the geneticssociety

www.nature.com/hdy

ARTICLE

Check for updates

Quantifying the effects of computational filter criteria on the accurate identification of de novo mutations at varying levels of sequencing coverage

Mark Milhaven[1,2], Aman Garg[1], Cyril J. Versoza[1,2] and Susanne P. Pfeifer[1,2]✉

The rate of spontaneous (de novo) germline mutation is a key parameter in evolutionary biology, impacting genetic diversity and contributing to the evolution of populations and species. Mutation rates themselves evolve over time but the mechanisms underlying the mutation rate variation observed across the Tree of Life remain largely to be elucidated. In recent years, whole genome sequencing has enabled the estimation of mutation rates for several organisms. However, due to a lack of community standards, many previous studies differ both empirically – most notably, in the depth of sequencing used to reliably identify de novo mutations – and computationally – utilizing different computational pipelines to detect germline mutations as well as different analysis strategies to mitigate technical artifacts – rendering comparisons between studies challenging. Using a pedigree of Western chimpanzees as an illustrative example, we here quantify the effects of commonly utilized quality metrics to reliably identify de novo mutations at different levels of sequencing coverage. We demonstrate that datasets with a mean depth of ≤ 30X are ill-suited for the detection of de novo mutations due to high false positive rates that can only be partially mitigated by computational filter criteria. In contrast, higher coverage datasets enable a comprehensive identification of de novo mutations at low false positive rates, with minimal benefits beyond a sequencing coverage of 60X, suggesting that future work should favor breadth (by sequencing additional individuals) over depth. Importantly, the simulation and analysis framework described here provides conceptual guidelines that will allow researchers to take study design and species-specific resources into account when determining computational filtering strategies for their organism of interest.

Heredity (2025) 134:273–279; https://doi.org/10.1038/s41437-025-00754-0

INTRODUCTION

Mutations spontaneously occurring in the germline (*i.e.*, de novo mutations) are a double-edged sword in evolution and medicine. On the one hand, mutations can be beneficial, facilitating adaptation to local or changing environments (see the reviews of Fan et al. 2016 and Harris et al. 2020); on the other, mutations can be deleterious, causing genetic disease and developmental disorders (see the review of Acuna-Hidalgo et al. 2016) and contributing to the mutational load in species threatened with extinction (see the review of Agrawal and Whitlock 2012). Consequently, characterizing the rates and patterns by which mutations arise is an important endeavor both evolutionarily – to improve our understanding of the demographic and adaptive history of populations including their historical sizes and inferring the timing of speciation events – and biomedically – to advance our knowledge of the genetic underpinnings of heritable disease.

After decades of research focused on the indirect estimation of mutation rates – using both genetic screens for monogenic Mendelian mutations underlying diseases with major phenotypic effects and phylogenetic approaches relying on differences observed between species (see the review of Pfeifer 2020) – recent advances in high-throughput sequencing have enabled the direct, and relatively unbiased, identification of de novo mutations by searching whole-genome sequences of trios (parents and their offspring) for mutations that constitute Mendelian violations. Using this strategy, germline mutation rates have been estimated for both humans (Roach et al. 2010; Conrad et al. 2011; Campbell et al. 2012; Kong et al. 2012; Michaelson et al. 2012; Jiang et al. 2013; Besenbacher et al. 2015; Francioli et al. 2015; Yuen et al. 2015; Goldmann et al. 2016; Rahbari et al. 2016; Wong et al. 2016; Jónsson et al. 2017; Maretty et al. 2017; Turner et al. 2017; Sasani et al. 2019; Kessler et al. 2020) and closely-related non-human primates (Venn et al. 2014; Pfeifer 2017a; Tatsumoto et al. 2017; Thomas et al. 2018; Besenbacher et al. 2019; Wang et al. 2020, Wu et al. 2020; Bergeron et al. 2021; Campbell et al. 2021; Yang et al. 2021; Versoza et al. 2024) as well as several other vertebrates (Smeds et al. 2016; Feng et al. 2017; Milholland et al. 2017; Martin et al. 2018; Koch et al. 2019; Lindsay et al. 2019; Wang et al. 2022a, b; Bergeron et al. 2023) and invertebrates (Keightley et al. 2014, 2015; Liu et al. 2017). These studies, providing the first direct empirical insights into the germline mutation rate across the Tree of Life, confirmed earlier observations that rates of mutation vary markedly not only between species but also between individuals and populations of the same species (see the review of Baer et al.

[1]School of Life Sciences, Arizona State University, Tempe, AZ 85281, USA. [2]Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85281, USA. Associate editor: Louise Johnson. ✉email: susanne@spfeiferlab.org

Received: 25 August 2023 Revised: 24 February 2025 Accepted: 24 February 2025
Published online: 13 March 2025

2007) – however, the factors driving this variation remain largely to be elucidated.

In addition to the biological and life history factors likely at play, recent work by Bergeron and colleagues (2022) demonstrated that methodological differences in computational pipeline design can affect the reliable detection of de novo mutations from high-throughput sequencing data and thus ultimately mutation rates estimated. Specifically, using a trio of rhesus macaques (*Macaca mulatta*) sequenced at medium to high coverage (40X to 70X per individual), the authors observed a nearly two-fold variation in mutation rate estimates (ranging from 0.46×10^{-8} to 0.85×10^{-8} per base pair per generation) due to differences in computational filtering strategies necessary to distinguish genuine, but often extremely rare, de novo mutations from orders of magnitude more frequent sequencing errors as well as alignment, variant calling, and genotyping artefacts. As studies published to date by different research groups have applied a variety of custom computational pipelines to identify high-confidence de novo mutations (see Supplementary Table 1 in Bergeron et al. 2022), this sensitivity to the implemented filtering strategies inherently limits comparison across studies and organisms.

This, in turn, highlights the pressing need of a standardized community-level consensus and best practices in the development and comparison of computational pipeline designs. Bergeron and colleagues (2022) made an important step towards this goal – however, as their study was based on empirical data for which the "ground truth" was unknown, their evaluations were limited to 43 candidate sites that were validated via PCR amplification and Sanger sequencing. As such, the performance of different computational filters and their thresholds to reliably detect de novo mutations while mitigating false positives remains to be comprehensively benchmarked at a genome-wide scale. Such a benchmarking experiment necessarily requires knowledge of a "ground truth" dataset (see the discussion in Pfeifer 2021), *i.e.*, a dataset for which the positions of all de novo mutations are known a priori – something which is generally not feasible in a laboratory setting due to the limited number of sites that can practically be manually validated. Moreover, despite continuously decreasing costs, high-throughput sequencing of large pedigrees remains an expensive endeavor for many scientific laboratories, making it necessary to carefully consider study design, most notably in terms of the depth of sequencing coverage needed to obtain high-confident de novo mutation call sets. Yet, similar to computational pipeline designs, no community standard currently exists for pedigree-based mutation rate studies with regards to sequencing coverage, with previous study designs ranging from relatively low (< 20X) to ultra-high (> 150X) coverage (Supplementary Fig. S1).

To provide recommendations for future germline mutation rate studies, we here developed a simulation and analysis framework to quantify the effects of commonly applied computational filter criteria and their thresholds on the accurate identification of de novo mutations at varying levels of sequencing coverage, and provide guidance with regards to the minimum coverage necessary for reliable de novo mutation calling from Illumina sequencing data (the *de facto* standard in the field). Using a trio of chimpanzees as an illustrative example, this framework serves as a conceptual guideline for in silico benchmarking for future pedigree-based mutation studies.

MATERIALS AND METHODS

Quantifying the effects of computational filter criteria and thresholds on the accurate identification of de novo mutations at varying levels of sequencing coverage requires a "ground truth" dataset (see the discussion in Pfeifer 2021). To obtain such a dataset, reads were simulated from polymorphism-aware, haplotype-resolved reference assemblies ("*Real data*") and de novo mutations were spiked in at the species-specific mutation rate ("*Simulated data*"). Thereby, the incorporation of polymorphisms acts as "noise" in the simulated data, complicating both the accurate mapping of reads and the identification of de novo mutations in a manner similar to that of genuine resequencing data (Pfeifer 2017b).

Real data

To incorporate realistic levels of polymorphisms in the simulated reads, variants were called from a trio of Western chimpanzees (*Pan troglodytes verus*) previously sequenced to ultra-deep coverage (> 150X) on an Illumina HiSeq 2000 platform (Tatsumoto et al. 2017). Specifically, sequencing data was downloaded from the DNA Data Bank of Japan (BioProject: PRJDB3537) and run through a modified version of the Genome Analysis Toolkit (GATK) Germline Short Variant Discovery pipeline (van der Auwera and O'Connor 2020), using GATK v.4.1.8.1 with default parameters unless noted otherwise. In brief, downloaded .fastq files were intermittently converted to unmapped .bam files (FastqToSam) to mark adapter sequences (MarkIlluminaAdapters). The resulting .bam files were converted back to .fastq files (SamToFastq) before mapping the reads to the species-specific reference assembly (panTro6; Kronenberg et al. 2018) downloaded from NCBI GenBank (accession number: GCA_002880755.3) using BWA-MEM v.0.7.17 (Li and Durbin 2009). Mappings were merged back with the original unmapped .bam files (MergeBamAlignment) to preserve metadata. Next, duplicate reads were marked (MarkDuplicates), indels re-aligned (RealignerTargetCreator and IndelRealigner), and base quality scores recalibrated (BaseRecalibrator and ApplyBQSR) using a previously published variant catalogue of five Western chimpanzees (Prado-Martinez et al. 2013) to mask out sites of expected variation. A second round of duplication marking was performed (MarkDuplicates) prior to calling (HaplotypeCaller) and jointly genotyping variants (GATK v.3.7.0 GenotypeGVCFs), assuming a species-specific heterozygosity rate of 8×10^{-4} (' *--heterozygosity 0.0008* '; Prado-Martinez et al. 2013). The dataset was limited to autosomal biallelic single nucleotide polymorphisms (SNPs) with genotype information in all individuals (SelectVariants with the ' *--restrict-alleles-to BIALLELIC* ', ' *--select-type-to-include SNP* ', and ' *-select 'AN == 6'* ' flags). SNPs were filtered following the GATK Best Practices for hard-filtering germline short variants (applied filter criteria: QD < 2.0, SOR > 3.0, FS > 60.0, MQRankSum < −12.5, and ReadPosRankSum < −8.0; with acronyms as defined by the GATK package) and Mendelian violations removed (FindMendelianViolations). The resulting dataset contained 4,634,632 high-quality autosomal biallelic SNPs, with a transition-transversion ratio of 1.98, similar to previous observations in the species (Auton et al. 2012).

In order to create a polymorphism-aware, haplotype-resolved reference assembly for each individual in the trio, the variant dataset was first phased using BEAGLE v.4.0 (Browning and Browning 2007) and phased variants were then embedded within the species-specific reference assembly (panTro6) using the *vcf2fasta* command built-in vcflib v.1.0.2 (Garrison et al. 2022). In other words, each individual was represented in a haplotype-resolved manner (*i.e.*, by two assemblies, one per haplotype including the corresponding phased variants). Including this haplotype structure in the simulation scheme as detailed below is crucially important as variant callers (such as the Genome Analysis Toolkit used in this study) rely on haplotype information to accurately call and genotype variants.

Simulated data

For each individual in the trio, 10 replicates of 100 bp paired-end reads were simulated from these polymorphism-aware, haplotype-resolved reference assemblies using Mason v.2.0.9 (Holtgrewe 2010) – a simulator that well-mimics genomic characteristics of empirical datasets (for a performance evaluation of several popular short-read simulators, see Milhaven and Pfeifer 2023) – with the '*-oa*' flag enabled to output a "golden" (ground truth) set of mapped reads that specify the regions where the reads originated from. Thereby, sequencing errors were introduced in the simulated reads using Mason's default error model (Holtgrewe 2010). Parents were simulated to 100X coverage whereas the offspring was simulated to a higher coverage (120X) to allow for binomial sampling of reads from regions containing de novo mutations to mimic patterns of genuine heterozygote sites (Supplementary Fig. S2). Next, de novo mutations were introduced in the offspring by randomly sampling sites on the autosomes at the species-specific mutation rate of ~10^{-8} per base pair per generation (Venn et al. 2014; Tatsumoto et al. 2017; Besenbacher et al. 2019) using an in-house script (mutator.py) and introducing mutations at these sites with an allele balance of 1:1 (reference allele vs. alternative allele) using jvarkit v.2021.10.13 (https://github.com/

lindenb/jvarkit) (see Supplementary Table S1 for the sequence coordinates of the de novo mutations that were introduced in each replicate). In order to implement a computational de novo mutation detection pipeline similar to those developed in earlier studies (for an overview, see Bergeron et al. 2022), datasets were converted to .fastq (using GATK v.4.1.8.1 SamToFastq; van der Auwera and O'Connor 2020) – the *de facto* standard format used by many sequencing centers.

De novo mutation detection pipeline

For each replicate, simulated reads were mapped to the species-specific reference assembly (panTro6) using BWA-MEM v.0.7.15 (Li and Durbin 2009) before marking duplicate reads (GATK v.4.1.0.0 MarkDuplicates; van der Auwera and O'Connor 2020). Mapped reads were down-sampled to coverages ranging from 10X to 100X per individual in 10X increments using a previously established pipeline (Milhaven and Pfeifer 2023). For each down-sampled dataset, base quality scores were recalibrated (GATK v.4.1.0.0 BaseRecalibrator and ApplyBQSR), excluding loci known to vary in the population (Prado-Martinez et al. 2013). Mapping accuracy was assessed by comparing the mappings of the simulated reads including the spiked-in de novo mutations in the offspring with the "golden" (ground truth) dataset using an in-house script (mapping_stats.py) (Supplementary Table S1).

Variants were called (GATK v.4.1.8.1 HaplotypeCaller), assuming a species-specific heterozygosity rate of 8×10^{-4} (' *--heterozygosity 0.0008* '; Prado-Martinez et al. 2013). To obtain a high-confidence dataset, only reads with a minimum mapping quality of 40 (corresponding to a probability of 99.99% that a read was mapped correctly) were considered during variant calling (' *--minimum-mapping-quality 40* ') and down-sampling of reads that shared the same start position was disabled to take all high-quality reads at any given site into account (' *--max-reads-per-alignment-start 0* '). In addition, as reads were simulated without PCR bias, the ' *--pcr-indel-model* ' was set to ' *NONE* ' to disable PCR error correction on variant likelihoods. Next, variants were jointly genotyped (GATK v.3.7.0 GenotypeGVCFs), and the dataset was limited to biallelic SNPs with genotype information in all individuals (GATK v.4.1.8.1 SelectVariants with the ' *--restrict-alleles-to BIALLELIC* ', ' *--select-type-to-include SNP* ', and ' *--select 'AN == 6* ' flags). From this dataset, de novo mutation candidates were identified by selecting sites at which the offspring was heterozygous despite both parents being homozygous for the reference allele (GATK v.4.1.8.1 SelectVariants ' *--select 'vc.getGenotype("Offspring").isHet()' --select 'vc.getGenotype("Sire").isHomRef()' --select 'vc.getGenotype("Dam").isHomRef()'* ').

Computational filter criteria and thresholds

With de novo mutation candidates on hand, it was next necessary to apply computational filter criteria to differentiate genuine de novo mutations from technical artefacts resulting from both sequencing errors as well as errors introduced by the de novo mutation detection pipeline (mapping, variant calling, and genotyping) (Pfeifer 2021). In order to determine suitable computational filtering strategies at varying levels of sequencing coverage, the sensitivity and specificity of several commonly used computational filter criteria and thresholds were assessed (thresholds were varied between the minimum and maximum values listed in Supplementary Table S2): (1) the probability that a site is variable among the trio (variant confidence; QUAL), (2) the probability that a site is genotyped correctly in an individual of the trio (genotype quality; GQ), (3) the scaled depth of coverage at a site in the trio (depth; DP), (4) the alternative allele depth in the parents (allele depth; AD), (5) the proportion of reads that support the alternative allele relative to the total depth of coverage at a site in the offspring (allele balance; AB), and (6) GATK Best Practices hard-filter criteria for germline short variants (filter criteria: QD < 2.0, SOR > 3.0, FS > 60.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0; with acronyms as defined by the GATK package). The effect of removing sites known to segregate in the population was assessed based on a recently published variant catalogue of 11 Western chimpanzees (Brand et al. 2022). In addition to these commonly used computational filter criteria, we tested whether the information obtained during the variant calling step could be harnessed to distinguish between genuine de novo mutations and technical artefacts. Specifically, to improve accuracy in regions exhibiting evidence of variation, GATK's HaplotypeCaller builds a graph-based de novo assembly from the sequencing reads to construct candidate haplotypes, against which it then re-aligns the reads to call variants and assign genotypes (van der Auwera and O'Connor 2020). We postulated that the depth of coverage in such reassembled regions in the parents – which, at sites of genuine de novo mutations, should be homozygous for the reference allele and thus, not display any variation – might be a suitable additional filter (referred to from hereon as "reassembly" filter) to help tease apart genuine de novo mutations from technical artefacts.

Each computational filter criterion was first assessed individually in order to determine the best threshold (defined here as the threshold that mitigated the largest number of false positives while retaining the majority of genuine de novo mutations) and benchmark its sensitivity and specificity at varying levels of sequencing coverage (see Supplementary Figs. S3-S12 for replicates 1-10). Afterward, the best-performing filter criteria and thresholds were applied sequentially in order of their effectiveness (see Supplementary Table S3 for the best thresholds in replicates 1-10 and Supplementary Table S4 for the best thresholds across replicates) to obtain a high-confidence set of de novo mutation candidates for each replicate (see Supplementary Figs. S13-S22 for information regarding the datasets resulting from this sequential application of the best performing filter criteria in replicates 1-10).

Visual curation

The validation of de novo mutation candidates is a critical aspect of mutation rate studies given the often high false positive rates. Given that PCR validation and Sanger sequencing is both costly and difficult in non-model organisms, many previous studies employ alternative validation strategies; for example, the visual inspection of a high-confidence set of de novo mutation candidates to distinguish genuine de novo mutations from false positives based on read alignments and available variant calling / genotyping information. As an illustrative example of the feasibility of this approach, the candidate de novo mutations from the highest coverage (100X) data were curated visually using the Integrative Genomics Viewer v.2.14.0 (IGV; Robinson et al. 2011) focusing on candidate sites and surrounding 20 bp regions (IGV screenshots are provided as Supplementary Material at the GitHub repository: https://github.com/PfeiferLab/DNM_coverage). To avoid any potential biases, visual curation was performed independently by three lab members without prior knowledge of whether a site contained a genuine de novo mutation.

RESULTS AND DISCUSSION

In order to study the effects of computational filter criteria and thresholds on the detection of de novo mutations at varying levels of sequencing coverage, 10 replicates of short-read Illumina sequencing data (the *de facto* standard in the field) were simulated from polymorphism-aware, haplotype-resolved reference assemblies for a trio of Western chimpanzees (*Pan troglodytes verus*) to mean depths ranging from 10X to 100X per individual (in 10X increments) and de novo mutations were introduced at random in the offspring at a species-specific mutation rate of $\sim 10^{-8}$ per base pair per generation, resulting in an average of 52 de novo mutations (range: 35-65 de novo mutations) across the autosomal genome per replicate (Supplementary Table S1). After mapping the simulated reads of each individual to the species-specific reference genome (panTro6), variants were called and genotyped using the Genome Analysis Toolkit following standard best practices in the field (for details, see "Materials and Methods"). Next, candidate de novo mutations were identified as SNPs at which the offspring was heterozygous despite both parents being homozygous for the reference allele.

Notably, none of the lowest (10X) coverage datasets captured all introduced de novo mutations (see Supplementary Figs. S3-S12 for replicates 1-10) – with the missing de novo mutations either having been mis-called as homozygous for the reference allele or alternative allele in the offspring, or being absent from the dataset as their supporting reads exhibited mapping quality scores below the calling threshold – strongly suggesting that a mean depth of 10X is insufficient to comprehensively identify de novo mutations at a genome-wide scale. Similarly, the majority of the replicates at 20X coverage missed one or more discoverable de novo mutations. All other datasets (30X-100X) contained all discoverable de novo mutations, *i.e.*, all de novo mutations with the exception of those randomly introduced in inaccessible (gap)

regions of the reference genome and those mis-called / mis-genotyped by the variant caller (4 and 5 out of 524 de novo mutations introduced across the 10 replicates, respectively; Supplementary Table S1). However, the number of false positives – that is, segregating polymorphisms or sequencing errors mis-classified as de novo mutations – varied widely, between ~75k in the raw (unfiltered) low (10X) coverage dataset (standard deviation [sd]: 1,183), ~1.3k in the raw medium (50X) coverage dataset (sd: 35), and ~750 in the raw high (100X) coverage dataset (sd: 29; Supplementary Figs. S3-S12), highlighting the importance of sequencing depth to limit spurious de novo mutation calls.

To reduce the number of false positives in the call set, candidate sites were filtered based on a variety of commonly utilized quality statistics and sequence metrics. Out of the tested computational filter criteria and thresholds (Supplementary Table S2), the genotype quality (GQ) score – assessing the probability that a site is genotyped correctly in an individual – showed the largest effect on reducing the false positive rate, with the best threshold (defined here as the threshold that filtered out the largest number of false positives while retaining the majority of genuine de novo mutations in a call set; see Supplementary Table S3 for the thresholds applied to replicates 1-10 and Supplementary Table S4 for the best threshold across replicates) reducing the false positive rate in the low (30X), medium (50X), and high (100X) coverage datasets by an average of 93.3% (sd: 3.4%), 97.8% (sd: 1.9%), and 96.6% (sd: 0.9%), respectively. Relatedly, the variant confidence (QUAL) score – assessing the probability that a site is variable among the trio – exhibited the second largest effect for datasets with a minimum coverage of 30X, lowering the false positive rate in the 30X, 50X, and 100X coverage datasets by 44.5% (sd: 13.7%), 75.5% (sd: 6.3%), and 87.3% (sd: 2.3%), respectively. However, as each individual read contributes to the score, scores reported by the variant caller appeared artificially inflated in the high coverage ($\geq$ 70X) datasets. Furthermore, the application of a QUAL filter negatively impacted the low coverage (10X and 20X) datasets by removing genuine de novo mutations, likely due to the small number of reads at each site (Supplementary Figs. S3-S12).

Although a combined filtering on GQ and QUAL scores was sufficient to eliminate the vast majority (mean: 98.5%; sd: 0.7%) of false positives in the highest coverage (100X) datasets (Supplementary Figs. S13-S22), additional filter criteria were required for all other datasets. Specifically, both the lower threshold of the overall depth of coverage (DP_{min}) – reflecting the power to accurately determine the genotype of an individual at a given site – and the lower threshold of the proportion of reads that support the alternative allele in the offspring relative to the total depth of coverage at a given site (AB_{min}) played an important role in further decreasing the false positive rates. Additional filtering using the alternative allele depth in the parents (AD) – guarding against the alternative allele being carried by one of the parents and hence the candidate site not constituting a Mendelian violation – as well as upper thresholds on the allele balance in the offspring (AB_{max}) and the overall depth of coverage (DP_{max}) – guarding against mis-genotyped sites and mis-called variants in regions of unresolved paralogs in the reference genome (for a discussion, see Pfeifer 2017b) – aided the further reduction of false positives, particularly in the low coverage (10X to 30X) datasets (Supplementary Figs. S13-S22). Information about variants known to segregate in the population (here using a recently published variant catalogue of 11 Western chimpanzees; Brand et al. 2022) was also effective in removing spurious de novo mutation candidates (Supplementary Figs. S3-S12) – however, the application of this filter criterion did not provide any additional benefit beyond the commonly used computational filter criteria described above. Moreover, it should be noted that this procedure may lead to the exclusion of genuine de novo mutations, particularly in genomic regions experiencing high mutation rates (such as CpG sites; Hwang and Green 2004) where the assumption of an infinite sites model is likely violated. In contrast, the application of GATK's Best Practice hard-filter criteria, frequently utilized to obtain high-quality germline variant calls (see Supplementary Table 1 in Bergeron et al. 2022), led to the exclusion of genuine de novo mutations in several of the 10X and 20X datasets.

Overall, a joint application of these commonly used computational filter criteria (Fig. 1, and see Supplementary Table S4 for the best thresholds across replicates) successfully reduced the false positive rates to an average of 24.5% in the 40X dataset, 14.2% in the 50X, and between 9.1% and 11.5% in the 60X to 100X datasets (Fig. 2a, and see Supplementary Table S5 for details regarding each individual replicate). The additional application of the newly developed reassembly filter further aided the reduction of false positive rates to an average of 22.6% in the 40X dataset, 12.1% in

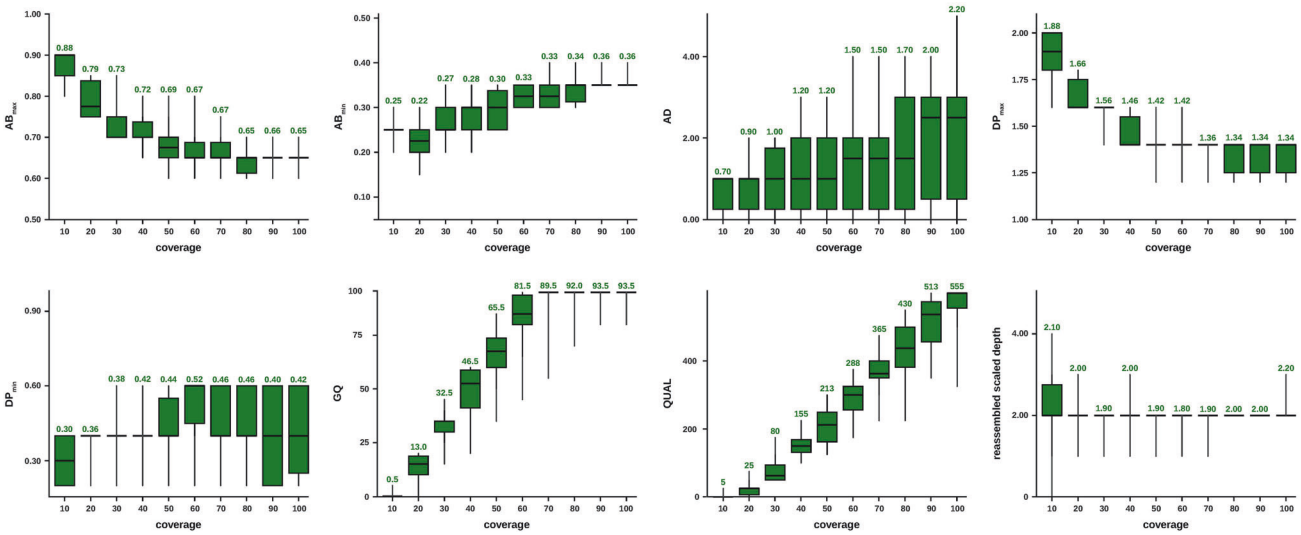

Fig. 1 Computational filter criteria and thresholds that mitigated the largest number of false positives while retaining the majority of genuine de novo mutations at varying levels of sequencing coverage across replicate runs. Filter criteria include a minimum/maximum allele balance ($AB_{min/max}$) filter in the offspring, an alternative allele depth (AD) filter in the parents, a minimum/maximum depth of coverage ($DP_{min/max}$) filter in the trio, a genotype quality (GQ) filter in the trio, a variant confidence (QUAL) filter in the trio, and a reassembly filter in the parents (additional details are provided in Supplementary Table S4).

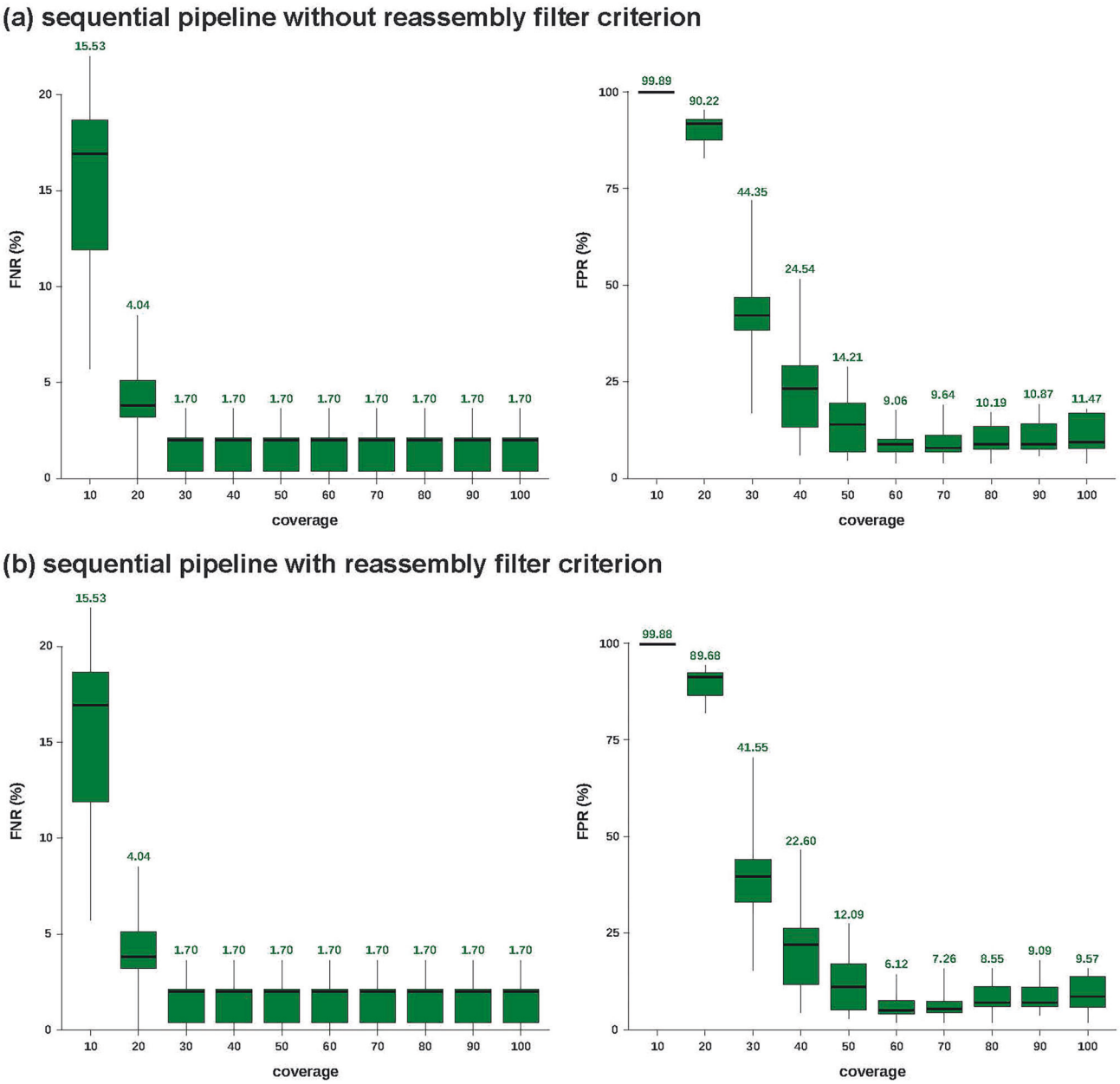

Fig. 2 Misclassification rates. False negative rates (FNRs) and false positive rates (FPRs) after the sequential application of the best computational filter criteria and thresholds at varying levels of sequencing coverage (**a**) without and (**b**) with the newly developed reassembly criterion (additional details regarding each individual replicate are provided in Supplementary Table S5).

the 50X, and between 6.1% and 9.6% in the 60X to 100X datasets (Fig. 2b). Interestingly though, the false positive rates slightly increased in datasets with coverages above 60X, likely due to the larger number of reads supporting spurious de novo mutation candidates, thus falsely boosting the confidence of the variant caller, particularly in (frequently misaligned) repetitive regions of the genome. In contrast, despite the application of several filter criteria, false positive rates remained high in the low coverage datasets at an average of 99.9%, 90.2%, and 44.4% at a mean depth of 10X, 20X, and 30X, respectively. Furthermore, the false negative rates at the lowest coverages (10X and 20X) was non-negligible (on average 15.5% and 4.0%, respectively).

Taken together, our study suggests that datasets with mean depths of ≤ 30X are ill-suited for the detection of de novo mutations due to both high false negative and high false positive rates, the latter of which can only be partially mitigated by commonly applied quality statistics. In contrast, medium coverage (40-60X) data together with minimal filtering enables the comprehensive and reliable identification of de novo mutations at low false positive rates, resulting in high-confidence candidate sets that can be further validated experimentally using an orthogonal sequencing technology (such as Sanger sequencing) or visually by inspecting the genomic regions of interest. For the sake of an example, three researchers independently visually investigated the high-confidence set of de novo mutation candidates in the highest coverage (100X) data of this study. All investigators correctly distinguished between genuine de novo mutations and technical artifacts due to sequencing, mapping, calling, and genotyping errors frequently located in low-complexity and repetitive regions of the genome (Supplementary Fig. S23 provides an example of the IGV screenshots used for visual curation; the complete series of screenshots is provided as

Supplementary Materials at the GitHub repository: https://github.com/PfeiferLab/DNM_coverage), demonstrating the power of manual curation for validation. Lastly, confirming earlier work suggesting that increased sequencing depth might exhibit marginal effects on mitigating false positive calls in empirical data (Koch et al. 2019; Wu et al. 2020), no improvement in de novo mutation calling was observed at higher levels of coverage (70-100X), indicating that a point of diminishing returns likely exists for any given study design. We thus recommend future studies to favor breadth (by sequencing additional individuals) rather than depth beyond a mean coverage of 40-60X.

Closing thoughts

Advances in high-throughput sequencing have enabled the direct detection of de novo mutations from whole-genome pedigree data – however, methodological differences and a lack of community standards currently prohibit meaningful comparisons across studies and organisms. Looking forward, it will likely remain challenging to achieve consistency in the field, particularly with regards to study design, which depends upon both resources available for the species of interest – most notably, the contiguity, completeness, and correctness of the reference genomes used in the analyses which varies between model and non-model organisms – as well as economic factors – such as the costs associated with sampling and sequencing. Nevertheless, the conceptual guidelines presented here using a trio of Western chimpanzees as a case study provide important in silico benchmarks for future pedigree-based mutation studies. Specifically, the developed simulation and analysis framework will allow researchers to stratify the performance of computational filter criteria by study design – taking into account potential biases arising from library preparation, sequence platform-specific read lengths and error rates, as well as depth of coverage – and publicly available genomic resources for any species of interest.

Data archiving

The versions, settings, and parameters of the software used in this study are described in the Materials and Methods section; all custom scripts used in the analyses are available at the GitHub repository: https://github.com/PfeiferLab/DNM_coverage. Analyses were based on whole-genome sequencing data of a trio of Western chimpanzees available from the DNA Data Bank of Japan (BioProject: PRJDB3537) and the chimpanzee reference assembly (panTro6) available from NCBI GenBank (accession number: GCA_002880755.3).

REFERENCES

Acuna-Hidalgo R, Veltman JA, Hoischen A (2016) New insights into the generation and role of de novo mutations in health and disease. Genome Biol 17(1):241

Agrawal AF, Whitlock MC (2012) Mutation load: the fitness of individuals in populations where deleterious alleles are abundant. Annu Rev Ecol Evol Syst 43:115–135

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T et al. (2012) A fine-scale chimpanzee genetic map from population sequencing. Science 336(6078):193–198

Baer CF, Miyamoto MM, Denver DR (2007) Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet 8(8):619–631

Bergeron LA, Besenbacher S, Bakker J, Zheng J, Li P, Pacheco G et al. (2021) The germline mutational process in rhesus macaque and its implications for phylogenetic dating. GigaScience 10(5):giab029

Bergeron LA, Besenbacher S, Turner T, Versoza CJ, Wang RJ, Price AL et al. (2022) The Mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. Elife 11:e73577

Bergeron LA, Besenbacher S, Zheng J, Li P, Bertelsen MF, Quintard B et al. (2023) Evolution of the germline mutation rate across vertebrates. Nature 615(7951):285–291

Besenbacher S, Hvilsom C, Marques-Bonet T, Mailund T, Schierup MH (2019) Direct estimation of mutations in great apes reconciles phylogenetic dating. Nat Ecol Evol 3(2):286–292

Besenbacher S, Liu S, Izarzugaza JM, Grove J, Belling K, Bork-Jensen J et al. (2015) Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat Commun 6:5969

Brand CM, White FJ, Rogers AR, Webster TH (2022) Estimating bonobo (*Pan paniscus*) and chimpanzee (*Pan troglodytes*) evolutionary history from nucleotide site patterns. Proc Natl Acad Sci USA 119(17):e2200858119

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81(5):1084–1097

Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L et al. (2012) Estimating the human mutation rate using autozygosity in a founder population. Nat Genet 44(11):1277–1281

Campbell CR, Tiley GP, Poelstra JW, Hunnicutt KE, Larsen PA, Lee HJ et al. (2021) Pedigree-based and phylogenetic methods support surprising patterns of mutation rate and spectrum in the gray mouse lemur. Heredity 127(2):233–244

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F et al. (2011) Variation in genome-wide mutation rates within and between human families. Nat Genet 43(7):712–714

Fan S, Hansen ME, Lo Y, Tishkoff SA (2016) Going global by adapting local: a review of recent human adaptation. Science 354(6308):54–59

Feng C, Pettersson M, Lamichhaney S, Rubin CJ, Rafati N, Casini M et al. (2017) Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. Elife 6:e23907

Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I et al. (2015) Genome-wide patterns and properties of de novo mutations in humans. Nat Genet 47(7):822–826

Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P (2022) A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. PLoS Comput Biol 18(5):e1009123

Goldmann JM, Wong WS, Pinelli M, Farrah T, Bodian D, Stittrich AB et al. (2016) Parent-of-origin-specific signatures of de novo mutations. Nat Genet 48(8):935–939

Harris RB, Irwin K, Jones MR, Laurent S, Barrett RDH, Nachman MW et al. (2020) The population genetics of crypsis in vertebrates: recent insights from mice, hares, and lizards. Heredity 124(1):1–14

Holtgrewe M (2010) Mason: a read simulator for second generation sequencing data. Dissertation, Freie Universität Berlin.

Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA 101(39):13994–14001

Jennewein DM, Lee J, Kurtz C, Dizon W, Shaeffer I, Chapman A et al. (2023) The Sol Supercomputer at Arizona State University. Practice and experience in advanced research computing, 296–301.

Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X et al. (2013) Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. Am J Hum Genet 93(2):249–263

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E et al. (2017) Parental influence on human germline de novo mutations in 1548 trios from Iceland. Nature 549(7673):519–522

Keightley PD, Ness RW, Halligan DL, Haddrill PR (2014) Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. Genetics 196(1):313–320

Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J et al. (2015) Estimation of the spontaneous mutation rate in *Heliconius melpomene*. Mol Biol Evol 32(1):239–243

Kessler MD, Loesch DP, Perry JA, Heard-Costa NL, Taliun D, Cade BE et al. (2020) De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. Proc Natl Acad Sci USA 117(5):2560–2569

Koch E, Schweizer RM, Schweizer TM, Stahler DR, Smith DW, Wayne RK et al. (2019) De novo mutation rate estimation in wolves of known pedigree. Mol Biol Evol 36(11):2536–2547

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488(7412):471–475

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS et al. (2018) High-resolution comparative analysis of great ape genomes. Science 360(6393):eaar6343

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25(14):1754–1760

Lindsay SJ, Rahbari R, Kaplanis J, Keane T, Hurles ME (2019) Similarities and differences in patterns of germline mutation between mice and humans. Nat Commun 10(1):4053

Liu H, Jia Y, Sun X, Tian D, Hurst LD, Yang S (2017) Direct determination of the mutation rate in the bumblebee reveals evidence for weak recombination-

associated mutation and an approximate rate constancy in insects. Mol Biol Evol 34(1):119–130
Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P et al. (2017) Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. Nature 548(7665):87–91
Martin HC, Batty EM, Hussin J, Westall P, Daish T, Kolomyjec S et al. (2018) Insights into platypus population structure and history from whole-genome sequencing. Mol Biol Evol 35(5):1238–1252
Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X et al. (2012) Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151(7):1431–1442
Milhaven M, Pfeifer SP (2023) Performance comparison of six popular short-read simulators. Heredity 130(2):55–63
Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J (2017) Differences between germline and somatic mutation rates in humans and mice. Nat Commun 8:15183
Pfeifer SP (2021) Studying mutation rate evolution in primates – the effects of computational pipeline and parameter choices. GigaScience 10(10):giab069
Pfeifer SP (2017a) Direct estimate of the spontaneous germ line mutation rate in African green monkeys. Evolution 71(12):2858–2870
Pfeifer SP (2017b) From next-generation resequencing reads to a high quality variant data set. Heredity 118(2):111–124
Pfeifer SP (2020) Spontaneous mutation rates. In Ho SYW (ed) The Molecular Evolutionary Clock. Theory and Practice. Springer Nature, pp. 35–44
Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B et al. (2013) Great ape genetic diversity and population history. Nature 499(7459):471–475
Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Turki SA et al. (2016) Timing, rates and spectra of human germline mutation. Nat Genet 48(2):126–133
Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328(5978):636–639
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G et al. (2011) Integrative genomics viewer. Nat Biotechnol 29(1):24–26
Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB et al. (2019) Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. Elife 8:e46922
Smeds L, Qvarnström A, Ellegren H (2016) Direct estimate of the rate of germline mutation in a bird. Genome Res 26(9):1211–1218
Tatsumoto S, Go Y, Fukuta K, Noguchi H, Hayakawa T, Tomonaga M et al. (2017) Direct estimation of de novo mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing. Sci Rep. 7(1):13561
Thomas GWC, Wang RJ, Puri A, Harris RA, Raveendran M, Hughes DST et al. (2018) Reproductive longevity predicts mutation rates in primates. Curr Biol 28(19):3193–3197
Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC et al. (2017) Genomic patterns of de novo mutation in simplex autism. Cell 171(3):710–722
van der Auwera GA, O'Connor BD (2020) Genomics in the cloud: using Docker, GATK, and WDL in Terra (1st Edition). O'Reilly Media.
Venn O, Turner I, Mathieson I, de Groot N, Bontrop R, McVean G (2014) Strong male bias drives germline mutation in chimpanzees. Science 344(6189):1272–1275
Versoza CJ, Ehmke E, Jensen JD, Pfeifer SP (2024) Characterizing the Rates and Patterns of De Novo Germline Mutations in the Aye-Aye (Daubentonia madagascariensis). Mol Biol Evol 42(3):msaf034. https://doi.org/10.1093/molbev/msaf034
Wang RJ, Thomas GWC, Raveendran M, Harris RA, Doddapaneni H, Muzny DM et al. (2020) Paternal age in rhesus macaques is positively associated with germline mutation accumulation but not with measures of offspring sociability. Genome Res 30(6):826–834
Wang RJ, Peña-Garcia Y, Bibby MG, Raveendran M, Harris RA, Jansen HT et al. (2022) Examining the effects of hibernation on germline mutation rates in grizzly bears. Genome Biol Evol 14(10):evac148
Wang RJ, Raveendran M, Harris RA, Murphy WJ, Lyons LA, Rogers J et al. (2022b) De novo mutations in domestic cat are consistent with an effect of reproductive longevity on both the rate and spectrum of mutations. Mol Biol Evol 39(7):msac127
Wong WS, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC et al. (2016) New observations on maternal age effect on germline de novo mutations. Nat Commun 7:10486
Wu FL, Strand AI, Cox LA, Ober C, Wall JD, Moorjani P et al. (2020) A comparison of humans and baboons suggests germline mutation rates do not track cell divisions. PLoS Biol 18(8):e3000838
Yang C, Zhou Y, Marcus S, Formenti G, Bergeron LA, Song Z et al. (2021) Evolutionary and biomedical insights from a marmoset diploid genome assembly. Nature 594(7862):227–233
Yuen RK, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N et al. (2015) Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med 21(2):185–191

ACKNOWLEDGEMENTS

The authors would like to thank Colin Brand for sharing their previously generated variant catalogue of Western chimpanzees and members of the Pfeifer Lab for their help with the visual curation of IGV screenshots. Computations were performed on the Sol Supercomputer at Arizona State University (Jennewein et al. 2023).

AUTHOR CONTRIBUTIONS

SPP conceived and designed the study. MM, AG, and CJV conducted read simulations and analyzed the data. MM and SPP wrote the manuscript with input from all authors. SPP obtained research funding.

FUNDING

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM151008 to SPP. MM, AG, and CJV were supported by the National Science Foundation CAREER Award DEB-2045343 to SPP. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41437-025-00754-0.

Correspondence and requests for materials should be addressed to Susanne P. Pfeifer.