applied biosystems

Integrated Forensic DNA Data Analysis and Management – A scalable enterprise solution for forensic laboratories

Jianye Ge, Allan T. Minn, Jie Deng, Makesh Karpagavinayagam, Narasimhan Rajagopalan
Human Identification Division, Thermo Fisher Scientific, South San Francisco, CA, 94080, USA

Introduction

Over the past 25 years various robust and reliable DNA typing technologies for human identity testing have been implemented. These technologies enable analyses of minute quantities of DNA and provide a resolving power in many forensic cases. Recently, the introduction of Next Generation Sequencing (NGS) technologies has ignited a revolution forensic science. NGS technologies use a massively parallel sequencing strategy with high coverage, low error, and high throughput of specified targets to provide DNA sequencing data with unprecedented capacity and speed at a reduced cost. In this fashion, a large battery of autosomal, Y chromosome, and X chromosome Short Tandem Repeat (STR) and Single Nucleotide Polymorphism (SNP) markers, as well as mitochondrial genome, can be analyzed in a single multiplex. Therefore, NGS technologies are compatible with the current STR based forensic standards and also provide extra power to support new applications with the markers beyond STRs.

On the other side, with the constant growth of DNA testing in criminal investigations, many forensic laboratories struggle to deal with the rapidly increasing data sets and records associated with these growing case loads. Forensics laboratories use a variety of software and systems in the laboratory pertaining to data creation, management and analysis. Typically, most of these systems function independently and data generated from these various systems are later patched together through the use of predefined spreadsheet templates housed in the laboratory information management system. In addition, many of the current available analysis tools for local database searching and relationship testing are difficult to use leading to challenges in adopting them for regular use in the laboratory.
Thermo Fisher Scientific is developing a software solution specifically to address the data management and analysis difficulties currently challenging forensic DNA laboratories. This software supports the data generated from both CE and NGS technologies, and provides functionalities beyond conventional forensic applications.

NGS platform architecture

The NGS platform software is built on top of Life Technologies™ unified software architecture. It uses best of breed Java® enterprise and server deployable technologies to enable its high levels of flexibility, scalability and user definable configuration. Some of key technical features of the software architecture are:

- Support data generated from both CE and NGS technologies
- Cloud based system with highly secured data protection
- Standard conforming web interfaces that enable client access through standard web browsers
- Internationalization and localization through deployable language packs
- Modular application development and deployment

Data management

The data management system centralizes and stores DNA workflow data, facilitates integration of your Thermo Fisher Scientific DNA workflow instrumentation and helps you manage your laboratory more effectively and efficiently. This system is based on a highly scalable and hierarchy based database which allows the user to view, sort and search their workflow data in many views including case level, evidence or subject level, sample level or profile level. It can automatically pulls and stores raw, processed and final report data into the system. In addition, the tool is able to store SOPs, protocols, instrument data and kit data to allow for more efficient day to day workflow management and worksheet generation. Tools have been developed and integrated in the data management system to read and convert the output (i.e., BAM files) of the NGS instruments to standard forensic data formats with our newly developed tools. This forensic data includes conventional autosomal STRs, Y-STRs, and X-STRs, as well as new-fashioned SNPs, Indels, mitochondrial sequences, microRNA sequences, etc. With this data, a variety of forensic applications are supported.

Support Forensic Applications

1. Direct Search. The Direct Search application enables the creation and rapid searching of segmentable locally stored profile databases. The search is supported for all types of forensic data (e.g., STR, SNP, Y-STR, etc.) in the database.

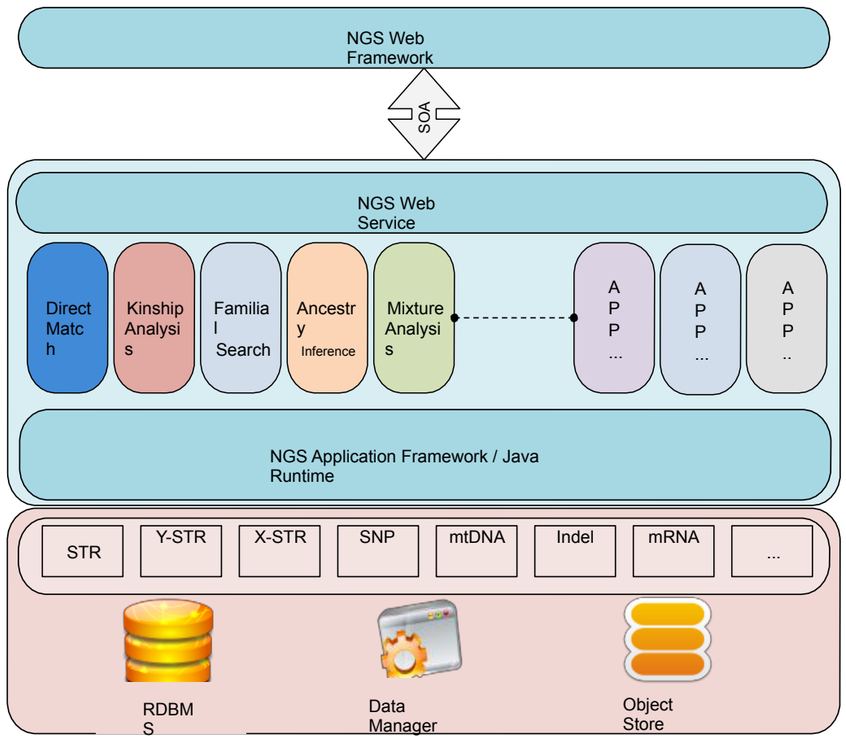

Figure 1: NGS Software Platform Schema
The NGS platform software is built on top of Life Technologies™ scalable unified software architecture which is highly modular allowing the development of new tools, and infrastructure in the same environment.

The tool also allows the user to define basic search criteria, such as loci match stringency (high, moderate, and low) for autosomal STR, as well as a minimum number of matching loci. It is optimized to allow for the searching of millions of profiles records housed in the database within seconds.
2.Kinship Analysis. The Kinship Analysis application is a simple, graphic user interface driven tool that allows for the rapid analysis of simple paternity testing, as well as more complex kinship cases, multi-generation cases, and other challenging cases involved in missing persons or DVI forensic cases. Both unlinked STRs and SNPs can be used in kinship analysis. The tool is driven and set up by a simple graphic user interface which allows the user to draw pedigree trees, assign profiles as well as configure analysis settings and reporting formats. The software tool allows the user to configure the general report format, headers and some content to meet their laboratory or regional specific requirements. Familial search is also supported by this kinship analysis engine with the filtering and ranking functions added.
3.Mixture Analysis. The Mixture Analysis application provides Bayesian interpretation of the DNA mixture samples. It will generally follow the recommendation of the International Society of Forensic Genetics on mixture interpretation for STRs. SNP based mixture analysis will also be supported.
4.Ancestry Inference. Ancestry Inference application can provide the investigative leads on which population an individual may belong to or which area an individual may come from. The current system is built based on autosomal SNP and is able to visualize with plug-in. The tool will be expanded to support STR markers.
5.Future applications. The scalable system was designed to support new applications in the future, such as tissues and fluids identification, human physical appearance, etc. The identification of human body fluids or tissues through methylation or mRNA-based profiling is very useful for forensic investigations. The physical appearance of the criminals could be inferred with the SNP profiling from the crime scene samples. Both of these applications can be extremely useful tools to generate investigative leads.

Current Status

1.Direct match. Both autosomal STRs and SNPs are supported in direct match. Users can customize the search criteria. Millions of records can be searched in a very efficient way (Figure 2).
2.Kinship analysis. Users can delineate both Prosecution Hypothesis (H0 or PH) and Defense Hypothesis (H1 or DH) with graphic tools. For each hypothesis, the likelihood ratio of each locus can be calculated and multiplied together, assuming the markers are independent from each other.
3.Familial search. The familial search tool uses the same engine as the kinship analysis module, but it is a large scale of pairwise kinship analysis (i.e., parent-child and full sibling) to search against DNA databases.
4.Ancestry inference. The current ancestry inference tool uses SNPs and their frequencies in different populations to estimate the likelihood that an individual is from a specific population. The results are visualized in the to facilitate the interpretation.

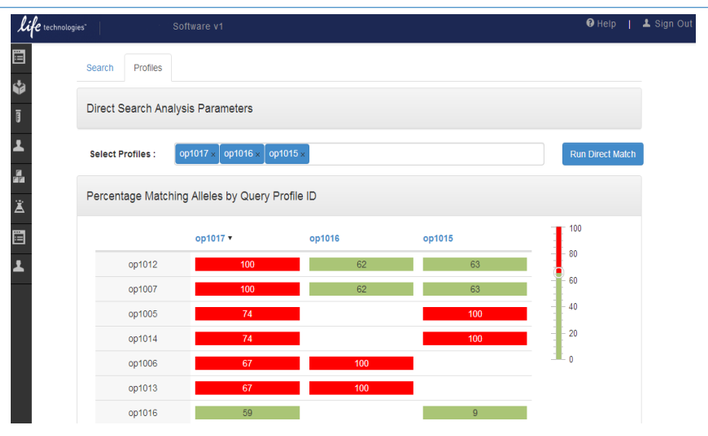

Figure 2. Current status of direct match

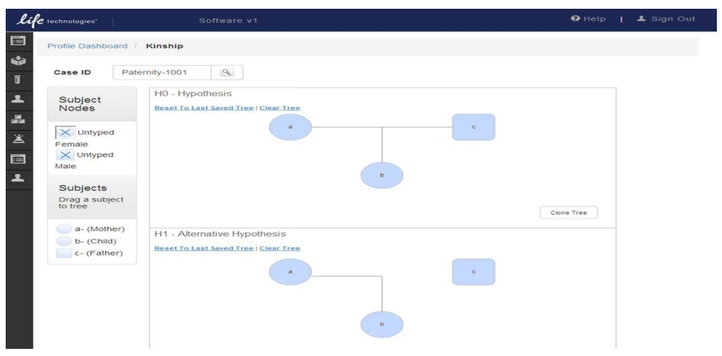

Figure 3. Current status of kinship analysis

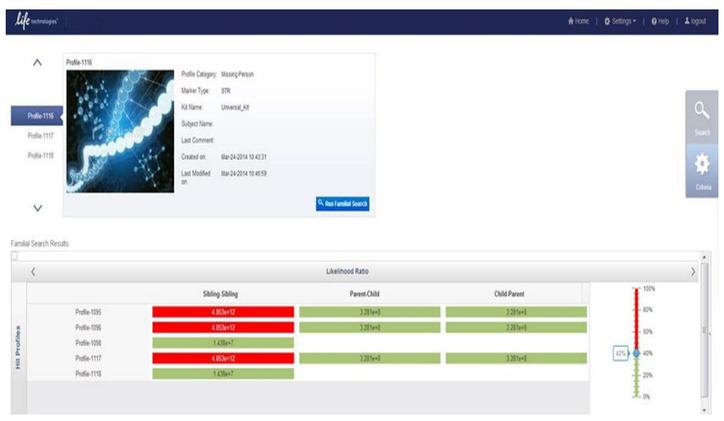

Figure 4. Current status of familial search

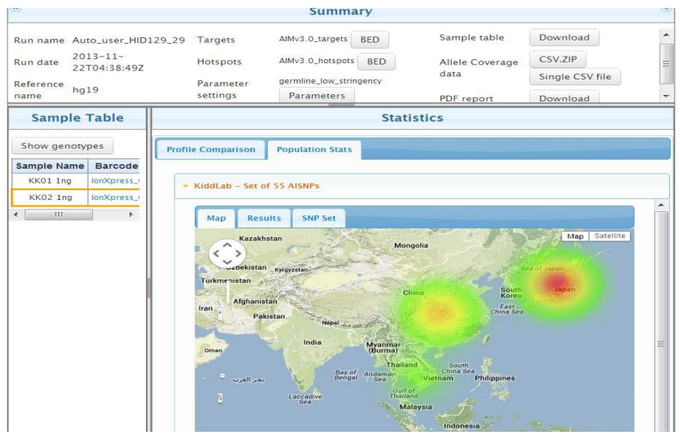

Figure 5. Current status of ancestry inference

Summary

The NGS data management and analysis platform allows the forensic laboratories a single integrated platform to centralize data creation, management, storage and analysis in the laboratory. It allows for high scalability and has been demonstrated to efficiently store and rapidly search sample databases containing millions of data records. The three major analysis tools used to support local database searching, relationship testing, and ancestry inference are highly configurable and enable users to define analysis settings, search parameters and define segmented profile databases. Other applications (e.g., tissues and fluids identification, human physical appearance, etc.) can also be supported and are under development.

Primary Author Contact Information:
Email: Jianye.Ge@lifetech.com, Ph: +001 650 872 7263

ThermoFisher SCIENTIFIC