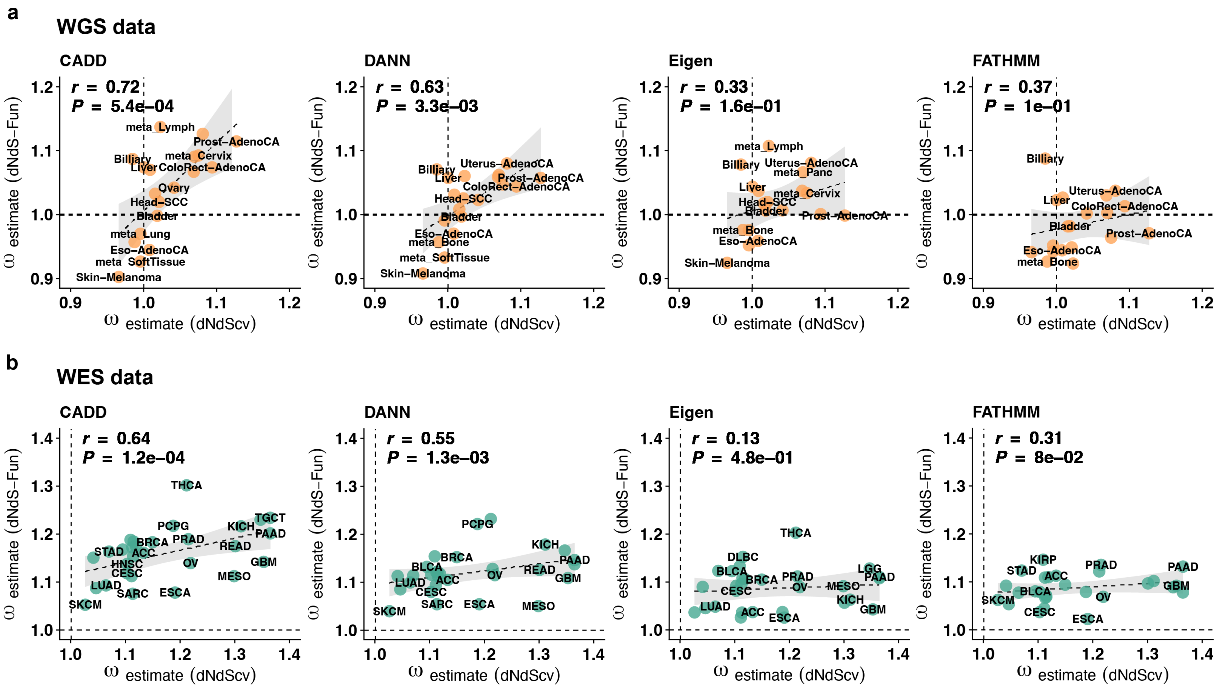
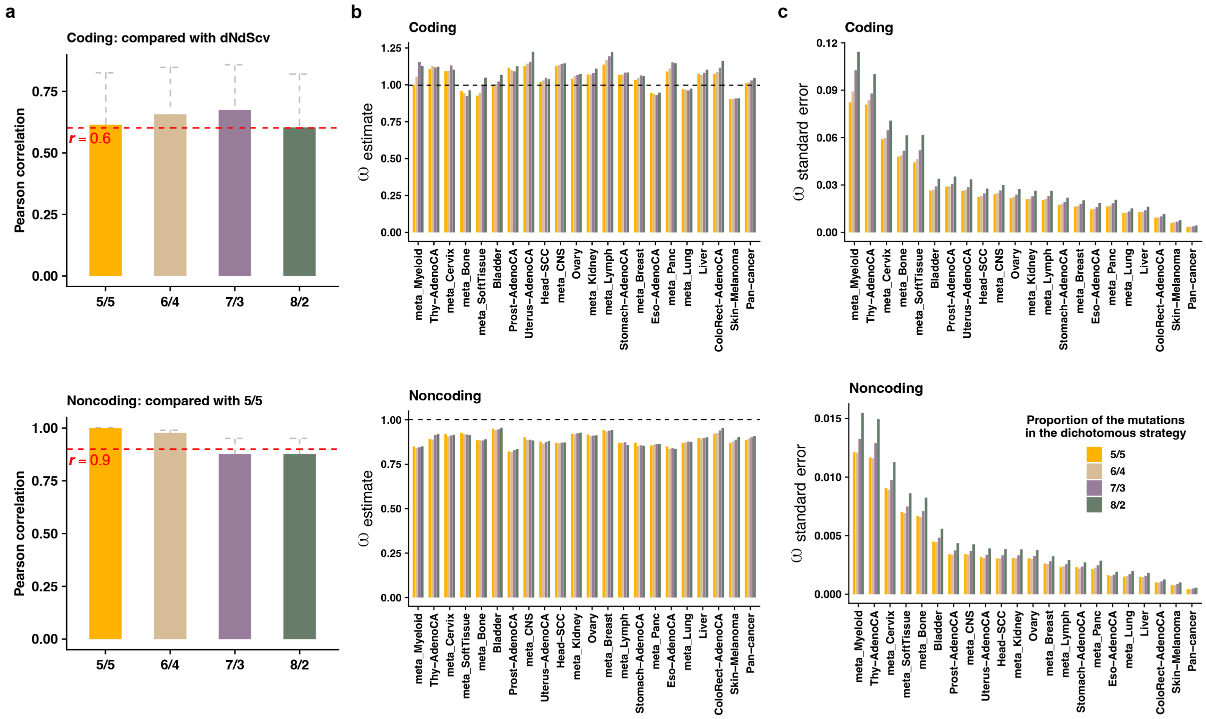
Characterizing selection signatures in coding and noncoding regions of 14,886 cancer genomes

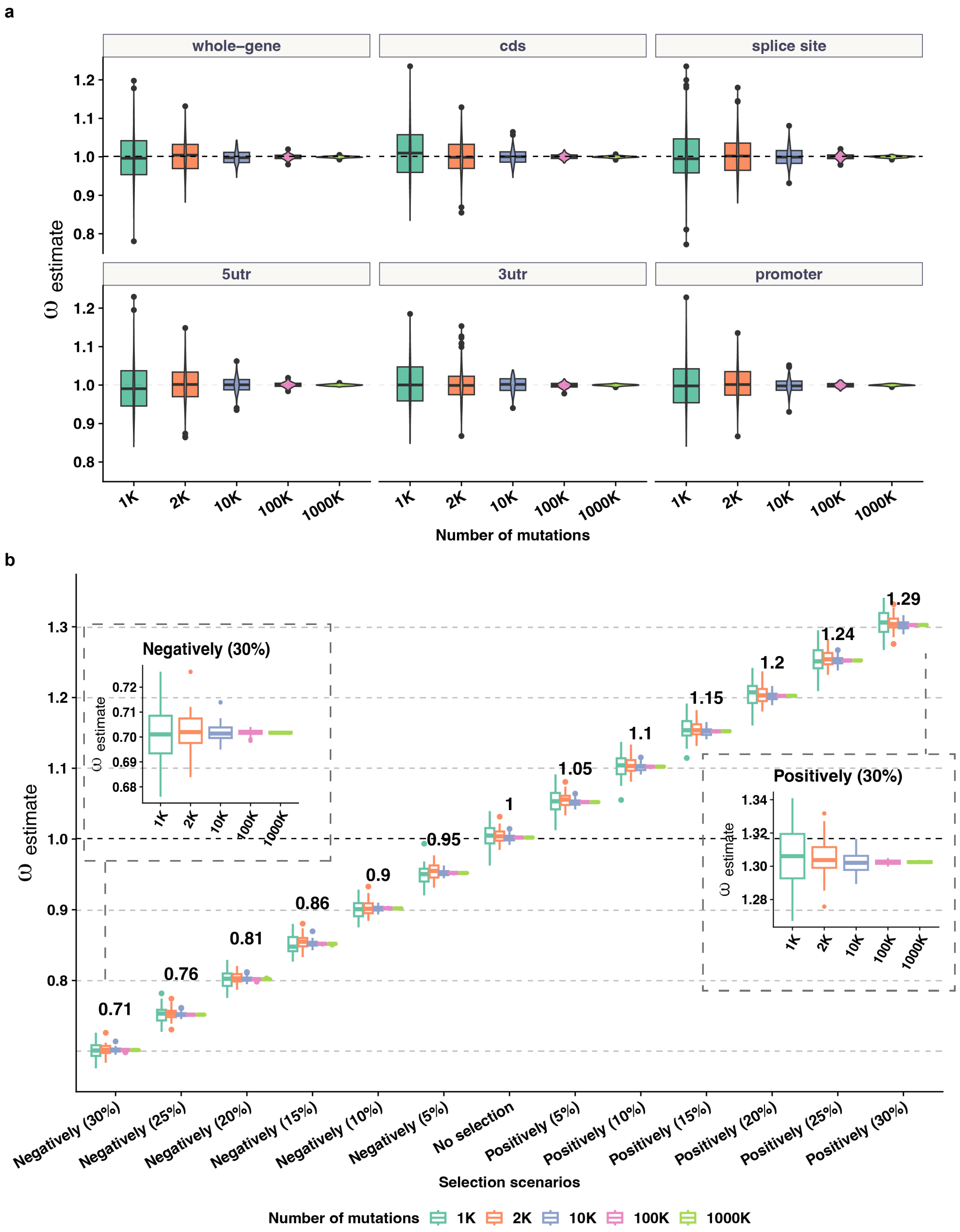
Zheng *et al*.



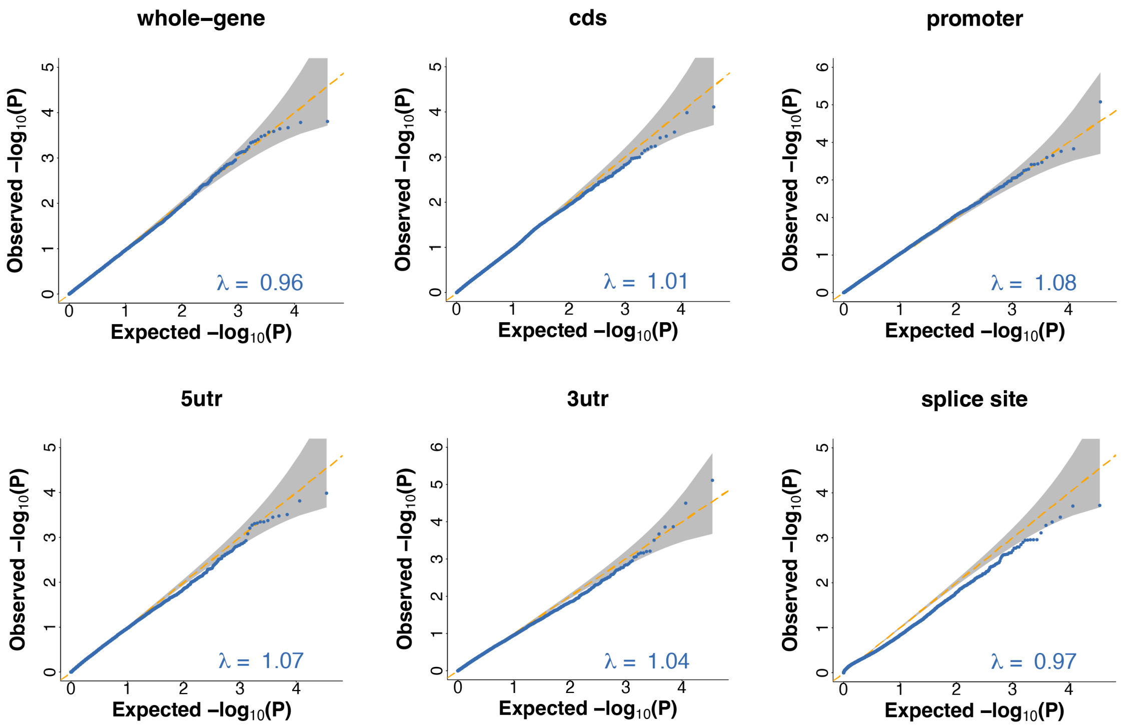
Supplementary Figure 1. Comparison of *ω* estimates between dNdS-Fun and dNdScv. The plot illustrates the correlation between *ω* estimates obtained from dNdS-Fun and dNdScv, using only coding mutations from the PCAWG dataset (panel a) and the TCGA dataset (panel b). The dNdS-Fun models are constructed using CADD, DANN, Eigen, and FATHMM scores and are displayed sequentially from left to right. The Pearson correlation coefficient (r) and *P*-value are shown in the top left corner of each plot. The grey shaded area represents the 95% CI.



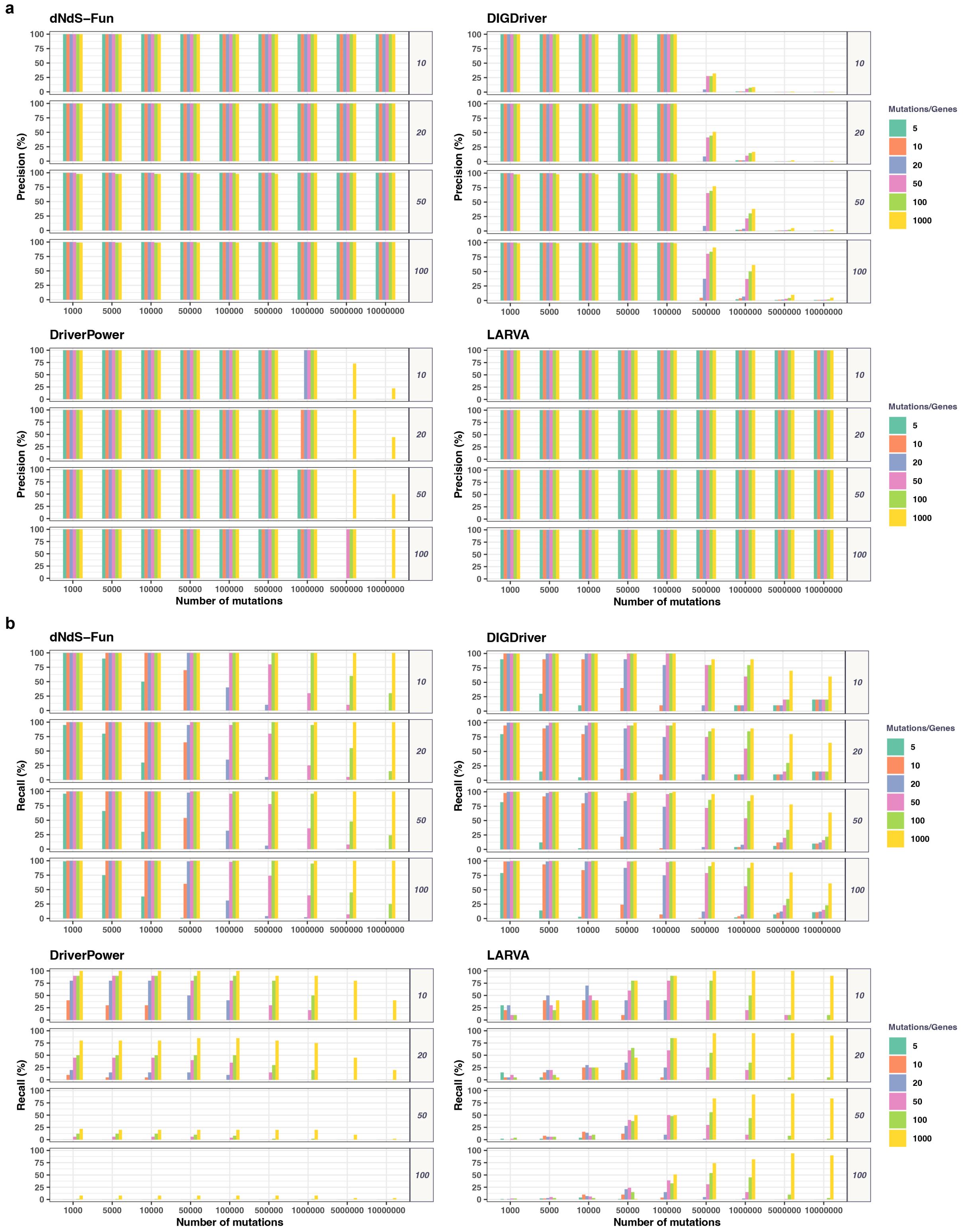
Supplementary Figure 2. Estimation of *ω* under different dichotomization thresholds. a) Comparison of *ω* estimates between dNdS-Fun with different dichotomization thresholds (i.e., 5:5, 6:4, 7:3, and 8:2) and dNdScv in coding regions (top panel), and between dNdS-Fun with a threshold of 5:5 and dNdS-Fun with other thresholds in noncoding regions (bottom panel). The dichotomization thresholds—5:5, 6:4, 7:3, and 8:2—represent the ratios of less functional to more functional genomic sites. b) Estimates of *ω* across different cancer types using dNdS-Fun with different thresholds. c) Standard errors of the *ω* estimate across various cancer types using dNdS-Fun with different thresholds.



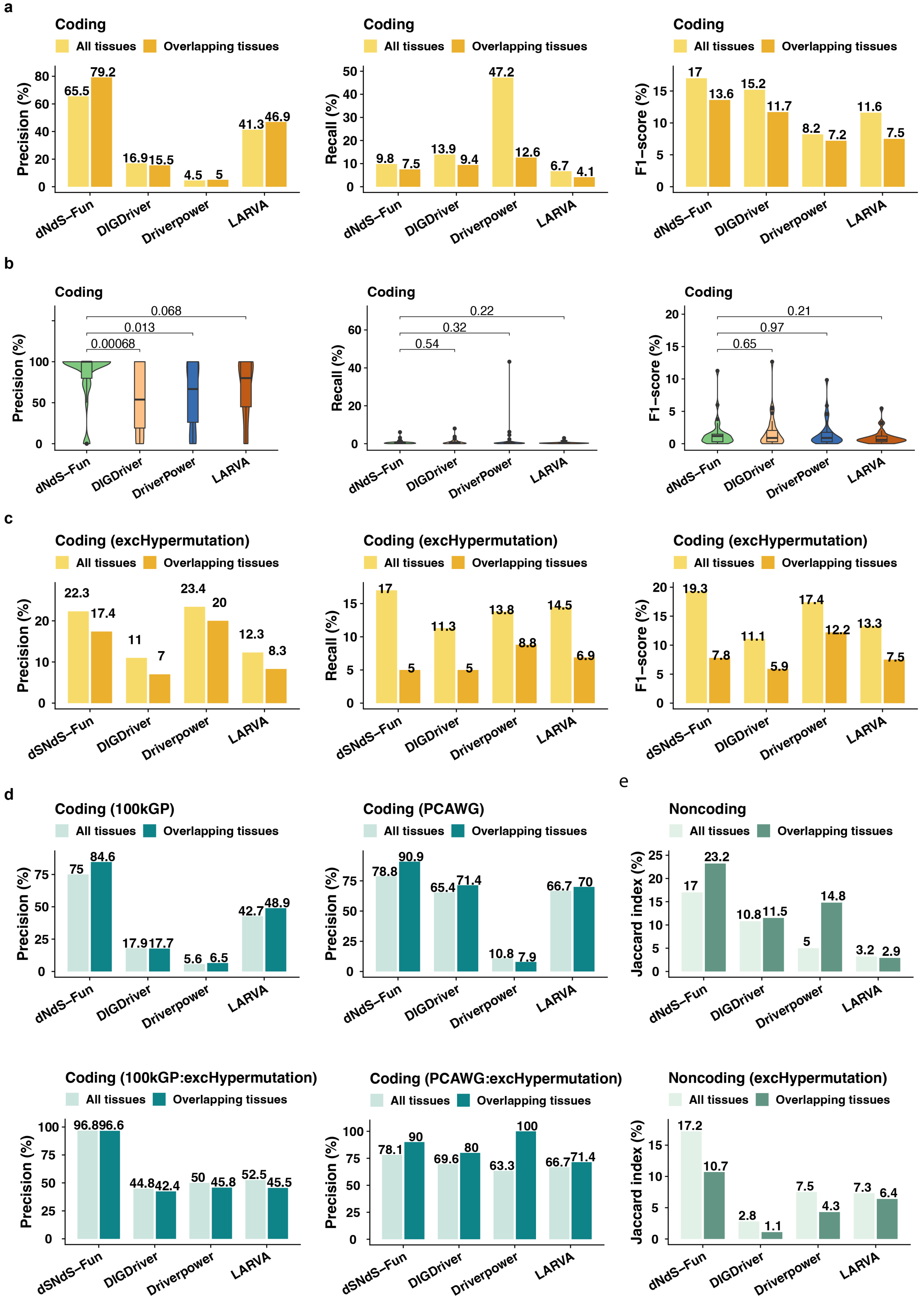
Supplementary Figure 3. Simulations under the null and selection models. a) Estimates of ω from dNdS-Fun using data simulated under the null model. Each box plot represents the results from 300 simulation replicates, categorized by the number of mutations. b) Estimates of ω from dNdS-Fun using data simulated under various selection models. Each boxplot shows the results from 50 simulation replicates. The median estimate of ω for each scenario is displayed above the corresponding box plots, with values in the parentheses indicating the proportion of mutations under selection.



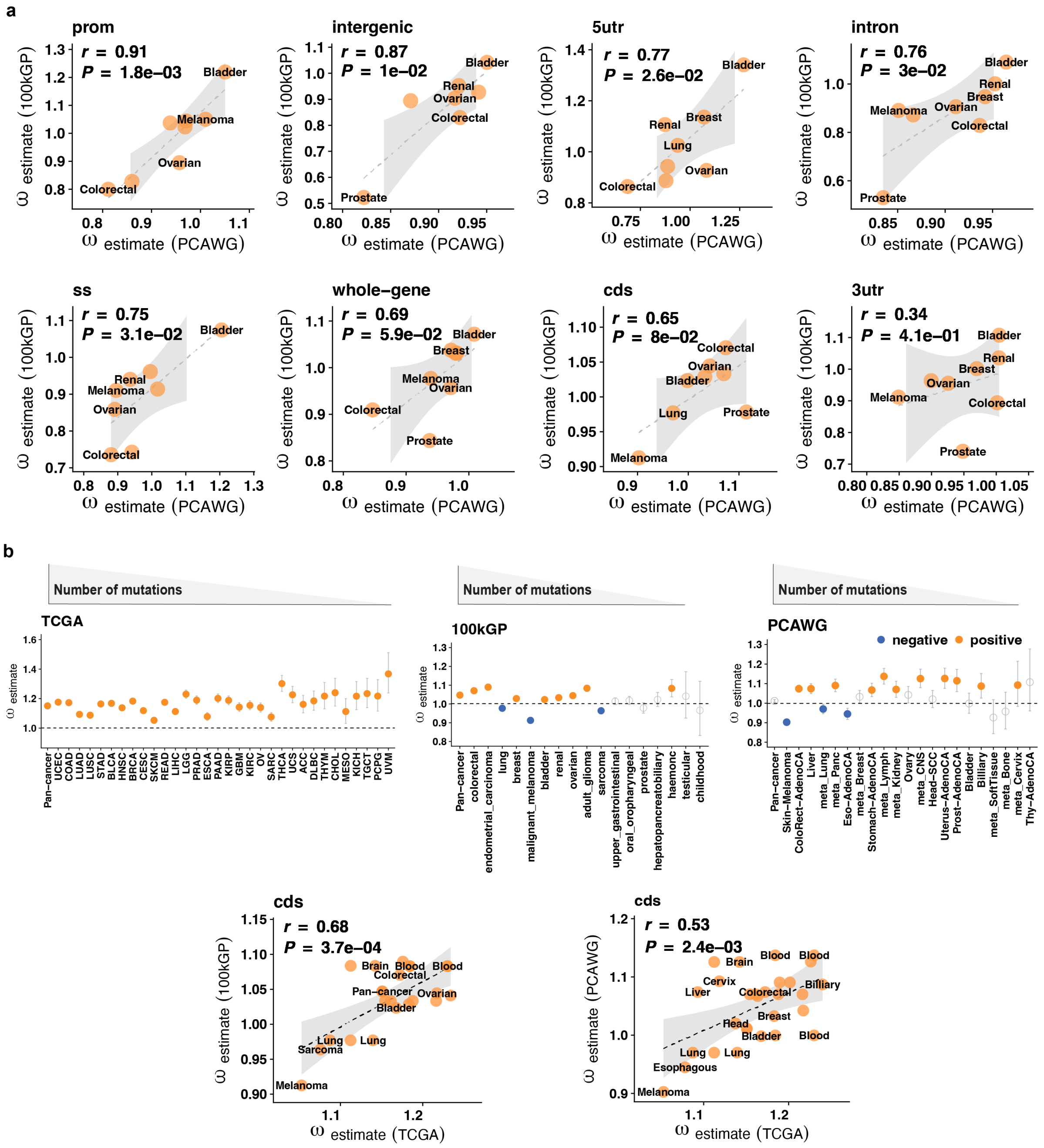
Supplementary Figure 4. Quantile-Quantile (QQ) plots of the dNdS-Fun test-statistics across genes or genomic elements in simulations under the null model. The QQ plots compared the distribution of observed *P*-values for individual genes or genomic elements against those expected under the null hypothesis. Each dot represents an individual gene or genomic element. The grey shaded area denotes the 95% CI. *λ* represents the median of the observed test statistics divided by its expected value under the null hypothesis.



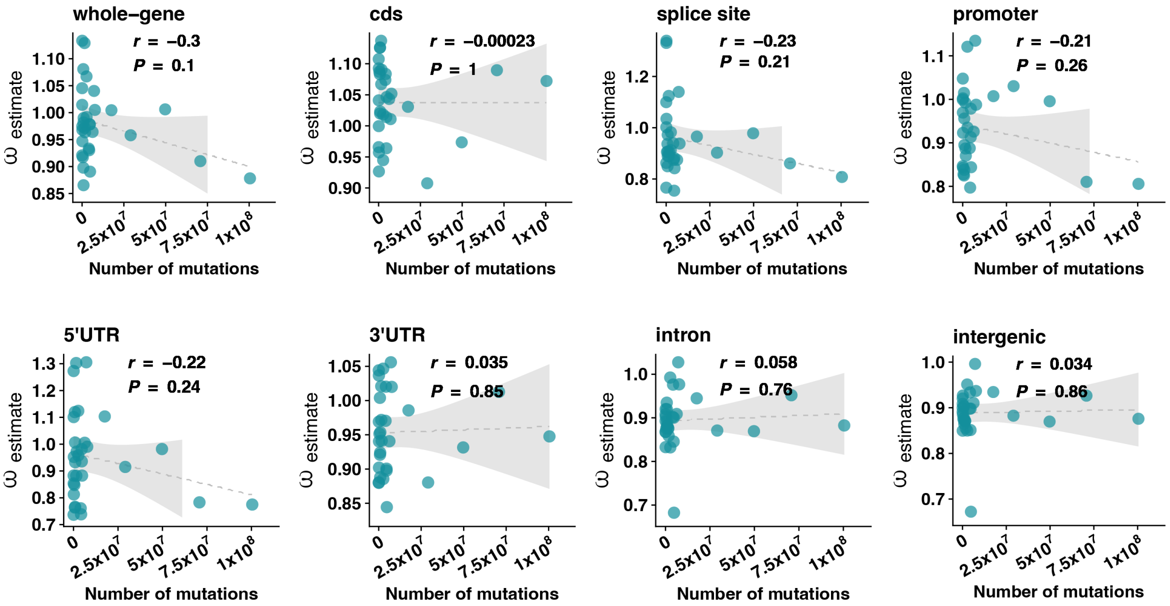
Supplementary Figure 5. Comparison of the performance of dNdS-Fun with DIGDriver, DriverPower, and LARVA using simulated data. The simulations varied in the number of driver genes, driver mutations per gene, and total number of mutations to assess each method’s performance. Panel a displays the precision of each method, while panel b presents the recall, with both metrics plotted against the total number of mutations. The simulated driver genes serve as the gold standard for evaluating performance.


a
Coding
All tissues
Overlapping tissues
Precision (%)
0
20
40
60
80
65.5
79.2
16.9
15.5
4.5
5
41.3
46.9
dNdS–Fun
DIGDriver
Driverpower
LARVA
Coding
All tissues
Overlapping tissues
Recall (%)
0
10
20
30
40
50
9.8
7.5
13.9
9.4
47.2
12.6
6.7
4.1
dNdS–Fun
DIGDriver
Driverpower
LARVA
Coding
All tissues
Overlapping tissues
F1–score (%)
0
5
10
15
17
13.6
15.2
11.7
8.2
7.2
11.6
7.5
dNdS–Fun
DIGDriver
Driverpower
LARVA
b
Coding
Precision (%)
0
50
100
0.068
0.013
0.00068
dNdS–Fun
DIGDriver
DriverPower
LARVA
Coding
Recall (%)
0
20
40
60
0.22
0.32
0.54
dNdS–Fun
DIGDriver
DriverPower
LARVA
Coding
F1–score (%)
0
5
10
15
20
0.21
0.97
0.65
dNdS–Fun
DIGDriver
DriverPower
LARVA
c
Coding (excHypermutation)
All tissues
Overlapping tissues
Precision (%)
0
10
20
22.3
17.4
11
7
23.4
20
12.3
8.3
dSNdS–Fun
DIGDriver
Driverpower
LARVA
Coding (excHypermutation)
All tissues
Overlapping tissues
Recall (%)
0
5
10
15
17
5
11.3
5
13.8
8.8
14.5
6.9
dSNdS–Fun
DIGDriver
Driverpower
LARVA
Coding (excHypermutation)
All tissues
Overlapping tissues
F1–score (%)
0
5
10
15
20
19.3
7.8
11.1
5.9
17.4
12.2
13.3
7.5
dSNdS–Fun
DIGDriver
Driverpower
LARVA
d
Coding (100kGP)
All tissues
Overlapping tissues
Precision (%)
0
25
50
75
75
84.6
17.9
17.7
5.6
6.5
42.7
48.9
dNdS–Fun
DIGDriver
Driverpower
LARVA
Coding (PCAWG)
All tissues
Overlapping tissues
Precision (%)
0
25
50
75
78.8
90.9
65.4
71.4
10.8
7.9
66.7
70
dNdS–Fun
DIGDriver
Driverpower
LARVA
e
Noncoding
All tissues
Overlapping tissues
Jaccard index (%)
0
5
10
15
20
25
17
23.2
10.8
11.5
5
14.8
3.2
2.9
dNdS–Fun
DIGDriver
Driverpower
LARVA
Coding (100kGP:excHypermutation)
All tissues
Overlapping tissues
Precision (%)
0
25
50
75
100
96.8
96.6
44.8
42.4
50
45.8
52.5
45.5
dSNdS–Fun
DIGDriver
Driverpower
LARVA
Coding (PCAWG:excHypermutation)
All tissues
Overlapping tissues
Precision (%)
0
25
50
75
100
78.1
90
69.6
80
63.3
100
66.7
71.4
dSNdS–Fun
DIGDriver
Driverpower
LARVA
Noncoding (excHypermutation)
All tissues
Overlapping tissues
Jaccard index (%)
0
5
10
15
17.2
10.7
2.8
1.1
7.5
4.3
7.3
6.4
dNdS–Fun
DIGDriver
Driverpower
LARVA

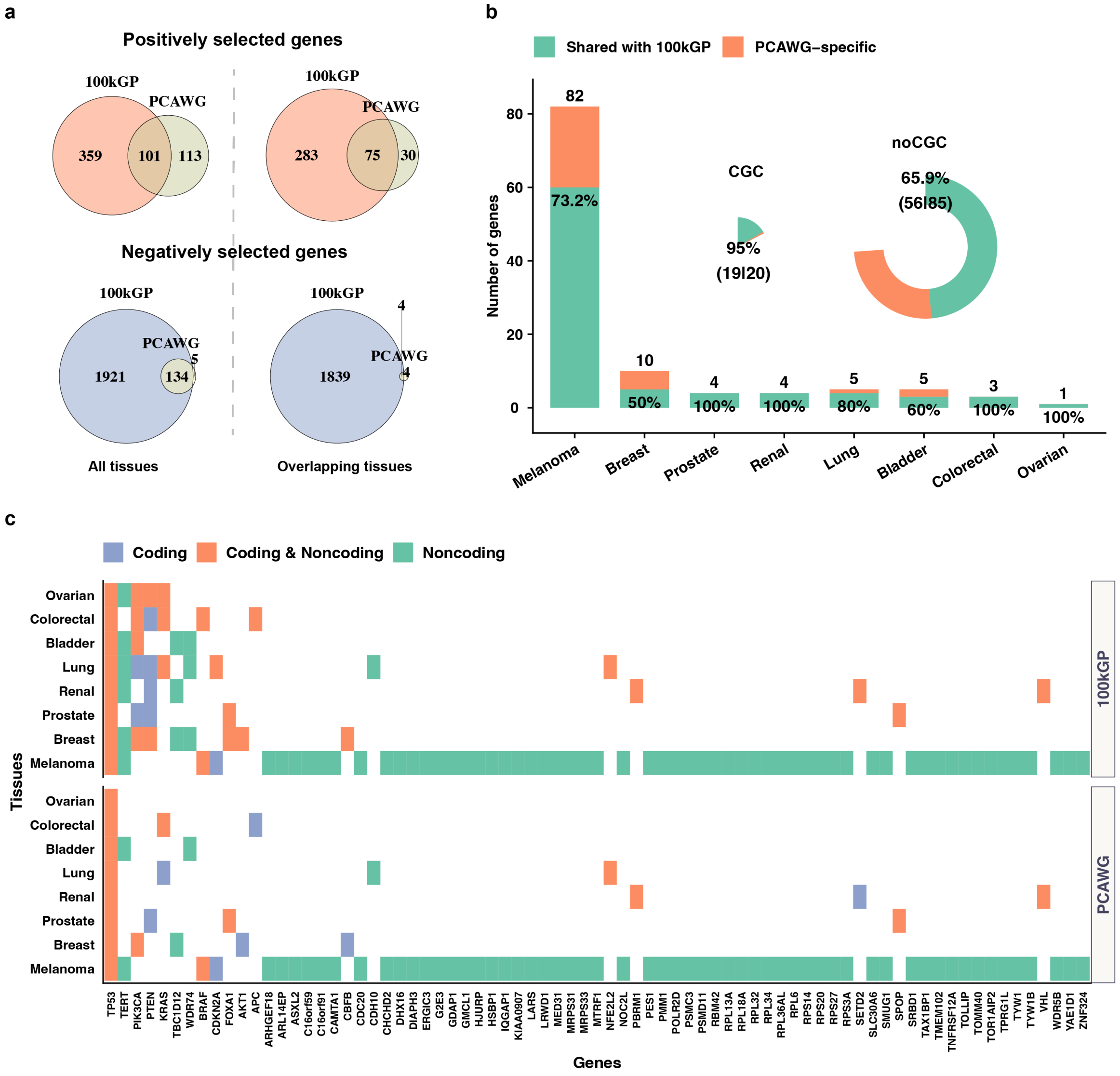
Supplementary Figure 6. Comparison of the performance of dNdS-Fun with DIGDriver, DriverPower, and LARVA using real data. a) Precision, recall, and F1-score for genes identified across all tissues and in eight tissues that overlap between 100kGP and PCAWG (i.e., melanoma, breast, bladder, colorectal, prostate, renal, lung, and ovarian cancers), hereafter referred to as overlapping tissues. b) Precision, recall, and F1-score for genes identified in each individual tissue, with *P*-values computed using Student's *t*-test to compare dNdS-Fun with the other methods. c) Precision, recall, and F1-score for genes identified across all tissues, and in overlapping tissues, excluding hypermutator cancers (i.e., colorectal, melanoma, and endometrial cancers). d) Precision of genes identified in all tissues and in overlapping tissues from 100kGP and PCAWG, before and after excluding hypermutator cancers. e) Jaccard similarity of genes identified between 100kGP and PCAWG, before and after excluding hypermutator cancers.



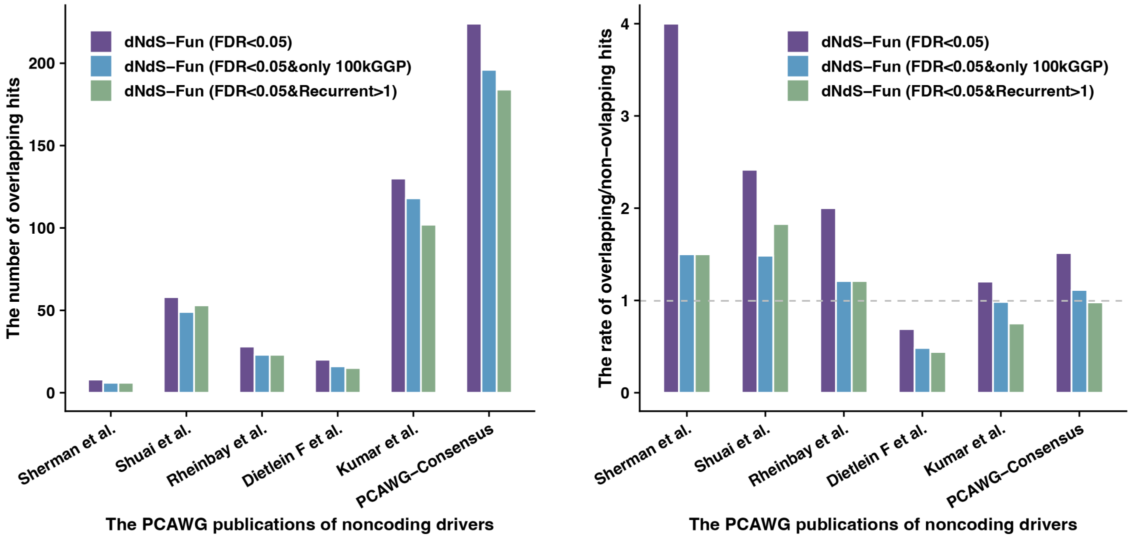
Supplementary Figure 7. Comparison of *ω* estimates between different cohorts and sequencing platforms. a) Comparisons of *ω* estimates between 100kGP and PCAWG across different functional categories. b) The top panel displays selection signatures in coding regions across the TCGA, 100kGP, and PCAWG datasets. The bottom panel compares the *ω* estimates in coding regions between different sequencing platforms: WGS (100kGP and PCAWG) vs. WES (TCGA). The Pearson correlation coefficient (r) and *P*-value are shown in the top left corner of each plot. The grey shaded area represents the 95% CI.



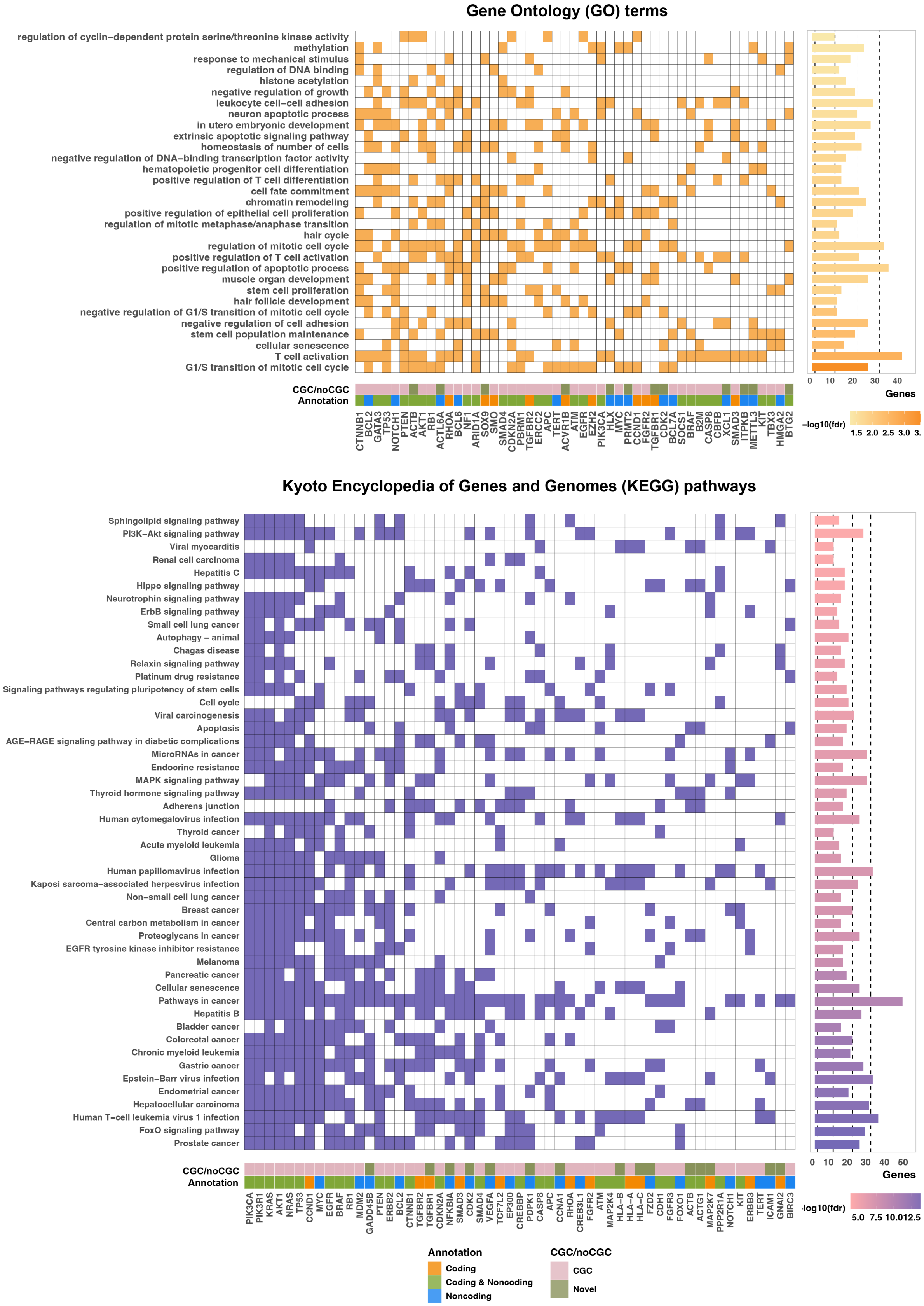
Supplementary Figure 8. Correlation between number of mutations and *ω* estimate across cancer types in each of the eight functional categories. Each dot represents the *ω* estimate from a cancer type. The Pearson correlation coefficient (r) and *P*-value are shown in the top right corner of each plot. The grey shaded area represents the 95% CI.



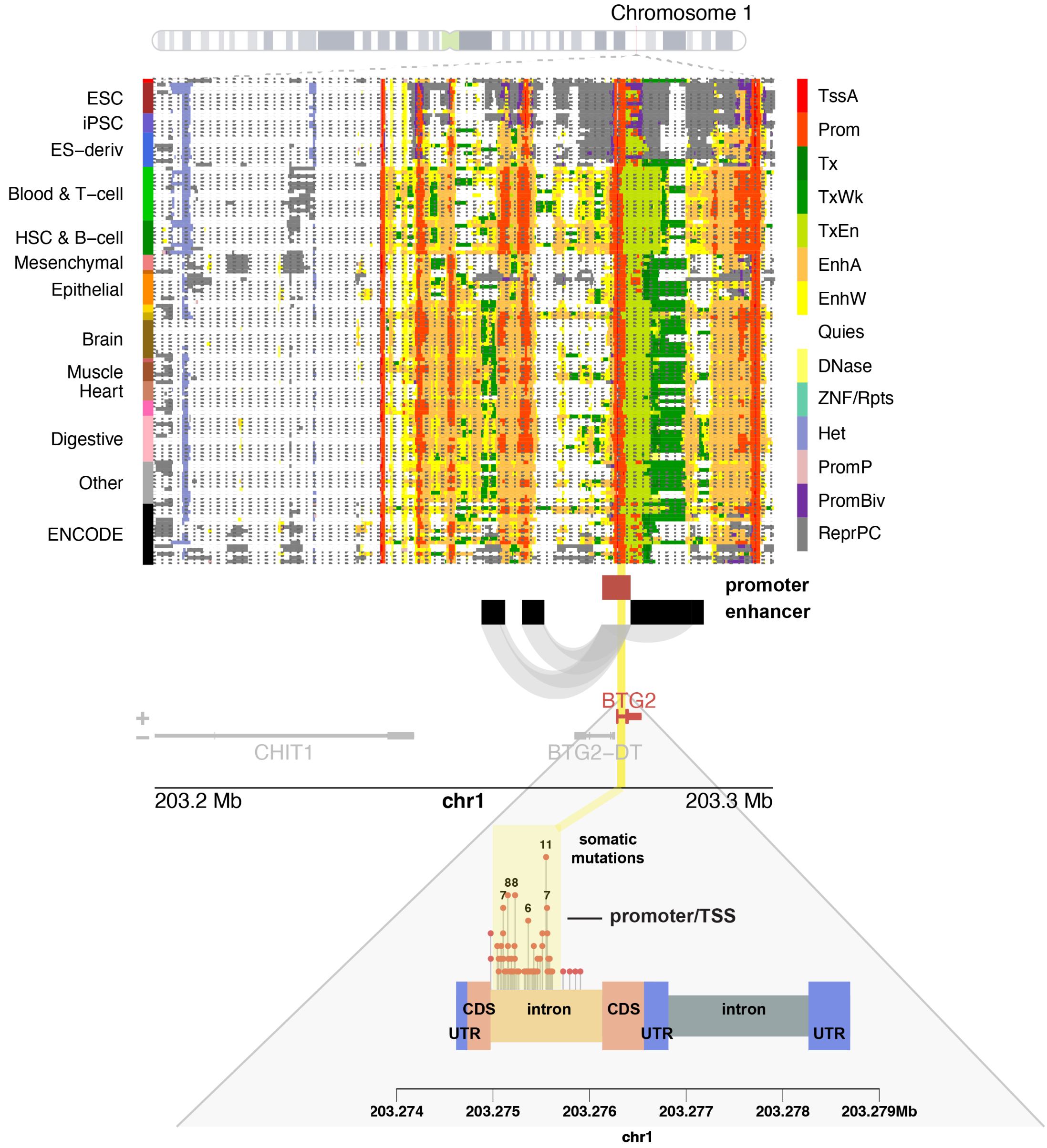
Supplementary Figure 9. Comparison of genes under selection identified in the 100kGP and PCAWG datasets. a) The number of overlapping positively and negatively selected genes identified in both datasets. b) Replication in the 100kGP dataset for genes under selection identified in PCAWG, categorized by their inclusion in the CGC. c) Evidence of selection for genes identified in both the 100kGP and PCAWG datasets.



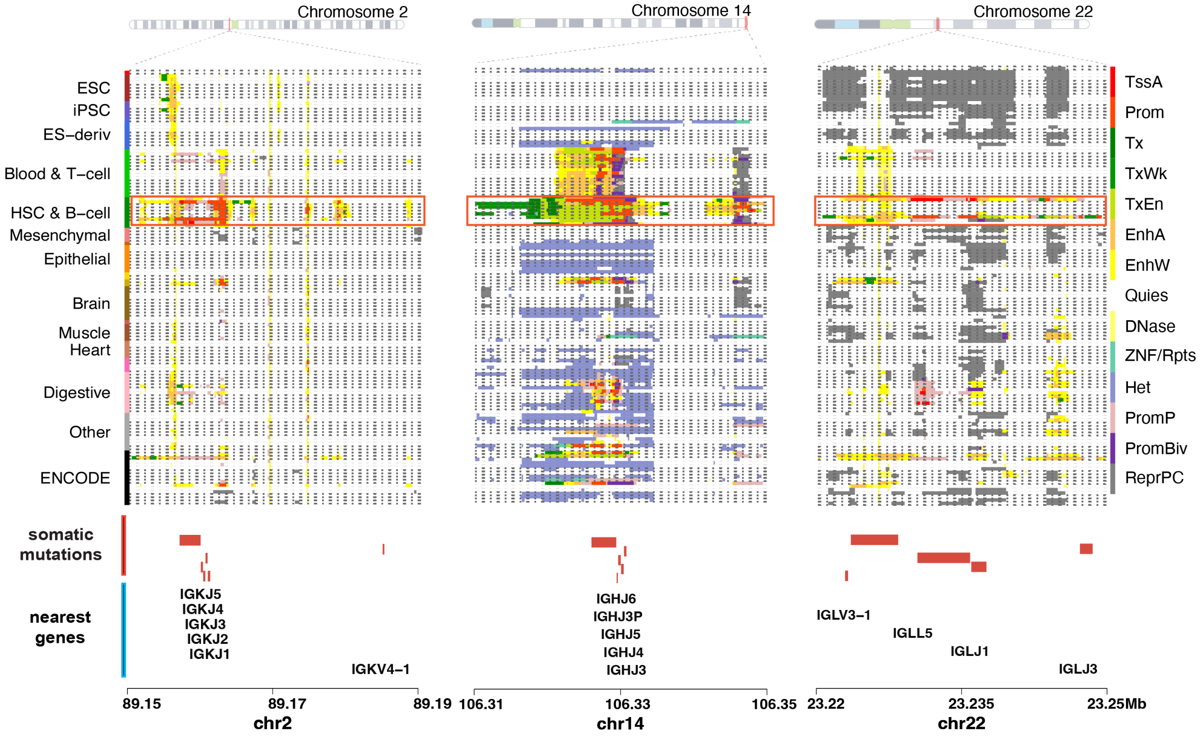
Supplementary Figure 10. Replication of known noncoding drivers in positively selected genes as identified by dNdS-Fun. The known noncoding drivers were previously reported in papers from PCAWG and are detailed in **Supplementary Table 9**. This analysis includes genes with mutations in key regulatory regions: promoters, splice sites, 5’UTRs and 3’UTRs. Each hit represents a unique genomic element-tumor pair where replication of the known noncoding driver was observed.


Gene Ontology (GO) terms
regulation of cyclin−dependent protein serine/threonine kinase activity
methylation
response to mechanical stimulus
regulation of DNA binding
histone acetylation
negative regulation of growth
leukocyte cell−cell adhesion
neuron apoptotic process
in utero embryonic development
extrinsic apoptotic signaling pathway
homeostasis of number of cells
negative regulation of DNA−binding transcription factor activity
hematopoietic progenitor cell differentiation
positive regulation of T cell differentiation
cell fate commitment
chromatin remodeling
positive regulation of epithelial cell proliferation
regulation of mitotic metaphase/anaphase transition
hair cycle
regulation of mitotic cell cycle
positive regulation of T cell activation
positive regulation of apoptotic process
muscle organ development
stem cell proliferation
hair follicle development
negative regulation of G1/S transition of mitotic cell cycle
negative regulation of cell adhesion
stem cell population maintenance
cellular senescence
T cell activation
G1/S transition of mitotic cell cycle
CGC/noCGC
Annotation
CTNNB1
BCL2
GATA3
TP53
NOTCH1
PTEN
ACTB
AKT1
RB1
ACTL6A
RHOA
BCL6
NF1
ARID1A
SOX9
SMO
SMAD4
CDKN2A
PBRM1
TGFBR2
ERCC2
APC
TERT
ACVR1B
ATM
EGFR
EZH2
PIK3CA
HLX
MYC
PRMT2
CCND1
FGFR2
TGFBR1
CDK2
BCL7A
SOCS1
BRAF
B2M
CASP8
CBFB
XCL1
SMAD3
ITPKB
METTL3
KIT
TBX3
HMGA2
BTG2
0
10
20
30
40
Genes
−log10(fdr)
1.5
2.0
2.5
3.0
3.
Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways
Sphingolipid signaling pathway
PI3K−Akt signaling pathway
Viral myocarditis
Renal cell carcinoma
Hepatitis C
Hippo signaling pathway
Neurotrophin signaling pathway
ErbB signaling pathway
Small cell lung cancer
Autophagy − animal
Chagas disease
Relaxin signaling pathway
Platinum drug resistance
Signaling pathways regulating pluripotency of stem cells
Cell cycle
Viral carcinogenesis
Apoptosis
AGE−RAGE signaling pathway in diabetic complications
MicroRNAs in cancer
Endocrine resistance
MAPK signaling pathway
Thyroid hormone signaling pathway
Adherens junction
Human cytomegalovirus infection
Thyroid cancer
Acute myeloid leukemia
Glioma
Human papillomavirus infection
Kaposi sarcoma−associated herpesvirus infection
Non−small cell lung cancer
Breast cancer
Central carbon metabolism in cancer
Proteoglycans in cancer
EGFR tyrosine kinase inhibitor resistance
Melanoma
Pancreatic cancer
Cellular senescence
Pathways in cancer
Hepatitis B
Bladder cancer
Colorectal cancer
Chronic myeloid leukemia
Gastric cancer
Epstein−Barr virus infection
Endometrial cancer
Hepatocellular carcinoma
Human T−cell leukemia virus 1 infection
FoxO signaling pathway
Prostate cancer
CGC/noCGC
Annotation
PIK3CA
PIK3R1
KRAS
AKT1
NRAS
TP53
CCND1
MYC
EGFR
BRAF
RB1
MDM2
GADD45B
PTEN
ERBB2
BCL2
CTNNB1
TGFBR2
TGFBR1
CDKN2A
NFKBIA
SMAD3
CDK2
SMAD4
VEGFA
TCF7L2
EP300
CREBBP
PDPK1
CASP8
APC
CCNA1
RHOA
CREB3L1
FGFR2
ATM
MAP2K4
HLA−B
HLA−A
HLA−C
FZD2
CDH1
FGFR3
FOXO1
ACTB
ACTG1
MAP2K7
PPP2R1A
NOTCH1
KIT
ERBB3
TERT
ICAM1
GNAI2
BIRC3
0
10
20
30
40
50
Genes
-log10(fdr)
5.0
7.5
10.0
12.5
Annotation
Coding
Coding & Noncoding
Noncoding
CGC/noCGC
CGC
Novel

Supplementary Figure 11. Enrichment analysis for positively selected genes identified by dNdS-Fun. The plot shows the significant GO terms and KEGG pathways. The bar plot on the right shows the number of identified genes associated with each term or pathway. Classification under 'Known/Novel' distinguishes between genes that are listed in the CGC and those that are not. The 'Annotation' classification details the type of mutation evidence observed: coding only, noncoding only, or both coding and noncoding.



Supplementary Figure 12. Selection signatures in the *BTG2* gene. The top track shows the 14 chromatin state annotations across 127 samples from the Roadmap Epigenomics Mapping Consortium (REMC), representing various primary cells and tissue types, each row color-coded according to chromatin state. The middle track shows the Hi-C interaction loops, collected from two blood cell samples, indicating regulatory interactions. The bottom part of the figure illustrates recurrent somatic mutations located in the intronic regions of the *BTG2* gene as identified in the PCAWG dataset, with numbers indicating the mutation count at each genomic site. The "promoter/TSS" annotation for this intronic region is inferred based on the chromatin marks shown.



Supplementary Figure 13. Selection signatures in the Immunoglobulin (Ig) gene loci. The top track displays the 14 chromatin state annotations from 127 samples provided by the REMC, representing different primary cells and tissue types. Each chromatin state is indicated by a unique color. Notably, items related to HSC and B-cell are highlighted in red. The bottom track shows the regions with evidence of positive selection identified in lymphoma cancers. These selected elements are annotated with the nearest genes located within the Ig gene loci, indicated at their corresponding positions on the genome.