MICROBIAL GENOMICS

RESEARCH ARTICLE

Bush, *Microbial Genomics* 2020;6

DOI 10.1099/mgen.0.000434

Read trimming has minimal effect on bacterial SNP-calling accuracy

Stephen J. Bush*

Abstract

Read alignment is the central step of many analytic pipelines that perform variant calling. To reduce error, it is common practice to pre-process raw sequencing reads to remove low-quality bases and residual adapter contamination, a procedure collectively known as 'trimming'. Trimming is widely assumed to increase the accuracy of variant calling, although there are relatively few systematic evaluations of its effects and no clear consensus on its efficacy. As sequencing datasets increase both in number and size, it is worthwhile reappraising computational operations of ambiguous benefit, particularly when the scope of many analyses now routinely incorporates thousands of samples, increasing the time and cost required. Using a curated set of 17 Gram-negative bacterial genomes, this study initially evaluated the impact of four read-trimming utilities (Atropos, FASTP, Trim Galore and Trimmomatic), each used with a range of stringencies, on the accuracy and completeness of three bacterial SNP-calling pipelines. It was found that read trimming made only small, and statistically insignificant, increases in SNP-calling accuracy even when using the highest-performing pre-processor in this study, FASTP. To extend these findings, >6500 publicly archived sequencing datasets from *Escherichia coli*, *Mycobacterium tuberculosis* and *Staphylococcus aureus* were re-analysed using a common analytic pipeline. Of the approximately 125 million SNPs and 1.25 million indels called across all samples, the same bases were called in 98.8 and 91.9% of cases, respectively, irrespective of whether raw reads or trimmed reads were used. Nevertheless, the proportion of mixed calls (i.e. calls where <100% of the reads support the variant allele; considered a proxy of false positives) was significantly reduced after trimming, which suggests that while trimming rarely alters the set of variant bases, it can affect the proportion of reads supporting each call. It was concluded that read quality- and adapter-trimming add relatively little value to a SNP-calling pipeline and may only be necessary if small differences in the absolute number of SNP calls, or the false call rate, are critical. Broadly similar conclusions can be drawn about the utility of trimming to an indel-calling pipeline. Read trimming remains routinely performed prior to variant calling likely out of concern that doing otherwise would typically have negative consequences. While historically this may have been the case, the data in this study suggests that read trimming is not always a practical necessity.

DATA SUMMARY

All analyses conducted in this study used publicly available third-party software. All data and parameters necessary to replicate these analyses are provided within the article or through supplementary data files. A total of >6500 Sequence Read Archive (SRA) sample accession numbers, representing Illumina paired-end sequencing data from *Escherichia coli*, *Mycobacterium tuberculosis* and *Staphylococcus aureus*, and used to evaluate the impact of FASTQ pre-processing, are listed in Tables S1, S2 and S3 (available with the online version of this article). This study makes use of Illumina, Oxford Nanopore Technologies (ONT) and PacBio sequencing data from a study by De Maio *et al.* [1], available via the National Center for Biotechnology Information (NCBI) SRA under BioProject PRJNA422511 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA422511). The associated Illumina/ONT and Illumina/PacBio assemblies are available via FigShare (https://doi.org/10.6084/m9.figshare.7649051). The Illumina reads were used in a previous evaluation of 209 SNP-calling pipelines [2], for

Received 08 September 2020; Accepted 25 November 2020; Published 11 December 2020

Author affiliations: [1]Nuffield Department of Medicine, University of Oxford, Oxford, UK.

***Correspondence:** Stephen J. Bush, stephen.bush@ndm.ox.ac.uk

Keywords: read pre-processing; read trimming; SNP calling; variant calling.

Abbreviations: NCBI, National Center for Biotechnology Information; NIHR, National Institute for Health Research; ONT, Oxford Nanopore Technologies; SRA, sequence read archive.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Ten supplementary figures and eight supplementary tables are available with the online version of this article.

which the SNP call truth sets, also used here, are available via the GigaDB repository (http://dx.doi.org/10.5524/100694).

Impact Statement

Short-read sequencing data is routinely pre-processed before use, to trim off low-quality regions and remove contaminating sequences introduced during its preparation. This cleaning procedure – read trimming – is widely assumed to increase the accuracy of any later analyses, although there are relatively few systematic evaluations of trimming strategies and no clear consensus on their efficacy. In this study, real sequencing data from 17 bacterial genomes was used to show that several commonly used read-trimming tools, used across a range of stringencies, had only a minimal, statistically insignificant, effect on later SNP calling. To extend these results, >6500 publicly archived sequencing datasets were re-analysed, calling SNPs and indels both with and without any read trimming. Using a common analytic pipeline, it was found that of the approximately 125 million SNPs within this dataset, 98.8% were identically called irrespective of whether raw reads or trimmed reads were used. Taken together, these results question the necessity of read trimming as a routine pre-processing operation.

INTRODUCTION

Read alignment is the central step of many analytic pipelines that perform variant calling. To reduce error, it is common practice to pre-process raw sequencing reads to remove low-quality bases and residual adapter contamination, a procedure collectively known as trimming. This is because, assuming Illumina sequencing data, errors are non-randomly distributed over the length of the read, clustering towards the 3′ end (which is also where adapters are located). These poorer-quality flanking regions are frequently trimmed to leave only the higher-quality internal bases.

Numerous pre-processing tools ('read trimmers') exist for this purpose, which often simultaneously perform both quality- and adapter-trimming. However, previous studies differ as to whether the effect of trimming on downstream SNP calling is generally beneficial [3, 4], generally minimal [5], or conditional on the genome and aligner used [6]. Similarly, quality trimming was reported to have little effect on the completeness of *de novo* genome assembly [7] and be detrimental to *de novo* transcriptome assembly unless comparatively gentle (i.e. trimming on the basis of quality score <5) [8] although for assembly purposes, adapter trimming is generally recommended [9]. Taken together, the benefits of trimming do not appear universal and in many situations may not be realized at all.

As sequencing datasets increase both in number and size, it is worthwhile reappraising computational operations of ambiguous benefit, particularly when the scope of many analyses now routinely incorporate thousands of samples. To that end, this study evaluates the effect of several read-trimming strategies on the subsequent accuracy and completeness of various bacterial variant-calling pipelines. Four commonly used read trimmers – FASTP [10], Trimmomatic [11], Atropos [12] and TrimGalore (the latter two employing Cutadapt [13]), each of which applies a different strategy to adapter detection and removal – were evaluated across a range of stringencies.

To assess the effect of pre-processing stringency upon the precision (positive predictive value) and recall (sensitivity) of SNP calling, I quality- and adapter-trimmed 17 sets of 150 bp Illumina HiSeq 4000 paired-end reads before calling SNPs relative to a set reference genome using three different pipelines (the pairwise combination of one read aligner, BWA-MEM [14], and three variant callers, LoFreq [15], mpileup [16] and Strelka [17]). These reads represent environmentally sourced samples of the genera *Citrobacter*, *Klebsiella*, *Escherichia* and *Enterobacter*, and were obtained from a previous study [1] and curated for use in a large-scale comparison of bacterial SNP-calling pipelines [2]. The three pipelines chosen for the present study were previously found to be among the highest performing when tested on divergent bacterial data [2].

To determine the effect of trimming upon indels as well as SNPs, and upon a larger-scale dataset, I applied the highest-performing strategy identified for the Gram-negative dataset to a set of >6500 publicly archived *Escherichia coli*, *Mycobacterium tuberculosis* and *Staphylococcus aureus* sequencing reads. This represents a substantive, and diverse, range of Illumina sequencing platforms, library preparation strategies, read lengths, insert sizes and coverage depths. In this dataset, I have no a priori knowledge of which SNP and indel calls are accurate, although for the purpose of this analysis this is not relevant – the intention was to compare two sets of variants called before and after a common set of pre-processing steps were applied, and so I was not interested in the correctness of each call but whether the same calls were made in each condition. This dataset is sufficiently large (approximately 125 million SNPs and 1.25 million indels) that I can generalize about the benefits of read trimming as a routine procedure.

METHODS

Evaluating the effect of read trimming on SNP-calling accuracy

To evaluate the effect of read trimming upon SNP calling, I first required a truth set of SNPs against which comparisons could be made. These were generated as previously described [2] and briefly recapitulated here. Firstly, a dataset was obtained from a published study [1] that comprised 17 parallel sets of 150 bp Illumina HiSeq 4000 paired-end short reads, and both Oxford Nanopore Technologies (ONT) and SMRT Pacific Biosciences (PacBio) long reads for four *Enterobacter* spp., four *Klebsiella* spp., four *Citrobacter* spp. and three *E. coli* (all environmentally sourced), plus subcultures of stocks from two reference strains, *Klebsiella pneumoniae*

subsp. *pneumoniae* MGH 78578 and *E. coli* CFT073. Sample accession numbers are listed in Table S4.

To assess how divergence between the source and reference genomes affected the performance of SNP-calling pipelines, I used this Gram-negative dataset to generate truth sets of SNPs. To do so, whole-genome alignments were made between each closed assembly and each species' reference genome using both nucmer (from MUMmer v4.0.0beta2) [18] and Parsnp v1.2 [19] with default parameters, with common SNPs then identified within one-to-one alignment blocks. The set of SNPs identically called by both nucmer and Parsnp in both *de novo* assemblies were considered the truth set and used to evaluate the performance of each trimmer/aligner/caller pipeline, as follows.

I first adapter- and quality-trimmed each of the 17 sets of Illumina reads using four different read trimmers – Atropos v1.1.25 [12], fastp v0.20.1 [10], TrimGalore v0.5.0 (https://github.com/FelixKrueger/TrimGalore, accessed 1 April 2020) and Trimmomatic v0.38 [11] – as well as retaining an 'untrimmed' control. Read trimmers often have rich feature sets, although my concern with this study was not to systematically evaluate the broad range of parameters by which a read may be trimmed but to assess, in general, the added-value benefit of trimming when a minimum-effort application was made for each program. In this respect, I considered trimming to encompass two simultaneous pre-processing operations: the removal of adapter sequence (using default parameters where possible) and 3′ quality-trimming (across a range of stringencies).

I configured each trimmer to automatically detect and remove adapter sequence, where this was the default setting (fastp, TrimGalore), or to remove the Illumina universal adapter, should specification be required (Atropos, Trimmomatic). As Illumina reads degrade in quality towards the 3′ end, 3′ trimming is by far the most commonly applied pre-processing operation. Alongside adapter removal, I removed trailing bases should they fall below Phred quality thresholds of 2, 5, 10, 20 or 30 (the Phred scale being logarithmic, this represents base-call accuracies of 37, 68, 90, 99 and 99.9%, respectively). I also required a minimum post-trimming read length of 50 bp and, where possible, for each trimmer to output both paired and unpaired reads, i.e. those reads where both ends of a pair are retained after trimming, and those where one end is discarded (by default, fastp and TrimGalore discard the entire pair if only one end meets the acceptance criteria). Finally, I did not explicitly disable any filter criterion implemented automatically, considering this to be the default recommendation for general-purpose use. The specific parameters used for each trimmer are detailed in Table S5.

Using the trimmed (or, as a control, untrimmed) Illumina reads, and the reference genome for each species (listed in Table S4), SNPs were then called using the pairwise combination of the aligner bwa-mem v0.7.17 [14] with three variant callers, LoFreq v2.1.2 [15], mpileup v1.7 [16] and Strelka v2.9.2 [17], each used with default parameters. Each pipeline applied a common set of post-processing steps: BAM files were cleaned, sorted, had duplicate reads marked and were indexed using Picard Tools v2.17.11 [20], and VCF records were regularized using the vcfallelicprimitives module of vcflib v1.0.0-rc2 (https://github.com/ekg/vcflib, accessed 1 April 2020) to ensure different representations of the same variant were presented in the same way (this module does so by, for example, left-aligning indels and splitting complex variants into individual VCF records). Finally, VCF records were filtered using BCFtools v1.7 [16] to retain only biallelic SNPs with call quality >20 and >5 reads mapped at that position, >75% of which, including at least one in each direction, supporting the alternative allele (as in [21], and broadly similar to those recommended by a previous study for maximizing SNP-calling precision [22]).

To evaluate each pipeline, I calculated precision (positive predictive value), recall (sensitivity) and F-score, a summary measure which considers precision and recall with equal weight, producing a value between 0 and 1 (perfect precision and recall). Precision was calculated as TP/(TP+FP), recall as TP/(TP+FN) and F-score as 2×[(precision×recall)/(precision+recall)], where TP, FP and FN are the number of true-positive, false-positive and false-negative SNP calls, respectively.

The command lines used for these pipelines were also previously implemented [2] within a suite of Perl scripts (so as to handle subsidiary data manipulation operations and calculate summary statistics), available at https://github.com/oxfordmmm/GenomicDiversityPaper. The SNP call truth sets are available via the GigaDB repository at http://dx.doi.org/10.5524/100694.

Finally, when SNP calling using publicly sourced *E. coli*, *M. tuberculosis* and *S. aureus* data (see below), I used only one of the above pipelines, fastp (with trailing Q <20) /bwa-mem/mpileup, aligning all reads relative to *E. coli* K-12 substrain MG1655 (RefSeq assembly accession no. GCF_000005845.2), *M. tuberculosis* H37Rv (RefSeq assembly accession no. GCF_000195955.2) and *S. aureus* subsp. *aureus* NCTC8325 (RefSeq assembly accession no. GCF_000013425.1). This fastp/bwa-mem/mpileup pipeline was as previously used upon the Gram-negative dataset, although to reduce runtime was modified to omit the BAM file cleaning (i.e. Picard CleanSam), VCF regularization and VCF filtering steps (regularization in any case only necessary when comparing VCF records produced by different variant callers). Command lines are detailed in Table S5.

Evaluating the effect of trimming upon a diverse range of publicly archived sequencing datasets

To obtain a broad range of sequencing data across multiple laboratories, I downloaded the daily updated National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) BioProject summary file (*n*=417689 BioProjects; ftp://ftp.ncbi.nlm.nih.gov/bioproject/sumrefseqmary.txt, accessed 22 March 2020), parsing it to extract BioProject IDs with a data type of 'genome sequencing' and associated NCBI taxonomy IDs of 562 (*E. coli*), 1773 (*M. tuberculosis*)

and 1280 (*S. aureus*). These are 'top level' taxonomy IDs, encompassing samples for which an additional level of strain-specificity could not be made. These top level IDs were chosen so that, alongside the criteria detailed below, a set of samples would be obtained that was large enough to draw sound conclusions (*E. coli*, *M. tuberculosis* and *S. aureus* being three of the more commonly sequenced bacteria), but not so large as to be computationally expensive to analyse [which would be the case if obtaining all whole-genome sequencing (WGS) data from every *E. coli*, *M. tuberculosis* and *S. aureus* strain]. These three species were chosen to represent different degrees of genomic diversity, from, broadly speaking, high (*E. coli*) to low (*M. tuberculosis*). I used the Entrez Direct suite of utilities (https://www.ncbi.nlm.nih.gov/books/NBK179288/, accessed 1 May 2019) to associate each BioProject ID with a list of SRA sample and run IDs (a 'RunInfo' file). RunInfo files were parsed to retain only those runs where 'Platform' was 'ILLUMINA', 'Model' (i.e. sequencer) contained 'HiSeq' or 'MiSeq' (all of which use TruSeq3 adapters), 'LibrarySource' was 'GENOMIC', 'LibraryStrategy' was 'WGS', 'LibraryLayout' was 'PAIRED', 'LibrarySelection' was 'RANDOM', 'avgLength' was ≥150 (i.e. mean read length of 150 bp) and 'spots' was >1 and <5 (i.e. approximating a read depth of >1 and <5 million reads). The upper limit on read depth was chosen to minimize the computational cost of processing, although read depth remains sufficiently high as to not compromise SNP calling. For instance, as the *E. coli* genome is approximately 5 million bp in length, sequencing 1–5 million 150 bp paired-end reads represents a mean base-level coverage of approximately 60- to 300-fold. These criteria generated sets of 1661 *E. coli*, 4416 *M. tuberculosis* and 1156 *S. aureus* sequencing reads, detailed in Tables S1, S2 and S3, respectively.

Publicly archived sequencing data often does not contain computationally accessible metadata regarding any informatic pre-processing and nor is it immediately apparent whether pre-processing has already been performed. As such, I excluded from consideration all samples where <1% of the original bases could be trimmed, reasoning that in these circumstances the sample had already been pre-processed using broadly similar trimming criteria to those applied here. The final *E. coli* dataset comprised 1606 samples, representing a particularly diverse set of strains. Relative to the *E. coli* reference, K-12 substrain MG1655, the majority of samples contained between approximately 10000 and 100000 SNPs, and 100 and 500 indels, although with outliers containing as few as 15 SNPs and 3 indels and as many as 308000 SNPs (>103 million SNPs, and >500000 indels, in total), with a mean of 64476 SNPs and 316 indels per sample (Table S6). The final *M. tuberculosis* dataset comprised 3946 samples, incorporating subsets from two large-scale studies from South Africa (representing >2000 of the original 4416 samples) [23] and Canada (>1000 samples) [24]. Each sample contained between approximately 170 and 2500 SNPs, and 20 and 500 indels (approx. 5 million SNPs and approx. 0.5 million indels in total). Across all samples, the mean number of SNPs and indels relative to the *M. tuberculosis* reference, H37Rv, was 1238 and 121, respectively (Table S7). The final *S. aureus* dataset comprised 1100 samples, each of which contained between approximately 1000 and 60000 SNPs, and 70 and 550 indels, relative to the *S. aureus* reference genome, NCTC8325 (mean 15167 SNPs and 223 indels per sample, with approx. 17 million SNPs and approx. 250000 indels in total) (Table S8).

RESULTS AND DISCUSSION

Read trimming has minimal effect on SNP-calling accuracy

It has previously been described how the performance of a bacterial SNP-calling pipeline is affected by divergence between the genome from which reads are sequenced and the genome to which reads are aligned, the latter often being the NCBI 'reference genome', a high-quality (albeit often arbitrary) species representative [2]. To do so, I obtained data from a diverse range of environmentally sourced Gram-negative bacteria (as described in [1] and summarized in Table S4) and used it to generate truth sets of SNPs for evaluation purposes, as previously detailed [2] (and briefly recapitulated in Methods). I chose three SNP-calling pipelines found to be generally higher performing, even when reads were aligned to particularly divergent genomes [2]: the pairwise combination of the aligner BWA-MEM [14] with three different variant callers, LoFreq [15], mpileup [16] and Strelka [17], each used with a common set of post-alignment processing operations.

I used these three pipelines to call SNPs across the range of Gram-negative bacteria, both in the absence of read pre-processing and after prefixing to each pipeline one of four different read trimmers: Atropos, FASTP, TrimGalore and Trimmomatic. Each trimmer was configured to remove adapter sequence as well as to trim bases from the 3′ end of each read, on the basis of Phred quality score (Q) thresholds 2, 5, 10, 20 and 30 (the Phred scale being logarithmic, this represents base call accuracies of 37, 68, 90, 99 and 99.9%, respectively). In total, this dataset contained 1071 records, comprising 17 species × 4 read trimmers × 5 Phred score thresholds × 3 variant callers, plus 17 species × 3 variant callers provided untrimmed data (i.e. not including the latter, 50 records per trimmer/Phred threshold combination).

I found that across the full range of genomes and aligner/caller combinations, trimming had minimal effect on overall SNP-calling accuracy, with only miniscule, insignificant, changes observed in F-score relative to untrimmed data (these changes visible only in the third decimal place of F-score; see Fig. 1, Table S5). It appeared that the highest-performing trimming strategy employed FASTP, which, relative to untrimmed data, produced consistent, albeit statistically insignificant, increases in F-score irrespective of trailing Phred threshold (the median F-score using untrimmed data was 0.9579 and, using FASTP with a threshold of Q <20, 0.958; Mann–Whitney U P=0.976). This pattern appears driven by relatively small decreases in precision (Fig. S1) compensated by slightly larger increases in recall (Fig. S2), which suggests that FASTQ pre-processing

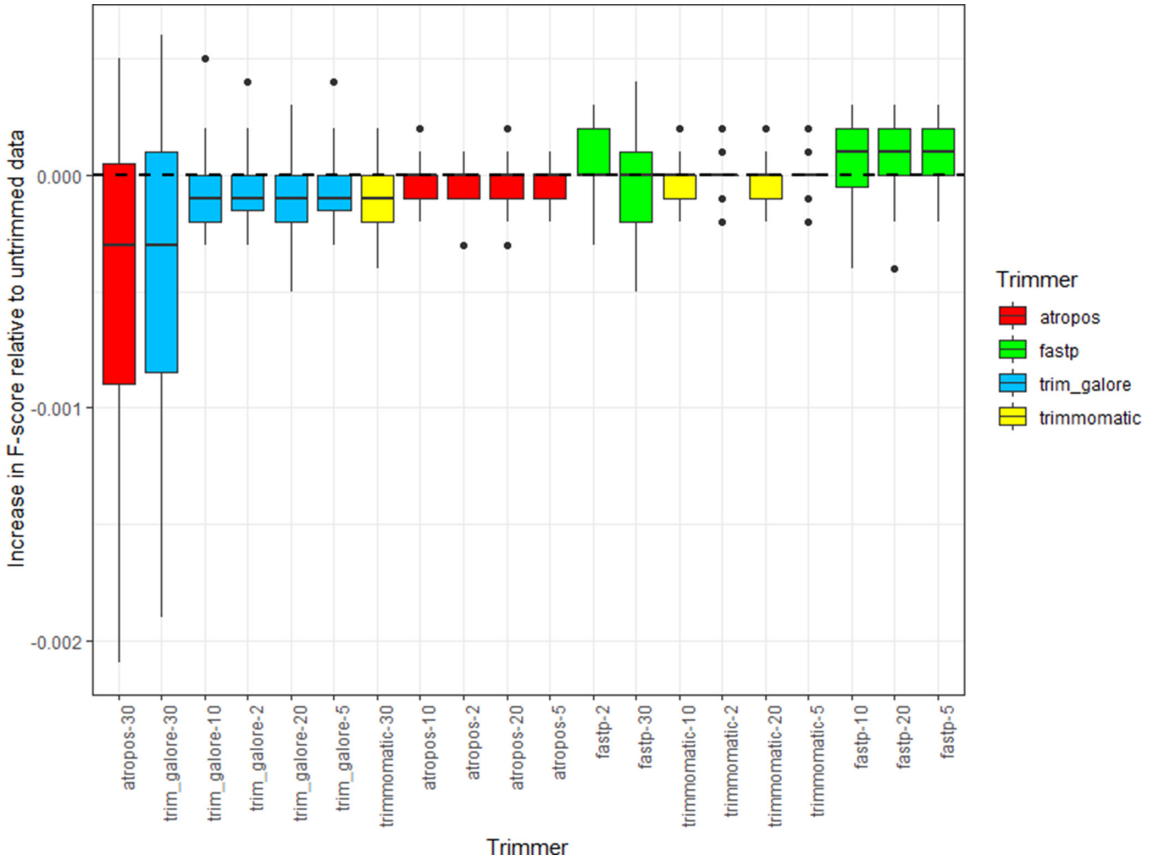

Fig. 1. Effect of read trimming upon F-score, a measure of overall SNP-calling accuracy, in a curated Gram-negative dataset. Median difference in F-score per trimmer relative to untrimmed data, across a range of trimming stringencies (i.e. varying the Phred score threshold for trimming 3′ bases). Boxes represent the interquartile range of the F-score, with midlines representing the median. Upper and lower whiskers extend, respectively, to the largest and smallest values no further than 1.5× the interquartile range. Data beyond the ends of each whisker are outliers and plotted individually. Columns are ordered according to median F-score and coloured according to the trimmer used. The dashed line *y*=0 is marked in black. The raw data for this figure are available in Table S5. Boxplots showing the effect of read trimming upon precision and recall are shown, respectively, in Figs S1 and S2. Note that FASTP implements quality filters other than 3′ trimming by default, which for the data in this figure were retained. A version of this figure with these filters disabled is available in Fig. S3.

enables a small proportion of reads to map that otherwise would not, allowing lower-coverage SNPs to be called that would otherwise be omitted. It is important to note that FASTP differs from the other trimmers in that, unless explicitly disabled, it implements several filters by default – that is, it performs more quality-trimming operations than simply 3′ trimming (discussed below). Notably, 3′ trimming alone (as implemented by the other three trimmers) appears to uniformly, albeit marginally, improve precision (Fig. S1), although at the expense of recall (Fig. S2) and thereby overall F-score (Fig. 1). This is also the case for FASTP if repeating the analysis with the default filters disabled (Fig. S3).

Nevertheless, if considering SNP calling in absolute terms, these results suggest that there is no overt disadvantage to trimming, but little substantive benefit either. To ensure this conclusion is not generalized on the basis of a small number of samples, I selected one SNP-calling pipeline (BWA-MEM/mpileup) and then applied one of the higher-performing trimming strategies (FASTP with trailing Q <20) to hundreds of publicly sourced, and genomically diverse, *E. coli* samples (Table S1), calling SNPs in all cases relative to a single reference genome and applying a common set of post-processing operations to restrict analysis only to higher-confidence calls (see Methods).

I found that trimming, in general, improved the proportion of reads that could be aligned (Fig. 2a), although SNP calling – that is, the interpretation of those alignments – was not substantially altered. In 1579 of the 1606 *E. coli* samples, >99% of SNPs could be identically called irrespective of read trimming (>99.9% in 385 samples, 24% of the total) (Fig. 2b, Table S6). Of the total set of >103 million *E. coli* SNPs, 99.85% were identically called irrespective of any read trimming. It is important to clarify that by identically called I am referring only to whether the base calls are the same, e.g. G/A, and do not distinguish here between a G/A homozygote and a G/A heterozygote, this being a matter of VCF post-processing criteria (see below).

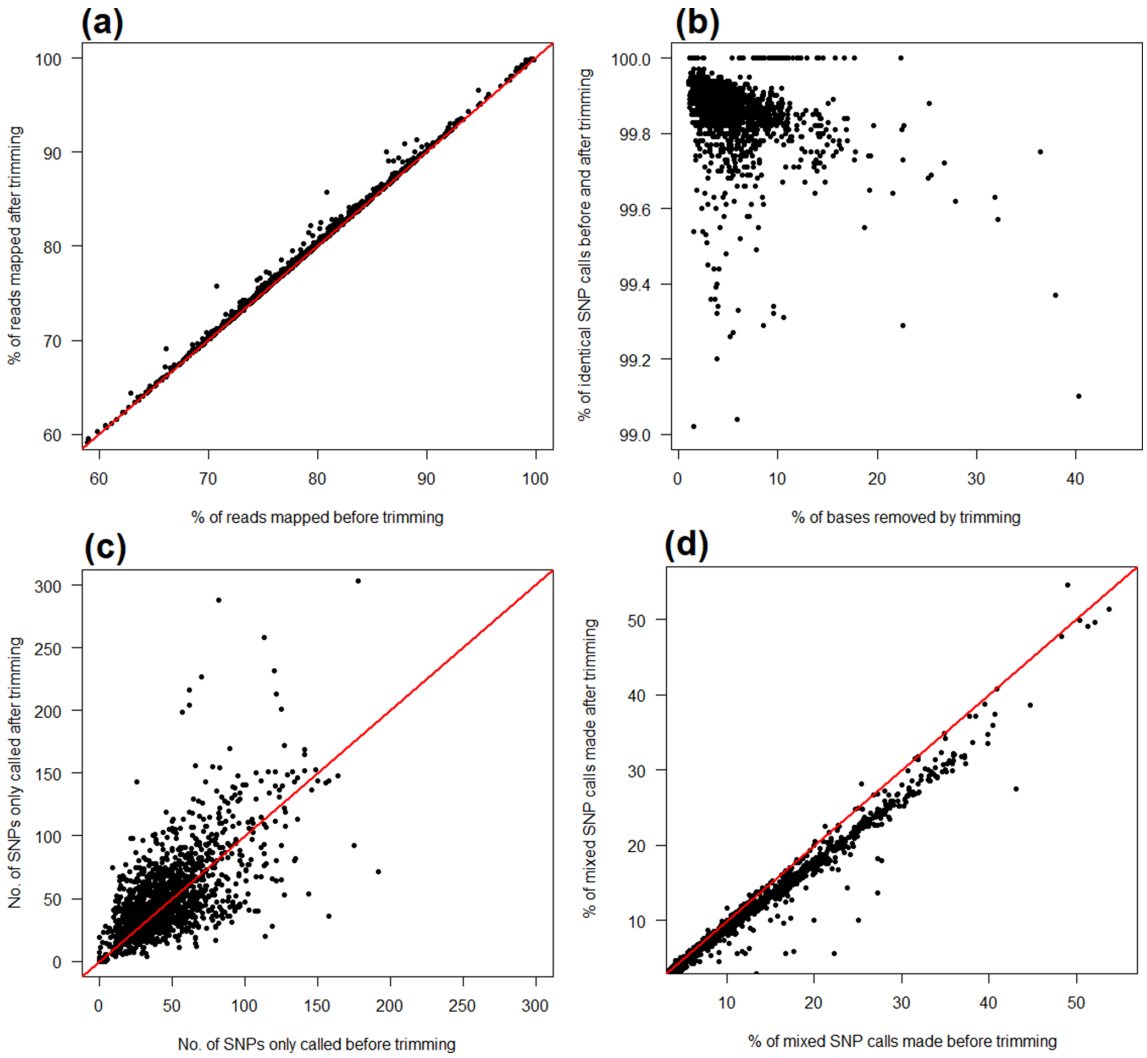

Fig. 2. Effect of read trimming upon SNP calls made using publicly archived *E. coli* sequencing data. Trimming marginally increases the proportion of successfully aligned reads (a), although the interpretation of those alignments (i.e. SNP calling) is not substantially altered, with the vast majority of SNPs (>99%) called irrespective of trimming (b). This value is 100% for a number of samples containing very few SNPs (approximately 15) relative to the *E. coli* reference genome. A relatively small number of SNPs (in the majority of cases <200) are only called when using either raw or trimmed data, but not both (c). The proportion of mixed SNP calls, considered a proxy of false-positive calling, decreases when using trimmed data (d). The raw data for this figure are available in Table S6 and represent 1606 *E. coli* samples, with a mean of 64476 SNPs per sample. The red lines denote *y=x*.

The negative correlation between the percentage of bases removed by trimming and the percentage of identically called SNPs (Fig. 2b) likely reflects reduced sequencing depth in the more extreme cases – some samples are evidently lower quality (according to the filter criteria applied), with >20% of bases trimmed. In one of the more extreme cases, sample SRS3938880 (collected during a transnational survey of veterinary pathogens [25]), 40% of its original set of 1000 million bases were discarded, although compared to the raw data, 99.1% of the total set of 53701 SNPs could still be called (Table S6).

In absolute terms, it also appears that approximately as many SNPs per sample were called only when using trimmed reads as when using raw reads (Fig. 2c). However, even if I was to assume that every SNP called only when using raw reads was erroneous (that is, it is an error that trimming will resolve), this remained <200 SNPs for 1598 of 1606 samples (Table S6).

Although I was unable to ascertain the accuracy of any of the SNP calls in the publicly sourced dataset, and thereby which were true positives, I can nevertheless assume that mixed calls (those where <100% of reads mapped at that position support the variant allele) are a proxy of false positives, as in a previous study [3]. As the proportion of mixed SNP calls is marginally higher for raw than trimmed data, this suggests that SNP calling using raw data may introduce a number of errors that trimming would resolve (the median percentage

of mixed SNP calls is 10.26% when using untrimmed data, and 9.05% when using trimmed data; Mann–Whitney U P=9.1×10^{-7}; Fig. 2d).

Taking the above results together, I can conclude that compared to raw reads, the use of trimmed reads rarely changes the variant base called (see Fig. 2b), but does alter the level of support for a given variant in approximately 1% of cases. However, it is worth noting that the identification of (potentially spurious) mixed calls is by definition a VCF post-processing operation and so could be applied regardless of any FASTQ pre-processing, and at varying levels of stringency.

E. coli is a characteristically diverse species, and so in this analysis the mean number of SNPs per sample was relatively high, at approx. 70000 (Table S6). A consequence of this genomic diversity is that it minimizes the variance introduced by small differences in SNP calling, before and after trimming, which in another situation could be critical. To explore the effect of trimming on a clonal system, where only a small number of SNPs are expected (and so trimming-associated differences would have greater impact), I repeated the analysis using 3946 publicly sourced *M. tuberculosis* samples (Tables S2 and S7).

Unlike *E. coli*, which is sufficiently diverse that only approximately 75% of reads per sample could be aligned to the reference genome (this proportion marginally increased after trimming; Fig. 2a), virtually all *M. tuberculosis* reads could be successfully aligned to the reference, H37Rv, regardless of trimming (although, as with *E. coli*, trimming also increased the proportion of aligned reads, as shown in Fig. 3a). Similar results were observed for *M. tuberculosis* as for *E. coli* with, in the majority of cases, >98% of SNPs identically called irrespective of read trimming (Fig. 3b), a small number of SNPs (often <50) only called when using raw reads or trimmed reads, but not both (Fig. 3c), and a small but significant decrease in the number of mixed calls made after trimming (the median percentage of mixed calls is 18.9% when using untrimmed data, and 14.1% when using trimmed data; Mann–Whitney U P <2.2×10^{-16}; Fig. 3d). Quantitatively similar results were again observed if expanding the analysis to incorporate 1100 publicly archived *S. aureus* samples (Tables S3 and S8, Fig. S4), with the exception that there was no significant decrease in the number of mixed calls made by trimming (the median percentage of mixed calls is 12.7% when using untrimmed data, and 12.6% when using trimmed data; Mann–Whitney U P=0.48; Fig. S4). Across the combined set of approximately 125 million SNPs from all three species, 98.8% of SNPs could be identically called irrespective of any read trimming.

Read trimming has slightly greater effect on indel calling compared to SNP calling

For the purpose of this study, my primary focus has been SNP calling, although the same pipelines simultaneously call indels. Read trimming appears to have a slightly greater effect on indel compared to SNP calling, although I interpret this data with caution. Across the combined set of approximately 1.25 million indels from the 1606 *E. coli*, 3946 *M. tuberculosis* and 1100 *S. aureus* samples, 91.9% of indels could be identically called irrespective of any read trimming (Tables S6, S7 and S8), a lower proportion than observed with SNPs. However, there are fewer indels per sample than there are SNPs (in some cases, as few as three), so it is important to note that measurements of the percentage of 'identically called indels' are inherently noisier. Furthermore, it is reasonable to believe that compared to SNP calls, a greater proportion of the indel calls are incorrect, irrespective of whether reads are trimmed or not. This reflects the technical difficulty in indel calling, which often necessitates dedicated algorithms (see performance reviews [26, 27]). In my dataset, the majority of indel calls are mixed (Fig. S5), suggesting a greater degree of ambiguity in real alignment around indels than SNPs, and thereby false-positive calling (across the three species, 93.1% of the total indel calls are mixed, even when using trimmed reads; Tables S6, S7 and S8). I conclude that broadly similar findings can be seen with indels as with SNPs – that trimming has little effect on the accuracy and completeness of a set of variant calls – although note that as indel calling is more technically challenging than SNP calling, it may be argued that any means of reducing error, including trimming, is of value.

Recommendations for read trimming prior to SNP calling

I found that read trimming, in general, had minimal effect on the performance of a BWA-MEM/mpileup SNP-calling pipeline (a finding corroborated with other pipelines; discussed below). This suggests that when analysing isolated bacterial strains, routine read trimming is not always a practical necessity. If the purpose of SNP calling is to construct large-scale bacterial phylogenies (for instance, to infer transmission), then trimming appears of little value given the majority of SNPs in a sample are identically called regardless of whether reads are trimmed or not. By contrast, if reducing the likelihood of even a small number of false-positive calls is essential (for instance, when predicting antimicrobial resistance) then trimming may yet prove critical: not all SNPs are equally important when it comes to drawing biological conclusions. Supporting this point, I found that after re-analysing a large number of publicly archived datasets there were small (approx. 1%) but significant decreases in the proportion of mixed SNP calls made when using trimmed compared to untrimmed data (Figs 2d and 3d), these considered a proxy for false-positive calls.

Further to this, I found that some SNPs were only called when using trimmed reads. It is possible these SNPs reside within complex regions of the genome more susceptible to error-prone alignment, and which trimming can help mitigate. Nevertheless, when considering the *E. coli* dataset (n=1606 samples), the number of SNPs only called using trimmed reads is low in absolute terms (<200 per sample for the majority of samples, collectively representing 83820 calls; Fig. 2c). This set of 83820 calls comprises 31164 unique SNPs, of which 24681 (79%) were called only in one sample (Table S6). This high proportion suggests that trimming does not for the most part help align reads around specific SNPs,

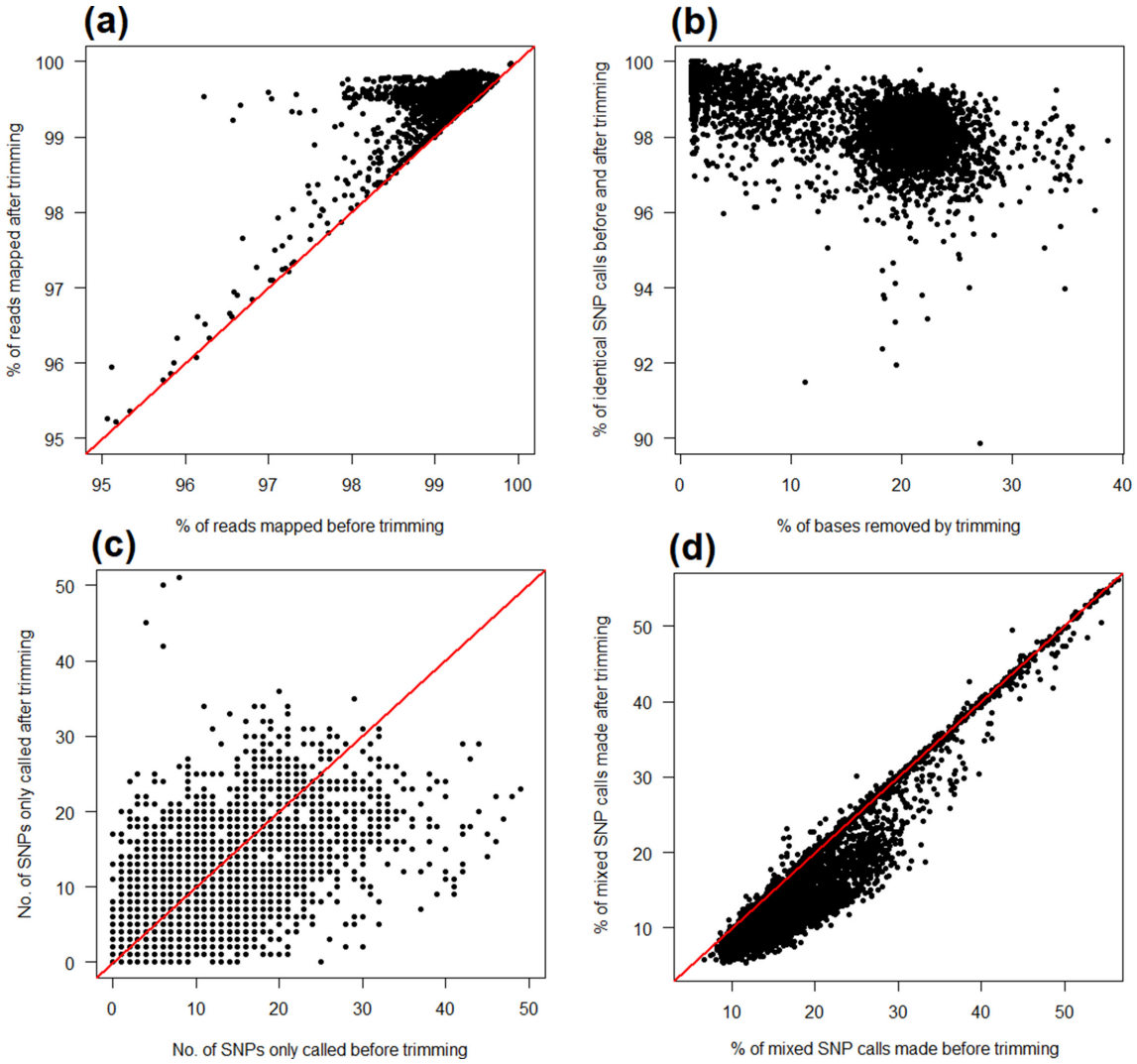

Fig. 3. Effect of read trimming upon SNP calls made using publicly archived *M. tuberculosis* sequencing data. This figure recapitulates patterns seen in Fig. 2 and illustrates the effect of read trimming upon SNP calls made in a clonal species, *M. tuberculosis*, for which relatively high alignment accuracy is expected and the impact of misalignment (i.e. false-positive SNP calls) accordingly greater. Trimming marginally increases the proportion of successfully aligned reads, albeit from a high baseline value, >98% (a). The vast majority of SNPs (>98%) are nevertheless called irrespective of any trimming (b). A relatively small number of SNPs (often <40) are only called when using either raw or trimmed data, but not both (c). The proportion of mixed SNP calls, considered a proxy of false-positive calling, decreases when using trimmed data (d). The raw data for this figure are available in Table S7 and represent 3946 *M. tuberculosis* samples, with a mean of 1238 SNPs per sample. The red lines denote *y=x*.

those found within regions more prone to alignment error (because if it did, SNPs only detectable using trimmed reads would be called across multiple samples). Nevertheless, some exceptions were apparent, including 5 SNPs within a 46 bp region (bases 607640–607686), that were each called in >100 samples, but only after trimming (Table S6).

Many of the justifications for routine read trimming relate to its simplicity: it is an easily performed procedure. In this respect, if it is only necessary to process a limited number of samples and computational resources are not at a premium, there appears little detriment to trimming – albeit, as I have shown, not necessarily much benefit either. However, if computational resources are at a premium, then various factors may be taken into consideration before attempting to trim reads. Firstly, when a trimmer is provided paired-end reads as input, it can produce both paired- and single-end reads as output, the latter when only one end of a pair is discarded. Directing paired and unpaired reads to separate sets of output files is necessary as, in general, a read aligner, if aligning paired-end reads, requires an identical number of reads in each input FASTQ. However, as numerous aligners do not allow the simultaneous provision of paired-end and unpaired input (with some exceptions, such as HISAT2 [28]), any subsequent alignment step would need to be run

twice and the output BAM files merged, an additional (and, therefore, time-consuming) set of operations. The default parameters of FASTP and TrimGalore are to output only the trimmed paired-end reads which, although convenient, does discard a small proportion of otherwise useable single-end data. I found, however, that these single-end reads were often few in number and negligibly informative, and so in many cases could reasonably be dispensed with. I re-processed the 1606 *E. coli* samples, discarding those reads unpaired after trimming, and found negligible difference in the total number of SNPs called using only paired-end reads and using all reads (the median number of SNPs called when only using paired-end reads and when using all reads are 55137 and 55147, respectively; Mann–Whitney U P=0.871; Fig. S6). There was also a small decrease in the number of mixed calls made when discarding the unpaired reads, although this difference was not significant (the median percentages of mixed calls made when retaining and discarding unpaired reads were 9.03 and 8.51%, respectively; Mann–Whitney U P=0.055) (Fig. S6).

Secondly, one needs to consider the computational cost of pre-processing. If calling SNPs for a large number of samples, this cost may be weighed negatively against the relative speed and simplicity of post-processing a VCF record – for instance, by masking repetitive regions and applying positional filters (both quick operations), SNP-calling precision can easily be increased [22]. I reasoned that as trimming had minimal effect on the overall number of SNP calls made, then if it was to be performed, it should at least be done quickly. While no in-depth assessment of runtime was made in this study, my experience was that there was very little reason to use a comparatively slow trimmer, and so of those considered here, FASTP, previously benchmarked as up to 5× faster than Trimmomatic and Cutadapt [10] (and with a richer feature set), was greatly preferred. When processing the Gram-negative dataset using a BWA-MEM/mpileup pipeline, I found that trimming with FASTP increased the total runtime by either 5 or 27 min, should reads unpaired by the trimming process be discarded or retained, respectively (Fig. S7).

While trimming extends the runtime of a given pipeline, it could conceivably 'save' downstream memory as a consequence of there being fewer reads to align. However, in practice, I did not find a discernible impact. For example, the memory usage (more specifically, maximum resident set size) of BWA-MEM when aligning the untrimmed and FASTP-trimmed reads (n=1591328 and 1509306 pairs, respectively) from the *K. pneumoniae* reference strain MGH78578 was reduced from 114.8 to 113.1 Mb, a negligible difference.

In terms of user convenience, both FASTP and TrimGalore also automatically detect adapter sequence from the reads themselves, whereas Atropos and Trimmomatic require that the adapter be specified, assuming it is known (while Atropos can also predict which adapter sequence is present in a set of reads, this function is included as a separate module; unlike TrimGalore and FASTP, Atropos does not perform adapter detection and removal simultaneously). Other FASTQ pre-processing procedures are also in routine use, such as sequencing error correction [29–32] and the depletion of human contaminants [33], and although the impact these have upon SNP calling is beyond the scope of this study, I cannot exclude the possibility that they confer greater benefit than simply trimming reads. As such, despite the results of the present study, I cannot simply recommend that prior to SNP calling no FASTQ pre-processing operations be performed at all.

Finally, one needs to consider an appropriate set of trimming criteria. It was not my intention to exhaustively test all possible ways to trim a read and so I restricted analysis to 3′ quality-trimming across a range of stringencies, by far the most commonly applied operation (because Illumina reads degrade in quality towards the 3′ end). It was found that the greatest apparent benefit to SNP-calling performance (evaluated as an increase in F-score) was when using FASTP to trim 3′ bases at Phred score thresholds of 20 or lower (as illustrated in Fig. 1), although it should be re-iterated that the difference in F-scores when SNP calling using untrimmed data, and data trimmed using these approaches, was statistically insignificant.

It is also important to note while the four read trimmers used in this study share the same core functionality of adapter- and quality-trimming, they differ in several respects, having different feature sets and, in the case of FASTP, implementing several additional filters by default (which I did not explicitly disable). Irrespective of additional 3′ quality-trimming, by default, FASTP automatically detects and remove adapters, discards reads with >5 N bases, performs polyG trimming (should it detect NextSeq or NovaSeq input), and requires a minimum 'qualified quantity' per read, i.e. that >40% of the bases in each read have Phred >15 (disabling the qualified quantity filter renders the performance of FASTP similar to other trimmers, especially in terms of precision; see Fig. S3). That more filters are applied by default explains why FASTP is the only trimmer for which recall (and thereby F-score) was found to increase after trimming (Fig. S2). By contrast, the other three trimmers, in only performing 3′ trimming, more directly affect an improvement to precision (Fig. S1), although as this occurs at the expense of recall, the overall effect on F-score appears detrimental (note, however, that in absolute terms, these differences were statistically insignificant; Fig. 1).

For any given trimming program, it is possible to apply the same set of filters, as well as to add others, such as trimming leading as well as trailing bases or to clip reads should the mean quality within a sliding window advanced from the 5′ to 3′ end fall below a minimum threshold (a feature of Trimmomatic and FASTP, but not TrimGalore or Atropos). However, although drawing from a richer set of filters seems intuitively superior, in practice they are unlikely to make a substantive difference to the set of SNPs called. I have already shown that even after removing >20% of the (lower-quality) bases from *E. coli*, *M. tuberculosis* and *S. aureus* reads, essentially the same variant bases are still called. It is reasonable to believe that extra filters would for the most part act as complementary methods of removing

the same set of lower-quality bases. By way of illustration, I re-calculated F-score for the set of Gram-negative bacteria after pre-processing each sample using FASTP with four parameters in addition to those used previously (the default settings, plus 3′ quality-trimming using a threshold of Q20; see Table S5). Compared to the previous use of FASTP (i.e. to perform adapter-trimming, 3′ quality trimming, and to require a minimum read length of 50 bp and a qualified quantity of bases), the addition of four extra parameters, including base correction and a 'low complexity' filter, made no significant difference to overall SNP-calling performance (the correlation between F-scores when using the two sets of filters was Spearman's *rho*=0.999, $P < 2.2\times10^{-16}$; Fig. S8). It is important to note that the pipeline used to process the Gram-negative data applied numerous VCF filtering criteria to reduce error (see Methods). These post-alignment operations could conceivably have a similar effect on performance as pre-alignment read trimming, resulting in similar F-scores for the trimmed and untrimmed data. To account for this, I also applied this expanded set of FASTP parameters to the set of 1606 *E. coli* samples (Table S6). For this analysis, I directly compared the contents of the untrimmed and trimmed VCF records, without any VCF filtering (see Methods). Consistent with the above findings, I found no significant difference in the percentage of SNPs identically called before and after trimming, although there was a significant – and pronounced – reduction in the percentage of mixed calls made (Fig. S9). This suggests that when reducing SNP-calling error, it is not necessarily 3′ trimming that adds the greatest value, but a combined set of read pre-processing operations – which, in Fig. S9, include the base correction function of FASTP.

For simplicity, I have used trimming to refer to all functions performed by a read pre-processing tool, which for convenience I call a read trimmer. However, it is important to note that many trimmers are multipurpose and simultaneously clip (cut reads at a point other than their end), trim (remove bases from the ends) and perform other functions. This latter category of functions can include procedures for 'read scrubbing' [34], the removal or revision of low-quality segments from a read. The base correction feature of FASTP could be considered an example of this.

These functions are important because SNP calling is not always performed on isolated strains, and the absolute number of true SNP calls is not necessarily the most critical user requirement. Multiple strains of a particular species may be sequenced simultaneously (for instance, from patient blood cultures), and if the purpose of doing so is to assess the differences between them, significantly reducing even a small number of mixed calls – which these functions facilitate – may be beneficial. For the most part, however, the data suggests that read trimming (used in the general sense to mean trimming, clipping, scrubbing or any other pre-processing) makes minimal difference to the set of SNPs called, and to whether they are considered homozygous or heterozygous.

In general, overly conservative trimming parameters may also prove counterproductive. Especially stringent quality-trimming (for instance, Q <30) could artificially reduce coverage depth by discarding a larger proportion of reads, as well as shortening many more. In the absence of an additional filter on the basis of minimum read length, reads that are too short are also more likely to be misaligned. This is consistent with a previous study that explored the effects of RNA-seq read trimming upon gene expression estimates, finding that the majority of differences were driven by the spurious mapping of short reads, which could be mitigated by requiring a minimum read length [35]. The default minimum read lengths for FASTP and TrimGalore are 15 and 20 bp, respectively, both relatively short given current Illumina read lengths (>300 bp). In this study, I consistently required a minimum read length of 50 bp, one third of the length of the shortest raw read in any sample (150 bp). It would in principle be possible to use an aligner optimized for comparatively short reads, such as Bowtie [36], to perhaps more accurately map those truncated by stringent trimming, although far more pragmatic to simply abstain from such stringency to begin with. As Fig. 1 illustrates, overall SNP-calling performance can actually decrease when using overly conservative filters: relative to untrimmed data, F-score is notably reduced when using Atropos and TrimGalore with Q thresholds of 30, a consequence of marginally improving precision (Fig. S1) at the more substantial expense of recall (Fig. S2). Consistent with this, a previous study has advocated a more gentle trimming strategy (Q <5) as optimal across a range of metrics [8]. At least relative to especially stringent trimming, this conclusion appears supported by Fig. 1 – essentially the same distribution of F-scores can be seen for Q <5 as for Q <20. However, this is to not to say that, prior to SNP calling, gentle trimming is necessarily superior to no trimming at all, as in my initial Gram-negative dataset the absolute difference in F-score for untrimmed reads relative to reads trimmed using Q <5 was statistically insignificant (Mann–Whitney U *P*=0.98). However, when re-analysing the larger set of publicly archived *E. coli* reads, I found that the percentage of SNPs identically called before and after trimming was fractionally higher when trimming at Q <5 compared to Q <20 (median percentage 99.91 and 99.87%, respectively; Mann–Whitney U $P < 2.2\times10^{-16}$; Table S6). This suggests that more rigorous trimming does allow a small number of additional calls to be made – although whether these additional calls represent true SNPs not detectable using untrimmed or lightly trimmed data, or simply additional errors, cannot be determined. Regardless, the percentage of mixed calls was not significantly different between the Q <5 and Q <20 data (median percentage 9.02 and 9.05%, respectively; Mann–Whitney U *P*=0.96).

A final consideration is the choice of pipeline used to call SNPs. The data generated in this study employed the widely used BWA aligner, which 'soft clips' reads, in essence incorporating its own trimming step into the process of alignment (in the resulting BAM file, soft clips are seen as mismatch states in the alignment, comprising a contiguous section of the end of a read [37]). This approach is in contrast to end-to-end aligners, such as Bowtie2 [38], which by attempting to align every base of a read should be more greatly affected by trimming (as

by removing lower-quality bases, the read is more likely to successfully align). To demonstrate this, I re-analysed a subset (n=200) of the 1606 *E. coli* samples, substituting BWA for one of two alternative aligners, NextGenMap v0.5.5 [39] (which soft clips) and Bowtie2 v2.3.4.1 [38] (which does not), each used with default parameters. I found that, as with the BWA-based pipeline, the majority of SNPs were identically called irrespective of whether trimmed or untrimmed reads were used (>99% in the majority of samples, using either aligner; Fig. S10, Table S6). However, with Bowtie2, there appears to be a sharper decline in the percentage of identically called SNPs when a higher number of bases have been trimmed (Fig. S10). This suggests that trimming reads has greater effect upon SNP calling if the aligner does not perform soft clipping, as lower-quality samples (those with a higher proportion of bases trimmed) are more strongly affected. As such, a pragmatic recommendation is to use an aligner which performs on-the-fly trimming, or to trim reads if using an aligner which does not (such as Stampy [40]). There are many aligners in the former category, and so a dedicated trimming step is arguably redundant in many SNP-calling workflows. This is particularly apparent with more recent programs, and with long-read sequencing technologies. Many long-read workflows do not explicitly require trimmed input, for example in an evaluation of structural-variant-calling pipelines using Nanopore reads [41], or with the Shasta toolkit for *de novo* Nanopore read assembly [42]. Nevertheless, long-read trimmers are available, such as the NanoFilt module of NanoPack [43].

Many 'general purpose' pre-processing tools perform similar functions to those used in this study, including AdapterRemoval v2 [44], AfterQC [45], AlienTrimmer [46], Btrim [47], fastQ_brew [48], FastqPuri [49], ngsShoRT [50], PEAT [51], SeqPurge [52], SeqTrim [53], Skewer [54] and SOAPnuke [55], alongside more protocol-specific tools such as NxTrim [56] and NextClip [57], both of which were designed to remove adaptors from Illumina Nextera mate pairs. I did not seek to evaluate a comprehensive range of tools because my primary concern was not to identify the highest-performing read trimmer per se, but to explore the effect of trimming, in general, upon bacterial SNP-calling accuracy. I anticipate my findings would be generalizable to a broad range of read trimmers, as these essentially share the same core functionality although differ in runtime and memory use.

Conclusions

A simple means of improving any particular SNP-calling pipeline is to remove minimal-value operations, as this decreases the computational time and data manipulation required. The benefit of various operations may not be universally realized and so in many situations could prove an unnecessary computational expense. This has previously been demonstrated for several post-alignment processing steps, such as local indel realignment and base quality score recalibration, when calling variants from exome sequencing data [58]. A previous study also demonstrated that PCR duplicate removal had minimal effect on SNP calling and questioned its necessity as a routine procedure [59] (it has not escaped my notice that I duplicate-mask in my own pipelines; this perhaps reflects the ease with which such analytic habits become ingrained).

By re-analysing >6500 publicly archived sequencing datasets from *E. coli*, *M. tuberculosis* and *S. aureus*, I found that the retention of lower-quality bases and residual adapter contaminants had minimal effect upon SNP calling. Of the approximately 125 million SNPs called across all samples, 98.8% were identically called irrespective of whether raw reads or trimmed reads were used. This suggests that quality- and adapter-trimming, although routine FASTQ pre-processing operations, add relatively little value to a SNP-calling pipeline and may only be necessary if small differences in the absolute number of SNP calls are critical (such as, for instance, when predicting antimicrobial resistance).

As such, given the majority of read trimmers perform the same basic functions with comparable accuracy, there seems little practical reason to make use of any other than the fastest (of those in this study, FASTP [10]), if at all. My results also suggest that if pre-processing is performed then, in terms of optimal parameters, I would consider not being too conservative (trimming at a quality threshold of 20 or lower), not retaining reads unpaired as a consequence of trimming, and not relying on 3′ trimming alone (as shown with FASTP, there appeared greater performance when supplementing 3′ trimming with a minimum qualified quantity per read).

Read trimming remains routinely performed prior to SNP calling, likely out of concern that doing otherwise would typically have negative consequences. While historically this may have been the case, the data presented here suggests that read trimming is not always a practical necessity.

Funding information
This study was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) (grant HPRU-2012–10041) and supported by the NIHR Oxford Biomedical Research Centre. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The report presents independent research funded by the NIHR. The views expressed in this publication are those of the author and not necessarily those of the NHS (National Health Service), the NIHR, the Department of Health nor PHE.

Author contributions
S. J. B. conceived of and designed the study, performed all analyses and wrote the manuscript.

Conflicts of interest
The author declares that there are no conflicts of interest.

References

1. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND *et al*. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb Genom* 2019;5:e000294.

2. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N *et al*. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. *Gigascience* 2020;9:giaa007.
3. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One* 2013;8:e85024.
4. Farrer RA, Henk DA, MacLean D, Studholme DJ, Fisher MC. Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. *Sci Rep* 2013;3:1512.
5. Liu Q, Guo Y, Li J, Long J, Zhang B *et al*. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics* 2012;13:S8.
6. Pightling AW, Petronella N, Pagotto F. Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses. *PLoS One* 2014;9:e104579.
7. Yang S-F, Lu C-W, Yao C-T, Hung C-M. To TRIM or not to TRIM: effects of read trimming on the de novo genome assembly of a widespread East Asian passerine, the Rufous-Capped Babbler (Cyanoderma ruficeps Blyth). *Genes* 2019;10:737.
8. MacManes MD. On the optimal trimming of high-throughput mRNA sequence data. *Front Genet* 2014;5:13.
9. Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C *et al*. Ten steps to get started in genome assembly and annotation. *F1000Res* 2018;7:ELIXIR-148
10. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–i890.
11. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
12. Didion JP, Martin M, Collins FS. Atropos: specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* 2017;5:e3720.
13. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;17:10.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
15. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH *et al*. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 2012;40:11189–11201.
16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
17. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E *et al*. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* 2018;15:591–594.
18. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL *et al*. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol* 2018;14:e1005944.
19. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 2014;15:524.
20. Broad Institute. Picard: a set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF (http://broadinstitute.github.io/picard/). Cambridge, MA: Broad Institute; 2018.
21. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A *et al*. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *N Engl J Med* 2013;369:1195–1205.
22. Jia P, Li F, Xia J, Chen H, Ji H *et al*. Consensus rules in variant detection from next-generation sequencing data. *PLoS One* 2012;7:e38470.
23. Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P *et al*. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*. *Nat Med* 2016;22:1470–1474.
24. Guthrie JL, Delli Pizzi A, Roth D, Kong C, Jorgensen D *et al*. Genotyping and whole-genome sequencing to identify tuberculosis transmission to pediatric patients in British Columbia, Canada, 2005–2014. *J Infect Dis* 2018;218:1155–1163.
25. Ceric O, Tyson GH, Goodman LB, Mitchell PK, Zhang Y *et al*. Enhancing the one health initiative by using whole genome sequencing to monitor antimicrobial resistance of animal pathogens: Vet-LIRN collaborative project with veterinary diagnostic laboratories in United States and Canada. *BMC Vet Res* 2019;15:130.
26. Hasan MS, Wu X, Zhang L. Performance evaluation of indel calling tools using real short-read data. *Hum Genomics* 2015;9:20.
27. Li D, Kim W, Wang L, Yoon KA, Park B *et al*. Comparison of indel calling tools with simulation data and real short-read data. *IEEE/ACM Trans Comput Biol Bioinform* 2019;16:1635–1644.
28. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;37:907–915.
29. Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol* 2010;11:R116.
30. Marçais G, Yorke JA, Zimin A. QuorUM: an error corrector for Illumina reads. *PLoS One* 2015;10:e0130821.
31. Greenfield P, Duesing K, Papanicolaou A, Bauer DC. Blue: correcting sequencing errors using consensus and context. *Bioinformatics* 2014;30:2723–2732.
32. Yang X, Chockalingam SP, Aluru S. A survey of error-correction methods for next-generation sequencing. *Brief Bioinform* 2013;14:56–66.
33. Bush SJ, Connor TR, Peto TEA, Crook DW, Walker AS. Evaluation of methods for detecting human reads in microbial sequencing datasets. *Microb Genom* 2020;6:e000393.
34. LaPierre N, Egan R, Wang W, Wang Z. De novo Nanopore read quality improvement using deep learning. *BMC Bioinformatics* 2019;20:552.
35. Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* 2016;17:103.
36. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25.
37. Schröder J, Hsu A, Boyle SE, Macintyre G, Cmero M *et al*. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics* 2014;30:1064–1072.
38. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
39. Sedlazeck FJ, Rescheneder P, von Haeseler A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* 2013;29:2790–2791.
40. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 2011;21:936–939.
41. Zhou A, Lin T, Xing J. Evaluating nanopore sequencing data processing pipelines for structural variation identification. *Genome Biol* 2019;20:237.
42. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE *et al*. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* 2020;38:1044–1053.
43. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;34:2666–2669.
44. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes* 2016;9:88.
45. Chen S, Huang T, Zhou Y, Han Y, Xu M *et al*. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* 2017;18:80.

46. Criscuolo A, Brisse S. AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* 2013;102:500–506.
47. Kong Y. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 2011;98:152–153.
48. O'Halloran DM. fastQ_brew: module for analysis, preprocessing, and reformatting of FASTQ sequence data. *BMC Res Notes* 2017;10:275.
49. Pérez-Rubio P, Lottaz C, Engelmann JC. FastqPuri: high-performance preprocessing of RNA-Seq data. *BMC Bioinformatics* 2019;20:226.
50. Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code Biol Med* 2014;9:8.
51. Li Y-L, Weng J-C, Hsiao C-C, Chou M-T, Tseng C-W *et al*. PEAT: an intelligent and efficient paired-end sequencing adapter trimming algorithm. *BMC Bioinformatics* 2015;16 (Suppl. 1):S2.
52. Sturm M, Schroeder C, Bauer P. SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. *BMC Bioinformatics* 2016;17:208.
53. Falgueras J, Lara AJ, Fernández-Pozo N, Cantón FR, Pérez-Trabado G *et al*. SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics* 2010;11:38.
54. Jiang H, Lei R, Ding SW, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 2014;15:182.
55. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* 2018;7:gix120.
56. O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA *et al*. NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* 2015;31:2035–2037.
57. Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. NextClip: an analysis and read preparation tool for Nextera long mate pair libraries. *Bioinformatics* 2014;30:566–568.
58. Tian S, Yan H, Kalmbach M, Slager SL. Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics* 2016;17:403.
59. Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B *et al*. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics* 2016;17:239.