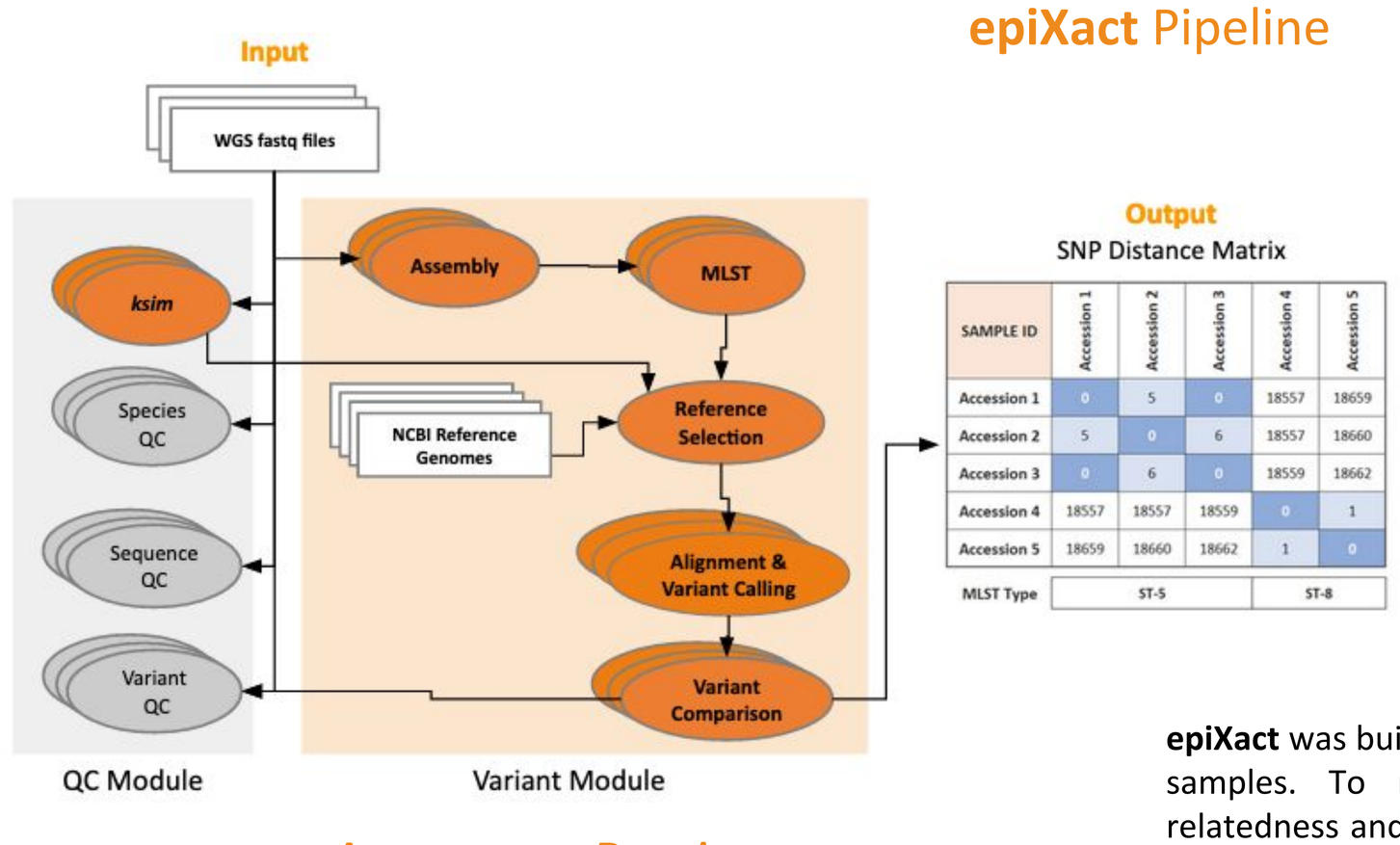
epiXact®: Rapid, precise and robust bacterial relatedness and outbreak detection from WGS data

Tim Farrell[1], Mohamad Sater[1], Ian Herriott[1], Febriana Pangestu[1], Jong Lee[1] and Miriam Huntley[1]
[1] Day Zero Diagnostics, Inc. (Boston, MA)

Introduction

Whole genome sequencing (WGS) is a well-established, high-resolution method for measuring pathogen relatedness to better understand infectious disease transmission. To date the lack of rapid, precise, and reliable computational workflows has been a major obstacle to the adoption of WGS for routine clinical applications. Single nucleotide polymorphism (SNP)-based analyses provide the highest resolution for measuring relatedness of bacterial pathogens. However, these methods can be difficult to implement with the reliability, speed, and scale needed to inform infection control decision-making. These obstacles become more significant for genomic surveillance systems, which require longitudinal analysis of larger numbers of samples. To enable the use of WGS for real-time determination of infectious disease outbreaks, we have developed **epiXact®**, an automated computational workflow that can rapidly and robustly detect pathogen relatedness from WGS data.

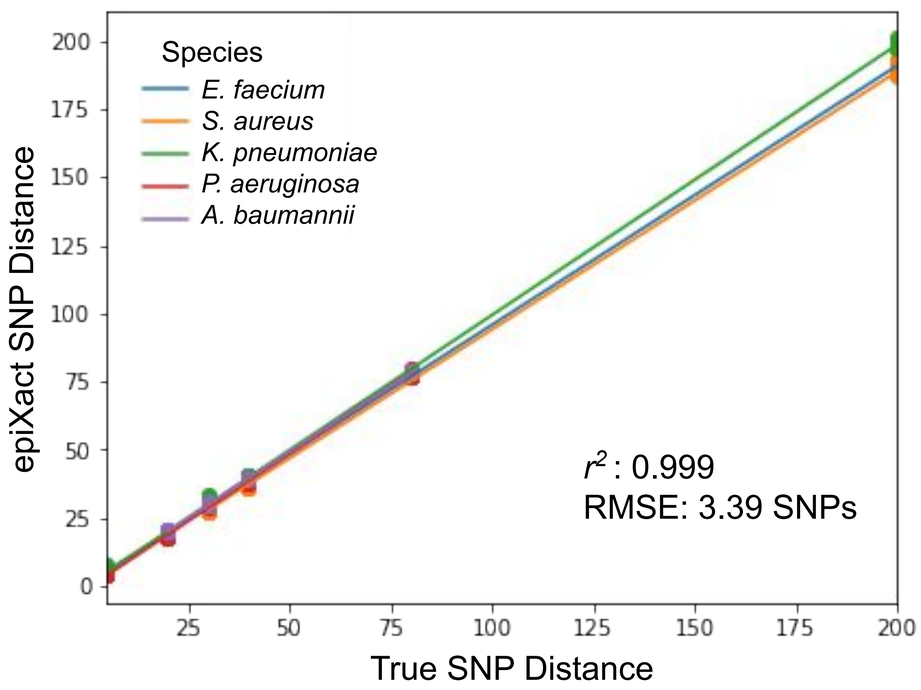
epiXact Pipeline



The **epiXact** pipeline begins with Illumina WGS data for the set of all bacterial samples suspected of involvement in an outbreak. The pipeline computes variants for each sample relative to a closely matching reference genome, and SNPs are compared between pairs of samples to quantify relatedness. The pipeline's final output is a matrix containing the SNP distances between all sample pairs. Samples with few SNP differences (highlighted in blue) are considered clonal while samples displaying large genomic distance can be ruled out from being related in an outbreak. Infection control specialists can use this to inform outbreak control measures.

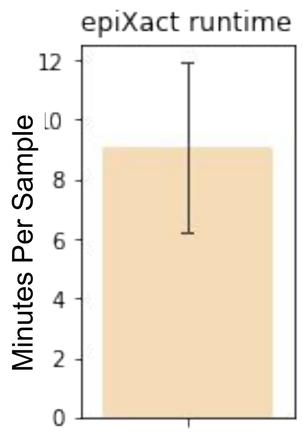
Accurate + Precise

We performed an *in silico* validation of epiXact on a set of 40 synthetic samples generated across 5 species. For each species we selected closed reference genomes from NCBI, introduced low numbers (5-200) of synthetic SNPs, and then generated synthetic Illumina reads from those genomes (with and without introduced SNPs). epiXact was used to determine the pairwise SNP distance between samples. We found epiXact achieved **high accuracy** (r^2: 0.999) and low error (*RMSE*: 3.39 SNPs) across the five species.



Fast + Scalable

epiXact demonstrated speed and scalability, capable of parallelized analysis over large datasets. Validation samples were processed using local pipeline execution in 6-12 minutes, with 9.1 minutes on average. Preliminary experiments with cloud native execution achieved ~2X faster processing times using cloud computing. Related experiments show the pipeline is capable of scaling to datasets with 500+ samples.



Robust

epiXact was built to withstand a variety of challenges expected of real-life samples. To minimize biases associated with reference genome relatedness and maximize resolution, epiXact uses a combined reference genome and *de novo* assembly approach. The pipeline makes use of ***ksim***, an in-house algorithm that we developed to rapidly compute relatedness between samples using a kmer similarity metric. *ksim* is used to select the most similar available reference genome for each isolate pair. When no suitable reference is available, the pipeline uses the *de novo* assemblies generated from the samples themselves as references, effectively enabling analysis that is agnostic to strain or species type. We have also outfitted epiXact with methods that make it robust to recombination events and mobile genetic elements, both of which commonly confound relatedness estimates. Quality control (QC) is also incorporated into the pipeline to automate checks for contamination and sequencing quality, and to cross-check SNP distance results using kmer relatedness via *ksim.*

Clinically Relevant

We have investigated a wide variety of suspected outbreaks (in both clinical and laboratory settings) at the request of partnering institutions, using the **epiXact** pipeline to determine sample genomic relatedness. In 24 recently examined suspected outbreaks, bacterial samples were sent to Day Zero Diagnostics for clonality assessment. Across these cases we sequenced and analyzed a total of 116 bacterial samples encompassing 12 species/pathogen types (e.g. MRSA, ESBL, CRE). 16 (66.6%) of the 24 cases were found to contain one or more clonal clusters, providing evidence for an outbreak, while the remaining 8 cases did not. In all cases we reported back results within 24-48 hours from sample receipt. Infection control specialists were able to use the clonality determination results to inform decisions regarding ward cleaning, staff screening, and equipment contamination.

Contact information:
Mohamad Sater, PhD mohamad@dayzerodiagnostics.com
Tim Farrell, MS tim@dayzerodiagnostics.com
epiXact Team epiXact@dayzerodiagnostics.com

DAYZERO DIAGNOSTICS