ISDS 2014 Conference Abstracts

Development of Genomic Surveillance Bioinformatics Modules

Eishita Tyagi[1], R. C. Hopkins*[1], Dan Baker[1], King Jordan[2] and Shiyuyun Tang[1, 2]

[1]Booz Allen Hamilton, Atlanta, GA, USA; [2]Georgia Institute of Technology, Atlanta, GA, USA

Objective

To develop a modular approach to infectious disease genomic analysis that can easily integrate with public health analytics systems. Using dynamic approaches to genomic sequence analysis, relevant whole genome data can be quickly and accurately visualized and correlated, using a minimum of computational resources. We propose to develop visualization modules that integrate disparate data sources including integrate geospatial location metadata with associated epidemiological factors to enable faster outbreak identification and enhance surveillance.

Introduction

Whole-genome sequencing of disease-causing organisms provides an unabridged examination of the genetic content of individual pathogen isolates, enabling public health laboratories to benefit from comparative analyses of total genetic content. Combining this information with sample metadata such as temporal, geospatial, morbidity, and mortality can greatly increase the efficacy of genomics analysis. However, with the vast amount of data generated by such techniques, meaningful, rapid, and accurate analysis that interprets and correlates nucleotide polymorphisms for public health practice presents many challenges. To this end we have created a modular genomics analysis toolkit that can easily integrate diverse data streams and couple analysis with an array of visualization platforms.

Methods

Using open source tools we have assembled an analysis package that automatically processes next generation sequencing (NGS) data from the ubiquitous Illumina MiSeq. FastQ files are uploaded, filtered, trimmed and assembled. The largest contiguous DNA assemblies are BLASTed (Basic Local Alignment Search Tool) to determine closest reference genome match in RefSeq to identify the species of any isolate sequenced. After reference determination, a custom gene by gene typing algorithm calculates the core genome alignment required for phylogenetic evolutionary analysis. This approach is based on the whole genome multiple sequence typing (wgMLST) approach that was developed to define a rapid universal identification and typing scheme for pathogens. Alternative genomic methods used to process NGS data for evolutionary analysis rely on first calculating high quality single nucleotide polymorphisms (SNPs) for all sequenced isolates with respect to the reference genome and then creating a phylogeny. These approaches however can be computationally expensive as the number of sequenced isolates increases. Our algorithm attempts to overcome these computational bottlenecks through the more efficient gene by gene typing approach. Additionally, a key component of our algorithm is a rapid tree construction module where we calculate the minimal set of genes that can effectively recreate the ideal (core genome) phylogeny at a user accepted threshold of consensus identity.

Results

This toolkit provides an automated analysis suite for processing isolate sequencing data directly from the Illumina MiSeq. Utilizing a minimal core genome algorithm simplifies the data sets and reduces overall compute time for even large data sets. Additional modules being developed utilize open source tools and common sequence formats to integrate evolutionary analysis results from quality scored whole genome sequences with geographical data in order to provide geospatial visualization of distinct and related isolates in an outbreak. Output data from the PERL modules is seamlessly integrated into open source C++ Qt libraries prepackaged to perform geospatial visualization and relatedness clustering using multidimensional scaling (MDS) approaches. Platform independent Qt libraries provide a cross-platform application framework for easy integration of these “genomic surveillance” modules into existing surveillance applications. The virtual overlay of phylogenetic relationships onto isolate maps provides population structure in epidemiological studies and provides a mechanism for rapid real time analysis of transmission chains and effective retrospective analysis of pathogen evolutionary trends.

Conclusions

Utilizing and analyzing raw whole genome sequence data directly from the Illumina MiSeq moves current capabilities one step closer to real-time infectious disease characterization. Minimal core gene alignment analysis allows for computation on systems commonly available to infectious disease laboratories, circumventing the need for computationally expensive analysis. These genomic methods, if implemented within existing public health laboratory response programs, promise to revolutionize the ability of the laboratory to provide information and evidence on the evolution, transmission and virulence for pathogenic organisms.

Keywords

Bioinformatics; Epidemiology; Next Generation Sequencing

*R. C. Hopkins
E-mail: hopkins_robert@bah.com