The SNVPhyl Pipeline

a) Pipeline Overview

b) Expanded workflow for variant detection

SNVPhyl uses SMALT (https://github.com/ekg/smalt) and SAMTools’ BCFtools (http://samtools.github.io/bcftools/) for read mapping and variant calling. c) SNV Table

Additional visualization and analysis is provided by GenGIS (http://kiwi.cs.dal.ca/GenGIS/Main_Page), a phylogeographic visualization tool. GenGIS interacts with IRIDA via a REST API to download phylogenomic trees and geographic metadata. This image shows the relationship between *Vibrio cholerae* isolates associated with or closely related by (PFGE) match to the 2010 cholera outbreak in Haiti (http://mbl.asm.org/content/4/4/s0039b-13). Isolates from Haiti and the Dominican Republic are coloured in red, Nepalese isolates are coloured in blue, and all other countries are shown in gray. Solid lines connect each isolate from a given country to its corresponding node in the phylogenomic tree. Pie charts show the proportion of isolates from each country associated with each strain of cholera. Phylogeographic images such as this are a powerful means of conveying the genomic epidemiology of an outbreak. In this example, the GenGIS image illustrates a cholera transmission event from Nepal to Hispaniola followed by a clonal expansion signifying an outbreak.

The SNVPhyl Pipeline. a) Pipeline Overview. Each set of reads is mapped to the reference sequence and variants are called using multiple variant calling methods. Variants from each read set are consolidated into a master file and a status for each variant is assigned. Only SNVs in the core genome are retained for further processing. Repeat region analysis is performed and optionally masked from downstream analysis. Informative SNVs are selected and a multiple meta-alignment (MMA) is built. The MMA is used to infer the phylogenetic relationship of the sequences under investigation. Phylogenomic trees are built using PhyML (http://atgc.lirmm.fr/phyml/) and stored for later review. Additional quality control steps can be applied for each pipeline.

The SNVPhyl Pipeline: a) Pipeline Overview. Each set of reads is mapped to the reference sequence and variants are called using multiple variant calling methods. Variants from each read set are consolidated into a master file and a status for each variant is assigned. Only SNVs in the core genome are retained for further processing. Repeat region analysis is performed and optionally masked from downstream analysis. Informative SNVs are selected and a multiple meta-alignment (MMA) is built. The MMA is used to infer the phylogenetic relationship of the sequences under investigation. Phylogenomic trees are built using PhyML (http://atgc.lirmm.fr/phyml/) and stored for later review. Additional quality control steps can be applied for each pipeline.

**Quality Control**

IRIDA implements an initial QA / QC step for acceptance of reads into the sequence repository. Reports are generated by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and stored for later review. Additional quality control steps can be applied for each pipeline.

**Data Provenance**

IRIDA implements an initial QA / QC step for acceptance of reads into the sequence repository. Reports are generated by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and stored for later review. Additional quality control steps can be applied for each pipeline.

**Phylogeographic Visualization**

Additional visualization and analysis is provided by GenGIS (http://kiwi.cs.dal.ca/GenGIS/Main_Page), a phylogeographic visualization tool. GenGIS interacts with IRIDA via a REST API to download phylogenomic trees and geographic metadata. This image shows the relationship between *Vibrio cholerae* isolates associated with or closely related by (PFGE) match to the 2010 cholera outbreak in Haiti (http://mbl.asm.org/content/4/4/s0039b-13). Isolates from Haiti and the Dominican Republic are coloured in red, Nepalese isolates are coloured in blue, and all other countries are shown in gray. Solid lines connect each isolate from a given country to its corresponding node in the phylogenomic tree. Pie charts show the proportion of isolates from each country associated with each strain of cholera. Phylogeographic images such as this are a powerful means of conveying the genomic epidemiology of an outbreak. In this example, the GenGIS image illustrates a cholera transmission event from Nepal to Hispaniola followed by a clonal expansion signifying an outbreak.

**Data Provenance**

IRIDA implements an initial QA / QC step for acceptance of reads into the sequence repository. Reports are generated by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and stored for later review. Additional quality control steps can be applied for each pipeline.

**Phylogeographic Visualization**

Additional visualization and analysis is provided by GenGIS (http://kiwi.cs.dal.ca/GenGIS/Main_Page), a phylogeographic visualization tool. GenGIS interacts with IRIDA via a REST API to download phylogenomic trees and geographic metadata. This image shows the relationship between *Vibrio cholerae* isolates associated with or closely related by (PFGE) match to the 2010 cholera outbreak in Haiti (http://mbl.asm.org/content/4/4/s0039b-13). Isolates from Haiti and the Dominican Republic are coloured in red, Nepalese isolates are coloured in blue, and all other countries are shown in gray. Solid lines connect each isolate from a given country to its corresponding node in the phylogenomic tree. Pie charts show the proportion of isolates from each country associated with each strain of cholera. Phylogeographic images such as this are a powerful means of conveying the genomic epidemiology of an outbreak. In this example, the GenGIS image illustrates a cholera transmission event from Nepal to Hispaniola followed by a clonal expansion signifying an outbreak.

**Acknowledgments**

IRIDA project funding is gratefully provided by Genome Canada, Genome BC and the federal Genomics R&D Initiative (GRDI) with additional support by SFU. The findings and conclusions in this poster are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.