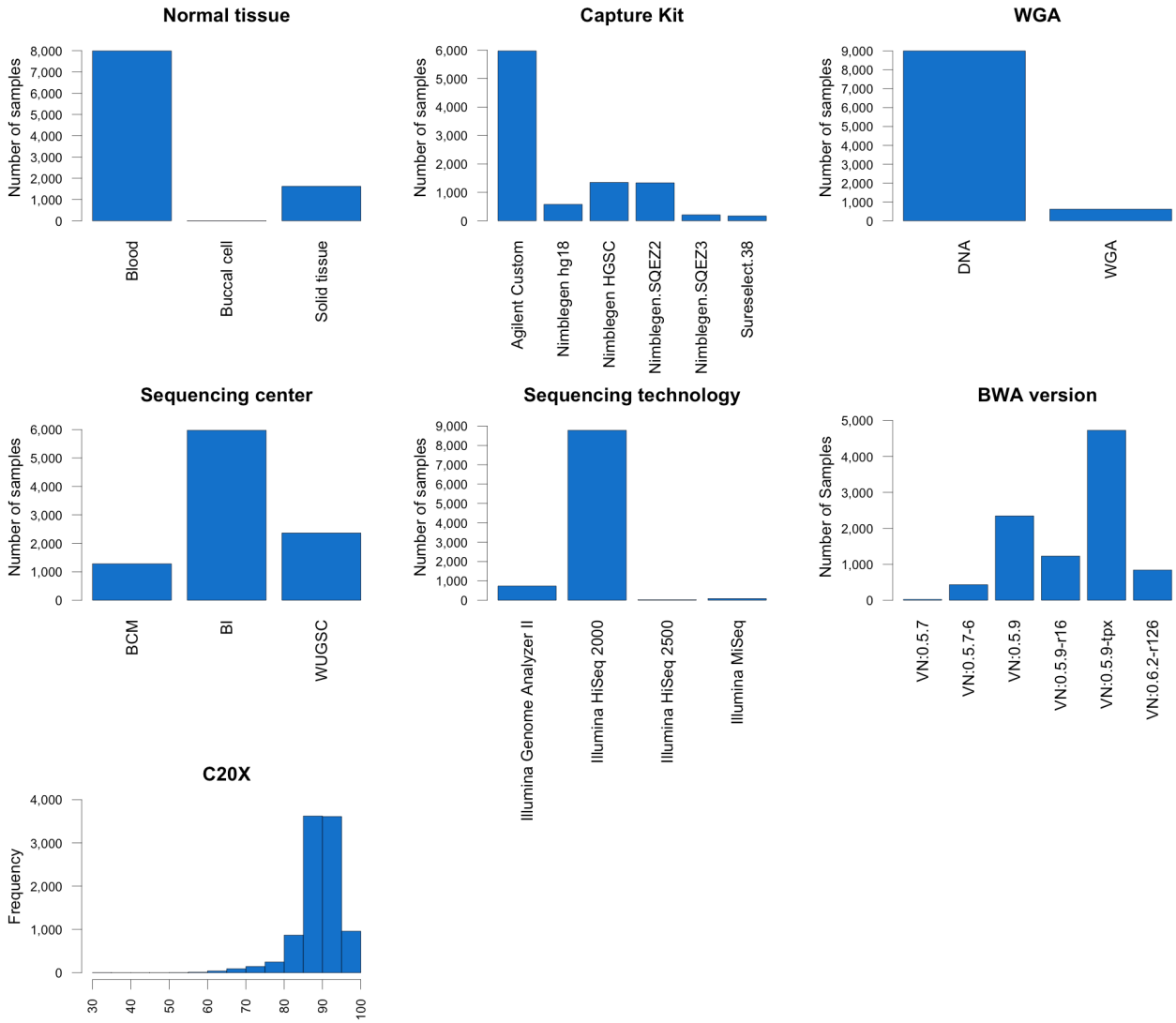
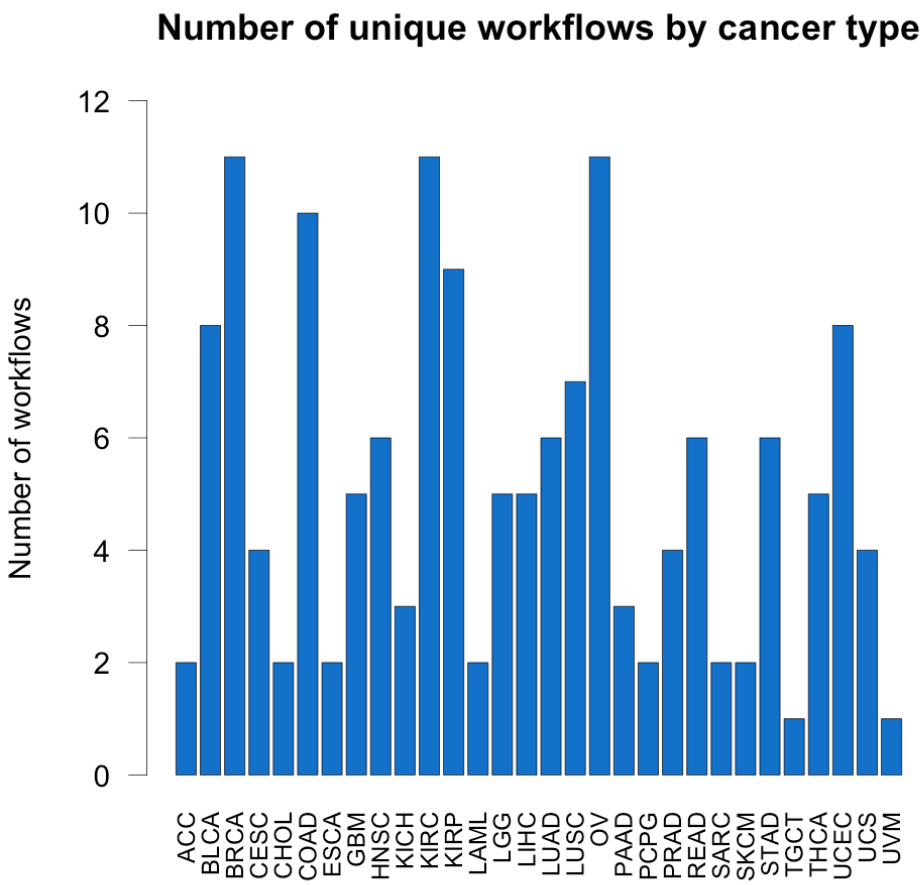
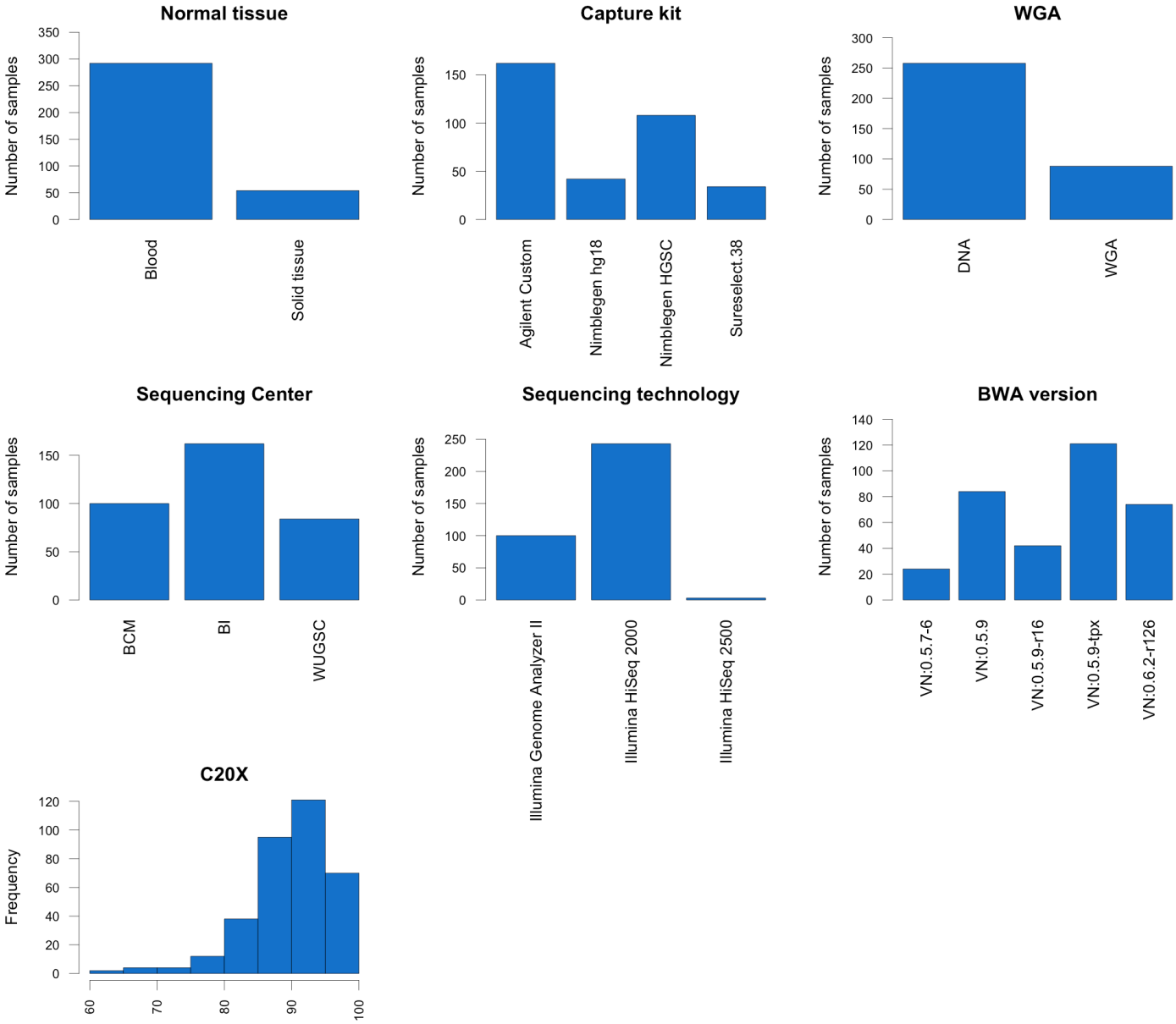


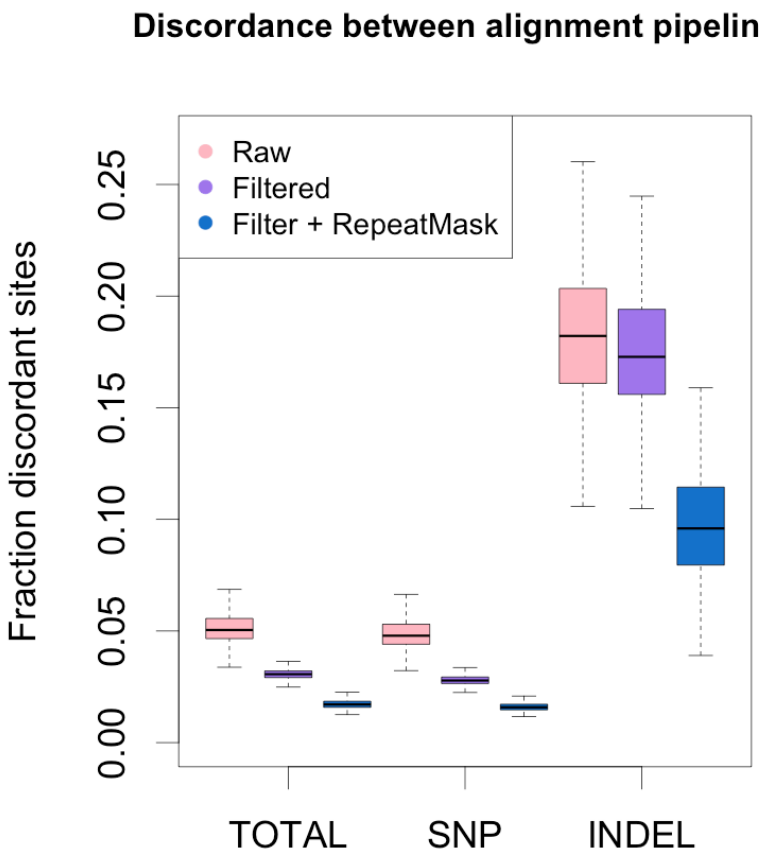
S1 Fig. Distribution of technical covariates for the pan-cancer cohort.
The distribution of the seven identified technical covariates for *n*=9618 TCGA WXS samples. Capture efficiency is measured as percentage of capture target area covered by at least 20 X read depth (denoted C20X)



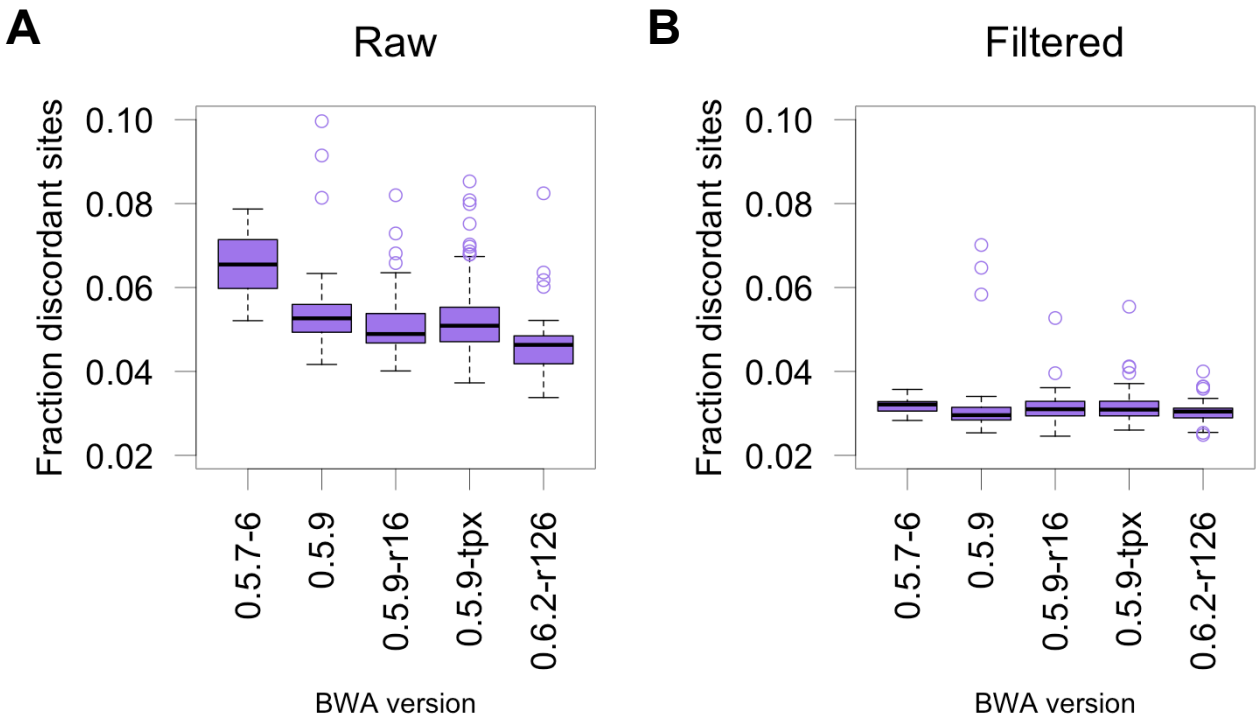
S2 Fig. Number of processing workflows used to generate TCGA WXS data.
The number of unique combinations of six technical factors (sequencing center, normal tissue, WGA, BWA version, capture kit, and sequencing technology) per cancer type



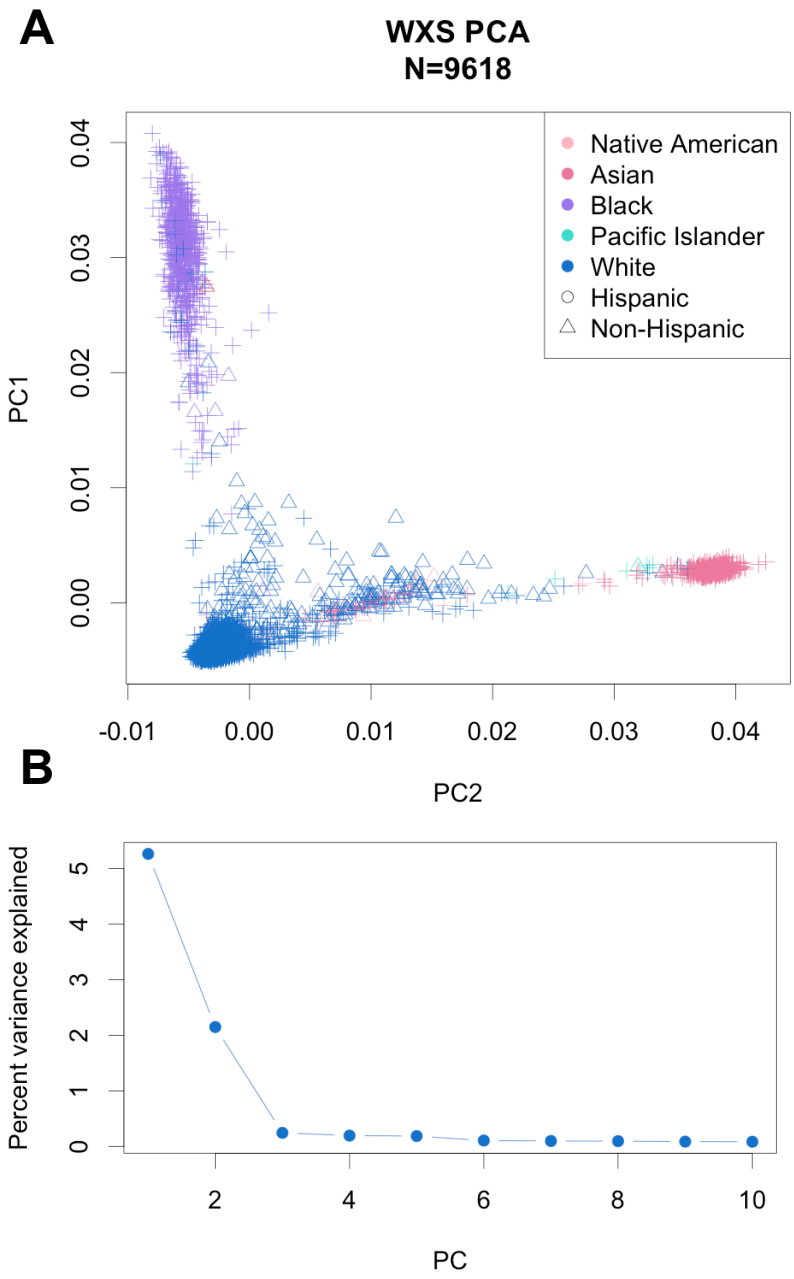
S3 Fig. Distribution of technical covariates for the NewAlign cohort.
The distribution of the seven identified technical covariates for *n*=345 TCGA WXS samples.



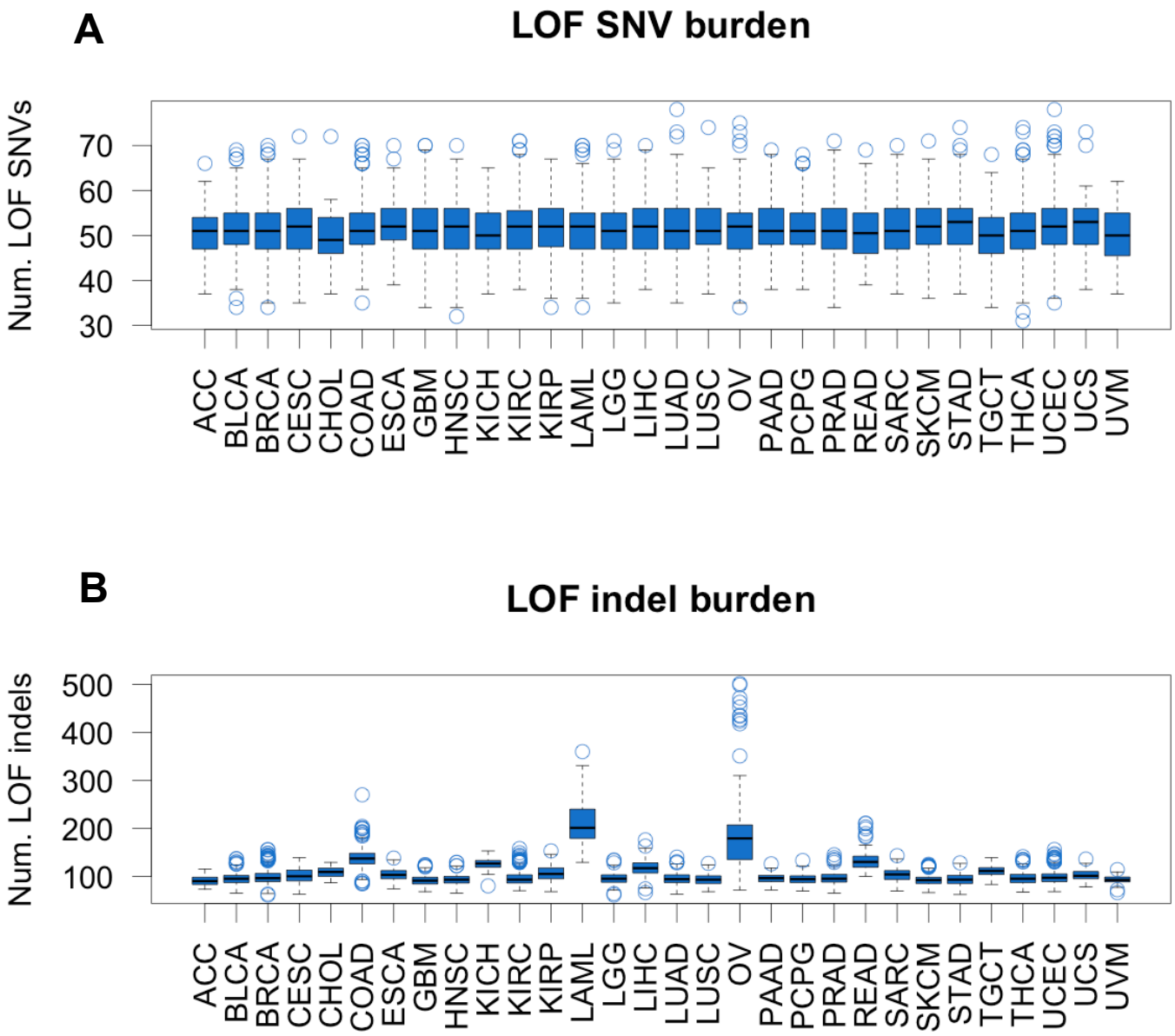
S4 Fig. Variant call discordance between NewAlign and OldAlign samples (*n*=345).
For filtered condition SNPs were filtered using GATK VQSR TS 99.5 and indels using GATK hardfilter. For filtered + RepeatMask condition variants in UCSC tracks RepeatMasker and Segmental Dups were excluded.



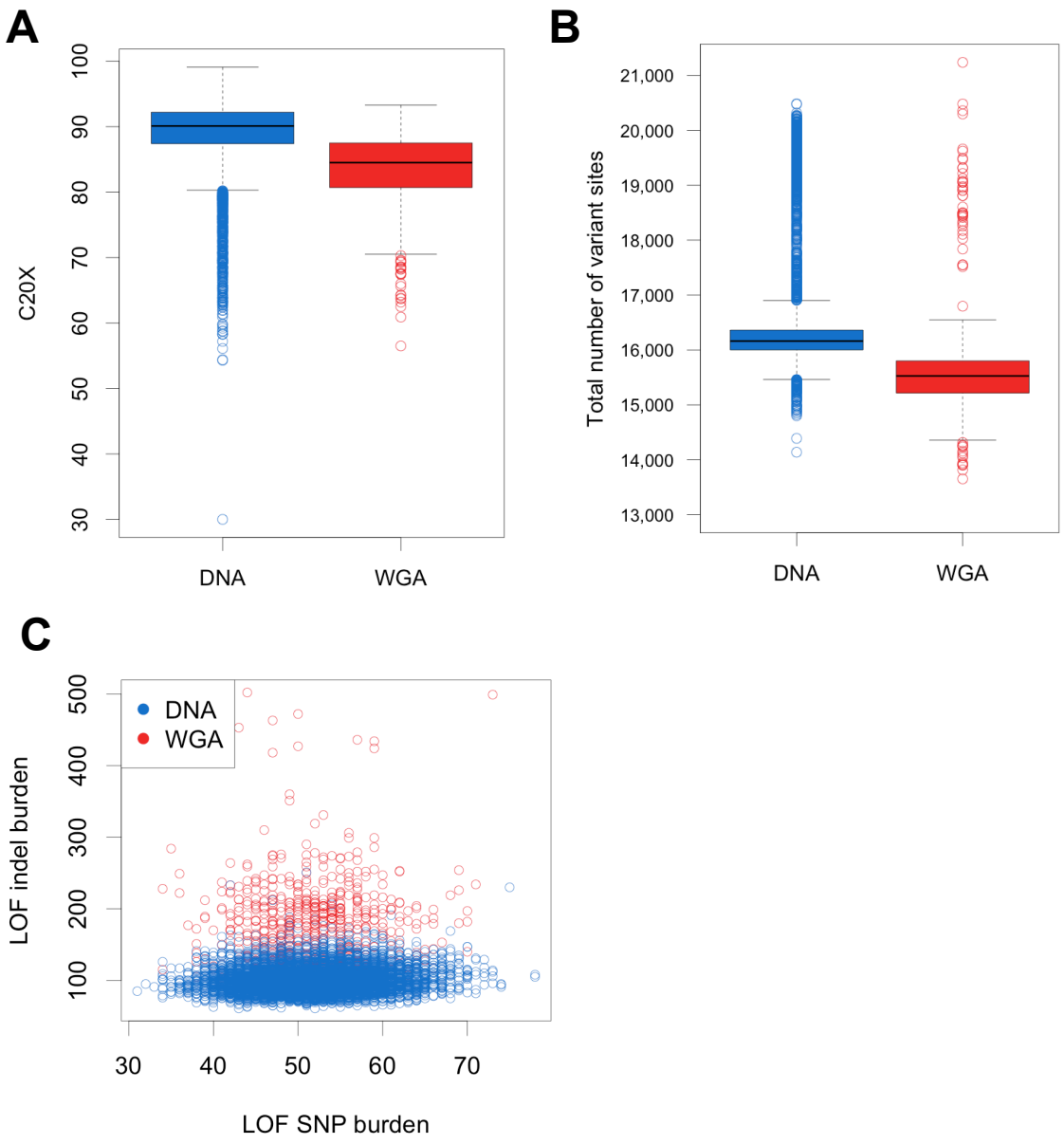
S5 Fig. Discordance with BAM realignment plotted by BWA version used to generate BAM file. (A) Raw VCF discordance between NewAlign and OldAlign samples plotted by BWA version. (B) Filtered VCF discordance between NewAlign and OldAlign samples plotted by BWA version. SNPs were filtered at GATK VQSR TS 99.5, indels with Hardfilters.



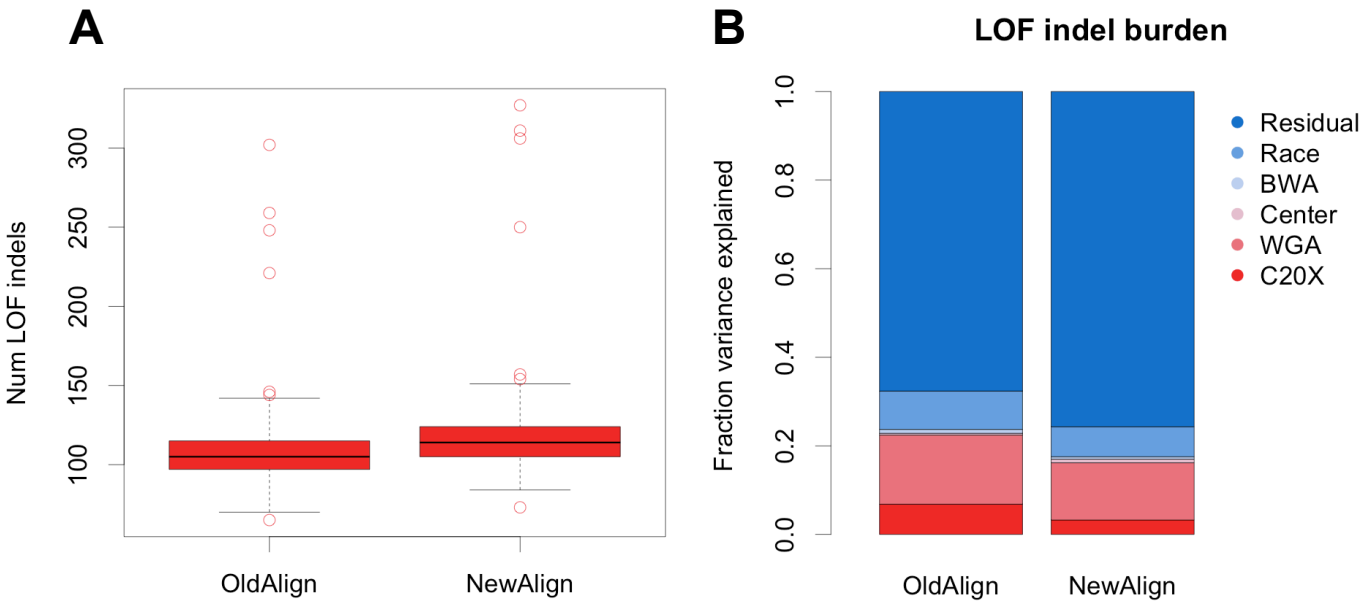
S6 Fig. PCA of common variants from pan-cancer VCF.
(A) Principal components 1 and 2 calculated from AF > 1% variants. Individuals colored by self report race. (B) Percent total variance explained by the top 10 principal components.



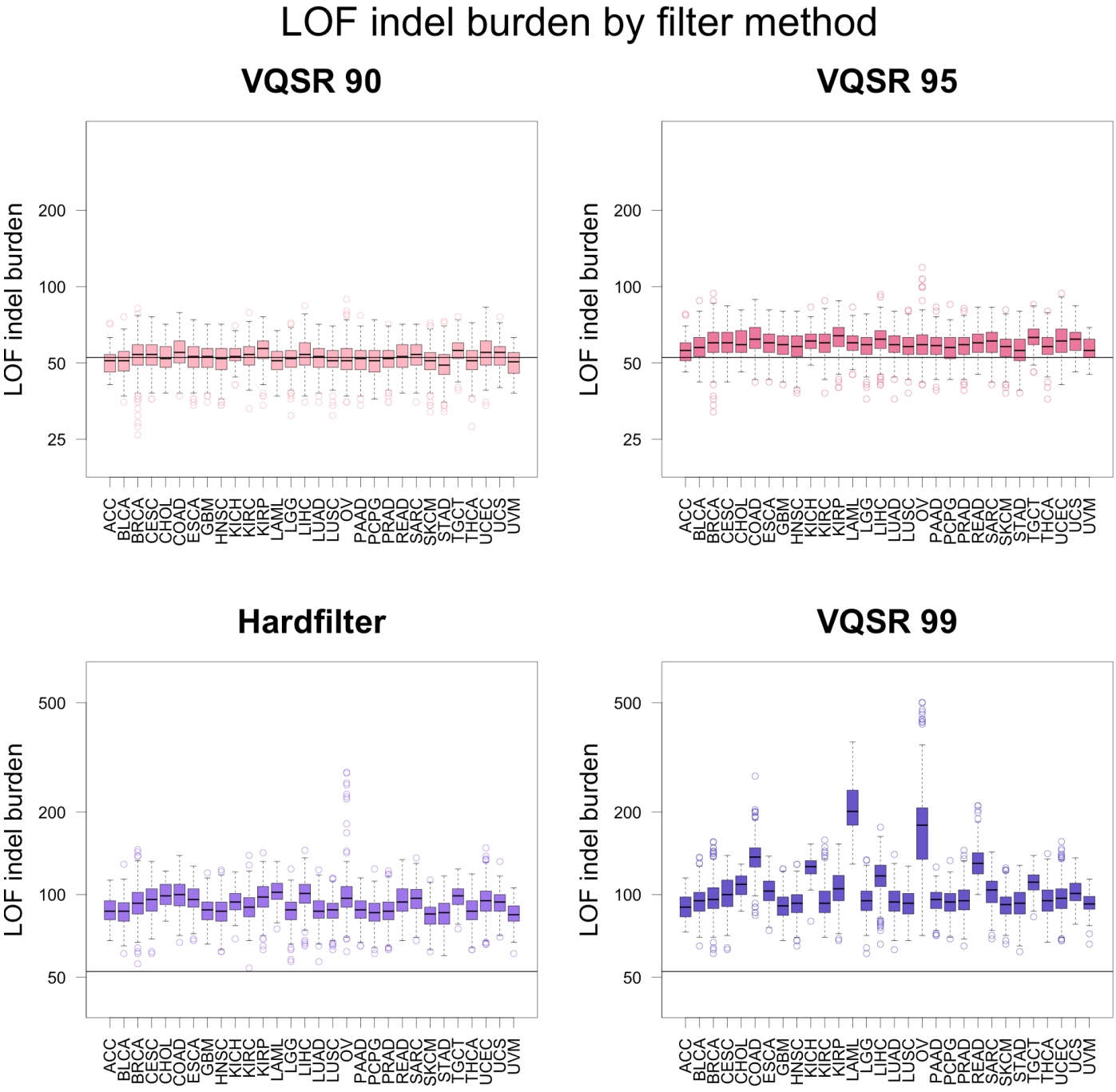
S7 Fig. LOF variant burden split by variant type.
(A) Individual LOF SNP burden plotted by cancer type. (B) Individual LOF indel burden plotted by cancer type



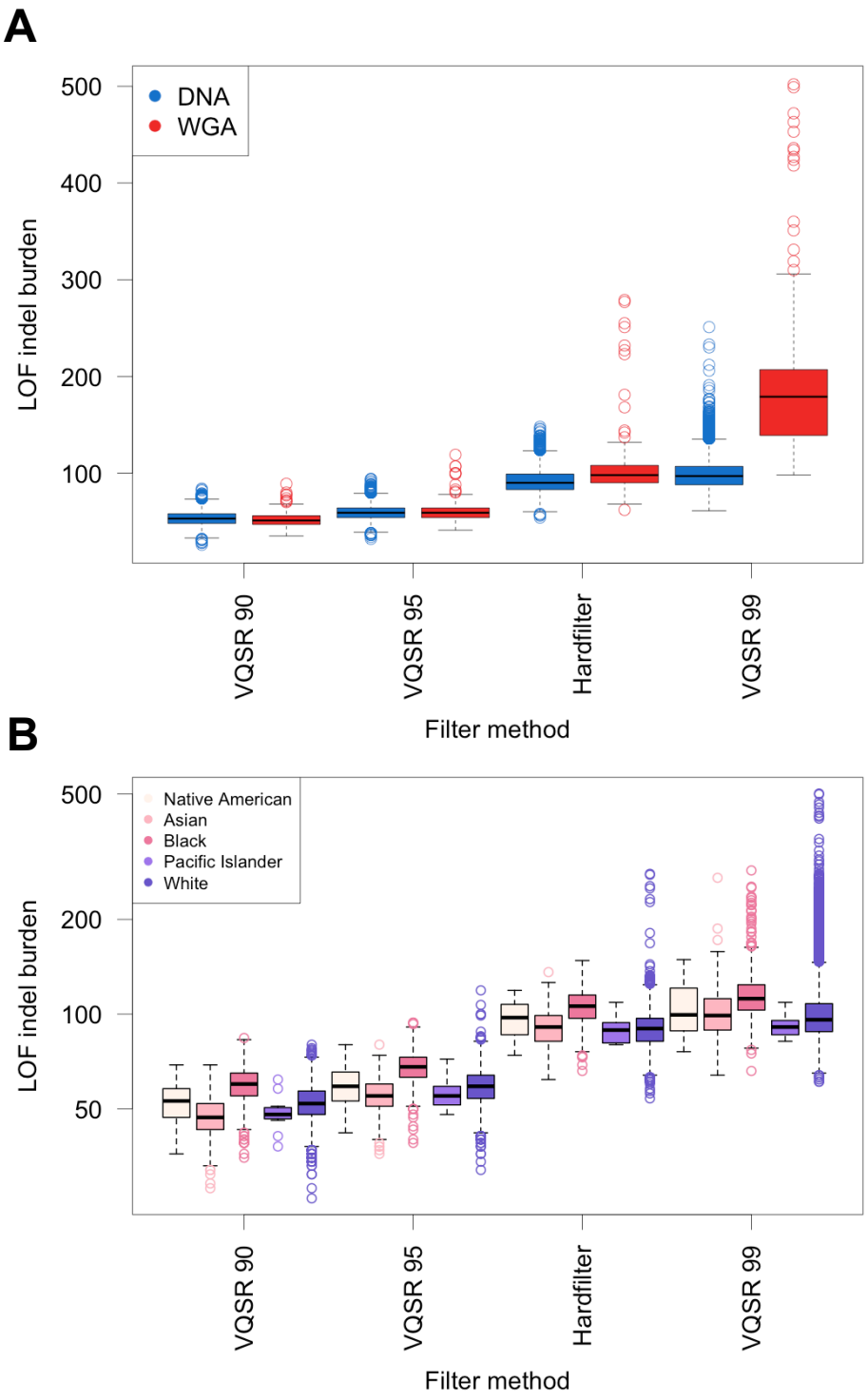
S8 Fig. Capture efficiency and variant profile of WGA samples.
(A) C20X plotted by WGA status. (B) Total number of variant calls plotted by WGA status. (C) Individual LOF indel burden vs. individual LOF SNP burden. Color indicates WGA status.



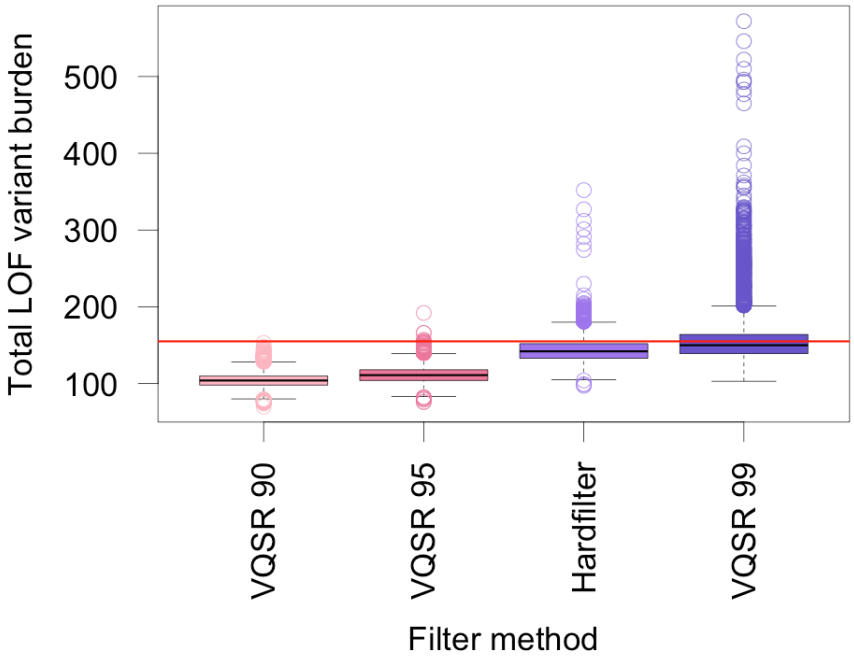
S9 Fig. Variance in LOF indel burden explained by technical factors in NewAlign cohort. (A) Number of LOF indels per individual in NewAlign and OldAlign pipelines. There were a median 8 more LOF indels in the NewAlign pipeline. Overall individual LOF indel burden was highly correlated between pipelines (Pearson R^2 = 0.947). (B) Percent of variation in individual LOF indel burden explained by technical covariates as assessed by ANOVA.



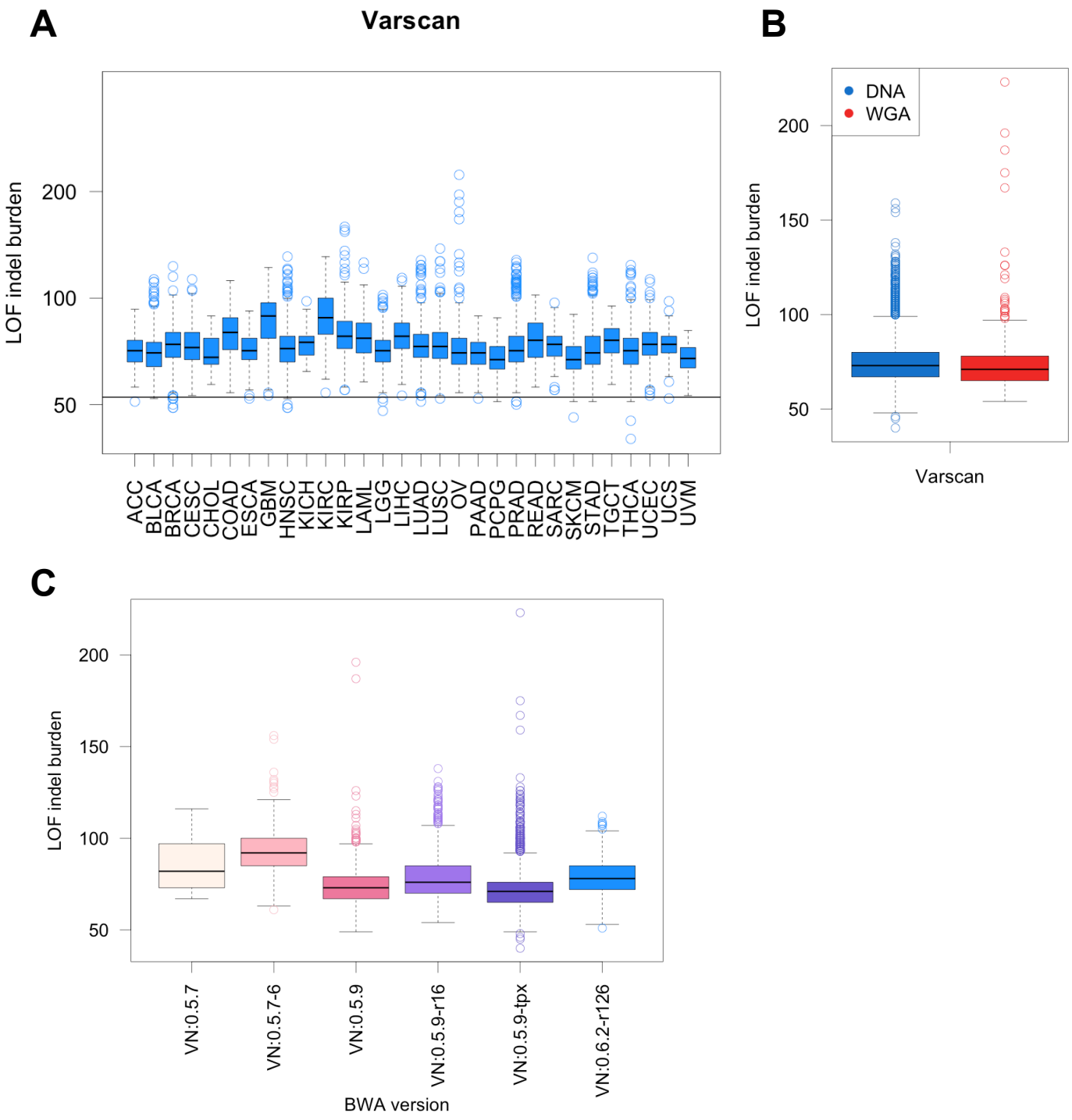
S10 Fig. Individual LOF indel burden by cancer type for all indel filtering methods tested. The black line represents the median LOF indel burden in the most stringent filter condition (VQSR 90) for comparison.



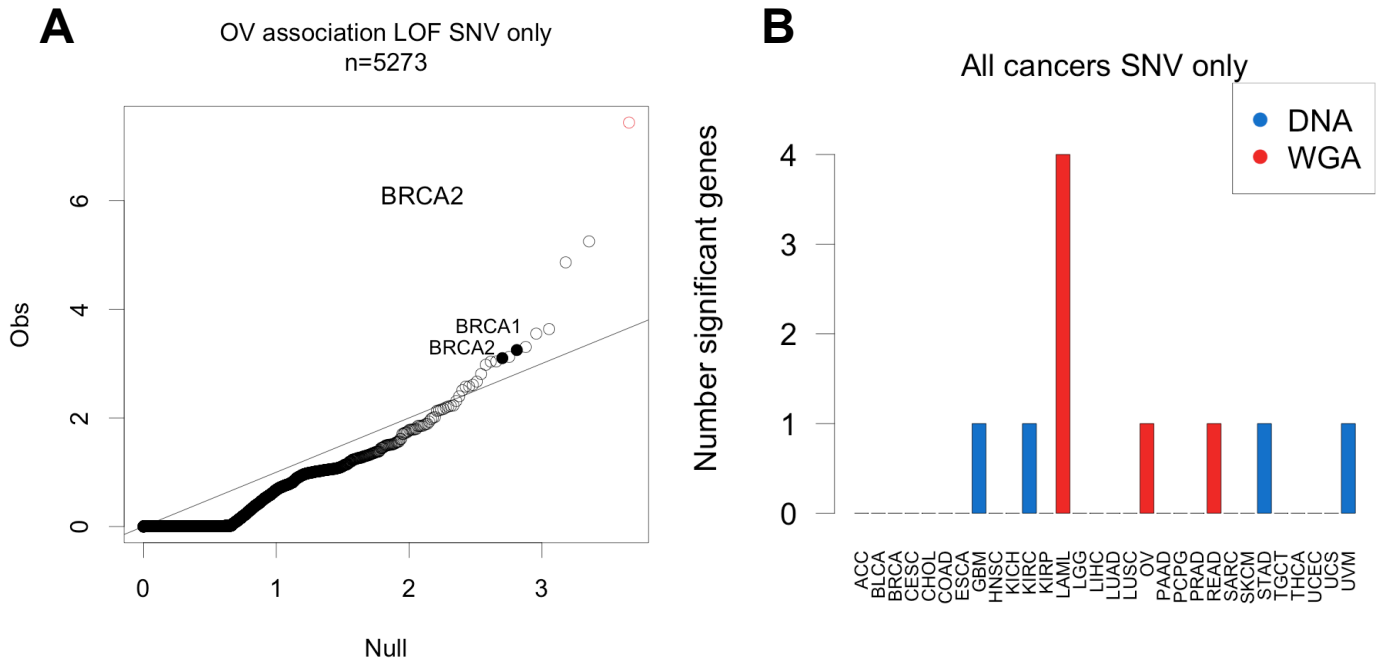
S11 Fig. LOF indel burden plotted by WGA status and race.
(A) Individual LOF indel burden by self-report race for each filter. (B) Individual LOF indel burden of WGA and DNA samples for each filter.



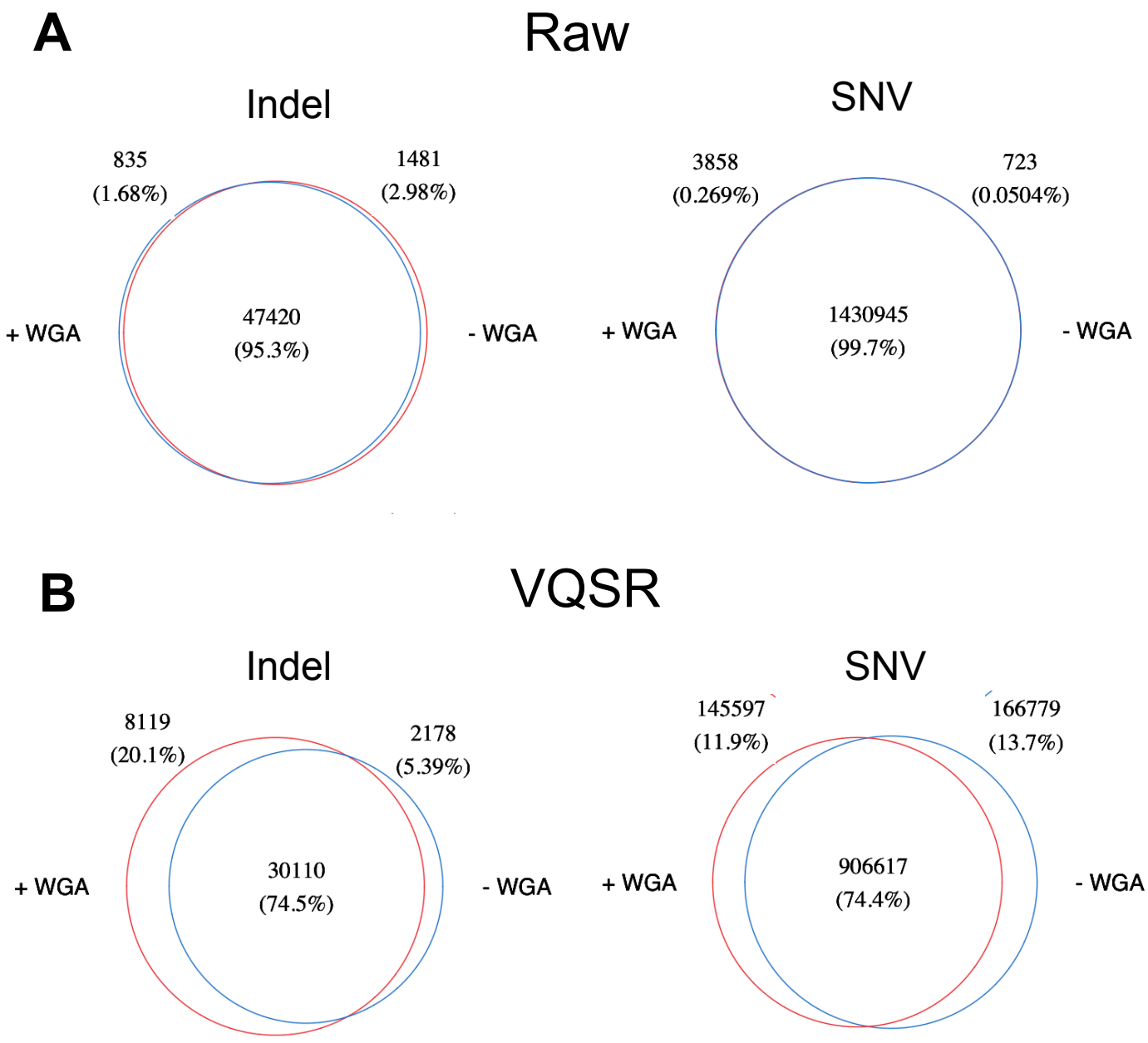
S12 Fig. Total LOF variant count for tested indel filter methods.
LOF variant count includes both SNV and indels.



S13 Fig. LOF indel burden from Varscan calls.
(A) Individual varscan LOF indel burden by cancer type. The black line represents the median LOF indel burden in the most stringent filter condition (VQSR 90) for comparison. (B) Individual varscan LOF indel plotted by WGA status. (C) Individual varscan LOF indel plotted by BWA version used to generate BAM file.



S14 Fig. Logistic regression analysis using only LOF SNVs.
(A) QQ plot from logistic regression association testing between germline LOF SNV burden and OV. N = number of genes tested. Red indicates associations significant $p < 1.61 \times 10^{-7}$. *BRCA1* association highlighted. (B) Number of genes significant $p < 1.61 \times 10^{-7}$ by logistic regression for all cancer types. Color indicates cancer types containing WGA samples.



S15 Fig. Overlap of variant calls after recalling variants excluding 614 WGA samples.
(A) Overlap of group-called raw variants between 9004 WXS DNA samples group-called in the full pan-cancer cohort (WGA+) or group-called omitting WGA samples (WGA-). (B) Overlap of the same samples after VQSR filtering. VQSR settings: SNP TS 99.5%, Indel TS 90.0%

S1 Table. Number of samples of each cancer type in the pan-cancer cohort

ACC	Adrenocortical carcinoma	89	**LUAD**	Lung adenocarcinoma	575
BLCA	Bladder Urothelial Carcinoma	416	**LUSC**	Lung squamous cell carcinoma	327
BRCA	Breast invasive carcinoma	849	**OV**	Ovarian serous cystadenocarcinoma	399
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	308	**PAAD**	Pancreatic adenocarcinoma	188
CHOL	Cholangiocarcinoma	49	**PCPG**	Pheochromocytoma and Paraganglioma	182
COAD	Colon adenocarcinoma	325	**PRAD**	Prostate adenocarcinoma	510
ESCA	Esophageal carcinoma	190	**READ**	Rectum adenocarcinoma	114
GBM	Glioblastoma multiforme	315	**SARC**	Sarcoma	259
HNSC	Head and Neck squamous cell carcinoma	585	**SKCM**	Skin Cutaneous Melanoma	472
KICH	Kidney Chromophobe	66	**STAD**	Stomach adenocarcinoma	485
KIRC	Kidney renal clear cell carcinoma	275	**TGCT**	Testicular Germ Cell Tumors	149
KIRP	Kidney renal papillary cell carcinoma	319	**THCA**	Thyroid carcinoma	529
LAML	Acute Myeloid Leukemia	123	**UCEC**	Uterine Corpus Endometrial Carcinoma	487
LGG	Brain Lower Grade Glioma	516	**UCS**	Uterine Carcinosarcoma	57
LIHC	Liver hepatocellular carcinoma	380	**UVM**	Uveal Melanoma	80

S2 Table. Size and overlap with Gencode exons for the six capture kits used to collect TCGA normal DNA samples.

Capture Kit	Size (MB)	Fraction Overlap With Gencode Exons	Number Samples
Agilent Custom	33	0.993	5793
Nimblegen SQEZ v2	36	0.996	1337
Nimblegen hg18	36	0.992	577
Nimblegen HGSC	37	0.997	1350
Nimblegen SQEZ v3	39	0.999	210
SureSelect 38	64	0.982	171
Intersection	27	0.977	

S3 Table. Variance in LOF indel burden explained by technical covariates each indel filtering approach.

GATK Indel VQSR TS 90.0				
	Sum Sq	**Df**	**F value**	**Pr(>F)**
C20X	1.31E+04	1	323.8629669	4.52E-71
WGA	2.39E+00	1	0.05899371	8.08E-01
Center	1.49E+03	2	18.33944372	1.13E-08
BWA	1.19E+03	5	5.86624688	2.04E-05
Race	5.60E+04	5	276.5583032	1.43E-274
Residuals	3.39E+05	8363		
GATK Indel VQSR TS 95.0				
	Sum Sq	**Df**	**F value**	**Pr(>F)**
C20X	15219.389	1	327.944016	6.31E-72
WGA	1361.419	1	29.335557	6.26E-08
Center	3258.38	2	35.105429	6.57E-16
BWA	1507.617	5	6.497157	4.89E-06
Race	68648.485	5	295.844466	2.06E-292
Residuals	388114.262	8363		
GATK Hardfilter				
	Sum Sq	**Df**	**F value**	**Pr(>F)**
C20X	50615.091	1	419.836036	4.45E-91
WGA	76075.977	1	631.025972	2.60E-134
Center	18981.98	2	78.724735	1.34E-34
BWA	4239.435	5	7.032952	1.44E-06
Race	150187.094	5	249.150811	6.51E-249
Residuals				
GATK Indel VQSR TS 99.0				
	Sum Sq	**Df**	**F value**	**Pr(>F)**
C20X	52930.43	1	153.90042	4.95E-35
WGA	3744887.28	1	10888.62716	0.00E+00
Center	383585.43	2	557.65614	4.53E-228
BWA	169507.9	5	98.57217	2.76E-101
Race	146904.86	5	85.42806	7.59E-88
Residuals	2876257.21	8363		
Varscan				
	Sum Sq	**Df**	**F value**	**Pr(>F)**
CX20	17843.14	1	167.73824	5.34E-38
WGA	4169.66	1	39.19778	4.02E-10
Center	5477.7	2	25.74714	7.12E-12
BWA	178965.82	5	336.48128	0.00E+00
Race	72980.97	5	137.21464	2.80E-140
Residuals	889613.31	8363		