

Information theoretic alignment free variant calling

Justin Bedo, Benjamin Goudey, Jeremy Wazny, Zeyu Zhou

While traditional methods for calling variants across whole genome sequence data rely on alignment to an appropriate reference sequence, alternative techniques are needed when a suitable reference does not exist. We present a novel alignment and assembly free variant calling method based on information theoretic principles designed to detect variants have strong statistical evidence for their ability to segregate samples in a given dataset. Our method uses the context surrounding a particular nucleotide to define variants. Given a set of reads, we model the probability of observing a given nucleotide conditioned on the surrounding prefix and suffixes of length k as a multinomial distribution. We then estimate which of these contexts are stable intra-sample and varying inter-sample using a statistic based on the Kullback-Leibler divergence.

The utility of the variant calling method was evaluated through analysis of a pair of bacterial datasets and a mouse dataset. We found that our variants are highly informative for supervised learning tasks with performance similar to standard reference based calls and another reference free method (DiscoSNP++). Comparisons against reference based calls showed our method was able to capture very similar population structure on the bacterial dataset. The algorithm's focus on discriminatory variants makes it suitable for many common analysis tasks for organisms that are too diverse to be mapped back to a single reference sequence.

Information theoretic alignment free variant calling

Justin Bedř^{*1,2}, Benjamin Goudey^{1,3}, Jeremy Wazny¹, and Zeyu Zhou^{1,4}

¹IBM Research – Australia, Level 5, 204 Lygon St Carlton, VIC 3053 Australia

²Dept. of Computing and Information Systems, University of Melbourne, Parkville, VIC 3010 Australia

³Centre For Epidemiology and Biostatistics, University of Melbourne, Parkville, VIC 3010 Australia

⁴School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010 Australia

May 2, 2016

Abstract

While traditional methods for calling variants across whole genome sequence data rely on alignment to an appropriate reference sequence, alternative techniques are needed when a suitable reference does not exist. We present a novel alignment and assembly free variant calling method based on information theoretic principles designed to detect variants have strong statistical evidence for their ability to segregate samples in a given dataset. Our method uses the *context* surrounding a particular nucleotide to define variants. Given a set of reads, we model the probability of observing a given nucleotide conditioned on the surrounding prefix and suffixes of length k as a multinomial distribution. We then estimate which of these contexts are stable intra-sample and varying inter-sample using a statistic based on the Kullback–Leibler divergence.

The utility of the variant calling method was evaluated through analysis of a pair of bacterial datasets and a mouse dataset. We found that our variants are highly informative for supervised learning tasks with performance similar to standard reference based calls and another reference free method (DiscoSNP++). Comparisons against reference based calls showed our method was able to capture very similar population structure on the bacterial dataset. The algorithm's focus on discriminatory variants makes it suitable for many common analysis tasks for organisms that are too diverse to be mapped back to a single reference sequence.

1 Introduction

Many sequencing studies begin by the transformation of raw sequence data to relatively few features, usually single-nucleotide variants. Typically, this is done by aligning the individual sequence reads to a reference genome to identify single nucleotide differences from the reference.

Although straightforward, the genome alignment approach has several shortcomings:

- a suitable reference may not exist; this is especially important for unstable genomes such the anuploid genomes frequently encountered in cancer (Beroukhim, Mermel, Porter, et al., 2010), and also for some organisms with large genetic diversity such as bacteria (Ochman, Lawrence, and Groisman, 2000);

*corresponding author

- 36 • selecting a reference may be difficult when there is uncertainty about what has been
37 sampled; and
- 38 • it performs poorly when a sample contains significant novel material, i. e., sequences that
39 are not simple variations of the reference.

40 Existing reference-free approaches are either based on assembly (Li, 2012), which possibly
41 introduces misassembly biases, or on searching for structural motifs within a universal de Bruijn
42 graph of all samples (Peterlongo, Schnel, Pisanti, et al., 2010; Iqbal, Caccamo, Turner, et al.,
43 2012; Uricaru, Rizk, Lacroix, et al., 2015) that correspond to simple variants.

44 We present a variant calling algorithm to generate features from unaligned raw reads. Rather
45 than attempting to identify all genetic variation within a given set of samples, we instead focus on
46 selected variants that have strong statistical evidence for their ability to segregate samples in
47 a given dataset. Such variants form useful features for many tasks including genomic prediction
48 of a given phenotype, modelling population structure or clustering samples into related groups.

49 Our method uses the *context* surrounding a particular nucleotide to define variants. Given
50 a set of reads, we model the probability of observing a given nucleotide conditioned on the
51 surrounding prefix and suffix nucleotide sequences of length k as a multinomial distribution. We
52 then estimate which of these contexts form potential variants, i. e., those that are stable intra-
53 sample and varying inter-sample, using a statistic based on the Kullback–Leibler divergence.
54 Given this list of candidate variants, we call those variants by maximum likelihood of our
55 multinomial model.

56 Furthermore, we show that the size of the context k can be chosen using the minimum
57 message length principle (Wallace and Boulton, 1968) and that our context selection statistic is
58 γ -distributed. Consequently, k can be determined from the data and the contexts surrounding
59 variants can be selected with statistical guarantees on type-1 errors.

60 The utility of variant calling method was evaluated through simulation experiments and
61 empirical analysis of a pair of bacterial datasets and a mouse dataset. Through simulations we
62 showed the method has good power and false positive rate for detecting variants, though the
63 ability to detect rare variants required high depth and large number of samples.

64 Our empirical results indicated our variants are highly informative for antimicrobial resistance
65 phenotypes on the bacterial datasets and were able to accurately capture population structure.
66 On the mouse dataset, the variants were also found to be good for modelling coat colour.
67 Further investigations of the variants found for the bacterial dataset using a known reference
68 sequence revealed variants associated with *boxB* repeat regions, a repeat previously used for
69 population structure mapping (Rakov, Ubukata, and Robinson, 2011), suggesting the model can
70 generate features for more complex genetic elements. These results suggest the variants are
71 capturing genotypic variation well and can model heritable traits in different organisms. Our
72 proposed method will be of strongest utility when modelling of population structure, phylogenetic
73 relationships or phenotypes from genotype for large scale datasets of organisms with either
74 variable genomes (as is the case for many bacteria), or those lacking a reference genome.

75 2 Methods

76 Our variant calling method comprises two steps: modelling the probability that a base is observed
77 in a sample given the surrounding context; and determining which contexts surround variable
78 bases in a population represented by several samples. The former provides a mechanism to
79 call variants in a sample given a set of contexts, and the latter determines the set of contexts
80 associated with variants.

81 2.1 Variant calling

82 We consider the case of variant calling directly from a collection of reads. Let random variable
83 x_{ij} taking values in $\{A, C, G, T\}$ denote the j^{th} nucleotide of the i^{th} read, with $1 \leq i \leq n$ and
84 $1 \leq j \leq m_i$ the number of reads and nucleotides in the read i .

85 **Definition 1 (k -context)** *The k -context around a nucleotide j consists of a k -prefix sequence*

$$\pi_k(x_i, j) := [x_{i(j-k)}, x_{i(j-k+1)}, \dots, x_{i(j-1)}].$$

86 *and a k -suffix sequence*

$$\sigma_k(x_i, j) := [x_{i(j+1)}, x_{i(j+2)}, \dots, x_{i(j+k)}].$$

87 *Contexts that consist of only the prefix/suffix sequences are suffix/prefix-free.*

88 **Definition 2 (k -context probability)** *The k -context probability is the probability of observing a*
89 *base at a particular position given the context, that is*

$$P(x_{ij} | \pi_k(x_i, j), \sigma_k(x_i, j)).$$

90 The k -context probabilities can be estimated from the data by maximising a pseudolikelihood.
91 Let $f(b, \pi_k, \sigma_k) := 1 + \sum_{ij} \mathbb{I}[x_{ij} = b \wedge \pi_k = \pi_k(x_i, j) \wedge \sigma_k = \sigma_k(x_i, j)]$ denote the counts of how
92 often b was observed with k -prefix π_k and k -suffix σ_k in the read set x , where $\mathbb{I}[\cdot]$ is the Iverson
93 bracket. Here the pseudocount encodes a weak uniform prior. The probability density estimate
94 of observing a base b in context (π_k, σ_k) is then given by

$$\hat{P}(b | \pi_k, \sigma_k) := \frac{f(b, \pi_k, \sigma_k)}{\sum_{b'} f(b', \pi_k, \sigma_k)}.$$

95 The suffix/prefix free densities are thus

$$\hat{P}(b | \pi_k) = \sum_{\sigma_k} \hat{P}(b | \pi_k, \sigma_k) \text{ and } \hat{P}(b | \sigma_k) = \sum_{\pi_k} \hat{P}(b | \pi_k, \sigma_k).$$

96 Given a context (π_k, σ_k) , the base can be called as $\arg \max_b \hat{P}(b | \pi_k, \sigma_k)$, and similarly for
97 prefix/suffix free densities.

98 2.2 Variant finding

99 Determining the list of variants consists of determining which contexts (π_k, σ_k) surround a
100 variable base in our population, then call the base for each variant-defining context and each
101 sample. We consider inter-sample variants and not intra-sample variants; we are interested in
102 finding contexts which define variants that differ amongst samples and are not attributable to
103 noise. In this section, we develop a statistic based on the Kullback–Leibler (KL) divergence that
104 achieves these two points.

105 Let \mathcal{X} be a set of samples, each consisting of a collection of reads as defined above. For
106 each $x \in \mathcal{X}$, we refer to the j^{th} nucleotide of the i^{th} read as x_{ij} , the number of reads in the sample
107 as n_x , and the number of nucleotides in read x_i as m_{x_i} . Similarly to the previous section, we
108 denote $f_x(b, \pi_k, \sigma_k)$ as the frequency of observing base b given context (π_k, σ_k) for sample x . As
109 before, a pseudocount is used when estimating f_x to encode a uniform prior.

110 The KL divergence measure provides a way of quantifying the differences between two
111 probability distributions. We will develop a statistic based upon the KL-divergence that compares
112 the individual sample distributions of nucleotide occurrence for a given context with a global
113 expected distribution. Contexts that significantly diverge from the global expected distribution
114 surround a site which is variant in the population sample.

115 **Definition 3 (Kullback–Leibler divergence)** Let P and Q be two discrete probability densities
 116 over the domain \mathcal{Y} . The Kullback–Leibler (KL) divergence is

$$P(\cdot) \parallel_{kl} Q(\cdot) := \sum_{y \in \mathcal{Y}} P(y) \log \frac{P(y)}{Q(y)}.$$

117 **Definition 4 (Total divergence)** The total divergence for a given context (π_k, σ_k) is estimated
 118 as the total KL divergence between the samples in the dataset \mathcal{X} and the expected probability
 119 distribution given the context:

$$D_{\mathcal{X}}(\pi_k, \sigma_k) := \sum_{x \in \mathcal{X}} \hat{P}_x(\cdot | \pi_k, \sigma_k) \parallel_{kl} Q(\cdot | \pi_k, \sigma_k),$$

120 where

$$\hat{P}_x(\cdot | \pi_k, \sigma_k) := \frac{f_x(b, \pi_k, \sigma_k)}{\sum_{b'} f_x(b', \pi_k, \sigma_k)}.$$

121 denotes the probability density estimated for sample x and context (π_k, σ_k) and

$$Q(b | \pi_k, \sigma_k) := \frac{\sum_{x \in \mathcal{X}} f_x(b, \pi_k, \sigma_k)}{\sum_{x \in \mathcal{X}, b'} f_x(b', \pi_k, \sigma_k)}.$$

122 The total divergence statistic is proportional to the expected KL-divergence between a sample
 123 and the global expected probability distribution. To see why this statistic is robust to noise
 124 consider the case where variation is due purely to noise. As the noise distribution is independent
 125 of sample, it will be well modelled by the expected distribution Q and therefore the divergence
 126 between each sample and Q will be small. Conversely, if variation is due to samples being drawn
 127 from two or more latent probability densities, then Q will be an average of these latent densities
 128 and divergence will be high.

129 The next theorem is crucial for determining when a particular divergence estimate indicates
 130 a significant divergence from the expected distribution Q . Using this theorem, we can use
 131 hypothesis testing to select which contexts are not well explained by Q . These contexts not well
 132 explained by Q are *variant* and we call them as in section 2.1.

133 **Theorem 5** Under random sampling from Q , D follows a γ distribution.

134 The proof of this theorem is trivial given a well known result regarding the G-test (see Sokal
 135 and Rohlf (1994)):

136 **Lemma 6** Let f_x be a frequency function and $g := E[f_x]$. The G-test is

$$G := \sum_{x \in \mathcal{X}} \sum_{b \in \{A, T, C, G\}} f_x(b, \pi_k, \sigma_k) \log \left(\frac{f_x(b, \pi_k, \sigma_k)}{g(b, \pi_k, \sigma_k)} \right).$$

137 Under the null hypothesis that f_x results from random sampling from a distribution with expected
 138 frequencies g , G follows a χ^2 distribution with $3|\mathcal{X}|$ degrees of freedom asymptotically.

139 From this lemma, the proof of theorem 5 follows easily:

140 **Proof** D is proportional to the G-test. As the G-test is χ^2 -distributed, D is γ -distributed. ■

141

142 Clearly our statistic D is very similar to G , but has an important property: D is invariant to
 143 coverage. As D operates on estimates of the probability rather than the raw counts, changes
 144 in coverage are effectively normalised out. This is advantageous for variant discovery as it
 145 avoids coverage bias and allows variants to be called for (proportionally) low-coverage areas, if
 146 statistical support for their variability in the population exists.

147 To select contexts a γ distribution is fitted to the data. For the results in our experiments, we
 148 used a Bayesian mixture model with a β prior over the mixing weights whereby each context
 149 could originate from the null (γ) distribution or from a uniform distribution. The mixing weights
 150 were then used to determine if a context is not well supported by the null distribution. Such a
 151 model comparison procedure has several advantages and directly estimates the probabilities of
 152 support by the data for each context (Kamary, Mengersen, Robert, et al., 2014), providing an
 153 easily interpretable quantity.

154 2.3 Choosing context size

155 The problem of choosing context size k is difficult; if too large then common structures will not
 156 be discovered, and if too small then base calling will be unreliable. We propose to choose k
 157 using the *minimum message length* principle (Wallace and Boulton, 1968).

158 Consider a given sample x . The message length of a two-part code is the length of the
 159 compressed message plus the length of the compressor/decompressor. In our case, the length of
 160 the compressed message is given by the entropy of our above probability distribution:

$$L(x; \hat{P}(\cdot | \pi_k, \sigma_k)) := - \sum_{ij} \log \hat{P}(x_{ij} | \pi_k, \sigma_k).$$

161 The compressor/decompressor is equivalent to transmitting the counts for the probability distri-
 162 bution. This can be thought of as transmitting a k length tuple of counts. Let $N = \sum_i (m_i - 2k)$ be
 163 the total number of contexts in the read set (i. e., the total number of prefix and suffix pairs in the
 164 data). Thus, $\binom{N+4^{2k}-1}{4^{2k}-1}$ count distributions are possible amongst the number of total prefix and
 165 suffix pairs ($4^k \times 4^k = 4^{2k}$ distinct prefix/suffix pairs), giving a total message length of

$$ML(x; \hat{P}(\cdot | \pi_k, \sigma_k)) := L(x; \hat{P}(\cdot | \pi_k, \sigma_k)) + \log \binom{N+4^{2k}-1}{4^{2k}-1}.$$

Approximating the R.H.S using Stirling's approximation and dropping constant terms yields

$$ML \approx L(x; \hat{P}(\cdot | \pi_k, \sigma_k)) + \frac{(2N + 2^{4k+1} - 1) \log(N + 2^{4k})}{2} - \frac{(4k(2^{4k+1} - 1) + 1) \log 2}{2}.$$

For suffix free densities the message length simplifies to

$$\begin{aligned} ML(x; \hat{P}(\cdot | \pi_k)) &:= L(x; \hat{P}(\cdot | \pi_k)) + \log \binom{N+4^k-1}{4^k-1} \\ &\approx L(x; \hat{P}(\cdot | \pi_k)) + \frac{(2N + 2^{2k+1} - 1) \log(N + 2^{2k})}{2} - \frac{(2k(2^{2k+1} - 1) + 1) \log 2}{2}, \end{aligned}$$

166 and similarly for prefix free.

167 2.4 Prefix/suffix free contexts

168 The method we have presented so far has been developed for any contexts defined by any
169 combination of prefix and suffix. The question of whether prefix/suffix-free contexts or full
170 contexts (both prefix and suffix) naturally arises. The decision depends on the type of variants of
171 interest: using full contexts will restrict the variants to single nucleotide variants (SNV), while
172 one sided contexts allow for more general types of variants such as insertions and deletions. Full
173 contexts also have less power to detect variation caused by close-by SNVs; two SNVs in close
174 proximity will create several different contexts when modelling with both prefixes and suffixes.
175 It is also worth remarking that the choice between prefix and suffix free contexts is immaterial
176 under the assumption of independent noise and sufficient coverage. Thus, our experiments
177 concentrate on suffix-free contexts as it is the more general case.

178 2.5 Reference-based variant calling

179 To compare the ability of our proposed method to a reference-based approach, we have processed
180 all datasets using a standard mapping-based SNP calling pipeline. Using SAMtools v1.2-34, raw
181 reads from each sample were mapped to the relevant reference sequence and sorted. The mapped
182 reads are then further processed to remove duplicates arising from PCR artefacts using Picard
183 v1.130 and to realign reads surrounding indels using GATK v3.3-0. Pileups are then created
184 across all samples using SAMtools and SNPs are called using the consensus-method of BCFtools
185 v1.1-137. The resulting SNPs were then filtered to remove those variants with phred-scaled
186 quality score below 20, minor allele frequency below 0.01 or SNPs that were called in less than
187 10% of samples.

188 3 Results

189 3.1 Simulation study

190 We first investigate the power and the false positive rate (FPR) of our method by simulations
191 as minor allele frequency (MAF), sequencing depth, and sample size are varied. A total of
192 3,000 contexts per sample, of which one was a variant site with two possible alleles across the
193 population, were simulated by sampling counts from a multinomial distribution. This corresponds
194 to a simulating a SNP, indel or any other variant whose first base, i. e., the base directly following
195 the context, is bi-allelic. Each context was simulated with a sequencing read error of 1% by
196 sampling from a multinomial distribution, with the total number of simulations per context
197 determined by the specified sequencing depth. Variants were determined by fitting a gamma
198 distribution and rejecting at a level of $p < 0.05$ corrected for multiple testing by Bonferroni's
199 method. This procedure is repeated 1,000 times for each combination of simulation parameters.

200 Figure 1 and fig. 2 shows the results of the simulation. With a depth of 25 our method is
201 able to recover the variant site with high power when the MAF is 20% or higher, even with few
202 samples (50). The FPR was also well controlled, but reduces sharply with moderate depth (>25)
203 at 100 samples, and is low at most depth for 1,000 samples. Identification of rare variants at low
204 sample sizes (1% MAF at 100 samples) is not reliable, however rare variants are still identifiable
205 with high power at high depth and samples (depth greater than 64 and 1,000 samples).

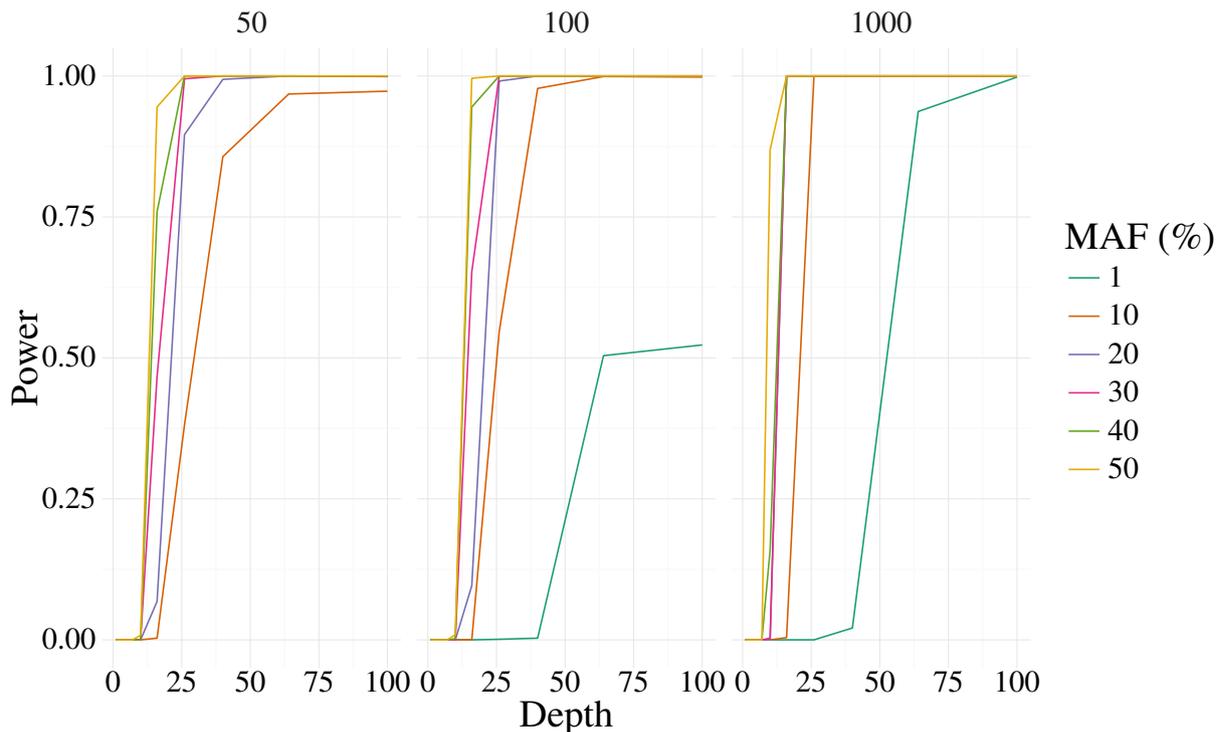


Figure 1: Power curves for 3000 simulated contexts with a single variant context for varying depth and sample size (panels). The bi-allelic variant context was simulated 1,000 times and curves show the mean of the 1,000 simulations. The error for the mean is less than 3% in all cases.

3.2 Empirical experiments

We also evaluated our method on three different datasets: two datasets are of *Streptococcus pneumoniae* bacteria, one collected in Massachusetts (Nicholas J Croucher, Finkelstein, Pelton, et al., 2013) and the other in Thailand (Chewapreecha, Harris, Nicholas J Croucher, et al., 2014); and one mouse dataset (Fairfield, Gilbert, Barter, et al., 2011). The two *S. pneumoniae* datasets comprise 681 and 3,369 samples sequenced using Illumina sequencing technology. The Jax6 mouse dataset (Fairfield, Gilbert, Barter, et al., 2011) contains sequenced exomes of 16 inbred mouse lines.

All experiments were conducted with suffix-free contexts and only contexts present across all samples were evaluated for variants. Our method identified 40,071 variants in the Massachusetts dataset, 57,050 in the Thailand dataset, and 50,000 in the mouse dataset. We refer to these as *KL variants*.

We also compare our method with a mapping-based SNP calling approach on the *S. pneumoniae* datasets. Using sequence for *S. pneumoniae* ATCC 700669 (NCBI accession NC_011900.1) as a reference, there were 181,511 and 251,818 SNPs called for the Massachusetts and Thailand datasets. To be comparable with the resulting binary SNPs calls, we transform our multi-allelic variants to binary variants with the major allele being one and other alleles being zero.

Finally, we compare our results with variants called by another reference-free caller DiscoSNP++ (Uricaru, Rizk, Lacroix, et al., 2015) (v2.2.1). DiscoSNP++ finds 8,728 variants for the Massachusetts *S. pneumoniae* data, and 290,615 variants for Jax6. DiscoSNP++ results are not available on the Thailand dataset as the software fails to run in reasonable time on such a large dataset.

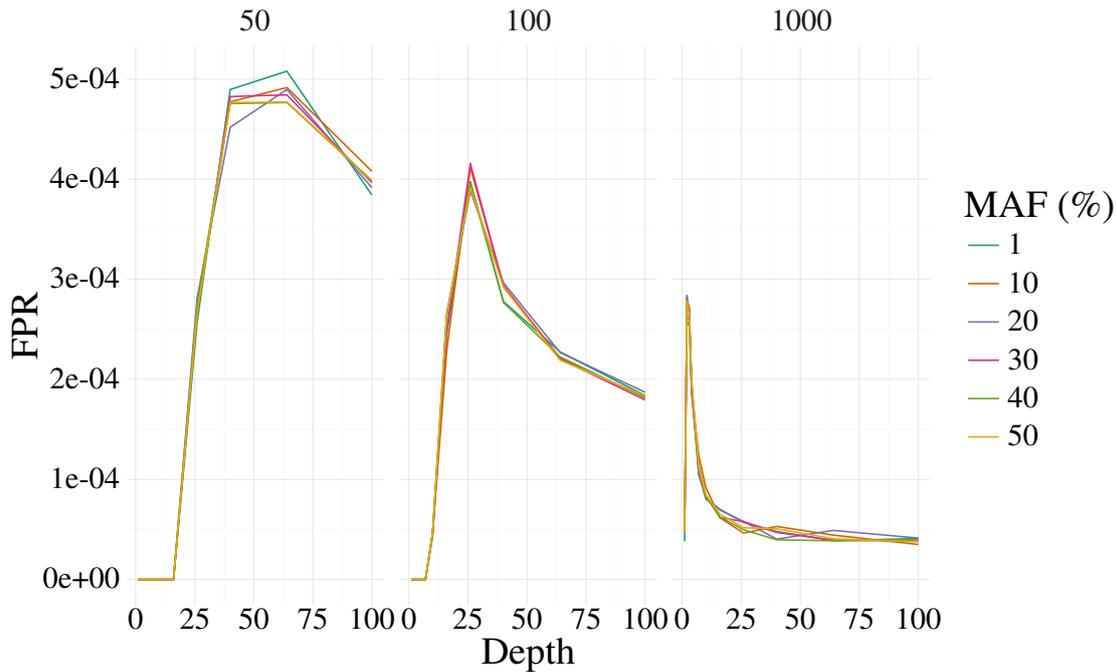


Figure 2: False positive rate for 3000 simulated contexts with a single variant context for varying depth and sample size (panels) as described in fig. 1. The error for the mean is less than 3% in all cases.

228 3.3 Message lengths

229 Our first experiment investigated the optimal k resulting from our message length criterion (see
 230 section 2.3). Figure 3 shows the results of various contexts sizes on three samples, one from
 231 each of the Massachusetts *S. pneumoniae*, Thailand *S. pneumoniae* and Jax6 mouse data. Both *S.*
 232 *pneumoniae* samples had the shortest message length at $k = 14$, and the 129S1/SvImJ mouse
 233 line had the shortest message length at $k = 15$.

234 To evaluate the stability of the message length criterion, the optimal k according to message
 235 length was calculated on all samples from the Massachusetts data (table 1). Most samples (83%)
 236 had an optimal length of $k = 14$, with the remainder being optimal at $k = 13$. Investigation into
 237 the singleton sample with minimal length at $k = 9$ revealed a failed sequencing with only 18,122
 238 reads present. We also evaluated all samples present in the Jax6 dataset and found all samples
 239 had minimal message length at $k = 15$. The stability of k is therefore high and we use $k = 14$
 240 for the two *S. pneumoniae* datasets and $k = 15$ for the Jax6 mouse dataset henceforth in all
 241 experiments.

Table 1: Proportion of samples in Massachusetts data by optimal k .

Optimal k	Count
9	1
13	113
14	567

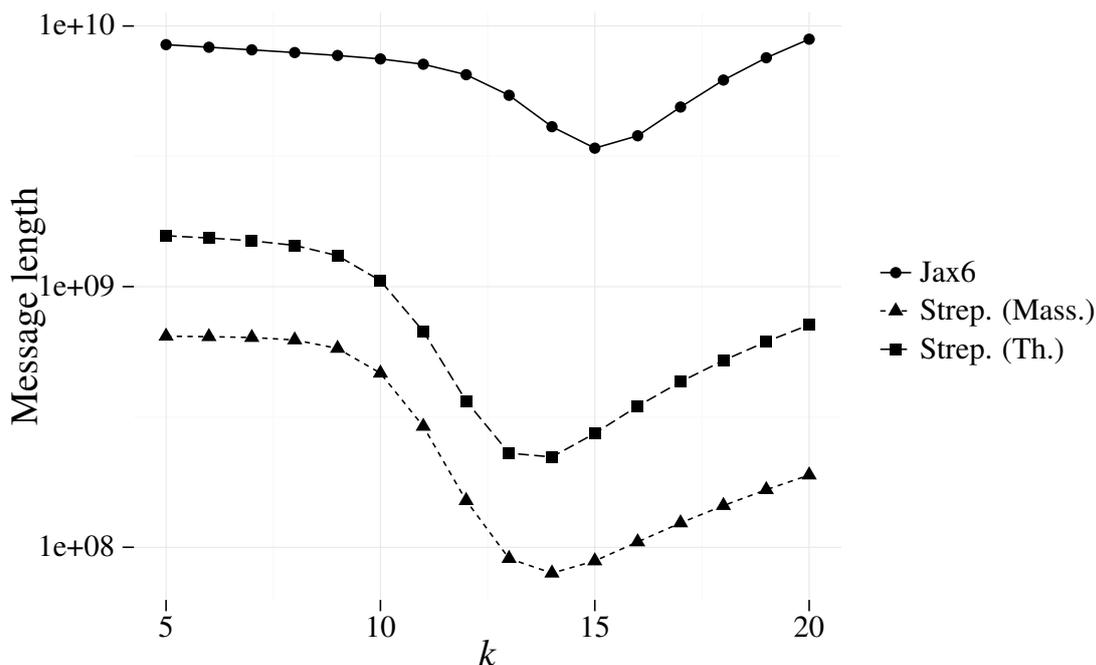


Figure 3: Message length for prefix-only contexts on two *S. pneumoniae* samples from the Massachusetts and Thailand datasets, and the 129S1/SvImJ mouse line from the Jax6 dataset. The optimal k under the MML framework is $k = 14$ for the *S. pneumoniae* datasets and $k = 15$ for Jax6.

3.4 Supervised learning performance

To investigate the robustness of our variants for genomic prediction tasks, we evaluated the ability of variants called on the Massachusetts *S. pneumoniae* dataset for the prediction of Benzylpenicillin resistance under different training and testing scenarios across the two *S. pneumoniae* datasets. Each sample was labelled as resistant if the minimum inhibitory concentration exceeded $0.063 \mu\text{g/mL}$ (Chewapreecha, Marttinen, Nicholas J. Croucher, et al., 2014). In all tasks, a support vector machine (SVM) (Schlkopf and Smola, 2001) was used to predict resistance from the variants, and the performance measured using the Area under the Receiver Operating Characteristic (AROC).

Table 2 shows the results of the experiments. Each row indicates what dataset models were trained on and the columns denote the testing dataset. For intra-dataset experiments (i. e., the diagonal), AROC was estimated using 10-fold cross validation.

Our variants are clearly capturing the various resistance mechanisms, as evident by the strong 10-fold cross validation predictive performance. In comparison to the traditional pipeline and DiscoSNP++ features (on Massachusetts data only) also performed well. Given the high level of accuracy, the three methods do not differ significantly in performance.

The model trained using our variants on the Massachusetts data is moderately predictive on the Thailand dataset. Conversely, the model from the Thailand dataset can also moderately predict resistance in the Massachusetts data, but to a lesser degree. One possible explanation for this limited predictive ability is the existence of resistance mechanisms unique to each dataset, hence a model trained on one dataset will not capture unobserved mechanisms and consequently the model is unable to predict resistance arising from these unknown mechanisms. This hypothesis is supported by the strong performance observable on the diagonal: when

Table 2: AMR prediction results using KL variants. Variants were discovered only on the Massachusetts dataset and then called on both Massachusetts and Thailand datasets. Each row indicates what dataset models were trained on and the columns denote the testing dataset. Numbers are the Area Under the Receiver Operating Characteristic (AROC). The AROC was estimated using 10-fold cross-validation within datasets. The numbers in parentheses are the performance when predicting on standard SNP calls (S) derived through a traditional alignment pipeline and DiscoSNP++ (D) calls. DiscoSNP++ results are not available on the Thailand dataset as the software fails to run in reasonable time on such a large dataset.

Training dataset	Massachusetts	Thailand	All
Massachusetts	95.6 (S: 94.4,D: 96.6)	81.3 (S: 88.6)	
Thailand	72.5 (S: 66.8)	97.6 (S: 97.6)	
All			97.1

265 combining both datasets and performing cross-validation, the performance is high.

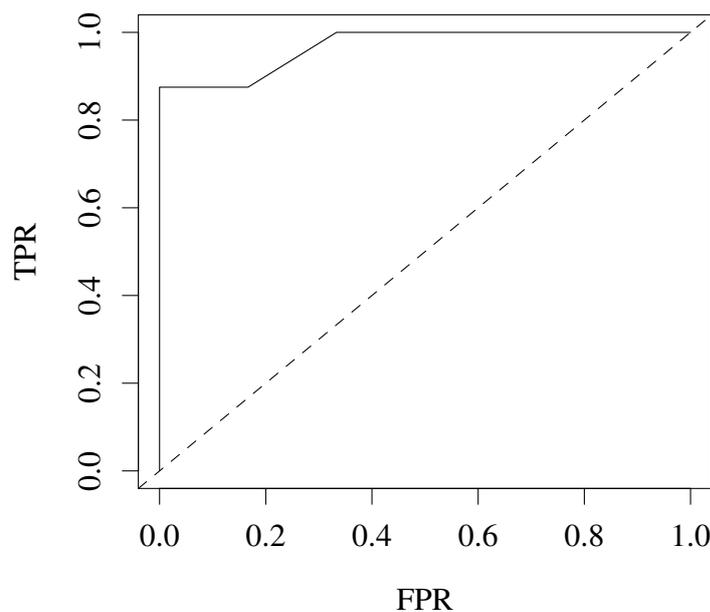


Figure 4: ROC produced from leave-one-out cross-validation performance predicting agouti coat colour from KL variants on Jax6 mouse dataset. AROC is 96%.

266 We also evaluated our variants for predicting coat colour on the Jax6 mouse dataset (Fairfield,
 267 Gilbert, Barter, et al., 2011). As few samples are available (14 labelled samples), we reduced
 268 the problem to a 2-class classification problem, classifying coat colour into agouti or not. This
 269 led to a well balanced classification problem with 8 samples in the agouti classes and 6 not.
 270 The performance for this task was estimated at 96% AROC using leave-one-out (LOOCV)
 271 cross-validation, suggesting the variants are also predictive of heritable traits in higher level
 272 organisms. Figure 4 shows the ROC for this classification problem.

273 3.5 Population structure

274 Finally, we investigate the population structure captured by KL variants and the SNP calls on the
275 Massachusetts dataset. The population structures were estimated using Principle Component
276 Analysis (PCA), a common approach whereby the top principal components derived across all
277 genetic variants reflect underlying population structure rather than the studied phenotype of
278 interest (Price, Zaitlen, Reich, et al., 2010). Five sub-populations (clusters) were identified using
279 k -means on the first two principal components from the SNP data. Projecting those 5 clusters
280 on to the principal component scores of our variants (fig. 5) results in highly concordant plots.
281 Four out of the five clusters can be easily identified using our variants, indicating the detected
282 variation preserves population structures well.

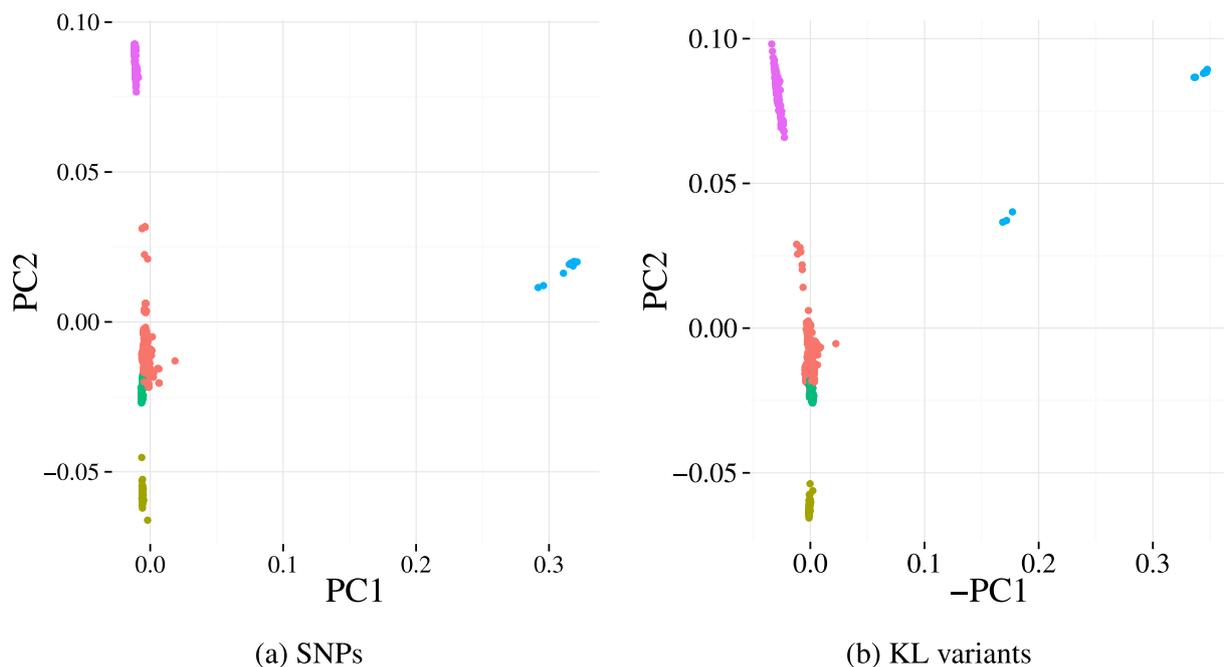


Figure 5: First two principal components derived from alignment-based SNP calls (left) and from variants detected by our method (right) applied to the Massachusetts *S. pneumoniae* dataset. Each point represents a sample and the colours denotes the cluster assignment determined by k -means clustering. The similar pattern of samples in each plot indicates that the same population structure signal is detected by the two variant detection methods.

283 A canonical correlation analysis (CCA) was performed to further assess the similarities
284 between the two feature sets. Regularisation was used to find the canonical vectors as the
285 cross-covariance matrices are singular for our dataset. As there are significantly more features
286 than samples, regularised CCA was used and the correlation between projections estimated using
287 100 samples of leave-one-out bootstrap (Hastie, Tibshirani, and Friedman, 2011). We found the
288 first three components explain all the variance (99%), with the first component alone explaining
289 76%. Therefore, both mapping-based SNPs and KL variants are largely capturing the same
290 variance on the Massachusetts data.

291 3.6 Analysis of contexts

292 To further elucidate the type of variants that are being discovered by our method, we aligned
293 the significant contexts from the Massachusetts dataset to the *S. pneumoniae* reference. Of the

Table 3: Correlation coefficients for first 5 CCA components, estimated using 10-fold cross-validation on Massachusetts data.

Component	Correlation coefficient ($\pm 95\%$ CI)
1	0.873 ± 0.014
2	0.880 ± 0.006
3	0.877 ± 0.007
4	0.862 ± 0.007
5	0.867 ± 0.008

294 contexts, less than 1% failed to align, 41% aligned in a single location, and the remainder aligned
295 in two or more locations.

296 One context aligned in 82 different locations in the reference genome. Further investigation
297 revealed the context corresponds to a *boxB* repeat sequence. Such repeats have previously
298 been used to identify population structure of *S. pneumoniae* isolates carrying the 12F serotype,
299 supporting our population structure findings (Rakov, Ubukata, and Robinson, 2011). This
300 suggests the variants may be tagging more complex structural elements than just single nucleotide
301 variants.

302 4 Conclusions

303 We presented a novel reference-free variant detection method for next-generation sequence data.
304 Our method has the advantage of no tuning parameters, rapid calling of known variants on new
305 samples, and may be suited for targeted genotyping once a known set of variants are obtained.

306 Simulation experiments showed the method is relatively robust and has good power and
307 FPR to detect common variants, but for rare variants the power was lower and a high depth and
308 number of samples were required to reliably detect them.

309 In a typical genomic prediction setting the method was able to predict heritable phenotypes
310 on both a bacterial dataset (anti-microbial resistance) and on a mouse dataset (coat-colour). On
311 the *S. pneumoniae* datasets, our method was shown to have similar performance to a standard
312 alignment-based SNP calling pipeline, with its requirements for a suitable reference genome.
313 Moreover, the method was shown to capture the same population structure on the Massachusetts
314 Streptococcus bacterial datasets as an alignment-based variant calling approach. These results
315 show our method is capable of capturing important genomic features without a known reference.

316 As with other reference-free variant calling methods, interpretation of the detected variants is
317 more difficult compared to a mapping-based approach as called variants are reported without
318 positional information. One approach to obtain such annotations is to map the variant and its
319 context back to a given reference. Given that most sequences with a length greater than 15bp that
320 exist in a given bacterial reference will have a unique mapping, many variants could be easily
321 mapped back. However, such information is unlikely to exist for variants that do not occur in the
322 reference, or may be misleading for variants that arise through complicated procedures such as
323 horizontal gene transfer. Alternatively, variants and their context could be examined via BLAST
324 searches to determine whether these sequences correspond to previously identified genes or other
325 genomic features.

326 In our experiments we used a combination of these approaches to investigate some of the
327 variants found on the bacterial dataset. We identified contexts that mapped to numerous locations
328 in the reference genome and then used BLAST to identify the likely origin of the sequence.

329 Through this method, variants associated with boxB repeat sequence were found, suggesting our
330 method is capturing variance associated with complex structures.

331 We envisage that the method proposed here could be used to conduct a rapid initial analysis
332 of a given dataset, such as species identification, outbreak detection or genomic risk prediction.
333 Our method also enables analysis of data without a suitable reference while still avoiding the
334 computationally expensive step of assembly. Furthermore, our method scales linearly with the
335 total number of reads, allowing application to large datasets.

336 The statistical framework established in this work is quite general and could be expanded
337 in several ways. While we have examined only single nucleotide variants within this work,
338 insertions and deletions could be explicitly modelled within this framework at the cost of
339 increased computational expense. It may also be possible to model other types of variants, such
340 as microsatellites, provided that a suitable representation for them could be found.

341 Acknowledgements

342 We thank Thomas Conway and Noel Faux for helpful discussions.

343 References

- 344 Beroukhim, Rameen, Craig H. Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry
345 Donovan, Jordi Barretina, Jesse S. Boehm, Jennifer Dobson, Mitsuyoshi Urashima, Kevin
346 T. Mc Henry, Reid M. Pinchback, Azra H. Ligon, Yoon-Jae Cho, Leila Haery, Heidi Greulich,
347 Michael Reich, Wendy Winckler, Michael S. Lawrence, Barbara A. Weir, Kumiko E. Tanaka,
348 Derek Y. Chiang, Adam J. Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld,
349 Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic J. Kaye, Hideofumi
350 Sasaki, Joel E. Tepper, Jonathan A. Fletcher, Josep Tabernero, José Baselga, Ming-Sound
351 Tsao, Francesca Demichelis, Mark A. Rubin, Pasi A. Janne, Mark J. Daly, Carmelo Nucera,
352 Ross L. Levine, Benjamin L. Ebert, Stacey Gabriel, Anil K. Rustgi, Cristina R. Antonescu,
353 Marc Ladanyi, Anthony Letai, Levi A. Garraway, Massimo Loda, David G. Beer, Lawrence
354 D. True, Aikou Okamoto, Scott L. Pomeroy, Samuel Singer, Todd R. Golub, Eric S. Lander,
355 Gad Getz, William R. Sellers, and Matthew Meyerson (2010). “The landscape of somatic
356 copy-number alteration across human cancers”. In: *Nature* 463.7283, pp. 899–905.
- 357 Chewapreecha, Claire, Simon R Harris, Nicholas J Croucher, Claudia Turner, Pekka Marttinen,
358 Lu Cheng, Alberto Pessia, David M Aanensen, Alison E Mather, Andrew J Page, Susannah J
359 Salter, David Harris, Francois Nosten, David Goldblatt, Jukka Corander, Julian Parkhill,
360 Paul Turner, and Stephen D Bentley (2014). “Dense genomic sampling identifies highways
361 of pneumococcal recombination”. In: *Nature genetics* 46.3, pp. 305–309.
- 362 Chewapreecha, Claire, Pekka Marttinen, Nicholas J. Croucher, Susannah J. Salter, Simon
363 R. Harris, Alison E. Mather, William P. Hanage, David Goldblatt, Francois H. Nosten,
364 Claudia Turner, Paul Turner, Stephen D. Bentley, and Julian Parkhill (2014). “Comprehensive
365 identification of single nucleotide polymorphisms associated with beta-lactam resistance
366 within pneumococcal mosaic genes”. In: *PLoS genetics* 10.8, e1004547.
- 367 Croucher, Nicholas J, Jonathan A Finkelstein, Stephen I Pelton, Patrick K Mitchell, Grace M Lee,
368 Julian Parkhill, Stephen D Bentley, William P Hanage, and Marc Lipsitch (2013). “Population
369 genomics of post-vaccine changes in pneumococcal epidemiology”. In: *Nature genetics* 45.6,
370 pp. 656–663.

- 371 Fairfield, Heather, Griffith J Gilbert, Mary Barter, Rebecca R Corrigan, Michelle Curtain, Yuem-
372 ing Ding, Mark D'Ascenzo, Daniel J Gerhardt, Chao He, Wenhui Huang, Todd Richmond,
373 Lucy Rowe, Frank J Probst, David E Bergstrom, Stephen a Murray, Carol Bult, Joel Richard-
374 son, Benjamin T Kile, Ivo Gut, Jorg Hager, Snaevor Sigurdsson, Evan Mauceli, Federica
375 Di Palma, Kerstin Lindblad-Toh, Michael L Cunningham, Timothy C Cox, Monica J Jus-
376 tice, Mona S Spector, Scott W Lowe, Thomas Albert, Leah Donahue, Jeffrey Jeddloh, Jay
377 Shendure, and Laura G Reinholdt (2011). "Mutation discovery in mice by whole exome
378 sequencing". In: *Genome Biology* 12.9, R86. ISSN: 1465-6906. DOI: 10.1186/gb-2011-
379 12-9-r86.
- 380 Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2011). *The Elements of Statistical*
381 *Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in*
382 *Statistics)*. 2nd ed. 2009. Corr. 7th printing 2013. Springer. ISBN: 9780387848570.
- 383 Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean (2012). "De novo
384 assembly and genotyping of variants using colored de Bruijn graphs". In: *Nature genetics*
385 44.2, pp. 226–232.
- 386 Kamary, K., K. Mengersen, C. P. Robert, and J. Rousseau (2014). "Testing hypotheses via a
387 mixture estimation model". In: *ArXiv e-prints*. arXiv: 1412.2044 [stat.ME].
- 388 Li, Heng (2012). "Exploring single-sample SNP and INDEL calling with whole-genome de novo
389 assembly". In: *Bioinformatics* 28.14, pp. 1838–1844.
- 390 Ochman, Howard, Jeffrey G Lawrence, and Eduardo A Groisman (2000). "Lateral gene transfer
391 and the nature of bacterial innovation". In: *Nature* 405.6784, pp. 299–304.
- 392 Peterlongo, Pierre, Nicolas Schnel, Nadia Pisanti, Marie-France Sagot, and Vincent Lacroix
393 (2010). "Identifying SNPs without a reference genome by comparing raw reads". In: *String*
394 *Processing and Information Retrieval*. Springer, pp. 147–158.
- 395 Price, Alkes L, Noah A Zaitlen, David Reich, and Nick Patterson (2010). "New approaches to
396 population stratification in genome-wide association studies". In: *Nature Reviews Genetics*
397 11.7, pp. 459–463.
- 398 Rakov, A. V., K. Ubukata, and D. A. Robinson (2011). "Population structure of hyperinvasive
399 serotype 12F, clonal complex 218 Streptococcus pneumoniae revealed by multilocus boxB
400 sequence typing." In: *Infect. Genet. Evol.* 11.8, pp. 1929–1939. DOI: 10.1016/j.meegid.
401 2011.08.016.
- 402 Schlkopf, Bernhard and Alexander J. Smola (2001). *Learning with Kernels: Support Vector*
403 *Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine*
404 *Learning)*. 1st. The MIT Press. ISBN: 9780262194754.
- 405 Sokal, Robert R. and F. James Rohlf (1994). *Biometry: The Principles and Practices of Statistics*
406 *in Biological Research*. 3rd. W. H. Freeman. ISBN: 9780716724117.
- 407 Uricaru, Raluca, Guillaume Rizk, Vincent Lacroix, Elsa Quillery, Olivier Plantard, Rayan Chikhi,
408 Claire Lemaitre, and Pierre Peterlongo (2015). "Reference-free detection of isolated SNPs".
409 In: *Nucleic acids research* 43.2, e11–e11.
- 410 Wallace, C S and D M Boulton (1968). "An information measure for classification". In: *The*
411 *Computer Journal* 11.2, pp. 185–194.