

<https://doi.org/10.1038/s41698-025-00902-z>

Benchmarking mouse contamination removing protocols in patient-derived xenografts genomic profiling



Mukund Bhandari^{1,5}, Funan He^{1,2,5}, Anna Rogojina¹, Fuyang Li^{1,3}, Yi Zou¹, Jing Jiang^{1,2}, Zhao Lai^{1,3,4}, Peter Houghton^{1,3,4}, Raushan T. Kurmasheva^{1,3,4}, Yidong Chen^{1,2,4}, Xiaojing Wang^{1,2,4} ✉ & Siyuan Zheng^{1,2,4} ✉

Patient-derived xenograft (PDX) models are widely used in cancer research. Genomic and transcriptomic profiling of PDXs are inevitably contaminated by sequencing reads originated from mouse cells. Here, we examine the impact of mouse read contamination on RNA sequencing (RNAseq), Whole Exome Sequencing (WES), and Whole Genome Sequencing (WGS) data of 21 PDXs. We also systematically benchmark the performance of 12 computational protocols for removing mouse reads from PDXs. We find that mouse read contamination increases expression of immune and stromal related genes, and inflates the number of somatic mutations. However, detection of gene fusions and copy number alterations is minimally affected by mouse read contamination. Using gold standard datasets, we find that pseudo-alignment protocols often demonstrate better prediction performance and computing efficiency. The best performing tool is a relatively new tool Xengsort. Our results emphasize the importance of removing mouse reads from PDXs and the need to adopt new tools in PDX genomic studies.

Patient-Derived Xenograft (PDX) models are an important model system for preclinical and molecular cancer research. PDXs are generated by implanting human tumor cells or tumor fragments into host animals, often mice. The engraftment and subsequent passaging provide an avenue for preserving tumor tissue in vivo, an application particularly important for research on rare cancers¹. The use of PDXs to advance precision oncology entails genomic and transcriptomic sequencing to catalogue their genetic and transcriptomic alterations. However, analysis of PDX sequencing data is confounded by sequencing reads that originate from host cells^{2–4}. With small sample sizes, previous studies have shown that mouse reads can significantly increase artificial mutation calls and skew gene expression levels if not filtered^{5–7}. These observations underscore the importance of removing host-derived reads from genomic data when characterizing PDX models.

To address this need, several computational tools have been developed to distinguish between human and mouse reads^{6–11}. These tools generally fall into two categories, alignment-dependent and pseudo-alignment. Alignment-dependent tools use existing aligners to map sequencing reads to human and mouse reference genomes, and then separate reads into source

organisms based on alignment information such as alignment scores, edit distance, mapping quality, etc^{6,7,11}. Pseudo-alignment tools typically use methods such as k-mers to match and separate sequencing reads^{8–10}. Performance of these tools has been benchmarked^{6–8}, but these benchmarking efforts were rarely undertaken by a third party, and for alignment-dependent tools, few benchmarking studies considered the choice of aligners. Moreover, these efforts should be regularly undertaken to keep the community up to date on new developments.

In this study, we assembled whole exome sequencing (WES), low pass whole genome sequencing (WGS), and mRNA sequencing (RNAseq) of 21 PDX models generated from childhood solid tumors¹². The matched normal samples were also available for the WES data. We compiled seven popular tools, and for alignment-dependent tools, we included multiple aligners that were recommended to work with the tool (Table 1). Most of these protocols were tested on both DNA and RNA sequencing data, except when the aligner was designed for only one data type. We first examined the impact of mouse read contamination on genomic data including gene expression, mutations, gene fusions, and copy number alterations. We then benchmarked the protocols for accuracy and computing efficiency. Our

¹Greehey Children's Cancer Research Institute, San Antonio, TX, USA. ²Department of Population Health Sciences, San Antonio, TX, USA. ³Department of Molecular Medicine, San Antonio, TX, USA. ⁴Mays Cancer Center, University of Texas Health at San Antonio, San Antonio, TX, USA. ⁵These authors contributed equally: Mukund Bhandari, Funan He.

✉ e-mail: wangx11@uthscsa.edu; zhengs3@uthscsa.edu

Table 1 | Overview of the tools benchmarked in this study

Protocols	Input	Aligner	Output	Algorithm	Language	Release
Disambiguate	BAM, name sorted	STAR, HISAT, BWA	BAM	Mapping score and quality	Python, C++	2016
Bamcmp	BAM, name sorted	STAR, HISAT, BWA	BAM	Mapping score and quality	C++	2017
Xenofilter	BAM, coordinate sorted	STAR, HISAT, BWA	BAM	Mapping score and quality	R	2018
Xenome	Fastq	NA	Fastq	K-mer	C++	2012
BBSplit	Fastq	BBMAP	Fastq	K-mer	Java	2020
Xengsort	Fastq	NA	Fastq	K-mer	Python, numba	2021

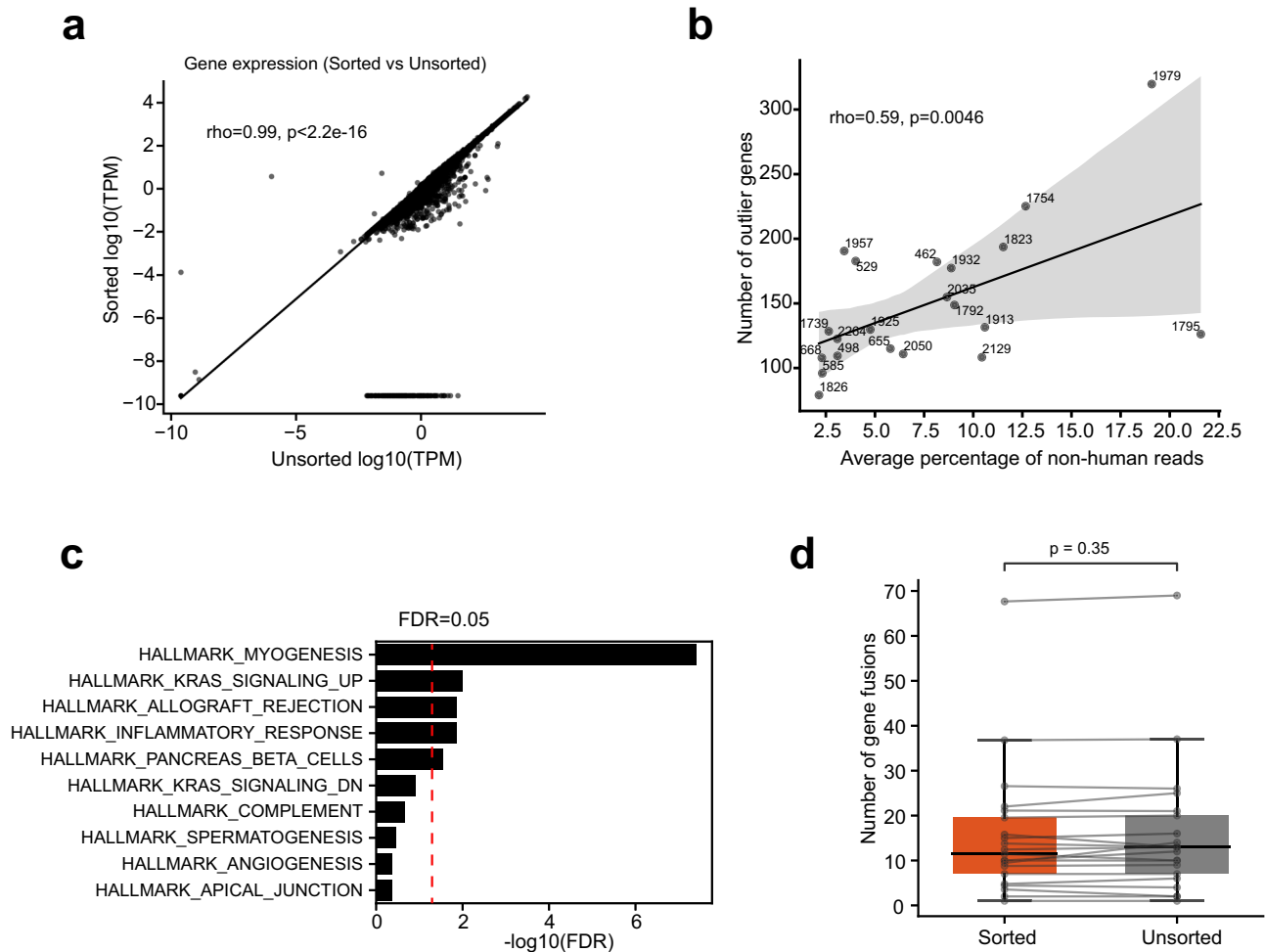


Fig. 1 | Impact of mouse reads on transcriptomic profiles. a An example of expression correlation between sorted and unsorted samples. Each dot is a gene. X axis, log scale expression in unsorted; Y axis, log scale expression in sorted. **b** Correlation between levels of mouse read contamination (x axis) and the number of genes detected as regression outliers (y axis). The x and y axis values are calculated

by averaging outputs from the protocols tested in the study. **c** Enrichment of outlier genes in hallmark gene sets. X axis, FDR in $-\log_{10}$ scale. Dashed line indicates $\text{FDR}=0.05$. **d** No statistical difference is found between the number of fusions between sorted and unsorted samples. Each dot represents one sample. P value is calculated using paired t-test.

results show that a newer tool Xengsort outperformed other established tools in prediction accuracy and computing efficiency.

Results

Impact of mouse read contamination on PDX transcriptomic data

To examine the impact of mouse read contamination on gene expression, we first compared gene expression values with and without removing mouse reads. For simplicity, we call a sample sorted if the mouse reads were removed, and unsorted if otherwise. Most genes showed highly similar expression levels between sorted and unsorted samples across the tools (Fig. 1a, Supplementary Fig. 1a; spearman correlation ρ range, 0.980–0.999; median, 0.997). However, some genes showed a higher expression in

unsorted samples, suggesting their expression was inflated by mouse reads. Next, we identified these genes for each sample as regression outliers (Methods). Different numbers of outlier genes were detected based on different protocols (Supplementary Fig. 1b), suggesting varying performance of the protocols. On average, the higher mouse read contamination was, the more outlier genes were detected (Fig. 1b; spearman correlation $\rho=0.59$, $p=0.0046$). These observations indicate that mouse reads inflated the expression of these genes.

To better understand the genes whose expression is susceptible to mouse read contamination, we identified 821 genes that were determined as outliers in at least two samples and tested functional enrichment of these genes (Methods). The outlier genes were significantly enriched in five gene

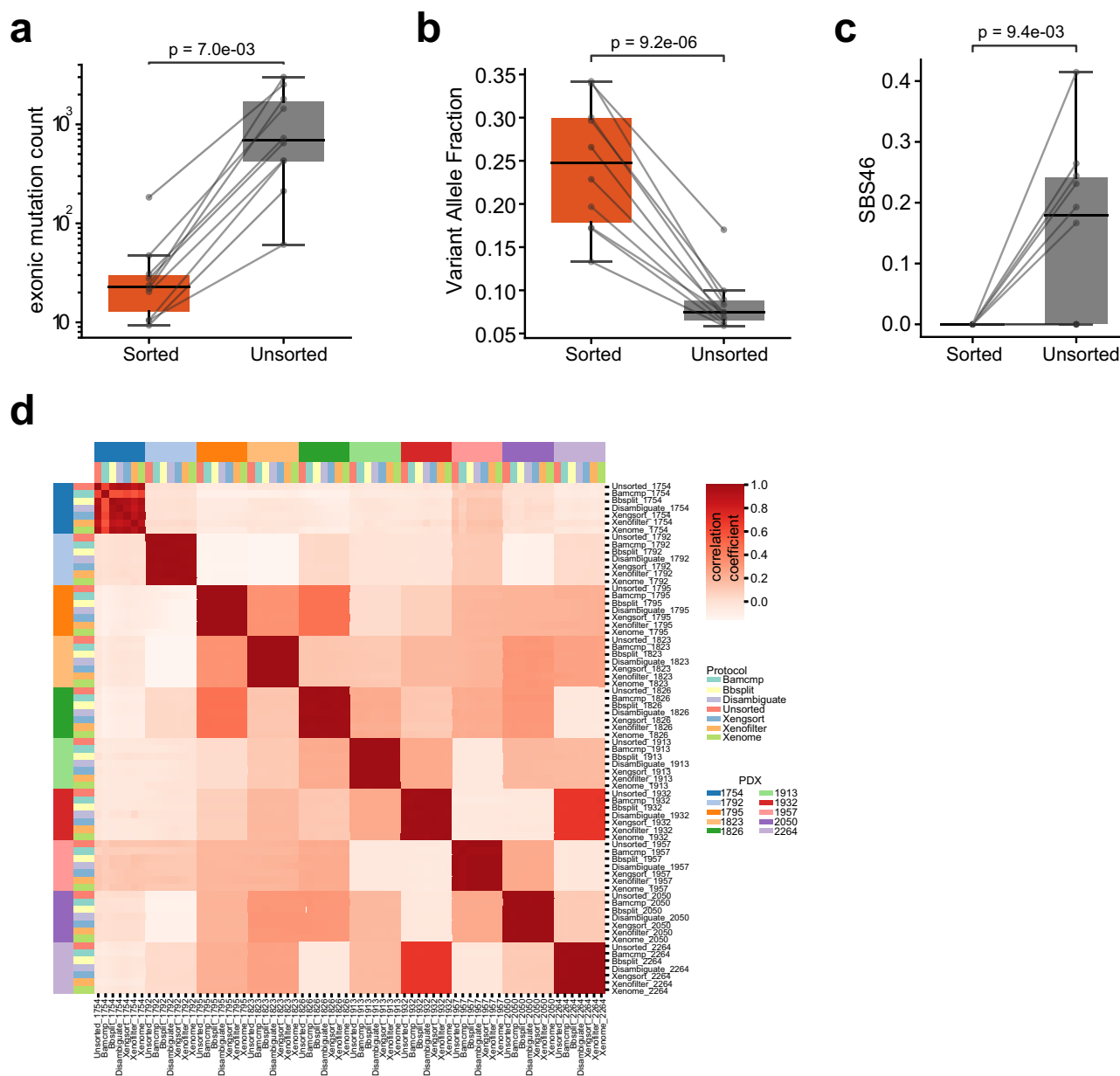


Fig. 2 | Impact of mouse reads on genomic profiles. a The number of somatic exonic mutations is significant higher in unsorted samples. $P = 7.0 \times 10^{-3}$, paired t -test. **b** Mutations called in unsorted samples have lower variant allele fraction. Each dot represents the average of mutations from one sample. $P = 9.2 \times 10^{-6}$, paired t -test. **c** SBS46 exposure is higher based on mutations detected in unsorted samples. Each

dot represents the average of mutations from one sample. $P = 9.4 \times 10^{-3}$, paired t -test. **d** Copy number profiles are highly similar between sorted and unsorted samples. Each block represents correlations of copy number profiles derived from applying different protocols to one sample. Unsorted sample is positioned first in each block.

sets (Fig. 1c; FDR < 0.05), which can be roughly divided into stroma related (myogenesis, pancreas beta cells), immune cell related (allograft rejection, inflammatory response), and interestingly, KRAS signaling (KRAS signaling up). Enrichment of immune and stromal related genes amongst outlier genes is consistent with the notion that mouse immune cells and stromal cells are the main source of mouse read contamination in PDXs.

We next tested the consistency of the pathway enrichment for the outlier genes across the samples. We used the hallmark gene signatures from MSigDB¹³. For each gene signature, we compared signature score between sorted and unsorted samples. The same gene signatures including myogenesis, allograft rejection, KRAS signaling were among the top differential signatures. Importantly, these signatures exhibited almost ubiquitous higher scores in the unsorted samples (Supplementary Fig. 1c).

On the other hand, genes that were most resistant to mouse read contamination were enriched in gene sets that typically identify with cancer

cells including cell division, TP53 pathway, DNA repair, and cell metabolism (Supplementary Fig. 1d).

Next, we evaluated the effect of mouse read contamination on gene fusion identification. We observed slightly fewer gene fusions after removing mouse contamination, but the difference was not statistically significant (Fig. 1d; $p = 0.35$, paired t -test). The choice of mouse read removal protocol had little effects (Supplementary Fig. 1e). Thus, these results suggest mouse read contamination has limited impact on fusion identification.

Finally, we examined mouse read contamination in PDX single cell RNA sequencing (scRNAseq) data. Unlike bulk samples, sequencing reads in scRNAseq data are attributable to individual cells, which are derived from either human or mouse. This dichotomy suggests that read classification can be used to identify human or mouse cells. We applied Xengsort and Disambiguate to the scRNAseq data of a PDX sample. Indeed, the fraction of

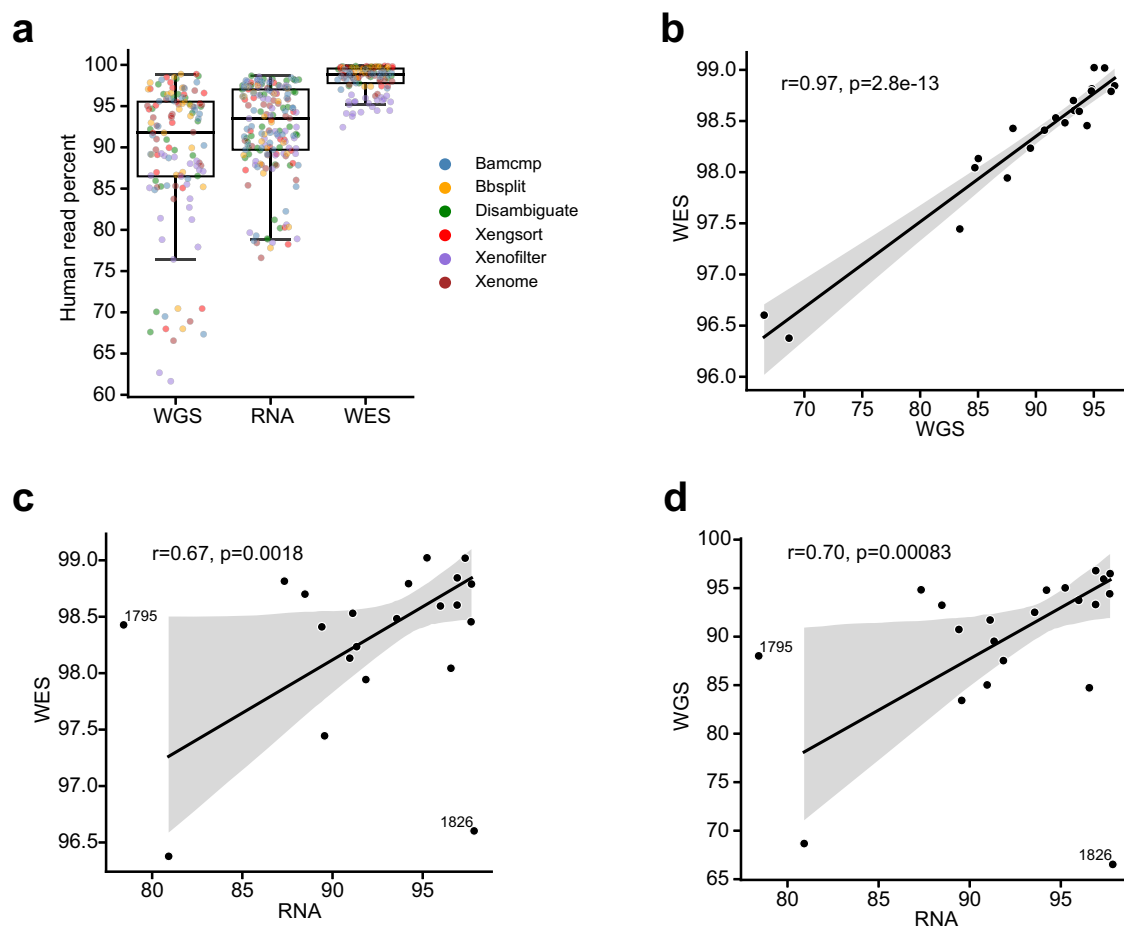


Fig. 3 | Mouse read contamination across data modalities. **a** Percentage of human reads in WGS, WES, and RNAseq data. Each dot represents one sample. **b** Pearson correlation between human read percentages in WGS (x axis) and WES data (y axis). **c** Pearson correlation between human read percentages in RNAseq (x axis) and WES

data (y axis). **d** Pearson correlation between human read percentages in RNAseq (x axis) and WGS data (y axis). Patient 1795 and 1826 were not included for calculating correlation coefficient.

mouse (or human) reads in individual cells followed a clear bimodal distribution (Supplementary Fig. 2a). In our sample, only 0.6% of the cells showed a difference in mouse/human read fraction less than 10%. Disambiguate and Xengsort reported highly consistent results regarding mouse read fraction (Supplementary Fig. 2b). These results suggest the tools can be effectively used on scRNAseq data despite their initial intent on bulk samples.

Impact of mouse read contamination on PDX genomic data

We examined the impact of mouse read contamination on somatic mutation calling using WES data. Without mouse reads removal, the number of exonic mutations detected in PDXs was on average about 30 times higher than those after mouse reads removal (mean: 1,127 vs 38; $p = 0.007$, paired *t*-test; Fig. 2a, Supplementary Fig. 3a). Thus, the presence of mouse reads can introduce abundant mutation artifacts in the PDX.

To provide insights into the mutation artifacts caused by mouse read contamination, we identified them as those uniquely detected in unsorted samples. As a reference, mutations detected in sorted samples were considered bona fide somatic mutations. Consistent with the idea that the artifacts were caused by relatively few mouse reads, variant allele fraction of the mutation artifacts was on average below 0.1, much lower than that of bona fide mutations ($p = 9.2e-6$, paired *t*-test; Fig. 2b, Supplementary Fig. 3b) while the sequencing coverages at the mutation loci were largely similar (Supplementary Fig. 3c). Pairwise comparisons of mutation artifacts from genetically unrelated PDXs showed an average overlap of 12.6% (Jaccard distance, Supplementary Fig. 3d and 3e). This was much higher than the overlap of bona fide mutations

(0.63%, $p = 2.83e-11$, paired *t*-test; Supplementary Fig. 3e), suggesting a substantial proportion of the artifacts can be recurrently detected in unrelated PDXs. Such recurrence may be leveraged for filtering purposes.

We then did a mutational signature analysis using mutations found in sorted and unsorted samples. Of all the signatures we detected, only SBS46, a signature that was thought to originate from sequencing artifacts¹⁴, showed a significant difference in signature exposure (0 in sorted vs. 0.18 in unsorted, $p = 0.009$; Fig. 2c).

We observed similar copy number patterns in sorted and unsorted samples. Except for one sample, copy number profiles between sorted and unsorted samples were nearly identical (Fig. 2d, Supplementary Fig. 4) suggesting a minor effect by mouse read contamination on copy number profiles.

Mouse read contamination affects sequencing data differently

We next sought to understand susceptibility of the three sequencing data modalities to mouse read contamination. We applied the mouse read removal protocols to WGS, WES, and RNAseq data. Averaging all the tested protocols across the samples, WES data showed higher proportions of human reads than WGS and RNAseq data (median, WES, 98%; WGS, 93%; RNAseq, 94%; Fig. 3a), likely because human exon-based enrichment during WES library preparation excluded some mouse sequences. More reads were classified as ambiguous in RNAseq than in WES and WGS (Supplementary Fig. 5). The proportions of human reads reported in the three data types were significantly correlated (Fig. 3b–d), especially between WES and WGS ($R = 0.97$, $p = 2.8e-13$; Fig. 3b), suggesting mouse contamination is dictated at the sampling rather than the sequencing level.

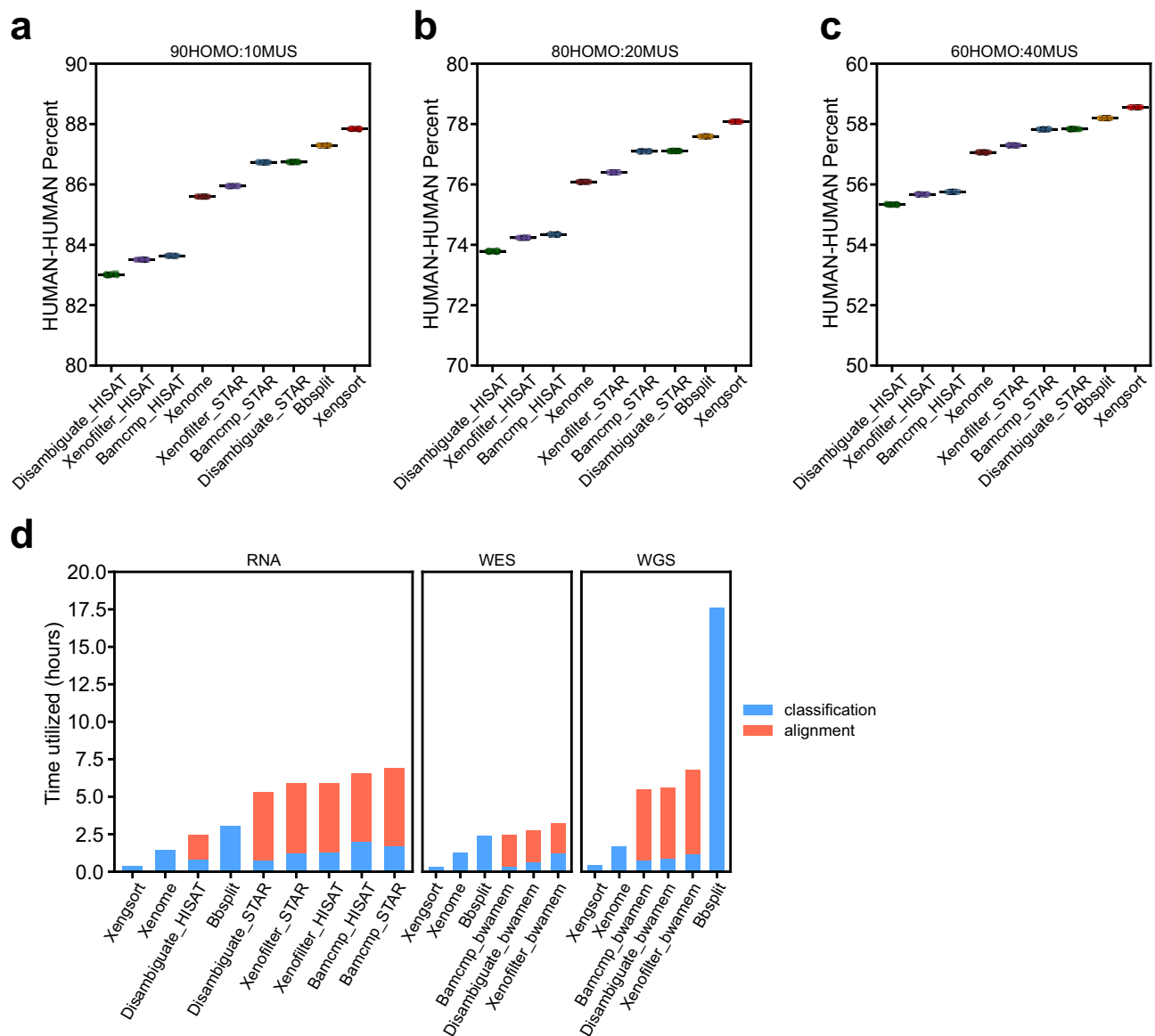


Fig. 4 | Benchmarking accuracy and computing efficiency. **a** Percentage of correctly classified human reads in ground truth dataset comprising 90% human reads and 10% mouse reads. Y axis, fraction of correctly classified human reads over all the

reads. **b** Same to **a** but with 80% human reads and 20% mouse reads. **c** Same to **a** but with 60% human reads and 40% mouse reads. **d** Computing speed of the tested protocols. Y axis, time consumed with 4 CPUs.

Benchmarking the performance of mouse contamination removal protocols

To benchmark the performance of the mouse contamination removal protocols, we synthesized 'gold standard' datasets by mixing human and mouse RNA sequencing reads at varying proportions (9:1, 8:2, and 6:4). To mimic PDX sequencing, we generated mouse reads by sequencing mouse blood samples with human sequencing kits. Since the organism of origin was known for each read, we could test the protocols for their ability to correctly classify the reads. Overall, all the protocols we tested recovered human reads with high accuracy across the three synthetic datasets (Fig. 4a–c). The pseudo-alignment tool Xengsort showed the best performance, followed closely by BBSplit. The choice of aligners also had an impact on performance. For instance, Disambiguate and Bamcmp showed better prediction accuracies when coupled with STAR than HISAT.

We next analyzed misclassified reads. Mouse reads misclassified as human origin were the remaining contamination in the data. In the 9:1 mix dataset, the maximum 0.1% of the total reads, reported by BBSplit, belonged to this category (Supplementary Fig. 6a). Even in the 6:4 mix dataset where more mouse reads were present, only the maximum 0.4% of the total reads were of mouse origin.

Human reads that were misclassified as of mouse origin represented data loss since they would be removed as mouse reads. In all three mix datasets, less than 1.7% of the total reads belonged to this category (Supplementary Fig. 6b).

To understand if the choice of the mouse reference genome would have any impact on read classification, we ran Xengsort and Disambiguate on three PDXs using genomes of four mouse strains (A/J, BALB/C, C57BL/6NJ, NOD/ShiLtJ) and GRCm38/mm10 (Supplementary Table 1). We found that both tools generated highly similar classification results based on the five reference genomes. The difference of reads classified as of human or mouse origin was no more than 0.5% using different reference genomes by both tools. These results suggest the choice of the mouse reference genome does not have a significant impact on read classification.

Finally, we compared computing time for the mouse contamination removal protocols across the three data types. Here, the running time also included the alignment step if an aligner was integral to the protocol. When tested with four 3.4 GHz CPUs, the alignment-dependent protocols generally took a longer running time than pseudo-alignment protocols, mostly because the alignment step took longer (Fig. 4d, Supplementary Fig. 7). The

pseudo-alignment tool was the fastest followed by Xenome. However, even Xenome took more than two times longer than Xengsort. Thus, Xengsort is a highly efficient and accurate tool.

Discussion

In this study we systematically evaluated the impact of mouse read contamination on genomic and transcriptomic profiling of PDXs. We showed that expression of stromal and immune related genes could be artificially increased if mouse reads were not removed. Mouse read contamination also substantially inflated the number of exonic mutations, though WES was less susceptible to mouse contamination than WGS and RNAseq. We observed a limited impact of mouse read contamination on detection of gene fusions and copy number alterations. Altogether, these results suggest removing mouse reads from PDX sequencing data is necessary.

We benchmarked common mouse reads removal protocols using gold standard datasets. We found all protocols were effective at removing mouse reads. The performance difference was minor, usually within 5% of the human reads. A pseudo-alignment tool Xengsort outperformed other tools in both prediction accuracy and computing efficiency. The improvement of computing efficiency is particularly noticeable, as Xengsort uses only a fraction of time compared with other tools. Because adopting more efficient tools can save a significant amount of computational resource, we recommend the PDX research community to use Xengsort for removing mouse reads from genomic and transcriptomic sequencing data.

Many studies have shown that tumor cells interact with immune and stromal cells. A recent study showed tumor purity, due to immune and stromal cell infiltration, is largely retained in PDXs¹⁵. This observation suggests that PDXs can be a useful model to study tumor-immune/stromal cell interactions. Removing signals of immune and stromal cells from PDX sequencing data obviously ablates the opportunity to study these interactions. However, should one use mouse reads to study the interactions, technical confounders need to be carefully gauged as standard library preparation kits for PDX genomic profiling including bulk sample, single cell, and spatial omics sequencing, are designed for human sequences and thus a drop in efficiency is expected for capturing mouse sequences. Tissue dual RNA sequencing can be an alternative¹⁶, but proof-of-principle studies in PDXs are currently lacking. Provided with proper tools, any interactions observed between human tumor cells and mouse immune and stromal cells must be validated in the human tumor microenvironment.

In summary, we analyzed the impact of mouse read contamination on PDX genomic profiling and benchmarked the performance of popular mouse read removal tools. We showed that Xengsort outperformed other tools in computing efficiency and classification accuracy.

Methods

Sample collection

Sequencing data used in this study were generated at Genome Sequencing Facility of Greehey Children's Cancer Research Institute. Details of pediatric tumor collection, patient-derived xenografts (PDX) generation, DNA and RNA sequencing library preparation were done by the collaborative effort of many labs and were described in a previous study¹². Here, we only included 21 PDXs with matched germline tissue, so we can better evaluate the performance of downstream analyses after mouse reads removal.

Generation of benchmarking data

Benchmarking data were generated by mixing mouse and human RNA-sequencing data. Buffy coat of NSG mice (host) and patient normal tissue (1823_PT) was used for mouse and human RNA-seq library preparation respectively following the protocols described above. Sequencing was done with NovaSeq platform with 100PE (paired end) run, generating 64,877,973 mouse and 51,057,950 human reads pool. Trim Galore¹⁷ (v0.6.10) was firstly used to remove the adapters and poor-quality reads from FASTQ files. These human reads and mouse reads were further tagged with "HUMAN" or "MOUSE" in the identifier line of FASTQ files. Next, we generated the human-mouse RNA-seq benchmarking data with a total of 10 million reads

by randomly selecting the human and mouse reads from the pool, and mixing them at different proportions (9:1, 8:2, and 6:4). 50 iterations of random sampling were taken for each proportion to create a total of 150 different benchmarking FASTQ files.

Mouse reads removal

We included three pseudo-alignment and three alignment-based protocols to evaluate their performance of mouse reads removal. Those pseudo-alignment protocols Xenome¹⁰, Bbsplit⁹ (v38.84), and Xengsort⁸ (v2.0.5) were directly implemented to remove mouse reads. Note that BBSplit uses BBMap as its aligner, but since BBMap uses k-mers for matching, we still consider BBSplit a pseudo-alignment tool. For those alignment-based protocols, FASTQ files were separately aligned to the mouse reference genome (GRCm38, GENCODE vM19) and human reference genome (GRCh38, GENCODE v29)¹⁸. BWA-MEM¹⁹ (v2.2.1) was used to align DNA sequencing data, HISAT2²⁰ (v2.1.0) and STAR²¹ (v2.7.10b) were used to align RNA sequencing data to the reference genome. Disambiguate¹¹ (v1.0), Xenofilter⁷, and Bamcmp⁶ were then used on those aligned BAM files to remove mouse reads contamination from PDXs. All these tools were executed by default parameters, and the codes were available: https://github.com/mukund-bhandari/PDX_Sorting_Protocols. Pseudo-Alignment and alignment-based tools were also applied to RNA-seq benchmarking dataset. To evaluate the performance of different mouse read removal tools, we further compared the human and mouse reads IDs of those output BAM or FASTQ files to those of the original human-mouse mixture. SAMtools²² (v1.13) was used to extract sequence read IDs for those BAM files, and a custom script was used to extract sequence read IDs for those FASTQ files.

RNAseq data analysis

After mouse reads removal, Kallisto²³ (v0.46.2) was applied to the human FASTQ files to calculate transcript per million (TPM). For alignment-based protocols, SAMtools was firstly applied to convert human BAM output to FASTQ format, which was further applied to Kallisto. Meanwhile, Kallisto was also applied to raw PDX FASTQ files before mouse reads removal for evaluating the performance of different mouse reads removal tools.

We compared the gene expression matrix between those with and without mouse contamination removal for each sample. Linear regression was applied, standard deviation (SD) and the residual value of each gene were calculated to evaluate the consistency of different gene expression matrices. Genes with residual values larger than 2*SD were regarded as outlier genes. Correlation between sorted and unsorted gene expression was calculated by spearman correlation. When correlating the number of outlier genes and the proportion of non-human reads, we used the average from all protocols across the samples for both variables.

Gene set enrichment analysis

Gene outliers were identified as common genes that were found in at least two different unsorted patients. Gene enrichment was applied to these genes to identify the pathways that were highly affected by mouse reads contamination. Besides, gene enrichment analysis was also applied to genes that were not identified as outlier genes among all the unsorted samples. When evaluating the consistency of pathway enrichment pattern for outlier genes, we scored Hallmark gene signatures from MSigDB¹³ using ssGSEA²⁴. Signature scores were compared between sorted and unsorted samples using Wilcoxon rank sum test.

Gene fusion identification

STAR-Fusion²⁵ (v1.13.0) was used to detect gene fusion, and in silico validation using FusionInspector²⁶ was also performed. Considering the fusion junction and spanning reads might be classified as unmapped and human-mouse ambiguous reads, we combined the unmapped reads, human reads and ambiguous reads after removing the mouse reads. For alignment-based protocols, we firstly combined those BAM files, that then converted the combined file into the FASTQ format using SAMtools for further analyses.

DNA data preprocessing

After mouse reads removal, GATK²⁷ (v4.4.0.0) best practice workflow was applied to remove duplicates and recalibrate for all PDX, paired PT and germline BAM files. For pseudo-alignment protocols, FASTQ files of human reads were re-mapped to human reference genome using BWA-MEM. For alignment-based protocols, BAM files of human reads were used for further analysis.

Somatic mutations

MuTect2 (GATK v4.2.3.0) was used to identify somatic single nucleotide polymorphisms (SNPs), multiple nucleotide polymorphisms (MNPs), and indels from the whole exome sequencing (WES) data. The Bcftools²⁸ (v1.15.1) tool was used to filter variants and indels and only those that passed quality filters were selected. To remove potential germline variants, we annotated the somatic mutation candidates using ANNOVAR²⁹, and removed those variants in gnomAD 3.0. The remaining mutations were further used for comparing the performance of different mouse reads removal protocols.

Mutation signature identification

Based on exonic mutations, SigProfilerMatrixGenerator was used for generating mutational signatures of each patient. For protocols with more than 20 exonic mutations, SigProfilerAssignment was applied to mutation profile to extract the mutational signature. Next, we identified the proportion of signature SBS46 that significantly increased in unsorted samples.

Somatic copy number alterations

CNVkit³⁰ (v0.9.10) was applied to detect the somatic copy number. We firstly generated the reference CNVkit file based on germline WES BAM files. Then, CNVkit batch and call were used to identify the copy number variations. To better compare the copy number differences between different protocols, we converted the CNVkit segmentation files into 1 Mb bin copy number ratio matrix, and compared the copy number profile between sorted and unsorted groups.

Time utilization

For time calculation, all mouse contamination removal protocols were run with default parameters on Microsoft Azure platform. Four different PDX samples were selected and executed with 4 3.4 GHz CPUs on a computing node with a memory of 342 Gb.

Data availability

The PDX sequencing data is available at EGA European Genome-Phenome Archive (EGAS00001006710). The mouse sequencing data generated with human library preparation kit is made available on Zenodo (doi: 10.5281/zenodo.14775572).

Received: 6 December 2024; Accepted: 5 April 2025;

Published online: 17 April 2025

References

- Yang, J. et al. PCAT: an integrated portal for genomic and preclinical testing data of pediatric cancer patient-derived xenograft models. *Nucleic Acids Res.* **49**, D1321–D1327 (2021).
- Schneeberger, V. E., Allaj, V., Gardner, E. E., Poirier, J. T. & Rudin, C. M. Quantitation of Murine Stroma and Selective Purification of the Human Tumor Component of Patient-Derived Xenografts for Genomic Analysis. *PLOS ONE* **11**, e0160587 (2016).
- Lin, M.-T. et al. Quantifying the relative amount of mouse and human DNA in cancer xenografts using species-specific variation in gene length. *BioTechniques* **48**, 351–355 (2010).
- Rogojina, A. et al. Comprehensive characterization of patient-derived xenograft models of pediatric leukemia. *iScience* **26**, 108171 (2023).
- Jo, S.-Y., Kim, E. & Kim, S. Impact of mouse contamination in genomic profiling of patient-derived models and best practice for robust analysis. *Genome Biol.* **20**, 231 (2019).
- Khandelwal, G. et al. Next-Generation Sequencing Analysis and Algorithms for PDX and CDX Models. *Mol. Cancer Res.* **15**, 1012–1016 (2017).
- Kluin, R. J. C. et al. Xenofilter: computational deconvolution of mouse and human reads in tumor xenograft sequence data. *BMC Bioinforma.* **19**, 366 (2018).
- Zentgraf, J. & Rahmann, S. Fast lightweight accurate xenograft sorting. *Algorithms Mol. Biol.* **16**, 2 (2021).
- BBMap. SourceForge <https://sourceforge.net/projects/bbmap/> (2024).
- Conway, T. et al. Xenome—a tool for classifying reads from xenograft samples. *Bioinformatics* **28**, i172–i178 (2012).
- Ahdsmäki, M. J., Gray, S. R., Johnson, J. H. & Lai, Z. Disambiguate: An open-source application for disambiguating two species in next generation sequencing data from grafted samples. *F1000Research*, **5**, 2741 (2017).
- He, F. et al. Genomic profiling of subcutaneous patient-derived xenografts reveals immune constraints on tumor evolution in childhood solid cancer. *Nat. Commun.* **14**, 7600 (2023).
- Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
- Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
- Qian, W. et al. Tumor Purity in Preclinical Mouse Tumor Models. *Cancer Res. Commun.* **2**, 353–365 (2022).
- Nuss, A. M. et al. Tissue dual RNA-seq allows fast discovery of infection-specific functions and riboregulators shaping host-pathogen transcriptomes. *Proc. Natl. Acad. Sci. USA* **114**, E791–E800 (2017).
- Krueger, F. et al. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path. Zenodo <https://doi.org/10.5281/zenodo.7598955> (2023).
- Harrow, J. et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760–1774 (2012).
- Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://doi.org/10.48550/arXiv.1303.3997> (2013).
- Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
- Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
- Haas, B. J. et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 213 (2019).
- Haas, B. J. et al. Targeted in silico characterization of fusion transcripts in tumor and normal tissues via FusionInspector. *Cell Rep. Methods* **3**, 100467 (2023).
- DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
- Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
- Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
- Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Comput. Biol.* **12**, e1004873 (2016).

Acknowledgements

We thank all the patients who consented to the use of their samples in research. This work is supported by CPRIT (RP220599, RP160716). F.L. is supported by R01CA258381 (to P.J.H.). We also acknowledge the support from GCCRI. Z.L. is supported by NIH NCI R50CA265339. The Sequencing data used in the study were partly generated at The Greehey Children's Cancer Research Institute (GCCRI) Genome Sequencing Facility (GSF). GSF is supported by UT Health San Antonio, NIH-NCI P30 CA054174 (Cancer Center at UT Health San Antonio) and NIH Shared Instrument grant S10OD030311 (S10 grant to NovaSeq 6000 System), and CPRIT Core Facility Award (RP220662). Z.L. is supported by NCI R50 CA265339.

Author contributions

Conceptualization: S.Z., C.Y., W.X. Supervision: S.Z., X.W. Experimental Design: S.Z., M.B., F.H., W.X. Data Analysis: S.Z., M.B., F.H., W.X., J.J. Sequencing Data Generation: Z.L., Y.C., Y.Z., A.R., F.L. PDX models: P.J.H., R.K.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41698-025-00902-z>.

Correspondence and requests for materials should be addressed to Xiaojing Wang or Siyuan Zheng.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025