CANcer Variant Allele Sequencing (CANVAS): A computational pipeline for the quantitation of ultra-low frequency hotspot mutations in myeloid neoplasm-associated genes via targeted error-corrected sequencing of human peripheral blood DNA

Page McKinzie, Jennifer Faske, Lascelles Lyn-Cook, Jr., and Meagan Myers
Division of Genetic and Molecular Toxicology, National Center for Toxicological Research, U.S. FDA, Jefferson, AR, USA

Abstract

Error-corrected sequencing describes a set of sequencing protocols that focus on mitigating the effects of induced errors in calling variants and is often used for evaluating somatic mutations associated with carcinogenesis. There are several laboratory techniques that can be used which must also be matched with a data processing computational workflow that processes the specialized raw sequencing data to highly accurate consensus sequences, allowing for the observation of smaller effect sizes between a control set and an exposed or treated set of samples. The current report of CANVAS describes the preparation of uniquely barcoded DNA libraries of synthesized sequences in known amounts and previously evaluated genomic DNA (gDNA) and the computational workflow that uses mostly standard sequencing software tools to output sequences corrected for errors introduced during PCR, library preparation and sequencing. Mutant fraction samples were generated using in vitro constructed standards to represent ratios of 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5} for a set of known blood-based cancer driver mutations and gDNA of samples previously evaluated with ACB-PCR and ddPCR were also prepared and sequenced. The performance of CANVAS was evaluated using both the results of the mutant fraction standard samples and by comparison of the gDNA evaluation to the quantitative range of ACB-PCR and ddPCR (10^{-1} to 10^{-5}). The target amplification adds molecular tags to both ends of target sequence and the CANVAS program takes advantage of these unique identifiers to provide highly accurate counts of sequences. The logic of CANVAS removes the introduction of unidentifiable nucleotides ("N") by comparing sequences with the same UID as whole words rather than single letters, making it less noisy and using less computational time. The results show this method accurately evaluates the abundance of mutations at each position over the range of 10^{-1} to 10^{-5} as determined using the constructed mutant fraction samples and comparison to results obtained from orthogonal methods. Application to future avenues of research include iPSC-derived biologic product safety assessment, and blood-based biomarkers of risk for cancer (*e.g.* therapy-related or after Phase I clinical trials) and noncancer diseases of aging and inflammation.

Introduction

Quantitative measurement of low frequency cancer- and disease-associated gene sequences, such as myeloid neoplasm-associated mutations, could advance risk assessment and benefit future monitoring paradigms to facilitate personalized health care for patients at risk. Low frequency mutations can be quantified by unique labeling of sample DNA molecules during library preparation that allows bioinformatic correction of sequencing errors in mutation counts. This report evaluates a targeted amplicon panel with a ligation-based library preparation technique along with a linux-based computation workflow made to resist obsolescence. It was evaluated by several methods, including a standard curve test, comparison to other sensitive methods of mutation quantitation and replicate measurements, comparison to a previously published workflow, and optimization of DNA polymerase. These results show that this approach correctly measures DNA mutation with a faster and easily parallelized computation program, providing a highly sensitive and quantitative means to assess the induction of new mutation and/or clonal expansion of preexisting mutant cells in human blood.

Table 1. Myeloid neoplasm-associated hotspot mutation panel

Gene	Hotspot codon	# of amplicons	Cell function
DNMT3A	R882	1	Epigenetic modification
IDH1	R132	1	
IDH2	R140 R172	2	
SETBP1	S867 S868 G870 T873	1	
JAK2	V617	1	Cell signaling
FLT3	D835	1	
KIT	N822 D816	1	
KRAS	G12 G13	1	
NRAS	G12 G13 Q61	2	
PTPN11	D61 A72 E76	1	
**RUNX1*	S141 R162 R166 D198 R201 R204	2	Transcription regulation
SF3B1	H662 K666 K700	2	RNA splicing regulation
TP53	R175 V216 Y220 Y234 M237 G244 R248 R273 C277 R282	4	Cell cycle/DNA damage response

*hotspot codons added after literature review

Panel consists of 13 genes, >40 hotspot codons and 20 amplicons and was designed via a systematic analysis of the COSMIC and ClinVar for the top 20 most mutated genes and their respective hotspot drivers present in blood cancers of myeloid origin, including acute myeloid leukemia (AML), therapy-related AML, myelodysplastic syndrome (MDS), and AML-associated with MDS (≥ 1% of neoplasms of each type).

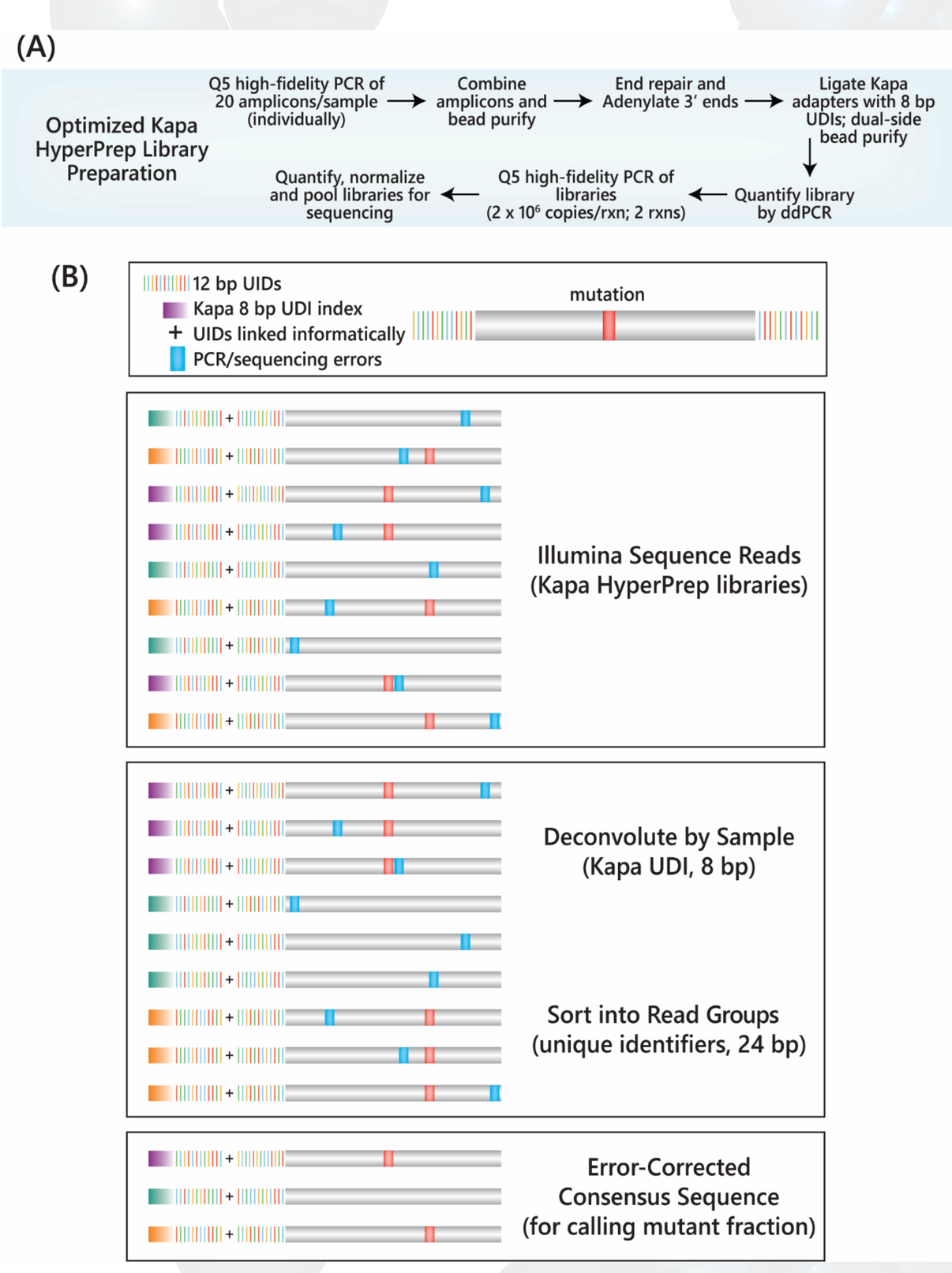


Figure 1. Overview of target amplification and library preparation (A) and ecNGS labeling with UIDs and adapter indices to construct ECCS using CANVAS (B).

CANVAS

Figure 2. CANVAS Computational Workflow

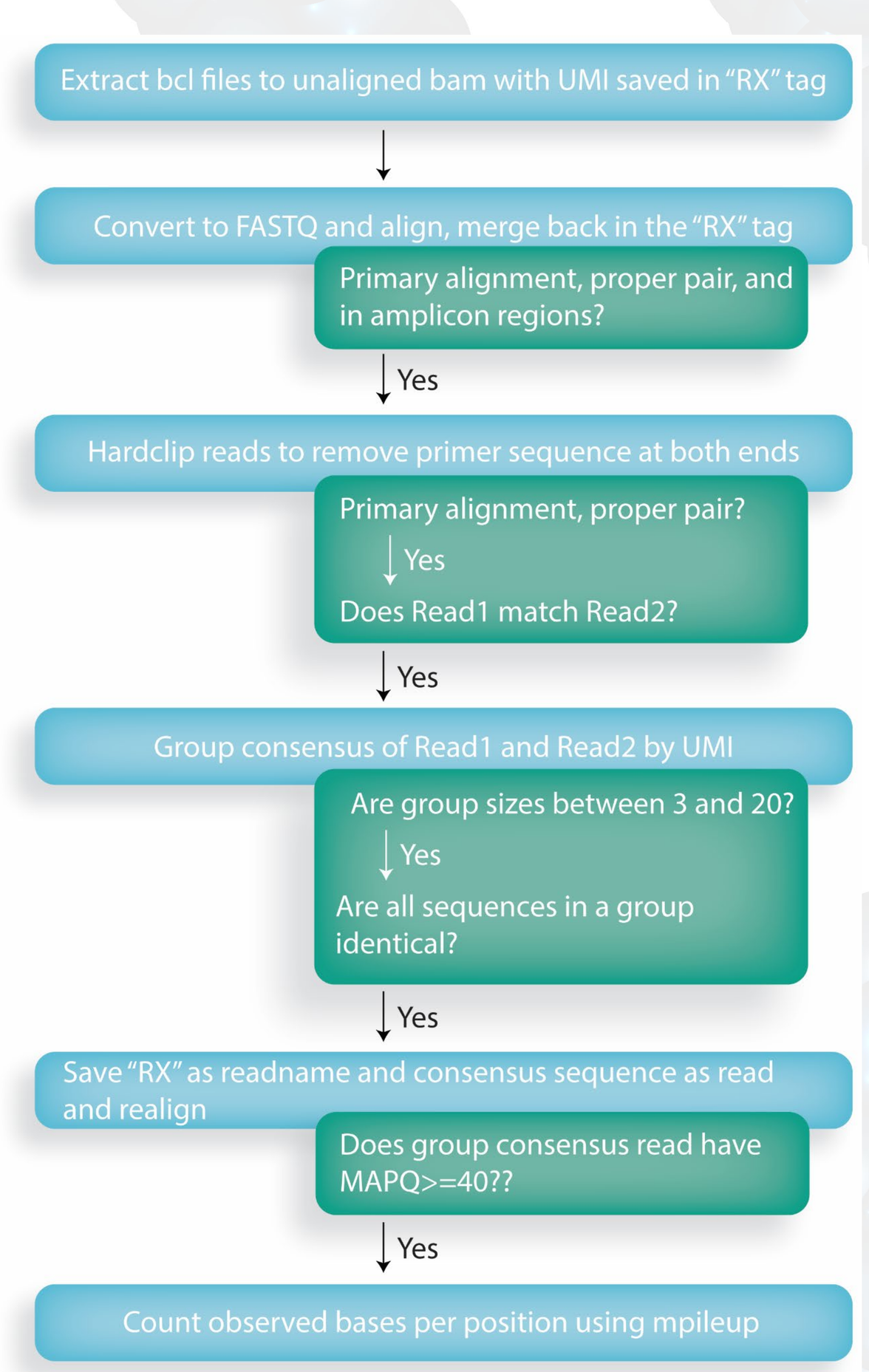


Table 2. Comparisons between the CANVAS and commonly used Kennedy *et al.* pipeline.

Component	Kennedy *et al.* pipeline†	CANVAS
Portability	• Requires fixed version of python (interpreter language)	• Uses bash scripting and commonly used, stable tools (*i.e.* bwa, samtools, etc.) to make robust to obsolescence due to programming language updates
Time	• ~4.5 hrs per sample; 1 sample at a time, single computer	• ~4 hrs per flow cell on HPC or ~4 hrs per sample using the option selection to use a personal computer
Evaluation	• Input files are fastq • UMIs determined from reads in fastq files, removed from sequence, and added to read name • Reads separated by read lengths and aligned for each length • Observed bases for each position is counted from the samtools mpileup output	• Input files are BCL files • UMIs determined during Picard tools basecall extraction of bcl files and saved as RX tag in an unaligned bam (ubam) • All reads in ubam are converted to fastq and aligned • UMI RX tag merged into aligned bam and read overhands hard clipped • Sequences filtered for read 1 matching read 2 • Group by tags, keep groups with 100% of reads matching in sequence, make ubam with UID as read name and consensus as read • Realign and count bases observed using samtools mpileup
Differences	• Because strand consensus sequences are determined by position, non-informative bases can be output (as a count of "N" bases) • Read1 and read2 are not used to confirm sequence of read	• Strand consensus sequence is determined by entire sequence prior to mpileup and counting, removing output of non-informative bases • The final alignment can be easily filtered for mapping quality to decrease counting of potential pseudogene amplifications • Annotates all observed mutations (using ANNOVAR and RefSeq database

†https://github.com/Kennedy-Lab-UW/Duplex-Seq-Pipeline

Results

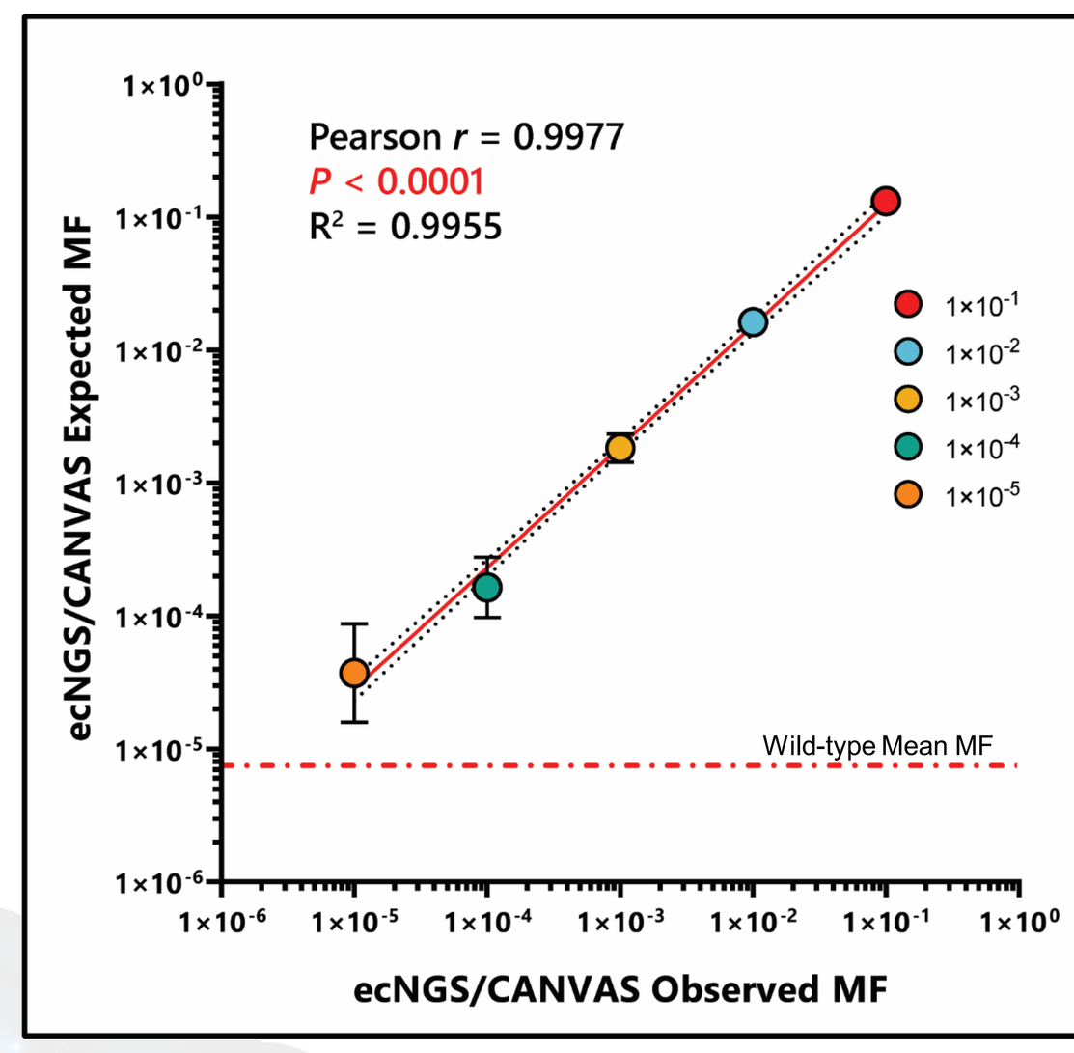


Figure 3. Mutant fraction (MF) standard curve demonstrating linearity and precision of ecNGS/CANVAS MF measurements down to 1×10^{-5}. MF standard curve consisting of 15 hotspot mutations in 8 myeloid neoplasm-associated genes quantified by ecNGS/CANVAS. MF samples of 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5} were constructed using synthesized and cloned wild-type and mutant IDT gBlocks™ containing a set of known blood-based cancer mutations.

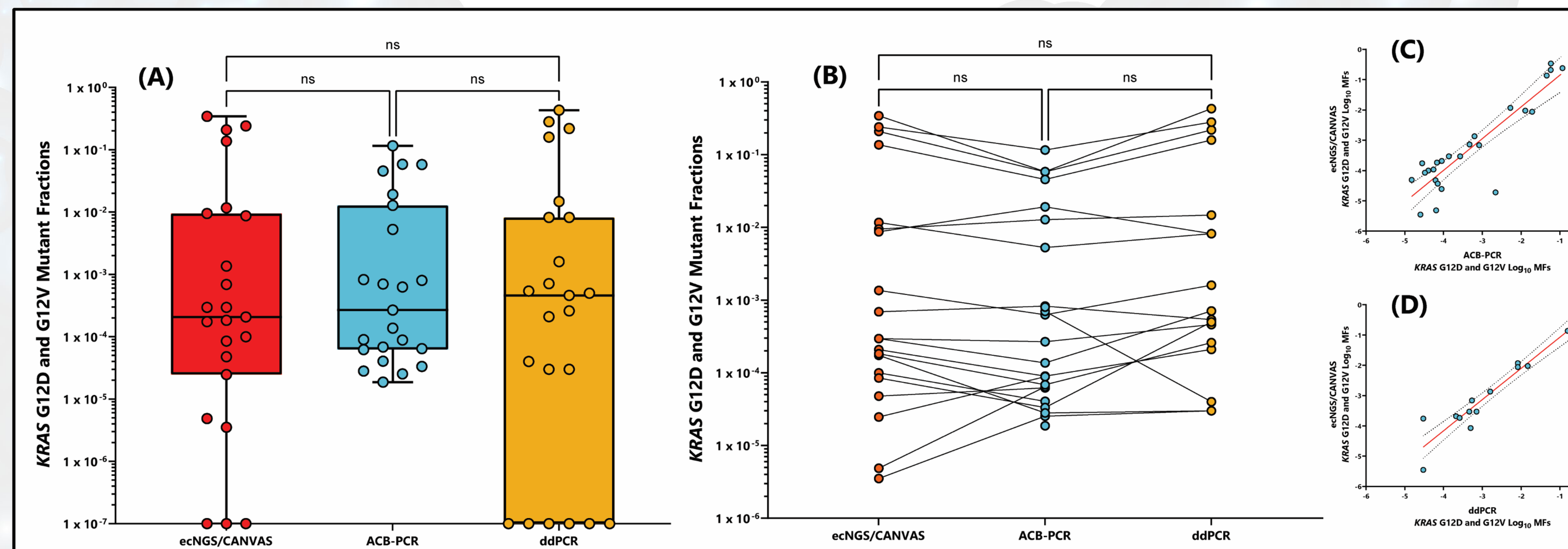


Figure 4. Validation of ecNGS/CANVAS MF measurements through concordance with orthogonal methods. *KRAS* G12D and G12V MFs quantified in lung adenocarcinoma gDNAs (n = 9) by each ecNGS/CANVAS, Allele-Specific Competitor Blocker PCR (ACB-PCR) and Droplet Digital PCR (ddPCR). MFs plotted at 1×10^{-7} indicate no calls in the ecNGS/CANVAS or no calls or MFs below the limit of detection in ddPCR (~2×10^{-4} for *KRAS* G12D and ~1×10^{-5} for *KRAS* G12V) (A). Paired MFs quantified by each method are depicted (B); Friedman test, P = 0.3932 (*n.s.*). Linear regression of *KRAS* G12D and G12V MFs quantified; Spearman r = 0.8013 and 0.9536; P < 0.0001, respectively (C and D).

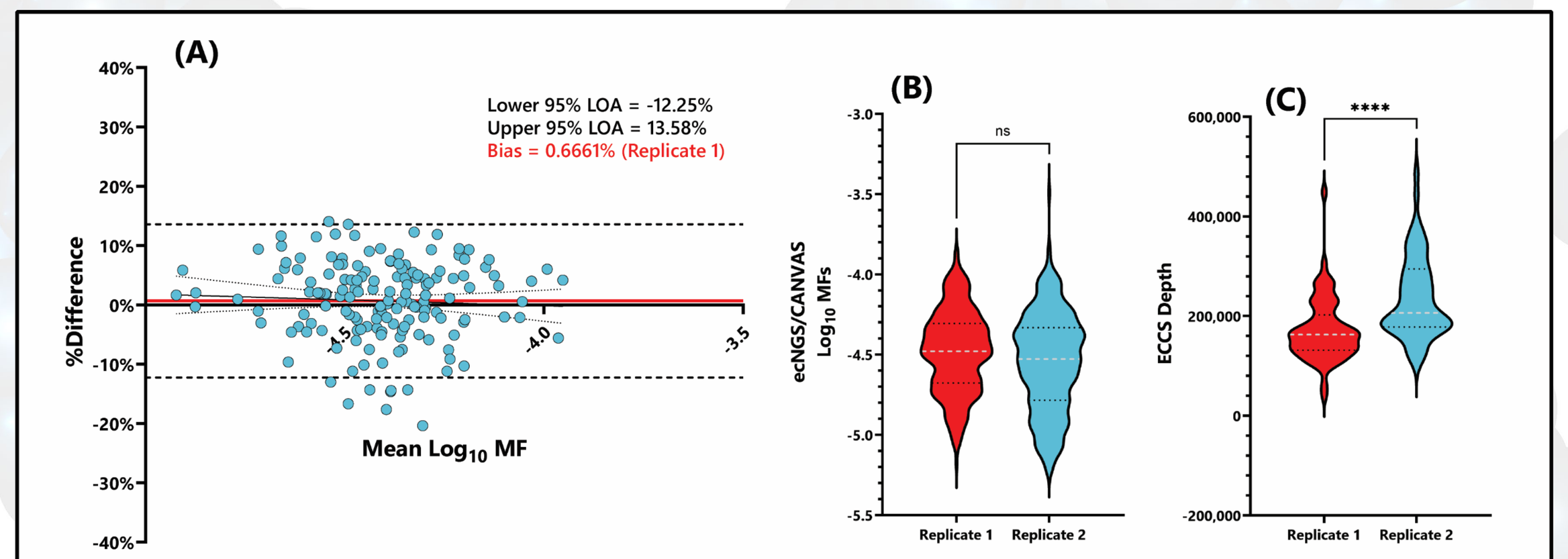


Figure 5. Reproducibility of ecNGS/CANVAS is demonstrated by agreement and minimal bias of MF measurements in replicate samples. Bland-Altman plots depicting %bias and the 95% limits of agreement (LOA) for ecNGS/CANVAS replicate MF measurements (A), their respective MFs (*n.s.*) and their error-corrected consensus sequence (ECCS) depths; Wilcoxon signed-rank test, P < 0.0001 (B).

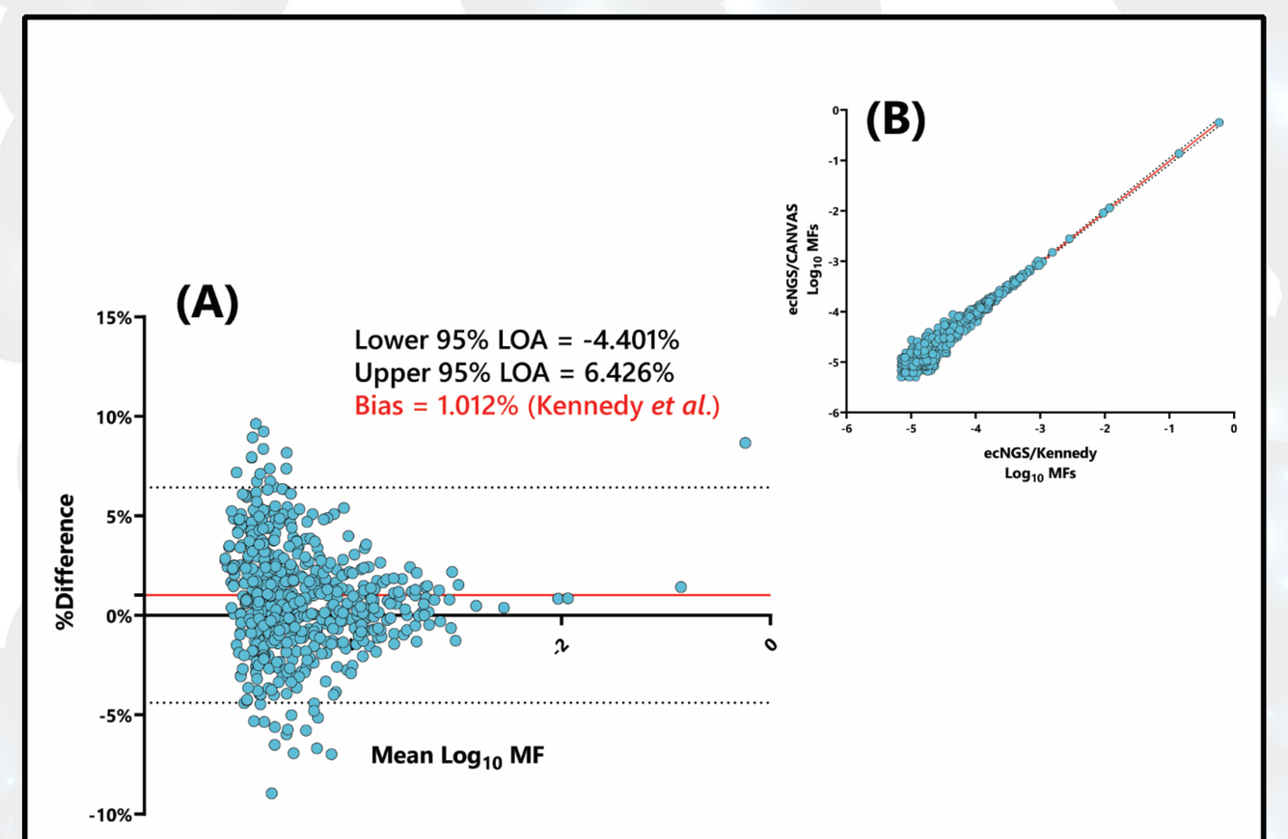
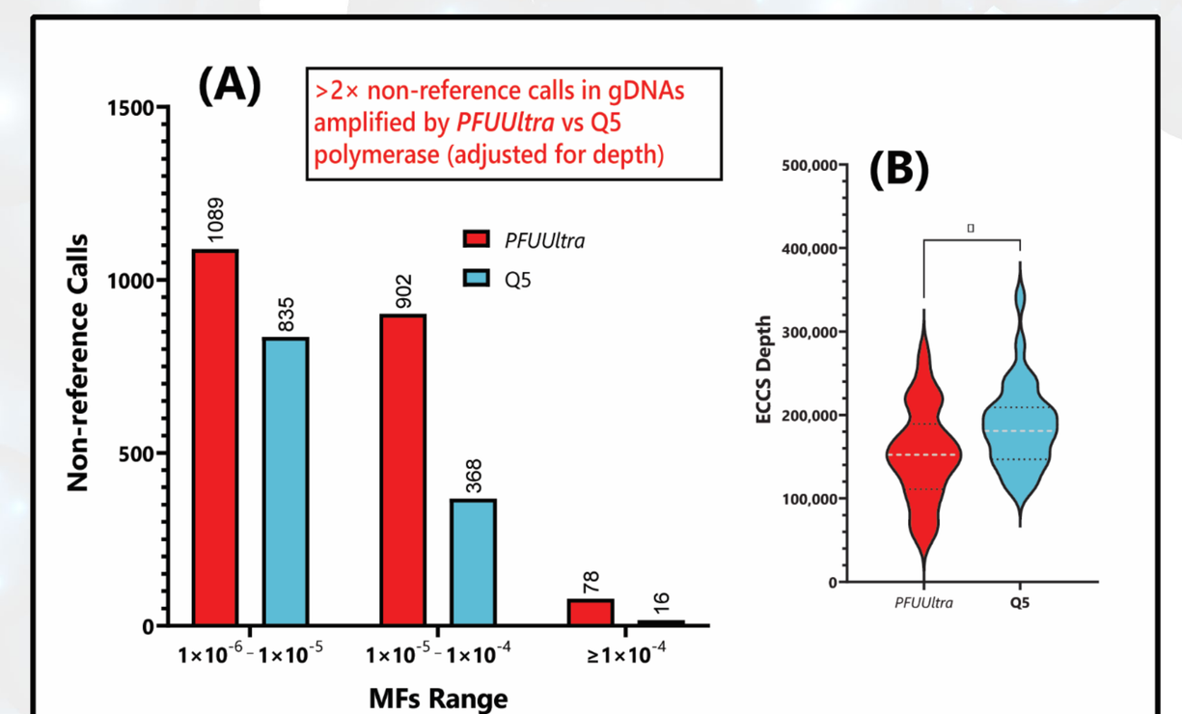


Figure 6. Agreement and correlation of MFs reported by CANVAS and Kennedy *et al.* computational workflows demonstrates concordance. Comparison of the agreement (A) and correlation (B) of reported MFs by CANVAS and Kennedy *et al.* pipelines. Spearman r = 0.8813; P < 0.0001.



f *PFUUltra* polymerase for target tly more non-reference lymerase, despite 16% lower th for PFU. ecNGS/CANVAS reported non-reference calls in lung adenocarcinoma (n = 4) and PBMC (n =) gDNAs amplified by Q5 or *PFUUltra* polymerase. Fisher's exact test, P < 0.0001 (A) and their respective ECCS depths. Wilcoxon signed-rank test, P < 0.0001 (B).

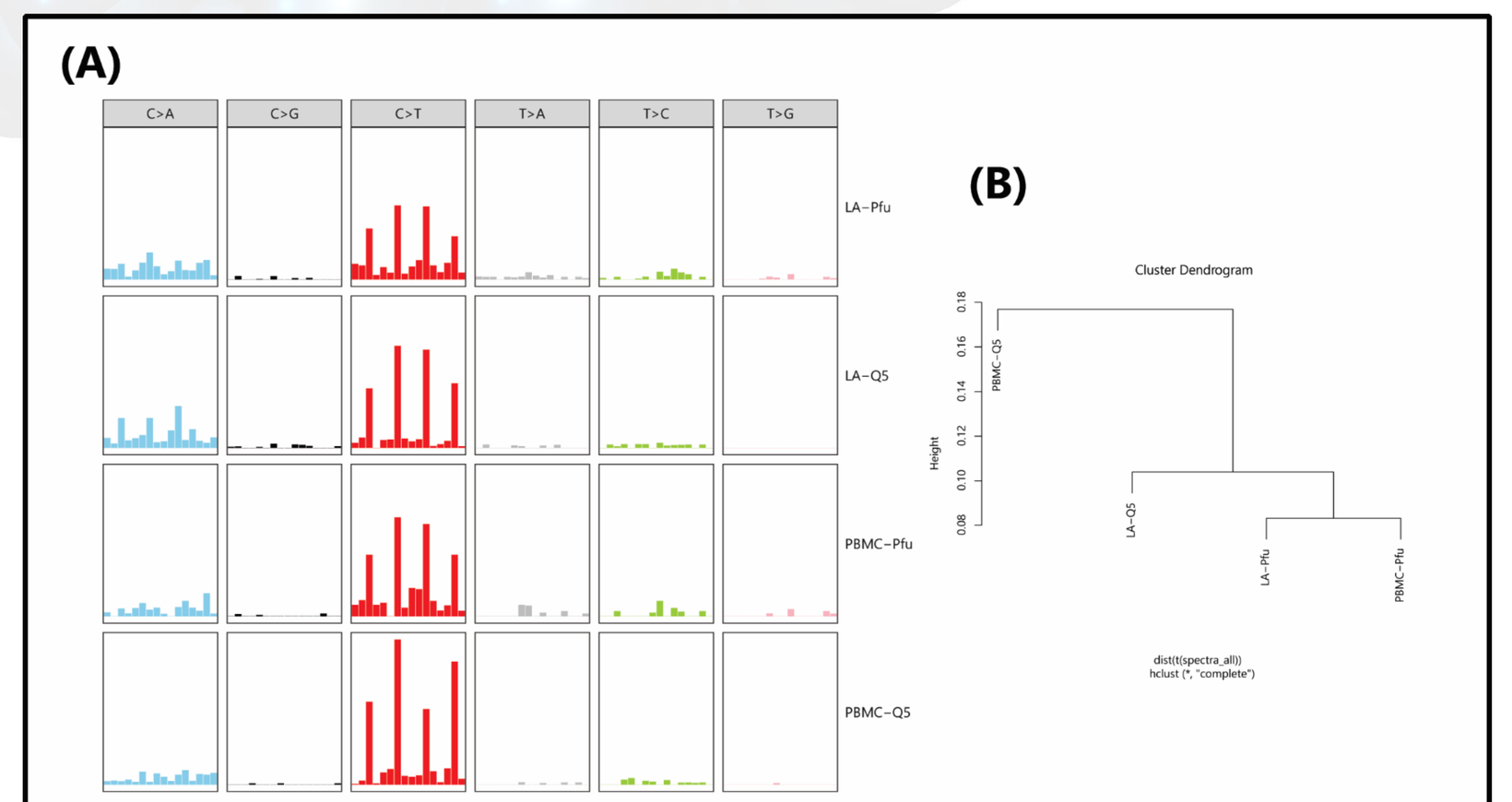


Figure 8. Mutational spectra and hierarchical clustering shows the polymerase used is more important than the cell type or DNA extraction used on mutational signatures. Mutational spectra (normalized) (A) and corresponding dendrogram of the hierarchical clustering (B) of ecNGS/CANVAS reported mutations in lung adenocarcinoma (extracted by phenol chloroform; n = 3) and PBMC gDNAs (extracted by salt precipitation; n = 2) amplified by Q5 or *PFUUltra* polymerase.

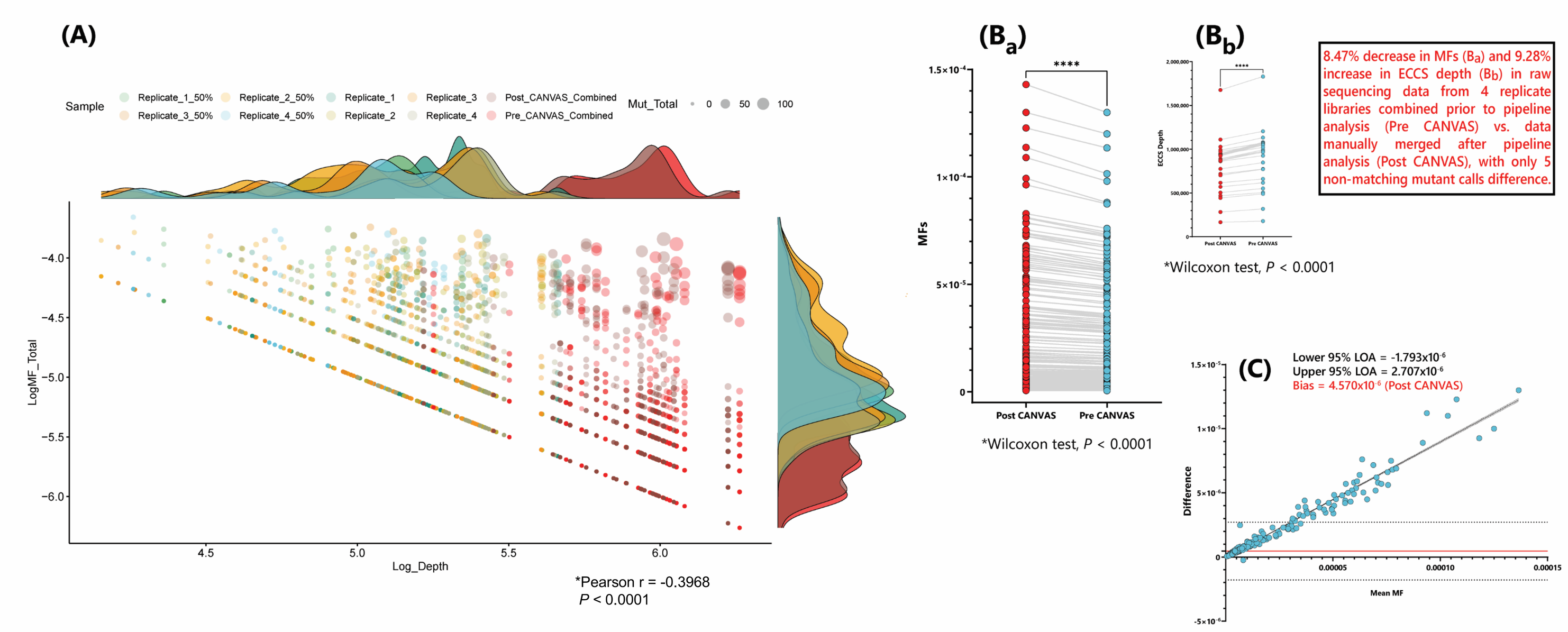
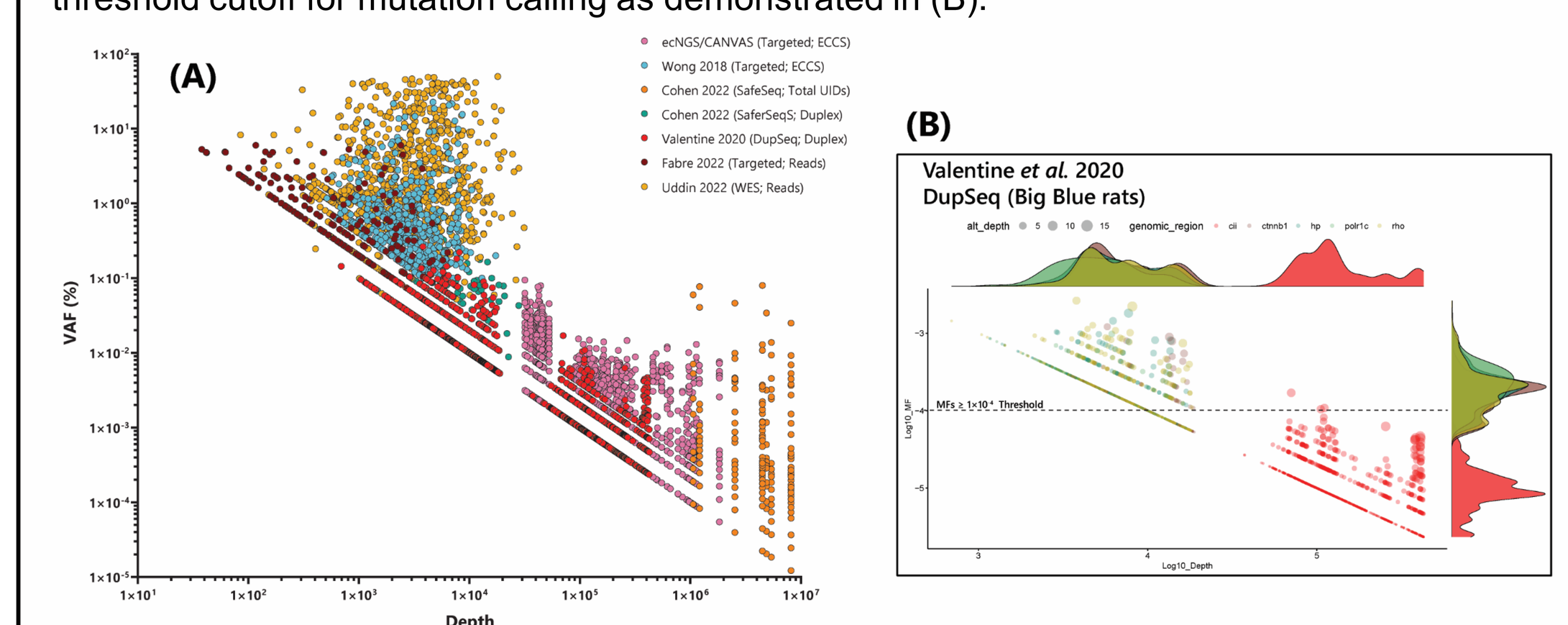


Figure 9. MF quantitation is biased and is a function of ECCS depth, with significant increases in measured MF values as depth decreases (A). With increasing depth, the power to detect a mutant molecule increases. We postulate the probability to call a mutant is disproportional to calling a wild-type, resulting in progressively lower MFs as depth increases. This is demonstrated by comparing the number of non-reference calls, MFs (B_a) and ECCS depths (B_b) of 4 independent libraries combined prior to the CANVAS pipeline (Pre CANVAS) to that of the data manually merged after the CANVAS pipeline (Post CANVAS). Interestingly, we find a strong proportional bias in MFs pre- and post-CANVAS, with the bias increasing as MF values increase (C).

Sample names in (A):
- Replicate libraries (Replicate_1 - Replicate_4)
- 50% random downsampling (Replicate_1_50% - Replicate_4_50%)
- Independent MFs combined manually after the CANVAS pipeline (Post_CANVAS_Combined)
- Combining the raw sequencing data prior to the CANVAS pipeline (Pre_CANVAS_Combined)

Figure 10. Impact of depth on reported MFs is ubiquitous across sequencing studies. Reported MFs and depth was collected from published studies that used different sequencing techniques and depth measurements (A). The impact of this bias could be substantial if quality metrics, such as depth criteria, are not assured, especially in the case when implementing a threshold cutoff for mutation calling as demonstrated in (B).



Conclusions

CANVAS can accurately quantify mutations from amplicon-based sequencing methods with single or dual-end UIDs and unequal lengths of UIDs.

CANVAS reduces counting of artifactual variants by comparing read 1 and read 2, and by deriving consensus from the entire read as an entity before counting variants at each position.

ecNGS/CANVAS correctly measures MFs by comparison to other MF quantitation methods, replicate measurements, a standard curve test, and comparison to a previously published workflow. ecNGS/CANVAS also showed better performance with Q5 than with *PFUUltra* polymerase.

ecNGS/CANVAS demonstrated an effect of depth on MF results, which is ubiquitous across sequencing studies and techniques. This bias has largely been under-appreciated in low-frequency ecNGS studies, indicating the need for incorporating depth criteria in the quality metrics in ecNGS studies, and ideally, statistical and mathematical modeling approaches to mitigate this bias.

Future applications for the ecNGS/CANVAS myeloid neoplasm-associated hotspot mutation panel include: 1) discovery of predictive marker(s) of risk in cancer patients undergoing chemotherapy, 2) an early indicator of genotoxic risk in Phase I clinical trials, 3) a predictive and/or prognostic marker of risk in diseases of autoimmunity, aging and inflammation, and/or 4) a surveillance tool for somatic mutation in induced pluripotent stem cell culture models.