RESEARCH

Deconvoluting the Diversity of Within-host Pathogen Strains in a Multi-Locus Sequence Typing Framework

Guo Liang Gan[1*], Elijah Willie[1*], Cedric Chauve[2] and Leonid Chindelevitch[1†]

Full list of author information is available at the end of the article
[*]Equal contributor [†]Corresponding author

Abstract

Background: Bacterial pathogens exhibit an impressive amount of genomic diversity. This diversity can be informative of evolutionary adaptations, host-pathogen interactions, and disease transmission patterns. However, capturing this diversity directly from biological samples is challenging.

Results: We introduce a framework for understanding the within-host diversity of a pathogen using multi-locus sequence types (MLST) from whole-genome sequencing (WGS) data. Our approach consists of two stages. First we process each sample individually by assigning it, for each locus in the MLST scheme, a set of alleles and a proportion for each allele. Next, we associate to each sample a set of strain types using the alleles and the strain proportions obtained in the first step. We achieve this by using the smallest possible number of previously unobserved strains across all samples, while among the unobserved strains we try to use those which are as close as the observed ones as possible, at the same time respecting the allele proportions as closely as possible. We solve both problems using mixed integer linear programming (MILP). Our method performs accurately on simulated data and generates results on a real data set of *Borrelia burgdorferi* genomes suggesting a high level of diversity for this pathogen.

Conclusion: Our approach can apply to any bacterial pathogen with an MLST scheme, even though we developed it with *Borrelia burgdorferi*, the etiological agent of Lyme disease, in mind. Our work paves the way for robust strain typing in the presence of within-host heterogeneity, overcoming an essential challenge currently not addressed by any existing methodology for pathogen genomics.

Keywords: Multi-Locus Sequence Typing; Bacterial diversity; Integer Linear Programming

1 Background

The study of bacterial pathogens has revealed an impressive genetic diversity that had not been fully suspected prior to the advent of genome sequencing technologies. This diversity may indicate an adaptive response to challenges such as the variability in host genetics, environmental conditions, and, in the case of pathogens affecting humans, the introduction of antibacterial drugs [1, 2, 3, 4].

One bacterial pathogen that is particularly well-known for its genetic diversity is *Borrelia burgdorferi*, the etiological agent of Lyme disease. It has been found that up to six genetically different strains can affect a single host [5, 6]. Furthermore, this diversity may result from both clonal evolution within the host as well as multiple infection events [7]. Unfortunately, techniques such as bacterial culture are

difficult to apply to reveal the whole range of diversity in bacteria like *B. burgdorferi*, a situation common to many bacterial pathogens. Next-generation sequencing (NGS) techniques such as whole-genome sequencing (WGS) with short reads have revolutionized our ability to investigate the genomic diversity of bacteria and other organisms [8]. Recently, an adaptation of WGS technology to *B. burgdorferi*, called whole-genome capture, has been proposed which is able to reliably filter out irrelevant DNA (such as host DNA) [9]. This novel approach for generating sequence data for *B. burgdorferi* nicely complements a highly reproducible strain-typing scheme known as multi-locus sequence typing (MLST), which has been developed and found to be useful for different pathogens in a number of contexts [10]. MLST is a summary of the bacterial genotype using the alleles of several (typically 6 to 9) housekeeping genes, which may be further grouped into closely related strain types. In the case of *B. burgdorferi*, several hundred strain types have been characterized using the MLST scheme developed in [11], while only 56 fully sequenced *B. burgdorferi* genomes[1] are currently available in the NCBI databases. MLST strain types thus provide a finer-grained picture of the strain diversity of this pathogen, which motivates the need for developing novel diversity estimation methods that combine NGS data and the wealth of strain types already characterized by MLST.

In principle, this problem is a special instance of estimating the diversity and abundance of microbial strains from metagenomics data, a problem for which several accurate methods have recently been developed (e.g. [12, 13]). *De-novo* methods, such as DESMAN [12], cannot take advantage of known reference strains, alleles and are likely to be confounded by the high similarity observed between strain types. Other methods such as strainEST [13] can consider a large set of reference genomes, which in our case can be defined by the concatenated allele sequences of the known *B. burgdorferi* strains types, but again, their diversity models are not well adapted to handle the very high similarity between strain types. Moreover, none of the reference-based methods consider the detection of novel strain types.

We introduce the first paradigm for extracting MLST information in the presence of within-host heterogeneity, which is also able to simultaneously take multiple samples into account and detect novel strains. Our method is based on mixed integer linear programming (MILP), and consists of two main stages. It starts by filtering the short reads in each sample, selecting those that closely match known alleles in at least one of the housekeeping genes in the MLST scheme, and then assigns fractional abundances to each allele of each gene, ensuring that as few such alleles as possible are used that can explain the data well. In the second stage, it assigns combinations of these alleles, with corresponding proportions, to each sample, while maximizing the usage of known strains and minimizing the number of novel strains, a parsimony-based approach that has been shown to perform well in related contexts [14].

We evaluate our approach on simulated samples and find that it is accurate in identifying both the fractional allele composition at each housekeeping gene, as well as the complete strain types present in each sample. We then apply it to a dataset of 24 real tick samples containing *B. burgdorferi* extracted via whole-genome capture, and find a substantial amount of diversity, as well as a number of new strains. In conclusion, our work provides a robust and reproducible pipeline for accurate strain

[1] `https://www.ncbi.nlm.nih.gov/genome/genomes/738`, accessed January 30, 2018.

typing via MLST is achievable from WGS data even in the presence of substantial within-host heterogeneity.

2 Methods

Terminology. An *MLST scheme* is composed of a set of loci together with a database of known alleles for each locus [15]. An *allele distribution* for a given locus is a set of alleles for this locus together with a proportion assigned to each allele; the proportions must be non-negative and add up to 1. A *strain type* is an assignment of a specific allele to each gene of the MLST scheme. A *strain type distribution* is a set of strain types together with a proportion assigned to each strain type; the proportions must once again be non-negative and add up to 1. A *sample* is a WGS dataset obtained from a single host that contains the sequence data from one or several pathogen strains present in the host (see Fig. 1).

Data. In the present work we use the traditional *B. burgdorferi* MLST scheme [11] composed of 8 housekeeping genes having a combined total of 1,726 known alleles[2]. For each locus, the various known alleles differ from one another primarily by single nucleotide polymorphisms (SNPs), with small indels also appearing in 4 out of the 8 genes. The number of known strain types is 753.

Problems and contribution overview. The problems we address in this work take as input (1) an MLST scheme together with databases of known alleles and strain types and (2) WGS data for a set of samples, that are mapped, using a short-read mapper of choice, onto the database of known alleles for the provided MLST scheme. It then proceeds in two stages, each addressing a specific problems:

- The Allele Diversity Problem. For a given sample and a given locus of the MLST scheme, given the mappings of DNA reads onto the known alleles for this locus, detect the alleles present in the sample and the corresponding allele distribution.
- The Strain Diversity Problem. Given a set of samples and an allele distribution for each locus at each sample, compute a strain type distribution per sample that requires the smallest number of novel strain types *among all considered samples*, which are closely distant from known strains.

2.1 The Allele Diversity Problem

We formulate the problem of allele detection as a variant of the Set Cover problem as follows. The input of the Allele Diversity Problem (ADP) is composed of a set of m reads $\mathcal{R} = \{r_1, \dots r_m\}$, a set of n alleles $\mathcal{A} = \{a_1, \dots, a_n\}$ for the chosen locus, and a set of mappings of the reads onto the alleles, encoded by a matrix M, where m_{ij} is the sum of the normalized Phred scoresof the mismatched bases in the mapping of read r_i onto allele a_j (we set it to ∞ if r_i does not map onto a_j). For instance, if read r_i maps to allele a_j with 2 mismatches with base quality scores of 60 and 80, respectively, then $m_{ij} = \frac{60-33}{126-33} + \frac{80-33}{126-33} = 0.796$. Each allele a_j implicitly defines a subset of $\mathcal{R}$ (the reads aligning with the allele), with each read r_i being weighted by m_{ij}. Informally, we then aim at selecting a subset of alleles covering the set of

[2] `https://pubmlst.org/borrelia`

reads, while minimizing the sum of the number of required alleles and the sum of the corresponding weights. The ADP is thus very similar to the Uncapacitated Facility Location Problem, and we discuss this observation in Supplementary Material.

Formally, we define an edge-weighted bipartite graph whose vertex set is $\mathcal{R} \cup \mathcal{A}$ and whose incidence matrix is M. A *read cover* is a subset of edges of this graph such that each read belongs to exactly one edge; the cost of a read cover is the number of allele vertices it is incident to plus the sum of the weights of the edges in the cover. The ADP aims at finding a read cover of minimum weight, the allele vertices incident on the edges of the cover representing the selected alleles.

Theorem 1 *The Allele Diversity Problem is NP-hard.*

The proof of Theorem 1 relies on a reduction from the 3-dimensional matching problem and is provided in Supplementary Material. Before describing our ILP we comment on the relevance of our formulation for selecting a set of alleles from short reads. Our objective function aims to minimize the sum of the number of alleles and the weight of each read based on the Phred scores; the latter part aims at explaining the data (reads) using as few errors/mismatches as possible, accounting for the base quality score of the mismatches, while the former part ensures that an allele is not introduced unnecessarily to reduce the contribution of the mismatches and their quality for a small number of reads. Our experiments on simulated data show that this objective function leads to extremely accurate results.

An Integer Linear Program for the Allele Diversity Problem. First we introduce the following notation: $R_j = \{r_i : m_{ij} \neq \infty\}$ represents the set of reads mapping onto allele a_j (i.e. covered by allele a_j), and $M_i = \{m_{ij} | 1 \leq j \leq n\} - \{\infty\} = \{q_{i1}, ..., q_{i|M_i|}\}$ represents the distinct summed Phred scores for read r_i. The decision variables of the ILP are:

- $x_j = 1$ if allele a_j is chosen, and 0 otherwise.
- $y_{ik} = 1$ if a mapping of read r_i with score q_{ik} is chosen, and 0 otherwise.

The objective function is $\min \left(\sum_{i=1}^{|\mathcal{R}|} \sum_{k=1}^{|M_i|} q_{ik} \cdot y_{ik} + \sum_{j=1}^{n} x_j \right)$.

Finally, the constraints of the ILP are the following ones:

- If $y_{ik} = 1$, there exists some allele a_j onto which r_i maps with score q_{ik}.
- There is a unique score with which read r_i is mapped onto the selected alleles.

These constraints can be translated as follows:

$$\sum_{\{j \mid r_i \in R_j, m_{ij} = q_{ik}\}} x_j \geq y_{ik} \, \forall \, i, k \qquad \sum_{k=1}^{|M_i|} y_{ik} = 1 \, \forall \, i.$$

Post-processing. If the above 0-1 ILP has multiple optimal solutions, we resort to a likelihood based method to select one, precisely GAML [16], a probabilistic model for genome assembly. Given a set of solutions where each solution represents a set of alleles, we measure the likelihood of observing the set of reads given a solution and pick the solution which maximizes the likelihood criterion. If there are multiple solutions maximizing the likelihood criterion, we pick one arbitrarily.

Computing allele proportions. Finally, once the alleles have been identified for a given locus, we compute the proportion of each allele. The principle is to assign a weight to each allele based on the read mappings (edges) selected by the ILP, and to normalize these weights to obtain proportions. First, we filter out all reads that map equally well (i.e. with the same score k) onto all selected alleles. Then every chosen allele gets and initial weight 0. Next, for every non-discarded read, say r_i, we consider all the alleles it maps onto with optimal score (say q_{ik} if $y_{ik} = 1$); assuming there are h such alleles, we increase the weight of each by $1/h$. We then normalize the weights of the alleles to define their respective proportions.

2.2 The Strain Diversity Problem

Once, for each sample, alleles present in the sample and their proportions have been identified, this information is passed to the second stage of the pipeline. Its goal is to compute strain types and proportions in all samples *jointly*, minimizing the number of novel strains required to explain the given allele distributions plus an error term measuring the total discrepancy between each given allele proportion and the proportions of strains having this allele. The rationale behind minimizing the number of new strains is driven by parsimony considerations; we would like to explain the data present in all samples using known strains as much as possible. The error terms allow some flexibility to modify the allele proportions by bounding each error to be $\leq \epsilon$ (in our analysis we set the bound to $\epsilon = 0.1$, or 10%).

The Strain Diversity Problem: problem definition and tractability. The Strain Diversity Problem (SDP) can be defined as follows. It takes as input four elements: (1) the set $G_{ij} = \{g_{ij1}, g_{ij2}, \ldots\}$ of all alleles selected for locus j in sample i (2) the set $P_{ij} = \{p_{ij1}, p_{ij2}, \ldots\}$ of proportions of these alleles, (3) a database Ω of known strain types, (4) an error bound $\epsilon \in [0, 1]$. From now, we assume that there are ℓ loci and m samples.

From this input, we generate, for each sample i the set of all possible strain types defined as the Cartesian product $G_{i1} \times G_{i2} \times \ldots \times G_{i\ell}$ which we denote by $V_i = \{V_{i1}, V_{i2}, \ldots, V_{iH_i}\}$ with $H_i = \prod_{j=1}^{\ell} |G_{ij}|$. We also denote by K the number of strain types that appear in at least one V_i and we define the set $\mathcal{S} = \{S_1, \ldots, S_K\}$ of all such strain types. We assign a weight w_j to each $\mathcal{S}_j \in \mathcal{S}$, where $w_j = N \cdot \min_{\{s \in \Omega\}} d(s, \mathcal{S}_j)$, where d is the edit distance metric and $N = \frac{1}{\max_j w_j}$, a normalizing term. These weights measure the distance to the closest known strain, so strains belonging to Ω are assigned a weight 0.

A solution to the SDP is fully described by assigning to every strain type V_{ih} from V_i a proportion π_{ih} for this strain type in sample i (where π_{ih} is 0 if the strain type is deemed to be absent from sample i). A strain type from $\mathcal{S} \setminus \Omega$ is said to be present in a solution if it is given a non-zero proportion in at least one sample; we denote by $\mathcal{S}_n$ the set of such novel strain types. The cost of a solution is then defined as

$$\sum_{\{h | \mathcal{S}_h \in \mathcal{S}_n\}} w_h + \sum_{i,j} e_{ij} \tag{1}$$

where the latter term of the cost represents the deviation from the input alleles proportions for sample i at locus j. This cost function penalizes the introduction

of highly different novel strains from known strains and the error introduced in the proportions of the selected alleles. The SDP aims at finding a solution of minimum cost, *i.e.* one that explains the provided allele distributions as much as possible with known strains and novel strains which are closer to known strains, and also sticks to the desired proportions as closely as possible. As expected, this problem is intractable; its decision version is proven to be NP-complete in Supplementary Material, by a reduction from the 3-partition problem.

Theorem 2 *The Strain Diversity Problem is NP-hard.*

An MILP for the Strain Diversity Problem. We now describe an MILP that solves the SDP. The decision variables of the MILP are the following:

- Binary variables a_k, $1 \leq k \leq K$, where $a_k = 1$ if strain type S_k is chosen to explain the observed allele distribution in at least one sample, and 0 otherwise.
- Proportion variables π_{ih} encoding the proportion of strain type V_{ih} in sample i; their values are constrained to be in $[0, 1]$.
- Variables $e_{ijk} \in [0, \epsilon]$ encoding the absolute error of the observed proportion p_{ijk} of allele g_{ijk} for locus j in sample i from the assigned proportions, in sample i, of the strain types containing this allele.

The objective function of the MILP is

$$\min \left(\sum_{\{k \mid S_k \notin \Omega\}} w_k a_k + \sum_{i,j,k} e_{ijk} \right) \tag{2}$$

Finally the constraints of the MILP are the following:

- For any allele $g_{ijk} \in G_{ij}$, the sum of the proportions of the strain types from V_i that contain this allele, denoted ν_{ijk}, belongs to $\left[p_{ijk} - \epsilon, p_{ijk} + \epsilon\right]$.
- For each sample i, the strain type proportions must form a distribution: $\sum_{h=1}^{H_i} \pi_{ih} = 1$.
- If the assigned proportion for some strain type $V_{ih} = S_k$ in a sample i is non-zero, then S_k must be chosen: $a_k \geq \pi_{ih}$.
- Conversely, if a strain is chosen, it must be assigned a non-zero proportion:

$$0 \leq a_k - \frac{1}{|\{\pi_{ih} \mid V_{ih} = S_k\}|} \cdot \sum_{\{(i,h) | V_{ih} = S_k\}} \pi_{ih} \leq 1 - \delta$$

  where δ is a tolerance chosen to match the smallest allowed proportion; we use $\delta = 0.001$. This constraint is needed because the binary decision variables for the usage of existing strains have coefficient 0 in the objective function, so setting these variables to 1 will not incur any cost in the objective function. If we do not impose such a constraint, we could end up with an incorrect solution where some existing strains have zero proportions, while the strain usage variables are set to 1,which would then need to be post-processed. Including the constraint eliminates the possibility of such a spurious solution.
- The absolute error between the input proportion and the assigned proportion for allele g_{ijk} for locus j in sample i: $e_{ijk} = |p_{ijk} - \nu_{ijk}|$. This is encoded by the following 2 constraints: $e_{ijk} \geq T_{ijk} - p_{ijk}$ and $e_{ijk} \geq p_{ijk} - T_{ijk}$ where

$T_{ijk} = \sum_{\{k \mid g_{ijk} \in V_{ik}\}} \pi_{ik}$. Note that since e_{ijk} is part of the objective function, it will be equal to the error in any optimal solution.

2.3 Implementation.

All scripts are written in Python 2.7. Both ILPs are formulated and solved using the Python API of IBM's CPLEX 12.6.3.0. For the ADP, each sample and each locus may require a different number of variables in the ILP. To evaluate the practical resources requirements of our ILP, we choose the sample SRR2034336, which has the largest number of reads among our samples. The average number of variables across each gene for this sample is 20,112, the maximum RAM usage is ∼1.5GB, and the time taken for all 8 genes is ∼ 33 minutes on a 4 CPUs Intel® Xeon® machine. The total time taken for each sample is presented in Supplementary Material. For the MILP solving the SDP on all 30 samples, there are a total of 21,885 variables, with 10,682 strain type variables, 10,795 proportion variables and 408 error variables. Due to the computational complexity of the MILP, we output a solution as long as the relative gap tolerance is within 10% and after a time limit of 24 hours. Our code is publicly available at `https://github.com/WGS-TB/MLST`.

2.4 Data simulation

Given the absence of benchmarks available for estimating diversity at the level of pecision considered in this work, we conducted several simulations. All reads are simulated as using ART [17], following the characteristics of the reads from the real data set described in Section 3.2.

ADP simulation. For each locus of the *Borrelia* MLST scheme, we drew a random number $k \in [2, 7]$, selected a random allele from the database and selected $k - 1$ other alleles, each at edit distance from the first chosen one at most d (a given parameter). Next, we randomly assigned proportions to each selected allele, which sums up to 1, then generate reads with coverage c. To align the simulated reads to the alleles of the database, we used Bowtie v0.12.7 [18]. We used parameters $c \in \{30, 100, 300\}$ and $d \in \{5, 10, 15, 20, 25\}$ and for each combination of parameters we ran 40 simulations. For this experiment, we compared our results with the results obtained with Kallisto [19] a recent method for isoform abundance estimation that has also been applied to metagenomics.

SDP simulation For this simulation we selected random strain types distributions and tested the ability of our SDP method to recover the true diversity given perfect alleles calls. We considered 5 different mechanisms to generate strain types distributions. EvoMod1: We select a random existing strain S, which is then mutated $m = 2$ times to obtain a new strain S', where each mutation results in an allele which has edit distance at most $d = 15$ from the original allele in S. The total number of strains simulated is 2 (1 existing and 1 novel). EvoMod2: We repeat EvoMod1 in parallel from two starting existing strains. The total number of strains simulated is 4 (2 existing and 2 novel). EvoMod2e/EvoMod2n: We apply EvoMod2 then remove a random existing/novel strain. EvoMod3: we apply EvoMod2, then apply a recombination event on two randomly chosen strains out of the 4 available strains. For all experiments, we assign random proportions to the chosen strains.

Full pipeline simulation. We generated strain type distributions as in the SDP simulations above, then generated reads as in the ADP simulations. The generated reads were then fed to the ADP solver, and the ADP results were provided as input to the SDP solver. We compared our pipeline with strainEST [13], a recent method to estimate the strain composition and abundance in metagenomics datasets. However, strainEST does not predict novel strain types. Hence, to complement EvoMod1, 2, 2e and 2n, we added an additional simulation where we randomly pick $k = \{1, 2\}$ existing strains and assign them random proportions.

Statistics. For each experiment, we recorded the following statistics: Precision, Recall and Total Variation Distance. Precision and recall are defined as usual, respectively $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$, where TP, FP, FN are the number of true positive calls, false positive calls, and false negative calls. The Total Variation Distance (TVD) [20, p. 50] is defined as $TVD = \frac{1}{2}\sum_{a \in S} |Pred(a) - True(a)|$, where $Pred$ and $True$ are the predicted distribution and the true distribution, respectively, and S is the set of all possible outcomes. TVD basically describes the average amount of distribution to "move" from $Pred$ to $True$ and conversely.

The statistics described above rely on a stringent measure of accuracy in calling alleles, strain types or proportions. For example, a novel strain type called which differs from the true simulated strain type by a single SNP would be considered as a False Positive. To account for this, we considered 3 additional statistics: Earth-Mover's distance (EMD), soft-precision and soft-recall. Soft precision and soft recall are similar to precision and recall, however, a strain is considered a TP if it differs from the true strain type by at most 5 SNPs. EMD [21] is similar in principle to TVD but is finer as it considers the edit distances between strains and is commonly used in genomics to evaluate haplotype reconstruction methods [22]. We provide a full definition in Supplementary Material.

3 Results

3.1 Simulated data

ADP simulation. Table 1 shows the performance of our method. Overall, our method obtained very high precision and recall statistics. Compared to Kallisto (results presented in Supplementary material) our method performs better in terms of precision and comparable in TVD, while Kallisto recall is better. Gene-by-gene boxplots for our method and Kallisto are available in Supplementary Material.

SDP and full pipeline simulation. The results are presented in Table 2. Given perfect input data, our SDP algorithm performed extremely well for each mechanism, maintaining a precision and recall of almost 75% with EvoMod3, that involves recombination. For the full pipeline simulation, our pipeline performs extremely well on ADP, which aligns well with the observations in ADP simulation. However, the full pipeline performance suffered in the SDP. Soft precision and recall are still high, but exact precision and recall are much lower. We can observe a dramatic impact on the SDP due to relatively small errors in the ADP (*i.e.* wrong allele identification or discrepancy in allele proportion estimation).

Comparison with strainEST. We compared our methods to strainEST in full pipeline simulation with 2 set of different experiments: (1) benchmark simulation where only existing strains are simulated (2) 4 different evolutionary mechanisms, where novel strains are involved. We refer the readers to the Supplementary Material for detailed results, but our method outperforms strainEST in all situations.

3.2 Application to real data

The sequencing data we analyzed are from 24 tick samples infected with *B. burgdorferi*, collected using the standard tick dragging method [23] in 2007 from 8 different sites in Vermont, New York, Massachusetts and Connecticut. For each tick sample, the *B. burgdorferi* genome was captured as described in [9]. The sequencing data is composed of 2×76bp paired-end reads and the number of read pairs ranges from $2.7 \cdot 10^4$ to $2.7 \cdot 10^6$ over all tick samples (coverages ranging from 5X to 500X).

Based on the output of the pipeline, 60 novel and 10 existing strains were inferred to be potential candidates to explain the strain diversity in this large sample of ticks. The total error component of the MILP for solving the SDP amounts to 1.258, or an average of 0.05 per sample. The total proportion of new strains is 14.67 in these 24 samples. On average, each new strain constitutes roughly 40% of the diversity of its sample. For each sample having novel strains, 76% of its genotype is composed of novel strains. Fig. 2 further illustrates the diversity, showing a wide range of strain composition in each of the 30 samples, with an average of 3 strains and a maximum of 9 strains infecting each sample, consistent with previous reports [5]. This suggest that the diversity of the *B. burgdorferi* strain types might be much larger than what was known so far. To further refine our analysis, Fig. 3 illustrates the distribution of strains types in the 30 tick samples and the respective contribution to the total diversity of each strain type. Although we observe that 2 of the 10 detected existing strains are present in more than one sample, only 5 out of the 60 novel strains appear in more than one sample.

It is striking to observe that most strain types appear in exactly one tick sample each. Also we can observe that for 11 samples, we do not detect any existing strains. This suggests that some of these strain types could have been improperly called, and that the correct call should have been another strain type, extremely close to this one in terms of sequence similarity; a reasonable cause for such errors could be a mistake while solving the ADP, in which case a wrongly called allele could be very similar to the correct allele. Due to possibility of wrong allele calls leading to introducing novel strains, we also computed a minimum spanning tree (MST) of the 70 strains found in these 24 samples, with edge connecting 2 strains having weight defined by the edit distance between the sequences of the alleles over the 8 genes of the MLST scheme. The MST figures are provided in Supplementary Material.We can observe clusters of predicted strains that are very close to each other, such as, for example, a cluster of 8 novel strains and 2 existing strains that are all within edit distance 5 of the nine other strains. This suggests, in line with the level of precision and recall we observe in our simulations, that some of these strains might result from a limited level of erroneous allele calls, away by a couple of SNPs from the correct call, that result in this apparent high diversity.

4 Discussion

We presented an optimization-based pipeline for estimating the within-host strain diversity of a pathogen from WGS data analyzed in the MLST framework. This is a specific instance of estimating the diversity of a bacterial pathogen from metagenomics data, focusing on within-host diversity and taking advantage of the availability of a large database of known MLST strain types.

Our approach is composed of two main steps, each of a different nature; the first step detects the alleles present in a sample from the sequence data, while the second step estimates the strain diversity based on the output of the first one. In both steps we follow a parsimonious approach that aims at explaining the input using as few alleles or novel strains as possible. The main contribution of our work is the formulation and the solution of the Strain Diversity Problem for a group of samples. The main challenge of this problem is the need to consider a potentially large set of samples at once. While this leads to a relatively complex MILP, with a large number of variables (whose number is guided by the number of potentially present novel strain types), we believe that the ability to consider a large set of samples at once is an important part of the model, for example for analyzing sequencing data from pathogen hosts originating from a single geographical area. Our work shows that this problem, despite its complexity, can actually be solved using reasonable computational resources and good accuracy.

Our experiments on real data suggest avenues for future research; in particular, the multiplicity of optimal solutions; this is obviously problematic as calling a wrong allele in a single sample during the first step might force the MILP computing the strain types to introduce a new strain type. We can observe in our results on real data groups of very closely related strain types, sometimes differing by a single SNP, which likely results from this issue. At the moment, our approach to this problem is to post-process the result of our pipeline to identify clusters of closely related strains, but other more principled approaches should be explored. Notwithstanding the aforementioned issues, our experiments suggest a strikingly high diversity in our dataset of 24 tick samples. This is not altogether surprising since the library of known strains might be limited, and within-host (or, more precisely, within-vector) evolution might result in the presence of a number of strains that only differ by a small number of SNPs in one or two loci of the MLST scheme.

Our work is, to our knowledge, the first comprehensive approach to the problem of reference-based detection of pathogen diversity in a collection of related samples that considers novel strain types. Our two-step pipeline, based on the principle of parsimony implemented through (mixed) integer linear programming, appears to perform extremely well on simulated data and produces reasonable results on a real dataset. We expect that both our approach and our publicly available pipeline will participate to the development of accurate and efficient tools for quantifying the within-host diversity of bacterial pathogens.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
LC designed the project, all authors designed the methods, GLG and EW implemented the methods and ran the experiments, all authors analyzed the results and wrote the paper.

Acknowledgements
The authors would like thank Maria Diuk-Wasser, Katharine Walter and Ben Adams for suggesting the problem as well as helpful discussions with regards to the data provenance and analysis. LC acknowledges support from NSERC, CIHR, Genome Canada and the Sloan Foundation. CC acknowledges the support of NSERC. GLG was partially funded by an NSERC CREATE scholarship. EW was partially funded by an SFU KEY fellowship.

Author details
[1]School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby (BC), Canada.
[2]Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby (BC), Canada.

References

1. Didelot, X., Walker, A.S., Peto, T.E., Crook, D.W., Wilson, D.J.: Within-host evolution of bacterial pathogens. Nat Rev Microbiol **14**(3), 150–162 (2016)
2. Cadena, A.M., Fortune, S.M., Flynn, J.L.: Heterogeneity in tuberculosis. Nat Rev Immunol **17**, 691 (2017). doi:10.1038/nri.2017.69
3. Tyler, A.D., Randell, E., Baikie, M., Antonation, K., Janella, D., Christianson, S., Tyrrell, G.J., Graham, M., Van Domselaar, G., Sharma, M.K.: Application of whole genome sequence analysis to the study of *Mycobacterium tuberculosis* in Nunavut, Canada. PLOS ONE **12**(10), 0185656 (2017). doi:10.1371/journal.pone.0185656
4. Alizon, S., de Roode, J.C., Michalakis, Y.: Multiple infections and the evolution of virulence. Ecol Lett **16**(4), 556–567 (2013). doi:10.1111/ele.12076
5. Strandh, M., Råberg, L.: Within-host competition between *Borrelia afzelii* ospc strains in wild hosts as revealed by massively parallel amplicon sequencing. Philos Trans R Soc Lond B Biol Sci **370**(1675) (2015). doi:10.1098/rstb.2014.0293
6. Brisson, D., Baxamusa, N., Schwartz, I., Wormser, G.P.: Biodiversity of *Borrelia burgdorferi* strains in tissues of Lyme disease patients. PLOS ONE **6**(8), 22926 (2011). doi:10.1371/journal.pone.0022926
7. Walter, K.S., Carpi, G., Evans, B.R., Caccone, A., Diuk-Wasser, M.A.: Vectors as epidemiological sentinels: Patterns of within-tick *Borrelia burgdorferi* diversity. PLOS Pathog **12**(7), 1005759 (2016). doi:10.1371/journal.ppat.1005759
8. Lynch, T., Petkau, A., Knox, N., Graham, M., Domselaar, G.V.: A primer on infectious disease bacterial genomics. Clin Microbiol Rev **29**(4), 881–913 (2016). doi:10.1128/cmr.00001-16
9. Carpi, G., Walter, K.S., Bent, S.J., Hoen, A.G., Diuk-Wasser, M., Caccone, A.: Whole genome capture of vector-borne pathogens from mixed DNA samples: a case study of *Borrelia burgdorferi*. BMC Genomics **16**(1) (2015). doi:10.1186/s12864-015-1634-x
10. Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G.: Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. PNAS **95**(6), 3140–3145 (1998)
11. Margos, G., Gatewood, A.G., Aanensen, D.M., Hanincova, K., Terekhova, D., Vollmer, S.A., Cornet, M., Piesman, J., Donaghy, M., Bormane, A., Hurn, M.A., Feil, E.J., Fish, D., Casjens, S., Wormser, G.P., Schwartz, I., Kurtenbach, K.: MLST of housekeeping genes captures geographic population structure and suggests a european origin of *Borrelia burgdorferi*. PNAS **105**(25), 8730–8735 (2008). doi:10.1073/pnas.0800323105
12. Quince, C., Delmont, T.O., Raguideau, S., Alneberg, J., Darling, A.E., Collins, G., Eren, A.M.: DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol **18**(1), 181 (2017). doi:10.1186/s13059-017-1309-9
13. Albanese, D., Donati, C.: Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat Commun **8**(1), 2260 (2017). doi:10.1038/s41467-017-02209-5
14. Chindelevitch, L., Colijn, C., Moodley, P., Wilson, D., Cohen, T., Else, e.: ClassTR: Classifying within-host heterogeneity based on tandem repeats with application to *Mycobacterium tuberculosis* infections. PLOS Comput Biol **12**(2), 1–16 (2016). doi:10.1371/journal.pcbi.1004475
15. Page, A.J., Alikhan, N.-F., Carleton, H.A., Seemann, T., Keane, J.A., Katz, L.S.: Comparison of Multi-Locus Sequence Typing software for Next Generation Sequencing data. Microb Genom **3**, 000124 (2017). doi:10.1099/mgen.0.000124
16. Boža, V., Brejová, B., Vinař, T.: GAML: genome assembly by maximum likelihood. Algorithms Mol Biol **10**(1), 18 (2015). doi:10.1186/s13015-015-0052-6
17. Huang, W., Li, L., Myers, J.R., Marth, G.T.: ART: a Next-Generation Sequencing read simulator. Bioinformatics **28**(4), 593–594 (2012). doi:10.1093/bioinformatics/btr708
18. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10**(3), 25 (2009). doi:10.1186/gb-2009-10-3-r25
19. Bray, N.L., Pimentel, H., Melsted, P., Pachter, L.: Near-optimal probabilistic RNA-seq quantification. Nat Biotech **34**(5), 525–527 (2016). doi:10.1038/nbt.3519
20. Levin, D.A., Peres, Y., Wilmer, E.L.: Markov chains and mixing times (2009)
21. Peleg, S., Werman, M., Rom, H.: A unified approach to the change of resolution: space and gray-level. IEEE Transactions on Pattern Analysis and Machine Intelligence **11**(7), 739–742 (1989). doi:10.1109/34.192468
22. Knyazev, S., Tsyvina, V., Melnyk, A., Artyomenko, A., Malygina, T., Porozov, Y.B., Campbell, E., Switzer, W.M., Skums, P., Zelikovsky, A.: CliqueSNV: Scalable reconstruction of intra-host viral populations from NGS reads. bioRxiv (2018). doi:10.1101/264242

23. Falco, R.C., Fish, D.: A comparison of methods for sampling the deer tick, *Ixodes dammini*, in a Lyme disease endemic area. Exp Appl Acarol **14**(2), 165–173 (1992). doi:10.1007/BF01219108

Figures

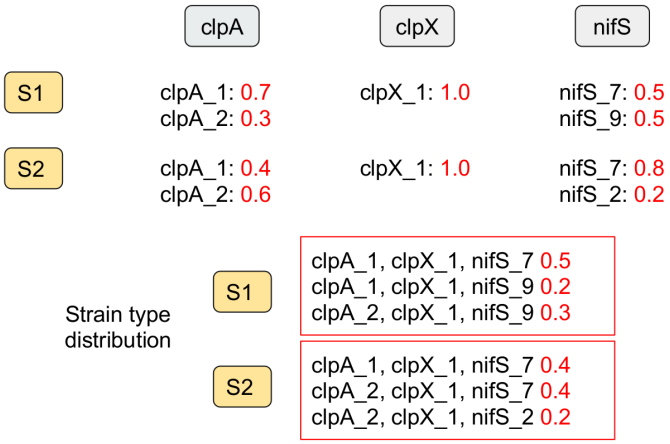


Figure 1 A dataset with two samples and an MLST scheme of three loci (genes clpA, clpX, nifS). The strain type distributions require 5 different strains as the strain (clpA_1,clpX_1, nifS_7) appears in both distributions.

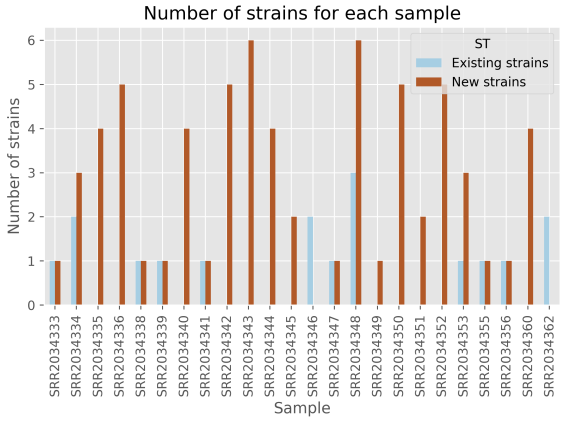


Figure 2 Distribution of the number of existing and novel strains per tick sample.

Tables

Prec.	clpA	clpX	nifS	pepX	pyrG	recG	rplB	uvrA
AVG	0.99	0.98	0.96	0.96	0.97	0.98	0.99	0.98
SD	0.009	0.012	0.024	0.016	0.024	0.013	0.007	0.011
Recall								
AVG	0.95	0.94	0.90	0.94	0.92	0.95	0.94	0.96
SD	0.022	0.027	0.05	0.034	0.032	0.028	0.043	0.026
TVD								
AVG	0.077	0.080	0.119	0.087	0.110	0.082	0.089	0.069
SD	0.015	0.010	0.039	0.024	0.019	0.028	0.030	0.020

Table 1 Average and standard deviation of precision, recall and TVD for each gene of the *Borellia* MLST scheme across all parameters combination.

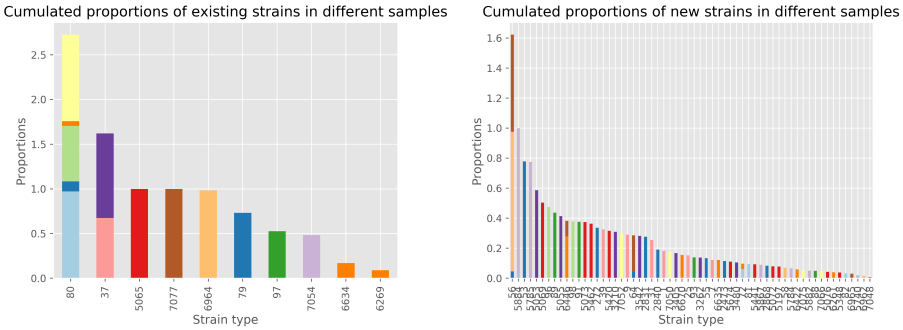


Figure 3 (Left) Cumulated proportion of the 10 existing strains in all 24 samples (In each bar, different colors represent different samples). (Right) Similar graph for the 60 novel strains.

	Soft-Precision	Soft-Recall	EMD	Precision	Recall	TVD
EM1	0.98 ± 0.11	0.96 ± 0.13	0.64 ± 1.7	0.85 ± 0.28	0.86 ± 0.23	0.15 ± 0.29
EM2	0.96 ± 0.12	0.98 ± 0.076	0.71 ± 1.18	0.81 ± 0.21	0.88 ± 0.14	0.17 ± 0.22
EM2e	0.98 ± 0.11	0.97 ± 0.1	0.34 ± 0.81	0.91 ± 0.20	0.92 ± 0.17	0.1 ± 0.23
EM2n	0.96 ± 0.13	0.95 ± 0.12	0.6 ± 1.35	0.86 ± 0.23	0.88 ± 0.16	0.14 ± 0.25
EM3	0.90 ± 0.17	0.88 ± 0.13	4.6 ± 7.58	0.76 ± 0.21	0.76 ± 0.17	0.22 ± 0.24

	ADP-Precision	ADP-Recall	ADP-TVD
EM1	0.96 ± 0.07	0.91 ± 0.09	0.07 ± 0.058
EM2	0.93 ± 0.07	0.91 ± 0.07	0.26 ± 0.16
EM2e	0.93 ± 0.08	0.91 ± 0.08	0.34 ± 0.25
EM2n	0.92 ± 0.09	0.9 ± 0.09	0.34 ± 0.25
EM3	0.94 ± 0.07	0.92 ± 0.08	0.29 ± 0.15

	Soft-Precision	Soft-Recall	EMD	Precision	Recall	TVD
EM1	0.96 ± 0.14	0.99 ± 0.079	4.1 ± 7.0	0.44 ± 0.34	0.58 ± 0.40	0.62 ± 0.37
EM2	0.79 ± 0.21	0.91 ± 0.16	68.8 ± 74.6	0.32 ± 0.19	0.44 ± 0.27	0.78 ± 0.2
EM2e	0.72 ± 0.24	0.88 ± 0.22	98.9 ± 89.4	0.36 ± 0.26	0.5 ± 0.30	0.72 ± 0.26
EM2n	0.76 ± 0.23	0.9 ± 0.19	98.6 ± 90	0.36 ± 0.25	0.52 ± 0.30	0.71 ± 0.24
EM3	0.68 ± 0.20	0.79 ± 0.2	83.7 ± 64	0.29 ± 0.2	0.35 ± 0.22	0.83 ± 0.16

Table 2 Average and standard deviation of different statistics for each evolutionary mechanisms. (Top) SDP simulation (Middle/Bottom) Full pipeline simulation: (Middle) ADP statistics, (Bottom) SDP statistics.

Deconvoluting the Diversity of Within-host Pathogen Strains in a Multi-Locus Sequence Typing Framework Supplementary Material

Guo Liang Gan [1], Elijah Willie [1], Cedric Chauve [2] and Leonid Chindelevitch [1]
1. School of Computing Science, Simon Fraser University, Burnaby, B.C. V5A 1S6, Canada.
2. Department of Mathematics, Simon Fraser University, Burnaby, B.C. V5A 1S6, Canada.

Complexity of the Allele Diversity Problem

Proof of NP-completeness of the Allele Diversity Problem

We transform the Allele Diversity Problem into a decision problem as follows.

Input:

- a set of reads $\mathcal{R} = \{r_1, r_2, \ldots\}$;
- a set of n alleles $\mathcal{A} = \{a_1, \ldots, a_n\}$ for the chosen locus;
- a target cost C.

Question: Is there a read cover whose cost is at most C?

It is clear that this problem is in NP, since given a particular subset $\mathcal{B}$ of alleles, it is easy to check, in polynomial time, that they form a read cover and that the cost of this read cover is at most C.

To show its NP-completeness we use a reduction from the 3-dimensional matching problem, formulated as follows:

Input:

- three disjoint sets, X, Y and Z, each of size n;
- a collection of $m \geq n$ triples $\{(x_i, y_i, z_i)\}_{1 \leq i \leq m}$, with $x_i \in X, y_i \in Y, z_i \in Z \; \forall \; i$.

Question: Is there a three-dimensional matching, i.e. a subset $M \subseteq \{1, 2, \ldots, m\}$ with $|M| = n$ such that $\cup_{i \in M} \{x_i, y_i, z_i\} = X \cup Y \cup Z$ (i.e. the corresponding triples contain each element exactly once)?

From a given input to the 3-dimensional matching problem we construct an instance of the allele diversity problem (with reads over a binary alphabet) as follows:

Select $3n$ binary strings, each of length $L-2$, such that the Hamming distance between any pair of them is at least 1 (for instance by using a Hamming code), pad them with a single 0 at the beginning and at the end to a total length L, and associate exactly one of them to each of the elements of $X \cup Y \cup Z$. For an element e denote by $s(e)$ the associated binary string.

Let the set of reads be $\mathcal{R} := \{s(e)\}_{e \in X \cup Y \cup Z}$. There are exactly $3n$ reads. We assume that the base quality scores at all the positions of all the reads equal the maximnum possible base quality score, so that after scaling, each mismatch contributes exactly 1 to the objective function.

For each triplet (x_i, y_i, z_i) we construct an allele of length $5L$, $s(x_i)1^L s(y_i)1^L s(z_i)$, where 1^L is the string of L 1's and the product is via concatenation. In other words, $\mathcal{A} := \{s(x_i)1^L s(y_i)1^L s(z_i)\}_{1 \leq i \leq m}$. Lastly, we set $C := n$. Since L can be chosen to be linear in n, and the codewords can be constructed in linear time with respect to n as well, this reduction can be performed in polynomial time.

We now prove the following

Claim: There is a read cover with cost at most n if and only if there is a three-dimensional matching.

Proof: By construction, $s(e)$ maps to any allele based on a triplet containing e with no errors, and to any allele based on a triplet not containing e with at least 1 error (since the encodings of the elements differ in at least one position, and a mapping overlapping with the "sentinel" string 1^L also incurs at least one mismatch due to the presence of a 0 at either end). It follows that we can get a score of at most n if we pick the n alleles corresponding to a three-dimensional matching, since there are no errors on any of the reads and n alleles are used in this case.

On the other hand, suppose that we do get a score of at most n. Since each allele can explain at most 3 reads with no errors, the case of no errors requires at least n alleles for the $3n$ reads, and the only way that this can be achieved with exactly n alleles is if each allele explains three different reads with no errors, meaning that the triples corresponding to the n chosen alleles form a three-dimensional matching. If there are $k > 0$ errors, then at most $n-k$ alleles can be used. It follows that at most $3(n-k)$ of the $3n$ reads incur no errors, and the remaining $3k$ reads must each incur at least one error, so the number of errors incurred must be at least $3k$, contradicting the hypothesis that there were only k errors. This proves the claim.

The ADP and the Uncapacitated Facility Location Problem

Our formulation of the ADP is very similar to that of the Uncapacitated Facility Location Problem (UFLP), in which a set of customers needs to be served by facilities placed in a subset of a finite set of possible locations, such that the total cost of opening the facilities plus the distances between each customer and the closest facility is minimized. Indeed, if we consider reads to be customers, alleles to be facilities, and distances to be the sum of normalized base quality scores of the corresponding mappings, the ADP becomes a special case of the UFLP. However, our problem is somewhat simpler than the general UFLP because the set of possible distances is limited. We exploit this fact in our ILP formulation which, unlike the formulation for the UFLP, has fewer variables and constraints.

Related to the Uncapacitated Facility Location Problem, already known to be NP-hard, our NP-hardness proof is of independent interest because it actually produces a set of reads and a set of alleles that have the desired distances, whereas not every instance of the UFLP may be obtainable from actual reads and alleles.

Alternative Formulation of the Allele Diversity Problem

The seemingly intuitive way to formulate the ADP is to introduce a decision variable b_{ij} representing the assignment of read r_i to allele a_j. Here we present the alternative formulation of ADP based on this idea as an integer linear program. Given there are n alleles and m reads, we denote the set of reads $R = \{r_1, ..., r_m\}$ and the set of alleles $A = \{a_1, ..., a_n\}$. Also, m_{ij} is the score for mapping read r_i to allele a_j, the entries of the matrix input for ADP as described previously. Next, the decision variables and objective function are as follows:

- $z_j = 1$ if allele a_j is chosen, and 0 otherwise.
- $b_{ij} = 1$ if the mapping of read r_i to allele a_j is chosen and 0 otherwise.
- $\min \left(\sum_{i=1}^{m} \sum_{j=1}^{n} m_{ij} \cdot b_{ij} + \sum_{k=1}^{n} z_k \right)$

The constraints are:

- All reads are covered: $\sum_{j=1}^{n} b_{ij} = 1 \quad \forall i$
- If read r_i is assigned to allele a_j, then allele a_j must be chosen: $z_j \geq b_{ij} \quad \forall i, j$

Proof of the Equivalence of Both Formulations

Denote our formulation as (1) and the alternative formulation as (2). We prove that both formulations are equivalent by showing that a feasible solution of (1) can be translated to a feasible solution of (2), with both

having the same objective value and vice versa.

("⇒") Prove that a feasible solution of (1) can be translated into a feasible solution of (2). Given a feasible solution $s_1 = (\mathbf{x}; \mathbf{Y})$ where $\mathbf{x} = (x_j)$ and $\mathbf{Y} = (y_{ik})$. We denote a solution for (2), $s_2 = (\mathbf{z}; \mathbf{B})$ where $\mathbf{z} = (z_j)$ and $\mathbf{B} = (b_{ij})$. Also, recall that the matrix input for the ADP is $M = (m_{ij})_{m \times n}$.

- For all j, assuming the indices for both refer to the same allele, let $z_j = x_j$.
- For all i, due to the 2nd constraint in (1), $\exists!\ k^*$ such that $y_{ik^*} = 1$ and $y_{ik} = 0\ \forall k \neq k^*$. Denote $H_{ik^*} = \{j \mid m_{ij} = q_{ik^*} \wedge x_j = 1, j \in \{1, ..., n\}\}$. Choose an index $j^* \in H_{ik^*}$, let $b_{ij^*} = 1$ and $b_{ij} = 0$ for all $j \in H_{ik^*} - \{j^*\}$. For $j \notin H_{ik^*}$, let b_{ij}=0.

Claim: s_2 is a feasible solution for (2) with same objective value as (1).

Proof:

- Constraint 1: $\sum_{j=1}^{n} b_{ij} = 1\ \forall i$. This is true by construction of b_{ij}.
- Constraint 2: $z_j \geq b_{ij}\ \forall i, j$. If $b_{ij} = 1$, then $j = j^* \Rightarrow j \in H_{ik^*} \Rightarrow x_j = 1 = z_j \Rightarrow z_j \geq b_{ij}$. The case for $b_{ij} = 0$ is trivial as z_j is always ≥ 0.
- Objective value: The summation of the allele variables is the same in both formulations. For the other summation term:
$$\sum_{i=1}^{|\mathcal{R}|} \sum_{k=1}^{|M_i|} q_{ik} y_{ik} = \sum_{i=1}^{|\mathcal{R}|} \left(q_{ik^*} y_{ik^*} + \sum_{k \neq k^*} q_{ik} y_{ik} \right)$$
$$= \sum_{i=1}^{|\mathcal{R}|} \left(m_{ij^*} b_{ij^*} + \sum_{j \in \{1,..,n\} - \{j^*\}} m_{ij} b_{ij} \right) \quad (b_{ij} = 0\ , j \neq j^*)$$
$$= \sum_{i=1}^{m} \sum_{j=1}^{n} m_{ij} b_{ij}$$

("⇐") Prove that a feasible solution of (2) can be translated to a feasible solution of (1). Given a feasible solution $s_2 = (\mathbf{z}; \mathbf{B})$.

- Let $x_j = z_j$
- For all i, due to first constraint of (2), $\exists! j \in \{1, ..., n\}$, say j^*, where $b_{ij^*} = 1$. Also, say the corresponding score $q_{ik^*} = m_{ij^*}$. Let $y_{ik^*} = 1$, $y_{ik} = 0\ \forall q_{ik} \in M_i - \{q_{ik^*}\}$.

Claim: s_2 is a feasible solution for (2) with same objective value as (1).

Proof:

- Constraint 1: If $y_{ik} = 1$, $k = k^*$. Due to how we construct the solution, $\exists j^*$ such that $b_{ij^*} = 1 \wedge m_{ij^*} = q_{ik^*}$. Due to the 2nd constraint of (2), $x_j^* = z_j^* \geq b_{ij^*} = 1 = y_{ik^*}$. The case of $y_{ik} = 0$ is trivial.
- Constraint 2: $\sum_{k=1}^{|M_i|} y_{ik} = \left(y_{ik^*} + \sum_{k \neq k^*} y_{ik} \right) = (1 + 0) = 1$
- Objective value: The summation of the allele variables is the same. For the other summation term:
$$\sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} b_{ij} = \sum_{i=1}^{m} \left(b_{ij^*} w_{ij^*} + \sum_{j \neq j^*} w_{ij} b_{ij} \right)$$
$$= \sum_{i=1}^{m} \left(q_{ik^*} y_{ik^*} + \sum_{k \neq k^*} q_{ik} y_{ik} \right) \quad (\text{as } y_{ik} = 0 \text{ for } k \neq k^*)$$
$$= \sum_{i=1}^{|\mathcal{R}|} \sum_{k=1}^{|M_i|} q_{ik} y_{ik}$$

Proof of NP-completeness of the Strain Diversity Problem

We turn the Strain Diversity Problem into a decision problem by putting the objective function value into the input. This decision problem version is then formulated as follows:

Input:

- The set $G_{ij} = \{g_{ij1}, g_{ij2}, \ldots\}$ of all alleles selected for locus j in sample i, together with the set $P_{ij} = \{p_{ij1}, p_{ij2}, \ldots\}$ of proportions of these alleles;
- A database Ω of known strain types;
- An error bound $\epsilon \in [0, 1]$;
- A target cost C.

Let the set of all possible strain types for sample i, the Cartesian product $G_{i1} \times G_{i2} \times \ldots \times G_{i\ell}$, be denoted by $V_i = \{V_{i1}, V_{i2}, \ldots, V_{iH_i}\}$ with $H_i = \prod_{j=1}^{\ell} |G_{ij}|$. For simplicity, assuming the weights for each novel strain are set to 1.

Question: Is there a set of proportions π_{ih} for each strain type $V_{ih} \in V_i$ in each sample i, satisfying the validity constraint

$$\nu_{ijk} := \sum_{(i,k)|g_{ijk} \in V_{ik}} \pi_{ik} \in \left[p_{ijk} - \epsilon, p_{ijk} + \epsilon\right] \; \forall \; g_{ijk} \in G_{ij},$$

such that the total cost is at most equal to the target C, i.e.

$$|\{V_{ih} | \pi_{ih} > 0\} - \Omega| + \sum_{i,j,k} |p_{ijk} - \nu_{ijk})| \leq C?$$

This problem is clearly in NP because given a set of proportions π_{ih}, it is easy to check, in linear time in the size of the input, that the validity constraints are satisfied, and that the cost is at most C.

To show its NP-completeness we use a reduction from the 3-partition problem, formulated as follows:

Input: A multiset of $n = 3m$ positive integers $x_1, x_2, \ldots, x_n$ with sum $S = mB$ for some integer B, and satisfying $\frac{B}{4} < x_i < \frac{B}{2}$ for each $1 \leq i \leq n$.

Question: Does there exist a partition of this multiset into m disjoint triples $T_1, T_2, \ldots, T_m$, such that the sum of each triple is exactly B?

From a given input to the 3-partition problem we construct the following instance of SDP. Let us choose $\epsilon = \frac{1}{4m}$. Let us look at a single sample with two loci, 1 and 2, where the first locus has the $n = 3m$ possible alleles $g_1, g_2, \ldots, g_n$ and the second one has the m possible alleles $h_1, h_2, \ldots, h_m$.

Let the target allele distribution for locus 1 be g_1 with proportion $\frac{x_1}{S}$, g_2 with proportion $\frac{x_2}{S}, \ldots, g_n$ with proportion $\frac{x_n}{S}$, and let the target allele distribution for locus 2 be h_i with proportion $\frac{1}{m}$ for each $1 \leq i \leq m$.

Finally, let the library strains be $\Omega := \emptyset$, and let the target cost be $C := n$.

It is clear that the transformation is polynomial in the input size. We now prove the following

Claim: The original problem has a "yes" answer if and only if the transformed version has cost at most C.

Proof: Suppose that there is a partition of the multiset into disjoint triples $T_1, T_2, \ldots, T_m$. Then we can take the 3 integers, say $x_{j_1}, x_{j_2}, x_{j_3}$, in a triple T_j, and use the 3 new strains $g_{j_1} - h_j$ with proportion $\frac{x_{j_1}}{S}$, $g_{j_2} - h_j$ with proportion $\frac{x_{j_2}}{S}$, and $g_{j_3} - h_j$ with proportion $\frac{x_{j_3}}{S}$, to cover the allele h_j completely, with no error (since $\frac{x_{j_1}}{S} + \frac{x_{j_2}}{S} + \frac{x_{j_3}}{S} = \frac{B}{S} = \frac{1}{m}$ by construction). Thus, we need a total of $n = 3m$ strains to cover all the alleles in both loci with no error, so the cost of our solution is indeed n.

Conversely, suppose that there is a solution with cost at most n. Note first that because $\epsilon = \frac{1}{4m} = \frac{B}{4Bm} = \frac{B}{4S} < \frac{x_i}{S}$ for any $1 \leq i \leq n$, it follows that no allele in locus 1 can be eliminated from the solution while satisfying the validity constraint. It follows that at least $n = 3m$ strains are needed in any valid solution, so the cost cannot be smaller than n.

Furthermore, in order to get a cost of exactly n, there needs to be no error on the proportions and exactly $n = 3m$ strains, so each allele in locus 1 must be in exactly one strain. Thus, the allele h_j for locus 2 must be associated with a subset T_j of alleles of locus 1, and the T_j must be disjoint and therefore form a partition of the alleles of locus 1. Since $\frac{B}{4} < x_i < \frac{B}{2}$ for each i, we must have exactly 3 alleles in each subset T_j, since

adding up 2 of the $\frac{x_i}{S}$ results in a sum below the target proportion $\frac{B}{S} = \frac{1}{m}$ for h_j, and adding up 4 of them results in a sum above it, thus incurring a non-zero error on the proportions.

By considering the sum of the proportions in each triple T_j containing three alleles, say $g_{j_1}, g_{j_2}, g_{j_3}$, we see that we must have

$$\frac{x_{j_1}}{S} + \frac{x_{j_2}}{S} + \frac{x_{j_3}}{S} = \frac{1}{m} = \frac{B}{S} \Rightarrow x_{j_1} + x_{j_2} + x_{j_3} = B.$$

As the T_j are disjoint, this proves that the answer to the original problem is "yes", completing the proof.

WGS data preprocessing.

Before running the pipeline, we map the reads of each sample to a database of alleles for each gene. We use the memory efficient short read mapper Bowtie (v0.12.7) for read mapping. Next, we use *SAMTOOLS* (v1.3.1) to process the *SAM* file and extract information such as reads that are mapped as a pair, the alleles that are involved in the mappings, the number of mismatches, the base qualities, and the position of the mismatches.

Definition of Earth-Mover's Distance

Let $\mathcal{T} = \{T_i, t_i\}_{i=1}^{|\mathcal{T}|}$ be the set of true strains T_i with its respective proportion t_i and $\mathcal{P} = \{P_j, p_j\}_{i=1}^{|\mathcal{P}|}$ be the set of predicted strains P_j with predicted proportion p_j. Also, denote d_{ij} as the edit distance between T_i and P_j. EMD measures the minimum amount of work to transform the predicted distribution defined by $\mathcal{P}$ into true distribution defined by $\mathcal{T}$, where the amount of work is also weighted by distances d_{ij}. It involves solving a transportation problem which can be modeled as a network flow problem, where the problem is to obtain a flow $F = \{f_{ij}\}$ which minimizes $\sum_{ij} f_{ij} d_{ij}$ and the $EMD = \frac{\sum_{ij} f_{ij} d_{ij}}{\sum_{ij} f_{ij}}$. It can be thought as a network flow problem in the following sense: we assign fractions of each P_i to any T_j (flow f_{ij}), constrained by the maximum amount of fractions that P_i and T_j can give and take, which is p_i and t_j (capacities), where the assignments f_{ij} are weighted by d_{ij}.

Results of Kallisto on ADP simulation

As Kallisto introduces large number of alleles with extremely small abundances, we only retained alleles with larger than 1% proportion.

Prec.	clpA	clpX	nifS	pepX	pyrG	recG	rplB	uvrA
AVG	0.97	0.94	0.89	0.93	0.93	0.89	0.95	0.93
SD	0.014	0.014	0.027	0.03	0.02	0.021	0.012	0.023
Recall								
AVG	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99
SD	0.004	0.005	0.003	0.006	0.006	0.011	0.006	0.005
TVD								
AVG	0.029	0.041	0.085	0.046	0.0047	0.068	0.032	0.05
SD	0.011	0.015	0.028	0.022	0.018	0.018	0.011	0.022

Table 1: Kallisto: Average and standard deviation of precision, recall and TVD for each gene of the *Borellia* MLST scheme across all parameters combination.

Results of strainEST on full pipeline simulation

Recall that full pipeline simulation involves generating reads after strains are chosen, and our pipeline first solves the ADP problem then the SDP problem, by taking results of ADP as an input. StrainEST infers strains and abundances, and we compute the statistics for ADP based on the strains and abundances inferred. In this section, we present the results for our algorithm and strainEST on 2 different set of experiments: (1) benchmark simulation: Only existing strains are chosen for simulation, where we consider $k = \{1, 2\}$ number of strains (2) EvoMods simulation: 4 evolutionary mechanisms. **Remark**: As strainEST can only take a maximum of 100 genomes, we randomly choose 100 existing strains as the database for simulation.

Results on benchmark simulation

	Our Algorithm			StrainEST		
	Precision	Recall	TVD	Precision	Recall	TVD
k=1	0.96 ± 0.057	0.88 ± 0.08	0.052 ± 0.052	0.67 ± 0.29	0.5 ± 0.47	0.4 ± 0.33
k=2	0.94 ± 0.061	0.9 ± 0.06	0.33 ± 0.23	0.62 ± 0.24	0.5 ± 0.23	0.4 ± 0.274

Table 2: Comparison between our algorithm and strainEST on the SDP problem

	SoftPrecision	SoftRecall	EMD	Precision	Recall	TVD
k=1, our method	0.99 ± 0.08	1 ± 0	0.623 ± 0.625	0.48 ± 0.39	0.68 ± 0.47	0.406 ± 0.42
k=1, strainEST	0.35 ± 0.43	0.48 ± 0.5	8.07 ± 7.39	0.24 ± 0.4	0.3 ± 0.46	0.4 ± 0.33
k=2, our method	0.90 ± 0.18	0.99 ± 0.08	79.2 ± 90.3	0.4 ± 0.25	0.6 ± 0.31	0.67 ± 0.24
k=2, strainEST	0.37 ± 0.32	0.49 ± 0.34	12.5 ± 8.7	0.2 ± 0.24	0.28 ± 0.32	0.714 ± 0.352

Table 3: Comparison between our algorithm and strainEST on the ADP problem

Results on 4 EvoMods

The results for different evolutionary mechanisms presented for our algorithm are different from those presented in the Results section, due to the database size constraint by strainEST.

	StrainEST			Our Algorithm		
	Precision	Recall	TVD	Precision	Recall	TVD
EM1	0.55 ± 0.26	0.51 ± 0.34	0.5 ± 0.29	0.97 ± 0.06	0.92 ± 0.08	0.066 ± 0.059
EM2	0.64 ± 0.16	0.57 ± 0.24	0.42 ± 0.23	0.93 ± 0.07	0.91 ± 0.07	0.252 ± 0.16
EM2e	0.52 ± 0.16	0.46 ± 0.24	0.51 ± 0.22	0.94 ± 0.06	0.92 ± 0.06	0.277 ± 0.162
EM2n	0.61 ± 0.18	0.56 ± 0.23	0.41 ± 0.22	0.93 ± 0.07	0.92 ± 0.07	0.275 ± 0.168

	Soft-Precision	Soft-Recall	EMD	Precision	Recall	TVD
EM1	0.42 ± 0.46	0.28 ± 0.27	13.6 ± 10.9	0.27 ± 0.41	0.18 ± 0.24	0.83 ± 0.26
EM2	0.45 ± 0.31	0.33 ± 0.18	13.6 ± 7.84	0.3 ± 0.31	0.21 ± 0.18	0.79 ± 0.26
EM2e	0.22 ± 0.26	0.18 ± 0.17	17.86 ± 9.04	0.13 ± 0.23	0.1 ± 0.15	0.912 ± 0.178
EM2n	0.43 ± 0.32	0.38 ± 0.23	12.7 ± 7.68	0.26 ± 0.3	0.22 ± 0.22	0.8 ± 0.251

	Soft-Precision	Soft-Recall	EMD	Precision	Recall	TVD
EM1	0.87 ± 0.29	0.91 ± 0.27	3.5 ± 3.28	0.41 ± 0.34	0.54 ± 0.39	0.643 ± 0.37
EM2	0.74 ± 0.23	0.88 ± 0.21	55.58 ± 62	0.28 ± 0.23	0.39 ± 0.3	0.813 ± 0.195
EM2e	0.69 ± 0.27	0.83 ± 0.28	61.23 ± 65.32	0.28 ± 0.21	0.38 ± 0.28	0.766 ± 0.218
EM2n	0.79 ± 0.21	0.94 ± 0.15	60.6 ± 67	0.33 ± 0.24	0.47 ± 0.31	0.742 ± 0.243

Table 4: Average and standard deviation of different statistics for full pipeline simulation of 4 evolutionary mechanisms. (Top) ADP comparison (Middle) strainEST on SDP (Bottom) Our method on SDP

Accession numbers for the real data samples

The following table lists the SRA accession numbers of all the samples used in this study.

SRR2034333	SRR2034334	SRR2034335	SRR2034336	SRR2034338
SRR2034339	SRR2034340	SRR2034341	SRR2034342	SRR2034343
SRR2034344	SRR2034345	SRR2034346	SRR2034347	SRR2034348
SRR2034349	SRR2034350	SRR2034351	SRR2034352	SRR2034353
SRR2034355	SRR2034356	SRR2034360	SRR2034362	

Table 5: SRA accession numbers for all samples used in this study

Minimum Spanning Tree of Strains Found in 24 Samples

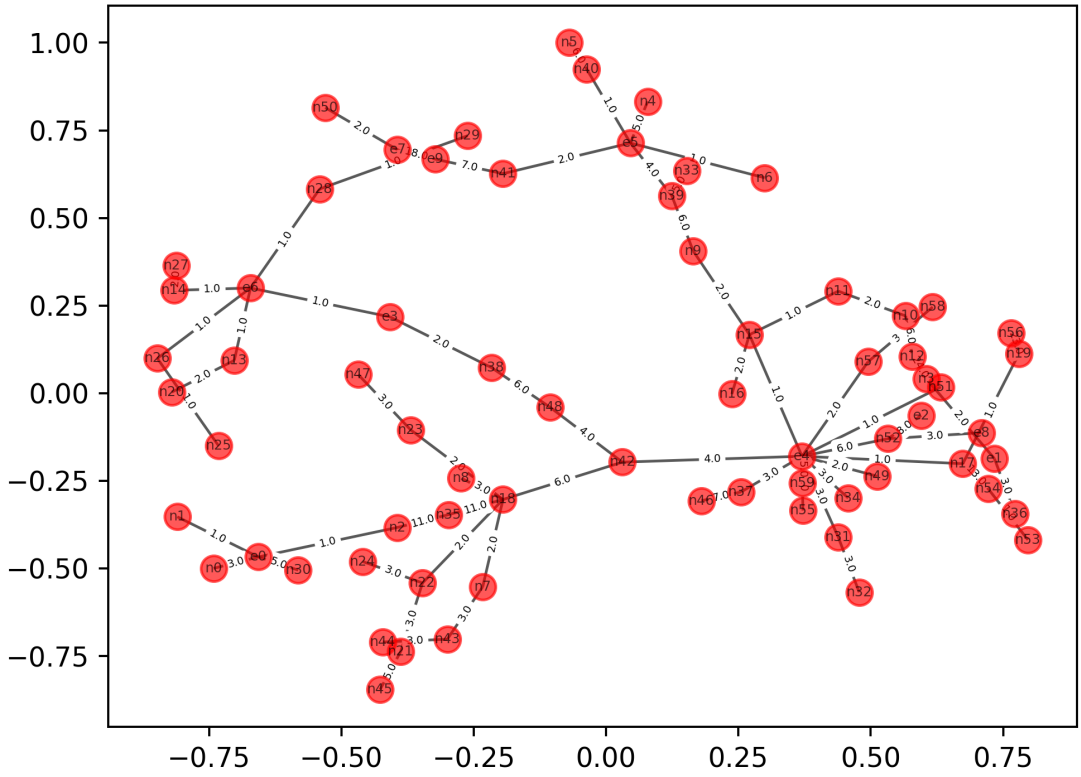


Figure 1: Nodes with prefix 'n' represents novel strains, 'e' represents existing.

We can observe clusters of strains which are very close to each other, for example, the circled tree branch in the upper left, consists nodes n25, n26, e6, e3, n13, n20, n14, n27, n28, n29. These nodes can be grouped as a cluster with strains of edit distance at most 5 bases. It is worth taking note that this distance is taken across 8 genes where each is approximately 580 bases long. Furthermore, mapping back to samples which contain these strains, n13 and n14 are in sample SRR2034335; n25, n26, n27, n28, n29 are in sample SRR2034342. Similar observations were seen on nodes which are in the star structure circled in Figure2 where any 2 nodes in the structure differ by at most 11 bases. These observation indicates that either strains in these samples mutate into extremely similar strains, or it indicates that wrong allele call leads to relatively higher number of novel strains inferred by our method.

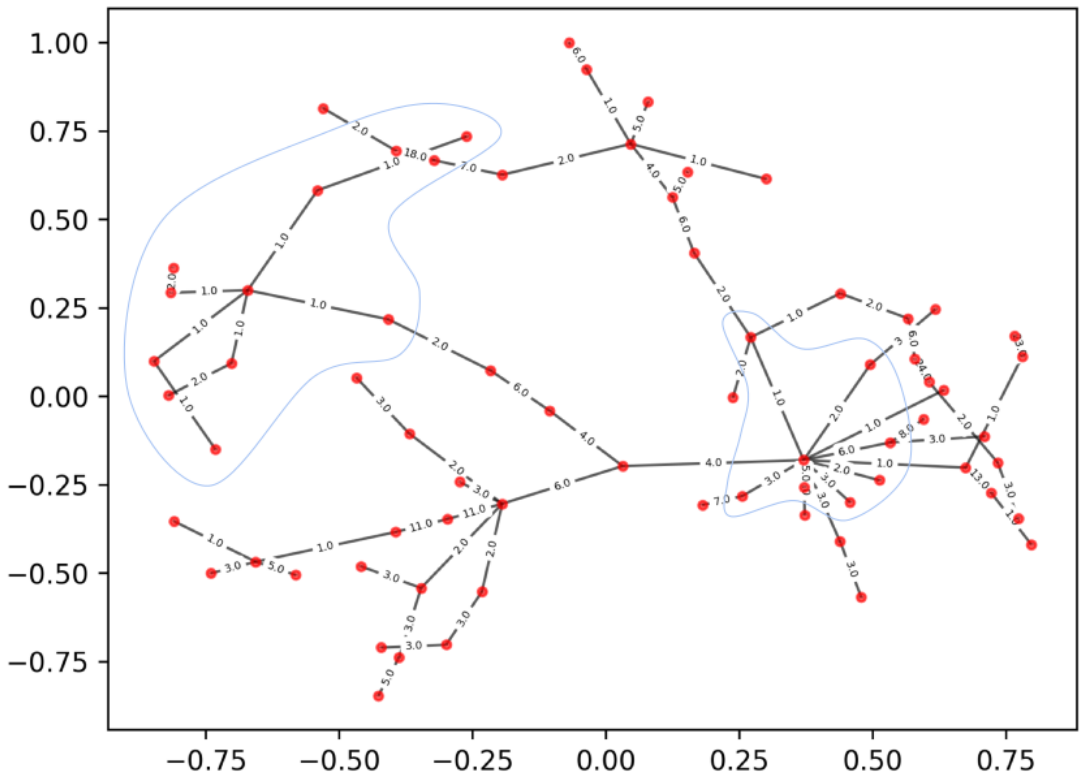
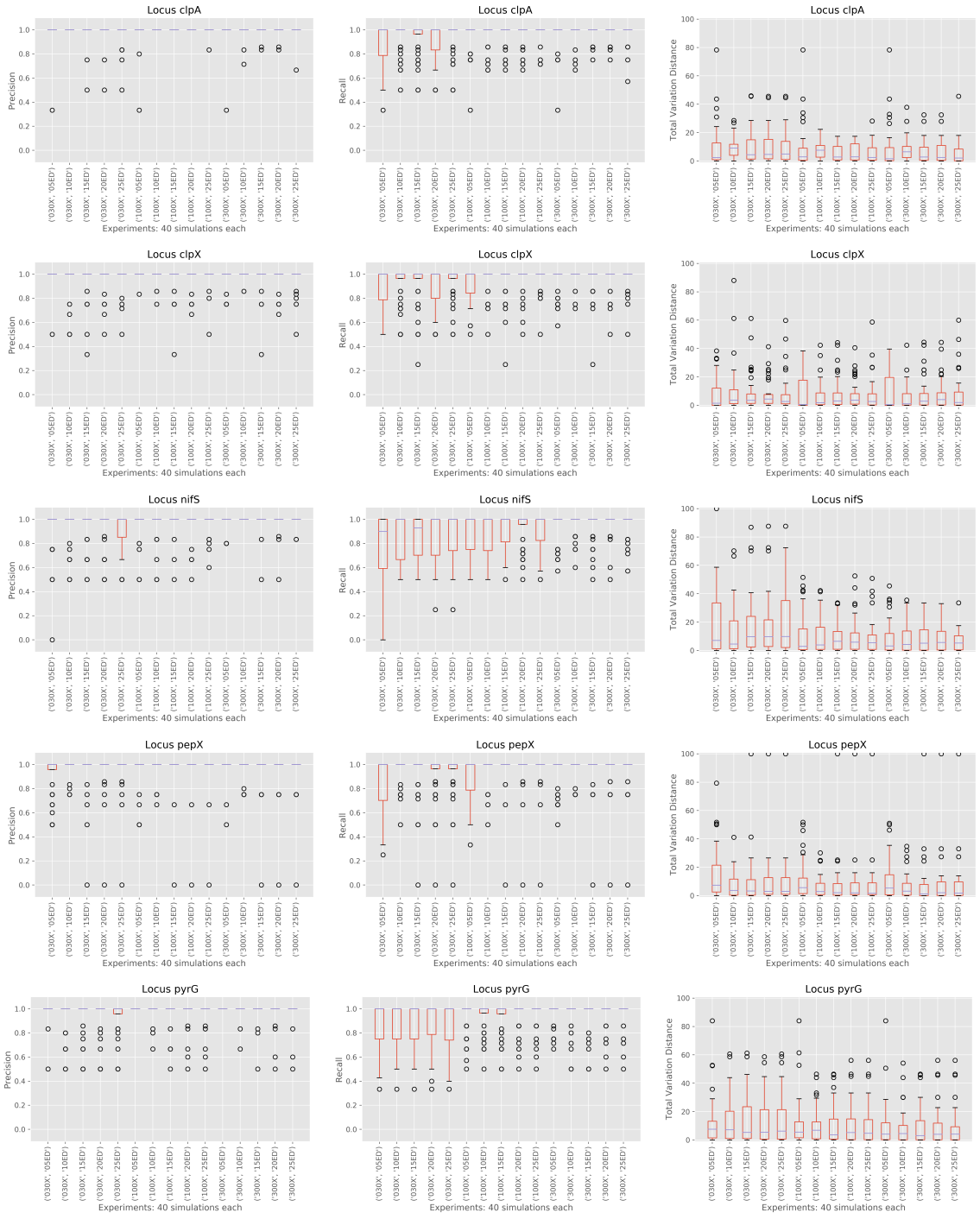


Figure 2: Minimum spanning tree of strains found in 24 samples.

Box Plots of Our Method on ADP simulation



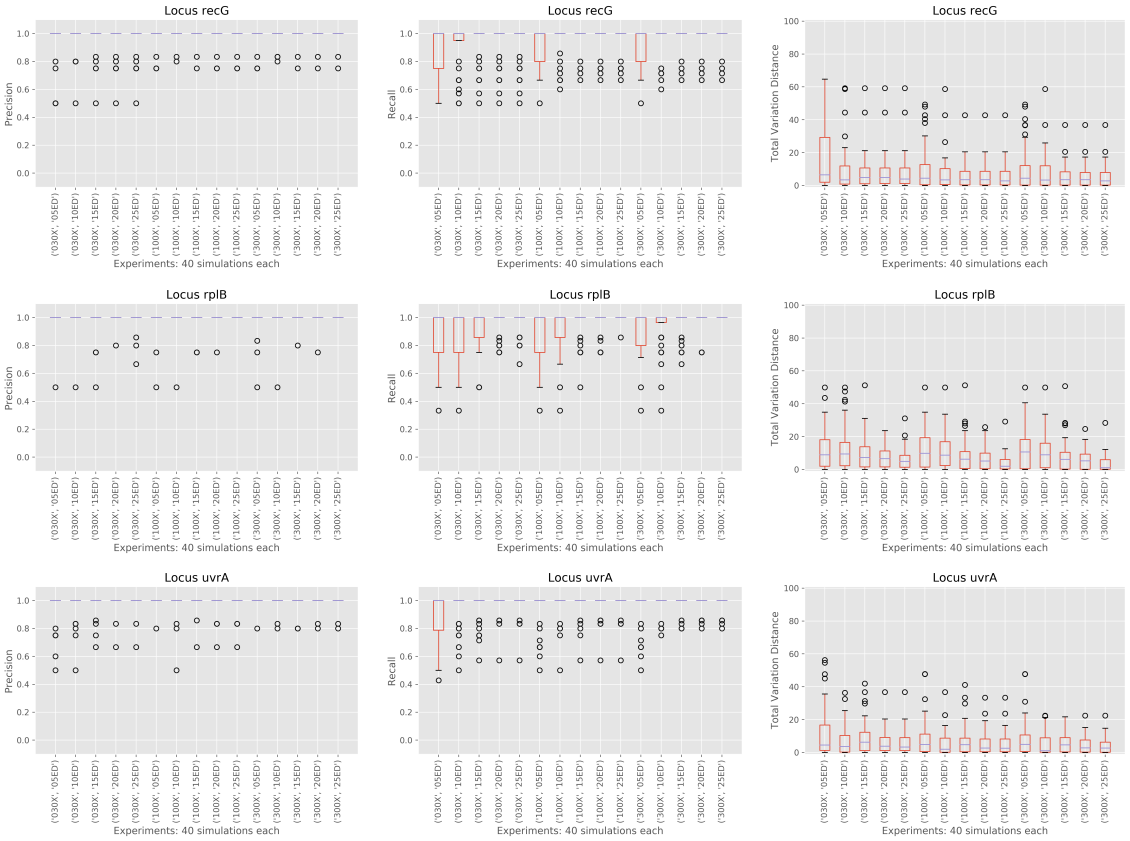
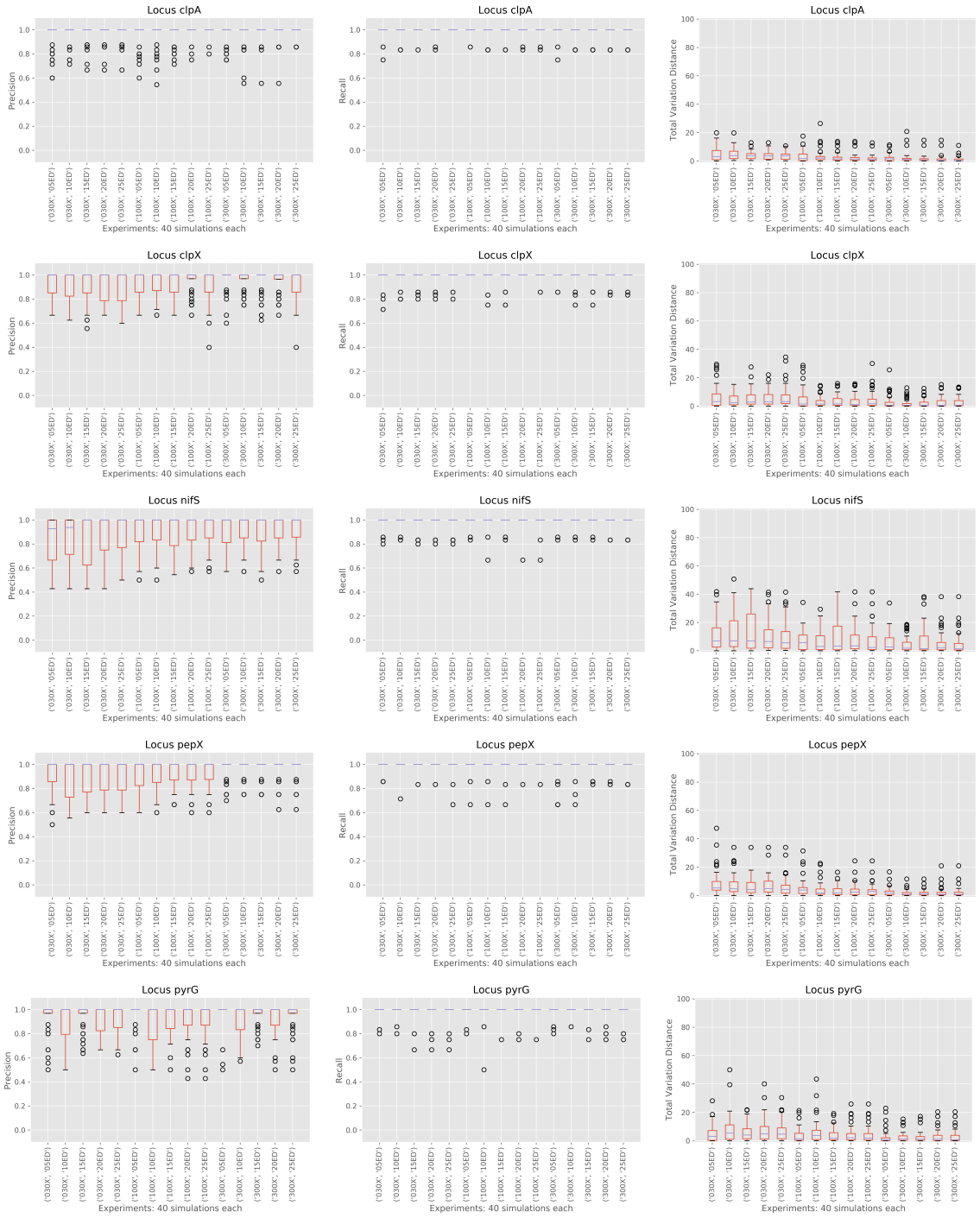


Figure 3: Box plots of Precision, Recall, and TVD of our method for different combinations of coverages and edit distance parameters over 40 simulations for 8 genes of the Borrelia MLST scheme

Box plots of Kallisto on ADP simulation



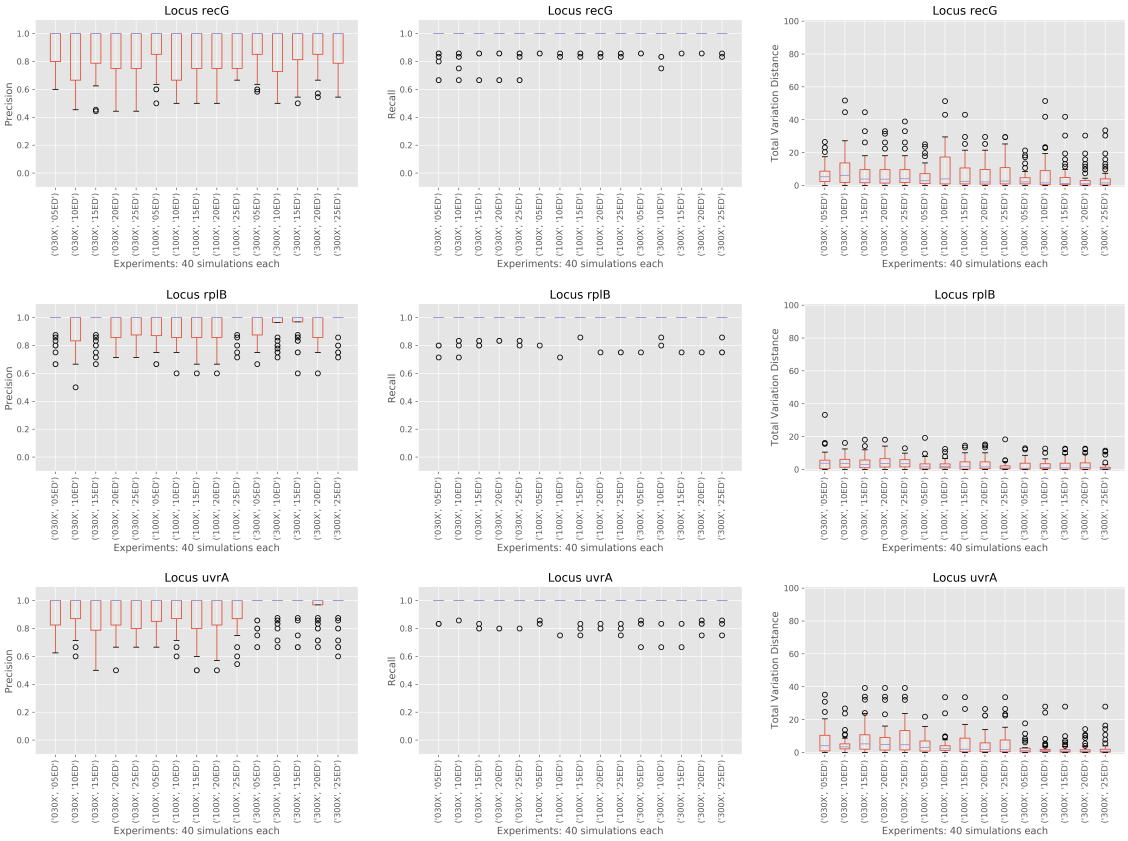


Figure 4: Box plots of Precision, Recall, and TVD of Kallisto for different combinations of coverages and edit distance parameters over 40 simulations for 8 genes of the Borrelia MLST scheme

Box Plots of Our Method and strainEST on Benchmark Simulation

Recall that the benchmark simulation is one of the simulations used to compare to strainEST.

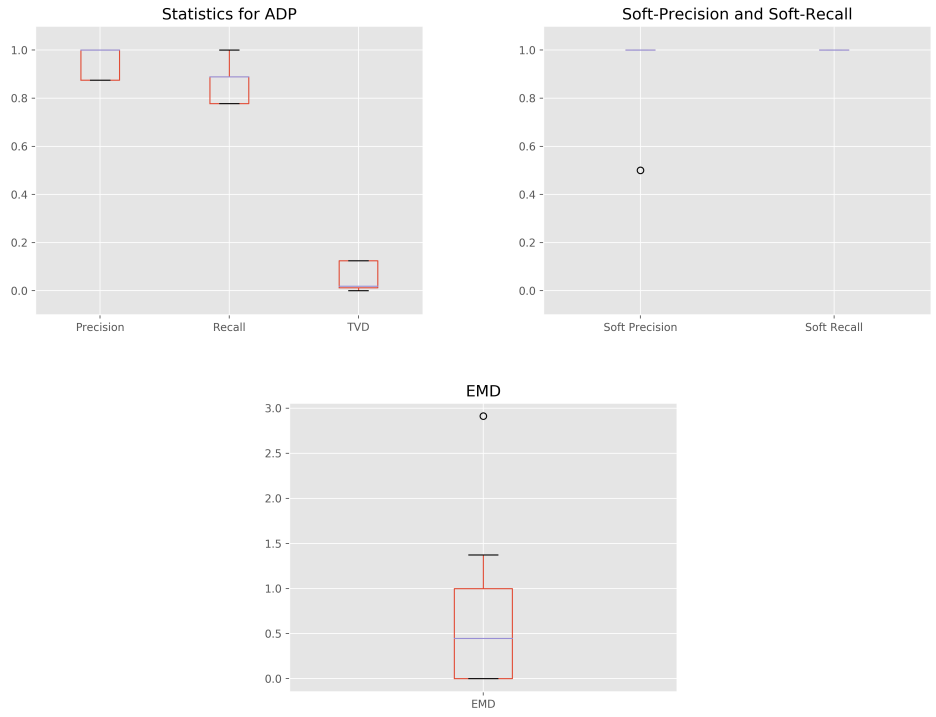


Figure 5: Our method: Box plots of (1) ADP statistics (2) Soft-Precision and Soft-Recall on SDP (3) EMD on SDP where k=1

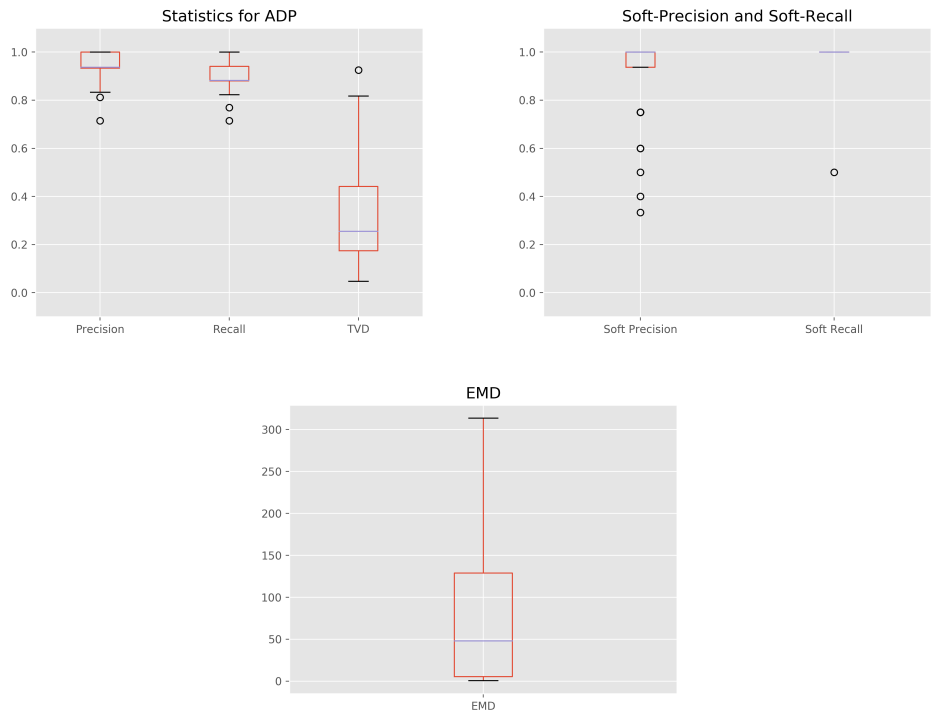


Figure 6: Our method: Box plots of (1) ADP statistics (2) Soft-Precision and Soft-Recall on SDP (3) EMD on SDP where k=2

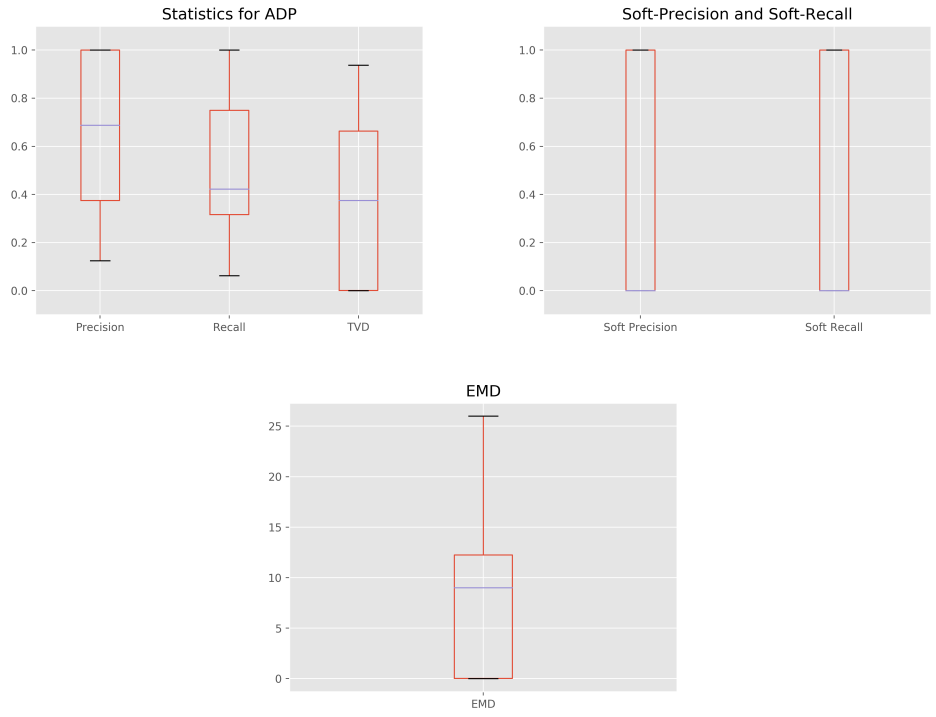


Figure 7: strainEST: Box plots of (1) ADP statistics (2) Soft-Precision and Soft-Recall on SDP (3) EMD on SDP where k=1

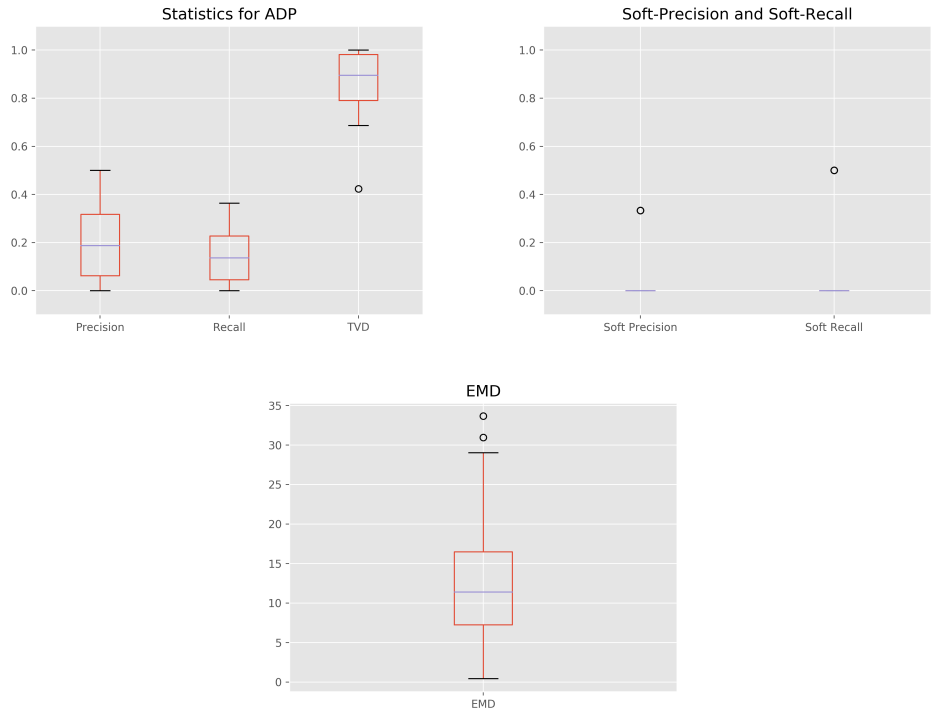


Figure 8: strainEST: Box plots of (1) ADP statistics (2) Soft-Precision and Soft-Recall on SDP (3) EMD on SDP where k=2