Bioinformatic Framework Streamlines Forensic Genomics on the MiSeq® System

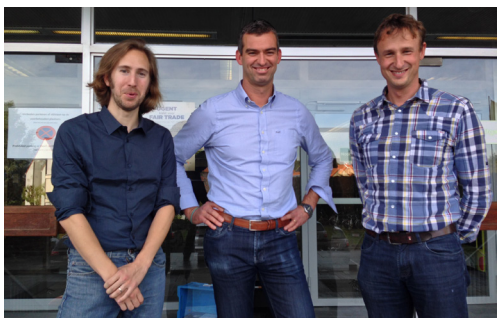
Developed by University of Ghent researchers, the MyFLq framework efficiently processes forensic DNA data without grouping to extract maximal information for automatically determined regions of interest.

Introduction

Researchers at Ghent University's Laboratory of Pharmaceutical Biotechnology are known for their proteome and genome research into the mechanisms of autoimmunity, yet it's not their only focus. The laboratory is also one of the leading centers of forensic genomics research in Belgium. On the surface, the two research areas may seem to have little in common, but they share common tools. "The genomics skill set we've developed for our disease research is enabling us to make significant advances in our forensics studies," according to Professor and Laboratory Director Dieter Deforce, Ph.D.

For more than 15 years, the laboratory has performed forensic analyses for Belgian police departments, moving from slab gel electrophoresis to capillary gel electrophoresis (CE) systems. After becoming familiar with next-generation sequencing (NGS) through genome, transcriptome, and proteome studies for their disease research, Dr. Deforce and his team decided to see if it held promise for forensics work. "Three years ago, NGS wasn't suited to perform forensic analyses," said Dr. Deforce. "We decided to investigate it for this application, with the hope that it might become a forensic tool in the near future."

With a focus on enabling the efficient analysis of NGS-generated forensic data, Dr. Deforce and his team began development of an open source software-based workflow tool. The current version of this tool is called the My-Forensic-Loci-queries (MyFLq) framework. Recently, it was used to analyze a MiSeq system data set of DNA mixtures[1] and it will soon be available in the BaseSpace® cloud computing environment.



The MyFLq workflow was developed by researchers at Ghent University's Laboratory of Pharmaceutical Biotechnology and led by Professor and Laboratory Director Dieter Deforce, Ph.D (middle), postdoctoral researcher, Filip Van Nieuwerburgh, Ph.D. (right), and Ph.D. student Christophe Van Neste (left).

iCommunity spoke with Dr. Deforce, postdoctoral researcher, Filip Van Nieuwerburgh, Ph.D., and Ph.D. student Christophe Van Neste about their research and the MyFLq workflow.

Q: What happened in 2010 that caused you to investigate NGS for forensics?

Dieter Deforce (DD): There are two things that are necessary for an NGS system to perform forensic analysis. The system must be capable of sequencing long read lengths to enable STR analysis, and it must be cost effective. In 2010, the Roche 454 GS FLX system was the only instrument that could deliver long read lengths, but it did so at a considerable sequencing cost and the turnaround time was long. Despite these drawbacks, we decided to move forward to see if it was technologically possible to perform forensic analyses on a 454.

Q: How did your team become involved in developing bioinformatics tools?

DD: Efficient bioinformatics tools are important for any application that generates a lot of data, something we recognized when conducting proteomics research. We realized that existing tools are not always sufficient, so for the last few years the bioinformaticians within our laboratory have collaborated to perform the hands-on work with our data sets.

Q: Why did you decide to continue development of your bioinformatics framework with the MiSeq system?

Christophe Van Neste (CVN): We initiated the work with the 454 as an exploration. It wasn't an ideal system for forensics. We knew that it wasn't going to work in practice. When the MiSeq system was announced, we realized that it offered a practical application of NGS in forensics that could be deployed in the near future."

DD: "We hoped that NGS technology would evolve, that the platforms would become less expensive and sequence long enough reads to enable STR analysis. The MiSeq system offers the long reads and fast turnaround times needed in a forensic laboratory. Its cost per sample is much lower than the 454.

Q: How did the 454 research inform development of the MyFLq framework?

CVN: Because the 454 has difficulty sequencing through homopolymer regions, we had to cluster the reads to produce a consensus sequence. There weren't many bioinformatics tools available at the time, so we created a bioinformatics software pipeline and used grouping tools made by other bioinformaticians[2].

From this experience, we realized that we should avoid clustering to obtain consensus sequences when investigating forensic mixtures. Clarity is lost, with SNPs or insertion/deletion alleles from a minor contributor resembling artifacts from major contributors and going undetected. PCR amplification errors can also resemble minor contributors. So it helped us understand what type of NGS system we needed and what data would be important to the forensic community.

> "The MiSeq system offers the long reads and fast turnaround times needed in a forensic laboratory. "

Q: What bioinformatic challenges did you face?

Filip Van Nieuwerburgh (FVN): The biggest challenge was determining how to parse out the sequences contributed by different individuals in a mixture. Looking at the whole-genome was unnecessary. We realized that we needed bioinformatics tools that looked at only the regions of interest, where individuals differ in their alleles for a locus. Any part outside of these regions was treated as a flanking region. We found that maximizing the flanking regions and removing them from reads eliminated noise and aided detection of alleles of low contributing donors.

Q: What makes up the MyFLq workflow?

FVN: It uses a MySQL database populated with reference alleles with automatically determined regions of interest, Python scripts to compare NGS STR sequences against the database, a method to assess the quality of an NGS-identified forensic locus, and another method to estimate whether an allele that isn't present in the database is a new allele or a sequencing error.

Q: Why did you choose to use open source software to develop MyFLq?

CVN: The goal was to create a straightforward application that was easy for anyone analyzing forensic data to use, not just bioinformaticians. Python is an easy programming language to understand. We chose it so that it could be applied in court. Anybody will be able to follow the workflow and understand what's being done. It's not hidden. The MyFLq framework has a Creative Commons open source license.

Q: What types of samples did you use in your study?

DD: We created DNA mixtures from two National Institute of Standards and Technology (NIST) standard reference materials (SRM 2391c: DNA A, DNA B) and three purified genomic DNA sources (K562, 9947A, 2800M; Promega). We used mixtures of four and five source DNAs, along with two single source (9947A, K562) samples. Amplicon libraries were generated from STR multilocus PCRs of the samples and sequenced on the MiSeq system; with FASTQ files generated automatically using MiSeq Reporter.

Q: How did the MiSeq system perform with the framework?

FVN: When we processed the 454 data, we had severe issues with the homopolymer regions and used a compression algorithm to eliminate any sequences with the same base. In the process, we lost some part of the information, but that's what we needed to do in order to retrieve the individual profiles.

The MiSeq system doesn't have a homopolymer issue. For some loci, there were issues in the flanks outside the region of interest. We solved that by using the compression algorithm in the flanks.

CVN: With the 454, the individual reads were not useful. The biggest advantage of the MiSeq system is that the raw reads are accurate enough to use them. It's absolutely necessary to have the individual reads. You learn so much more about the sample.

Q: Were you able to identify the different individuals within your sample mixture from the MiSeq data?

CVN: The minimal abundance threshold during data analysis was set to 0.5%. We were able to pick up the different contributors in the MiSeq data set.

DD: The experiment wasn't set up to determine the lower detection limit for minor alleles, so that's something we'd like to study in the future.

> "As part of making the MyFLq workflow easy to use, we're developing an Illumina BaseSpace app."

Q: What are the next steps in completing the workflow?

FVN: In addition to identifying the minimal abundance threshold, we're looking into defining the regions of interest more narrowly to exclude parts that carry no or little relevant information.

The forensics community expects to have a visual overview of samples, something we created manually for the paper. We're working on integrating visual forensic profiles into the workflow, so they are generated automatically.

Q: When will the MyFLq workflow be available in the BaseSpace cloud?

FVN: As part of making this workflow easy to use, we're developing an Illumina BaseSpace app that will be available by the beginning of 2014. Forensic researchers won't need to install any software to analyze their STR data files with the MyFLq BaseSpace app.

Q: How quickly can MyFLq analyze data?

CVN: The study was designed from a research point of view. We weren't really concerned with speed, so we analyzed everything

serially and it took about an hour. Once the MyFLq app is in BaseSpace, we will be able to analyze multiple samples in parallel and get results much faster than our current solution.

Q: What kind of reference database needs to be created to support the use of NGS in forensics?

DD: A reference allele database populated with sequences of all known STR alleles that exist in the general population. The only way to fill this database is to actually start sequencing samples. The MyFLq framework can be a resource in generating that database. It has a function to assess an unknown allele, determine whether it's in the database, whether it's a new allele or a sequencing error, and then add it into the database if it's a new allele. All of this can be performed in the cloud. That's why we felt it was important to create the BaseSpace app.

> "The MiSeq system doesn't limit us to the STR regions of interest, we can see everything between the Illumina adapters."

CVN: The reference allele database should also include the expanded information that we can see with Illumina technology. The MiSeq system doesn't limit us to the STR regions of interest, we can see everything between the Illumina adapters.

FVN: Creation of the database could begin now. We can fill the database with the *in silico* STR allele sequences that are known. It would be at least as good as the STR reference database or allelic ladders we use for CE.

Q. What is the value of using NGS for forensic DNA analysis?

FVN: With the multiplexing capability of NGS, it will be possible to combine the 50 or so different kinds of forensics kits that are now used in labs, such as STR, SNP, Y-STR, X-STR kits, into one assay. That will be great.

DD: I think NGS will become a real tool in forensic DNA analysis in the not too distant future. What's needed is the ability to analyze 48 or so samples a day, with a fast turnaround, and a low cost per sample—even lower than what is now attainable.

References

1. Van Neste C, Vandewoestyne M , Van Criekinge W, Deforce D, Van Nieuwerburgh F (2013) My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. Forensic Sci. Int. Genet.
2. Van Neste C, Van Nieuwerburgh F, Van Hoofstat D, Deforce D (2012) Forensic STR analysis using massive parallel sequencing. Forensic Sci. Int. Genet. 6: 810–818.

Illumina, Inc. • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 1470-2013-003 Current as of 12 November 2013

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACG
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA