`SplitStrains`, a tool to identify and separate mixed *Mycobacterium tuberculosis* infections from WGS data

Authors: Einar Gabbasov, Miguel Moreno-Molina, Iñaki Comas, Maxwell Libbrecht, Leonid Chindelevitch.

Introduction. Bacterial infections by pathogens such as *Mycobacterium tuberculosis* and *Clostridium difficile* often occur as mixed infections, whereby a single patient is infected by several different strains of the same organism. The identification of such mixed infections can be important for reasons including both patient-level decisions as well as public health measures. In the latter setting, if the tracing of the origins of the mixed infection is needed, it may be additionally required to separate the mixed infection into its constituent strains. The separation may also be informative when the mixed infection is heteroresistant, namely, when some, but not all, the strains are resistant to a particular antimicrobial drug. Moreover, a failure to identify the within-host pathogen diversity can lead to misdiagnosing a relapse and reinfection. However, so far, the problem of identifying mixed infections and separating them into their constituent strains has not received a sufficient amount of attention in the literature.

Although older techniques based on the detection of specific regions, such as VNTR (variable-number tandem repeats), are often able to detect such a mixed infection, this is not always the case with next-generation sequencing. The main challenge is that the presence of two alternative alleles in a given genomic position may signal a sequencing error as well as the presence of multiple strains. The key distinguishing feature of a mixed infection is the consistency of the fraction of the sample attributable to the sub-dominant strain(s) across most of the variable positions. Thus, depending on the depth of coverage, the similarity between the constituent strains and the proportions in which they are mixed, the problem of detecting and separating mixed strains may vary from straightforward to nearly infeasible.

Several methods for this problem have appeared over the past decade. Eyre et al. propose a `Mixed Infection Estimator`, a two-step maximum likelihood-based approach for mixture proportion estimation and mixed strain identification using a custom database (built for MLST based sequence types). Even though the paper presents results for *C. difficile*, the mixture estimation algorithm can be generalized to other pathogens such as *M. tuberculosis.* In order to differentiate pure and mixed infections the method computes a deviance statistic and uses it as a threshold for confirming mixed infection. However, the proposed algorithm can resolve at most two strains and there is no readily available compatible *M. tuberculosis* database. More recently, Sobkowiak et al. developed `MixInfect`, a method for mixed strain proportion estimation using a Bayesian model-based clustering technique. To distinguish between mixed and pure samples the tool measures the proportion of heterozygous calls to total SNPs and uses it as a thresholding value. While the algorithm can estimate mixture proportions it does not provide any functionality for resolving the constituent strains. The most recent method, `QuantTB` by Anyansi et al, relies on a specially constructed publicly available database of 2166 *M. tuberculosis* assemblies gathered from NCBI. The method provides mixture estimates of WGS samples as well as the identification of strains whose sequence is similar to at least two strains included in the database. To determine the constituent strains, the SNPs from a sample are compared against SNP sequences in the reference database. Based on the presence scores of every genome in the database, the algorithm determines how many constituent strains are present in a sample. However, such an approach does not generalize well to new data in situations where the underlying strains are not represented in the database, and thus its performance is highly dependent on the coverage provided by the database.

In this paper, we address this problem with a tool called `SplitStrains`, grounded in a rigorous statistical framework. It is based on formulating, for a given set of WGS reads, two alternative hypotheses, namely: the reads belong to a single strain (null hypothesis) or to a mixture of two or more strains (alternative hypothesis). We then use Expectation-Maximization algorithm to estimate the parameters of both hypotheses, and compare their likelihoods to draw a conclusion. As a result, we simultaneously obtain

- A call to decide whether the sample represents a simple or a mixed infection,
- A likelihood ratio (between the alternative and the null hypothesis) for the call, and
- If mixed, the proportion of each constituent strain and its identity defined by its SNPs (single-nucleotide polymorphisms) relative to a reference genome.

Our results on both simulated and real data show that `SplitStrains` is effective at identifying mixed infections even at a low depth of coverage (50X) and low genetic distance (100 SNPs) between strains. Moreover, `SplitStrains` outperforms previously published tools `Mixed infection estimator`, `MixInfect` and `QuantTB`. Furthermore, our results show that `SplitStrains` accurately separates the constituent strains provided that their proportions are not too close to each other and they are not too similar.

Methods

Algorithm workflow. For simplicity, a sample obtained via Whole Genome Sequencing will be called *mixed* if it contains multiple strains of the sequenced organism, and *pure* otherwise. The splitStrains algorithm classifies a sample as being mixed or pure. In the case when a sample is mixed, the algorithm detects the proportion of each strain and separates the reads according to which strain they belong to. In order to accomplish this, the algorithm proceeds through three stages.

First, SplitStrains uses the sample's single nucleotide polymorphisms (SNPs) to infer the parameters of a Gaussian or Binomial Mixture Model, which identifies the number and the proportions of the constituent simple strains. The likelihood ratio statistic produced in the process provides a rigorous quantification of the confidence about its status as a pure or mixed sample. The algorithm then uses the model's estimated parameters in a Naive Bayes classifier to assign each read to one strain. Finally, it produces Sequence Alignment/Map files for each constituent strain. The process is shown in Figure 1.

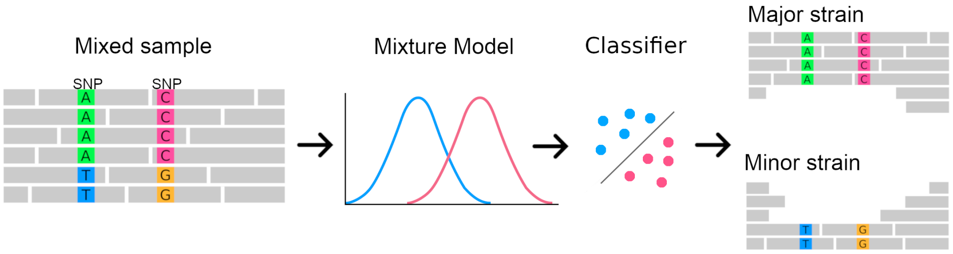

Figure 1: **SplitStrains workflow overview.**

Results

Mixed samples detection. We evaluate the performance of the `SplitStrains` algorithm by measuring its mixture proportion estimation and strain separation accuracy on 3 datasets with known strain proportions: *in vitro* artificially mixed samples and 2 *in silico* artificially generated and mixed samples. `SplitStrains` is able to correctly classify 91% of all samples.

Mixture proportion estimation. For each sample that is classified as mixed, we estimate the major strain proportion and compare it with the true proportion Figure 2. In general, the estimation is accurate up to a 90% major strain frequency, but starts to decrease as this frequency approaches 95%.

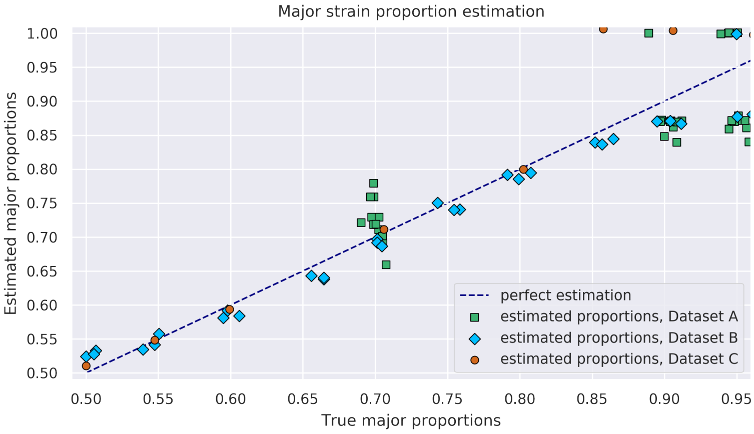

Figure 2: **Proportion estimation.** 74 mixed samples with their major proportion estimates.

Assignment of reads to constituent strains. Once the mixture model parameters have been estimated, the algorithm assigns each read containing one or more variable sites to a constituent strain using a Naive Bayes approach. Note that those reads that do not contain any variant sites or have zero mapping quality remain unassigned (i.e. we perform a partial, rather than complete, strain reconstruction). In Figure 3 below we respectively present a representative two-strain and three-strain confusion matrices to show the performance of this assignment.

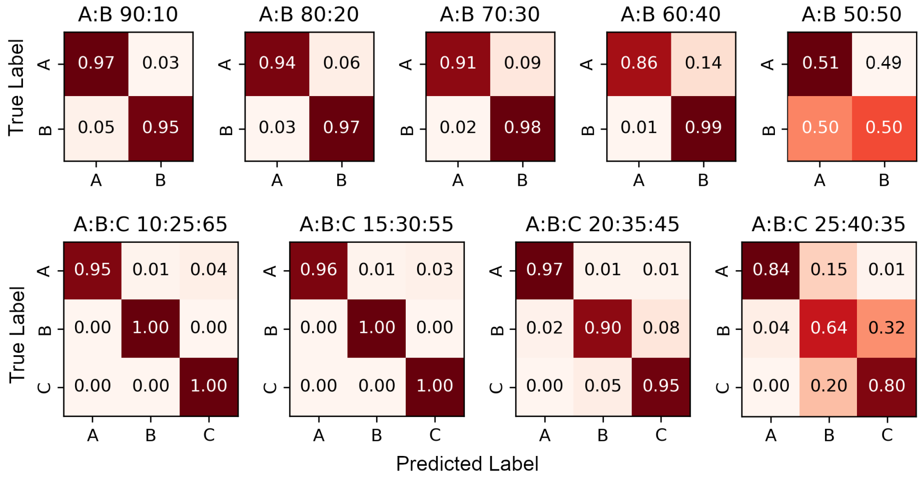

Figure 3: **Confusion matrices for 2-strain and 3-strain samples.** We denote each strain as A, B and C with the respective proportions displayed above each figure.

Strain genome reconstruction. Using the read assignments to the strains, the algorithm outputs a new alignment file for each strain. In order to further evaluate the accuracy of the assignment, we create a consensus sequence from each alignment file. We expect the consensus sequences to match the respective genomes of the constituent strains. As the genome of each constituent strain has the same number N of base substitutions relative to the reference genome, due to the way they are generated, the consensus sequences can have between 0 and N mismatches with the true sequences. In the case of the two-strain mixtures, our algorithm successfully separates the strains with major strain proportion varying from 55% to 90%. However, as the major strain proportion gets closer to 50%, correct assignment of reads becomes steadily more challenging, as shown in Figure 4.

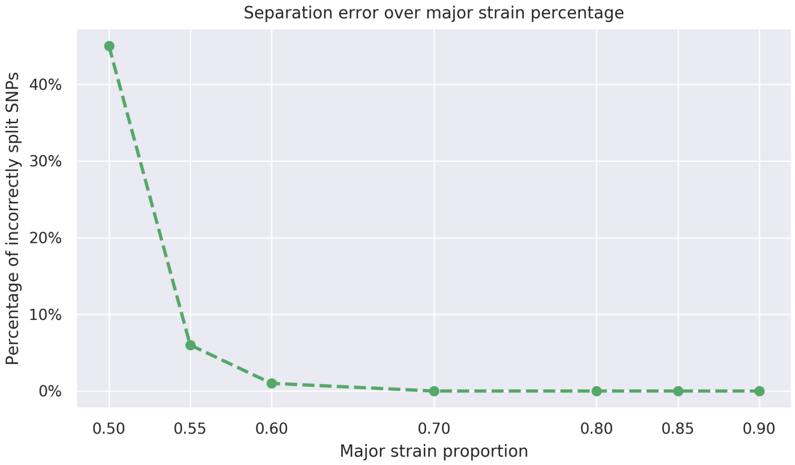

Figure 4: **Assignment error**. The proportion of mismatches due to the incorrect assignment of reads, among the bases where the strains differ from one another.

Comparison with other tools. `SplitStrains` consistently outperforms `MixInfect`, `QuantTB` and `Mixed Infection Estimator`.The Receiver Operating Characteristic (ROC) curve (Fig 5) shows the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The Area Under the Curve (AUC) quantifies how well the algorithm is able to distinguish between pure and mixed infections. The higher the AUC, the better the algorithm is at predicting the class of a sample. `SplitStrains` has the highest AUC (0.99) and can achieve close to 100% TPR with an FPR as low as 11%.

SplitStrains also has a proportion estimation error on each dataset that is consistently the lowest or second lowest among the tools, and it has the lowest error on the combined dataset. These results are summarized in Table 1.

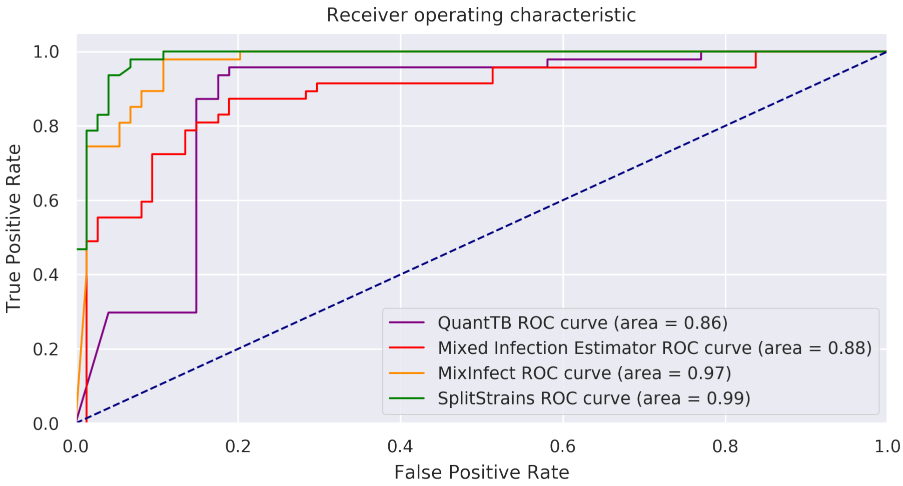

Figure 5: **ROC curves for the four tools**. SplitStrains achieves a higher area under the ROC curve than the other methods.

Table 1: **Root Mean Squared Error comparison across all datasets.**

Dataset	Size	SplitStrains	Mixed Infection Estimator	MixInfect	QuantTB
A	48	**0.056**	0.068	0.178	0.153
B	60	0.025	**0.018**	0.031	0.202
C	22	0.066	0.066	**0.041**	0.312
Combined	130	**0.047**	0.053	0.126	0.196

Conclusion In this abstract we introduced a novel algorithm, called `SplitStrains`, based on a rigorous statistical framework, for detecting multiple-strain infections, estimating the proportion of the major and minor strains, and partially reconstructing their sequences by assigning the reads that contain variants to one of these strains. In addition, `SplitStrains` is unique among existing methods in its ability to provide additional information, namely, the assignment of each read to one of the underlying strains, with a subsequent identification of their sequence if desired. Importantly, unlike `QuantTB` it does not rely on the knowledge of a large number of previously identified sequences, which is a clear advantage when investigating either a novel outbreak or an isolate originating from a data-poor setting. Furthermore, `SplitStrains` returns not only a call, but also a likelihood ratio, which is an indicator of the algorithm's confidence about the presence or absence of a mixed infection. We believe that, in situations where such information has either clinical or public health importance, the `SplitStrains` method will be a valuable addition to the existing collection of tools. In future work, we plan to extend `SplitStrains` to work with other bacterial pathogens as well as to improve its resolution, at least in datasets with high depth of coverage. Lastly, we plan to use `SplitStrains` as a preprocessing step in two pipelines - one for identifying related isolates in an outbreak, where mixed infections can mask such relatedness, and another one for predicting drug resistance, where mixed infections can impede a correct prediction when only the minor strain is drug-resistant.