Science Translational Medicine
AAAS

www.sciencetranslationalmedicine.org/cgi/content/full/10/457/eaar7939/DC1

Supplementary Materials for

A machine learning approach for somatic mutation discovery

Derrick E. Wood, James R. White, Andrew Georgiadis, Beth Van Emburgh, Sonya Parpart-Li, Jason Mitchell, Valsamo Anagnostou, Noushin Niknafs, Rachel Karchin, Eniko Papp, Christine McCord, Peter LoVerso, David Riley, Luis A. Diaz Jr., Siân Jones, Mark Sausen, Victor E. Velculescu*, Samuel V. Angiuoli*

*Corresponding author. Email: velculescu@jhmi.edu (V.E.V.); angiuoli@personalgenome.com (S.V.A.)

Published 5 September 2018, *Sci. Transl. Med.* **10**, eaar7939 (2018)
DOI: 10.1126/scitranslmed.aar7939

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at www.sciencetranslationalmedicine.org/cgi/content/full/10/457/eaar7939/DC1)

Table S8 (Microsoft Excel format). Performance results for cell lines with validated somatic mutations.
Table S9 (Microsoft Excel format). Unique and shared mutation load results for Cerebro and TCGA (MC3).
Table S10 (Microsoft Excel format). Concordance results for Cerebro and TCGA for driver oncogenes and tumor suppressor genes.
Table S11 (Microsoft Excel format). Comparison of mutation calls in immunotherapy publications and Cerebro reanalysis.
Table S12 (Microsoft Excel format). Evaluation of TruSeq false-positive calls using raw CS125 and Oncomine sequence data.
Table S13 (Microsoft Excel format). Comparison of mutation callers for clinical samples.
Table S14 (Microsoft Excel format). Shared and unique somatic mutation calls between Cerebro and TCGA.
Table S15 (Microsoft Excel format). Genomic ROIs used in clinical panel comparisons.

Supplementary Text
Materials and Methods

BWA MEM Alignment Command

```
bwa mem -t <total threads> <reference index> <R1 read pairs> <R2 read pairs>
| samtools view -@ <total threads> -Su - | samtools sort -@ <total threads> -
<output bam file>
```

Variant Caller Commands

MuTect1

```
java -Xmx24g -Djava.io.tmpdir=<temporary directory> -jar muTect-1.1.4.jar
--analysis_type MuTect
--reference_sequence <human reference>
--input_file:tumor <tumor bam>
--input_file:normal <normal bam>
--out <output file>
--coverage_file <output coverage file>
--vcf <output vcf file>
--num_threads <total threads>
--intervals <ROI bed file>
--enable_extended_output
```

MuTect2

```
java -Xmx32g -jar GenomeAnalysisTK.jar
-T MuTect2
-R <human reference>
-I:tumor <tumor bam>
-I:normal <normal bam>
-o <output vcf file>
-L <ROI bed file>
```

SomaticSniper

```
bam-somaticsniper -F vcf -f <human reference> <tumor bam> <normal bam>
<output vcf file>
```

Strelka

```
configureStrelkaWorkflow.pl --tumor=<tumor bam> --normal=<normal bam> --
ref=<human reference> --config=strelka_demo_config.ini --output-dir=<output
directory>
```

VarDict

```
vardict -G <human reference> -t -f 0.10 -N tumor_sample_name -b "<tumor bam>
|<normal bam>" -c 1 -S 2 -E 3 -g 4 <ROI bed file>  > <output file>
```

VarScan

```
java -jar VarScan.v2.4.2.jar somatic <normal mpileup file> <tumor mpileup
file>
--min-coverage 30
--min-var-freq 0.10
--min-freq-for-hom 0.75
--output-snp <outsnp snp file>
--output-indel <output indel file>
--normal-purity 1.00
--tumor-purity  0.50
```

VarDict Filters

According to its documentation (https://github.com/AstraZeneca-NGS/VarDict), VarDict has a "philosophy of 'calling everything'", which means that if one were to leave its results completely unfiltered, it would have an incredibly low PPV. The documentation points to a short article (http://bcb.io/2016/04/04/vardict-filtering/) written by Brad Chapman, one of the VarDict authors, that provides guidance on how to create well-performing filters for VarDict. That article describes two filters derived from manual examination of the DREAM Challenge data and mixtures of Genome in a Bottle sequences. These filters focus on low allele frequency variants that exhibit either low coverage or low "quality" (a metric combining base quality and mutant coverage).

Contribution of Cerebro and Dual Alignment to Mutation Filtering

To evaluate the contributions of dual alignment vs. the Cerebro random forest model, we examined a normal cell line that we had sequenced twice using NGS, and treated one sequencing run as a tumor and one as its matched normal before running them through our Cerebro pipeline. The pipeline's output included zero somatic mutations (as was expected). We then examined the internal output of the pipeline to determine how many candidate somatic coding mutations existed in the two aligners' datasets. In the BWA-MEM data, 303/616 (49.2%) candidate mutations were removed through the dual alignment concordance requirement; in the Bowtie2 data, 351/664 (52.8%) candidate mutations were removed by the concordance requirement. 313 candidate mutations were found in both aligners' output, and all received Cerebro scores lower than the 0.75 minimum required to be reported. From this analysis we conclude that in whole exome NGS data, dual alignment removes approximately half of the artifactual candidate somatic mutations that would be found by a single aligner, but hundreds of other artifacts will still require removal by filtering methods such as our Cerebro model.

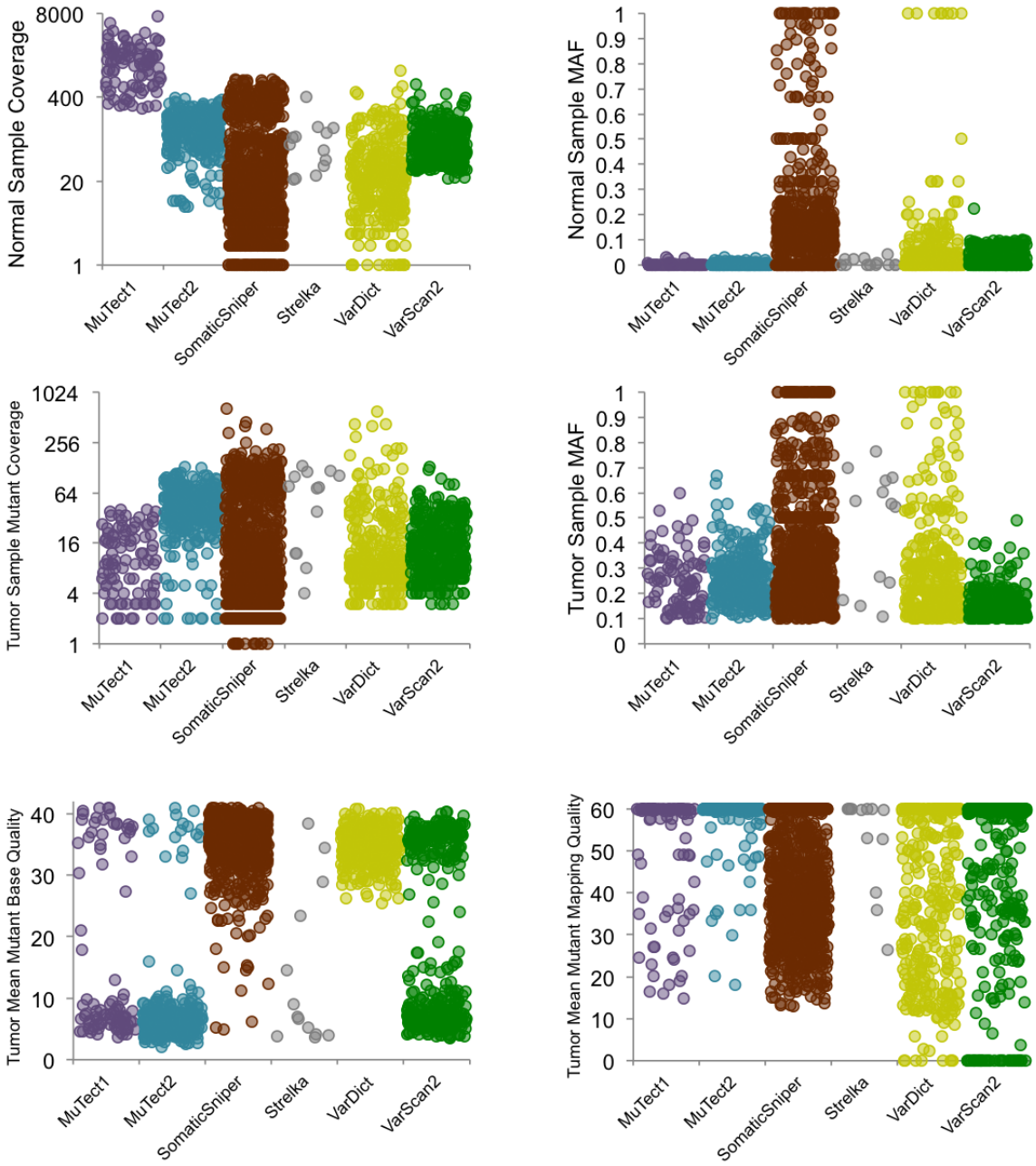

Fig. S1. False-positive evaluation for somatic mutation callers. Using results from simulated mutations in exomes, we display factors frequently associated with false positive mutation calls including tumor/normal coverage, tumor/normal mutant allele frequency (MAF), mean mutant base quality and mean mutant mapping quality. Each point represents a false positive mutation called by a specific program.

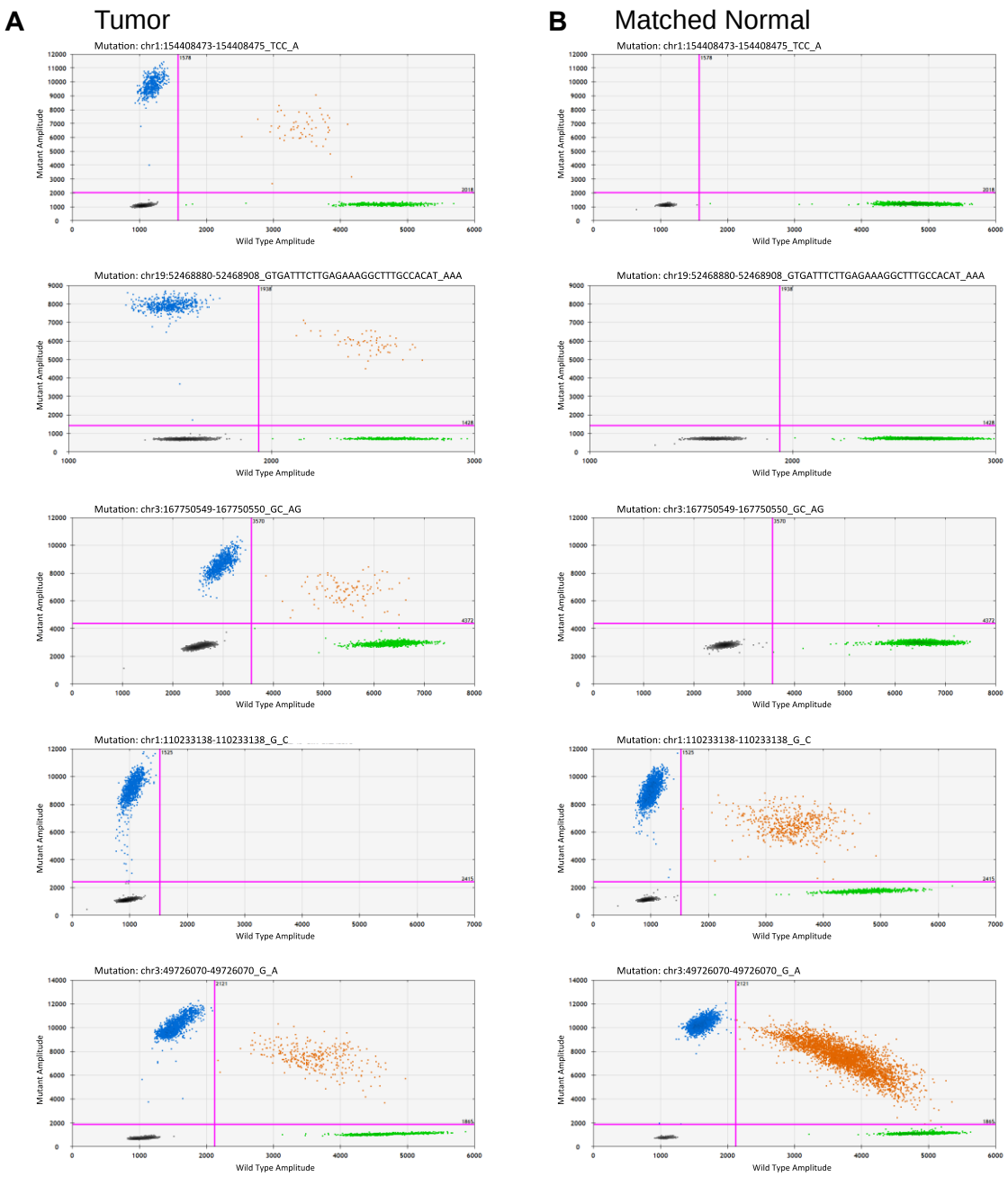

Fig. S2. ddPCR mutation validation analyses. Each panel is a unique sample and each point on the graphs represents the fluorescent signal from PCR amplified fragments in a droplet. Green droplets have been fluorescently tagged with a reference probe (HEX), blue droplets have been tagged with a mutant probe (FAM), and orange droplets have been tagged with both mutant and reference probes. Rows correspond to a variant where both the tumor (A) and matched normal sample (B) were tested. The top three rows indicate somatic mutations detected by Cerebro and validated by ddPCR. In each case, the tumor shows amplified mutant copies (blue) while the normal only shows amplified reference copies (green). For the bottom two rows, mutant copies (blue) were amplified in both the tumor and matched normal sample indicating that these variants were not tumor specific.

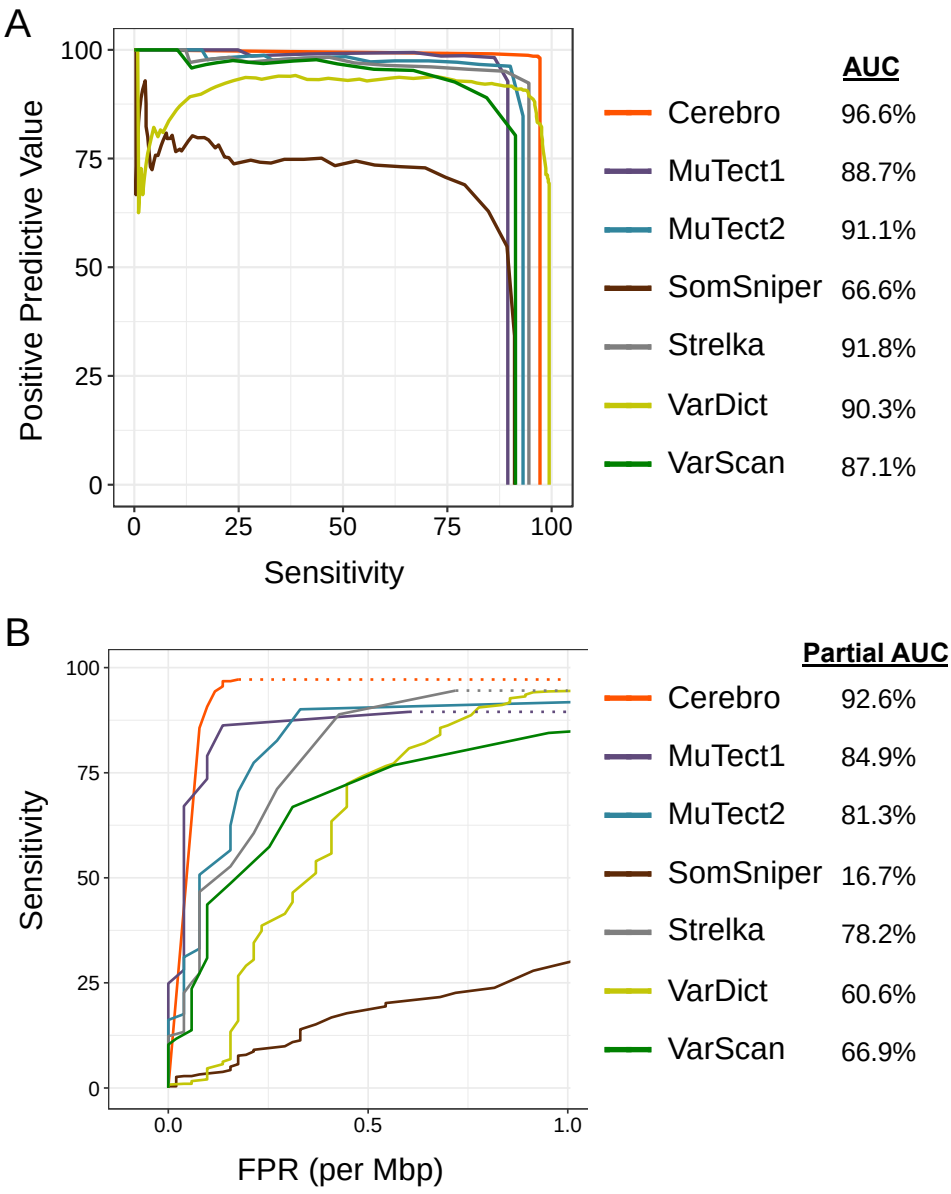

Fig. S3. Precision-recall and ROC curve analyses of Cerebro and other mutation callers using experimentally validated alterations. We observe that Cerebro outperforms other methods using both precision-recall (A, sensitivity vs. positive predictive value) and ROC analyses (B, false positive rate (per Mbp) vs. sensitivity).

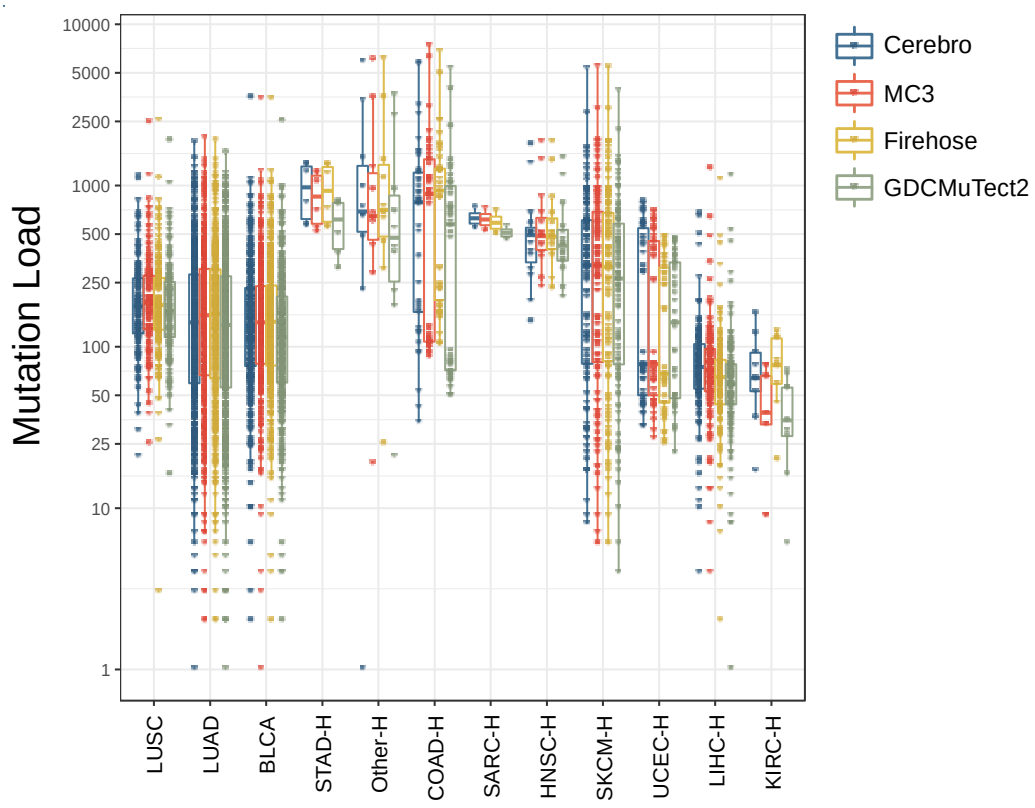

Fig. S4. Mutation loads of TCGA exomes using different mutation calling methods. Call sets include Cerebro, MC3 (PanCanAtlas), Broad Firehose, and MuTect2 (Genomic Data Commons). LUSC=lung squamous cell carcinoma; LUAD=lung adenocarcinoma; BLCA=bladder; STAD=stomach; COAD=colorectal; SARC=sarcoma; HNSC=head and neck squamous cell; SKCM=melanoma; UCEC=uterine; LIHC=liver; KIRC=kidney; -H indicates high mutation load enriched tumor sets.

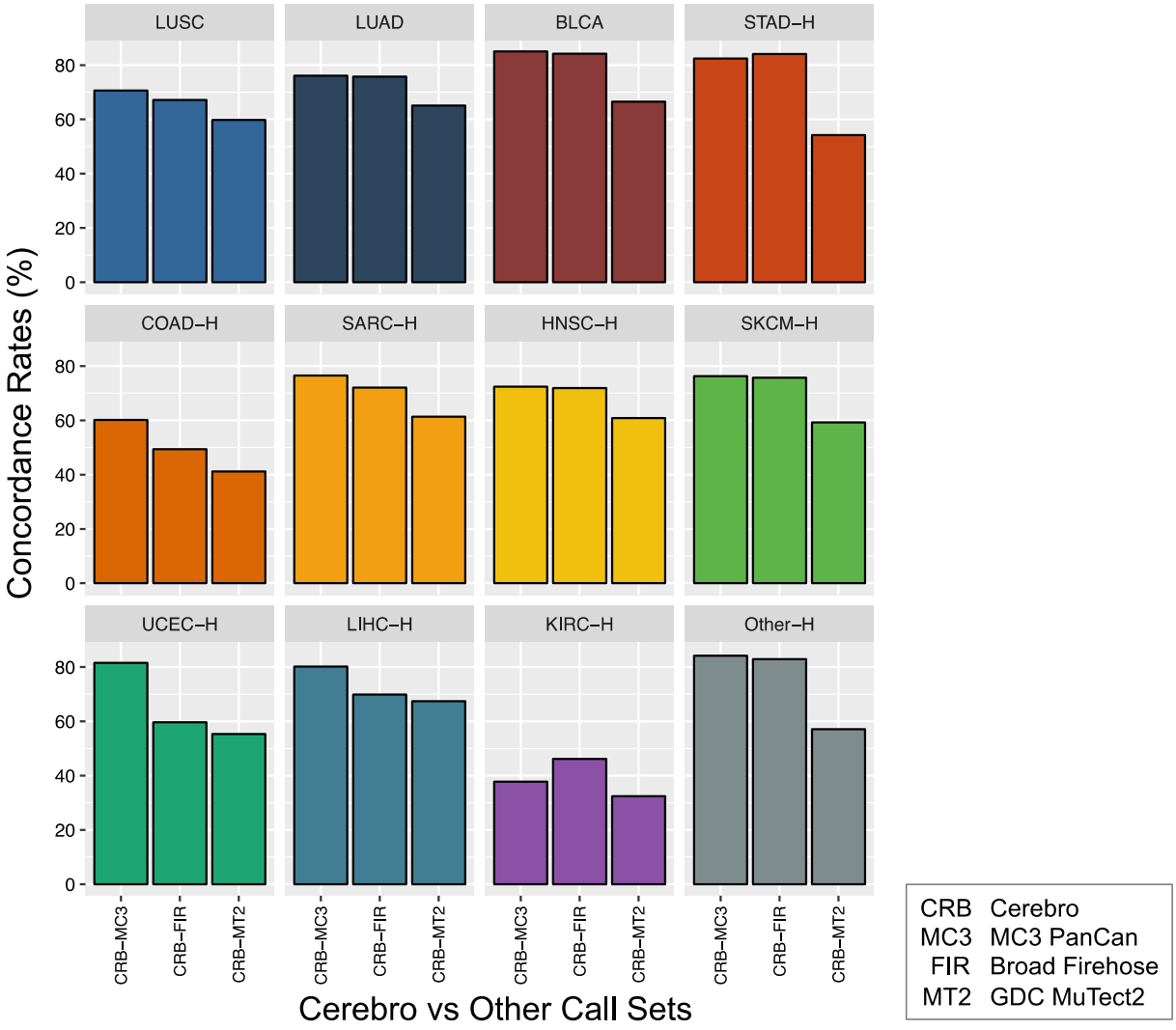

Fig. S5. Concordance rates (% of total mutations) of Cerebro compared to other mutation call sets for TCGA exomes. LUSC=lung squamous cell carcinoma; LUAD=lung adenocarcinoma; BLCA=bladder; STAD=stomach; COAD=colorectal; SARC=sarcoma; HNSC=head and neck squamous cell; SKCM=melanoma; UCEC=uterine; LIHC=liver; KIRC=kidney; -H indicates high mutation load enriched tumor sets. Overall, MC3 had the strongest concordance with the Cerebro call set.

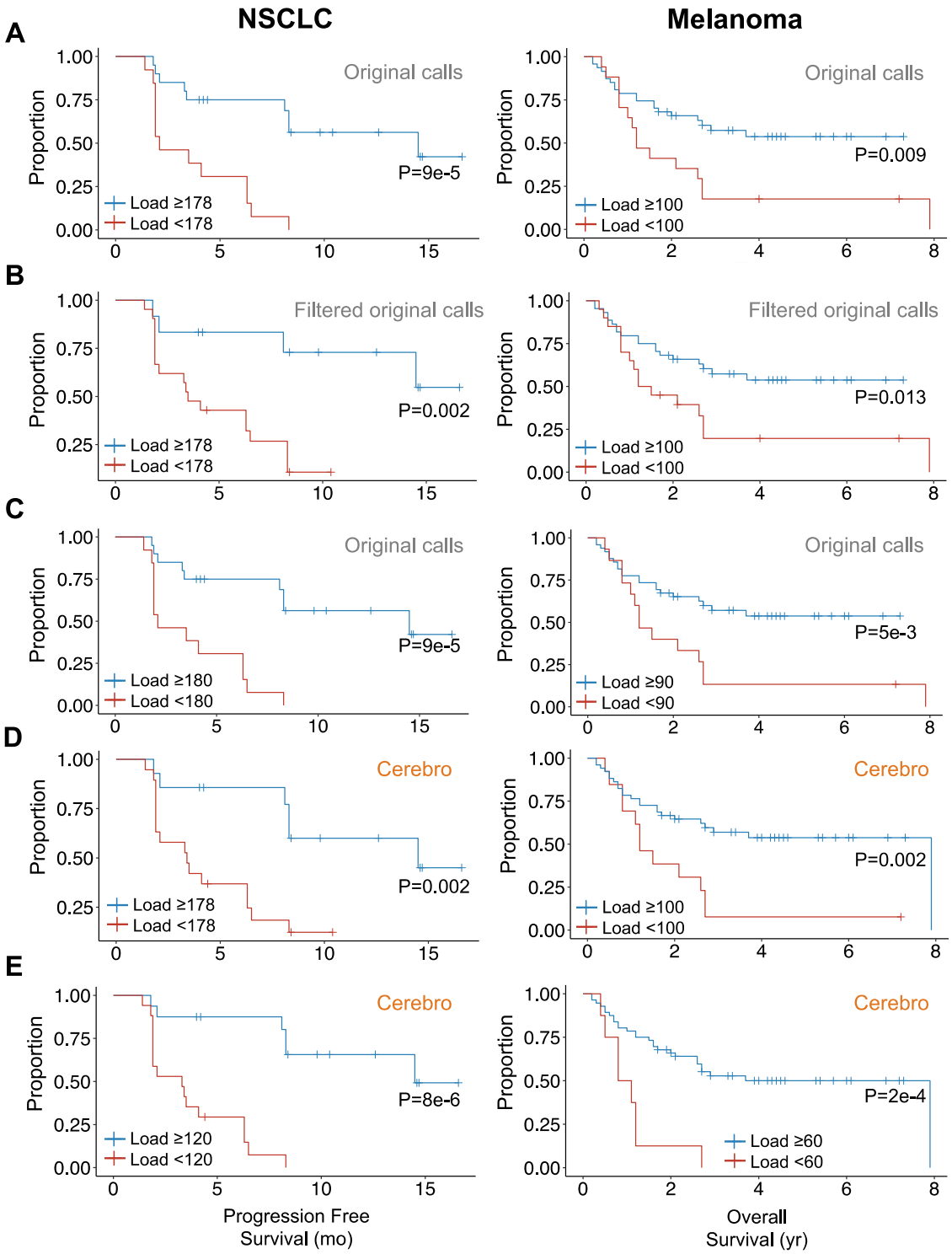

Fig. S6. Response to checkpoint inhibitors associated with mutational load. Comparison of Cerebro mutation calls with published calls associated with NSCLC (left panels) or melanoma (right panels). (**A**) Kaplan-Meier analysis of progression free survival (left) or overall survival (right) using tumor mutation loads from original publications; (**B**) original publication results filtered for problematic mutations (Figure 5B); **(C)** original publication results using optimal thresholds for survival prediction; (**D**) Cerebro calls

using same threshold as original publications; (**E**) Cerebro calls using optimal thresholds for survival prediction. Log-rank P-value shown for each survival plot.

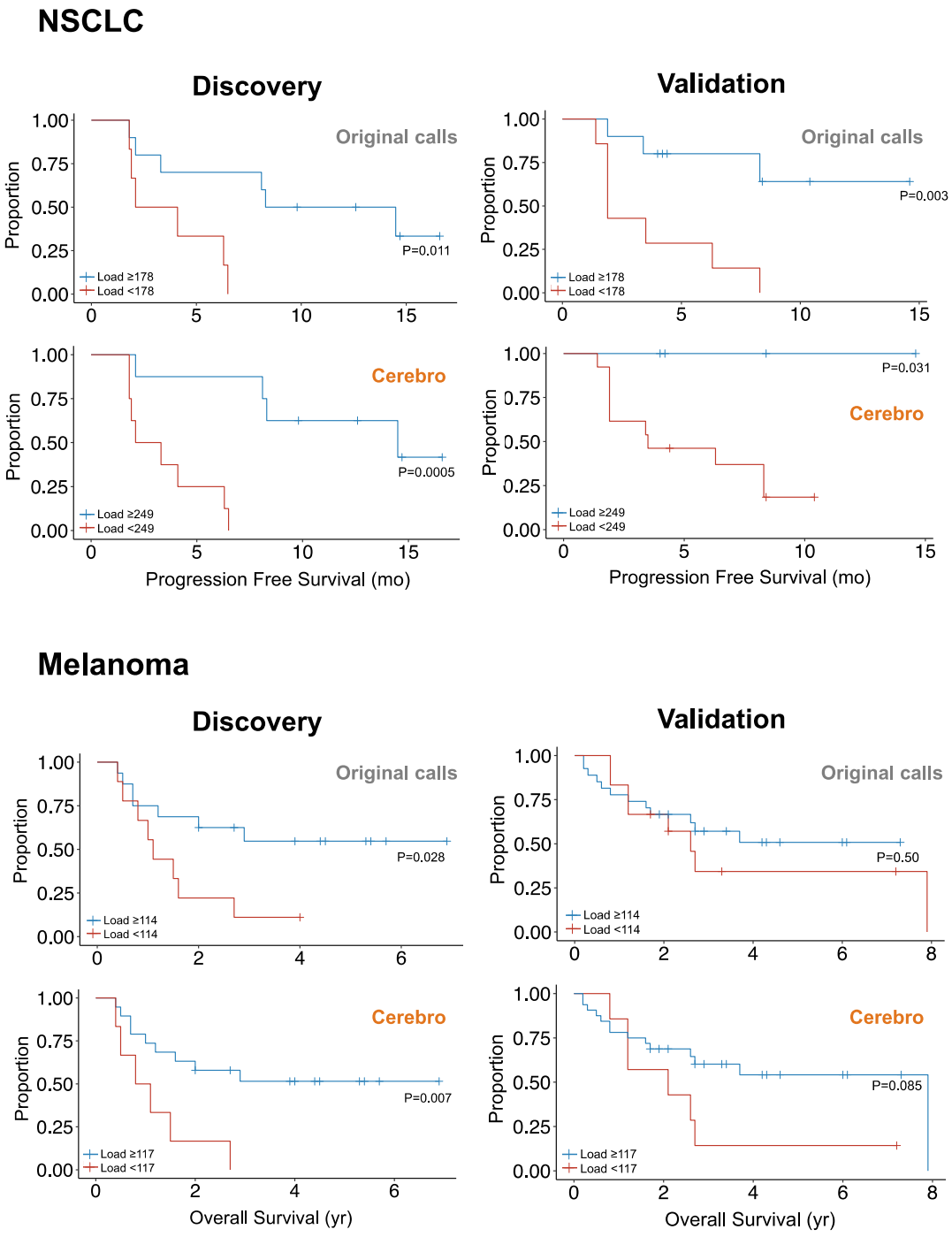

Fig. S7. Survival analysis by mutation load stratified by discovery and validation cohorts. Comparison of Cerebro mutation calls with published calls associated with NSCLC (top panels) or melanoma (bottom panels), stratified by discovery and validation cohort membership. In this analysis, we optimize the mutation load threshold using the same selection criteria as Rizvi et al. (NSCLC). Specifically, we select the mutation load value that maximizes sensitivity to detect a responder with the highest possible specificity value in the discovery cohort.

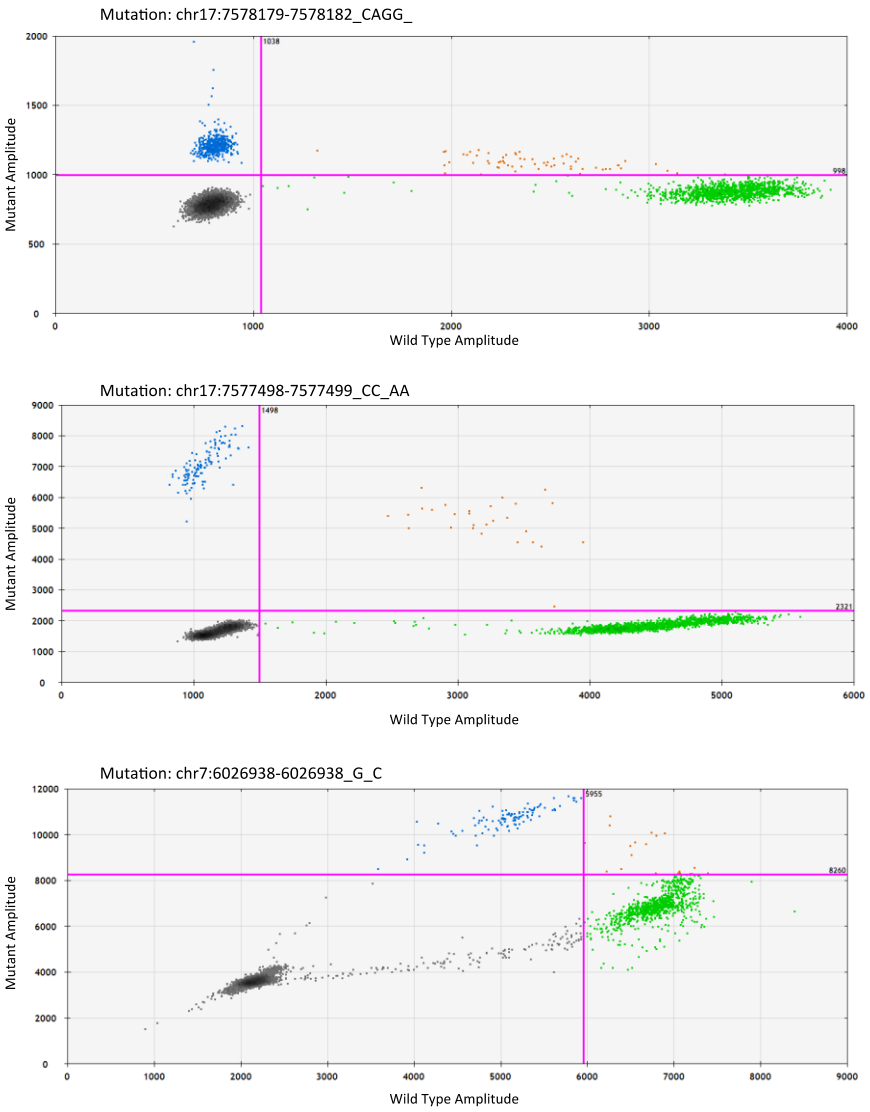

Fig. S8. Alterations confirmed by ddPCR in clinical NGS panel comparison. Each graph shows a unique alteration examined in a tumor sample with each point representing the fluorescent signal from PCR amplified fragments in a droplet. Green droplets have been fluorescently tagged with a reference probe (HEX), blue droplets have been tagged with a mutant probe (FAM), and orange droplets have been tagged with both mutant and reference probes. All three samples confirmed the presence of the mutant sequences identified by Cerebro.

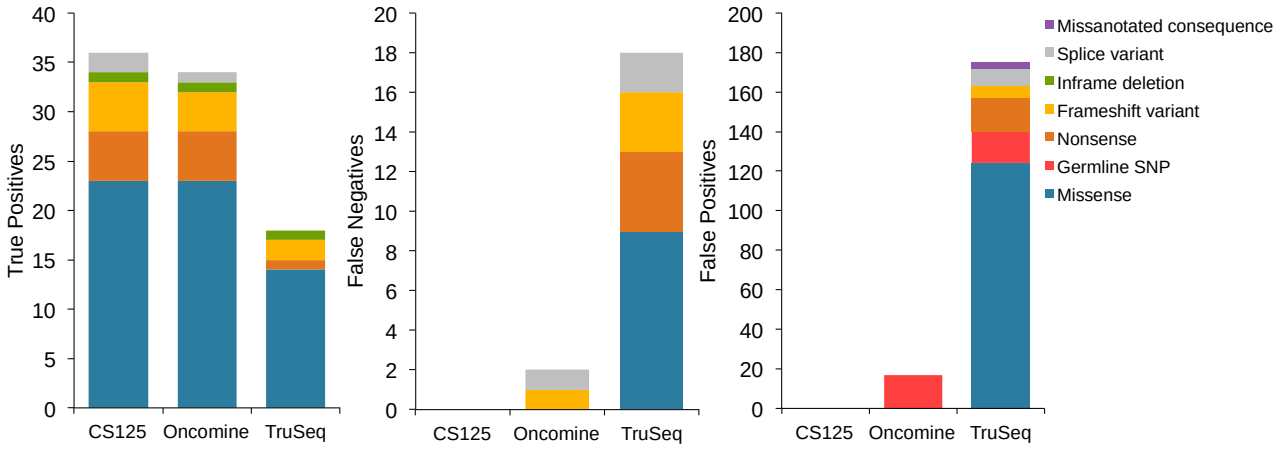

Fig. S9. Comparative results of three clinical sequencing panels. Reported results include true positives, false negatives and false positives. The shades in each category indicate the type of alteration observed.