Supplementary Figure 18

a Alignment & Pre-processing: Adapters Trimming, BWA-MEM, Sambamba, GATK v4. Mutational Calling: Mutect v2, Platypus, Strelka, Consensus Calls. Functional Assessment: Coding (CGI), Splicing (S-CAP), Chromatin (Sasquatch & DeepSEA), TFBS (deltaSVM). Copy Number Alterations: CNVkit.

b Sequencing Statistics: %On Target, %Unique, Number SNV (***), % 30X, % 100X, Purity.

c metastatic_site

d R2 = 0.9; 95% CI [0.87 − 0.92]; p = 1.6e−289

e P(SCC=−0.32; p=0.0013) M(SCC=−0.25; p=0.013)

f P(SCC=0.26; p=0.0098) M(SCC=0.021; p=0.84)
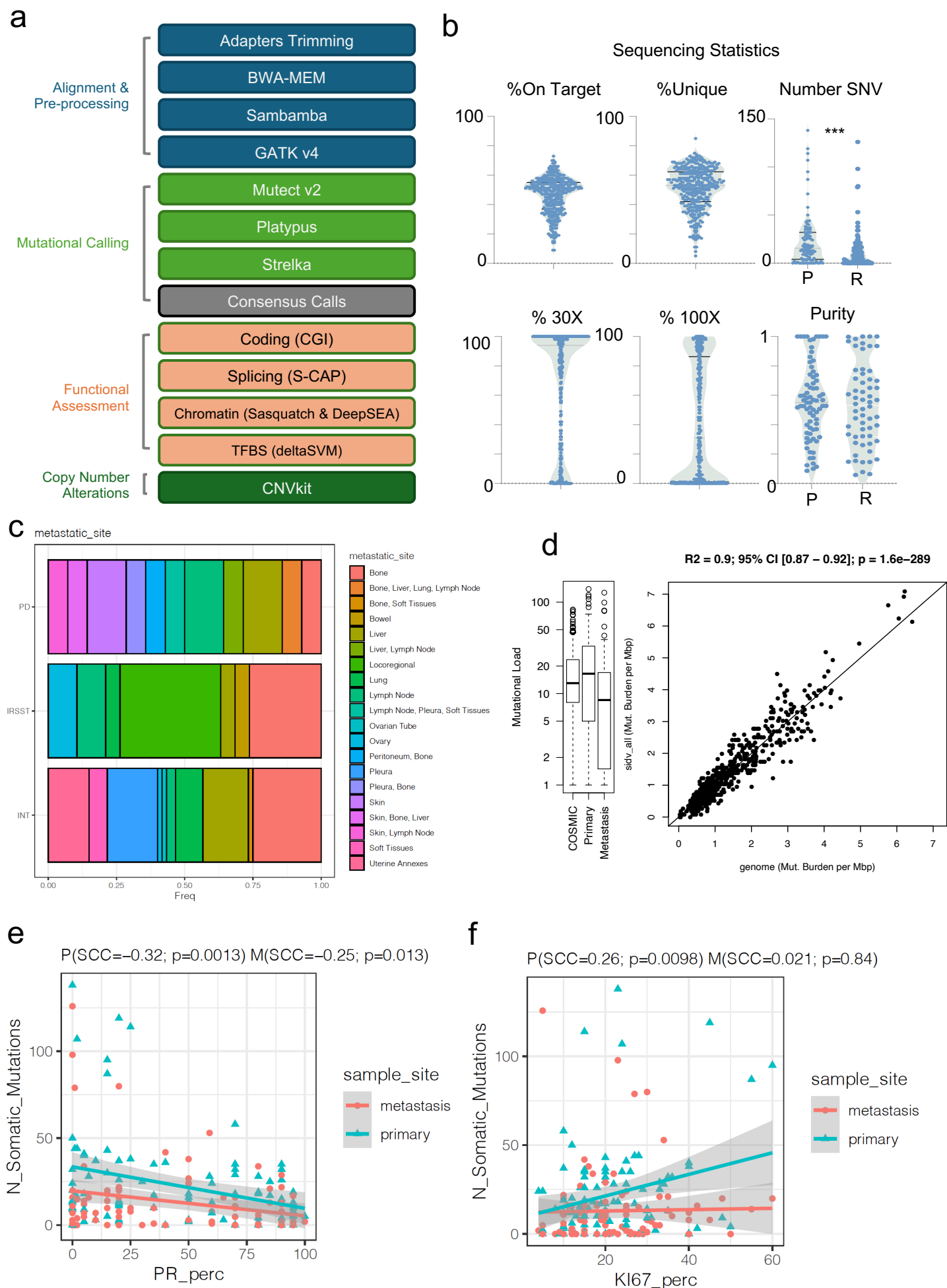
Supplementary Figure 18. SIDV design and results. (a) Schematic summarising the steps of the custom computational pipeline employed for the identification and functional annotation of the SIDV variants. (b) Summary of the sequencing statistics for the profiled samples. (c) Stacked bar plots showing the anatomic site of the profiled relapse, split by center. (d) Box plot showing the distribution of the overall mutational load per sample in the SID regions, as estimated in either the non-coding COSMIC, or in our cohort (separately for primary and metastatic samples). The companion scatterplot shows the correlation beetween the genome-wide estitate of mutational burden and the same estimate using only the mutations identified in SID regions, considering the WGS data available in the non-coding COSMIC. The statistics and statistical significance of this linear correlation is indicated on top of the plot. (e-f) Scatterplots showing the relationship between the number of somatic mutations detected per sample, and the indicated variables. For visualization purpose, least square regression models were trained separately for primary and metastatic samples. For quantifying the relationships, Spearman's correlation coefficients (SCC) are indicated on top of the plots, along with the corresponding p-values.