MICROBIAL GENOMICS

RESEARCH ARTICLE

Alkam *et al.*, *Microbial Genomics* 2021;7:000655

DOI 10.1099/mgen.0.000655

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Is amplification bias consequential in transposon sequencing (TnSeq) assays? A case study with a *Staphylococcus aureus* TnSeq library subjected to PCR-based and amplification-free enrichment methods

Duah Alkam[1], Thidathip Wongsurawat[2,3], Intawat Nookaew[3], Anthony R. Richardson[4], David Ussery[3], Mark S. Smeltzer[5] and Piroon Jenjaroenpun[2,3,*]

Abstract

As transposon sequencing (TnSeq) assays have become prolific in the microbiology field, it is of interest to scrutinize their potential drawbacks. TnSeq data consist of millions of nucleotide sequence reads that are generated by PCR amplification of transposon-genomic junctions. Reads mapping to the junctions are enumerated thus providing information on the number of transposon insertion mutations in each individual gene. Here we explore the possibility that PCR amplification of transposon insertions in a TnSeq library skews the results by introducing bias into the detection and/or enumeration of insertions. We compared the detection and frequency of mapped insertions when altering the number of PCR cycles, and when including a nested PCR, in the enrichment step. Additionally, we present nCATRAs - a novel, amplification-free TnSeq method where the insertions are enriched via CRISPR/Cas9-targeted transposon cleavage and subsequent Oxford Nanopore MinION sequencing. nCATRAs achieved 54 and 23% enrichment of the transposons and transposon-genomic junctions, respectively, over background genomic DNA. These PCR-based and PCR-free experiments demonstrate that, overall, PCR amplification does not significantly bias the results of TnSeq insofar as insertions in the majority of genes represented in our library were similarly detected regardless of PCR cycle number and whether or not PCR amplification was employed. However, the detection of a small subset of genes which had been previously described as essential is sensitive to the number of PCR cycles. We conclude that PCR-based enrichment of transposon insertions in a TnSeq assay is reliable, but researchers interested in profiling putative essential genes should carefully weigh the number of amplification cycles employed in their library preparation protocols. In addition, nCATRAs is comparable to traditional PCR-based methods (Kendall's correlation=0.896–0.897) although the latter remain superior owing to their accessibility and high sequencing depth.

Received 22 December 2020; Accepted 19 July 2021; Published 01 October 2021

Author affiliations: [1]Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, AR, USA; [2]Division of Bioinformatics and Data Management for Research, Research Group and Research Network Division, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand; [3]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA; [4]Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA; [5]Department of Microbiology and Immunology, University of Arkansas for Medical Sciences, Little Rock, AR, USA.

***Correspondence:** Piroon Jenjaroenpun, piroon.jen@mahidol.edu

Keywords: PCR-bias; PCR-free; TnSeq; transposon; Nanopore; nCATS; Cas9.

Abbreviations: CDS, coding sequence; CIP, calf intestinal alkaline phosphatase; CRISPR, clustered regularly interspaced short palindromic repeats; crRNA, CRISPR RNA; EDTA, ethylenediaminetetraacetic acid; gDNA, genomic DNA; GMM, gaussian mixture models; gRNA, guide RNA; IR, transposon inverted repeat; nCATRAs, nanopore Cas9-targeted transposon sequencing; NESTED, TnSeq library prepared by amplifying the transposon-genomic junctions via 25 PCR cycles followed by a nested reaction of 15 PCR cycles; OD, optical density; ONT, Oxford Nanopore Technologies; PAM, protospacer-adjacent motif; PBS, Phosphate-buffered Saline; PCR15, TnSeq library prepared by amplifying the transposon-genomic junctions via 15 PCR cycles; PCR25, TnSeq library prepared by amplifying the transposon-genomic junctions via 25 PCR cycles; PCR, Polymerase Chain Reaction; qRT-PCR, quantitative real-time polymerase chain reaction; ROI, region of interest; SNP, single nucleotide polymorphism; TA, thymine-adenine dinucleotide; TnSeq, transposon sequencing; TSB, tryptic soy broth; WGS, whole-genome sequencing.

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

DATA SUMMARY

Raw sequencing (fastq) files and processed (BedGraph) files detailing location and counts of mapped insertions are deposited in GEO (series entry: GSE162988; reviewer access token: cxeziussrdifvyr) and SRA (BioProject: PRJNA639565). The analysis pipeline scripts of TnSeq and nCATRAs were deposited in GitLab (https://gitlab.com/piroonj/ncatras_pipeline). The *Staphylococcus aureus* USA300 LAC genome is deposited in GenBank (CP055225: chromosome and CP055226: cryptic plasmid).

Impact Statement

With the ever-increasing accessibility to nucleotide sequencing data, the application of whole-genome screens which provide unprecedented insight into microorganisms has become routine in microbiology laboratories. Genome-wide transposon sequencing (TnSeq) is one such screen which profiles almost every non-essential gene in an organism. The prolific use of TnSeq has deepened our understanding of the intricacies of many microorganisms. Here we present the first interrogation into the existence and repercussions of the hypothetical PCR amplification bias in TnSeq. Additionally, we present, for the first time, a novel amplification-free TnSeq method - termed nCATRAs. With nCATRAs, we are now able to examine TnSeq libraries in their most native form, that is, without any artificial enrichment, paving the way for new studies of the intricacies of transposition events. The analyses and conclusions described in this work may guide investigators in implementing TnSeq with the utmost rigour, and may help maximize the information distilled from this assay.

INTRODUCTION

Transposon sequencing (TnSeq) is a genome-wide, high-throughput screen for genes contributing to bacterial fitness [1–4]. TnSeq couples genome-wide transposon mutagenesis with high-throughput deep sequencing of the transposon-genomic junctions (referred to herein as 'transposon insertions' or 'insertions') to identify genes which compromise bacterial fitness under a given experimental growth condition [5]. Several groups have successfully applied this tool to define putative essential genes of different species [5–9]. For bacterial pathogens, TnSeq has been used to characterize genes required for establishing successful infections [10, 11]. The TnSeq assay is conducted with a library consisting of a pool of cells mutagenized by transposon insertions. Each cell in the library is assumed to have one transposon insertion and the pool consists of thousands of transposon insertions at unique locations across the genome. The higher the number of distinct insertions, the more saturated the library. A highly saturated library increases the power of the assay as most non-essential genes would be represented with at least one transposon insertion and the contribution of all these genes to fitness in a given environment can be assessed. Transposon insertions are enriched by PCR amplification and sequenced before and after applying experimental selective pressure. The number of reads mapping to each insertion is then tallied and genes with significantly altered insertion frequency are prioritized and considered important for the microorganism's fitness in the experimental condition [5, 9].

As TnSeq has become a prolific tool in the microbiology field, it is worth dissecting its potential drawbacks and attempting to address them. The ultimate strength of this assay is in distilling genes whose disruption is most consequential to the organism from a diverse pool of thousands of disrupted genes. Thus, preserving the complexity of the TnSeq library by minimizing introduction of unintended experimental bias is important to achieving success with this approach. Upon examining the protocol for preparing the TnSeq library for next-generation sequencing, we wondered whether bias could be introduced during the enrichment of transposon insertions by PCR cycle number. We reasoned that amplification may contribute to this bias in light of the established issue of PCR jackpots, where certain regions are over-amplified not because they are present at a higher level in the sample but due to inherent properties of the PCR reaction [12–15]. To our knowledge, all TnSeq reports to date have enriched for the transposon insertions via PCR amplification. However, we have found no discussion on the potentially deleterious effects of PCR amplification bias on the results of a TnSeq assay.

Here, we investigated the extent to which PCR amplification contributes to skewing the results of the TnSeq assay using a previously-validated *S. aureus* library generated in the USA300 LAC background [10]. We began with sequencing and reconstructing the library parent strain, USA300 LAC, to produce a high-quality reference genome that is crucial for analyses. Next, we reasoned that bias introduced during PCR amplification would manifest in altered frequency of mapped transposon insertions upon changing the number of PCR cycles. To examine this, we enriched for transposon insertions using three methods which varied in the number of PCR amplification cycles. We then sequenced using next-generation sequencing and compared the number of mapped insertions resulting from each method. Further, we wondered whether eliminating the PCR amplification step would dramatically change the output of the TnSeq assay. To address this, we enriched for the transposon insertions via a PCR-free method, termed nanopore Cas9-targeted transposon sequencing or nCATRAs, and sequenced using Oxford Nanopore Technologies (ONT) - a protocol adapted from a recently published targeted enrichment method [16, 17].

METHODS

Bacterial strains and growth conditions

The USA300 LAC *S. aureus* strain was obtained from the laboratory of Dr. Anthony R. Richardson, thus ensuring that comparisons were made with the identical strain used

to generate the TnSeq library [10]. This strain was cultured in 5 ml Tryptic Soy Broth (TSB) overnight in 37 °C with continuous agitation and aeration prior to genomic DNA extraction [18]. The TnSeq library was previously generated in the USA300 LAC *S. aureus* strain as described by Grosser *et al.* [10]. All TnSeq libraries were cultured for in 37 °C with continuous agitation in 50 ml TSB with antibiotic selection for the transposon (erythromycin 5 µg ml^{-1}). Exponential phase TnSeq libraries were sampled at OD_{600}=1 (~10^8 colony forming units ml^{-1}), and stationary phase libraries were sampled from overnight cultures (grown for ~16 h) that were standardized to OD_{600}=10 (~10^9 colony forming units ml^{-1}). Cells were harvested by centrifugation, the pellets resuspended in TSB containing 25% glycerol and frozen for later extraction of genomic DNA.

Sequencing the USA300 LAC *S. aureus*

Genomic DNA was extracted using the NucleoBond HMW DNA kit (MACHEREY-NAGEL Inc., PA, USA) according to manufacturer's protocol with the additional step of incubating the previously frozen bacterial pellets (following thawing, a wash with PBS and centrifugation) in 560 µl Tris-EDTA (TE, pH 8.0) containing lysostaphin (1 µg $µl^{-1}$) prior to the enzymatic lysis step. Genomic DNA (gDNA) was sequenced by Psomagen, Inc. (Rockville, MD) using Illumina HiSeq paired-end with a read length of 100bps. gDNA from the same sample was also sequenced with Oxford Nanopore Technologies FLONGLE Flongle (ONT, United Kingdom) using the Rapid Sequencing Kit (SOK-RAD004) according to the manufacturer's protocol. Sequencing reads obtained from both platforms were used to create a *de novo* assembly of the complete genome, as described below in the 'Bioinformatic analyses' section.

Preparation of TnSeq Next-Generation sequencing libraries

Bacterial genomic DNA was extracted from cell pellets as described above. Extracted gDNA was then quantified and 45 µg from each sample was diluted in 150 µl Tris-EDTA (TE) before shearing to a peak fragment size of ~250 bp using a Covaris ME220 Focused-Ultrasonicator (Covaris, Woburn, MA). Next, 1 µg of sheared gDNA was subjected to the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA) protocol with key modifications to the PCR amplification step. Specifically, in the above-described PCR15 and PCR25 methods the olj511 forward primer specific to the transposon inverted repeat (IR) [10] and the NEBNext Indexed reverse primers obtained from NEBNext Multiplex Oligos for Illumina were used to amplify the transposon-genomic junctions and add the p5 and p7 illumina adaptors as well as a unique barcode to each sample by implementing either 15 or 25 PCR cycles respectively. The cycling conditions for this PCR were 98 °C for 30 s, 98 °C for 10 s followed by 65 °C for 75 s for 15 or 25 cycles, and 65 °C for 5 min. In the NESTED method, the junctions were amplified via a nested PCR consisting of two reactions, similarly to previous reports [10]. Here, the olj510 forward primer [10] specific to the transposon but upstream to the inverted repeat region and the NEBNext Indexed reverse primers obtained from NEBNext Multiplex Oligos for Illumina were used in the first PCR reaction. The cycling conditions for this PCR were 98 °C for 30 s, 98 °C for 10 s followed by 65 °C for 75 s for 25 cycles, and 65 °C for 5 min. In the second PCR reaction, the forward olj511 inverted region-specific primer and the same reverse primers used in the previous reaction were used to enhance the specificity towards the transposon and add the p7 illumina adaptor via 15 PCR cycles. The cycling conditions for this reaction were 98 °C for 30 s, 98 °C for 10 s followed by 65 °C for 75 s for 15 cycles, and 65 °C for 5 min. The quality of the libraries was then assessed via Agilent TapeStation (Santa Clara, CA) and quantitative real-time PCR (qRT-PCR) using the olj512 primer [10] specific to the transposon terminal inverted repeat. The validated libraries were then sequenced with the custom sequencing primer olj512 using an Illumina HiSeq 2500, Single-End 100, at the TUFTS University Core Facility. Each flow cell was loaded with six total multiplexed libraries. Thus, six samples prepared by each of the PCR methods (PCR15, PCR25 and NESTED) were loaded onto a separate flow cell. For the PCR15 libraries (three exponential and three stationary), 9 nM of DNA was loaded onto one lane of the flow cell. For PCR25 (three exponential and three stationary), 10 nM of DNA was loaded onto one lane of the flow cell. For NESTED (three exponential and three stationary), 10 nM of DNA of the libraries were loaded onto one lane of the flow cell. Thus, all six samples for a single method were loaded onto the same lane, but separate lanes (totaling three) were used for the different methods.

Nanopore Cas9-targeted transposon sequencing (nCATRAs)

The previously published nCATS protocol was modified in this paper to enrich for the transposon-genomic junctions such that only one end of the region of interest was cleaved as opposed to excision of the region of interest as described in the nCATS protocol [16]. In our modified nCATRAs method, gDNA extracted from a TnSeq library grown to stationary phase was sheared using Covaris g-tubes to 6000 bp, which resulted in DNA fragments between 6kb to 10kb in size. Of this sheared gDNA, 5 µg were used as starting material for each of three independent runs of this experiment. Each run was carried out as follows, two tubes each containing 6 µg of the fragmented gDNA were separately dephosphorylated using 3 µl Quick CIP (New England Biolabs, Ipswich, MA) and incubated at 37 °C for 15 mins. Two crRNA probes (5′-GTTATCTATTATTTAACGGG-3′); (5′-AGGATTCTA-CAAGCGTACCT-3′) targeting two sites on the 5′ end of the transposon were pooled to a final concentration of 1 µM and used to direct Cas9 (HiFi Cas9, Integrated DNA Technologies, Coralville, IA) to the region. Cas9-cleaved DNA fragments were ligated to the Oxford Nanopore adaptors and the Ligation Sequencing Kit (LSK-109) protocol was applied according to the manufacturer's protocol. The two starting tubes were pooled in the final elution step prior to loading onto the MinION flow cell. One MinION flow cell was used

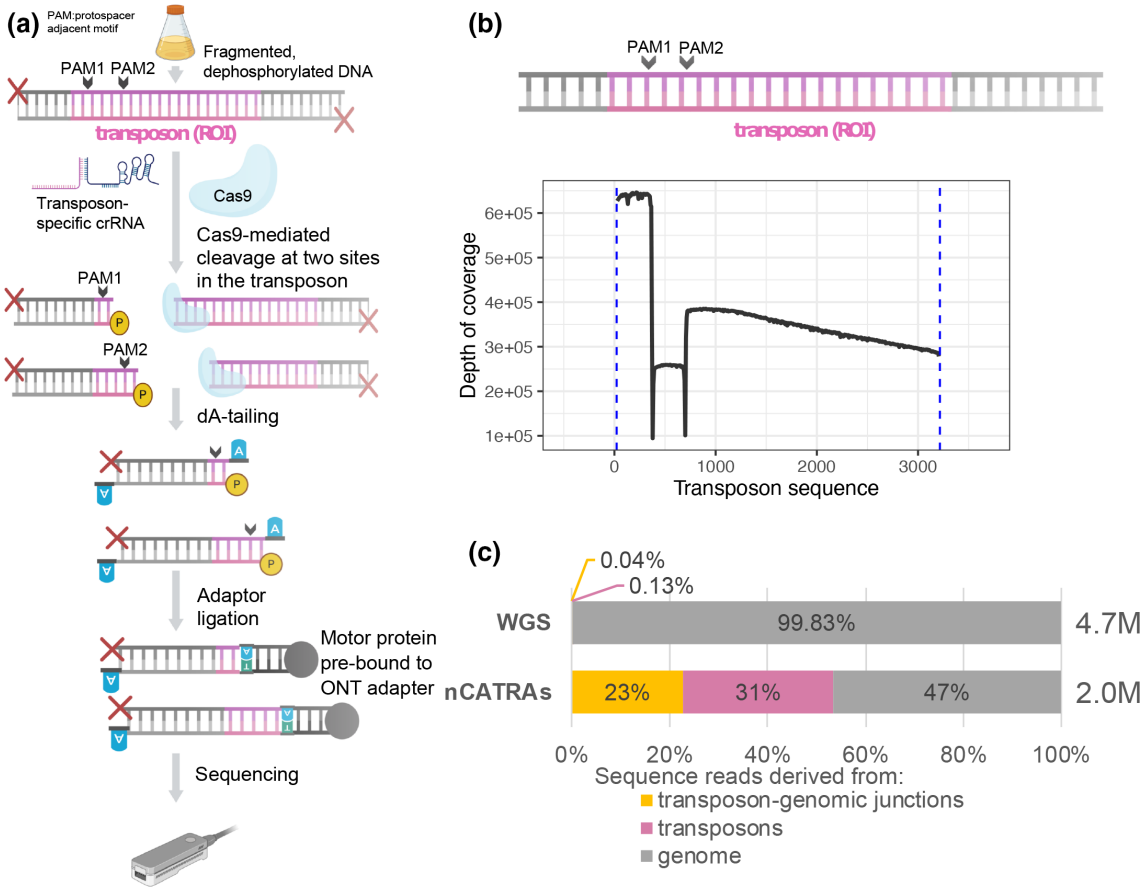

Fig. 1. nCATRAs: a novel amplification-free transposon enrichment and sequencing method. (a) Schematic showing the steps of the nCATRAs method - (adapted from [17]). The Cas9 nuclease is simultaneously targeted to two adjacent sites (in the same reaction tube) at the 5' end of the transposon (region of interest) releasing the transposon-genomic junctions that are subsequently sequenced using Oxford Nanopore Technologies (ONT). The resulting sequencing reads are mapped to the transposon in (b), which shows enrichment at the Cas9 cut sites, quantified in (c). Sequencing of the TnSeq library without nCATRAs enrichment, denoted as 'WGS' for whole-genome sequencing' emphasizes the necessity of enriching for the transposon as only 0.17% of the reads mapped to the transposon (sum of: 0.04% junctions; and 0.13% transposon sequence without genomic region attached). The nCATRAs method substantially enriched for the transposon as a total of 54% (sum of: 23% junctions; and 31% transposon sequence without genomic region attached) of the reads mapped to the Cas9 cut sites on the transposon. The plots represent read pooled from three experimental replicates. Panel (a) was created with BioRender.

per pool of two tubes. In total, we utilized three MinION flow cells - Figs 1 and 2 represent data aggregated from all three flow cells.

Bioinformatic analyses

de novo assembly and completion of the USA300 LAC *S. aureus* genome

ONT raw signals were base called and demultiplexed using Guppy v3.0.3 (https://community.nanoporetech.com). The ONT reads were then trimmed using Porechop (https://github.com/rrwick/Porechop), filtered using a mean quality score of 9 and a minimum read length of 2000 bases following the criteria described in [19]. The adapters and low-quality Illumina paired-end reads were trimmed using Fastp v0.19.5 [20]. The short and long reads (from Illumina and ONT sequencing, respectively) were used to assemble the complete genome *de novo* using UniCycler v0.4.4 [21]. Assembly errors were then corrected by using Snippy v4.4.0 (https://github.com/tseemann/snippy) to call for SNPs using the short and long reads as input data against the assembled genome as reference - SNPs found here are likely errors of the assembler, which were corrected by comparing sequencing read alignments to the assembly (using Integrative Genomics Viewer, IGV) and changed incorrect bases in the assembly to those appearing in the aligned reads. Furthermore, we verified that all thymine-adenine (TA) dinucleotide sites were intact (that is, no SNPs were detected in these sites) in our reference genome since the *S. aureus* TnSeq library was generated using a Mariner-based transposon that inserts at these sites and accurate identification of transposon insertions is contingent upon accurately-defined TA coordinates in the reference genome. This genome and the sequence of

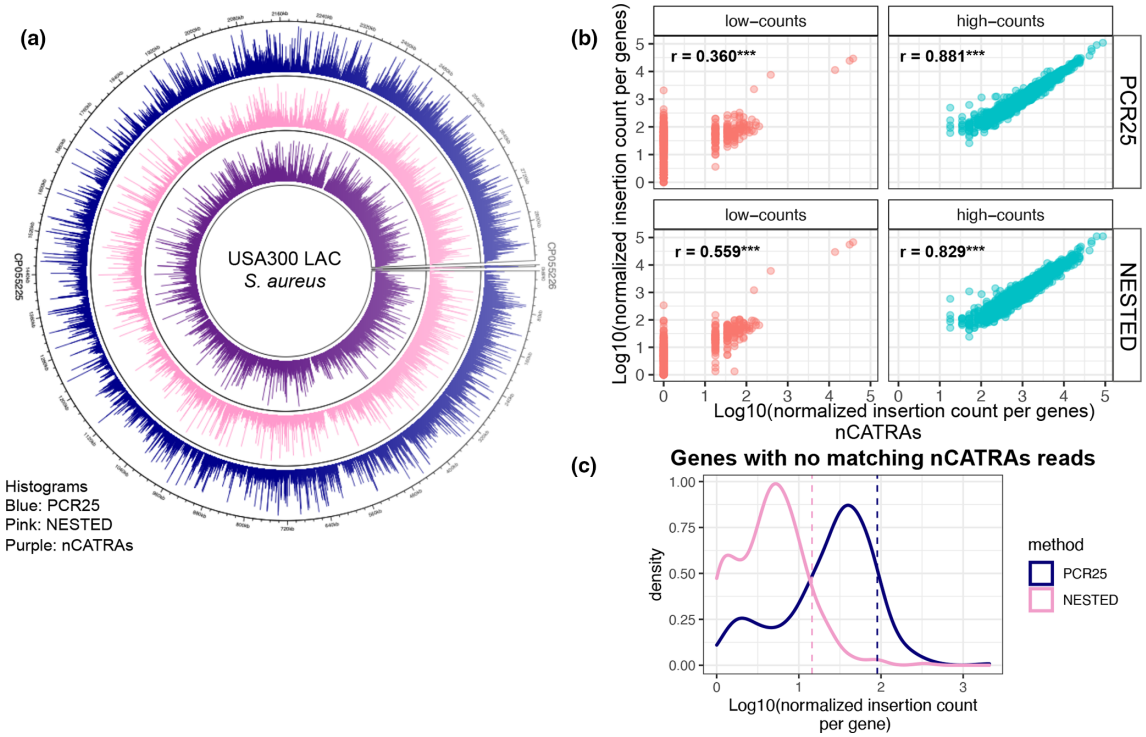

Fig. 2. The novel amplification-free TnSeq method (nCATRAs) is comparable to traditional PCR-based methods. (a) Distribution and frequency of mapped transposon insertions of samples prepared by either the PCR-based PCR25 (outer blue track), NESTED method (middle, pink track) or by the PCR-free nCATRAs method (inner, purple track). Each track represents a pool of three libraries grown to stationary phase and prepared by the same method. Scales were adjusted per track to illustrate the differences among the samples. (b) Kendall correlation of the normalized insertion counts in 'low-counts' genes (left) and 'high-counts' genes (right) between nCATRAs and the PCR-based methods. Here, the correlation plots were separated by the groups of genes defined in (Fig. 4a). (c) Genes with zero insertions detected by nCATRAs were extracted from the PCR25 (blue) and NESTED (pink) datasets and distribution of their insertion counts detected by each of the PCR-based methods was plotted. The dashed lines indicate the 90th percentile where 90% of the genes have insertion counts <15 (NESTED) and <100 (PCR25).

the cryptic plasmid encoded in this strain were deposited to GenBank (CP055225: chromosome and CP055226: cryptic plasmid). The genome sequences were annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v4.12 [22]. Default parameters were used for all software unless otherwise specified.

Mapping transposon insertions in short reads

The USA300 LAC *S. aureus* genome (described above) of the library parent strain was used as reference for analyses. In-house scripts were used for TnSeq analyses. Briefly, 'BWA-MEM' was used to align reads to the reference genome [23]. Aligned reads which started next to TA sites (i.e. the 5′ of the reads maps next to a TA site in the reference genome) (Fig. S1) were then further processed using Samtools [24] and Bedtools [25] to retrieve, count and sum all reads aligned to TA sites as detailed in the 'tnseq_illm' pipeline made available in (https://gitlab.com/piroonj/ncatras_pipeline). Next, the insertion counts per gene were normalized by estimateSizeFactors from the DESeq2 package [26]. This was done to allow for comparing the samples by correcting for differences in sequencing depths. The normalized counts of each samples were retrieved by the 'counts(data, normalized=TRUE)' command of DEseq2 package.

Sequencing depth analyses

The raw sequencing reads resulting from all libraries prepared via the same PCR-based method were pooled, generating three datasets. Reads from each of the pooled datasets were sampled by increasing increments of 50 000 reads. The insertions within each increment of sampled reads were mapped to the TA sites of the reference genome and enumerated (Fig. 3b, left panel). The total number of genes containing insertions was enumerated as well (Fig. 3b, right panel).

Correlation plots

Kendall's rank correlation (cor.test) in R v4.0.2 was used to assess the correlation of insertion counts per gene between noted samples.

Gaussian Mixture Models (GMM)

The insertion counts per gene were pooled by each PCR-based method and by growth phase, thus producing four datasets:

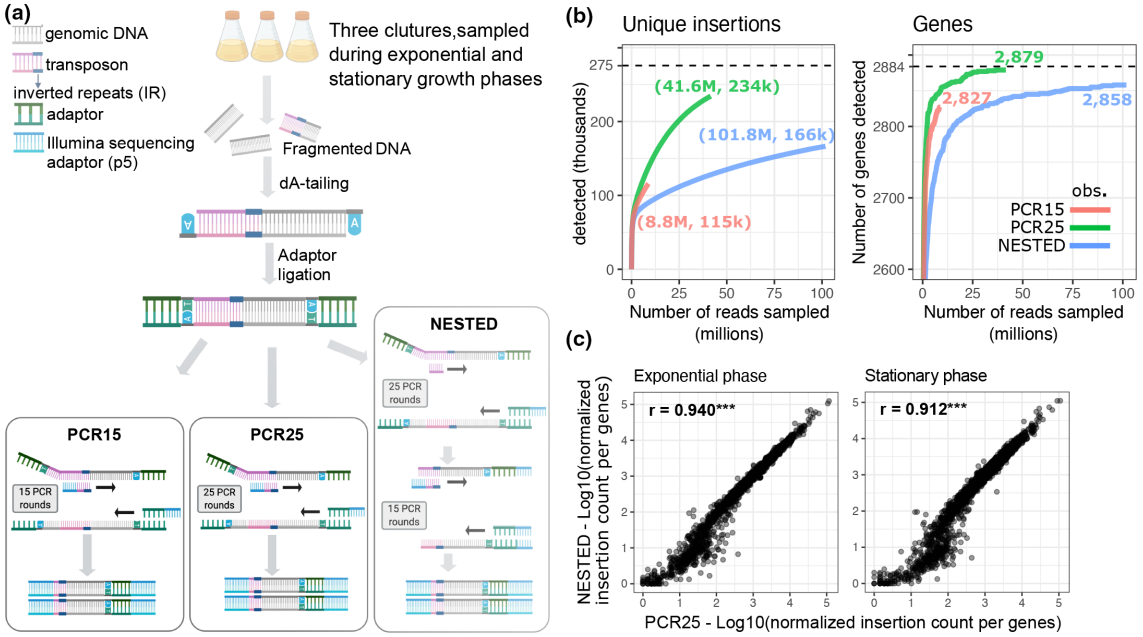

Fig. 3. Number of PCR cycles determines sensitivity of insertion detection and sequencing coverage of insertions. (a) Schematic describing three methods for TnSeq library preparation that vary in the number of PCR cycles and the inclusion of a nested PCR during the transposon insertion enrichment step. Created with BioRender. (b) Sequencing reads were pooled from all libraries (both exponential and stationary) prepared via the same PCR-based method described in (a). These pools were randomly sampled and the number of unique insertions (left) and number of genes with (at least one) mapped insertions (right) were determined using different sized subsets of sequence reads drawn randomly from the combined reads. The dashed lines indicate the maximum number of unique insertions and genes in the genome. The numbers in round brackets indicate the number of sequencing reads and the number of unique insertion from each method, respectively. (c) Kendall correlation plots of normalized transposon insertion counts per gene between PCR25 and NESTED. Each sample represents a pool of three TnSeq libraries grown independently and prepared via the same method outlined in (a). Here, the TnSeq libraries grown to either exponential (left) or stationary (right) phase are presented separately. Each dot represents a gene and is coloured in semi-transparent grey. Those which appear in a darker colour are high-density overlapped dots. Triple asterisks represent a *P*-value <2.2e-16. Axes are presented in log scale.

PCR25-exponential; NESTED-exponential; PCR25-stationary; NESTED-stationary. The counts were normalized with respect to library size using DESeq2 [26]. These datasets, along with: the $\log_2$ fold change (PCR25/NESTED) – exponential; $\log_2$ fold change (PCR25/NESTED) – stationary, were used as input to cluster the data based on the GMMs with three components using the Mclust R package [27]. This produced three subpopulations of genes grouped by levels (i.e. low, intermediate, and high) of the insertion counts detected in each gene (Fig. S5a). Since the distributions of the two gene groups with intermediate to high insertion counts were unaffected by alterations in PCR amplification (Fig. S5b), we decided to merge these two groups for simplicity. Subsequently, two groups of genes were retained: (1) the 'low-counts' group consisting of 510 genes that are sensitive to the number of PCR cycles; (2) the 'high-counts' group consisting of 2374 genes.

Mapping transposon insertions in long reads (nCATRAs)

Briefly, base calling and adapter trimming was performed using Guppy v3.4.5 (https://community.nanoporetech.com). Minimap2 v2.17-r941 was used for alignment and reads which mapped to both the transposon and the reference genome were considered for further analyses [28]. The presence of a TA site in the genomic region of these reads was verified to ensure these were transposon-genomic junctions and not chimeric by-products generated during library preparation. Reads mapped to the same gene were then enumerated using Bedtools v2.25.0. Code detailed in the 'tnseq_ont' pipeline (https://gitlab.com/piroonj/ncatras_pipeline).

Data visualization

All figures were plotted using the R package ggplot2 [29].

RESULTS

Sequencing and reconstruction of the *Staphylococcus aureus* USA300 LAC genome

The TnSeq library used in this study was previously generated in the USA300 LAC background [10, 18]. This library was created using the Mariner-type transposable element (transposon) which inserts at thymine-adenine dinucleotide (TA) sites. This TnSeq library was reported to comprise of more than 70000 unique transposon insertions, with one insertion every ~16 bp [10]. A crucial step in bioinformatic analyses

Table 1. The nucleotide sequence of the whole USA300 LAC *S. aureus* genome was determined using both Illumina and Oxford Nanopore Technologies methods. The short and long reads resulting from both platforms were used to assemble a closed circular genome *de novo*. In addition, we sequenced and assembled the cryptic plasmid within this strain. The chromosome is 2874400 bp with a cryptic plasmid of 3125 bp, totaling one TA dinucleotide every ~10 bp on average. The number of different annotated characteristics is shown, including TA dinucleotide sequences which are the target for the Mariner transposon used to create the transposon mutant library

Features annotated	Chromosome		Plasmid	
	no. of features	TA sites	no. of features	TA sites
CDSs	2803	223337	3	170
tRNAs	59	185	0	0
rRNAs	19	1609	0	0
ncRNAs	3	67	0	0
Intergenic regions	2460	50208	4	135
Total	5344	275406	7	305

CDS, Coding Sequence.

of TnSeq data is the mapping of transposon insertions to the parent strain genome. Therefore, a high-quality genome of the specific LAC parent strain used to generate the library is necessary for optimal analyses. We decided to sequence the exact parent strain used by the laboratory that generated the TnSeq library in order to generate a high-quality genome that ensures the most accurate analysis of our TnSeq data. To generate a complete circular bacterial chromosome, we sequenced the USA300 LAC strain using both next generation Illumina sequencing and single-molecule Oxford Nanopore sequencing. We assembled the genome via *de novo* assembly using both short and long reads obtained from the two sequencing platforms with additional error correction steps (see Methods). We obtained a high quality complete circular chromosome (2874400 bp) and a cryptic plasmid (3125 bp). The genome is deposited in GenBank (CP055225: chromosome and CP055226: cryptic plasmid) and is used as reference in all analyses conducted in this report (Table 1). Additionally, this genome is useful for other projects focused on the USA300 LAC *S. aureus* strain.

PCR cycle number determines detection and coverage of transposon insertions

We aimed to apply a TnSeq assay on *S. aureus* in an *in vivo* osteomyelitis model to elucidate genes critical for its fitness in this important clinical context. Capturing the diversity of a TnSeq library is important to distilling an effective list of significant genes with minimal false positives [5]. Given the complexity of our *in vivo* model, we first wanted to ensure that we minimized bias in every step of the assay. This led us to scrutinize the PCR amplification step employed to enrich transposon insertions. We reasoned that the PCR amplification step may bias TnSeq results in a way that would skew the identification of relevant genes. To test this hypothesis, we devised the experiment described in Fig. 3(a). Briefly, we independently cultured the same TnSeq library in rich media and collected samples for transposon sequencing at both the exponential and stationary growth phases. The purpose of assessing the libraries at two growth phases was to verify whether our observations relating to the effects of PCR regimen were consistent across different biological conditions. The extracted genomic DNA from each sample was then subjected to three separate protocols for enriching the insertions, with the variables between the protocols being the number of PCR cycles and the use of nested primers (Fig. 3a).

Specifically, the literature describes various methods for conducting the PCR-based enrichment step, with some amplifying the insertions via 18 to 25 PCR cycles (21-23) or by implementing a nested PCR reaction with a total of 40 cycles [10, 30], among other methods. With this in mind, we compared the number of mapped insertions identified following 15 vs. 25 cycles of PCR in the PCR15 and PCR25 methods, respectively, and after two PCR reactions consisting of 25 and 15 cycles of PCR using nested primers in the second round of amplification (the NESTED method) (Fig. 3a) [10]. To compare the methods in terms of sensitivity (number of detected insertions) and sequencing depth (the number of sequencing reads), we aggregated the sequence reads of all samples prepared by the same method - to maximize statistical power - and randomly sampled these reads in sequentially large increments to assess the number of unique insertions detected with increasing sequencing depths (Fig. 3b, left panel, Table S1, available in the online version of this article) [31]. Importantly, only reads whose 5′ start site mapped next to coordinates of a TA site in the reference genome were included in further analyses – giving us confidence that minimum to no background reads were considered (Fig. S1). The results demonstrate that the number of detected insertions (i.e. sensitivity) depends on the number of PCR cycles used during the enrichment step (Fig. 3b, left panel, solid lines). The maximum number of possible unique insertions is 275711, which is simply the number of all TA sites in the genome (when considering both chromosome and plasmid). PCR25 was the most sensitive method as it detected 234342 unique insertions, indicating that 84% of all TA sites

in the genome sustained a transposon insertion in this highly saturated TnSeq library. In comparison, the NESTED method detected 166511 (60%) and PCR15 detected 115930 (42%) unique insertions (Fig. 3b, left panel, solid lines). Differences in sequencing depth between the methods were also evident as NESTED yielded the highest sequencing depth with 101.8 million reads per sequencing flow cell (these are reads which mapped next to a TA site in the reference genome, as detailed in Fig. S1, Methods), whereas PCR25 yielded 41.6 million reads per flow cell (Fig. 3b, left panel, solid lines). PCR15, however, yielded a very low sequencing depth of 8.8 million reads per flow cell, which is inadequate for optimal detection of transposon insertions in a TnSeq assay. As can be deduced from the figure, PCR25 is most likely to reach sequencing saturation when employing relatively low sequencing depths, whereas NESTED requires deeper sequencing to detect all insertions in all TA sites.

Plotting the number of genes with mapped insertions against increasing sequencing depths shows that all three methods identify a comparable number of genes with insertions (i.e. mutants) despite the large disparity in the number of detected insertions per TA site (Fig. 3b, right panel). Specifically, the number of genes with mapped insertions identified with the PCR15, PCR25, and NESTED protocols were 2827, 2879, and 2858, respectively (Table S2). Considering that the total number of genes in LAC is 2884, it appears most genes sustain at least one transposon insertion (98.0, 99.8 and 99.1%, respectively). Importantly, identification of these genes reaches a plateau (that is, few additional mutants were detected with increasing increments of two million reads) in PCR25 and NESTED at similar sequencing depths (~22 and ~26 million reads, respectively). Thus, PCR25 is more sensitive in detecting genes with insertions while NESTED yields higher coverage of each gene (Fig. 3b, right panel). This indicates that despite the differences in sequencing saturation, both methods identify most mutants in the library. Given the low sequencing depth produced by libraries prepared via the PCR15 method (Fig. 3b, left panel) and the fact that the number of detected genes does not plateau with PCR15 (Fig. 3b, right panel), the PCR15 samples were excluded from further analyses.

We observed high Kendall's tau correlations of >0.9 among and between the three methods when comparing mapped insertion counts (i.e. mapped reads) per gene – this is true of libraries grown to either exponential or stationary phase, indicating the reproducibility of our PCR-based methods (Figs 3c and S2, Table S2). Taken together, these results suggest that altering the number of PCR cycles used to enrich for the transposon insertions impacts both the detection of unique insertions present and sequencing coverage. It is of note that we observed high correlations when considering all genes in the library (Fig. 3c). However, these correlations show divergence in genes with few sequence reads mapped to their TA sites. The number of reads mapped to TA sites per gene are effectively insertion counts and are referred to as such throughout the text. We therefore hypothesized that PCR amplification may impact the detection of transposon insertions in genes with few insertion counts (a minority of mutated genes represented in the library), whereas genes with high insertion counts (the majority of mutated genes) are not affected, accounting for the high overall correlations. This is explored in the following section.

Essentiality designation is impacted by PCR amplification

To investigate the hypothesis stated above, we clustered the PCR25 and NESTED datasets based on a Gaussian finite mixture model (GMM) using the Mclust R package [27]. The Mclust algorithm detected gene clusters according to their insertion counts and the log odd ratio of the same counts from PCR25 vs. NESTED (detailed in Methods). The results are shown in Fig. 4. As expected, there is a Gaussian distribution with a large peak where most genes have lots of insertions; there is also a second, small peak, which likely reflects another set of genes with relatively few insertions. We defined the groups observed by GMM as: 'high-counts' – all genes in this group show an intermediate to high number of mapped insertions (>200 insertion counts per gene); and 'low-counts' – most genes in this group show fewer mapped insertions (Fig. 4a, Table S3). The distribution of insertion counts per gene in the exponential and stationary phase datasets was similar. The majority of *S. aureus* genes clustered in the 'high-counts' group, which included 82% (2374 genes) of all genes, whereas the remainder 18% (510 genes) clustered in the 'low-counts' group. It is clear that the distribution of insertion counts per gene in the 'low-counts' group shifted further left upon increasing the number of PCR cycles (PCR25 vs. NESTED), indicating a lower number of insertion counts (Fig. 4a). This is in contrast to the 'high-counts' group which remained stable despite changes in PCR cycles. Quantification of these results confirmed that the overall insertion counts are significantly reduced in the 'low-counts' group upon increasing the number of PCR cycles and adding the nested PCR, whereas insertion counts in genes of the 'high-counts' group were not affected (Fig. 4b). Detailed analyses show that most of the insertion counts per gene in both the genes encoding protein (CDS) and genes encoding tRNAs, which belong to the 'low-counts' group, is reduced in NESTED compared to PCR25 (Fig. S3).

These trends are reflected in the Kendall's tau correlations. Specifically, when considering genes which fall into the 'low-counts' group, the correlation is poor (0.674, 0.598), whereas the correlation is much better (0.929, 0.895) when considering the genes within the 'high-counts' group for exponential and stationary phase libraries, respectively (Figs 4c and 2d). To quantify the variability between the PCR25 and NESTED methods, we plotted the distribution of fold change differences of insertion counts produced by PCR25 relative to NESTED (Fig. 4d). This revealed a median of 5.92-fold increased total insertion count in the 'low-counts' genes when processed by the PCR25 method compared to NESTED in stationary phase. This is slightly lower (median of 2.83-fold) in exponential phase samples,

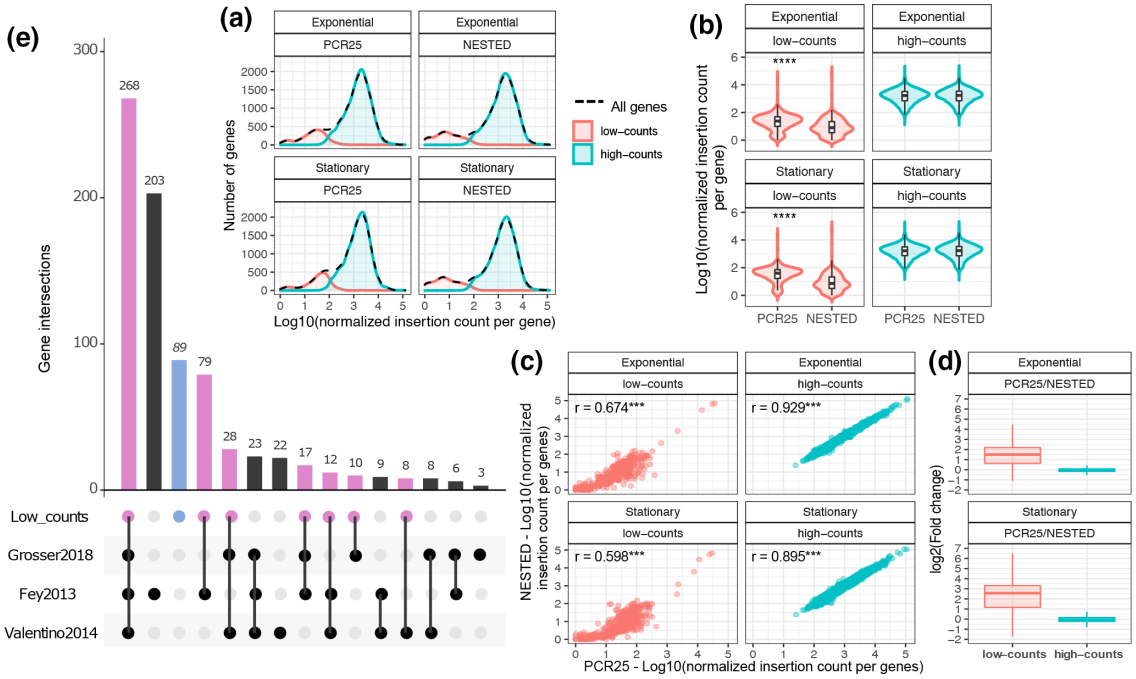

Fig. 4. NESTED identifies fewer insertion counts in candidate essential genes. (a) The insertion counts per gene of libraries prepared by either PCR25 or NESTED methods were fitted to a Gaussian Mixture Model (GMM). The GMM was generated for both exponential (top) and stationary (bottom) phase datasets. (b) Violin plots showing the distribution of insertion counts of genes belonging to either the 'low-counts' (left) or 'high-counts' (right) groups. Four asterisks represent *P*-value<2.2e-16 calculated by the Wilcoxon signed-rank test. (c) Kendall correlation plots of insertion counts in the 'low-counts' genes (left) and 'high-counts' genes (right). Three asterisks represent *P*-value <2.2e-16 (d) $\log_2$fold change of insertion counts in 'low-counts' genes (red) and 'high-counts' genes (green) assessed by PCR25 relative to NESTED. For a–d: each plot represents insertion counts from a pool of three TnSeq libraries grown independently to the noted growth phase and prepared via the same method outlined in (Fig. 3a). (e) UpsetR showing the intersections of genes which fall in each of the indicated categories. Categories include genes described as essential by: Grosser2018: [10]; Fey2013: [32]; and Valentino2014: [11]. Low_counts: genes with low insertion counts which were grouped by GMM described in (a). Intersections of the 'low-counts' and at least one other group are highlighted in pink and add up to 421 genes. The 'low-counts' genes not previously described as essential by any of the indicated groups are highlighted in blue and add up to 89 genes.

whereas the 'high-counts' group showed almost no difference (fold change median ~1) by either method (Fig. 4d). To investigate the importance of these observations, we attempted to identify trends among genes that fall within the 'low-counts' group. We speculated that this group may include genes essential for *S. aureus* survival. To explore this possibility, we compared the genes in the 'low-counts' group to three published *S. aureus* putative essential genes datasets [10, 11, 32] (Fig. 4e, Table S4). We find that the majority of the 'low-counts' genes were previously deemed essential with 421 (83%) of the genes in this group appearing in at least one of the reported putative essential gene datasets (Fig. 4e). We conclude that PCR amplification and the nested PCR biases the distribution of genes with few transposon insertion counts, and the majority of these genes are likely essential for *S. aureus* survival.

A novel PCR-free TnSeq for enrichment of transposon-genomic junctions

To further assess the effects of PCR amplification on the TnSeq assay, we devised and evaluated a PCR-free approach to detect and quantify transposon insertions in our TnSeq library. We termed this novel method nCATRAs (nanopore Cas9-targeted transposon sequencing). (Fig. 1a). The method was adapted from the previously described nCATS protocol used for target enrichment and subsequent sequencing via Oxford Nanopore Technologies (ONT) to enrich for the insertions in a pool of three TnSeq libraries grown to stationary phase in rich media (Fig. 1a) [16, 17]. We designed two guide RNAs (gRNAs), both specific to the transposon, one 322 bp and the other 648 bp away from the Inverted Repeat (IR). We directed the Cas9 nuclease to these two sites to introduce a single cut in DNA molecules encoding for the transposon at either of these two cut sites or both. This enhances enrichment efficiency as the number of enriched molecules encoding for the transposon insertions is increased by directing Cas9 to multiple sites on the same target [16]. DNA was then treated to adenylate 3′-OH group to which ONT adapters, that possess 3′-T cohesive ends, were ligated, and sequenced using MinION flow cells. The transposon insertions were successfully enriched as a high percentage of sequenced reads aligned

to the transposon at the Cas9 cut sites (Fig. 1b, quantified in Fig. 1c). This approach resulted in a 54% enrichment of the target transposon sequence over background genomic DNA, a substantial percentage considering that sequencing the TnSeq library without any prior enrichment step resulted in mapping just 0.17% of the reads the target transposon (Fig. 1c). However, our objective was to enrich for transposon-genomic junctions (i.e. insertions) and not merely the transposon. Therefore, when considering the insertions, we observed a 23% enrichment over background DNA and a total of 58849 mapped unique transposon insertions. We conclude that our nCATRAs PCR-free approach successfully enriches for the transposon insertions in the TnSeq library.

The PCR-free TnSeq (nCATRAs) method is comparable to traditional methods

We next compared nCATRAs to PCR-based methods in order to evaluate the effects of PCR amplification on the TnSeq assay results. The overall distribution and relative frequency profile of transposon insertions detected by nCATRAs is similar to the profiles produced by either the PCR25 or NESTED methods (cor >0.869) (Fig. 2A). Building on our observations that PCR amplification influences the distribution of genes with low transposon insertion counts (Fig. 4), we hypothesized that detection of insertions by nCATRAs would differ from PCR-based approaches primarily in the 'low-counts' group. Therefore, we compared nCATRAs to the PCR-based methods by Kendall's tau correlations of insertion counts per gene in each of the two groups defined in the previously defined 'low-counts' and 'high-counts' groups (Fig. 2b). Concurrent with our hypothesis, the correlations were poor (0.360: nCATRAs vs. PCR25, 0.559: nCATRAs vs. NESTED) when considering the 'low-counts' group (Fig. 2b, left panel). In contrast, the correlations between the nCATRAs and PCR-based methods were high (0.881, 0.829) when considering the 'high-counts' group of genes (Fig. 2b, right panel). It is of note that the nCATRAs method detected zero insertions in 362 genes due to the lower sequencing depths of the PCR-free approach, which is a much higher number than observed with any PCR-based method (Table S5). This may indicate that these 362 genes have very few insertions in the TnSeq library. To interrogate this, we plotted the distribution of insertion counts detected by PCR-based methods in these genes and demonstrate that insertion counts were low (<15 insertions per gene in NESTED; <100 insertions per gene in PCR25) in 90% of the 362 genes (Fig. 2c). We conclude that nCATRAs did not detect insertions in these genes because they indeed have few insertions, and not due to inherent randomness of the method. Nevertheless, when considering all genes, we find that the frequency of transposon junctions per gene detected by the PCR-free nCATRAs method highly correlates with that obtained by the PCR-based methods (Fig. S4). Taken together, our experiments with a PCR-free TnSeq method reinforce our conclusion that PCR amplification impacts transposon insertion detection in genes with few insertions, whereas the detection of insertions in the majority of genes is not affected by PCR-amplification bias. Moreover, we conclude that our PCR-free approach is valid and comparable to traditional, PCR-based approaches.

DISCUSSION

A major strength of the TnSeq assay is its sensitivity in detecting mutations that are pertinent in a niche condition. This sensitivity originates in the transposon insertion enrichment step, where the literature unanimously applies PCR amplification-based methods. Given the wide use of the TnSeq assay, and the lack of a comprehensive report comparing the potential bias produced by increased PCR amplification cycles, or lack thereof, we decided to investigate the possible effects of PCR amplification bias on TnSeq results. We explore this by implementing both PCR-based and novel PCR-free transposon enrichment methods and subsequently comparing the mapped insertion counts (i.e. number of sequence reads mapped next to TA sites of the reference genome) detected by each method. Before beginning our TnSeq assays, we decided to sequence the parent USA300 LAC *Staphylococcus aureus* strain – that is, the exact strain used by the laboratory which created the library [10]. This was necessary because changes in bacterial genomes occur with passage; and usage of an accurate reference genome is particularly important for TnSeq data analysis. For instance, transposon insertions may be called as putative essential if a deletion occurred in the reference genome. We reconstructed a high-quality genome which we used as reference in subsequent PCR-based and PCR-free assays. In our PCR-based enrichment methods, we subjected TnSeq libraries previously generated in the prominent bacterial pathogen, *S. aureus* [10], to three library preparation protocols that vary in: the number of PCR cycles used to amplify the transposon insertions; and in the inclusion of a nested PCR step. To verify whether the effects of PCR are consistent across different biological conditions, we applied these methods to TnSeq libraries grown to either exponential or stationary phase. It is if of note that we did identify mutants with altered fitness when grown to either exponential or stationary phase (Table S6). However, we do not discuss these differences here since the objective of leveraging two growth phases was to affirm that our conclusions regarding PCR effects stand across different biological conditions. Analysis of the relationship between unique transposon insertions detected and sequencing depth, revealed differences between the PCR-based methods. The low sequencing depth obtained by the PCR15 method indicates that the DNA amplified by this method retained high amounts of non-transposon background DNA resulting in reduced cluster generation when using transposon-specific sequencing primer on the Illumina platform. Thus, 15 PCR cycles are not sufficient to saturate the detection of insertions in the library as only 42% of the theoretical maximum unique insertion sites were detected. The PCR25 method was most sensitive as

this method detected unique insertions in 84% of all TA sites in the genome, indicating that our TnSeq library is particularly highly saturated. It may be that PCR25 detected all possible insertions since a fraction of TA sites (here, the remaining 16%) are expected to be transposon-intolerable. However, we speculate that not every insertion detected by this method necessarily correlates with viable mutants in the corresponding gene, as it has been suggested that dead cells carrying transposon insertions may be detected by the highly-sensitive PCR and subsequent next generation sequencing approaches [33]. In contrast, the method that incorporated the highest number of PCR amplification cycles along with a nested PCR, i.e. NESTED, produced the highest sequencing depth of insertions, which is advantageous in downstream TnSeq analyses as this allows for enhanced statistical power in determining genes of interest [31]. There is a trade-off between increased sequencing depth and reduced number of unique insertions and genes detected. Thus, we conclude that the number of PCR cycles should be adjusted based on amount of transposon in the sample. An optimal number of PCR cycles results in the saturation in detection of transposon insertions and genes with insertions.

Upon discerning the Kendall's tau correlations in Fig. 3(c), we noticed divergence of genes with low insertion counts whereas genes with high insertion counts were highly correlated. To closely examine this, we clustered the distribution of transposon insertion counts per gene from libraries prepared via the different PCR-based methods, specifically PCR25 and NESTED, using Gaussian finite mixture models (GMM) [27]. The GMM models produce two distinct groups of genes which further separate upon increasing the number of PCR cycles and adding a nested PCR to amplify the insertions. We defined the subpopulations of genes resulting from GMM based on their insertion counts. Detection of insertions in genes which fall within the 'low-counts' group is affected by PCR amplification bias since the distribution of insertion counts shifts and total insertion counts decrease in these genes upon increasing the number of PCR cycles and including a nested PCR. This is consistent across libraries grown to two different growth phases, giving us confidence in our conclusions. In contrast, insertion counts in genes of the 'high-counts' group were not affected by PCR amplification bias since their distribution was stable, their levels were unchanged and their correlation was high upon increasing the number of PCR cycles and including a nested PCR. As expected, the majority of genes previously described as essential for *S. aureus* survival are among the 'low-counts' group. While determination of gene essentiality is complex [33–35], we nonetheless conclude that the 'low-counts' group contains genes essential for *S. aureus* survival as 83% of the genes in this group were deemed essential by at least three other reports which incorporated independent studies to deduce essential genes [10, 11, 32]. High transposon insertion frequencies in non-essential genes (i.e. the 'high-counts' group) is expected since these genes are dispensable during growth in rich media and insertions in these genes are therefore tolerated well. Thus, our data indicate that detection of transposon insertions in an important group of genes is sensitive to the number of PCR cycles and the inclusion of a nested PCR, while the same does not apply to the majority of *S. aureus* genes which are presumably not essential in our growth conditions. We demonstrate that a higher number of PCR cycles with a nested PCR produces a sharper distinction between 'low-count' (putative essential) and 'high-count' (non-essential) genes. The NESTED method skews the data such that 'low-count' genes are less likely to be detected because the 'high-count' genes are overwhelmingly sequenced by this method. Our data suggest that researchers intending to investigate the putative essential genes of an organism using TnSeq should carefully weigh their PCR regimens including the number of PCR cycles and the inclusion of nested PCR used in their protocols. Too few PCR cycles (i.e. <15 cycles) provides insufficient sequencing depths for TnSeq data analysis. Too many PCR cycles (i.e. 25 cycles + 15 cycles of nested PCR – as in NESTED) may limit the detection of putative essential genes – which belong to the 'low-count' group – as these may be drowned out by the 'high-count' genes. Further, when applying TnSeq to search for conditionally essential genes, researchers may consider implementing the NESTED method as it provides higher sequencing depth per insertion, which bolsters statistical power needed for identifying differentially important genes. Indeed, we found that NESTED was most suitable for our experiments since it excelled at separating the 'high-count' genes from the high level of background genomic DNA in these complex samples [36].

We next sought to evaluate the effects of PCR amplification on the results of the TnSeq assay in a more direct manner by applying a PCR-free method for the enrichment of transposon insertions followed by sequencing using Oxford Nanopore Technologies (ONT). To this end we introduced the nCATRAs method for transposon-genomic junction enrichment [16]. Using this approach, we achieved (23%) enrichment of the transposon-genomic junctions and (54%) enrichment of the transposon over background genomic DNA. That is, the sequenced reads mapping to the transposon-genomic junctions comprised 23% of the total sequencing reads. These enrichment levels are substantial and are on par with results of a technique - published during the preparation of this manuscript - which employs a similar method to enrich for lentivirus insertion sites in mammalian cells [37]. The total number of unique insertions detected by this method was considerably lower than that detected by traditional PCR-based approaches, which is expected as the sequencing depth is lower since this method employed no amplification. Rather, the native DNA molecules were sequenced directly and the transposon insertions were enumerated following enrichment via Cas9 cleavage. The distribution of insertions detected by this method is comparable to that produced by PCR-based methods, lending to the validity of the PCR-free approach.

Further, we find a high correlation between the insertion counts detected by PCR-free and PCR-based approaches when considering either all genes or 'high-counts' genes only. However, in agreement with our findings from the PCR-based assays, we observe poor correlations between the PCR-free and PCR-based enrichment methods when considering 'low-counts' genes, further reinforcing our conclusions that detection of insertions within these genes is precarious and sensitive to PCR amplification bias and nested PCR. Interestingly, despite the low correlation in the 'low-counts' genes, we observed trends in the total insertion frequency per gene as the genes that had zero mapped insertions when processed by nCATRAs produced very low insertion counts in the PCR-based methods. This suggests that although this method did not detect the very low count insertions owing to lack of amplification, it nevertheless yields similar insertion profiles to PCR-based methods. In other words, genes which do not tolerate transposon insertions are underrepresented in both PCR-based and PCR-free output datasets. Thus, we present a novel, amplification-free method for detection of transposon insertions in a TnSeq library. We argue this method is valid and applicable to TnSeq assays since it produces results comparable to those observed by traditional PCR-based methods. Implementing nCATRAs gave us confidence in our biological conclusions from the TnSeq assay, as the results guided us in prioritizing the list of significant genes since those over-represented in both methods were likely true positives. A major limitation to nCATRAs is the low sequencing depth, which may lead to misinterpreting a missed insertion as putative essential when, in fact, it was missed during sequencing. Indeed, the low sequencing depth of the transposon-genomic junctions was our biggest experimental challenge. During our troubleshooting we learnt that this is attributed to poor depletion of background genomic DNA. We therefore attempted to enrich for the transposon-genomic junctions – and deplete background DNA - using various methods, including pull-down of a deactivated, biotinylated Cas9 targeted to the junctions; tweaking the de-phosphorylation step in the protocol by increasing concentration of the phosphatase and increasing incubation times (data not shown). However, we ultimately found that the highest enrichment percentage was achieved by fragmenting the starting DNA (see Methods) and introducing two adjacent cuts at one end of the junctions. This may be substantially improved by pairing our nCATRAs protocol with new algorithms that enable enrichment at the nanopore sequencing step, such that only DNA molecules encoding the target reads (in this case, the transposon-genomic junction) would be sequenced [38, 39]. In addition, the PCR-free TnSeq method offers unique advantages including sequencing the native DNA molecules which allows detection of DNA modifications.[40–42] This may further our understanding of transposition patterns as unexplained selectivity of both the Himar1 (used to generate the TnSeq library analysed in this study) and Tn5 transposases has been described, and DNA modifications have been eluded to as possible factors which dictate this selectivity [35]. Nevertheless, the PCR-based TnSeq methods remain superior due to accessibility, high sequencing depths and sensitivity.

Funding information
D.A., is currently supported by the National Science Foundation, award No. OIA-1946391. This work was supported by NIH grant R01-AI119380 to M.S.S. Additional support was provided by a generous gift from the Texas Hip and Knee Centre and research core facilities supported by the Centre for Microbial Pathogenesis and Host Inflammatory Responses (P20-GM103450), the Translational Research Institute (UL1TR000039). D.A. and experiments described in this work were also supported by a grant from the Arkansas Research Alliance awarded to D.U. and M.S.S. All bioinformatic analyses were conducted using the High-Performance Computing system at the University of Arkansas for Medical Sciences. National Institute of General Medical Sciences of the National Institutes of Health (P20GM125503) awarded to I.N.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Hayes F. Transposon-based strategies for Microbial functional genomics and proteomics. *Annu Rev Genet* 2003;37:3–29.
2. Judson N, Mekalanos JJ. Transposon-based approaches to identify essential bacterial genes. *Trends Microbiol* 2000;8:521–526.
3. McClintock B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* 1950;36:344–355.
4. van Opijnen T, Bodi KL, Camilli A. Tn-seq: High-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* 2009;6:767–772.
5. Cain AK, Barquist L, Goodman AL, Paulsen IT, Parkhill J, *et al*. A decade of advances in transposon-insertion sequencing. *Nat Rev Genet* 2020;21:526–540.
6. Barquist L, Boinett CJ, Cain AK. Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biol* 2013;10:1161–1169.
7. Shields RC, Jensen PA. The bare necessities: Uncovering essential and condition-critical genes with transposon sequencing. *Mol Oral Microbiol* 2019;34:39–50.
8. Coe KA, Lee W, Stone MC, Komazin-Meredith G, Meredith TC, *et al*. Multi-strain TN-SEQ reveals common daptomycin resistance determinants in *Staphylococcus aureus*. *PLoS Pathog* 2019;15:e1007862.
9. Santiago M, Lee W, Fayad AA, Coe KA, Rajagopal M, *et al*. Genome-wide mutant profiling predicts the mechanism of a lipid II binding antibiotic. *Nat Chem Biol* 2018;14:601–608.
10. Grosser MR, Paluscio E, Thurlow LR, Dillon MM, Cooper VS, *et al*. Genetic requirements for *Staphylococcus aureus* nitric oxide resistance and virulence. *PLoS Pathog* 2018;14:e1006907.
11. Valentino MD, Foulston L, Sadaka A, Kos VN, Villet RA, *et al*. Genes contributing to *Staphylococcus aureus* fitness in abscess- and infection-related ecologies. *mBio* 2014;5:e01729-14.
12. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, *et al*. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci U S A* 2013;110:19872–19877.
13. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, *et al*. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* 2012;109:14508–14513.
14. Kebschull JM, Zador AM. Sources of pcr-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* 2015;43:e143:21::.
15. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, *et al*. Analyzing and minimizing PCR amplification bias in illumina sequencing libraries. *Genome Biol* 2011;12:R18.
16. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, *et al*. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol* 2020;38:433–438.

17. Wongsurawat T, Jenjaroenpun P, De Loose A, Alkam D, Ussery DW, *et al*. A novel Cas9-targeted long-read assay for simultaneous detection of IDH1/2 mutations and clinically relevant MGMT methylation in fresh biopsies of diffuse glioma. *Acta Neuropathol Commun* 2020;8:87.
18. Boles BR, Thoendel M, Roth AJ, Horswill AR. Identification of genes involved in polysaccharide-independent *Staphylococcus aureus* biofilm formation. *PLoS One* 2010;5:e10146.
19. Jenjaroenpun P, Wongsurawat T, Udaondo Z, Anderson C, Lopez J, *et al*. Complete genome sequences of four isolates of vancomycin-resistant *Enterococcus faecium* with the VANA gene and two daptomycin resistance mutations, obtained from two inpatients with prolonged bacteremia. *Microbiol Resour Announc* 2020;9.
20. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–i890.
21. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
22. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, *et al*. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* 2016;44:6614–6624.
23. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.
24. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27:2987–2993.
25. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–842.
26. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
27. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *R J* 2016;8:289–317.
28. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.
29. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag; 2016.
30. Wilde AD, Snyder DJ, Putnam NE, Valentino MD, Hammer ND, *et al*. Bacterial hypoxic responses revealed as critical determinants of the host-pathogen outcome by TnSeq analysis of *Staphylococcus aureus* invasive infection. *PLoS Pathog* 2015;11:e1005341.
31. Chao MC, Abel S, Davis BM, Waldor MK. The design and analysis of transposon insertion sequencing experiments. *Nat Rev Microbiol* 2016;14:119–128.
32. Fey PD, Endres JL, Yajjala VK, Widhelm TJ, Boissy RJ, *et al*. A genetic resource for rapid and comprehensive phenotype screening of nonessential staphylococcus aureus genes. *mBio* 2013;4:e00537:00512.:.
33. Miravet-Verde S, Burgos R, Delgado J, Lluch-Senar M, Serrano L. FASTQINS and ANUBIS: Two bioinformatic tools to explore facts and artifacts in transposon sequencing and essentiality studies. *Nucleic Acids Res* 2020;48:e102.
34. Fang G, Rocha E, Danchin A. How essential are nonessential genes? *Mol Biol Evol* 2005;22:2147–2156.
35. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, *et al*. Comprehensive essentiality analysis of the *Mycobacterium tuberculosis* genome via saturating transposon mutagenesis. *mBio* 2017;8.
36. Alkam D, Jenjaroenpun P, Ramirez AM, Beenken KE, Spencer HJ, *et al*. The increased accumulation of *Staphylococcus aureus* virulence factors is maximized in a purR mutant by the increased production of SarA and decreased production of extracellular proteases. *Infect Immun* 2021;89.
37. van Haasteren J, Munis AM, Gill DR, Hyde SC. Genome-wide integration site detection using Cas9 enriched amplification-free long-range sequencing. *Nucleic Acids Res* 2020.
38. Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol* 2021;39:431–441.
39. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, *et al*. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol* 2021;39:442–450.
40. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, *et al*. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 2017;14.
41. Nookaew I, Jenjaroenpun P, Du H, Wang P, Wu J, *et al*. Detection and discrimination of DNA adducts differing in size, regiochemistry, and functional group by nanopore sequencing. *Chem Res Toxicol* 2020;33.
42. Jenjaroenpun P, Wongsurawat T, Wadley TD, Wassenaar TM, Liu J, *et al*. Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res* 2021;49.