher Eichman[c], Seth A. Faith[a,b,*]

[a] NC State University, Molecular Biomedical Sciences, 1060 William Moore Dr., Raleigh NC, 27607, United States
[b] NC State University, Forensic Sciences Institute, 1060 William Moore Dr., Raleigh NC, 27607, United States
[c] NC State University, College of Veterinary Medicine, Office of Information Technology, 1060 William Moore Dr., Raleigh NC, 27607, United States

ARTICLE INFO

Article history:
Received 16 March 2017
Received in revised form 24 July 2017
Accepted 6 August 2017
Available online 8 August 2017

Keywords:
Cloud
Bioinformatics
Microsatellite
Database
Sequencing
Security

ABSTRACT

Next-generation Sequencing (NGS) is a rapidly evolving technology with demonstrated benefits for forensic genetic applications, and the strategies to analyze and manage the massive NGS datasets are currently in development. Here, the computing, data storage, connectivity, and security resources of the Cloud were evaluated as a model for forensic laboratory systems that produce NGS data. A complete front-to-end Cloud system was developed to upload, process, and interpret raw NGS data using a web browser dashboard. The system was extensible, demonstrating analysis capabilities of autosomal and Y-STRs from a variety of NGS instrumentation (Illumina MiniSeq and MiSeq, and Oxford Nanopore MinION). NGS data for STRs were concordant with standard reference materials previously characterized with capillary electrophoresis and Sanger sequencing. The computing power of the Cloud was implemented with on-demand auto-scaling to allow multiple file analysis in tandem. The system was designed to store resulting data in a relational database, amenable to downstream sample interpretations and databasing applications following the most recent guidelines in nomenclature for sequenced alleles. Lastly, a multi-layered Cloud security architecture was tested and showed that industry standards for securing data and computing resources were readily applied to the NGS system without disadvantageous effects for bioinformatic analysis, connectivity or data storage/retrieval. The results of this study demonstrate the feasibility of using Cloud-based systems for secured NGS data analysis, storage, databasing, and multi-user distributed connectivity.

1. Introduction

Technology to analyze forensic evidence is evolving rapidly, including next-generation sequencing (NGS), also referred to as massively parallel sequencing (MPS). Since 2012, numerous studies have demonstrated the power of NGS in a forensic context by sequencing high density multiplexes of a variety of genetic data including: autosomal, Y and X-STRs, mitochondria, and single nucleotide polymorphisms (SNPs) for biogeographic ancestry, phenotype, identity and linage [1–10]. Thus, the field of forensics is likely to gravitate to NGS as a solution to investigate special cases that traditional DNA technologies cannot solve and possibly to conduct genetic analysis in the field/crime scene with portable, miniaturized genetic analysis devices. However, central to NGS technology, the volume (size) of data is immense and requires specialized computing approaches to analyze and manage the data. For example, the typical output of an Illumina MiSeq desktop sequencer is currently 50 million sequence reads, equivalent to 15 Gigabases [11]. Moreover, the trend in the field of NGS is larger data sets and smaller devices [12,13].

In the modern field of data science, big data is defined by the 5-Vs: volume, variety, velocity, value, and veracity [14]. As stated above, NGS technology applied to forensic DNA analysis satisfies the 3-principal V's of big data: variety, volume and velocity. Thus, to obtain precision and accuracy required in forensics, we must develop and validate tools to determine the *veracity* (truthfulness) and find *value*, as well as manage this data with effective storage, redundancy, security, and accessibility.

Fields outside of forensic science are finding extremely high value in implementing Cloud-based services. Many businesses, research organizations and government agencies have turned to the Cloud for access to the resources required to analyze and store

* Corresponding author at: NC State University, Molecular Biomedical Sciences, 1060 William Moore Dr., Raleigh, NC, 27607, United States.
E-mail addresses: sethadam30@gmail.com, safaith@ncsu.edu (S.A. Faith).

http://dx.doi.org/10.1016/j.fsigen.2017.08.008
1872-4973/

data in a secured environment [15]. Cloud computing is the on-demand delivery of compute power, database storage, applications, and other IT resources through a Cloud services platform via the internet with pay-as-you-go pricing. Many leading businesses are using the Cloud for website hosting, customer data centers, business analytics, mobile services, backup/recovery and transaction centers [16]. A research study has shown that companies implementing Cloud services grow an estimated 19.3 percent faster than their competitors and 86 percent of companies are spending at least part of their IT budgets on Cloud services [17]. Furthermore, Cloud services are increasingly more prevalent within the federal government and could be models for forensic laboratories. For example, Amazon Web Services (AWS) offers Federal Risk and Authorization Management Program (FedRAMP) compliant AWS GovCloud, Department of Defense (DoD) Cloud Computing Security Requirements Guide (SRG) Levels 2 and 4, National Institute of Standards and Technology (NIST) SP 800-53, and Criminal Justice Information Services (CJIS) services. Microsoft also offers Cloud services in the government sector; millions of people use Microsoft Cloud for Government that includes Azure Government, Office 365 Government, and Dynamics CRM Online Government.

Fortunately the field of NGS has seen multiple data processing tools developed to conduct bioinformatics analysis of NGS big-data [18–24]. These tools have helped establish a foundation for consistent bioinformatics analysis in forensics, however most are computer code written in high-level programming languages and require expertise and training to use. Some NGS instrument manufactures and software vendors have also produced software applications designed for laboratory scientists, but to date, none have complete flexibility in the methods to analyze data, nor provide a unified work environment where multiple users can analyze and share data for casework applications [9,8,25]. Further, these tools could benefit from following guidance by a working group in the International Society for Forensic Genetics (ISFG) for nomenclature of STRs and NGS data [26]. These guidelines encompass: reporting the core repeat and flanking sequences for STRs, reporting the sequence as the forward strand to a universal human genome reference (i.e., GRCh38), indexing the fragment (i.e., start and stop position on a chromosome), and maintaining backwards compatibility to existing STR databases with fragment size information. Thus, it is logical to construct a robust system that can meet the requirements of forensic DNA analysis of NGS data with the best practices defined to date.

This study was designed to evaluate the AWS Cloud for analysis of NGS data for forensic genetics. The major criteria evaluated include connectivity of devices and users, bioinformatics power, adherence to guidelines established to date, security, and end-user experience. The final product of this study is a demonstration of a secured, Cloud-based system for processing raw NGS data for autosomal, Y and X-STRs, conducting interpretations, and storing the data in a relational database that demonstrates the value and opportunity for forensic analysis in the Cloud.

2. Materials and methods

2.1. Sequencing

Genomic DNA samples from Standard Reference Material 2391c (National Institute of Standards and Technology, Gaithersburg, Maryland) were amplified with the PowerSeq™ Auto/Y System Prototype (Promega Corporation, Madison, Wisconsin) as described by the manufacturer and prior studies [2,4]. For Illumina Sequencing, post-PCR products were purified with Agencourt AMPure XP reagent (Beckman Coulter, Brea, California) using the epMotion 5075tc Liquid Handling Workstation (Eppendorf, Hamburg, Germany). The full volume of purified amplification product was used for automated library construction with the KAPA Hyper Prep Kit (Kapa Biosystems, Wilmington, Massachusetts), following the PCR-free protocol and using custom adapters manufactured by Integrated DNA Technologies (Coralville, Iowa). The Illumina-compatible libraries were quantified with the KAPA Library Quantification Kit (Kapa Biosystems) and sequenced on either the MiSeq (96 samples per run) or the MiniSeq (24 samples per run) sequencers (Illumina Inc., San Diego, California). MiSeq runs were performed with the MiSeq Reagent Kit v2 (300 cycles, single end sequencing), and MiniSeq reads were produced with the High Output (300 cycles, single end sequencing) Kit.

Library preparation for the sequencing runs on the MinION Mk1 B device (Oxford Nanopore Technologies, Oxford, United Kingdom) was performed using the manufacturer's Genomic DNA Protocol (SpotON R9 Flow Cell Version). A sample of 2391c – component B with a total concentration of 612 ng dsDNA post-PCR amplification with PowerSeq™ Auto/Y was used. Incubations for library preparation were performed in a Veriti 96-well thermal cycler (Applied Biosystems, Foster City, California) with reagents and values specific to manufacturer's protocol for the SQK-NSK007 kit. Product was quantified with QuBit dsDNA High Sensitivity Assay (ThermoFisher Scientific, Waltham, Massachusetts) and 0.2 pmoles was used to perform adapter ligation, library purification, and library elution based on the manufacturer's protocol for Genomic DNA (SpotON R9 Flow Cell Version). The Pre-Sequencing Mix (PSM), dsDNA library, was quantified with QuBit dsDNA High Sensitivity Assay. Two sequencing runs were conducted. The first run was performed on the SpotON R9 flow cell with two separate injections of 8.6 ng PSM over 25.5 h, and the MinKNOW (v.1.1.20) instrument control software protocol "map/NC_48Hr_Sequencing_Run_FLO_MIN105" and the Metrichor Agent (v 2.4.017) application "2D Basecalling RNN for SQK-NSK007" were used. The second run was performed on the SpotON R9.4 flow cell with two separate injections of 2.8 ng PSM over 10.1 h, and the MinKNOW (v.1.1.21 – updated) instrument control software protocol "map/NC_48Hr_Sequencing_Run_FLO_MIN106_SQK-007_plus_Basecaller.py," and the Metrichor Agent (v 2.42.2 – updated) application "2D Basecalling RNN for SQK-NSK007 plus Human Exome" were used.

2.2. Raw data pre-processing

Illlumina data, FASTQ files, were adapter trimmed and demultiplexed with the Illumina software installed on the sequencer and uploaded directly to the AWS environment. MinION FAST5 (hdf5 format) sequence data files produced via Metrichor were converted to FASTQ format using poretools [27]. The FASTQ files were parsed with a custom Python script to separate 1D and 2D reads. The 2D reads were uploaded to the AWS S3 bucket for further processing with the bioinformatics analysis tool.

2.3. Bioinformatics

To process raw data (FASTQ files), a data processing tool was developed as the backend bioinformatics engine. The logic of the algorithm is similar to that of a previously published method, STRait Razor [20], but has been optimized for high-throughput parallel processing in the Cloud using Python programming language (Version 2.7) and the BioPython package (version 1.68). The tool identifies target regions in the 5′ and 3′ STR flanks from a look up table of short DNA sequences (< 20 bps), adapted from Warshauer et al. (2015) and Parson et al. (2016), see Supplementary Table 2. The target regions are identified in the raw data using the 'pairwise' function of BioPython to accept aligned tags with greater than 73% homology (2–3 mismatches per tag)