Guzmán-Sáenz *et al. BMC Genomics* 2019, **20**(Suppl 2):194
https://doi.org/10.1186/s12864-019-5490-y

BMC Genomics

RESEARCH Open Access

Signal enrichment with strain-level resolution in metagenomes using topological data analysis

Aldo Guzmán-Sáenz[1], Niina Haiminen[1], Saugata Basu[2] and Laxmi Parida[1*]

From The 17th Asia Pacific Bioinformatics Conference (APBC 2019)
Wuhan, China. 14-16 January 2019

Abstract

Background: A metagenome is a collection of genomes, usually in a micro-environment, and sequencing a metagenomic sample *en masse* is a powerful means for investigating the community of the constituent microorganisms. One of the challenges is in distinguishing between similar organisms due to rampant multiple possible assignments of sequencing reads, resulting in false positive identifications. We map the problem to a topological data analysis (TDA) framework that extracts information from the geometric structure of data. Here the structure is defined by multi-way relationships between the sequencing reads using a reference database.

Results: Based primarily on the patterns of co-mapping of the reads to multiple organisms in the reference database, we use two models: one a subcomplex of a Barycentric subdivision complex and the other a Čech complex. The Barycentric subcomplex allows a natural mapping of the reads along with their coverage of organisms while the Čech complex takes simply the number of reads into account to map the problem to homology computation. Using simulated genome mixtures we show not just enrichment of signal but also microbe identification with strain-level resolution.

Conclusions: In particular, in the most refractory of cases where alternative algorithms that exploit unique reads (i.e., mapped to unique organisms) fail, we show that the TDA approach continues to show consistent performance. The Čech model that uses less information is equally effective, suggesting that even partial information when augmented with the appropriate structure is quite powerful.

Keywords: Metagenomics, Topological data analysis, Multi-mapping reads, False positives

Background

A metagenome is a collection of genomes in a micro-environment, such as the gut of an animal, bottom of an ocean, or soil. This captures the influence of the immediate environment on the phenotype of an organism. For instance, one of the factors in the safety of our food supply chain is knowing the microbiome in the food [1]. The state of disease and health of a host has been shown to be related to the microbiomes in its gut [2]. The sturdiness or weakness of a plant is shown to be related to its soil microbiome [3]. It is turning out that these microorganisms are perhaps playing a much bigger role than earlier anticipated. The DNA technology to capture these organisms also has been disruptive in the area of microbiology, i.e., each organism does not need to be cultured individually before sequencing but the entire volume of samples can be put through the sequencing process en masse [4]. The obvious advantage is that the recalcitrant organisms that were resistant to being cultured no longer pose a problem, as long as careful sample processing is performed to avoid sequencing biases [5]. However, there are two major challenges with the sequencing approach.

*Correspondence: parida@us.ibm.com
[1]Computational Biology Center, IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Firstly, the completeness and correctness of reference databases limits the power of detection. The database of reference sequences must systematically be updated when new genomes become available. Secondly, no matter how complete and accurate the databases are, there is the problem of correctly assigning sequencing reads when distinct organisms with very similar genomes are present. In this paper, we address the second challenge of accurately detecting the organisms present in a sample, given a database.

Characteristics of the short sequencing reads frequently results in them being mapped to multiple reference genomes in the database, even under very strict matching criteria. Thus when the database organisms are very similar to each other there is usually a substantial dearth of reads assigned to unique organisms. As a consequence, most solution pipelines yield mapping results that are riddled with false positives. In fact, in (microbial) simulation studies we find that often a large percentage of the predicted potential organisms, using standard pipelines from literature, are false positives. A recent benchmark study found the number of species reported for the same sample varying by orders of magnitude, depending on the classifier used [6]. Popular methods for metagenomic read classification include Kraken [7], CLARK [8], MetaPhlAn [9] and others.

Topological data analysis (TDA) is emerging as a promising approach for analyzing large genomic datasets for a variety of questions [10–12], with already demonstrated applications in evolutionary biology, cancer genomics, and analysis of complex diseases. TDA extracts information from the geometric structure of data; in our application the structure is defined by the relationships between sequencing reads and organisms in a reference database.

Based primarily on the patterns of co-mapping of reads to the organisms in a reference database, we use a subcomplex of a Barycentric subdivision complex to model the multi-way maps of reads, along with the extent of read coverage on the respective organisms. This subcomplex allows a natural mapping of our problem to homology computation and interpretation. We test this approach on the special scenario (dearth of uniquely mapped reads) and find that using an appropriate voting function on the typical bar diagrams of TDA, we can sort the organisms to separate true positives from false positives. Next, we test if a reduced information set, that is, only the number of reads without utilizing their lengths or coverage when combined with TDA is powerful enough to enrich for true positives. We observe success in this scenario as well, suggesting that the use of topology captures the non-obvious structure defined by the reads promiscuously mapping to multiple organisms. All the data used in the paper are simulated reads by necessity since it is the most reliable way of knowing the ground truth, to assess the results.

Methods

In this section we map the problem to topological data analysis, more specifically to bar diagrams arising from persistent homology.

Motivation and problem statement

Given a reference database and a collection of metagenomic reads from the sequencing process, we denote by Z the set of all possible organisms, resulting from a reads-to-organism-mapping (ROM) pipeline: $Z = \{X_1, X_2, ..., X_N\}$. We treat all the reads of the pipeline output equally, i.e., do not consider the extent of the match. The rationale is that these characteristics have already been used by the pipeline to produce the output. Hence we focus on the subsequent information gathered from the pipeline: the relationship between the reads an the organisms.

Consider the bipartite graph of reads and organisms, where an edge connects a read to the assigned organism, that captures the ROM relationships completely. But a much more natural fit is to a simplicial complex where a k-simplex, $k > 0$ (analogous to edge in a graph) is a read and the 0-simplex (analogous to a vertex in a graph) is an organism. Furthermore each of the the k-simplices are weighted, i.e., the weight is the number of reads that map to exactly these k organisms, which is not immediately apparent from the bipartite graph. The next question is whether the topological features of this naturally emerging simplicial complex can be exploited to solve the problem of false positives. In other words, is there some hidden characteristics of the ROM that separate the true from the false organisms.

In this paper, to make it more tractable, we simplify the problem as follows: Given the ROM pipeline output, obtain an order for the elements of Z as $X_{i_1}, X_{i_2}, ..., X_{i_N}$, such that the top of the list is enriched for true positives (TPs). In a practical scenario, a threshold can be used to snip away the bottom of the list to eliminate potential false positive candidates. Further, the retained true positives can be used to rescue and re-assign the reads that had been misassigned due to the false positives.

Model I: Barycentric subdivision

Each organism X_j has a collection of reads mapped to it. For each subset $Y = \{X_1, \ldots, X_k\} \in 2^{\mathcal{X}}$, we denote by S_Y the set of reads mapped to the organisms $X_1, X_2, ..., X_k$, and none other.

For instance, this implies that $S_{\{X_1,X_2\}} \cap S_{\{X_1,X_2,X_3\}} = \emptyset$. In general,

$$S_Y \cap S_{Y'} = \emptyset, \text{whenever} Y \neq Y'.$$

Definition 1 (area a_Y) *For $Y \in 2^{\mathcal{X}}$, the area a_Y is defined to be the sum total of all the lengths of the reads in S_Y.*

Definition 2 (depth $d_Y(X_j)$ for $Y \in 2^{\mathcal{X}}$ and $X_j \in Y$) *For $Y \in 2^{\mathcal{X}}$ and $X_j \in Y$ we denote the length of the genome of the organism X_j covered by the reads in S_Y by l_{Y,X_j}, and define*

$$d_Y(X_j) = \frac{1}{m^2}\left(\frac{a_Y}{l_{Y,X_j}}\right),$$

where $m = |Y|$.

This is motivated by the fact that if there is true overlap in the m organisms, then the depth d of coverage of each organism is magnified by some factor of m. Note that it is quite possible that for $Y = \{X_1, X_2, \ldots, X_m\}$, the values $d_Y(X_1), d_Y(X_2), d_Y(X_3), \ldots, d_Y(X_m)$ are all distinct.

Example 1 *Consider a set of 3 organisms $\mathcal{X} = \{A, B, C\}$. A graphical representation of the underlying mapped reads is shown in Fig. 1 and the values of d_Y for each set $Y \in 2^{\mathcal{X}}$ are given in Fig. 2 (top).*

The depth values were computed after simulating a small set of reads' placement on three reference genomes. For example, in this case no reads match both B and C (and not A), hence $d_{BC}(B) = d_{BC}(C) = 0$. However, some reads match all three organisms, hence $d_{ABC} > 0$. The values $d_A(A), d_B(B), d_C(C)$ indicate the depths of A, B, C based on their uniquely mapping reads.

Definition 3 ***(set Y^t for a $t \geq 0$)*** *For $Y \in 2^{\mathcal{X}}$ and $t \in \mathbb{R}, t \geq 0$, we denote*

$$Y^t = \{X_j \in Y \mid d_Y(X_j) \geq t\}.$$

Note that the sets $(X_i)_{1 \leq i \leq N}$ give a *cover* of the set of reads, while the sets $(S_Y)_{Y \in 2^{\mathcal{X}}}$ form a *partition* of the reads. This corresponds to the collection of organisms present at a time t, mapped to some stage of a filtration of a simplicial complex (see Definition 5).

Definition 4 *Given any finite set S, a simplicial complex with vertices in S, is a collection K of subsets of S, with the property that if $A \in K$, then for all $B \in 2^S$ with $B \subset A$, $B \in K$. A subcollection L of K is a subcomplex of K if L is a simplicial complex.*

In other words, K is closed under the operation of "taking subsets". This last property is crucial for the definition of the simplicial homology groups, $H_*(K)$, associated to a simplicial complex K.

Definition 5 *A finite filtered simplicial complex is a finite sequence $K_0 \subset K_1 \subset \cdots \subset K_N$ of simplicial complexes. In other words, K_i is a subcomplex of K_j for $i \leq j$. A filtration of a simplicial complex K is a filtered simplicial complex with $K_N = K$.*

This particular setup allows us to apply *persistent homology* to our problem.

To any finite simplicial complex K one can associate for each $j \geq 0$, a finite dimensional vector space $H_j(K, Q)$ (called the *j-th simplicial homology group of K*). Moreover, if we have a filtration of a simplicial complex K

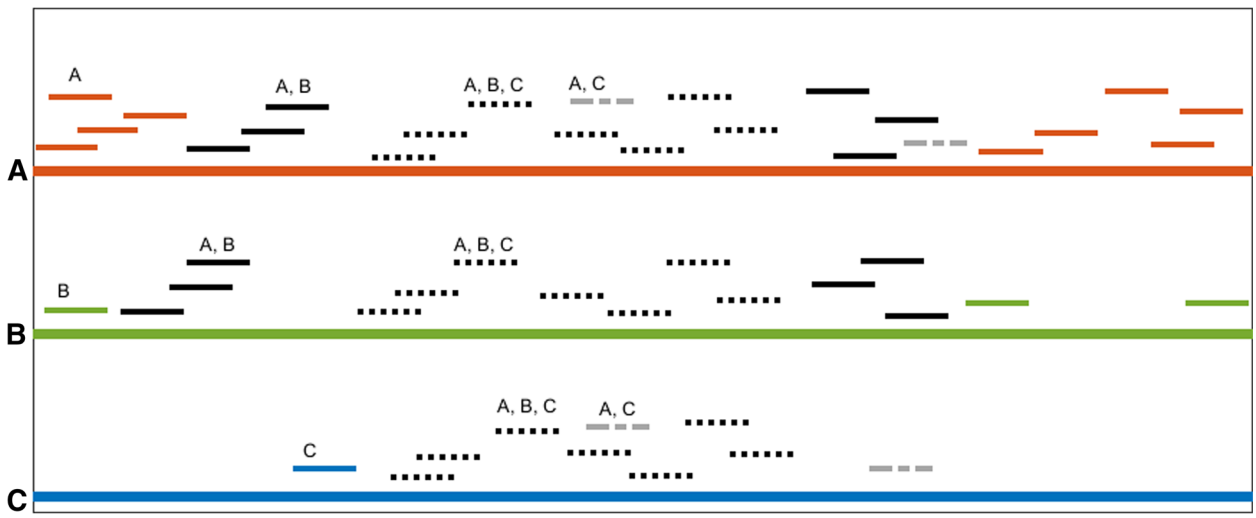

Fig. 1 A set of organisms {**a**, **b**, **c**} whose genomes are represented with a red, green, and blue line respectively. Reads mapped to the organisms are shown with shorter lines. The line colors and styles for the reads indicate unique and shared reads, according to the given labeling

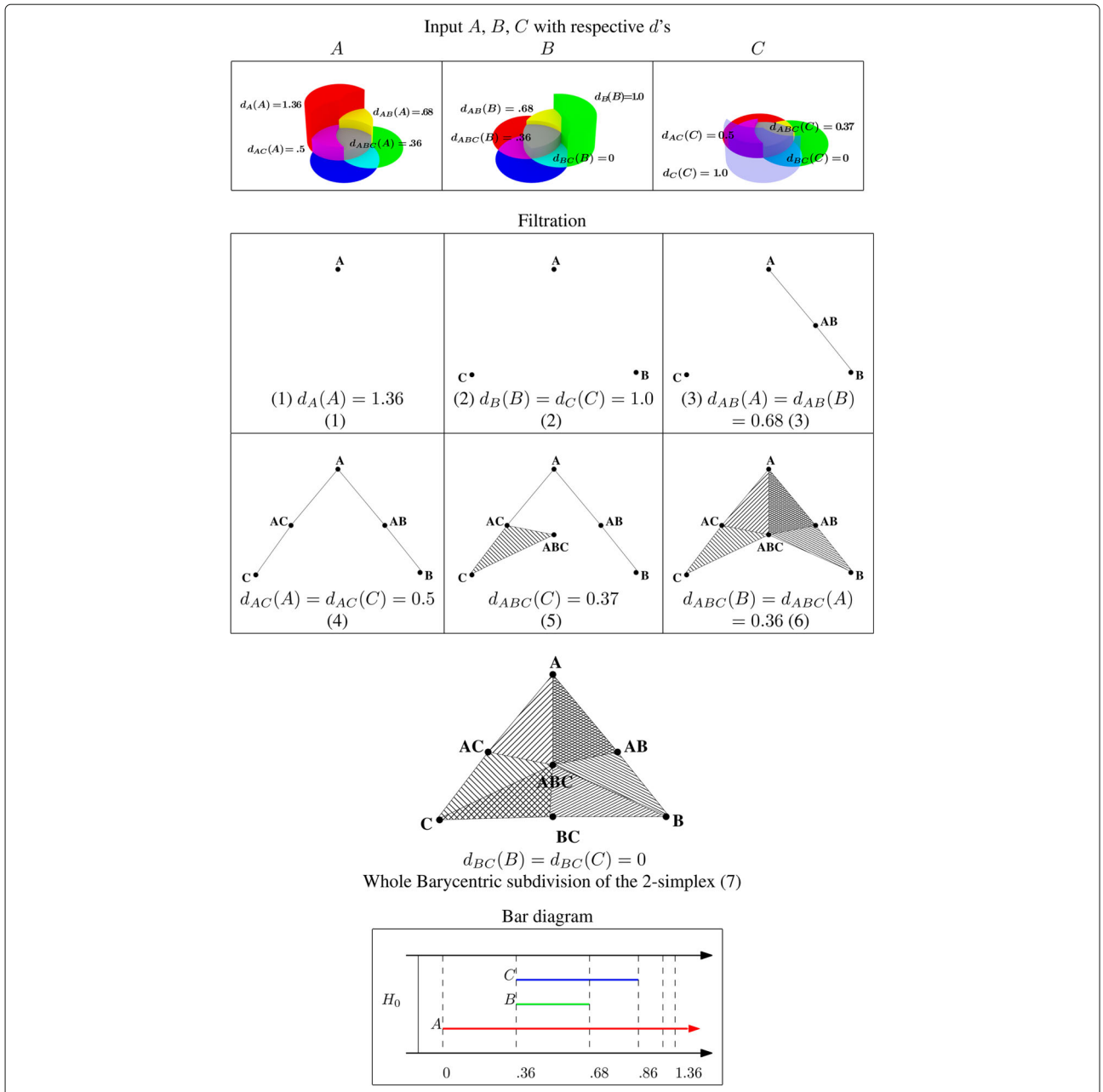

Fig. 2 Top: Values of depth *d* for the three organisms and read mapping shown in Fig. 1. Middle: Filtration of the Barycentric subdivision of the 2-simplex spanned by *A*, *B*, *C*. Bottom: The bar diagram in degree zero with organisms associated to bars

as above, then each inclusion $i_{s,t} : K_s \to K_t, s \leq t$ induces a *linear map*, $i_{s,t,*} : H_j(K_s, Q) \to H_j(K_t, Q)$. The images of these linear maps are usually called the *persistent homology groups* of the filtration, and their dimensions (i.e. the ranks of the maps $i_{s,t,*}$) determine the so called "bar diagram" associated to the filtration. Intuitively, the dimensions of the vector spaces $H_j(K, Q)$ (also called the *j*-th Betti number of the simplicial complex *K*) measure the number of independent *j*-dimensional cycles which are not boundaries of any $(j + 1)$-dimensional subcomplex of *K* (so called *j*-dimensional holes), and the *bar diagram* of the filtration is a record of the "times" of the births and deaths of these "homology classes", where we think of the sequence $(K_t)_{t\in[0,N]}$ as a complex growing with time *t*. Each bar in the bar diagram represents the interval in time in which a homology class *persisted*. We refer the reader to the book by Edelsbrunner and Harer [13] for further details about

persistent homology and its use in Topological Data Analysis.

However, notice that for each $t \in \mathbb{R}, t \geq 0$, the set system $(Y^t)_{Y \in 2^{\mathcal{X}}}$ does not necessarily satisfy the condition of being closed under taking subsets – and hence, does not necessarily form a simplicial complex with $\mathcal{X}$ as the set of vertices.

Instead, we utilize the following construction (a simplicial subcomplex of a *barycentric subdivision*, detailed at the beginning of this section) that does produce a simplicial complex (in fact a filtration of complexes), and which is also naturally aligned with the various functions $d_Y(\cdot)$ defined earlier.

Intuitively, the simplices of the barycentric complex correspond to chains of subsets of $\mathcal{X}$. For example, if $X_1, X_2, X_3 \in \mathcal{X}$, then the sequence $\{X_1\} \subset \{X_1, X_2\} \subset \{X_1, X_2, X_3\}$ is a chain in the partially ordered set $2^{\mathcal{X}}$ (ordered by inclusion). Given $t \geq 0$, we include the chain $\{X_1\} \subset \{X_1, X_2\} \subset \{X_1, X_2, X_3\}$ in the barycentric complex iff $\{X_1\}^t \cap \{X_1, X_2\}^t \cap \{X_1, X_2, X_3\}^t \neq \emptyset$.

Barycentric subdivision and its subcomplex of interest

We now define more precisely the barycentric complex, and the subcomplex we will use.

Let $\Delta_{\mathcal{X}}$ denote the $(|\mathcal{X}| - 1)$-dimensional simplex with vertex set $\mathcal{X}$.

A sequence $\boldsymbol{\sigma} = (Y_0, \ldots, Y_p)$, $Y_i \in 2^{\mathcal{X}}$ is called a *chain* of the poset $2^{\mathcal{X}}$ (ordered by inclusion) if $Y_0 \subsetneq Y_1 \subsetneq \cdots \subsetneq Y_p$. The first barycentric subdivision of $\Delta'_{\mathcal{X}}$ of $\Delta_{\mathcal{X}}$ (see Fig. 3) is then defined by

$$\Delta'_{\mathcal{X}} = \left\{ \mathbf{Y} = (Y_0, \ldots, Y_p) \mid \mathbf{Y} \text{is a chain in } 2^{\mathcal{X}} \right\}. \quad (1)$$

Now let $\mathcal{X}$ be a finite set, and let $\mathbf{d} = (d_Y : Y \to \mathbb{R})_{Y \in 2^{\mathcal{X}}}$ be a tuple of maps. For each $t \in \mathbb{R}$, we denote by $\Delta'_{\mathcal{X},\mathbf{d}}(t)$ the subcomplex of $\Delta'_{\mathcal{X}}$ defined as follows.

For $t \in \mathbb{R}$, and $\mathbf{Y} = (Y_0, \ldots, Y_p) \in \Delta'_{\mathcal{X}}$, let

$$\mathbf{Y_d}(t) = \bigcap_{0 \leq i \leq p} Y_i^t = \bigcap_{0 \leq i \leq p} \{X \in Y_i \mid d_{Y_i}(X) \geq t\}, \quad (2)$$

and set

$$\Delta'_{\mathcal{X},\mathbf{d}}(t) = \left\{ \mathbf{Y} = (Y_0, \ldots, Y_p) \in \Delta'_{\mathcal{X}} \mid \mathbf{Y_d}(t) \neq \emptyset \right\}. \quad (3)$$

We continue with Example 1 of three organisms A, B, C. In this case, the vertices of $\Delta'_{\mathcal{X}}$ are $\{A, B, C, AB, AC, BC, ABC\}$, where we write A for $\{A\}$, AB for $\{A, B\}$ and so on. The top part of Fig. 2 implies that we have only 7 different values for our functions d_Y. So we only need to specify the following collection of such $\Delta'_{\mathcal{X},\mathbf{d}}(t)$ (taking into account Remark 2 below):

$$\begin{aligned}
\Delta'_{\mathcal{X},\mathbf{d}}(0) &= \{[A]\} \\
\Delta'_{\mathcal{X},\mathbf{d}}(.36) &= \Delta'_{\mathcal{X},\mathbf{d}}(0) \cup \{[B], [C]\} \\
\Delta'_{\mathcal{X},\mathbf{d}}(.68) &= \Delta'_{\mathcal{X},\mathbf{d}}(.36) \cup \{[AB], [A, AB], [B, AB]\} \\
\Delta'_{\mathcal{X},\mathbf{d}}(.86) &= \Delta'_{\mathcal{X},\mathbf{d}}(.68) \cup \{[AC], [A, AC], [C, AC]\} \\
\Delta'_{\mathcal{X},\mathbf{d}}(.99) &= \Delta'_{\mathcal{X},\mathbf{d}}(.86) \cup \{[ABC], [C, ABC], [AC, ABC], [C, AC, ABC]\} \\
\Delta'_{\mathcal{X},\mathbf{d}}(1) &= \Delta'_{\mathcal{X},\mathbf{d}}(.99) \cup \{[A, ABC], [B, ABC], [AB, ABC], [A, AB, ABC], \\
&\qquad [A, AC, ABC], [B, AB, ABC]\} \\
\Delta'_{\mathcal{X},\mathbf{d}}(1.36) &= \Delta'_{\mathcal{X}}.
\end{aligned}$$

See Fig. 2 (middle) for a graphic representation of this filtered simplicial complex.

Fact 1 *$\Delta'_{\mathcal{X},\mathbf{d}}(t)$ is a simplicial complex with vertices in $2^{\mathcal{X}}$. Moreover, for any sequence $t_0 > t_1 > \ldots > t_N$,*

$$\Delta'_{\mathcal{X},\mathbf{d}}(t_0) \subset \Delta'_{\mathcal{X},\mathbf{d}}(t_1) \subset \cdots \subset \Delta'_{\mathcal{X},\mathbf{d}}(t_N).$$

Proof If $\mathbf{Y} = (Y_0, \ldots, Y_p) \in \Delta'_{\mathcal{X},\mathbf{d}}(t)$ is a chain, and $\mathbf{Z} = (Y_{i_0}, \ldots, Y_{i_q}), 0 \leq i_0 < \cdots < i_q \leq p$, is a subchain, then it is straightforward to check that

$$\mathbf{Y_d}(t) \subset \mathbf{Z_d}(t).$$

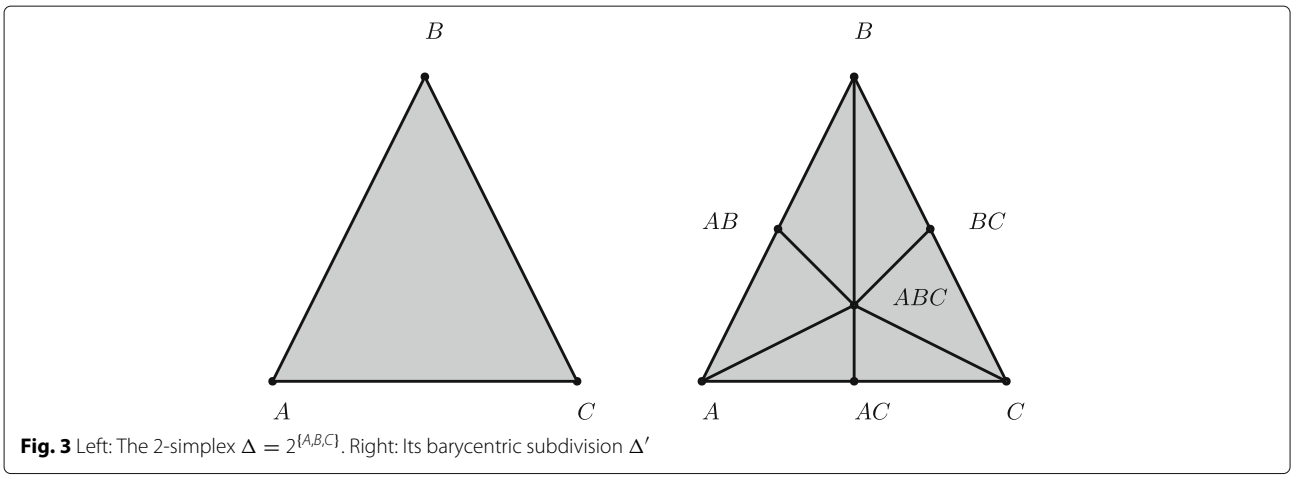

Fig. 3 Left: The 2-simplex $\Delta = 2^{\{A,B,C\}}$. Right: Its barycentric subdivision Δ'

Since $\mathbf{Y_d}(t)$ is not empty, $\mathbf{Z_d}(t)$ is not empty as well, so $\mathbf{Z} \in \Delta'_{\mathcal{X},\mathbf{d}}(t)$.

Now assume that $t_i > t_j$ for some $i,j \in \{0,\ldots,N\}$, and let $\mathbf{Y} \in \Delta'_{\mathcal{X},\mathbf{d}}(t_i)$. We have that if $\mathbf{Y_d}(t_i) \neq \emptyset$ then there exists some X such that for all Y_i in $\mathbf{Y}$ we have $d_{Y_i}(X) \geq t_i > t_j$, therefore $Y \in \Delta'_{\mathcal{X},\mathbf{d}}(t_j)$. □

Filtration

Armed with Fact 1 above, we have the following definition.

Definition 6 *Define the filtered simplicial complex*

$$\Delta'_{\mathcal{X},\mathbf{d}} = \bigcup_{t\in\mathbb{R}} \Delta'_{\mathcal{X},\mathbf{d}}(t). \tag{4}$$

This definition means that each of the simplicial complexes $\Delta'_{\mathcal{X},\mathbf{d}}(t)$ can be thought as a particular point in time of the filtration $\Delta'_{\mathcal{X},\mathbf{d}}$.

While the constructions so far deal with theoretical considerations, for actual computations we use the following remarks, which were in fact used to compute the filtration values for Example 1.

Remark 1 *Let* $\boldsymbol{\sigma} = (Y_0,\ldots,Y_p) \in \Delta'_{\mathcal{X}}$, *and let*

$$\tau = \bigcap_{Y\in\boldsymbol{\sigma}} Y.$$

Define t_0 *by*

$$t_0 = \max_{X\in\tau} \min_{Y\in\sigma} \{d_Y(X)\}. \tag{5}$$

Then t_0 *is the time of addition of* $\boldsymbol{\sigma}$ *to* $\Delta'_{\mathcal{X},\mathbf{d}}$.

Remark 2 *To make* $\Delta'_{\mathcal{X},\mathbf{d}}$ *covariant with respect to t, we use the change of variable* $t' = m - t$ *where m denotes the maximum value attained by the functions* d_Y.

From now on, unless stated otherwise, we will use the aforementioned change of variable.

Voting Scheme. Next we develop a voting scheme for all organisms that aims to solve the problem of ordering the organisms from true to false positives, using tools from algebraic topology. We refer the reader to Subsection 2.6 and Section 3 of [14] for the definition of persistent homology of a filtered simplicial complex, and a characterization in terms of $\mathcal{P}$-intervals respectively. In particular, using Corollary 3.1 of said article we have the following definition.

Definition 7 *Let K be a filtered simplicial complex and* $i \geq 0$ *an integer. We define the i-th degree bar diagram of K as the collection of* $\mathcal{P}$*-intervals associated to the i-th degree persistent homology of K.*

We now consider the persistent homology of the filtered complex $\Delta'_{\mathcal{X},\mathbf{d}}$. For $0 \leq i \leq |\mathcal{X}| - 1$ let $\mathcal{B}_i$ be a set of generators for the i-th degree persistent homology of $\Delta'_{\mathcal{X},\mathbf{d}}$, with fixed representing cycles, and $\mathcal{B} = \bigcup \mathcal{B}_i$. We can put these generators in a bijection with bars of the bar diagram of $\Delta'_{\mathcal{X},\mathbf{d}}$.

To help motivate the definition of the votings for each organism, we first list some of the components of the final formula. At each degree i, we assign to each organism X a collection of generators in $\mathcal{B}_i$ such that these generators witness X as a contributor to the features of our complex. For instance, for each X, consider all B such that X appears in its representing cycle (remember that such cycle is a linear combination of simplices, containing organisms as vertices). This gives rise to functions $F_i : \mathcal{X} \to 2^{\mathcal{B}_i}$.

For each generator $B \in \mathcal{B}$, let start(B) and end(B) denote its beginning and end as a bar. Consider now a strictly decreasing function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $f(x) \to 0$ as $x \to \infty$; this function will modulate the contributions of each bar to the voting value of a given organism along the filtration value, while a decreasing function $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ will modulate the contributions along homology degree.

With all this in place, the vote $\nu(X)$ for organism X is computed as follows:

$$\nu(X) = \sum_i g(i) \left(\sum_{B\in F_i(X)} \big(f(\text{start}(B)) - f(\text{end}(B))\big) \right) \tag{6}$$

Examples of f,g include $f(t) = \frac{1}{t+1}$ and $g(i) = \frac{1}{i+1}$, which is what we used to obtain the results in this paper.

Continuing our example from Figs. 1and 2 (bottom) shows the resulting bar diagram with associated generators after application of the mentioned procedure on the depth values. In this case there is only non-trivial homology in 0-th degree. After applying the voting scheme described above on the bar diagram shown in Fig. 2, the votes are $\nu(A) = 1.00, \nu(B) = 0.14, \nu(C) = 0.20$.

Model II: Čech complex

Recall that $Z = \{X_1, X_2, ..., X_N\}$ is the set of all possible organisms, resulting from an ROM pipeline. Let $\big(Y = S_{X_1X_2..X_k}\big) \in 2^Z$ be the set of reads associated with organisms X_1, X_2, ..., and X_k. Note that for all $i,j,k,...$,

$$|S_{X_i}| \geq |S_{X_iX_j}| \geq |S_{X_iX_jX_k}| \geq ... \tag{7}$$

The t (time) order filtration is $t_{\max}$ down to $t_{\min}$ in say steps of 1. A k-simplex on $X_1, X_2, ..., X_k, X_{k+1}$ at time t is introduced if $|S_{X_1X_2...X_kX_{k+1}}| \geq t$. Note that if the k-simplex on $X_1, X_2, ..., X_k, X_{k+1}$ belongs to the complex, then so does each $(k-1)$-simplex on $X_1, X_2, ..., X_k, X_{k+1}$ since, based on Eq. 7, for $1 \leq i \leq k$,

$|S_{Y_1Y_2Y_3...Y_k}| \geq |S_{X_1X_2X_3...X_kX_{k+1}}|$, where $Y_i \in \{X_1, X_2, X_3, ..., X_k, X_{k+1}\}$.

Each node or organism X gets a true-positive score $v(X)$ as follows: In the bar diagram, let b be the h-simplex $= X_0X_1...X_h$ with bar length denoted as len(b). In this model, the vote $v(X)$ for organism X is computed as follows:

$$v(X) = \sum_h \left(\sum_{b \in H_h, X \in b} h \times \text{len}(b) \right) \quad (8)$$

A simple example with four organisms is shown in Fig. 4.

Results

We applied the model on simulated shotgun sequencing reads from a collection of 36 recently published *Salmonella* genomes [15], in an effort to study the applicability of the approach to strain-level detection. We simulated 150 bp paired end reads, 100,000 per each input genome, with dwgsim [16] (with parameters $y0, e0.005, E0.005, d500, s0, r0.001, R0.15, X0.3$) from the 36 genomes. The reads were mapped to a database consisting of the same set of genomes using bowtie2 [17] (very-sensitive-local mode, searching for up to 101 hits per read). Read simulation and mapping was performed with the Metagenomics Computation and Analytics Workbench (MCAW) [18].

The mapped reads were processed with custom scripts to prepare the Barycentric depth values and Čech set sizes for each set of organisms. Concordance of paired reads was checked (both reads of a pair mapping to the same organism), a random representative selected if the organism had several hits from the same read, and only the best

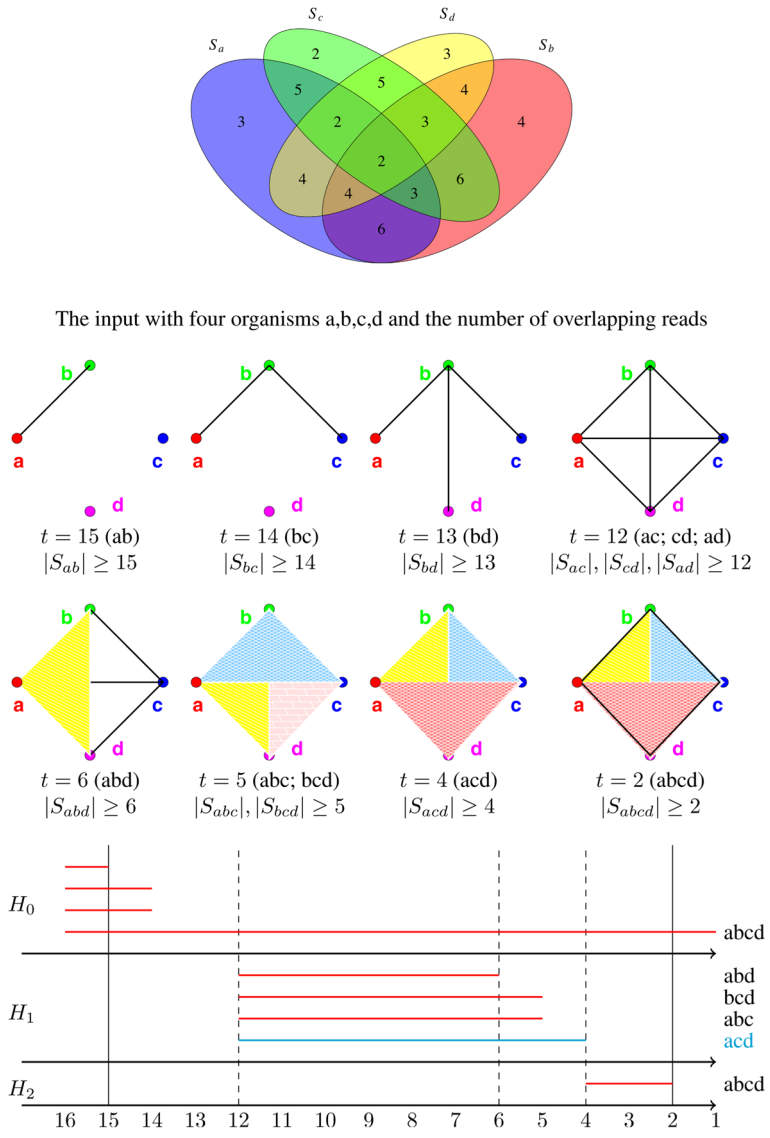

Fig. 4 Four organisms and the filtrations on the associated Čech complex with $t = 15$ down to $t = 2$

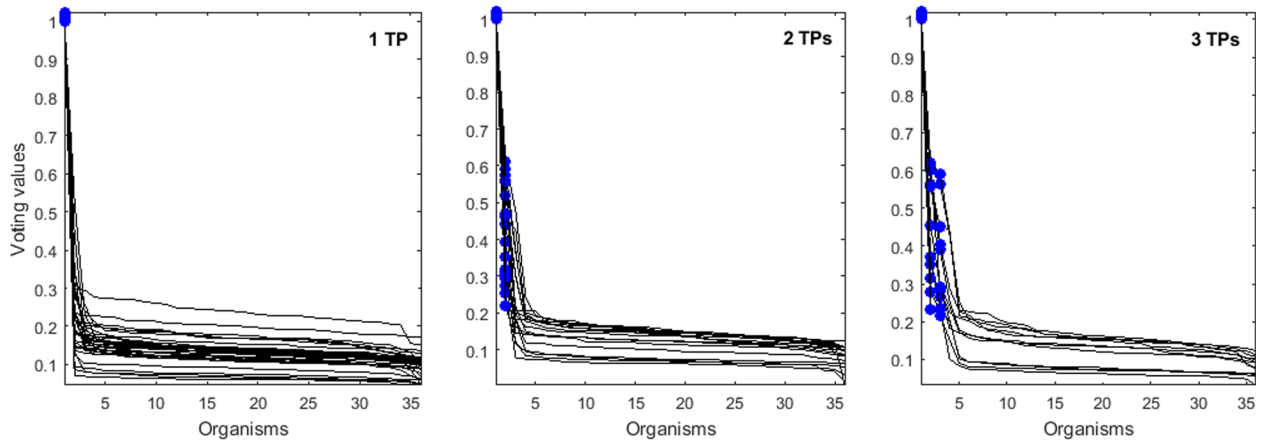

Fig. 5 Voting values (vertical axis) according to the Barycentric TDA method. The organisms (horizontal axis) are ordered according to their voting values from Eq. 6. Each line corresponds to a case where simulated reads from 1, 2, or 3 true positive (TP) input organisms were mixed together. The blue filled circles represent the voting values for the TPs

quality hits per read pair were used (based on sum of edit distances of reads in a pair). In this proof of concept we focused on the set of reads shared among 1 (unique reads), 2, or 3 organisms; the simulated data had 1–3 truly present organisms.

Indicative of the shared genome content between the closely related sequences, between only 2,087 to 92,176 of the 100,000 reads simulated per genome uniquely mapped to that genome after the process described above. Looking only at unique read counts per genome would certainly yield erroneous order of certain strains. For example, we observed a simulated mixture of 3 strains where two false positives (FPs) had more unique reads (2,596 and 2,427) than one of the true positives (TPs) (2,105). In this case, relying only on the number of unique reads would miss the truly present strain and falsely indicate the presence of missing ones. The observed unique reads for false positive organisms can arise from sequencing read errors, effects of the bioinformatics pipeline, and from subtle differences between the reference genomes and the observed genomes.

For analysis, the 36 *Salmonella* genomes were split into 18 non-overlapping sets of 2 strains each, and into 12 sets of 3 strains each. The TDA methodology, including the voting schemes described in Eqs. 6 and 8, was applied to the simulated reads from each set. The resulting voting values v for each organism were ordered from large (indicating truly present organisms) to small (indicating potential false positives).

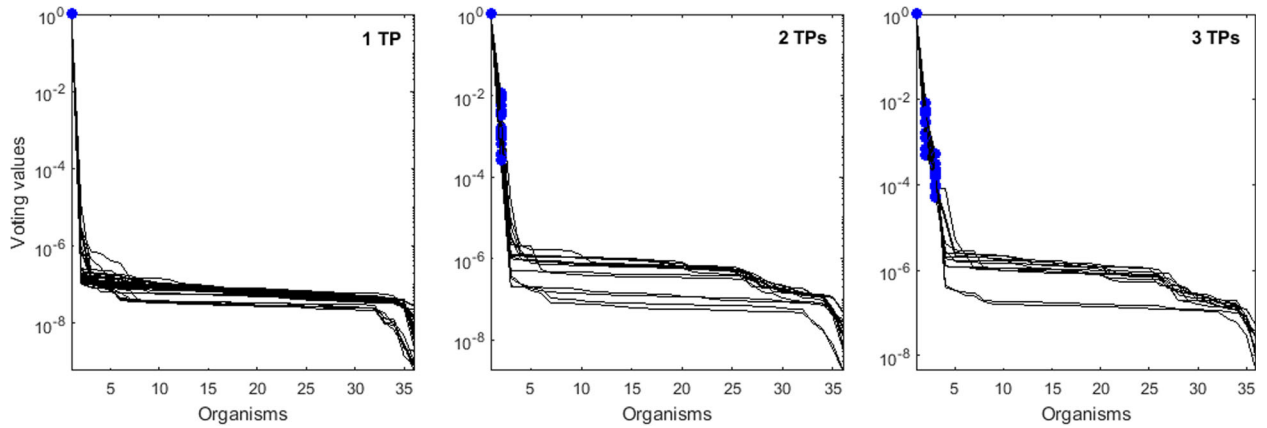

Fig. 6 Voting values (vertical axis) according to the Čech TDA method. The organisms (horizontal axis) are ordered according to their voting values from Eq. 8. Each line corresponds to a case where simulated reads from 1, 2, or 3 true positive (TP) input organisms were mixed together. The blue filled circles represent the voting values for the TPs

The voting values for simulated mixtures of 1, 2, or 3 *Salmonella* strains are visualized in Fig. 5 for the Barycentric approach and in Fig. 6 for the Čech approach. Ordering the organisms by voting values perfectly delineates the TPs from FPs with both methods. As a comparison, ordering the organisms by read coverage (computed by bedtools [19]) of the same reads as in the TDA input, the TPs could also be separated from FPs. However, as discussed earlier, using only the uniquely mapping reads would lead to false positives. We demonstrated that the TDA approach solves the problem of enriching the top of the list of voting values for truly present organisms.

Discussion and conclusions

In this proof of concept study we apply topological data analysis to the problem of separating signal from noise in the analysis of frequently multi-mapping metagenomic sequencing reads. Our approach is based on the construction of a particular subcomplex of a Barycentric subdivision complex, to rank-order the potential organisms and tease out the truly present ones.

The results from applying the approach on simulated genome mixtures show not just separation of signal from noise but also the potential for identifying microbes from metagenome samples, at strain level. We demonstrate the power of the TDA approach even in cases where alternative algorithms that exploit uniquely mapping reads fail. In fact, for the simple test cases the TPs bubble to the top from amongst a reference collection of highly related organisms, indicating promise of success for complicated real-life scenarios. The Čech model that uses less information is equally effective, suggesting that even partial information when augmented with the appropriate structure is quite powerful.

Additionally, the voting value curves show patterns of sharp decrease after the last true positive, suggesting automated calculation of cut-off thresholds. The same methodology could also be used to study higher taxonomic levels, e.g., separating true from false positive genera, families, etc. by modeling the read mappings across taxa.

Our next steps are to scale the implementation, and to apply it in the food safety as well as in the human health contexts. In both applications, a precise strain level assignment is of paramount importance.

Abbreviations

FP: False positive; ROM: Reads-to-organism mapping; TDA: Topological data analysis; TP: True positive

Acknowledgements

Not applicable.

Funding

Saugata Basu was partially supported by NSF Grant DMS-1620271. The publication charges were paid by the corresponding author's host institute.

Availability of data and materials

Not applicable.

About this supplement

This article has been published as part of *BMC Genomics Volume 20 Supplement 2, 2019: Selected articles from the 17th Asia Pacific Bioinformatics Conference (APBC 2019): genomics*. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-20-supplement-2.

Authors' contributions

LP conceived the model and designed the study. SB and AGS contributed towards the mathematical methodology. AGS designed and conducted the topological data analysis experiments. NH contributed to the methodology, and designed and conducted the metagenomic analysis. All authors contributed to writing the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Computational Biology Center, IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. [2]Department of Mathematics, Purdue University, West Lafayette, IN, USA.

Published: 10 April 2019

References

1. Kovac J, den Bakker H, Carroll LM, Wiedmann M. Precision food safety: A systems approach to food safety facilitated by genomics tools. Trends Anal Chem. 2017;96:52–61.
2. Hill-Burns EM, Debelius JW, Morton JT, et al. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. Mov Disord. 2017;32:5.
3. Finkel OM, Castrillo G, Paredes SH, Gonzalez IS, Dangl JL. Understanding and exploiting plant beneficial microbes. Curr Opin Plant Biol. 2017;38:155–63.
4. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: The Next Culture-Independent Game Changer. Front Microbiol. 2017;8:1069.
5. Costea PI, Zeller G, Sunagawa S, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;35:1069.
6. McIntyre ABR, Ounit R, Afshinnekoo E, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 2017;18:182.
7. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.
8. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16:236.
9. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9:8.
10. In: Przytyc ka T, editor. Research in Computational Molecular Biology. RECOMB 2015. Lecture Notes in Computer Science, vol 9029. Cham: Springer.
11. In: Natarajan R, Barua G, Patra MR, editors. Distributed Computing and Internet Technology. ICDCIT 2015. Lecture Notes in Computer Science, vol 8956. Cham: Springer.
12. Camara PG. Topological methods for genomics: Present and future directions. Curr Opin Syst Biol. 2017;1:95–101.

13. Edelsbrunner H, Harer JL. Computational Topology: An Introduction. Providence: American Mathematical Society; 2010.
14. Zomorodian A, Carlsson G. Computing Persistent Homology. Discrete Comput Geom. 2005;33:2.
15. Robertson J, Yoshida C, Gurnik S, Nash JHE. Completed Genome Sequences of Strains from 36 Serotypes of Salmonella. Genome Announc. 2018;6:3.
16. Homer N. DWGSIM: Whole Genome Simulator for Next-Generation Sequencing. 2010. https://github.com/nh13/DWGSIM. Accessed 23 Oct 2018.
17. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
18. Edlund SB, Beck KL, Haiminen N, Parida LP, Storey DB, Weimer BC, Kaufman JH, Chambliss D. Design of the MCAW compute service for food safety bioinformatics. IBM J Res Dev. 2016;5(/6):60.
19. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:6.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC