

 Open access • Posted Content • DOI:10.1101/2020.09.25.314252

## Detecting and phasing minor single-nucleotide variants from long-read sequencing data — [Source link](#)

Zhixing Feng, Jose C. Clemente, Brandon Wong, Eric E. Schadt

**Institutions:** Icahn School of Medicine at Mount Sinai, Johns Hopkins University

**Published on:** 27 Sep 2020 - bioRxiv (Cold Spring Harbor Laboratory)

**Topics:** Nanopore sequencing and Genetic heterogeneity

Related papers:

- [Detecting and phasing minor single-nucleotide variants from long-read sequencing data.](#)
- [NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks](#)
- [SVenX: A highly parallelized pipeline for structural variation detection using linked read whole genome sequencing data](#)
- [A Long-Read Sequencing Approach for Direct Haplotype Phasing in Clinical Settings.](#)
- [NanoCaller for accurate detection of SNPs and small indels from long-read sequencing by deep neural networks](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/detecting-and-phasing-minor-single-nucleotide-variants-from-30ccy8wz1e>

# Detecting and phasing minor single-nucleotide variants from long-read sequencing data

Zhixing Feng<sup>1,2\*</sup>, Jose Clemente<sup>1,2</sup>, Brandon Wong<sup>3</sup>, and Eric E. Schadt<sup>1,2,4</sup>

<sup>1</sup>Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at  
Mount Sinai, New York, NY, 10029, USA.

<sup>2</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai,  
New York, NY, 10029, USA.

<sup>3</sup>Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, 21218,  
USA.

<sup>4</sup>Sema4, Stamford, CT, 06902, USA.

\*To whom correspondence should be addressed. Email: zhixing.feng@mssm.edu

## Abstract

Cellular genetic heterogeneity is common in many biological conditions including cancer, microbiome, co-infection of multiple pathogens. Detecting and phasing minor variants, which is to determine whether multiple variants are from the same haplotype, play an instrumental role in deciphering cellular genetic heterogeneity, but are still difficult because of technological limitations. Recently, long-read sequencing technologies, including those by Pacific Biosciences and Oxford Nanopore, have provided an unprecedented opportunity to tackle these challenges. However, high error rates make it difficult to take full advantage of these technologies. To fill this gap, we introduce iGDA, an open-source tool that can accurately detect and phase minor single-nucleotide variants (SNVs), whose frequencies are as low as 0.2%, from raw long-read sequencing data. We also demonstrated that iGDA can accurately reconstruct haplotypes in closely-related strains of the same species (divergence  $\geq 0.011\%$ ) from long-read metagenomic data. Our approach, therefore, presents a significant advance towards the complete deciphering of cellular genetic heterogeneity.

## 25 Introduction

26 Cellular genetic heterogeneity is prevalent in multiple biological conditions. For example, the mi-  
27 crobiome contains multiple bacterial species with distinct genomes, and patients with infections may  
28 carry multiple bacterial strains. Likewise, in cancer, tumors are typically characterized by multiple  
29 cell types and cell lineages with different genomes. Deconvoluting such complex cellular genetic het-  
30 erogeneity is critical to basic biology and precision medicine. Minor variants, which are defined as the  
31 variants with frequencies lower than 10% in a cell population, play a central role in deciphering cellular  
32 genetic heterogeneity. Short-read genome sequencing can effectively characterize a large number of  
33 cells simultaneously but cannot phase minor variants directly due to the limitation of read length,  
34 which is generally under 300 bp<sup>1</sup>. Long-read sequencing, on the other hand, can be used to overcome  
35 this limitation. The latest long-read sequencing technologies, including those by Pacific Biosciences  
36 (PacBio) and Oxford Nanopore (ONT), enable sequencing more than 100 billion bases in a single run  
37 and yield reads with lengths that can exceed 10 kb<sup>2-4</sup>. These advantages make it feasible to adopt  
38 long-read sequencing to study cellular genetic heterogeneity in the microbiome, bacterial co-infection,  
39 and cancer in finer details. Because of its long read-length and high throughput, long-read sequencing  
40 has the potential to be applied to detect and phase minor variants at the single-molecule level without  
41 amplification. However, the error rate of raw long-read sequencing data is usually higher than 10%<sup>1,3</sup>,  
42 and makes it difficult to detect variants whose frequency is lower than the sequencing error rate.

43 Most of the existing methods to detect minor SNVs are based on short-read sequencing data<sup>5-14</sup>.  
44 The vast majority of these methods scan the reference genome and detect SNVs or other variants  
45 locus-by-locus. These methods cannot be used for long-read sequencing data because they are based  
46 on the error pattern of short-read sequencing data, which is different from long-read sequencing data.  
47 Researchers have also tried to leverage the information of multiple SNVs to increase detection accu-  
48 racy. V-Phaser and V-Phaser2<sup>15,16</sup>, which were designed for short-read sequencing data, use the joint  
49 probability of SNV pairs to detect SNVs. However, to avoid combinatorial explosion, they only use  
50 the joint probability of two SNVs. We will discuss the limitations of such a restriction for long-read  
51 sequencing and demonstrate how it leads to false negatives in Results.

52 There are several methods designed specifically to detect variants from long-read sequencing data.  
53 The GenomicConsensus module (<https://github.com/PacificBiosciences/GenomicConsensus>) de-  
54 veloped by PacBio generates a consensus sequence from the aligned PacBio reads and compares it to  
55 the reference genome to identify variants. Nanopolish<sup>17</sup> is a variant caller designed specifically for ONT

56 data, and Clairvoyante<sup>18</sup> is a deep-learning based tool for Illumina, PacBio, and ONT data. These  
57 methods assume that samples only have one or two haplotypes and therefore cannot be applied to  
58 detect minor variants. MinorSeq (<https://github.com/PacificBiosciences/minorseq>), developed  
59 by PacBio, is designed to detect minor variants but requires its input to be Circular Consensus Se-  
60 quencing (CCS) reads<sup>19</sup>. CCS is a special protocol of PacBio sequencing, which sequences each DNA  
61 molecule multiple times to increase accuracy. However, CCS reduces read length by 10 to 20 fold to  
62 achieve low error rates, and read length is critical to phasing minor SNVs. Recently, several tools  
63 have been developed to detect variants by leveraging haplotype information from long-read sequencing  
64 data<sup>20-22</sup>, but they also assume that the number of haplotypes is one or two. Thus, they cannot be  
65 applied to detect minor variants. To our best knowledge, there is currently no tool available to detect  
66 minor SNVs from raw data of long-read sequencing.

67 There are several short-read based methods available to phase minor SNVs<sup>23-29</sup>. These methods  
68 cluster the reads locally and phase distant SNVs, whose distances are longer than read length, using  
69 statistical models with strong assumptions. The major limitation of these methods is that they phase  
70 distant minor SNVs only based on indirect evidence because the read length is too short to span over  
71 the distant SNVs. This limitation can be overcome by using long-read sequencing data. The existing  
72 haplotyping methods for long-read sequencing data<sup>20-22</sup> assume there are only one or two haplotypes,  
73 and thus cannot be used to phase minor SNVs because the number of haplotypes is unknown.

74 To address the challenges of detecting and phasing minor SNVs, we developed a novel tool named  
75 iGDA (*in vivo* Genome Diversity Analyzer), which can accurately detect and phase minor SNVs, whose  
76 frequencies are as low as 0.2%. To detect minor SNVs, iGDA leverages the information of multiple  
77 loci without restricting the number of dependent loci, and uses a novel algorithm, Random Subspace  
78 Maximization (RSM), to overcome the issue of combinatorial explosion. To phase minor SNVs, iGDA  
79 uses a novel algorithm, Adaptive Nearest Neighbor clustering (ANN), which makes no assumption  
80 about number of haplotypes. To evaluate the performance of iGDA, we tested it on four pooled long-  
81 read sequencing datasets. The number of samples pooled in each dataset ranges from 65 to 755. The  
82 results demonstrate that iGDA can detect 85.8% to 96.7% of the real SNVs in these datasets at false  
83 discovery rate (FDR) lower than 1%. Finally, iGDA can phase minor SNVs at average accuracies range  
84 from 90.7% to 98.7%. We also tested iGDA on a pooled long-read metagenomic dataset consisting  
85 of 11 *Borrelia burgdorferi* strains and 744 other bacterial species, and discovered that the accuracy  
86 of iGDA is sufficient to reconstruct haplotypes in closely-related conspecific strains (strains belonging

87 to the same species) only using one reference genome. The divergences between the distinguishable  
88 conspecific strains are as low as 0.011%. These results shed light on tackling a number of challenges  
89 such as extracting strain-resolved genome sequences from long-read metagenomic data and identifying  
90 multiple strains in co-infection.

## 91 **Results**

### 92 **Detecting minor SNVs by leveraging information of multiple loci**

93 The major challenge of detecting minor SNVs is to distinguish between real SNVs and sequencing  
94 errors. It is especially difficult for raw data of long-read sequencing technologies, including those  
95 by PacBio and ONT, because they have relatively high error rates. However, we could leverage the  
96 fact that long reads can cover multiple SNVs to substantially increase detection accuracy. Intuitively,  
97 assuming that sequencing errors are independent, multiple sequencing errors are unlikely to repeatedly  
98 occur together on the same read. For example, in a pooled PacBio sequencing dataset consisting of  
99 186 *Bordetella* spp. samples (Figure 1A), the substitutions from the five marked loci occur together  
100 on 28 reads and there are 23,432 reads covering these five loci. The observed joint probability that  
101 these five substitutions occur together on the same read is  $28/23432 = 0.00119$ , while the expected  
102 joint probability is less than  $0.1^5 = 0.00001$  because substitution error rate of raw PacBio reads is  
103 less than 0.1 on this dataset (Figure 1B). The observed joint probability is over 100 times higher than  
104 the expected joint probability, so it is very likely that some of the five substitutions are real SNVs.  
105 However, the substitution rates of these five SNVs are 0.00569, 0.00845, 0.00748, 0.00960, and 0.00915  
106 respectively and it is difficult to distinguish them from sequencing errors only based on substitution  
107 rate (Figure 1B). Based on these observations, we propose a novel framework that uses conditional  
108 substitution rate instead of substitution rate to detect SNVs. In this framework, for each substitution,  
109 we adopt the maximal probability of observing the substitution conditional on observing substitutions  
110 at  $p$  other loci, defined as maximal conditional substitution rate, to detect whether the substitution  
111 is a real SNV. We call these  $p$  loci dependent loci. However, as the  $p$  dependent loci are unknown,  
112 it is infeasible to enumerate all combinations of these  $p$  loci to calculate the maximal conditional  
113 substitution rate due to high computational cost. As  $p$  is unknown, the number of combinations is  
114 about  $\sum_{p=1}^{2l} C_{2l}^p = 2^{2l} - 1$  for each locus if the average read length is  $l$ . We propose a novel algorithm  
115 called Random Subspace Maximization (RSM) to estimate the maximal conditional substitution rate

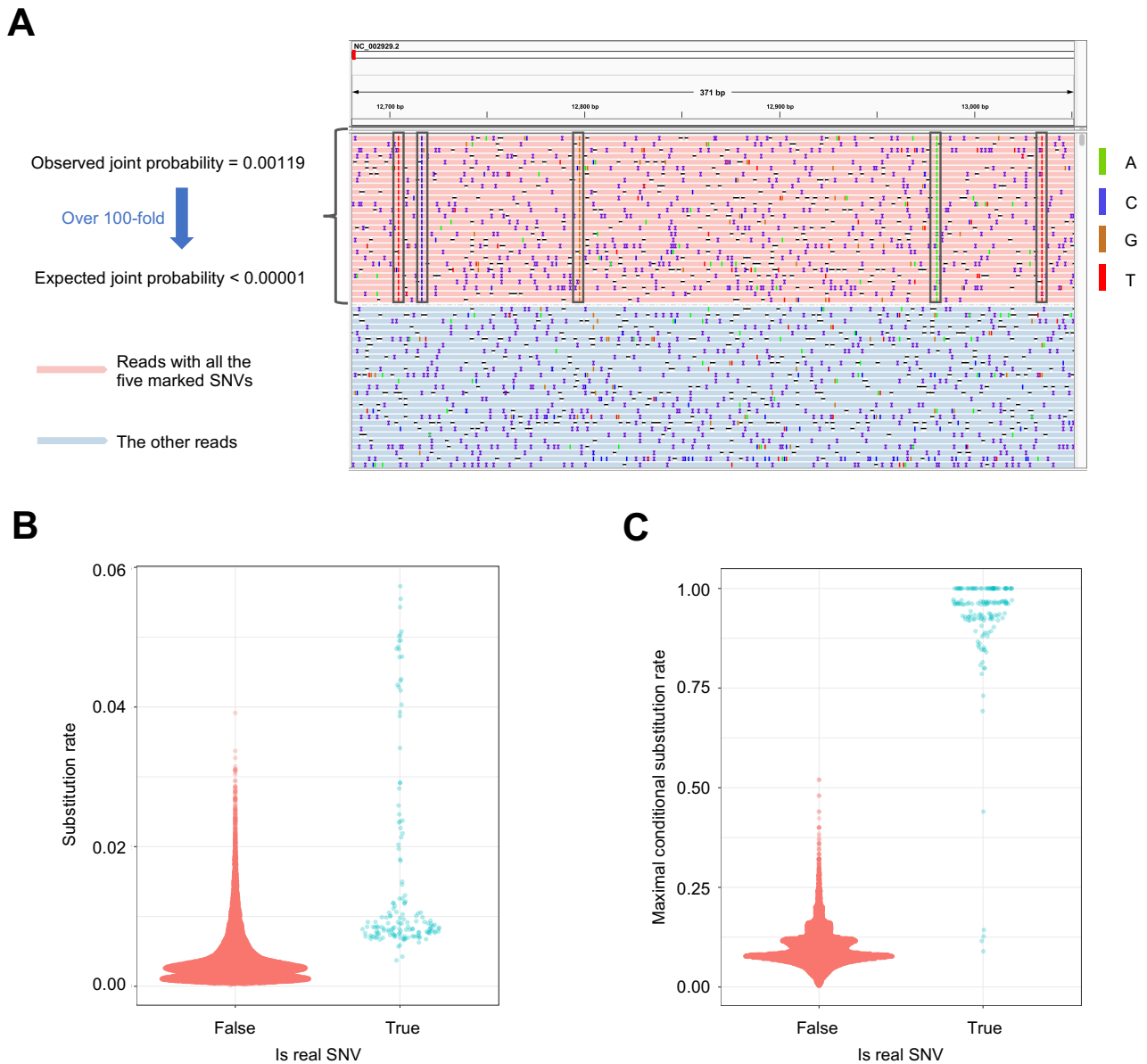


Figure 1: **SNVs are dependent on each other.** **A**, An IGV (Integrative Genomics Viewer)<sup>30</sup> snapshot demonstrating how to use the information of multiple loci to increase detection accuracy of SNVs. The number of reads containing the five SNVs marked by black boxes is 28 and the number of reads covering the five SNVs is 23,432. The observed and expected joint probabilities of the five SNVs are shown to the left of the IGV snapshot. Some reads are not shown in the figure due to the limit of figure size. **B**, The distribution of substitution rate on the *Bordetella* spp. data. No outlier is removed in the Sina plot. **C**, The distribution of maximal conditional substitution rate estimated by the RSM algorithm on the *Bordetella* spp. data. No outlier is removed in the Sina plot.

116 efficiently (Figure 2A-C) (details are in Methods). As shown in Figure 1C, on the *Bordetella* spp. data,  
117 the real SNVs and the sequencing errors are highly distinguishable based on the maximal conditional  
118 substitution rate calculated by the RSM algorithm.

119 It is very important to note that the number of dependent loci  $p$  should not be fixed. Figure  
120 S1 shows an example that fixing  $p$  can induce false negatives. In this example, the substitution  
121 at the locus 1 is independent with the substitutions at locus 2 and locus 3 respectively, but highly  
122 dependent on the combination of the substitutions at locus 2 and locus 3. Thus, the SNV at locus 1  
123 is difficult to be detected if  $p$  is fixed to 1, but is easy to be detected if there is no restriction on  $p$ .  
124 The existing algorithms V-Phaser and V-phaser2<sup>15,16</sup> were designed to identify minor variants from  
125 short-read sequencing data and only leveraged dependence between substitutions at two loci to avoid  
126 combinatorial explosion. This is equivalent to fixing  $p$  to 1, and making these algorithms unable to  
127 detect the SNVs in Figure S1. The proposed RSM algorithm has no restriction on  $p$  and can avoid  
128 combinatorial explosion.

129 If a SNV is the only SNV in the genome, we call it an orphan SNV. The proposed framework  
130 that uses conditional substitution rate to detect SNVs cannot detect orphan SNVs because its basic  
131 assumption is that there are multiple real SNVs in the same genome. We propose a single-locus based  
132 algorithm to overcome this limitation (Figure 2D). We discovered that substitution error rate is very  
133 different from locus to locus and it is highly predictable by sequence context (Figure 3). We trained  
134 a gradient boosting model<sup>31</sup> on independent public data and predicted substitution error rate for  
135 each locus. We then adopted a likelihood ratio test to compare the observed substitution rate to the  
136 predicted substitution error rate and reported a SNV if they are significantly different (details are in  
137 Methods).

### 138 **Phasing minor SNVs**

139 Intuitively, the reads of the same genome should be clustered together and the consensus sequence of  
140 each cluster can be used to phase minor SNVs. Herein, we propose a novel algorithm called Adaptive  
141 Nearest Neighbor (ANN) to cluster the reads and the consensus sequence of each cluster is called a  
142 draft contig (Figure 2E and Figure 2F) (details are in Methods). To reduce noise, loci with no detected  
143 SNVs are masked before applying ANN algorithm. A major advantage of ANN algorithm is that it  
144 can estimate the number of clusters automatically while clustering the reads. To reduce false positive  
145 rate of the draft contigs, we adopted a two-step filter to remove unreliable draft contigs (Figure 2G).

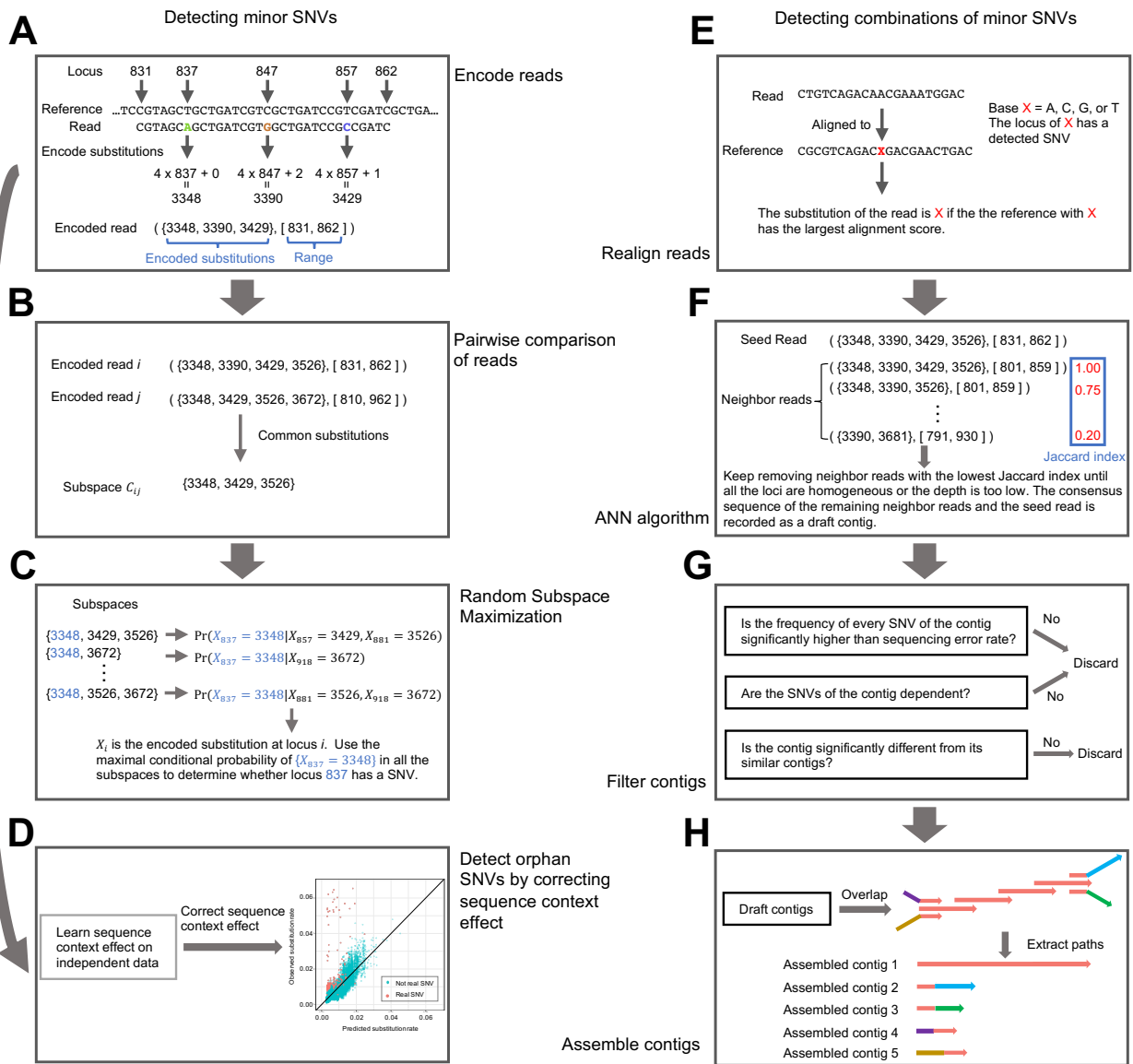


Figure 2: The main steps of iGDA. Details are in the Methods section



146 Intuitively, the SNVs in the same draft contig should be dependent with each other and the difference  
147 between two similar draft contigs should be statistically significant.

148 The lengths of the draft contigs are usually smaller than genome size. To maximize the range where  
149 the minor SNVs can be phased, we assemble the draft contigs using an algorithm inspired by overlap  
150 graph<sup>32</sup> (Figure 2H) (details are in Methods). The assembled draft contigs are called contigs.

## 151 **Evaluating performance on pooled PacBio sequencing data**

152 We constructed two datasets to test the accuracy of iGDA. The first dataset is a mixture of PacBio  
153 sequencing data of 186 *Bordetella* spp. samples, and the second dataset is a mixture of 155 *Escherichia*  
154 *coli* samples. The datasets have been previously published and their accession IDs in the SRA database  
155 (<https://www.ncbi.nlm.nih.gov/sra>) are listed in Table S1. The average sequencing depths of  
156 pooled data are 29,208x for *Bordetella* spp. and 19,175x for *Escherichia coli*. We downloaded the raw  
157 data in HDF format from SRA, and filtered the reads by requiring the estimated read quality (rq)  
158 greater than 0.75. The estimated read quality were extracted from the native HDF file. Bases with  
159 quality value (QV) less than a threshold were masked. We tested four thresholds, 0, 8, 10, and 12,  
160 respectively. We aligned the filtered reads to the reference genomes of *Bordetella pertussis* Tohama I  
161 (NCBI Reference Sequence ID is NC\_002929.2) for the *Bordetella* spp. data and *Escherichia coli* K12  
162 MG1655 (NCBI reference sequence ID is NC\_000913.3) for the *Escherichia coli* data by minimap2  
163 (version 2.12)<sup>33</sup> respectively. To minimize the alignment ambiguity caused by the aligner, we realigned  
164 the reads mapped to the negative strand by aligning their reverse complementary sequences. We  
165 only retained the reads aligned to the concatenated *rpoB* and *rpoC* region, which is highly conserved.  
166 The 1-based coordinates of the reference genomes is [11662, 20018] for *Bordetella pertussis* Tohama  
167 I and [4181245, 4189573] for *Escherichia coli* K12 MG1655. We pooled the realigned reads aligned  
168 to the concatenated *rpoB* and *rpoC* region for *Bordetella* spp. and *Escherichia coli* respectively to  
169 construct the two datasets. To evaluate accuracy of iGDA, we ran PacBio's genome consensus module  
170 (<https://github.com/pacificbiosciences/genomicconsensus>) on the aligned reads of each sample  
171 with default parameters to obtain the consensus genome sequences and SNVs. The union of the SNVs  
172 were used as benchmark to evaluate the accuracy of detecting SNVs. The genome sequence of an  
173 individual sample is defined as a real contig, and was used to evaluate the accuracy of contigs reported  
174 by iGDA. We merged samples (real contigs) with identical SNV profiles and calculated the relative  
175 abundances of the merged samples by the ratio between number of reads aligned to each sample and

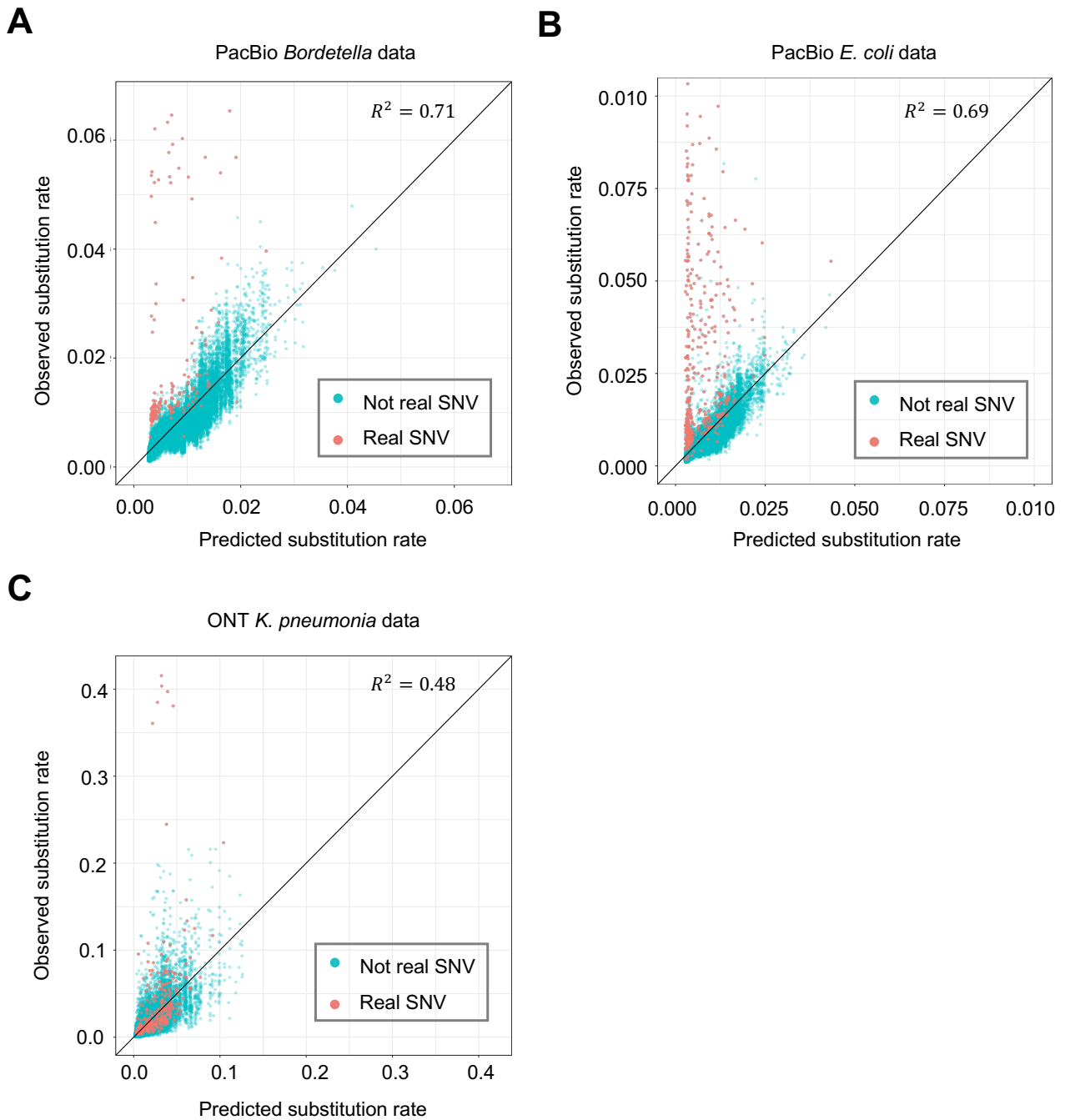


Figure 3: **Predicting substitution error rate by sequence-context-effect model trained on independent data.** **A**, Prediction of substitution error rate on the PacBio *Bordetella* spp. data. **B**, Prediction of substitution error rate on the PacBio *E. coli* data. **C**, Prediction of substitution error rate on the ONT *k. pneumoniae* data with DNA methylation masked.

176 the total number of aligned reads. The relative abundances of the samples distinct from the reference  
177 genome range from 0.25% to 3.05% for the *Bordetella* spp. data, and range from 0.30% to 1.92% for  
178 the *Escherichia coli* data. The average relative abundances are 0.82% and 0.74% for the *Bordetella*  
179 spp. data and the *Escherichia coli* data respectively.

180 For detecting minor SNVs, we tested three algorithms—a single-locus method (SL), which simply  
181 uses substitution rate of each locus to detect SNVs; a context-aware single-locus method (SLC), which  
182 uses substitution rate of each locus with correcting sequence-context effect (details are in Methods);  
183 and the proposed RSM algorithm—on these two test datasets. The results indicate that RSM algorithm  
184 greatly outperforms the two single-locus methods, and achieves a high accuracy (Figure 4A and Figure  
185 4B). With masking bases with QV lower than 8, iGDA detected 96.7% and 85.8% of the real SNVs  
186 at false discovery rate (FDR) lower than 1% for the *Bordetella* spp. data and *Escherichia coli* data  
187 respectively. Besides, correcting sequence-context effect substantially increases detection accuracy of  
188 the single-locus methods. The threshold of base QV also has minor impact on the accuracy. A non-zero  
189 threshold increases the accuracy on the *Bordetella* spp. data (Figure 4A), but decreases the accuracy  
190 on the *Escherichia coli* data (Figure 4B). This might be because masking bases with low QV removes  
191 some sequencing errors but reduces effective sequencing depth.

192 For phasing minor SNVs, we evaluated the ANN algorithm on these two datasets, where the bases  
193 with QV less than 8 were masked. The average accuracies (the maximal Jaccard index<sup>34</sup> with the real  
194 contigs) of the assembled contigs are 98.9% and 98.1% for the *Bordetella* spp. data and *Escherichia*  
195 *coli* data respectively (Figure 5A). Jaccard index between an iGDA-inferred contig and a real contig is  
196 the ratio between the number of shared SNVs and the total number of unique SNVs in their overlapped  
197 region. The IGV (Integrative Genomics Viewer)<sup>30</sup> snapshot of the contigs obtained from the *Bordetella*  
198 spp. data and the *Escherichia coli* data are shown in Figure 5B and Figure S2. The results show that  
199 the iGDA-inferred contigs match the real contigs very well, even for the real contigs with frequencies  
200 lower than 1%. In Figure 5B, there are five real contigs that are not detected by our algorithm. One  
201 of them has no SNV (the reference genome); two of them only have a single orphan SNV with very  
202 low frequency, which is hard for the RSM algorithm to detect; and two of them are highly similar to  
203 another genome. The results indicate that the minor SNVs can be phased effectively except for the  
204 genomes that have an orphan SNV or are highly similar to another genome.

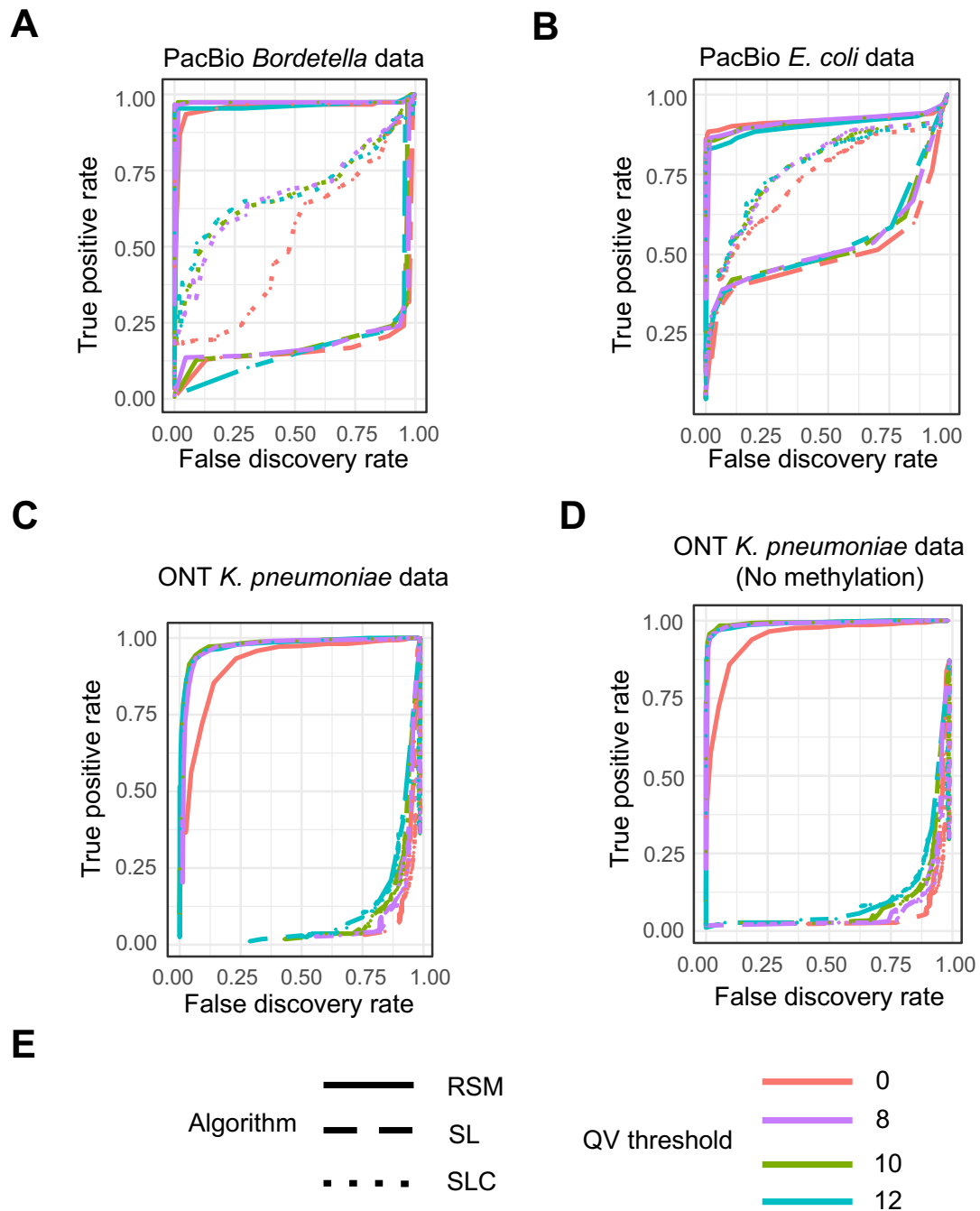


Figure 4: **The accuracy of detecting minor SNVs on pooled sequencing data.** **A**, The accuracy on PacBio *Bordetella* spp. data. **B**, The accuracy on PacBio *E. coli* data. **C**, The accuracy on ONT *K. pneumoniae* data. **D**, The accuracy on ONT *K. pneumoniae* data with DNA methylation masked. **E**, The legend of subfigures **A-D**. RSM = Random Subspace Maximization algorithm, SL = Single-Locus algorithm, SLC = Single-Locus algorithm with correcting sequence-context effect, and QV = Quality Value. True positive rate = number of correctly detected SNVs / number of real SNVs. False discovery rate = 1 - number of correctly detected SNVs / number of detected SNVs.

## 205 Evaluating performance on pooled ONT sequencing data

206 We tested iGDA on a dataset consisting of a mixture of ONT sequencing data of 65 *Klebsiella pneu-*  
207 *moniae* samples. The SRA IDs are listed in Table S2. We downloaded the raw data in fastq format  
208 from the SRA database (<https://www.ncbi.nlm.nih.gov/sra>), filtered and trimmed the reads using  
209 fastp<sup>35</sup>. The reads with average quality value (QV) less than 8 were discarded, and the first 50 bp  
210 and the last 200 bp were trimmed for each read. Similar to the PacBio data, we used four thresh-  
211 olds, 0, 8, 10, and 12 respectively, to mask bases with low QV. The reads were then aligned to the  
212 reference genome of *Klebsiella pneumoniae* subsp. *pneumoniae* HS11286 (NCBI reference sequence  
213 ID is NC\_016845.1). We realigned the reads mapped to the negative strand by aligning their reverse  
214 complementary sequences. We only retained the reads aligned to the concatenated *rpoB* and *rpoC*  
215 region, whose 1-based coordinate is [227354, 235682]. We then pooled the aligned reads to construct  
216 the testing data. To evaluate accuracy of iGDA, we downloaded assembly for each sample in the pooled  
217 data (Table S2) from NCBI (<https://www.ncbi.nlm.nih.gov/assembly>), and aligned the assembled  
218 genomes to the reference genome using MUMmer<sup>36</sup>. The union of the SNVs reported by MUMmer  
219 were used as benchmark to evaluate accuracy of detecting SNVs. The genome sequence of an indi-  
220 vidual sample is defined as a real contig, and was used to evaluate the accuracy of contigs reported  
221 by iGDA. We used the same method in the previous section to merge identical samples and obtain  
222 the relative abundance of each sample. The relative abundances range from 0.20% to 9.30%, and the  
223 average relative abundance is 3.20%.

224 Due to the unique sequencing mechanism of ONT, DNA methylation can affect the raw sequencing  
225 signal and substantially increase the base-calling error rate of methylated bases (Figure S3). The  
226 base caller used in the public ONT data in this study is Albacore (version 2.0) ([https://github.com/](https://github.com/Albacore/albacore)  
227 [Albacore/albacore](https://github.com/Albacore/albacore)). To avoid the impact of DNA methylation, we developed an algorithm to identify  
228 DNA methylation motifs in bacteria without using raw-signal of ONT data (details are in Methods).  
229 We masked loci within 5 bases to the DNA methylation motifs before applying iGDA to this dataset.

230 The result shows that the RSM algorithm substantially outperforms the single-locus methods to  
231 detect minor SNVs, and achieves a high accuracy (Figure 4C). With DNA methylation and bases with  
232 QV lower than 10 masked, iGDA detected 92.8% of the real SNVs at FDR lower than 1%. With  
233 masking no DNA methylation but masking bases with QV lower than 10, iGDA detected 41.3% of the  
234 real SNVs at FDR lower than 1%. Thus, masking DNA methylation increases the accuracy of the RSM  
235 algorithm (Figure 4D), which demonstrates the importance of removing DNA methylation or applying

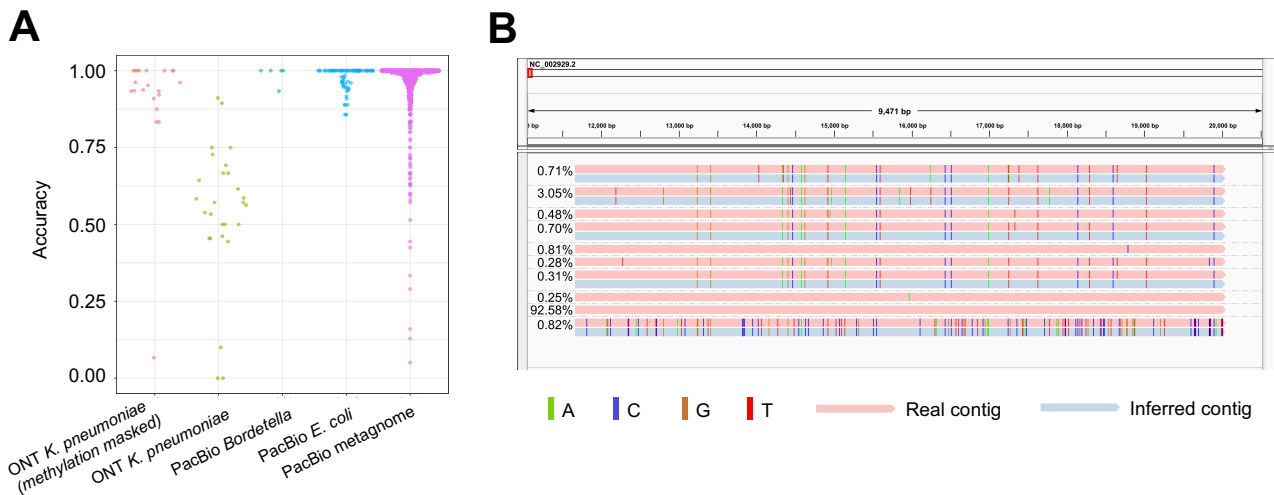


Figure 5: **The accuracy of phasing minor SNVs.** **A**, The sina plot of accuracy of phasing minor SNVs on the four testing datasets. **B**, The IGV snapshot of the contigs inferred by iGDA on the PacBio *Bordetella* spp. data. An inferred contig is grouped with its most similar real contig (measured by Jaccard index). Relative abundance is shown to the left of each contig.

236 a methylation-aware base caller to detecting minor SNVs from ONT data. Masking bases with low  
237 QV can substantially increase the accuracy and different thresholds have similar accuracies (Figure 4C  
238 and Figure 4D). In contrast to PacBio data, correcting sequence context does not significantly increase  
239 detection accuracy of the single-locus methods. We speculate that this is because the prediction power  
240 of sequence context on the ONT data is weaker than that on the PacBio data (Figure 3).

241 DNA methylation has a large impact on the accuracy of phasing minor SNVs. With masking loci  
242 affected by methylation and bases with QV lower than 10, the average accuracy of assembled contigs  
243 is 90.7% (Figure 5A). However, without masking loci affected by methylation, the average accuracy of  
244 assembled contigs is only 54.5% (Figure 5A). An IGV snapshot of methylation-masked contigs is shown  
245 in Figure S4. The result shows that the iGDA-inferred contigs match the real contigs very well with  
246 DNA methylation masked. It is critical to reduce the impact of DNA methylation by whole genome  
247 amplification or by adopting a methylation-aware base caller.

## 248 *De novo* identification of multiple *Borrelia burgdorferi* strains from long-read 249 metagenomic data

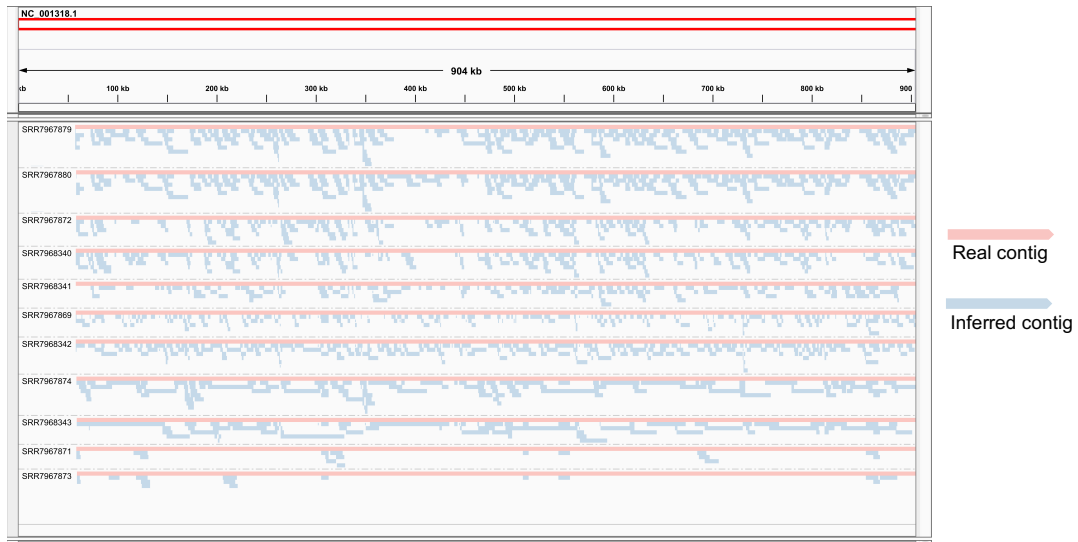
250 To test whether iGDA can be applied to identify multiple strains of the same species from metage-  
251 nomic data, we constructed a metagenomic dataset by mixing PacBio sequencing data of 11 *Borrelia*

252 *burgdorferi* strains, the causal agent of Lyme disease<sup>37</sup>, and 744 other bacterial samples. The SRA  
253 IDs, species, and strains are in Table S3. We filtered the reads by requiring read quality value greater  
254 than 0.75. Read quality (rq) was extracted from the native HDF files. Bases with QV less than 8  
255 were masked. We then aligned the reads to the reference genome of *Borrelia burgdorferi* B31 (NCBI  
256 reference sequence ID is NC\_001318.1), and realigned the reverse complementary of the reads mapped  
257 to the negative strand. To evaluate accuracy of iGDA, we assembled genome of each *Borrelia burgdor-*  
258 *feri* strain using flye<sup>38</sup>, and aligned the assembly to the reference genome using MUMmer<sup>36</sup> to obtain  
259 benchmark SNVs.

260 We ran iGDA on the realigned data and constructed 1,151 contigs. The average accuracy of  
261 the contigs is 95.0% (Figure 5A) and contig length is up to 139 kb. The IGV snapshots of the  
262 contigs reported by iGDA show that multiple strains of *Borrelia burgdorferi* can be clearly identified  
263 by iGDA (Figure 6A, Figure S5, and Figure S6). The minimal divergence of a region where the  
264 *Borrelia burgdorferi* strains can be distinguished is 0.011% (details are in Methods). To further evaluate  
265 the accuracy of iGDA, we performed MLST (Multilocus Sequence Typing)<sup>39</sup> on the contigs and the  
266 genome sequence of each strain using the database at <https://pubmlst.org/borrelia> (details are in  
267 Methods). In MLST, we aligned iGDA-inferred contigs and the genome sequence of each strain to the  
268 MLST database, consisting of known alleles of the eight house-keeping genes in *Borrelia* spp., to find  
269 the best matches. The result shows that most of the alleles that present in the genome sequence of  
270 each strain can be found in the iGDA-inferred contigs, and there is no false positive alleles (Figure  
271 6B). The alleles of the adjacent house-keeping genes, pyrG, recG, clpX, and pepX, can be phased by  
272 the contigs reported by iGDA (Figure 6B).

273 It is worth to note that some genome regions in Figure 6A are not covered by any contig. We call  
274 these regions missed regions, and call the SNVs not covered by any contig missed SNVs. We found  
275 that there are usually multiple strains that are highly similar to each other in the missed region. In  
276 the example shown in Figure S5, at least four samples have highly similar sequences in the missed  
277 region. Some missed regions have no SNV compared to the reference genome because iGDA does not  
278 report contigs with no SNV. In the example in Figure S6, samples SRR7967871 and SRR7967873 have  
279 several large missed regions, which have no SNV compared to the reference genome. To further assess  
280 the impact of highly similar strains on the performance of iGDA, we calculated Jaccard index of SNVs  
281 for each pair of the *Borrelia burgdorferi* samples, and found that some samples are highly similar to  
282 each other. The result in Figure S7A indicates that samples SRR7967879, SRR7967880, SRR7967872,

**A**



**B**

Distance	297 kb		120 kb		90 kb		8 kb		41 kb		18 kb		233 kb	
	nifS	clpA	rplB	pyrG	recG	clpX	pepX	uvrA						
SRR7967879	nifS_11	clpA_14	rplB_1	pyrG_1	recG_11	clpX_1	pepX_1	uvrA_10						
SRR7967880	nifS_11	clpA_14	rplB_1	pyrG_1	recG_11	clpX_1	pepX_1	uvrA_10						
SRR7967872	nifS_11	clpA_14	rplB_1 (99.84%)	pyrG_1	recG_1	clpX_1	pepX_1	uvrA_10						
SRR7968340	nifS_11	clpA_14	rplB_1	pyrG_1	recG_1	clpX_1	pepX_119	uvrA_10						
SRR7968341	nifS_11	clpA_14	rplB_1	pyrG_1	recG_1	clpX_1	pepX_1	uvrA_157						
SRR7967869	nifS_12	clpA_15	rplB_8	pyrG_1	recG_11	clpX_9	pepX_8	uvrA_16						
SRR7968342	nifS_12	clpA_15	rplB_8	pyrG_1	recG_11	clpX_9	pepX_8	uvrA_16						
SRR7967874	nifS_5	clpA_6	rplB_1	pyrG_1	recG_7	clpX_1	pepX_1	uvrA_8						
SRR7968343	nifS_4	clpA_10	rplB_1	pyrG_1	recG_6	clpX_5 (99.84%)	pepX_6	uvrA_6						
SRR7967871	nifS_1	clpA_1	rplB_1	pyrG_1	recG_1	clpX_1	pepX_1	uvrA_1						
SRR7967873	nifS_1	clpA_1	rplB_1	pyrG_1	recG_1	clpX_1	pepX_1	uvrA_1						

Alleles phased
Alleles phased

Detected allele  
 Missed allele

Figure 6: *De novo* identification of multiple *Borrelia burgdorferi* strains from PacBio metagenomic data. **A**, The IGV snapshot of the contigs inferred by iGDA from the metagenomic data. Each contig is grouped with its closest real contig (*B. burgdorferi* strain). **B**, MLST of *B. burgdorferi* in the metagenomic data. The columns are the alleles of the 8 house-keeping genes used in MLST. Each row is the alleles of the genome of each sample (strain). The row names are the SRA IDs of each sample. An allele is detected if it matches a contig inferred by iGDA. There are two alleles that have no 100% match in the MLST database, and their similarities to the closest alleles in the database are shown in the brackets. All the other alleles have a 100% match in the database.



283 SRR7968340 and SRR7968341 are highly similar to each other, and sample SRR7967869 is highly  
284 similar to sample SRR7968342. We constructed a new dataset where only one sample is retained out of  
285 the highly similar strains. Specifically, we excluded samples SRR7967879, SRR7967880, SRR7967872,  
286 SRR7968340 and SRR7968342 from the samples listed in Table S3, and reran iGDA on the new data.  
287 The result shows that the accuracy of each contig is not significantly changed by excluding highly  
288 similar strains (Figure S7B). However, the length of contigs and proportion of SNVs covered by contigs  
289 are substantially increased (Figure S7C and Figure S7D). The species other than *Borrelia burgdorferi*  
290 have limited impact on the results, because most of the reads from these species (Table S3) cannot be  
291 aligned to the reference genome of *Borrelia burgdorferi*, and 99.93% of the aligned reads are aligned to  
292 16S ribosomal RNA or 23S ribosomal RNA.

## 293 Discussion

294 We here present iGDA, a novel open-source tool implementing several innovative algorithms that can  
295 achieve a high accuracy for detecting and phasing minor SNVs. iGDA makes it feasible to study  
296 a number of previously challenging problems, such as constructing strain-level genome sequence in  
297 microbiome samples, and identifying genome sequence of pathogens in samples with co-infection. The  
298 RSM and ANN algorithms proposed in this work are generic methods and can be extended to apply  
299 to single-cell genome sequencing data or 10X genomics linked-read<sup>40</sup> data. In addition to genome  
300 sequencing, these algorithms have the potential to be applied in RNA sequencing data as well. For  
301 example, with an alternative preprocessing procedure, these algorithms can be used to decipher the  
302 heterogeneity of A-to-I RNA editing using long-read sequencing.

303 A major limitation of iGDA is that its high accuracy relies on the presence of multiple SNVs.  
304 Therefore, iGDA has reduced accuracy to detect orphan SNVs with very low frequency. Besides,  
305 presence of highly similar genomes will reduce accuracy of iGDA.

306 DNA methylation can induce correlated substitution errors on ONT data and reduce the accuracy  
307 of iGDA. Masking DNA methylation can increase the accuracy of iGDA on ONT data. Using whole  
308 genome amplification (WGA) to remove DNA methylation is a solution to this issue. Another solution  
309 is to use a base caller that can correct methylation induced error, but there is no such tool currently  
310 available according to our best knowledge.

311 In this work, we only detect minor SNVs because they are less affected by alignment ambiguity  
312 compared to insertions and deletions (Indel). Alignment ambiguity means an Indel might be located

313 to multiple loci in the genome but the corresponding alignment scores are equal. To extend our RSM  
 314 and ANN algorithms to detect minor Indels or other more complicated variants, an alternative way to  
 315 represent variants and alignments is needed.

## 316 Methods

### 317 Leveraging multiple loci to detect SNVs

318 For the  $i$ th aligned read, we encode its substitution at locus  $k$  of the reference genome by the following  
 319 formula:

$$s_{ik} = \begin{cases} 4k & r_{ik} \neq t_k, r_{ik} = A \\ 4k + 1 & r_{ik} \neq t_k, r_{ik} = C \\ 4k + 2 & r_{ik} \neq t_k, r_{ik} = G \\ 4k + 3 & r_{ik} \neq t_k, r_{ik} = T \\ \epsilon & r_{ik} = t_k \end{cases} \quad (1)$$

320 , where  $r_{ik}$  is the base (short for nitrogenous base) of the  $i$ th aligned read at locus  $k$ ,  $t_k$  is the base at  
 321 locus  $k$  of the reference genome and  $\epsilon$  is an empty element, which is formally defined by  $\{\epsilon\} = \emptyset$ . The  
 322 first locus of the reference genome is 0 throughout this paper unless otherwise stated. The  $i$ th read is  
 323 represented as a set of substitutions and its covering range (Figure 2A), and is denoted by

$$R_i = (S_i, [b_i, e_i]) \quad (2)$$

324 .  $b_i$  and  $e_i$  are the start and end loci of the region covered by the read respectively, and  $S_i$  is

$$S_i = \{s_{ib_i}, s_{ib_i+1}, \dots, s_{ie_i}\} \quad (3)$$

325 . The most intuitive way to detect SNVs is to use the substitution rate of each locus. Formally, we  
 326 denote the encoded substitution at locus  $k$  as a random variable  $X_k$ , and denote probability of the  
 327 event  $\{X_k = x_k\}$  as  $Pr(X_k = x_k)$ , where  $x_k \in \{4k, 4k + 1, 4k + 2, 4k + 3\}$ . Substitution rate is defined  
 328 as the estimated  $Pr(X_k = x_k)$ , which is

$$\hat{Pr}(X_k = x_k) = \frac{|\{i \mid x_k \in S_i\}|}{|\{i \mid k \in [b_i, e_i]\}|} \quad (4)$$

329 , where  $\{\cdot\}$  is a set and  $|\cdot|$  is the number of elements in a set. Intuitively, in equation (4), the numerator  
 330 is the number of reads with substitution  $x_k$  at locus  $k$ , and the denominator is the number of reads  
 331 covering locus  $k$ . Due to the high error rate of long-read sequencing data, it is inaccurate to detect minor

332 variants using substitution rate alone (Figure 1B). Herein, we leverage the information of multiple loci  
333 to increase the detection accuracy. Assuming sequencing errors are independent with each other, real  
334 SNVs are likely to be present if there are multiple reads containing the same set of substitutions (Figure  
335 1A). The conditional probability of  $\{X_k = x_k\}$  given other real SNVs of the same genome is therefore  
336 much larger than the marginal probability of  $\{X_k = x_k\}$  if  $x_k$  is a real SNV, because these real SNVs  
337 are positively dependent (Figure 1A and Figure 1C). Formally, the conditional probability of event  
338  $\{X_k = x_k\}$  given  $p$  other substitutions is defined as  $Pr(X_k = x_k | X_{g_1} = x_{g_1}, X_{g_2} = x_{g_2}, \dots, X_{g_p} = x_{g_p})$ ,  
339 which is estimated by

$$\hat{Pr}(X_k = x_k | X_{g_1} = x_{g_1}, X_{g_2} = x_{g_2}, \dots, X_{g_p} = x_{g_p}) = \frac{|\{i \mid \{x_k, x_{g_1}, x_{g_2}, \dots, x_{g_p}\} \subseteq S_i\}|}{|\{i \mid \{x_{g_1}, x_{g_2}, \dots, x_{g_p}\} \subseteq S_i, k \in [b_i, e_i]\}|} \quad (5)$$

340 . Intuitively, in equation (5), the numerator is the number of reads containing substitution  $x_k$  and the  $p$   
341 other substitutions, and the denominator is the number of reads that contain the  $p$  other substitutions  
342 and cover locus  $k$ . The  $p$  loci,  $g_1, g_2, \dots, g_p$  are called dependent loci. As  $x_{g_1}, x_{g_2}, \dots, x_{g_p}$ , and  $p$   
343 in equation (5) are unknown, the estimated maximal conditional probability of event  $\{X_k = x_k\}$  given  
344  $p$  other substitutions is used to detect SNVs, and is formally defined by

$$H(x_k) = \max_{p, x_{g_1}, x_{g_2}, \dots, x_{g_p}} \{\hat{Pr}(X_k = x_k | X_{g_1} = x_{g_1}, X_{g_2} = x_{g_2}, \dots, X_{g_p} = x_{g_p})\} \quad (6)$$

345 . The substitution  $x_k$  is detected as a real SNV if  $H(x_k)$  is larger than a threshold (0.65 in this study).  
346  $H(x_k)$  is also called maximal conditional substitution rate. To avoid high variance of the estimated  
347  $Pr(X_k = x_k | X_{g_1} = x_{g_1}, X_{g_2} = x_{g_2}, \dots, X_{g_p} = x_{g_p})$  (equation (5)), we require that  $|\{i \mid \{x_{g_1}, x_{g_2}, \dots, x_{g_p}\} \subseteq$   
348  $S_i, k \in [b_i, e_i]\}| \geq v_{min}$ , and  $v_{min} = 25$  in this study. Sequencing errors at multiple loci that are very  
349 close to each other might induce slightly dependent substitutions. To avoid the impact of dependent  
350 substitutions induced by sequencing errors, we require that locus  $k$  and loci  $g_1, g_2, \dots, g_p$  are not too  
351 close. Specifically, we require  $HD(k, g_s) \geq 15$  for any  $g_s \in \{g_1, g_2, \dots, g_p\}$ .  $HD(k, g_s)$  is the homopoly-  
352 mer distance between locus  $k$  and locus  $g_s$ , and is defined as the number of homopolymers between the  
353 two loci. A homopolymer is a set of consecutive identical bases, and a base with no identical adjacent  
354 bases is also defined as a special homopolymer with size equal to 1.

355 It is computationally infeasible to enumerate all combinations of  $p$  loci to estimate  $H(x_k)$  in equation  
356 (6). It is important to note that it is insufficient to detect SNVs accurately by restricting the number  
357 of dependent loci  $p$  to a certain number. In the example shown in Figure S1,  $H(x_k)$  fails to detect  
358 the real SNVs if  $p$  is restricted to 1. Likewise, we can also have similar examples if  $p$  is restricted to

359 another number greater than 1. In this work, we developed a novel algorithm called Random Subspace  
 360 Maximization (RSM) that can estimate  $H(x_k)$  efficiently without restricting  $p$ .

### 361 Detecting SNVs by RSM algorithm

#### 362 The greedy algorithm and its theoretical accuracy

363 We introduce a fast but inaccurate greedy algorithm to estimate  $H(x_k)$  (equation (6)), and then  
 364 improve its accuracy by Random Subspace Maximization (RSM) in the next section. To estimate  
 365  $H(x_k)$  for substitution  $x_k$  at locus  $k$ , we only need to consider dependent loci in range  $[k - t_l, k + t_r]$ ,  
 366 where

$$t_l = \max_t \{ |\{i \mid [k - t, k] \subseteq [b_i, e_i]\}| > 0 \}$$

$$t_r = \max_t \{ |\{i \mid [k, k + t] \subseteq [b_i, e_i]\}| > 0 \}$$

367 .  $[b_i, e_i]$  is the covering range of read  $R_i$  (equation (2)). We estimate  $Pr(X_k = x_k | X_g = x_g)$  by equation  
 368 (5) for each locus  $g \in [k - t_l, k + t_r] \cap \{k\}^c$  ( $\{\cdot\}^c$  is complement of a set), and sort the loci according  
 369 to  $Pr(X_k = x_k | X_g = x_g)$  in descending order. The sorted loci are denoted as  $\{s_1, s_2, \dots, s_{t_l+t_r}\}$ , and  
 370  $Pr(X_k = x_k | X_{s_{t-1}} = x_{s_{t-1}}) \geq Pr(X_k = x_k | X_{s_t} = x_{s_t})$ . We keep adding locus  $s_t$  to  $\{s_1, s_2, \dots, s_{t-1}\}$   
 371 if  $\hat{Pr}(X_k = x_k | X_{s_1} = x_{s_1}, X_{s_2} = x_{s_2}, \dots, X_{s_t} = x_{s_t}) > \hat{Pr}(X_k = x_k | X_{s_1} = x_{s_1}, X_{s_2} = x_{s_2}, \dots, X_{s_{t-1}} =$   
 372  $x_{s_{t-1}})$  and stop if otherwise.  $\hat{Pr}(X_k = x_k | X_{s_1} = x_{s_1}, X_{s_2} = x_{s_2}, \dots, X_{s_v} = x_{s_v})$  based on the final  $v$   
 373 selected loci  $\{s_1, s_2, \dots, s_v\}$  is used to estimate  $H(x_k)$ .

374 The naive greedy algorithm described above avoids combinatorial explosion but might have low  
 375 accuracy. We assume  $x_k, x_{g'_1}, x_{g'_2}, \dots, x_{g'_p}$  are  $p + 1$  real SNVs of the same genome, and  $x_{g'_1}, x_{g'_2}, \dots, x_{g'_p}$   
 376 are the only  $p$  substitutions that can maximize  $\hat{Pr}(X_k = x_k | X_{g_1} = x_{g_1}, X_{g_2} = x_{g_2}, \dots, X_{g_p} = x_{g_p})$ .  
 377 Formally,  $H(x_k) = \hat{Pr}(X_k = x_k | X_{g'_1} = x_{g'_1}, X_{g'_2} = x_{g'_2}, \dots, X_{g'_p} = x_{g'_p})$ , and  $\hat{Pr}(X_k = x_k | X_{g_1} =$   
 378  $x_{g_1}, X_{g_2} = x_{g_2}, \dots, X_{g_p} = x_{g_p}) < \hat{Pr}(X_k = x_k | X_{g'_1} = x_{g'_1}, X_{g'_2} = x_{g'_2}, \dots, X_{g'_p} = x_{g'_p})$  if  $\{g_1, g_2, \dots, g_p\} \neq$   
 379  $\{g'_1, g'_2, \dots, g'_p\}$ . Assuming  $k, g'_1, g'_2, \dots, g'_p$  are the only loci with real SNVs in  $[k - t_l, k + t_r]$ , we define  
 380 signal-to-noise ratio by

$$\rho_0 = Pr( \hat{Pr}(X_k = x_k | X_{g_s} = x_{g_s}) > \max_{x_{g_t}} \{ \hat{Pr}(X_k = x_k | X_{g_t} = x_{g_t}) \} )$$

381 , where  $x_{g_s} \in \{x_{g'_1}, x_{g'_2}, \dots, x_{g'_p}\}$  and  $g_t \notin \{g'_1, g'_2, \dots, g'_p\}$ .  $g_t \notin \{g'_1, g'_2, \dots, g'_p\}$  is equivalent to  $g_t \in$   
 382  $[k - t_l, k + t_r] \cap \{k, g'_1, g'_2, \dots, g'_p\}^c$ . For any locus  $g_s \in \{g'_1, g'_2, \dots, g'_p\}$ , the probability that it is selected  
 383 by the greedy algorithm is denoted as  $Pr(g_s \in \{s_1, s_2, \dots, s_v\})$ , where  $\{s_1, s_2, \dots, s_v\}$  is the  $v$  loci

384 selected by the greedy algorithm. Without loss of generality, assuming  $v \leq p$  and sequencing errors  
 385 are independent,

$$\begin{aligned} Pr(g_s \in \{s_1, s_2, \dots, s_v\}) &\leq Pr(g_s \in \{s_1, s_2, \dots, s_p\}) \\ &= \prod_{g_t \notin \{g'_1, g'_2, \dots, g'_p\}} Pr(\hat{Pr}(X_k = x_k | X_{g_s} = x_{g_s}) > \max_{x_{g_t}} \{\hat{Pr}(X_k = x_k | X_{g_t} = x_{g_t})\}) \\ &= \rho_0^{(t_l + t_r - p)} \end{aligned}$$

386 . The probability that the greedy algorithm correctly estimates  $H(x_k)$  is

$$\begin{aligned} Pr(H(x_k) = \hat{Pr}(X_k = x_k | X_{s_1} = x_{s_1}, X_{s_2} = x_{s_2}, \dots, X_{s_v} = x_{s_v})) &= Pr(\{g'_1, g'_2, \dots, g'_p\} \subseteq \{s_1, s_2, \dots, s_v\}) \\ &\leq Pr(g_s \in \{s_1, s_2, \dots, s_v\}) \\ &= \rho_0^{(t_l + t_r - p)} \end{aligned} \quad (7)$$

387 . According to inequation (7), assuming  $t_l \geq 2000$ ,  $t_r \geq 2000$ , and  $p = 1$ , which is a typical setting  
 388 for long-read sequencing data, the probability that the greedy algorithm correctly estimates  $H(x_k)$  is  
 389 less than  $3.5 \times 10^{-18}$  even if  $\rho_0 = 0.99$ . The key factor leading to the failure of the greedy algorithm  
 390 is selecting from too many loci ( $t_l + t_r$  loci). We propose a novel algorithm called Random Subspace  
 391 Maximization (RSM) to reduce the number of loci to be considered in the next section.

### 392 Improving accuracy of the greedy algorithm by Random Subspace Maximization

393 Firstly, we measure the similarity between two reads,  $R_i$  and  $R_j$ , by a modified Jaccard index<sup>34</sup>, which  
 394 is defined by

$$\text{Jaccard}(R_i, R_j) = \frac{|S_i \cap S_j|}{|(S_i \cup S_j) \cap [4 \max(b_i, b_j), 4 \min(e_i, e_j) + 3]|} \quad (8)$$

395 , where  $\text{Jaccard}(R_i, R_j) = 0$  if the denominator is 0. We require

$$|[\max(b_i, b_j), \min(e_i, e_j)]| \geq l_{\min}$$

396 where  $l_{\min}$  is the minimal length of the overlap region between the two compared reads. We used  
 397  $l_{\min} = 0.5(e_i - b_i)$  in this work. Intuitively, the Jaccard index between two reads is the ratio between  
 398 number of common substitutions shared by the two reads and the total number of substitutions of the  
 399 two reads in their overlapped region. Then, for a read  $R_i$ , we select  $w$  most similar reads according to  
 400 Jaccard index. For each read  $R_j$  in these  $w$  selected reads, we generate a set of substitutions shared

401 by  $R_i$  and  $R_j$ . Formally,

$$C_{ij} = S_i \cap S_j$$

402 .  $C_{ij}$  is called a subspace (Figure 2B), and we can generate  $w \times m$  subspaces if there are  $m$  reads.

403 We used  $w = 100$  in this work. For a substitution  $X_k \in C_{ij}$ , we estimate its maximal conditional

404 probability of  $\{X_k = x_k\}$  in subspace  $C_{ij}$ , which is defined by

$$\begin{aligned} H_{C_{ij}}(x_k) &= \max_{\{x_{g_1}, x_{g_2}, \dots, x_{g_p}\} \subseteq C_{ij}} \{ \hat{Pr}(X_k = x_k | X_{g_1} = x_{g_1}, X_{g_2} = x_{g_2}, \dots, X_{g_p} = x_{g_p}) \} \\ &= \max_{x_{g_1}, x_{g_2}, \dots, x_{g_p}} \left\{ \frac{|\{t \mid \{x_k, x_{g_1}, x_{g_2}, \dots, x_{g_p}\} \subseteq (S_t \cap C_{ij})\}|}{|\{t \mid \{x_{g_1}, x_{g_2}, \dots, x_{g_p}\} \subseteq (S_t \cap C_{ij}), k \in [b_t, e_t]\}|} \right\} \end{aligned} \quad (9)$$

405 , using the greedy algorithm described in the previous section by only considering the substitutions

406 in  $C_{ij}$ . Thus, compared to the original greedy algorithm, the number of loci to be considered is

407 substantially reduced. We then use

$$\hat{H}(x_k) = \max_{C_{ij}} (\hat{H}_{C_{ij}}(x_k)) \quad (10)$$

408 to estimate the maximal conditional probability of  $\{X_k = x_k\}$  defined by equation (6).  $\hat{H}_{C_{ij}}(x_k)$  is the

409 maximal conditional probability of  $\{X_k = x_k\}$  in subspace  $C_{ij}$  estimated by the greedy algorithm. The

410 whole procedure of estimating  $H(x_k)$  in the  $w \times m$  subspaces is called Random Subspace Maximization

411 (RSM) (Figure 2C).

## 412 Theoretical accuracy of RSM algorithm

413 Without loss of generality, we denote  $\{x'_{g_1}, x'_{g_2}, \dots, x'_{g_p}\}$  as the only set of substitutions that maxi-

414 mizes  $\hat{Pr}(X_k = x_k | X_{g_1} = x_{g_1}, X_{g_2} = x_{g_2}, \dots, X_{g_p} = x_{g_p})$ , and  $\Omega$  as the set of subspaces contain-

415 ing  $\{x_k, x'_{g_1}, x'_{g_2}, \dots, x'_{g_p}\}$ . For a subspace  $C_t \in \Omega$ , the probability that the greedy algorithm finds

416  $\{x'_{g_1}, x'_{g_2}, \dots, x'_{g_p}\}$  is denoted as  $Pr(\hat{H}_{C_t}(x_k) = H(x_k))$ , where  $H_k$  is defined by equation (6). The

417 probability that RSM algorithm finds  $\{x'_{g_1}, x'_{g_2}, \dots, x'_{g_p}\}$  is

$$\begin{aligned} Pr(\hat{H}(x_k) = H(x_k)) &= Pr(\cup_{C_t \in \Omega} \{\hat{H}_{C_t}(x_k) = H(x_k)\}) \\ &= 1 - Pr(\cap_{C_t \in \Omega} \{\hat{H}_{C_t}(x_k) \neq H(x_k)\}) \end{aligned} \quad (11)$$

418 . Assuming  $Pr(\hat{H}_{C_t}(x_k) = H(x_k)) > 0$ , and according to chain rule of joint probability,

$$\begin{aligned} &Pr(\cap_{C_t \in \Omega} \{\hat{H}_{C_t}(x_k) \neq H(x_k)\}) \\ &= Pr(\hat{H}_{C_1}(x_k) \neq H(x_k)) \prod_{t=2}^{|\Omega|} Pr(\hat{H}_{C_t}(x_k) \neq H(x_k) | \hat{H}_{C_{t-1}}(x_k) \neq H(x_k), \dots, \hat{H}_{C_1}(x_k) \neq H(x_k)) \end{aligned}$$

419 , where  $Pr(\hat{H}_{C_t}(x_k) \neq H(x_k) | \hat{H}_{C_{t-1}}(x_k) \neq H(x_k), \dots, \hat{H}_{C_1}(x_k) \neq H(x_k)) < 1$  if  $C_t \notin \{C_{t-1}, C_{t-2}, \dots, C_1\}$ .  
420 As sequencing depth increases,  $|\Omega|$  increases, and  $Pr(\cap_{C_t \in \Omega} \{\hat{H}_{C_t}(x_k) \neq H(x_k)\})$  converges to 0. Thus,  
421  $Pr(\hat{H}(x_k) = H(x_k))$  (equation (11)) converges to 1 as sequencing depth increases. Intuitively, with  
422 infinite sequencing depth, RSM algorithm is guaranteed to detect real SNVs correctly if these SNVs  
423 have larger maximal conditional probabilities than sequencing errors.

#### 424 **Detecting orphan SNVs by correcting sequence context effect**

425 As RSM algorithm requires multiple real SNVs, it can not detect orphan SNVs. An orphan SNV is  
426 the only SNV of the genome. We have to rely on the single-locus algorithm described in equation (4)  
427 to detect orphan SNVs. However, the substitution rate of a locus is not only affected by real SNVs  
428 but also affected by the sequence context of the locus. We built a gradient boosting<sup>31</sup> model to learn  
429 the sequence context effect and corrected it by the following likelihood ratio method (Figure 2D). For  
430 a substitution  $x_k$  at locus  $k$ , its likelihood ratio is

$$LR(x_k) = \frac{\text{Binomial}(t_k; n_k, p_1)}{\text{Binomial}(t_k; n_k, p_0)} \quad (12)$$

431 , where  $\text{Binomial}(x; n, p)$  is the probability mass function of binomial distribution with parameters  $n$   
432 and  $p$ , and

$$t_k = |\{i \mid x_k \in S_i\}|$$

$$n_k = |\{i \mid k \in [b_i, e_i]\}|$$

$$p_1 = \frac{t_k}{n_k}$$

$$p_0 = \text{Predicted sequencing error rate by sequence context}$$

433 . The substitution  $x_k$  is detected as a SNV if  $LR(x_k)$  is larger than a threshold. We used a threshold  
434 of 50 in this work. Calculation of  $p_0$  is introduced in the next section. To reduce false discovery rate,  
435 we also required a detected SNV has a substitution rate higher than 0.1 for PacBio data and 0.2 for  
436 ONT data respectively.

#### 437 **Modeling sequence context effect on sequencing error rate**

438 Error rate of long-read sequencing is strongly affected by sequence context (Figure 3). For locus  $i$ ,  
439 we define its one upstream homopolymer and one downstream homopolymer as its sequence context

440 (Figure S8). We adopted the gradient boosting model implemented by xgboost (version 0.90)<sup>31</sup> to  
441 predict substitution rate of each locus by its sequence context. For PacBio, we trained the model  
442 on a dataset consisting of 79 PacBio RS II runs with P6-C4 chemistry and a dataset consisting of  
443 24 PacBio RS II runs with P4-C2 chemistry respectively (SRA IDs of the data are listed in Table  
444 S4). As the sequence context effects on these two datasets are highly similar, we only used the model  
445 trained on the P6-C4 data for the analysis. For ONT, we trained the model on a dataset consisting  
446 of 8 MinION runs with R9.4 chemistry (SRA IDs of the data are listed in Table S4). We tuned three  
447 parameters in gradient boosting, step size (eta in xgboost), number of trees (num\_round in xgboost)  
448 and maximal depth of trees (max\_depth in xgboost) and used the parameters with the highest five-fold  
449 cross-validation accuracy (Table S5). We used  $R^2$  as the measurement of accuracy, which is defined by

$$R^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

450 where  $y_i$  is the substitution rate of a sequence context,  $\hat{y}_i$  is the predicted substitution rate,  $\bar{y}$  is the  
451 average substitution rate, and  $n$  is the number of unique sequence contexts. For PacBio, step size,  
452 number of trees and maximal depth of trees with the highest accuracy are 0.01, 2000 and 10 respectively.  
453 For ONT, step size, number of trees and maximal depth of trees with the highest accuracy are 0.1,  
454 2000 and 10 respectively.

455 We also masked bases with QV thresholds 8, 10 and 12, and trained three different models on the  
456 masked data. Each model is used in the detection algorithm which masks bases with the same QV  
457 threshold. In the case of not masking any base, we predicted substitution rate using the trained model  
458 on the three pooled sequencing datasets (Figure 3). The results show that substitution-error rate is  
459 strongly affected by sequence context and can be well predicted by our model.

## 460 **Phasing minor SNVs**

461 To detect whether multiple minor SNVs are from the same DNA molecule, we proposed a novel al-  
462 gorithm called Adaptive Nearest-Neighbors clustering (ANN). As the reads inevitably have errors, an  
463 intuitive way to phase minor SNVs is to cluster the reads and use the consensus sequences of each clus-  
464 ter to phase the minor SNVs. However, an intrinsic difficulty of clustering algorithms is to determine  
465 the number of clusters, which is unknown. The ANN algorithm can directly estimate the number of  
466 clusters from data.



## 467 Adaptive-Nearest-Neighbors clustering

468 Firstly, to reduce noise level, we only retain detected SNVs for each read. Formally, for read  $R_i$   
469 (equation (2)), we use

$$\tilde{S}_i = S_i \cap \{\text{Detected SNVs}\} \quad (13)$$

470 , where  $S_i$  is defined in equation (3).

471 The intuitive idea of ANN algorithm is that all loci should be homogeneous by piling up the reads  
472 in each cluster (Figure S9). A locus is homogeneous if it satisfies the following condition. For locus  $k$ ,  
473 its substitution rate satisfies

$$\hat{P}r(X_k = x_k) = \frac{|\{i \mid x_k \in \tilde{S}_i\}|}{\sum_{d=0}^3 |\{i \mid 4k + d \in \tilde{S}_i\}|} \in [0, p_{lim}] \cup [p_{lim}, 1] \quad (14)$$

474 , where  $x_k \in \{4k, 4k + 1, 4k + 2, 4k + 3\}$ . In this work, We set  $p_{lim} = 0.2$  for the PacBio data and  
475  $p_{lim} = 0.3$  for the ONT data. For a read  $i$  (called seed read), we sorted its  $q$  most similar reads  
476 according to Jaccard index (equation (8)), and kept discarding the most dissimilar one until all loci  
477 covered by the seed read are homogeneous or maximal coverage of the loci is smaller than a threshold  
478 (10 in this work) (Figure 2F). We recorded the consensus sequence as a draft contig if all the loci are  
479 homogeneous (Figure S9). We calculated the Jaccard index of each read with all the draft contigs,  
480 and assigned the read to the contig with the largest Jaccard index. A read is assigned to the reference  
481 genome if its largest Jaccard index is smaller than 0.5. The abundance of a contig is defined as the  
482 number of reads assigned to it.

483 A problem of the algorithm described above is that the alignment is affected by reference bias  
484 and homogeneous loci could be mistaken for heterogeneous loci. Reference bias is the phenomenon  
485 that the substitution rate of a real SNV at a homogeneous locus is significantly lower than 1 –  
486 substitution error rate (Figure S10A).

## 487 Reference bias and local realignment

488 For each detected SNV, we adopted standard Smith-Waterman algorithm implemented by SeqAn (ver-  
489 sion 2.4) (<https://www.seqan.de>) to realign reads to four modified reference sequences with A,C,G,  
490 or T at each locus with a detected SNV. The scores of match, mismatch, gap open, and gap extension  
491 are 2, -4, -4, and -2 respectively, and the score of a base aligned to base N or a masked low-QV base is 0.  
492 To avoid high computational cost, we only realigned 21 homopolymers whose center is the locus with

493 detected SNV. For each read, the modified base in the reference sequence with the highest alignment  
494 score is recorded as a substitution of the read (Figure 2E and Figure S11). We tested the realignment  
495 method on a single *Escherichia coli* dataset (SRA ID is ERS718594), which is presumably homoge-  
496 neous. The result shows that local realignment can substantially reduce reference bias (Figure S10B).  
497 The average substitution rate of loci with real SNVs is 84.8% before realignment, and the average  
498 substitution rate of loci with real SNV is 95.9% after realignment. We performed local realignment  
499 before ANN algorithm in our analysis.

### 500 **Filtering draft contigs**

501 To reduce false positive rate of the inferred draft contigs by ANN algorithm, we adopted a two-step  
502 algorithm to filter the draft contigs (Figure 2G). In the first step, we tested whether the frequency of  
503 each individual SNV in each contig is significantly higher than the sequencing error rate and whether  
504 SNVs in each contig are independent using Bayes factor. The contig is filtered if the frequency of  
505 any of its SNVs is not significant and its SNVs are independent (Figure S12A). In the second step,  
506 we compared the contigs pairwise, and the contig with lower abundance in each pair is filtered if the  
507 contigs are not significantly different according to Bayes factor (Figure S12B).

### 508 **Assembling draft contigs**

509 The length of the draft contigs obtained by ANN algorithm is usually smaller than genome size, except  
510 in a few cases like a virus genome. Therefore, we have to assemble the draft contigs to obtain the  
511 whole picture of the underlined genomes in the sequenced sample. We borrowed the idea of overlap  
512 graph<sup>32</sup> from *de novo* genome assembly to assemble the draft contigs. We denoted each draft contig  
513 as a vertex in a graph and compared the contigs pairwise. For a draft contig  $i$ , we linked it to another  
514 draft contig  $j$  by adding a edge from vertex  $i$  to vertex  $j$  if all the three criteria are met: 1) the two  
515 draft contigs are identical in their overlapped region; 2) the number of overlapped SNVs is more than  
516 50% of the number of SNVs in contig  $i$  or that in contig  $j$ , or the length of overlapped region is more  
517 than 50% of the length of contig  $i$  or that of contig  $j$ ; 3) the genome coordinate of the end locus of  
518 contig  $i$  is smaller than that of contig  $j$ . We then removed redundant edges by transitive reduction<sup>41</sup>  
519 (Figure S13A and Figure S13B). A contig is constructed by concatenating draft contigs which are in  
520 an unambiguous path. A path is an unambiguous path if the three criteria are met: 1) in-degree of  
521 the start vertex is not 1; 2) out-degree of the end vertex is not 1 or a daughter vertex of the end vertex

522 has more than one parental vertices; 3) in-degrees and out-degrees of the vertices other than the start  
523 vertex and the end vertex are 1 (Figure 2H and Figure S13C). We then filtered the contigs using the  
524 two-step filter introduced in the previous section. We calculated the Jaccard index of each read to all  
525 the contigs, and assigned the read to the contig with the largest Jaccard index. A read is assigned to  
526 the reference genome if its largest Jaccard index is smaller than 0.5.

## 527 **Detecting bacterial methylation motifs from ONT data without raw signal**

528 As the raw-signal files of ONT data are usually huge and not publicly available, we developed an  
529 algorithm to detect DNA methylation motifs without raw signal. For each individual ONT data file  
530 before pooling, we extracted the flanking sequences (40 bp long) of loci whose substitution rates are  
531 greater than 0.15, and detected motifs in the flanking sequences using the motif caller developed  
532 by PacBio (<https://github.com/PacificBiosciences/MotifMaker>)<sup>42</sup>. We only retained the motifs  
533 that matches the known bacterial methylation motifs in REBASE ([http://rebase.neb.com/rebase/rebase\\_methylase\\_recseqs.txt](http://rebase.neb.com/rebase/rebase_methylase_recseqs.txt))<sup>43</sup>. Thus, our methylation-motif detection algorithm is conservative  
534 and only detects known motifs. We only discovered two known motifs, CCWGG and CGCATC, on  
535 the ONT data. W represents A or T.

## 537 ***Borrelia* MLST**

538 We downloaded the allele sequences of the eight house-keeping genes from [https://pubmlst.org/bigssdb?db=pubmlst\\_borrelia\\_seqdef&page=downloadAlleles](https://pubmlst.org/bigssdb?db=pubmlst_borrelia_seqdef&page=downloadAlleles), and aligned them to the iGDA-inferred  
539 contigs and the genome sequence of each *Borrelia burgdorferi* strain using MUMmer (version 3)<sup>36</sup>. If  
540 a contig or genome sequence has no 100% match in the allele database, we reported the allele with the  
541 highest percent identity in the MUMmer output.

## 543 **Evaluating the minimal divergence that two conspecific strains can be distinguished**

544 We only retained the iGDA-reported contigs that is 100% identical to a true genome sequence and  
545 only has an unique closest true genome sequence. These retained contigs can be used to distinguish  
546 conspecific strains. We calculated the divergence between two contigs by

$$\text{Divergence}(\text{contig 1, contig 2}) = \frac{\text{number of different SNVs}}{\text{length of overlapped region}}$$

## 547 **Parameter setting in the third-party tools**

### 548 **flye**

549 In the PacBio metagenomic data, we used "flye -t 16 -pacbio-raw -g 2m".

### 550 **MUMmer**

551 We used "nucmer -c 150 -g 500 -l 12 -maxmatch" for alignment, and "show-snps -l -T -H" to obtain  
552 SNVs. To avoid the impact of repeats we used "mummerplot --filter" before "show-snps -l -T -H"  
553 for the metagenomic data.

## 554 **Software access**

555 iGDA is available at Anaconda Cloud <https://anaconda.org/zhixingfeng/igda>. Install Conda  
556 (<https://docs.conda.io/projects/conda/en/latest/user-guide/install/>) and type "conda in-  
557 stall -c zhixingfeng igda" to install iGDA and its dependencies. After installation, type "igda" for  
558 usage.

## 559 **Acknowledgement**

560 The project was supported by funds from the Steven & Alexandra Cohen Foundation.

## 561 **Author contributions**

562 Z.F. designed and implemented the computational models and algorithms of iGDA. B.W. proposed  
563 and tested the methods to improve speed of iGDA. Z.F. performed the data analysis with support  
564 from B.W.. Z.F. designed the experiments to evaluate iGDA on metagenomic data with the support  
565 from J.C. and E.E.S.. Z.F. wrote the manuscript with input from all authors.

## 566 **Competing interests**

567 E.E.S. is on the scientific advisory board of Pacific Biosciences.

## 568 References

- 569 1. Mantere, T., Kersten, S. & Hoischen, A. Long-read sequencing emerging in medical genetics.  
570 *Frontiers in Genetics* **10**, 1–14 (2019).
- 571 2. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis.  
572 *Genome Biology* **21**, 1–16 (2020).
- 573 3. Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing  
574 of mock microbial community standards. *GigaScience* **8**, 1–9 (2019).
- 575 4. Kingan, S. B. *et al.* A high-quality genome assembly from a single, field-collected spotted lanternfly  
576 (*Lycorma delicatula*) using the PacBio Sequel II system. *GigaScience* **8**, 1–10 (2019).
- 577 5. Bansal, V. A statistical method for the detection of variants from next-generation resequencing  
578 of DNA pools. *Bioinformatics* **26**, 318–324 (2010).
- 579 6. Wei, Z., Wang, W., Hu, P., Lyon, G. J. & Hakonarson, H. SNVer: A statistical tool for variant  
580 calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research*  
581 **39**, 1–13 (2011).
- 582 7. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation  
583 DNA sequencing data. *Nature Genetics* **43**, 491–501 (2011).
- 584 8. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer  
585 by exome sequencing. *Genome Research* **22**, 568–576 (2012).
- 586 9. Saunders, C. T. *et al.* Strelka: Accurate somatic small-variant calling from sequenced tumor-  
587 normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
- 588 10. Larson, D. E. *et al.* Somaticsniper: Identification of somatic point mutations in whole genome  
589 sequencing data. *Bioinformatics* **28**, 311–317 (2012).
- 590 11. Wilm, A. *et al.* LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering  
591 cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*  
592 **40**, 11189–11201 (2012).
- 593 12. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous  
594 cancer samples. *Nature Biotechnology* **31**, 213–219 (2013).
- 595 13. Shiraishi, Y. *et al.* An empirical Bayesian framework for somatic mutation detection from cancer  
596 genome sequencing data. *Nucleic Acids Research* **41**, e89 (2013).

- 597 14. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling  
598 variants in clinical sequencing applications. *Nature Genetics* **46**, 912–918 (2014).
- 599 15. Macalalad, A. R. *et al.* Highly sensitive and specific detection of rare variants in mixed viral  
600 populations from massively parallel sequence data. *PLoS Computational Biology* **8**, e1002417  
601 (2012).
- 602 16. Yang, X., Charlebois, P., Macalalad, A., Henn, M. R. & Zody, M. C. V-Phaser 2: Variant inference  
603 for viral populations. *BMC Genomics* **14**, 674 (2013).
- 604 17. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nature*  
605 *Methods* **14**, 407–410 (2017).
- 606 18. Luo, R., Sedlazeck, F. J., Lam, T. W. & Schatz, M. C. A multi-task convolutional deep neu-  
607 ral network for variant calling in single molecule sequencing. *Nature Communications* **10**, 1–11  
608 (2019).
- 609 19. Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient  
610 template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*  
611 **38**, e159 (2010).
- 612 20. Guo, F., Wang, D. & Wang, L. Progressive approach for SNP calling and haplotype assembly  
613 using single molecular sequencing data. *Bioinformatics* **34**, 2012–2018 (2018).
- 614 21. Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-  
615 molecule long read sequencing. *Nature Communications* **10**, 4660 (2019).
- 616 22. Ebler, J., Haukness, M., Pesout, T., Marschall, T. & Paten, B. Haplotype-aware diplotyping from  
617 noisy long reads. *Genome Biology* **20**, 1–16 (2019).
- 618 23. Zagordi, O., Bhattacharya, A., Eriksson, N. & Beerenwinkel, N. ShoRAH: Estimating the genetic  
619 diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* **12**, 119  
620 (2011).
- 621 24. Prospero, M. C. F. & Salemi, M. QuRe: Software for viral quasispecies reconstruction from next-  
622 generation sequencing data. *Bioinformatics* **28**, 132–133 (2012).
- 623 25. Töpfer, A. *et al.* Probabilistic inference of viral quasispecies subject to recombination. *Journal of*  
624 *Computational Biology* **20**, 113–123 (2013).

- 625 26. Giallonardo, F. D. *et al.* Full-length haplotype reconstruction to infer the structure of heteroge-  
626 neous virus populations. *Nucleic Acids Research* **42**, e115 (2014).
- 627 27. Prabhakaran, S., Rey, M., Zagordi, O., Beerenwinkel, N. & Roth, V. HIV haplotype inference  
628 using a propagating dirichlet process mixture model. *IEEE/ACM Transactions on Computational*  
629 *Biology and Bioinformatics* **11**, 182–191 (2014).
- 630 28. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotech-*  
631 *nology* **33**, 1045–1052 (2015).
- 632 29. Quince, C. *et al.* DESMAN: A new tool for de novo extraction of strains from metagenomes.  
633 *Genome Biology* **18**, 1–22 (2017).
- 634 30. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).
- 635 31. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. *Proceedings of the ACM*  
636 *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
- 637 32. Myers, E. W. Toward Simplifying and Accurately Formulating Fragment Assembly. *Journal of*  
638 *Computational Biology* **2**, 275–290 (1995).
- 639 33. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100  
640 (2018).
- 641 34. Jaccard, P. the Distribution of the Flora in the Alpine Zone. *New Phytologist* **11**, 37–50 (1912).
- 642 35. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor.  
643 *Bioinformatics* **34**, i884–i890 (2018).
- 644 36. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome biology* **5**, R12  
645 (2004).
- 646 37. Biesiada, G., Czepiel, J., Leśniak, M. R., Garlicki, A. & Mach, T. Lyme disease: Review. *Archives*  
647 *of Medical Science* **8**, 978–982 (2012).
- 648 38. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using  
649 repeat graphs. *eng. Nature biotechnology* **37**, 540–546 (2019).
- 650 39. Margos, G. *et al.* MLST of housekeeping genes captures geographic population structure and sug-  
651 gests a European origin of *Borrelia burgdorferi*. *Proceedings of the National Academy of Sciences*  
652 *of the United States of America* **105**, 8730–8735 (2008).

- 653 40. Mostovoy, Y. *et al.* A hybrid approach for de novo human genome sequence assembly and phasing.  
654 *Nature Methods* **13**, 587–590 (2016).
- 655 41. Myers, E. W. The fragment assembly string graph. *Bioinformatics* **21**, 79–85 (2005).
- 656 42. Clark, T. A. *et al.* Characterization of DNA methyltransferase specificities using single-molecule,  
657 real-time DNA sequencing. *Nucleic Acids Research* **40**, e29 (2012).
- 658 43. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and  
659 modification: Enzymes, genes and genomes. *Nucleic Acids Research* **43**, D298–D299 (2015).