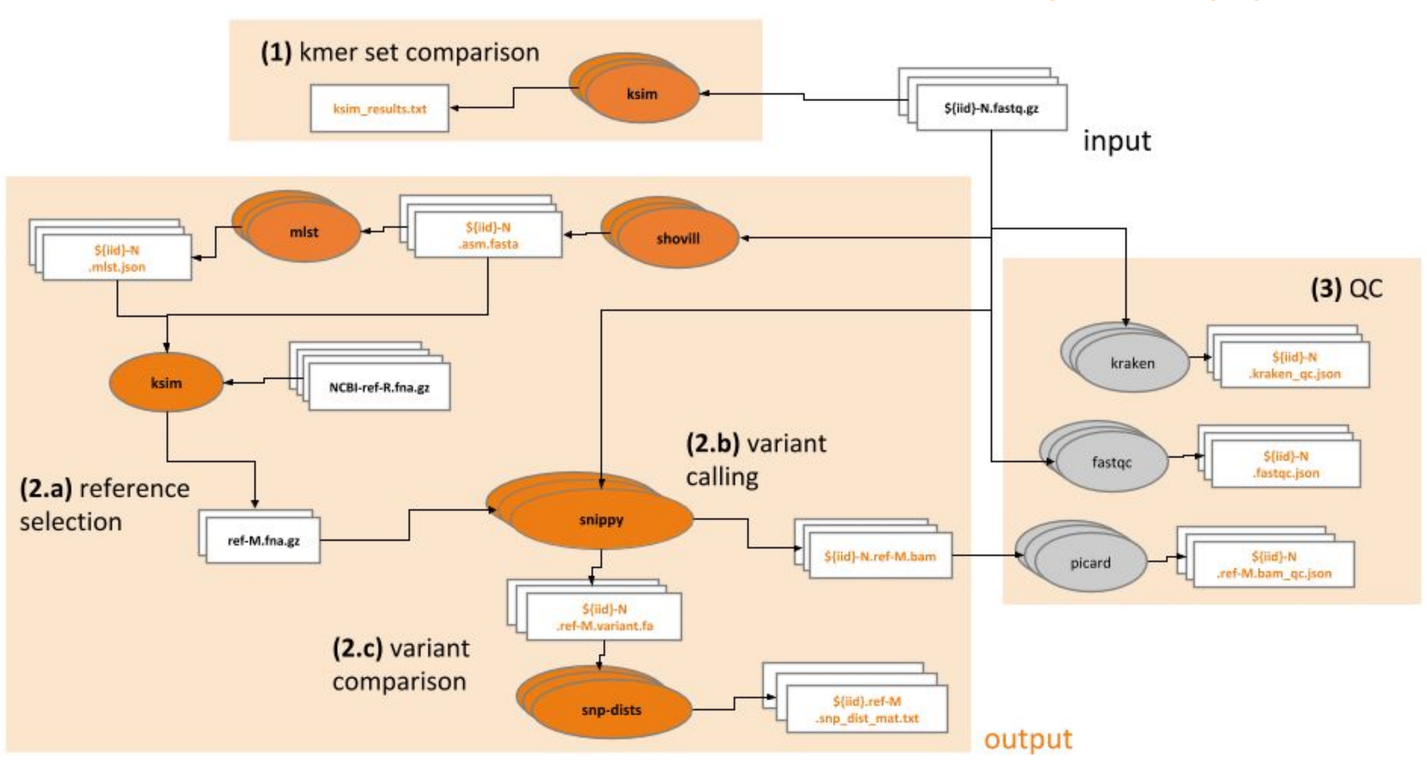
epiXact: Rapid, precise and robust bacterial relatedness and outbreak detection from WGS data

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 WGS being adopted more routinely in clinical settings. 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 analyzing larger numbers of samples over extended periods of time. 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



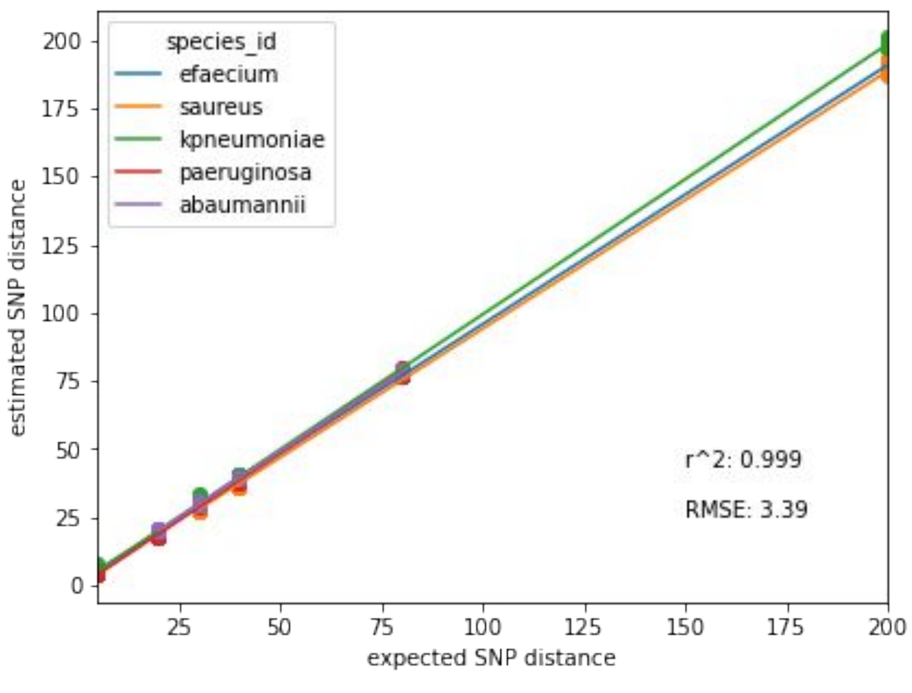
The **epiXact** (v2.1) pipeline consists of 3 (largely independent) modules:

1. kmer set comparison/ similarity
2. reference selection + variant calling/ comparison
3. quality control (QC)

Modules (1) and (3) are fast and primarily used to support results from module (2), which computes the primary results of the pipeline. In module (2), variants are called for each sample against a set of reference assemblies, selected based on genomic similarity to the sample set, and then compared between sample pairs to quantify relatedness at SNP-level resolution. The end result is a matrix containing pairwise SNP distances between all sample pairs.

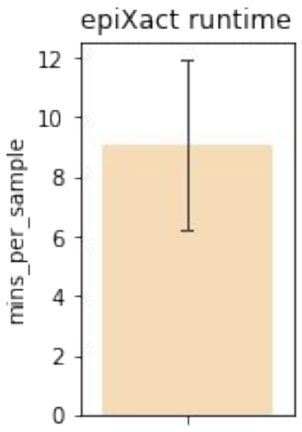
Accurate + precise

During *in silico* validation on a set of 40 samples across 5 species, **epiXact** achieved high accuracy (r^2: 0.999) and low error (*RMSE*: 3.39) when comparing the number of SNPs observed vs. the number expected. Samples for this validation were generated by introducing low numbers (0-200) of synthetic SNPs into closed NCBI genomes, generating synthetic Illumina reads from those genomes and then comparing read sets with SNPs to those without SNPs using the epiXact pipeline.



Fast + scalable

epiXact (v2.1) was designed for speed and scalability, capable of parallelized analysis over large datasets. Validation showed samples processed within 9.1 mins on average. However, this was achieved using local pipeline execution. Preliminary experiments with cloud native execution show we can easily achieve ~2X faster processing times using cloud computing. Related experiments show the pipeline is capable of scaling to datasets with 500+ samples.



Clinically relevant

To date, we have investigated 24 suspected outbreaks (in both clinical and laboratory settings), where we generate Illumina WGS data from isolates sent to us by partnering institutions and use the **epiXact** pipeline to estimate sample genomic relatedness. In total, we have analyzed 116 samples across 12 species types (e.g. MRSA, ESBL, CRE, etc.) and have detected clonal outbreaks in 66.6% of cases. Additionally, in all cases, we have reported back results in less than 48 hours.

# cases	# samples	# species types	# clonal outbreaks	# non-clonal outbreaks
24	116	12	16 (66.6%)	8 (33.3%)

Robust

To minimize biases associated with reference genome relatedness and maximize resolution, we equip **epiXact** (v2.1) with a combined reference alignment and *de novo* assembly approach. We use a kmer-based reference selection method (supported by a tool we developed called **ksim**) to select the most similar publicly 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 the pipeline with methods that make it robust to recombination events and mobile genetic elements, both of which commonly confound relatedness estimates.

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

DAYZERO DIAGNOSTICS