BMC Genomics

RESEARCH Open Access

Sequencing by binding rivals SMOR error-corrected sequencing by synthesis technology for accurate detection and quantification of minor (< 0.1%) subpopulation variants

Christopher J. Allender[1], Candice L. Wike[2], W. Tanner Porter[1], Dean Ellis[2], Darrin Lemmer[1], Stephanie J. K. Pond[2] and David M. Engelthaler[1*]

Abstract

Background Detecting very minor (< 1%) subpopulations using next-generation sequencing is a critical need for multiple applications, including the detection of drug resistant pathogens and somatic variant detection in oncology. A recently available sequencing approach termed 'sequencing by binding (SBB)' claims to have higher base calling accuracy data "out of the box." This paper evaluates the utility of using SBB for the detection of ultra-rare drug resistant subpopulations in *Mycobacterium tuberculosis* (*Mtb*) using a targeted amplicon assay and compares the performance of SBB to single molecule overlapping reads (SMOR) error corrected sequencing by synthesis (SBS) data.

Results SBS displayed an elevated error rate when compared to SMOR error-corrected SBS and SBB techniques. SMOR error-corrected SBS and SBB technologies performed similarly within the linear range studies and error rate studies.

Conclusions With lower sequencing error rates within SBB sequencing, this technique looks promising for both targeted and unbiased whole genome sequencing, leading to the identification of minor (< 1%) subpopulations without the need for error correction methods.

Keywords Sequencing by synthesis (SBS), Sequencing by binding (SBB), Tuberculosis (TB), Heteroresistance, Single molecule overlapping reads (SMOR), Ultra-rare mutation detection, NGS benchmarking

*Correspondence:
David M. Engelthaler
dengelthaler@tgen.org
[1] Pathogen and Microbiome Division, Translational Genomics Research Institute, 3051 W. Shamrell Blvd., Suite 106, Flagstaff, AZ 86005, USA
[2] Emerging Opportunities Division, Translational Genomics Research Institute, 445 N 5th Street, Phoenix, AZ, USA

Background

While next-generation sequencing (NGS) has revolutionized genomics in oncology and infectious disease, challenges remain in characterizing cellular heterogeneity, including the accurate measurement of minor populations (< 1%) due to the intrinsic error rates, particularly in 'sequencing by synthesis' (SBS) methods [1]. One critical use case for accurate measurement of minor populations (< 1%) is the early identification of drug resistant subpopulations of *Mycobacterium tuberculosis* (*Mtb*), a phenomenon known as heteroresistance. Tuberculosis remains a leading global infectious disease problem due to many challenges associated with detecting and treating drug resistant *Mtb* infections. Sequencing-based analysis of resistance causing mutation loci in drug resistance-related *Mtb* genes has revealed both dominant and minor drug resistant subpopulations in patients exhibiting drug resistant infections [2–5]; accurate detection and quantification of these subpopulations can be significantly affected by polymerase and sequencing error.

Multiple methods have been introduced to decrease sequencing error rates including unique molecular identifiers (UMIs) [6], duplex sequencing [7, 8] and others [9–11]. Current error correction methods, while powerful, introduce complexity to library prep and analysis, and often require over-sequencing. UMIs are an error correction tool commonly used in sequencing where barcodes are introduced by tagging molecules in the first cycle of PCR to identify and correct both PCR and sequencing errors. Single molecule overlapping reads (SMOR) is one error correction method and the focus of this study on *Mtb* amplicons, which requires overlapping read 1 (R1) and read 2 (R2) pairs and discards variants within overlapping reads where the base call differs between R1 and R2. We have previously shown that single molecule overlapping reads (SMOR) analysis reduces SBS sequencing error significantly and can be used to detect minor populations of drug resistant *Mtb* down to 0.1% of the total population [12–19]. Here, we assess the reliability of a novel 'sequencing by binding' (SBB) chemistry (Pacific Biosciences, PacBio) to sequence well defined *Mtb* drug resistance mutations in *katG* and *gyrA* genes (conferring isoniazid and fluoroquinolone resistance, respectively) down to 0.01% without employing additional error correction methods (e.g., UMIs or SMOR).

Methods

In order to assess the capabilities of targeted SBB sequencing to identify and quantify low to extremely low minor subpopulations, and compare to error-corrected and non-error-corrected targeted SBS sequencing, we created validated contrived mixtures, conducted PacBio SBB and Illumina SBS sequencing, and statistically compared sequencing outputs for sequencing accuracy and quantification precision. Methods in detail are as follows:

Contrived mixtures

Two plasmids were used as DNA template to create PCR products for making contrived drug resistant mutation mixtures at 10%, 1%, 0.1%, 0.01%, and 0.001%. A wildtype (WT) plasmid was created (Blue Heron Biotech, LLC) with *katG* and *gyrA* sequences that match the *Mycobacterium tuberculosis* H37Rv pan-susceptible strain. A second resistant (RS) plasmid was created (Blue Heron Biotech, LLC) with *katG* and *gyrA* mutations at *katG* g944c and *gyrA* a241g, which are mutations known to confer drug resistance in clinical isolates from tuberculosis patients [20]. Both plasmids were linearized with PciI restriction enzyme (New England Biolabs, Inc.) using manufacturer's conditions and diluted to 10^3 copies and used as DNA template for targeted sequencing PCR. This PCR contained primers (200 nM final conc.) with universal tails as previously described [13], Q5® Hot Start High-Fidelity 2X Master Mix (New England Biolabs, Inc.) (1x final conc.), betaine (MilliporeSigma) (1 M final conc.), and water to 30 µL. The cycling conditions were 98 °C for 1 min; 35 cycles of 98 °C for 15 s, 60 °C for 20 s, and 72 °C for 20 s; and 72 °C for 5 min. The WT PCR products were diluted with water to 400 uL to make enough volume for subsequent 10-fold dilution series. Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Inc.) quantification was used to measure, dilute, and confirm the RS PCR products were at the same concentration as the WT PCR products. With all PCR products at equimolar concentrations (i.e., for both *katG* and *gyrA*), three separate 10% mixtures were created by mixing 3 µL of RS PCR products into 27 µL of WT PCR products. Four additional 10-fold dilution series for each 10% mixture replicate were created using WT PCR products as the diluent. These universally-tailed PCR products (i.e., six different 10-fold dilution series) were purified with a 1.0x AMPure XP (Beckman Coulter, Inc) bead cleanup, eluted into 20 µl of water, processed further for either Illumina or PacBio SBB sequencing.

Illumina SBS sequencing

The purified PCR product mixtures were each barcoded with dual and unique 12 bp indexing primers in a second PCR that also added adapters for sequencing on an Illumina DNA sequencer platform. This PCR contained 2 µL of the previous product, primers (400 nM final conc.) with barcodes and adapters as previously described [13], KAPA HiFi Hotstart ReadyMix (Roche Molecular Systems, Inc.) (1x final conc.), and water to 50 µL. The cycling conditions were 98 °C for 2 min; 6 cycles of 98 °C for 30 s, 60 °C for 20 s, and 72 °C for 30 s; and

72 °C for 5 min. These PCR products were purified with a 0.8x AMPure XP (Beckman Coulter, Inc) bead cleanup and eluted into 40 µL of water. These individual libraries were quantified with a KAPA Library Quantification Low Rox Kit (Roche Molecular Systems, Inc.) on a QuantStudio™ 7 Flex Real-Time PCR System (Thermo Fisher Scientific, Inc.) and pooled in equimolar amounts before being combined with 20% PhiX sequencing control v3 (Illumina, Inc.), assessed with a High Sensitivity D5000 ScreenTape (Agilent Technologies, Inc.) on a 4200 TapeStation system (Agilent Technologies, Inc.), and finally sequenced on a MiSeq® System (Illumina, Inc.) with a 600-cycle MiSeq Reagent Kit v3 (Illumina, Inc.).

PacBio SBB sequencing

The individual Illumina libraries were converted using the PacBio Onso Conversion protocol (PacBio 102-529-500). Briefly, a p5/p7 library (5-100 fmol) were added to PCR conversion primers (2.5ul), PCR master mix (2X) and water up to 30ul. The cycling conditions were 98 °C for 30 s; 5 cycles of 98 °C for 10 s, 65 °C for 30 s, and 72 °C for 30 s; and 72 °C for 5 min. These PCR products were purified with a 1.6x AMPure XP (Beckman Coulter, Inc) bead cleanup and eluted into 52 µL of low TE. These individual libraries were quantified with a KAPA Library Quantification Low Rox Kit (Roche Molecular Systems, Inc.) on qPCR System and pooled in equimolar amounts before being combined with 10% Onso indexed library control (PacBio 102-529-900) assessed with a High Sensitivity D1000 ScreenTape (Agilent Technologies, Inc.) on a 4200 TapeStation system (Agilent Technologies, Inc.), and sequenced on a prototype Onso instrument in a single read, 150 cycle configuration (pre-commercial reagents).

Targeted amplicon sequencing analysis

fastQ.gz files were subsampled using "seqtk" (1.3) [21], leading to 20,000 or 100,000 single reads for SBB and 20,000 or 100,000 paired reads for SBS. After subsampling, the contrived population mixtures were analyzed using the Amplicon Sequencing Analysis Pipeline (ASAP) software (1.9.0) [13, 22]: a customized ASAP JavaScript Object Notation (JSON) file was created for these *katG* and *gyrA* nucleotide locations and the reads were trimmed of any adapter with any less than 80 nt being removed by bbduk [23]. Then the trimmed reads were aligned to the target amplicon references using Bowtie2 [24] and the resultant BAM files were analyzed for single nucleotide polymorphisms (SNPs) following the specifications in the JSON file and user defined thresholds including SMOR analysis. ASAP outputs an XML file containing all of the results. Utilizing an RStudio (2023.03.1) and R (4.3.0) installation on a high-performance computing cluster, the ASAP XML file was parsed using custom scripts. The "tidyverse" (2.0.0) [25], "ggpubr" (0.6.0) [26], and "Hmisc" (5.1-0) [27] packages were used within the analysis.

Full and linear range plots

For each method, gene, and read sampling by plotting the observed SNP percentage compared to the theoretical percentage. Linear models were used to calculate the R^2 values and to create 95% confidence intervals for the associated linear range. In addition, Supplemental Fig. 2 was created to identify the number of unique fragments needed to statistically validate a 0% SNP population. P-values were calculated for each SNP population and unique sequenced fragments using a binomial distribution accessed through the "binom.test" function within R.

Error rate analysis

Using the 100k read sampling files, regions where all three methods had more than 10,000X coverage were identified within *katG* and *gyrA*. Once these regions were identified, the per-base error rate was calculated across each base pair by classifying SNP errors as either transitions, transversions, insertions, or deletions. In addition to these classifications, the total error rate was also calculated.

$$Total\ Error\ Rate\ (\%)_{\ BP\ Position} = \left(\frac{Non\ Reference\ Calls_{\ BP\ Position}}{Total\ Depth_{\ BP\ Position}}\right) \times 100$$

After the per-base error rate was calculated, the mean error rate for each gene was calculated. In addition, due to the bounded nature of proportions, confidence intervals for error rate were calculated via nonparametric bootstrapping with the "Hmisc" package [27]. Finally, individual error rates were compared across methods using a t-test.

SNP specificity analysis

SNP specificity was measured by identifying false positive mutations for each sequencing method (SBS, SBS-SMOR, and SBB). Similar to the error analysis, the 100k read sampling files were used, and only regions with more than 1,000X coverage were used. Once regions were identified, mutations were classified into 6 groups (e.g., >10%, >1–10%, >0.1-1%, >0.01–0.1%, >0.001–0.01%, and <0.001%) based on the observed proportion. The average number of false positive SNPs per gene and sample were calculated along with a standard 95% confidence interval.

Results and discussion

We assess the capabilities of targeted SBB sequencing to identify and quantify rare subpopulations, compared to SMOR error-corrected and non-error-corrected targeted SBS sequencing. Thirty contrived *Mtb* mixtures were created to investigate the performance of these different sequencing and informatics approaches. 10%, 1%, 0.1%. 0.01%, and 0.001% proportions containing either *katG* g944c or *gyrA* a241g were made in triplicate using three different 10-fold dilution series, so each sample was generated in triplicate. In addition to these mixtures, the component samples were sequenced in singlet as controls representing 100% and 0% populations. Each mixture was sequenced with SBS and SBB (see Methods). The SBS sequencing data was analyzed both with no error correction (SBS) and using SMOR error correction (SBS-SMOR) and SBB was analyzed with no error correction. The fastQ files were downsampled for more consistent analysis.

For 100k read sampling, this yielded an average depth of ~180k for SBB (due to the paired reads), ~80k for SBS-SMOR, and ~97k for SBB (Supplemental Table 1). A higher percentage of sequenced reads were discarded during quality control and SMOR error correction than in the non-error corrected analysis. Thus, to remain comparative in read depth with the SBB chemistry, 8–13% additional paired reads may be required for the SBS-SMOR.

Even visually, substantial variation was seen in the total number of variants detected by each technology (i.e., SBS reads demonstrably accumulated more variants than SBB reads) (Fig. 1). We first evaluated the expected versus observed mutation proportions and their associated linear regressions at 20,000x depth, a typical depth for previous *Mtb* mutational analyses [12, 15, 17, 19], and 100,000x depth, to analyze ultra-low frequency mixtures. SBS, SBS-SMOR and SBB gave similar results for minor populations down to 0.1% for *katG* g944c, though SBB demonstrated an extended linear range for variant detection down to 0.01% (Supplemental Fig. 1A-C, 20,000x). The 100,000x depth improved the ability to detect the ultra-low variant frequencies for *katG* g944c down to 0.01% for SBS-SMOR and 0.001% for SBB (Supplemental Fig. 1D-F). For *gyrA* a241g, we observed similar regressions for all three conditions, with a linear range down to 0.1% (Supplemental Fig. 1G-H). The additional sequencing depth did not improve the ability to detect lower mutation frequencies for *gyrA* a241g with SBS, SBS-SMOR, or SBB (Supplemental Fig. 1I-J).

The sampling depth required to detect the lowest population proportion (i.e., 0.001%) at >95% confidence would require >376,435 original template reads, which is ~3.5 fold greater than the depth analyzed here, limiting the ability to assess performance of the 0.001% sample (Figure S2). Thus, a zero count within 100,000 reads for *gyrA* (i.e., 0/100,000 reads, variant not detected) with an expected minor population of 0.001% should be interpreted with caution. All three replicates for 0.001% frequency variants were detected for *katG* SBB (Supplemental Table 2, Supplemental Data 2), however, this dilution was not included in the linear regressions due to anticipated stochasticity within the 0.001% samples and within the 0% sample. The root cause for the difference in performance between the *gyrA* and *katG* amplicons is uncertain. Follow-up studies with the more recent versions of chemistry, more diverse samples, and an increased number of low percentage replicates should be prioritized to further understand these results.

We calculated the observed error rate across each sequencing method per position across conserved gene regions within the 100,000 read depth. Errors were defined as a deviation from the plasmid sequence which results from both the combination of polymerase errors in the PCR and sequencing errors. In this study, the two error mechanisms cannot be distinguished and both contribute to the analysis. This results in overstating the value of the sequencing error for all platforms, although a relative comparison can still be made. As the samples were amplified and then split for sequencing, with the SBB samples undergoing five additional cycles of error, the polymerase error will contribute slightly more to the SBB analysis, but is not corrected for in this analysis, and is assumed to be small. Within the genes in this study, SBS with no error correction had an average total error of 0.34% and is comparable to previous estimates that performed a more in-depth analysis [28] (Supplemental Table 3). The SNP observed errors are significantly higher than insertion and deletion observed errors in all three analyses (consistent with previous analysis of SBS [29], however, the finite sequence space utilized in this analysis should be generalized across larger regions with caution. Overall, SBS-SMOR and SBB showed an 8.3X and 8.5X reduction in observed errors, respectively, compared to SBS alone (Fig. 2), but the difference in SBB and SBS-SMOR observed error rates is not significant ($p=0.08$) (Supplemental Table 3).

In addition to the observed error rate analysis, we assessed the number of false-positive SNPs (compared to the plasmid reference) that were detected in each gene fragment and method across six thresholds. No method detected false positive SNPs at a frequency greater than 2.1%. False positives were substantially higher in uncorrected SBS reads when compared to SBS-SMOR and SBB (Supplemental Table 4). These results are consistent with past observations suggesting minor population detection without any error corrections is limited to between 1 and 10% for uncorrected SBS sequencing [13, 28]. SBS-SMOR had the lowest false positive SNP detections

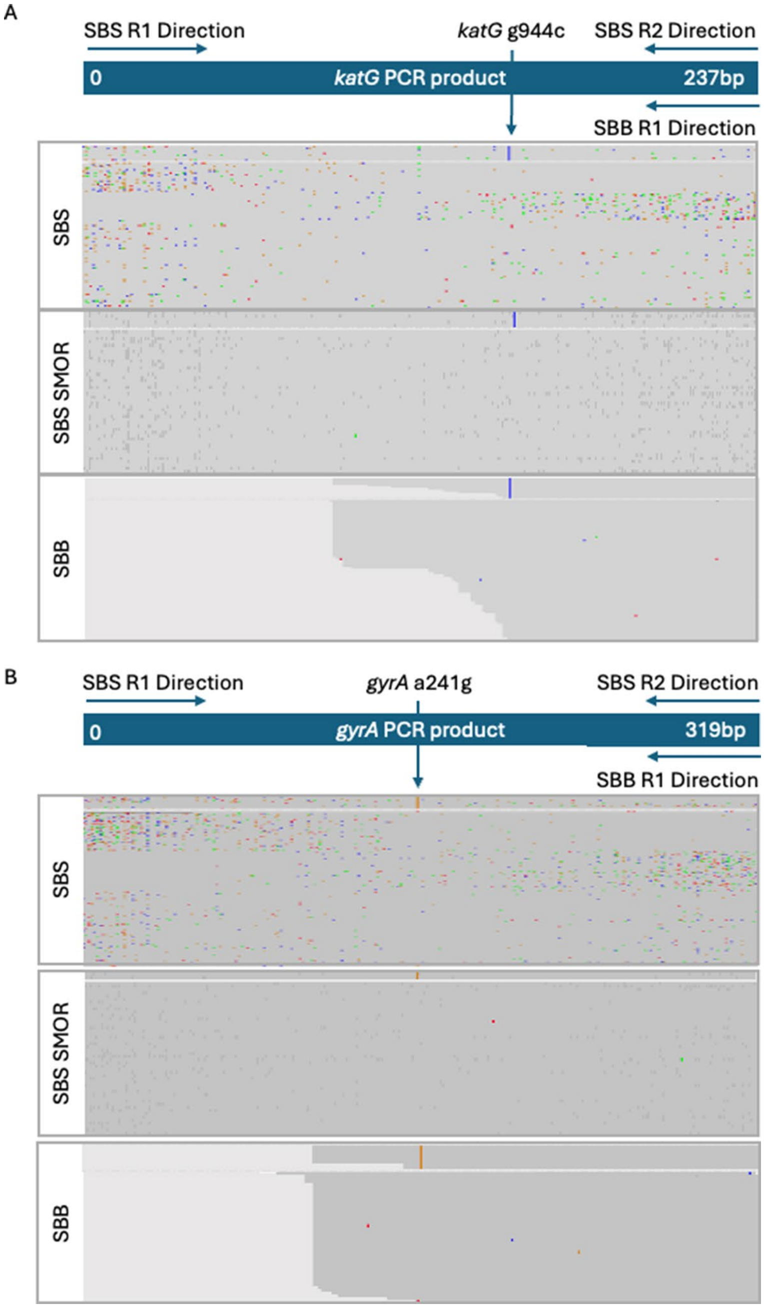

Fig. 1 Alignment of SBS, SBS-SMOR, and SBB reads for *katG* **(A)** and *gyrA* **(B)** Cartoon at the top of each IGV plot shows the relative position of the sequencing reads where SBS had both read 1 and read 2 each at 300 bps, while SBB has only read 1 from the 3' end with variable lengths up to 150 bps long. Images in IGV are representatives of the the 10% proportion mixture and each read is colored grey, in squished mode, and are grouped by *katG* PCR position 224 (gene mutation: g944c) or *gryA* PCR position 197 (gene mutation: a241g). Mutations compared to the plasmid reference include single nucleotide variants, A(Green), G (Orange), C (Blue), T (Red), and N (Dark Grey)

in the >0.1-1% range for both *gyrA* (2.6 SBS-SMOR vs. 10.8 in SBB) and *katG* (4.8 vs. 9.6, SBS-SMOR vs. SBB); however, this pattern did not continue below this cutoff range with SBB outperforming SBS-SMOR at lower thresholds. These false positives may be due to sequencing errors, but PCR errors and DNA replication errors during plasmid replication may also be factors at these lower thresholds and will need to be investigated in further studies.

Conclusions

Detecting rare genetic variants (i.e., <1% minor subpopulation) in next-generation sequencing data is a critical need for applications ranging from identifying antibiotic

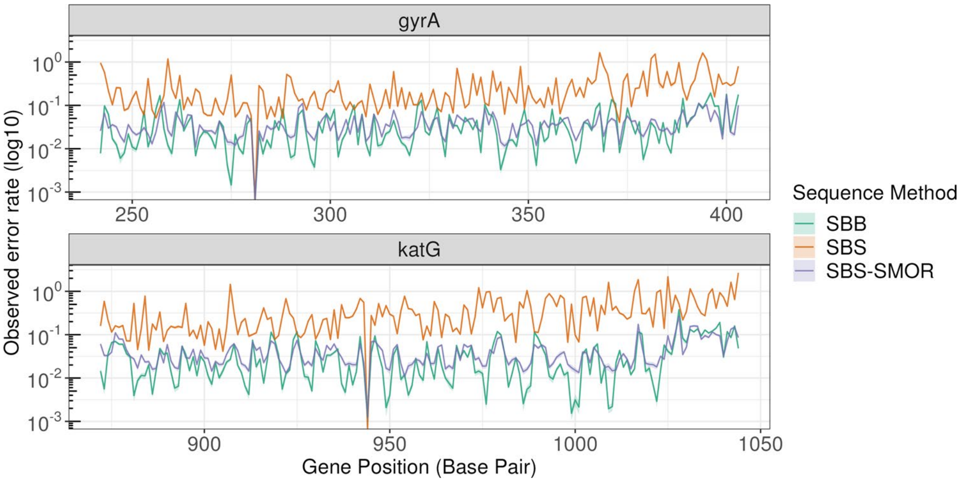

Fig. 2 Average observed % error rate (log 10) across *gyrA* and *katG* genes for comparable positions. The line represents the mean total error rate across samples at the specified position with a 95% confidence interval displayed as the lighter-colored ribbon

micro-heteroresistance in bacteria to detecting low-variant genetic mutations in oncology. Our results with contrived mixtures for an *Mtb* model system to characterize ultra-low genetic variants demonstrated SBB sequencing chemistry detected target SNPs down to 0.01% at 100,000x depth and 0.1% at 20,000x depth, without any error correction methods. Traditional SBS sequencing is unable to achieve this accuracy without the use of sophisticated error correction tools (e.g., SMOR, UMI, duplex, and others). Both SBB and SBS-SMOR analysis resulted in more than 8-fold decrease of overall error rate compared to SBS in this experiment. A broader study integration of error correction methods with SBB sequencing is outside the scope of this study, but should be investigated in future publications as it has the potential to significantly decrease the sequencing error rate even further. SBB sequencing looks very promising for both targeted and unbiased whole genome sequencing, as its innate accuracy can be used on non-amplicon reads to detect minor subpopulations with confidence to less than 1% of the total population.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10697-1.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Acknowledgements

We want to thank Nairi Pezeshkian and Daniel Nasko of Pacific Biosciences (Menlo Park, California), for their assistance in conducting SBB sequencing and initial data handling. We also thank James Schupp for his early discussions and recommendations at the inception of this study.

Author contributions

C.A. and C.W. contributed to study design, conducted research and analysis and wrote the main manuscript; D.E., D.L. and T.P. contributed to analysis and manuscript writing; S.P. contributed to study design, analysis, and manuscript writing; and D.E. contributed to study funding, study design, analysis, and manuscript writing. All authors reviewed the manuscript.

Funding

Support for this study was provided in part by NIH NIAID grants R01AI153213 (Engelthaler) and R01AI131939 (Engelthaler). SBB sequencing reagents and sequencing service was provided by Pacific Biosciences.

Data availability

All sequence data is available at NCBI: BioProject ID PRJNA1111863.

Declarations

Ethics approval and consent to participate

Not applicable since this work did not use any animal or human data or tissue.

Consent for publication

Not applicable since this work did not contain data from any individual person.

Competing interests

The authors declare no competing interests.

Received: 9 May 2024 / Accepted: 9 August 2024
Published online: 19 August 2024

References

1. Manley LJ, Ma D, Levine SS. Monitoring error rates in Illumina sequencing. J Biomol Tech. 2016;27:125–8.
2. Chakravorty S, Simmons AM, Rowneki M, Parmar H, Cao Y, Ryan J et al. The new Xpert MTB/RIF Ultra: improving detection of Mycobacterium tuberculosis and resistance to rifampin in an assay suitable for point-of-care testing. MBio. 2017;8.
3. Rinder H, Mieskes KT, Löscher T. Heteroresistance in Mycobacterium tuberculosis. Int J Tuberc Lung Dis. 2001;5:339–45.
4. Lee JH, Garg T, Lee J, McGrath S, Rosman L, Schumacher SG, et al. Impact of molecular diagnostic tests on diagnostic and treatment delays in tuberculosis: a systematic review and meta-analysis. BMC Infect Dis. 2022;22:940.

5. Soedarsono S, Mertaniasih NM, Hasan H, Kusmiati T, Permatasari A, Kusumaningrum D, et al. Line probe assay test in new cases of tuberculosis with rifampicin resistance not detected by Xpert MTB/RIF. Int J Mycobacteriol. 2022;11:429–34.
6. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. 2011;39:e81.
7. Stoler N, Arbeithuber B, Guiblet W, Makova KD, Nekrutenko A. Streamlined analysis of duplex sequencing data with Du Novo. Genome Biol. 2016;17:180.
8. Chen-Harris H, Borucki MK, Torres C, Slezak TR, Allen JE. Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs. BMC Genomics. 2013;14:96.
9. Sloan DB, Broz AK, Sharbrough J, Wu Z. Detecting rare mutations and DNA damage with sequencing-based methods. Trends Biotechnol. 2018;36:729–40.
10. Pel J, Choi WWY, Leung A, Shibahara G, Gelinas L, Despotovic M, et al. Duplex proximity sequencing (Pro-Seq): a method to improve DNA sequencing accuracy without the cost of molecular barcoding redundancy. PLoS ONE. 2018;13:e0204265.
11. Ren Y, Zhang Y, Wang D, Liu F, Fu Y, Xiang S, et al. SinoDuplex: an improved duplex sequencing approach to detect low-frequency variants in plasma cfDNA samples. Genomics Proteom Bioinf. 2020;18:81–90.
12. Metcalfe JZ, Streicher E, Theron G, Colman RE, Allender C, Lemmer D, et al. Cryptic Microheteroresistance explains Mycobacterium tuberculosis phenotypic resistance. Am J Respir Crit Care Med. 2017;196:1191–201.
13. Colman RE, Schupp JM, Hicks ND, Smith DE, Buchhagen JL, Valafar F, et al. Detection of low-level mixed-population drug resistance in Mycobacterium tuberculosis using high fidelity amplicon sequencing. PLoS ONE. 2015;10:e0126626.
14. Goossens SN, Heupink TH, De Vos E, Dippenaar A, De Vos M, Warren R et al. Detection of minor variants in Mycobacterium tuberculosis whole genome sequencing data. Brief Bioinform. 2022;23.
15. Metcalfe JZ, Streicher E, Theron G, Colman RE, Penaloza R, Allender C et al. Mycobacterium tuberculosis subculture results in loss of potentially clinically relevant heteroresistance. Antimicrob Agents Chemother. 2017;61.
16. Shin SS, Modongo C, Baik Y, Allender C, Lemmer D, Colman RE, et al. Mixed Mycobacterium tuberculosis-strain infections are associated with poor treatment outcomes among patients with newly diagnosed tuberculosis, independent of pretreatment heteroresistance. J Infect Dis. 2018;218:1974–82.
17. Wang Q, Modongo C, Allender C, Engelthaler DM, Warren RM, Zetola NM et al. Utility of targeted, Amplicon-based deep sequencing to detect resistance to first-line tuberculosis drugs in Botswana. Antimicrob Agents Chemother. 2019;63.
18. Whitfield MG, Engelthaler DM, Allender C, Folkerts M, Heupink TH, Limberis J, et al. Comparative performance of genomic methods for the detection of pyrazinamide resistance and heteroresistance in Mycobacterium tuberculosis. J Clin Microbiol. 2022;60:e0190721.
19. Engelthaler DM, Streicher EM, Kelley EJ, Allender CJ, Wiggins K, Jimenez D, et al. Minority Mycobacterium tuberculosis genotypic populations as an indicator of subsequent phenotypic resistance. Am J Respir Cell Mol Biol. 2019;61:789–91.
20. Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, Emerson C, et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. Eur Respir J. 2017;50:1701354.
21. Li H. seqtk. 2018.
22. Bowers JR, Lemmer D, Sahl JW, Pearson T, Driebe EM, Wojack B, et al. KlebSeq, a diagnostic tool for surveillance, detection, and monitoring of Klebsiella pneumoniae. J Clin Microbiol. 2016;54:2582–96.
23. Bushnell B. BBMap short read aligner, and other bioinformatic tools.
24. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
25. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. J Open Source Softw. 2019;4:1686.
26. Kassambara A, ggpubr. ggplot2 Based Publication Ready Plots. 2023.
27. Harrell FE. Hmisc: Harrell Miscellaneous. 2023.
28. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genom Bioinform. 2021;3:lqab019.
29. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016;17:125.

Publisher's Note