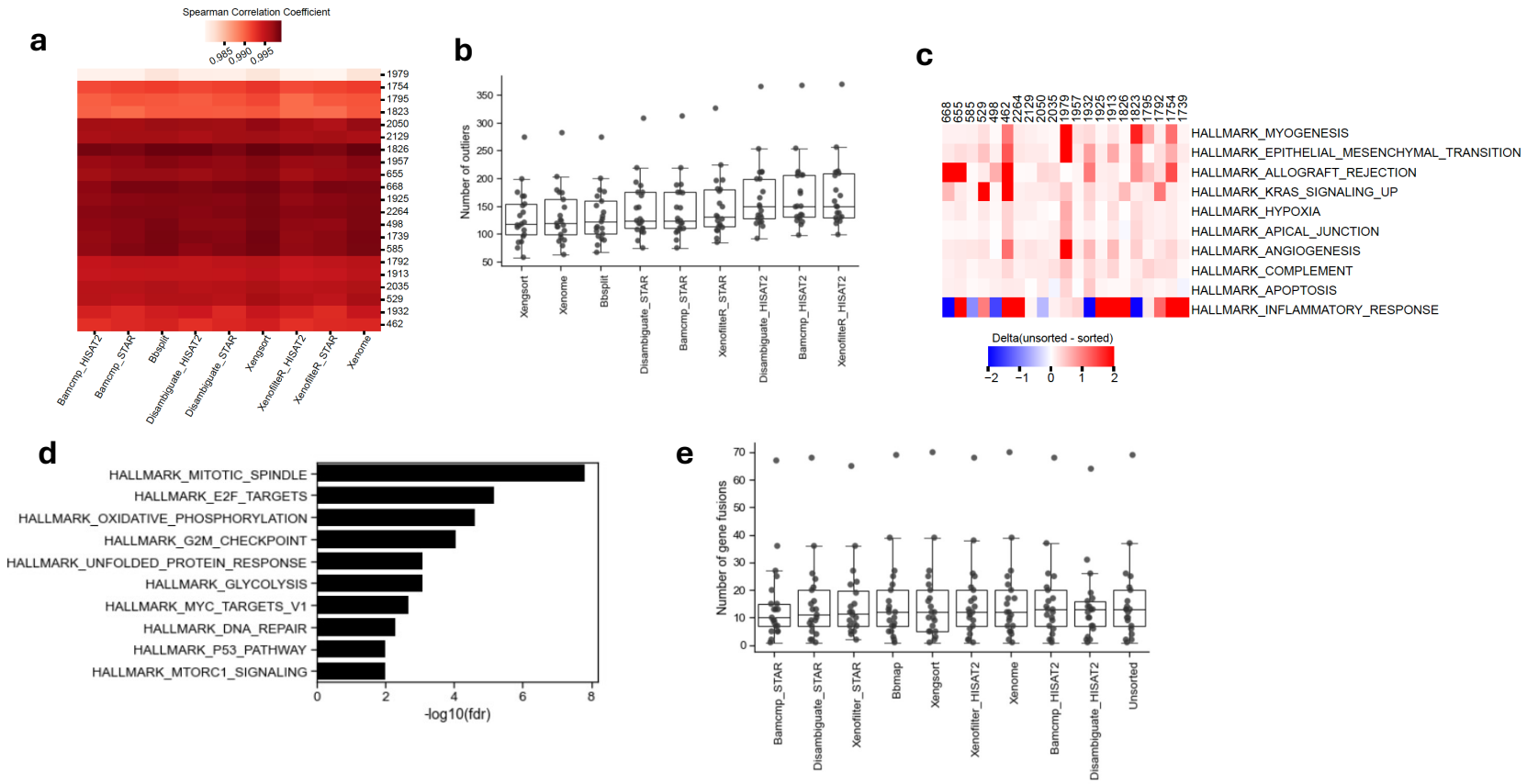
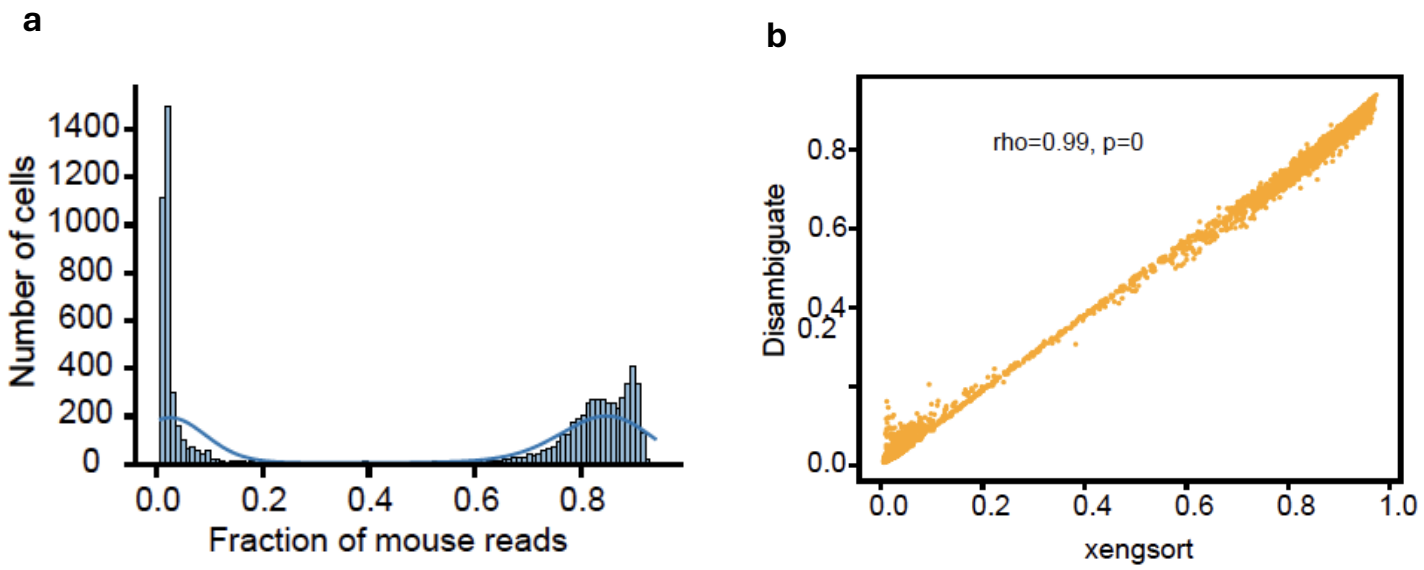
	Xengsort			*Disambiguate-STAR*		
1823	**human**	**mouse**	**ambiguous**	**human**	**mouse**	**ambiguous**
1823_aj	88.34%	9.77%	0.19%	84.91%	9.41%	5.68%
1823_balbcj	88.33%	9.81%	0.19%	84.98%	9.44%	5.58%
1823_c57bl6nj	88.46%	9.79%	0.20%	84.96%	9.38%	5.66%
1823_nodshiltj	88.19%	9.68%	0.22%	84.97%	9.23%	5.80%
1823_mm10	88.13%	10.11%	0.20%	84.53%	9.83%	5.65%
1913	**human**	**mouse**	**ambiguous**	**human**	**mouse**	**ambiguous**
1913_aj	89.25%	9.35%	0.12%	85.47%	8.99%	5.54%
1913_balbcj	89.27%	9.37%	0.12%	85.53%	8.95%	5.52%
1913_c57bl6nj	89.32%	9.35%	0.13%	85.50%	8.95%	5.55%
1913_nodshiltj	89.18%	9.32%	0.14%	85.52%	8.92%	5.56%
1913_mm10	89.07%	9.49%	0.12%	85.19%	9.18%	5.63%
1957	**human**	**mouse**	**ambiguous**	**human**	**mouse**	**ambiguous**
1957_aj	96.76%	2.11%	0.09%	93.09%	2.11%	4.80%
1957_balbcj	96.76%	2.13%	0.09%	93.14%	2.10%	4.76%
1957_c57bl6nj	96.88%	2.09%	0.10%	93.11%	2.05%	4.83%
1957_nodshiltj	96.65%	2.11%	0.09%	93.13%	2.08%	4.79%
1957_mm10	96.65%	2.16%	0.10%	92.83%	2.18%	4.98%

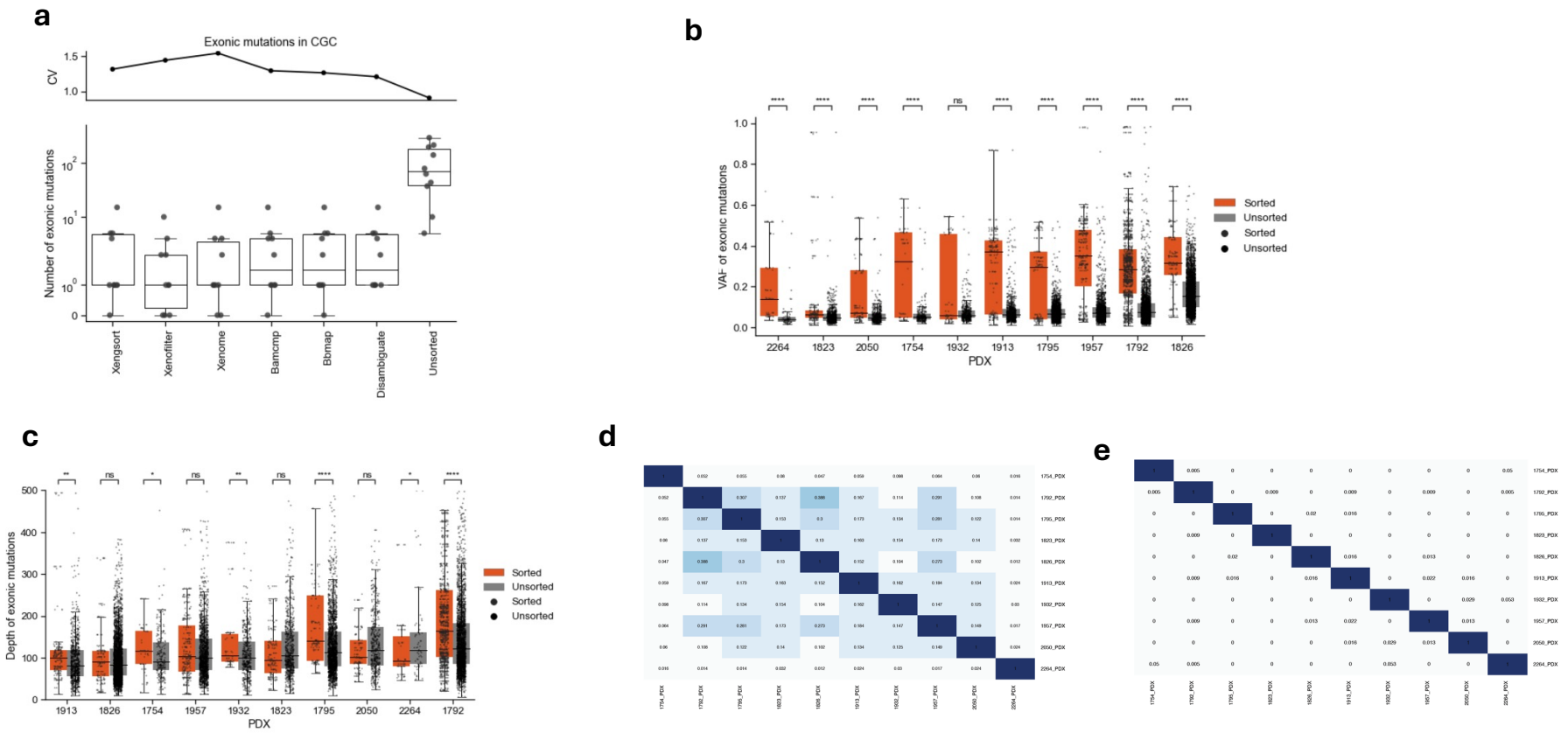
Supplementary Table 1. Comparison of classification results on three PDXs using different mouse reference genomes. 1823, 1913, 1957 are PDX numbers. Numbers in the table indicate the percentage of reads classified in each category (human, mouse, ambiguous).



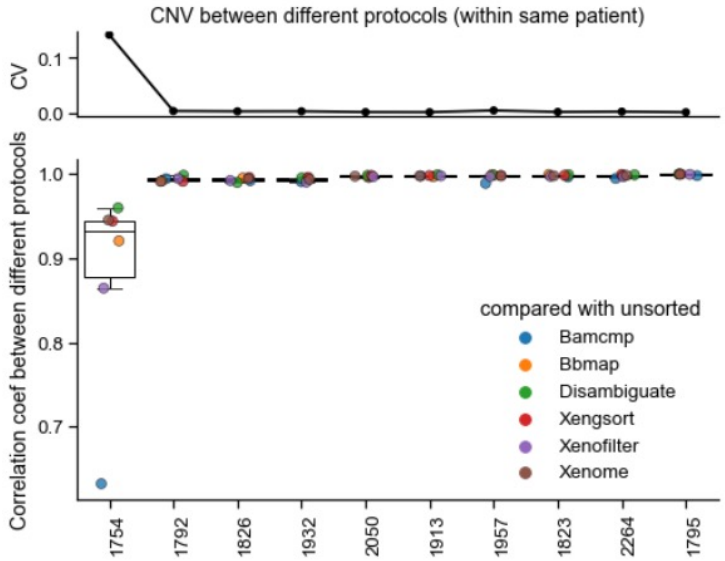
Supplementary Figure 1. Impact of mouse reads on transcriptomic profiles. a Expression correlation between unsorted and sorted samples across algorithms. Values of the heatmap show Spearman correlation coefficient. **b** The number of genes identified as regression outliers by comparing expression of sorted and unsorted samples. Each dot is a sample. **c** Difference in pathway scores between sorted and unsorted samples across the 21 PDXs. Red, pathway score higher in sorted samples. Blue, pathway score lower in sorted samples. Signature scores were calculated using ssGSEA. The top 10 signatures shown here are ordered by p values from Wilcox rank sum test. **d** Gene set enrichment of genes that are most resistant to mouse read contamination. **e** The number of gene fusions detected by different protocols. Each dot represents one sample.



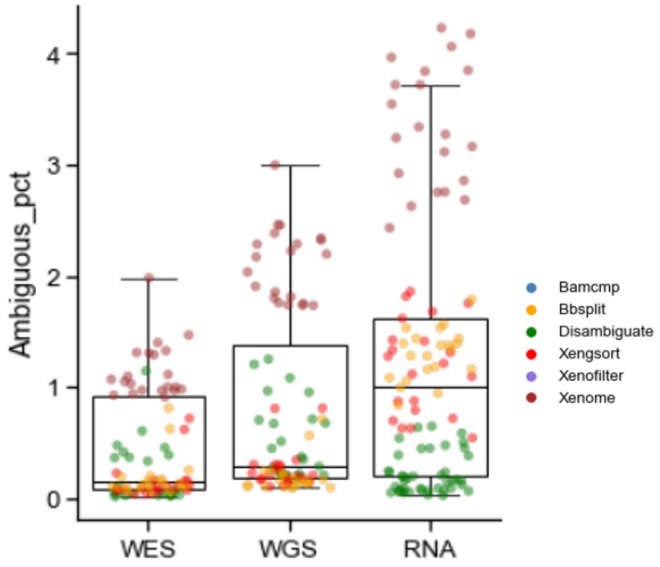
Supplementary Figure 2. Fraction of mouse reads in scRNAseq data of a PDX model. a The bimodal distribution indicates the origin of the cells. **b** Xengsort and disambiguate report highly consistent results. X axis, mouse read fraction reported by Xengsort. Y axis, mouse read fraction reported by Disambiguate. Each dot represents a cell. P value, Spearman correlation.



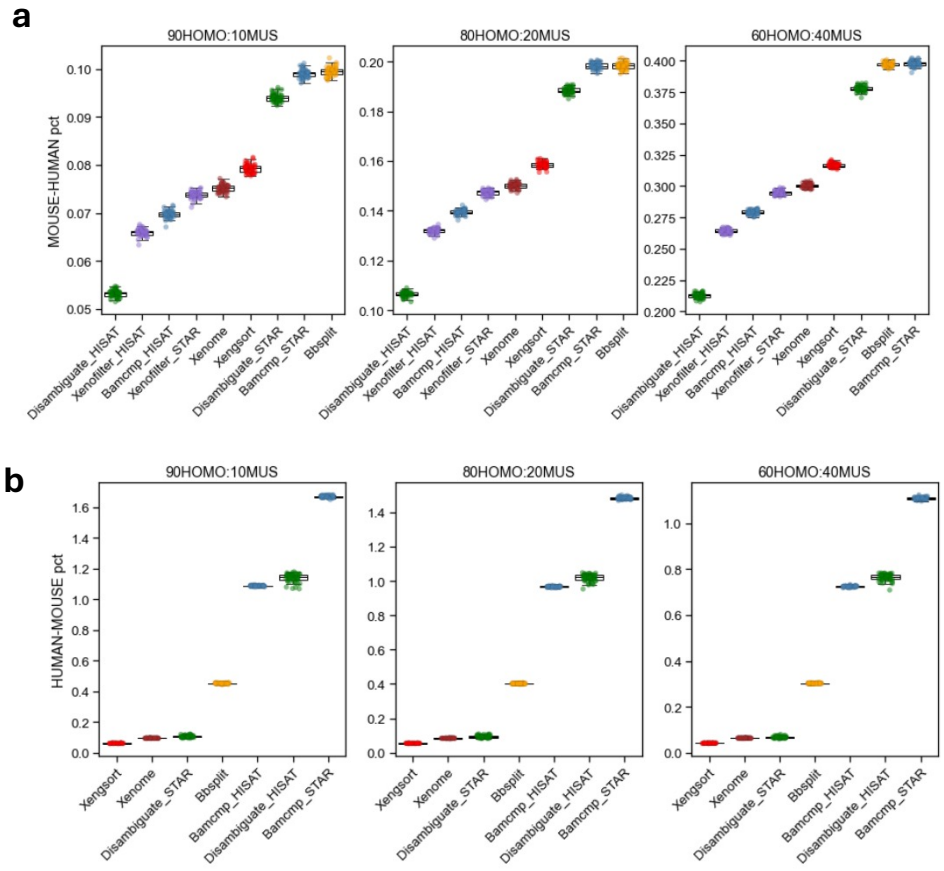
Supplementary Figure 3. Impact of mouse reads on genomic profiles. a The number of somatic exonic mutations detected in samples sorted by different protocols. **b** Variant allele fraction (VAF) of mutations detected in sorted and unsorted samples. Each dot represents a mutation. **c** Sequencing depth of mutations detected in sorted and unsorted samples. **d** Overlap of mutations (Jaccard distance) detected in unsorted samples. **e** Overlap of mutations (Jaccard distance) detected in sorted samples.



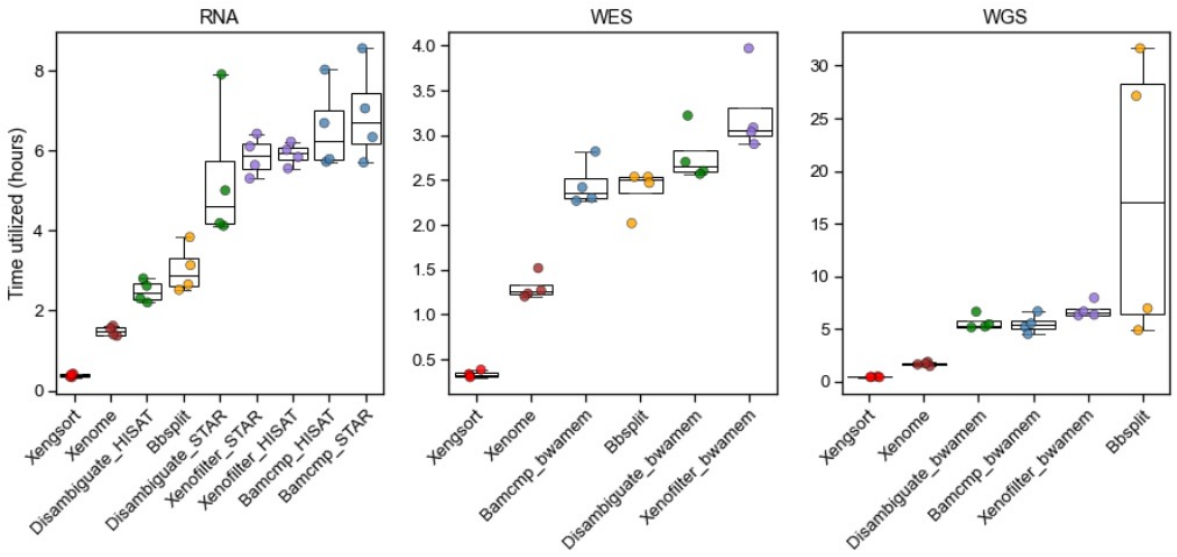
Supplementary Figure 4. Correlation of the copy number profiles between unsorted and sorted samples by different protocols. Each dot represents the correlation between the unsorted and the sorted by one protocol.



Supplementary Figure 5. Among the three data types, more reads are classified as ambiguous in RNAseq data. Each dot represents a sample. Y axis, percentage of ambiguous reads.



Supplementary Figure 6. Comparison of prediction accuracy in gold standard datasets. **a** Percentages of mouse reads classified as human in entire datasets in the three mixed datasets. **b** Percentages of human reads classified as mouse in entire dataset in the three mixed datasets.



Supplementary Figure 7. Running time of the computational protocols. Each dot represents one run. The running time can vary due to dynamic resource allocation on Microsoft Azure cloud platform.