Cell Systems, Volume *13*

Supplemental information

Allele-specific genomic data elucidate the role of somatic gain and copy-number neutral loss of heterozygosity in cancer

Yari Ciani, Tarcisio Fedrizzi, Davide Prandi, Francesca Lorenzin, Alessio Locallo, Paola Gasperini, Gian Marco Franceschini, Matteo Benelli, Olivier Elemento, Luca L. Fava, Alberto Inga, and Francesca Demichelis

This manuscript contains the following supplemental materials:

Supplemental Figures

-Supplementary Fig. 1 | Flowchart of the pipeline. Related to Figure 1.

-Supplementary Fig. 2 | Performance analysis of the pipeline. Related to Figure 1.

-Supplementary Fig. 3 | Genetic distance and inference of ethnicity. Related to Figure 1.

-Supplementary Fig. 4 | Distribution of genomic instability measures. Related to Figure 1.

-Supplementary Fig. 5 | Comparison between asP and ABSOLUTE ploidy. Related to Figure 1.

-Supplementary Fig. 6 | Association with prognosis of genomic instability measures. Related to Figure 1.

-Supplementary Fig. 7 | Examples of asCN classification in two samples. Related to Figure 1.

-Supplementary Fig. 8 | Clusters stability. Related to Figure 1.

-Supplementary Fig. 9| Allele-specific copy number data. Related to Figure 1.

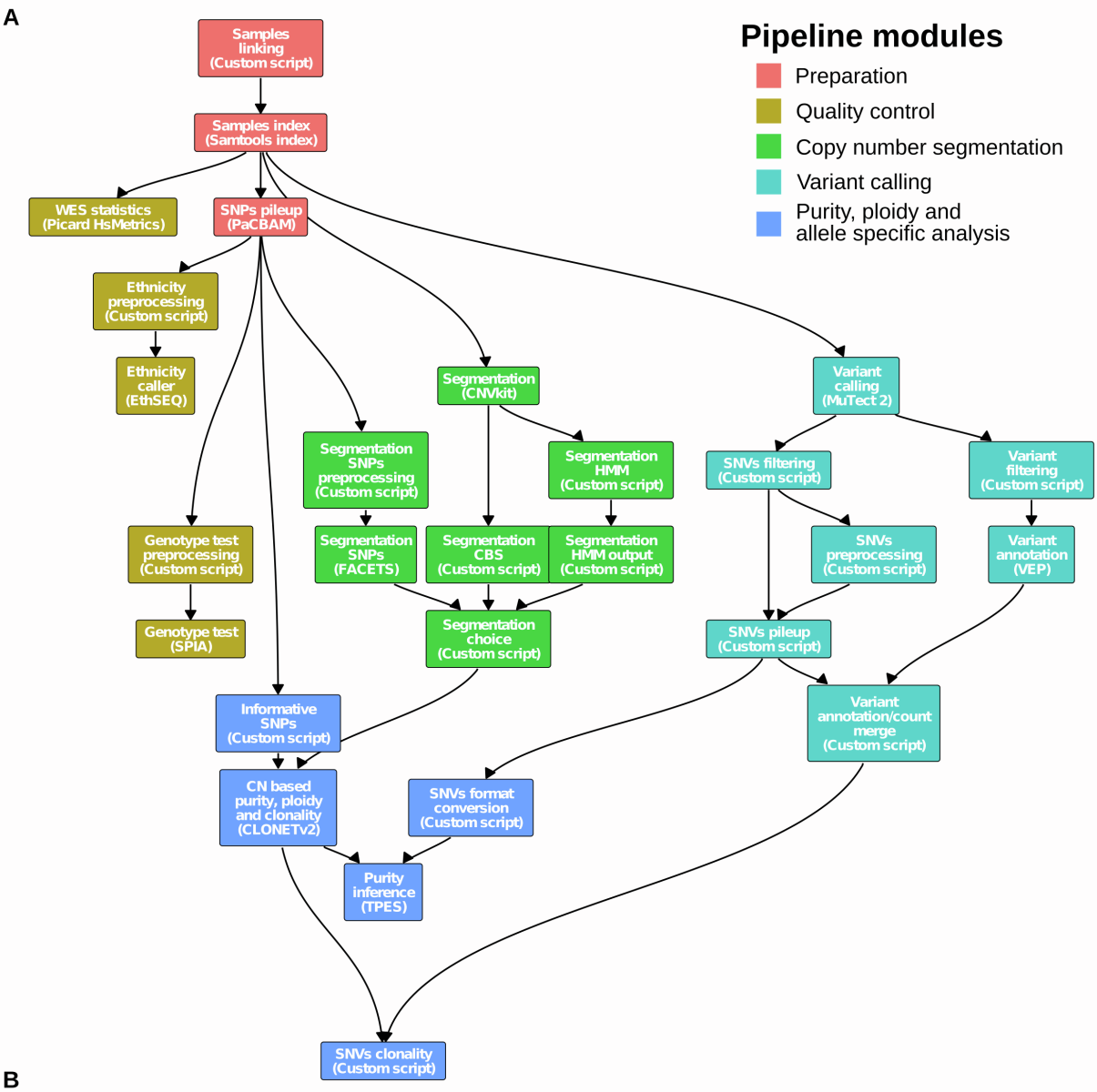
-Supplementary Fig. 10 | Effect of concomitant SNV and loss of wt copy of TP53 on target genes. Related to Figure 2.

-Supplementary Fig. 11 | Loss of Heterozygosity burden across TCGA studies. Related to Figure 3.

-Supplementary Fig. 12 | LOH events across samples and tumor types. Related to Figure 3.

-Supplementary Fig. 13 | Loss of Heterozygosity. Related to Figure 4.

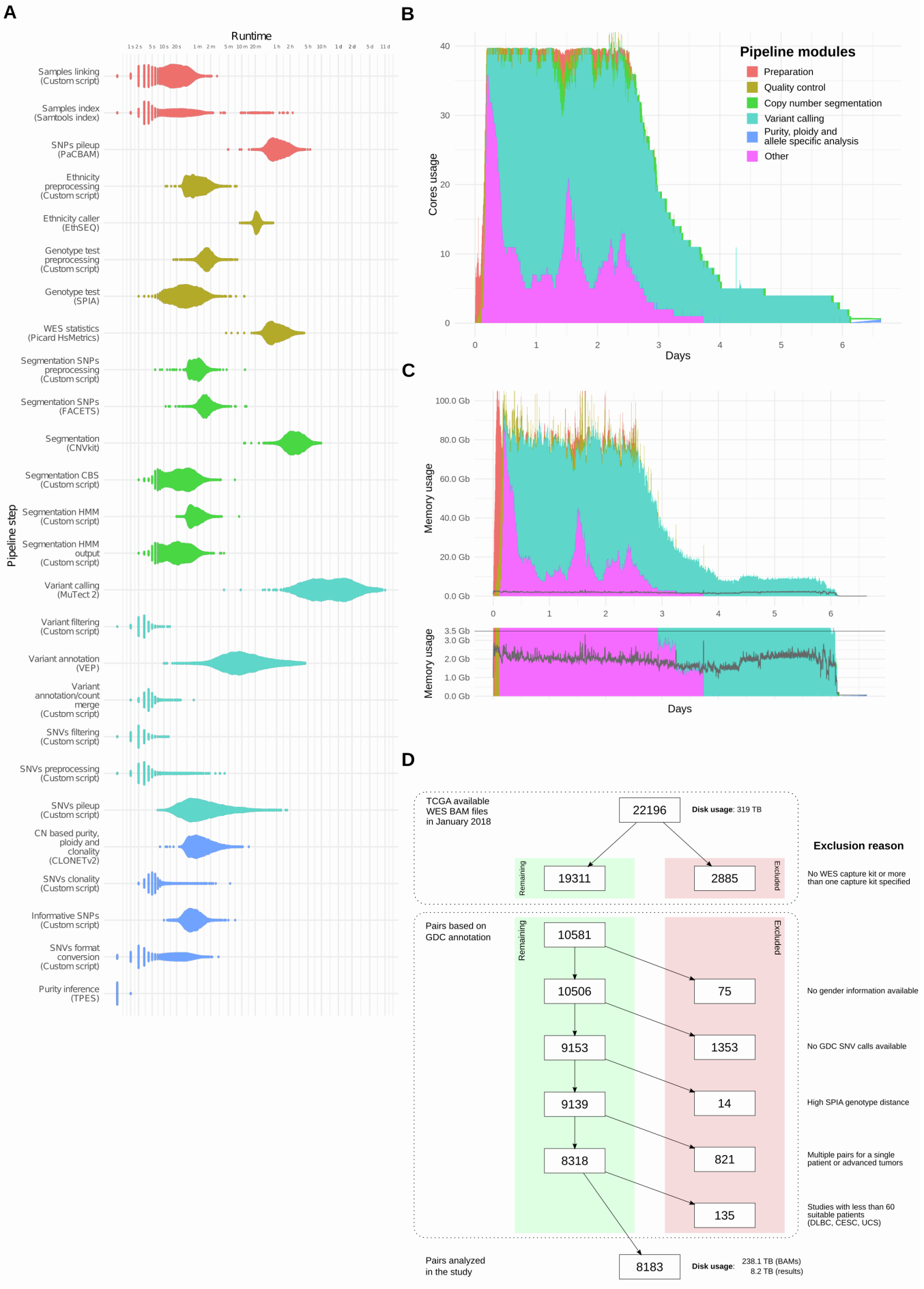
A



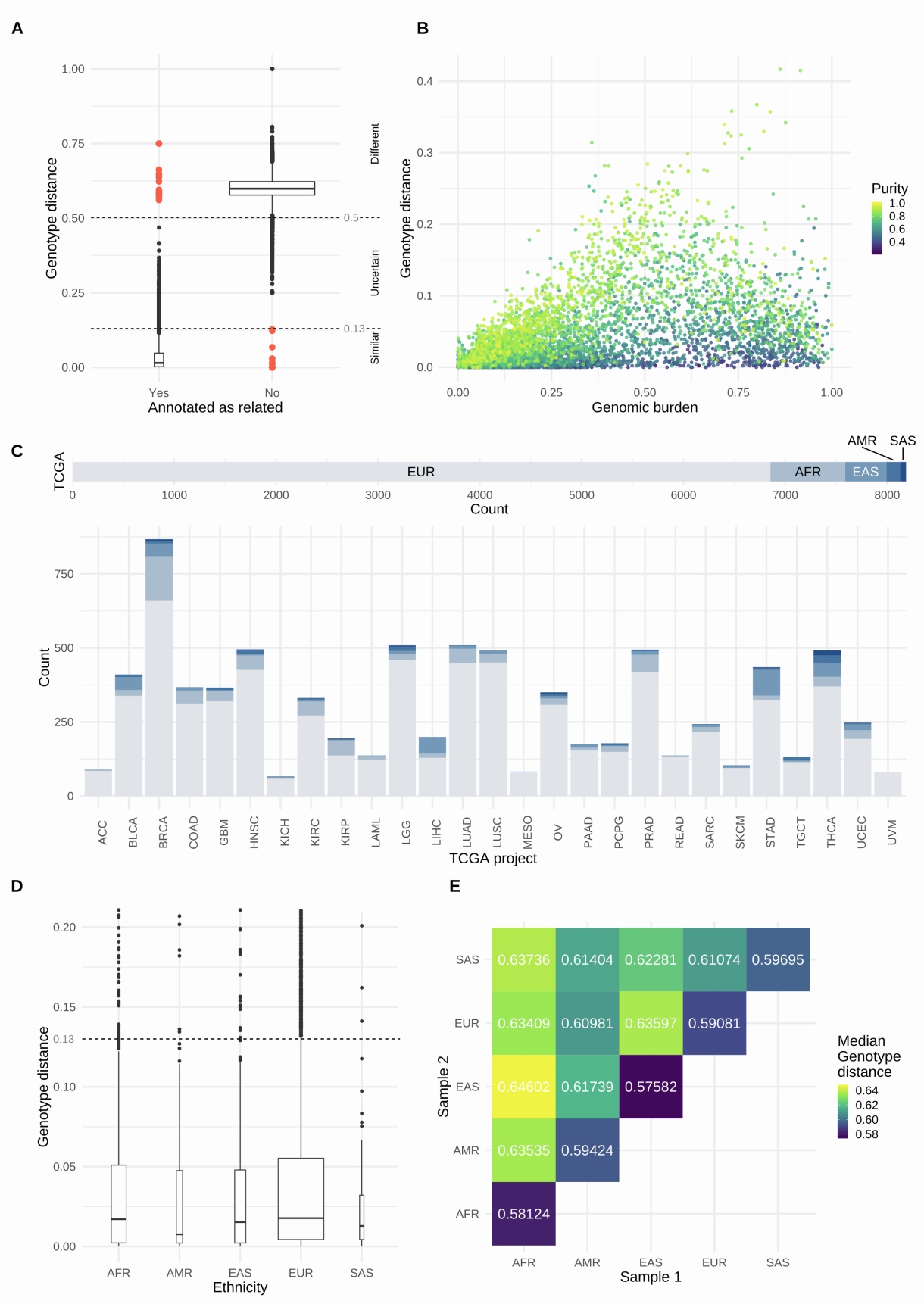
B

Tool	Publication	Link
Samtools	NA	http://www.htslib.org/
Picard	NA	http://broadinstitute.github.io/picard/
PaCBAM	Valentini et al. (2019)	https://doi.org/10.1186/s12864-019-6386-6
EthSEQ	Romanel et al. (2017)	https://doi.org/10.1093/bioinformatics/btx165
SPIA	Demichelis et al. (2008)	https://doi.org/10.1093/nar/gkn089
CNVkit	Talevich et al. (2016)	https://doi.org/10.1371/journal.pcbi.1004873
FACETS	Shen et al. (2016)	https://doi.org/10.1093/nar/gkw520
MuTect 2	NA	https://github.com/broadinstitute/gatk
VEP	McLaren et al. (2016)	https://doi.org/10.1186/s13059-016-0974-4
CLONETv2	Prandi et al. (2019 - 2014)	https://doi.org/10.1002/cpbi.81
TPES	Locallo et al. (2019)	https://doi.org/10.1093/bioinformatics/btz406

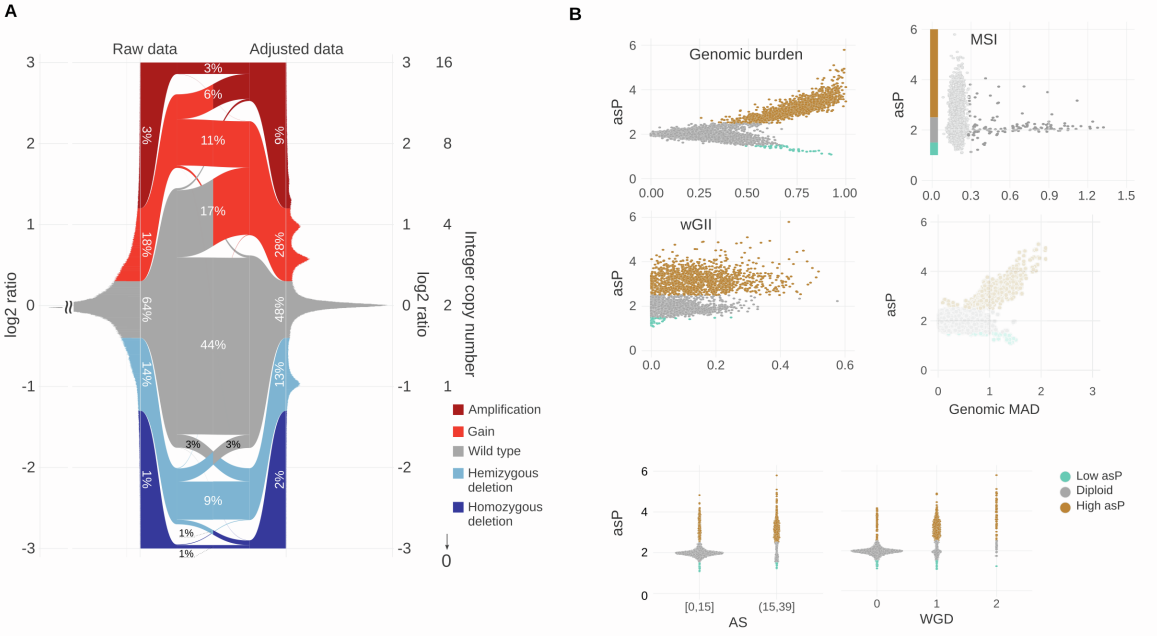
Supplementary Fig. 1 | **Flowchart of the pipeline. Related to Figure 1. A**, Overview of the tools included in the SPICE pipeline and dependencies between the tools. Color code denotes main analysis modules. Technical details at https://github.com/demichelislab/SPICE-pipeline-CWL. **B**, table showing references and links related to the tools used in the modules of the pipeline.


A
Runtime
1 s 2 s 5 s 10 s 20 s 1 m 2 m 5 m 10 m 20 m 1 h 2 h 5 h 10 h 1 d 2 d 5 d 11 d
Pipeline step
Samples linking (Custom script)
Samples index (Samtools index)
SNPs pileup (PaCBAM)
Ethnicity preprocessing (Custom script)
Ethnicity caller (EthSEQ)
Genotype test preprocessing (Custom script)
Genotype test (SPIA)
WES statistics (Picard HsMetrics)
Segmentation SNPs preprocessing (Custom script)
Segmentation SNPs (FACETS)
Segmentation (CNVkit)
Segmentation CBS (Custom script)
Segmentation HMM (Custom script)
Segmentation HMM output (Custom script)
Variant calling (MuTect 2)
Variant filtering (Custom script)
Variant annotation (VEP)
Variant annotation/count merge (Custom script)
SNVs filtering (Custom script)
SNVs preprocessing (Custom script)
SNVs pileup (Custom script)
CN based purity, ploidy and clonality (CLONETv2)
SNVs clonality (Custom script)
Informative SNPs (Custom script)
SNVs format conversion (Custom script)
Purity inference (TPES)
B
Cores usage
Days
Pipeline modules
Preparation
Quality control
Copy number segmentation
Variant calling
Purity, ploidy and allele specific analysis
Other
C
Memory usage
Days
D
TCGA available WES BAM files in January 2018
22196
Disk usage: 319 TB
Remaining
19311
Excluded
2885
Exclusion reason
No WES capture kit or more than one capture kit specified
Pairs based on GDC annotation
Remaining
Excluded
10581
10506
9153
9139
8318
8183
75
1353
14
821
135
No gender information available
No GDC SNV calls available
High SPIA genotype distance
Multiple pairs for a single patient or advanced tumors
Studies with less than 60 suitable patients (DLBC, CESC, UCS)
Pairs analyzed in the study
Disk usage: 238.1 TB (BAMs) 8.2 TB (results)

Supplementary Fig. 2 | **Performance analysis of the pipeline. Related to Figure 1. A**, Violin plot of the execution times of the SPICE pipeline colored by analysis module. **B**, Usage of the cores of the machine (40 cores) stratified by analysis step, including MuTect processing of reference model using panel of normal (N~200) (pink). **C**, Aggregate usage of memory of the machine stratified by analysis step (top). The gray line (magnified in the bottom part of the panel) reports the average usage of memory per core. **D**, The panel shows the selection steps performed. The top figure reports the total number of downloaded samples and number of samples excluded because of the WES kit (red). The bottom section of the figure refers to the number of sample pairs: left (green), number of pairs in the cohort at each step; central (red), number of excluded pairs; right, exclusion criteria.


A
1.00
0.75
0.50
0.25
0.00
Genotype distance
0.5
0.13
Different
Uncertain
Similar
Yes
No
Annotated as related
B
0.4
0.3
0.2
0.1
0.0
Genotype distance
0.00
0.25
0.50
0.75
1.00
Genomic burden
Purity
1.0
0.8
0.6
0.4
C
TCGA
EUR
AFR
EAS
AMR
SAS
0
1000
2000
3000
4000
5000
6000
7000
8000
Count
750
500
250
0
Count
ACC
BLCA
BRCA
COAD
GBM
HNSC
KICH
KIRC
KIRP
LAML
LGG
LIHC
LUAD
LUSC
MESO
OV
PAAD
PCPG
PRAD
READ
SARC
SKCM
STAD
TGCT
THCA
UCEC
UVM
TCGA project
D
0.20
0.15
0.13
0.10
0.05
0.00
Genotype distance
AFR
AMR
EAS
EUR
SAS
Ethnicity
E
SAS
EUR
EAS
AMR
AFR
0.63736
0.61404
0.62281
0.61074
0.59695
0.63409
0.60981
0.63597
0.59081
0.64602
0.61739
0.57582
0.63535
0.59424
0.58124
Sample 2
AFR
AMR
EAS
EUR
SAS
Sample 1
Median Genotype distance
0.64
0.62
0.60
0.58

Supplementary Fig. 3 | Genetic distance and inference of ethnicity. Related to Figure 1. A, Distribution of the SPIA genotype distance (**STAR Methods**) for all possible pairing of study samples stratified by reported annotation. Red dots are pairs whose distance deviates from the expected distance range, suggesting wrong matching between sample and patient in the annotation. In the right boxplot a subsample of 1 million points among of all non-matching (total: 167,588,199) is shown (n samples: 9,153). **B**, Effect of the copy number of the sample on the genotype distance between correctly paired tumor and normal samples (n: 4,950). **C**, Ethnicity of the whole cohort (EthSeQ, see **STAR Methods**) and stratified by tumor type (n: 8,183). **D**, Distribution of genotype distances of matching tumor-normal pairs stratified by ethnicity (n: 12,383). **E**, Median distances of non-matching pairs stratified by combinations of ethnicities (n: 167,588,199).

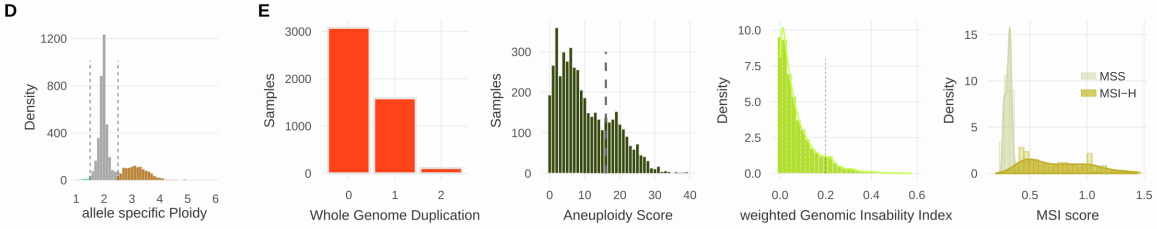


C

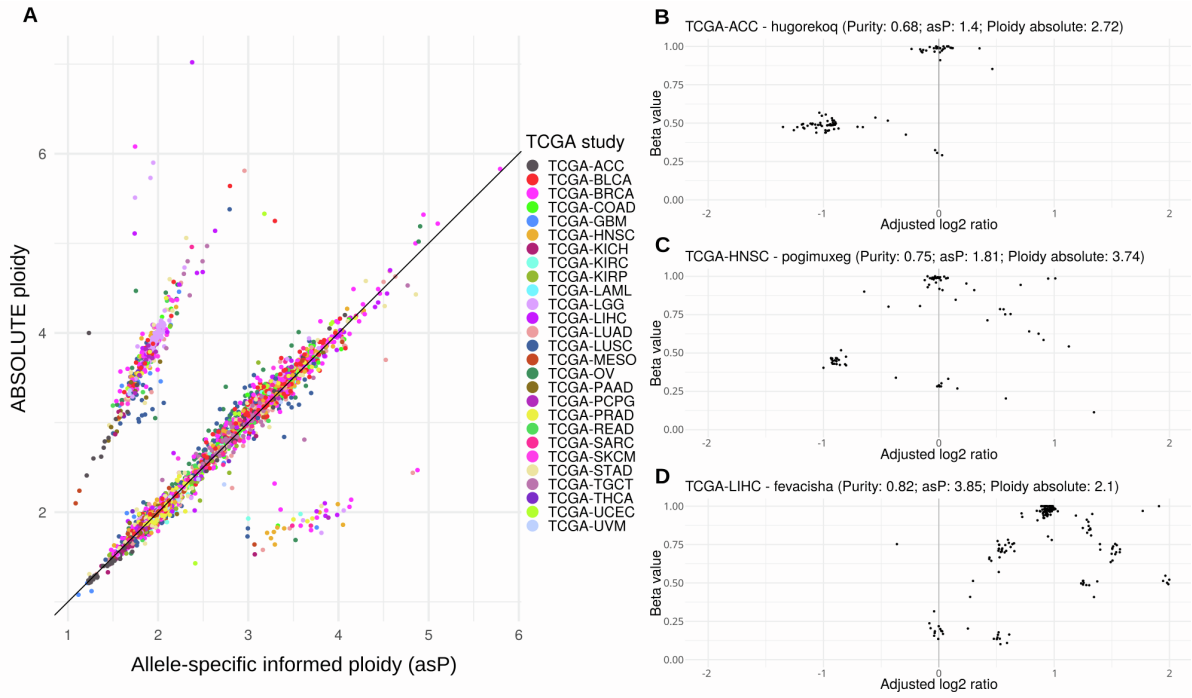
	Mouliere, et al., Sci Transl Med, 2018	*Beroukhim, et al.*, Nature, 2010	*Carter, et al.*, Nat Biotechnol, 2012	*Taylor, et al.*, Cancer Cell, 2018	*Burrell, et al.*, Nature, 2013	This study**
	MAD [0,∞)	GB [0,1]	WGD {0,1,2}	AS {0..39}	wGII [0,1]	asP [0,∞)
Genomic status of tumor cells						
Diploid	0	0	0	0	0	2
Triploid	0.58	1	1	39	0	3
Tetraploid	1	1	1	39	0	4
1 copy loss of 20% of each chr	0*	0.2	0	0	0.2	1.8
1 copy gain of 20% of chr1	0	~0.017	0	0	~0.005	~2.02
1 copy gain of 20% of each chr	0*	0.2	0	0	0.2	2.2
20 copies of chr1	0	~0.087	1	2	0	~3.56
Gain−rich chromothripsis (10% of genome involved)	0	0.1	0	0	~0.1	~2.1

*MAD >0 only if gained portion is telomeric.
**measures proportional to total DNA quantity, including *Carter, et al.*,

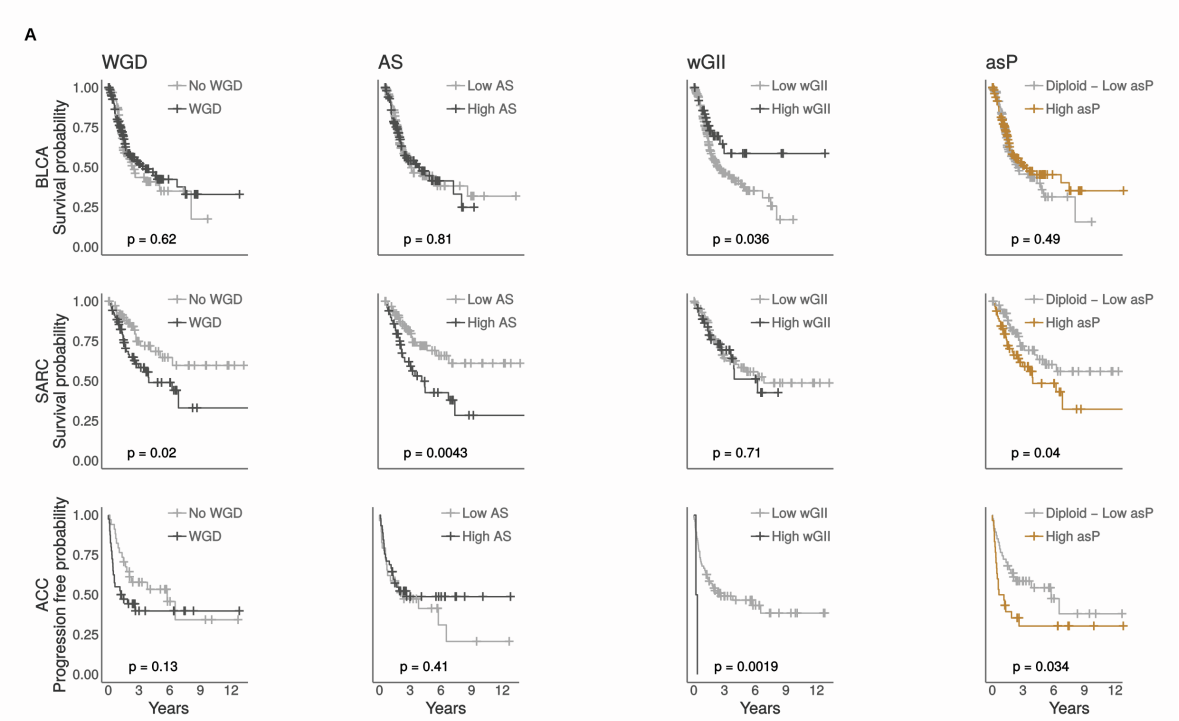
MAD: median absolute deviations
GB: genomic burden
WGD: whole genome duplication
AS: aneuploidy score
wGII: weighted genomic instability index
asP: allele-specific ploidy



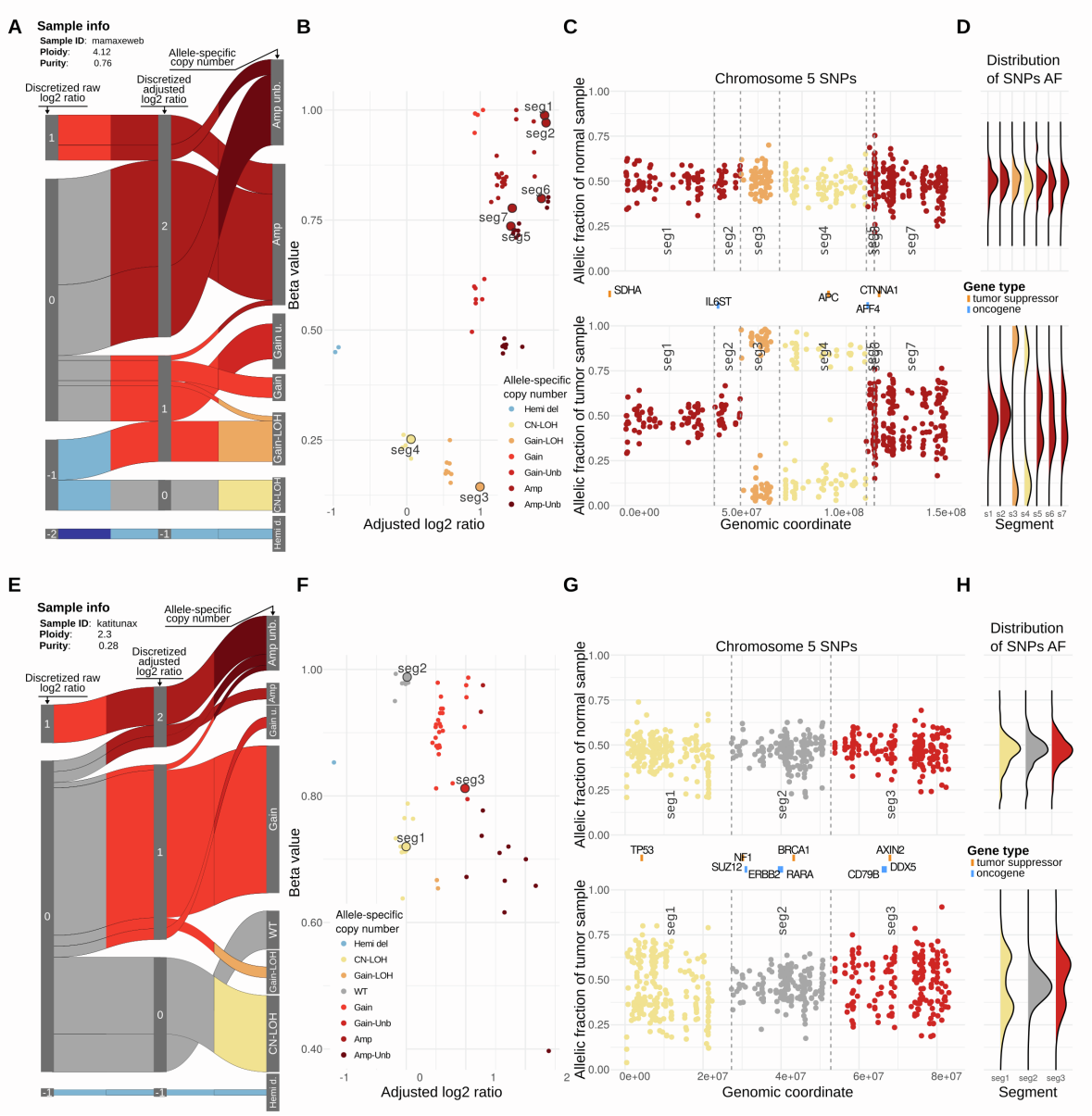
Supplementary Fig. 4 | Distribution of genomic instability measures. Related to Figure 1. A Sankey diagram linking the distribution of the raw log2 ratios with the purity and ploidy adjusted log2(tumor/normal) values for the study cohort. Percentages within the parallel sets plot provide relative fraction of genomic segments corresponding to discretized CN states. Peaks in the right-side distribution are annotated with corresponding CN state (far right column). Tumor ploidy and purity correction using de-facto standard thresholds (**STAR Methods**) led to the reclassification of 32% of the genomic segments, with significant increment of gains (from 18% to 28%, p-value<0.001, proportion test) and amplifications (from 3% to 9%, p-value<0.001, proportion test) (**Table S2**). Further, the number of homozygous deletions almost doubled upon data adjustment (from 4,241 to 9,031, p-value<0.001, proportion test), while a modest but significant reduction in the hemizygous deletions (from 14% to 13%, p-value<0.001, proportion test) was observed; vice-versa, 718 genomic segments previously marked as homozygous deletions (corresponding to 17% of the events) were re-classified as hemizygous deletion and 5,286 hemizygous deletions (corresponding to 7%) changed to homozygous deletions. **B**, Comparison of allele-specific ploidy (asP) with genomic MAD, AS and WGD. Each dot represents a sample color coded by ploidy status. **C**, Genomic indexes values for toy examples of a set of tumor genomic statuses. Each index has own peculiarities; for instance, GB, WGD, and AS don't distinguish triploid from tetraploid and wGII, by definition, is insensitive to whole genome events. **D**, Density distribution of allele-specific ploidy. Vertical dashed lines indicate values of 1.5 and 2.5, delimiting low asP and high asP status, respectively. **E**, Distribution of different genomic instability measures in our cohort. From left to right: whole genome duplication (n=4,780); aneuploidy score (n=4,780); weighted Genomic Instability Index (n=4,950); microsatellite instability score (n=4,945) stratified in microsatellite stable (MSS) and MSI high (MSI-H) samples. Vertical dashed lines separate low from high scores.



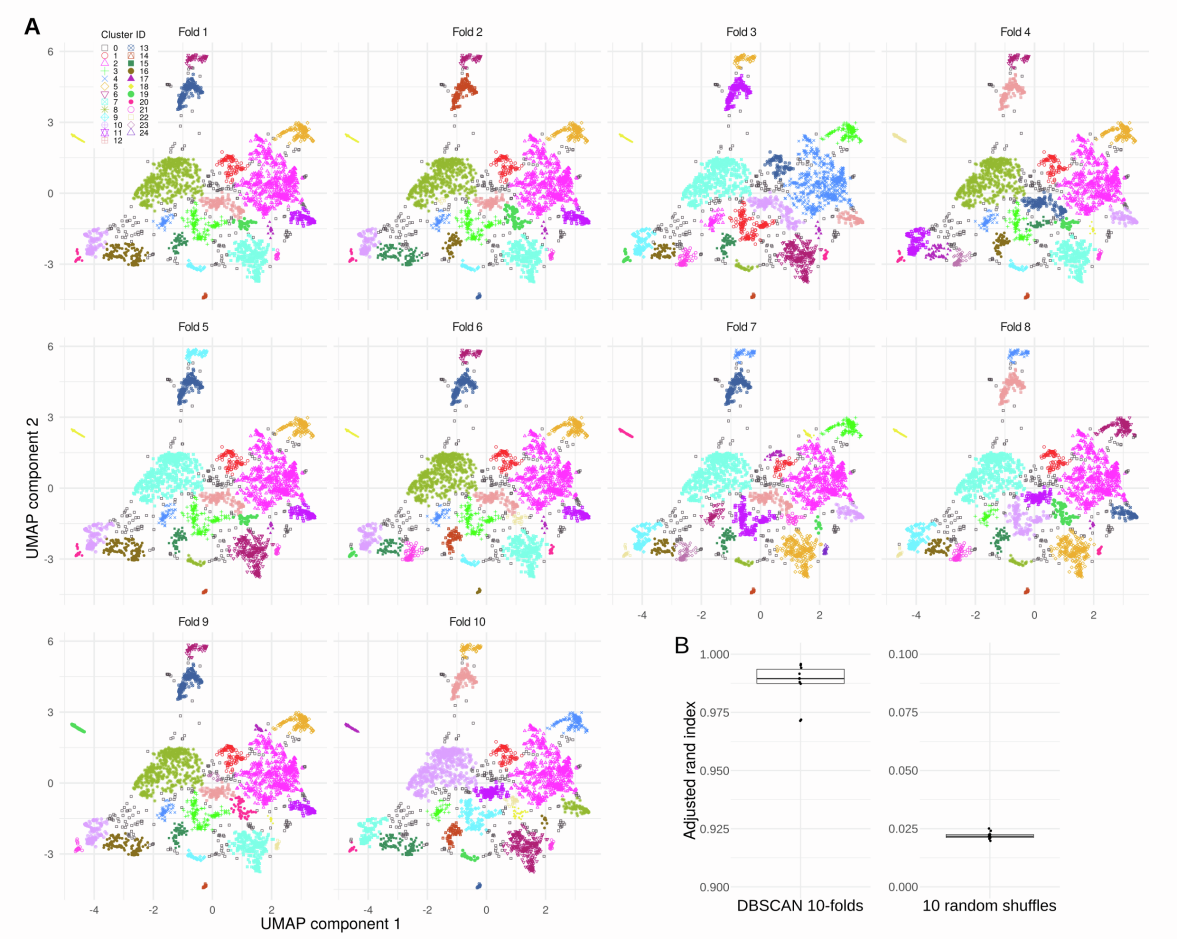
Supplementary Fig. 5 | Comparison between asP and ABSOLUTE ploidy. Related to Figure 1. **A**, Scatter plot of ploidy calls as reported by ABSOLUTE (y-axis) and by CLONETv2 (x-axis) (TCGA tumor data). There's good concordance between the methods, differences arise mainly in samples that one of the two methods estimate as high ploidy. **B** and **C,** details of two high purity tumor cases where ABSOLUTE calls ploidy >2, whereas CLONET v2 calls ploidy <2 and **D** shows a case where CLONET V2 calls high asP whereas ABSOLUTE returns a ploidy of ~2. scatterplot showing adjusted log2 ratio and Beta values of segments: each point represents a segment (labelled segments are shown in panels **F, G**) Beta values represent the fraction of reads in a segment equally representing the two parental alleles. asP values are calculated exclusively from asCN values while ABSOLUTE ploidy measures are derived from karyotypes also informed using the most common cancer karyotypes in a study cohort.



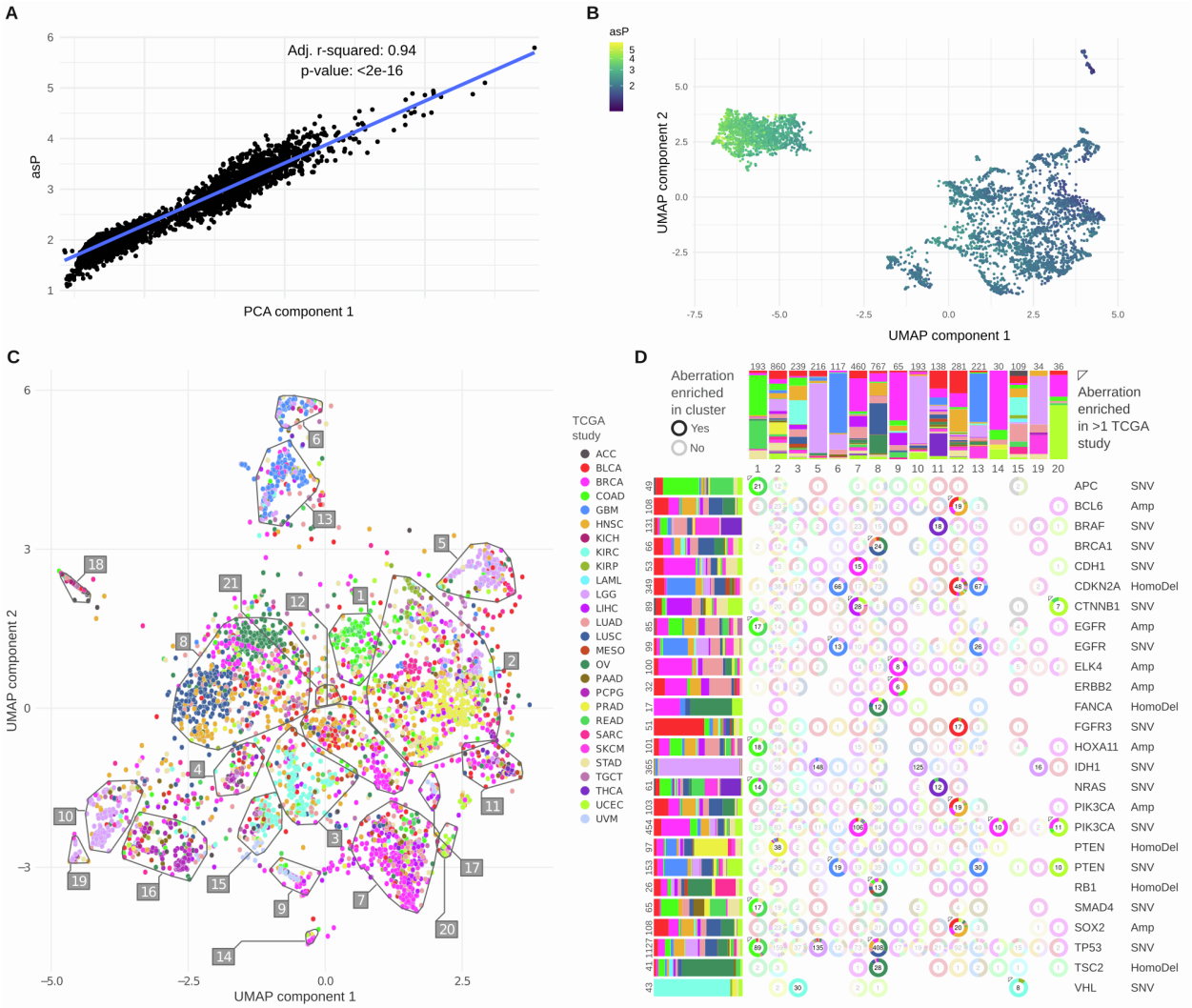
Supplementary Fig. 6 | Association with prognosis of genomic instability measures. Related to Figure 1. A, Kaplan-Meier curves of overall survival or progression free (for ACC) probabilities for different genomic measures. P-value of log-rank test statistics are reported. (GB: genomic burden, WGD: whole genome doubling, AS: aneuploidy score, wGII: weighted genome instability index).



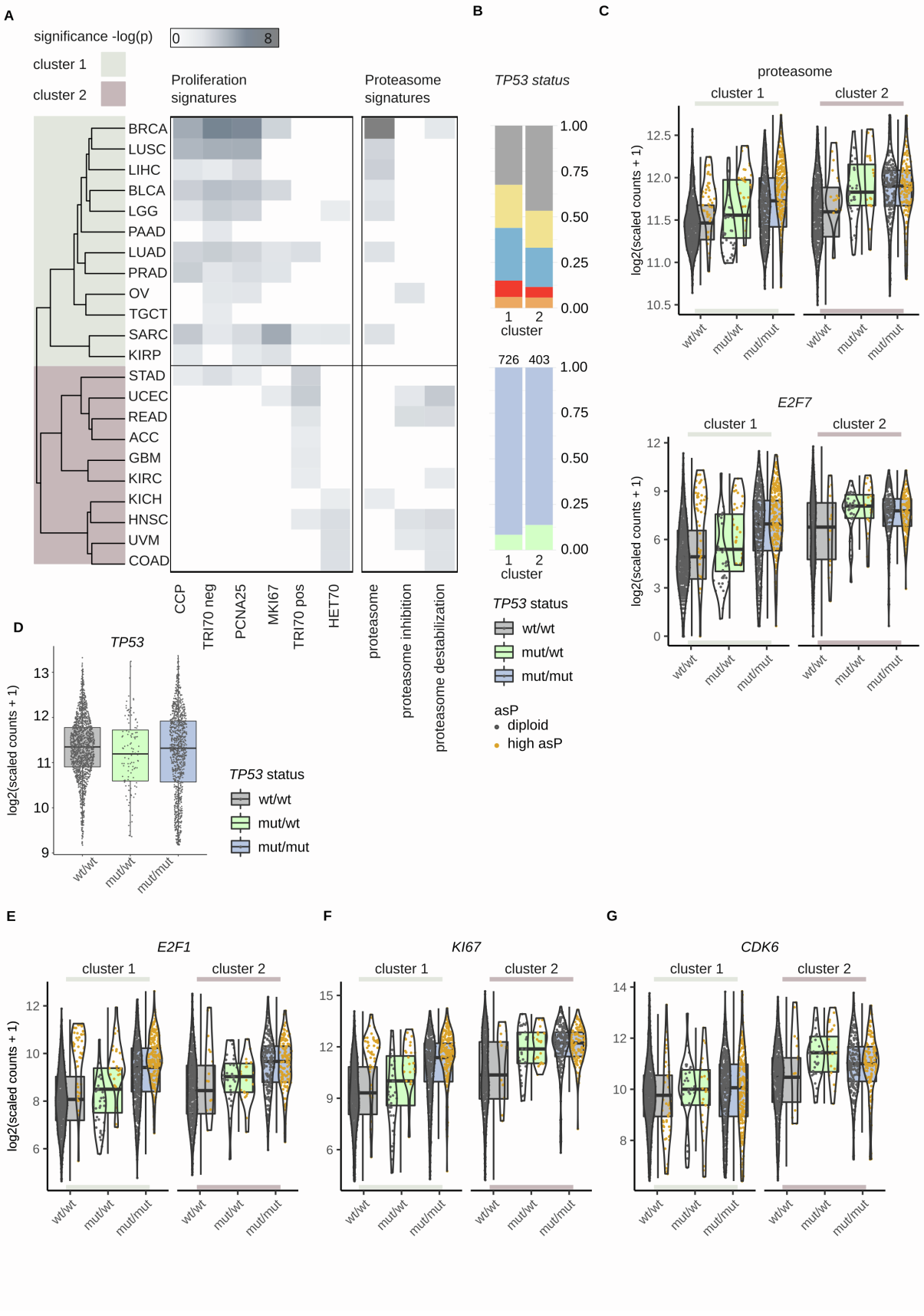
Supplementary Fig. 7 | Examples of asCN classification in two samples. Related to Figure 1. **A, E** Sankey diagram showing classification of segments based on discretized raw log2 (left), purity and ploidy adjusted log2 (middle), and allele-specific copy number (asCN, right). Detailed visualization of segments for the corresponding patients are shown in panels **B, C, D, F, G** and **h**. **B, F** scatterplot showing adjusted log2 and Beta space of segments: each point represents a segment, labelled segments are shown in panels **C, D** and **G, H.** Beta values are estimations of the fraction of reads equally representing the two parental alleles. **C, G** allelic fraction of informative SNPs on chr5 and chr17 in the normal sample (top panels) and tumor samples (bottom panels). **D**, **H**, distribution of allelic fractions in chr5 and chr17 in normal sample (top panels), and tumor samples (bottom panels) stratified by asCN (defined based on tumor sample).



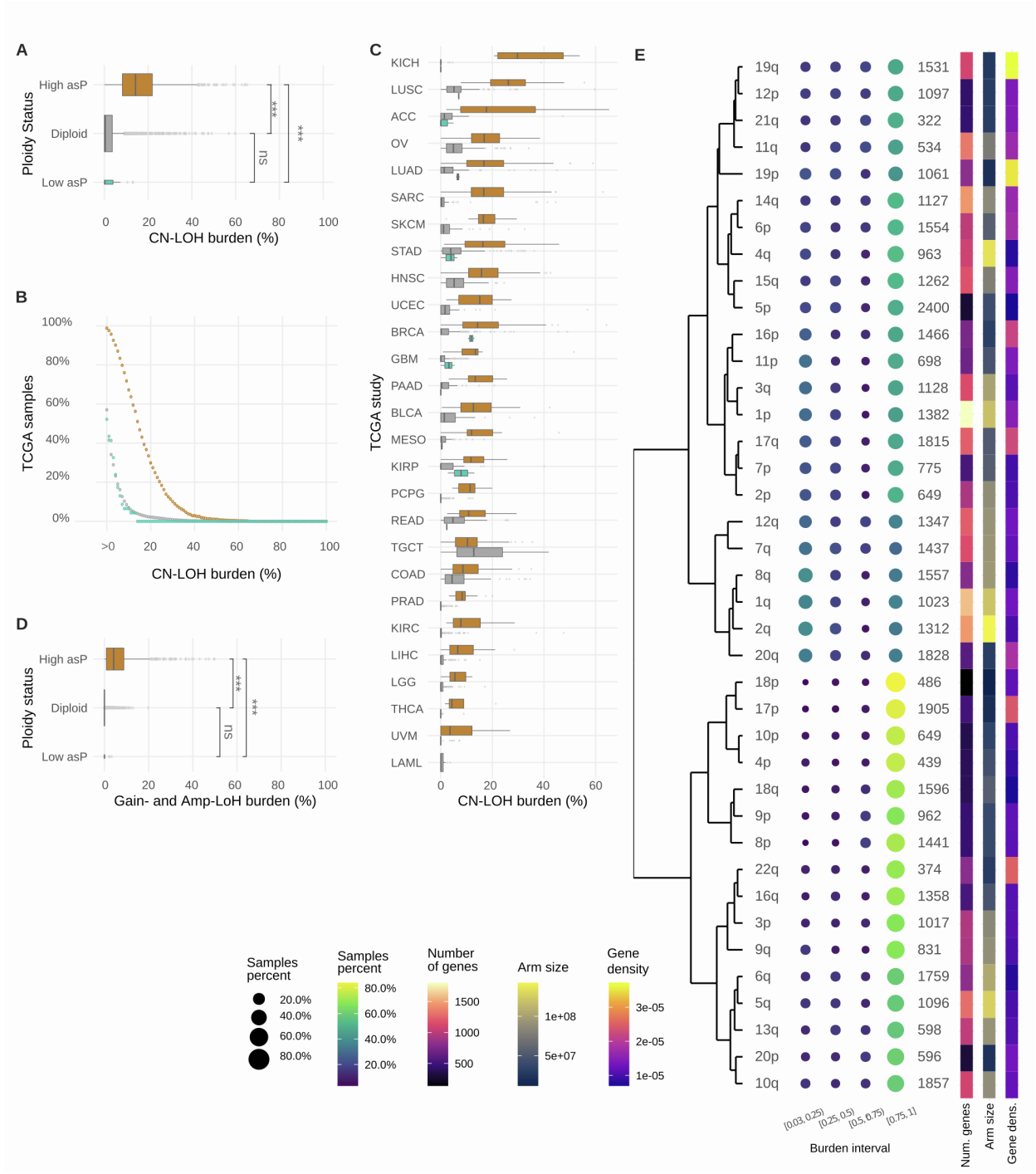
Supplementary Fig. 8 | Clusters stability. Related to Figure 1. A. Visualization of the clusters found by DBSCAN in each fold; the points excluded from each fold are not shown. **B**. On the left: boxplot of the Adjusted Rand Index (ARI) computed on each of the folds; on the right: boxplot of the ARI computed on 10 random shuffles of the cluster labels.



Supplementary Fig. 9 | Allele-specific copy number data. Related to Figure 1. A, scatter plot of the first component of the PCA of the asCN matrix of all the samples against the asP. This demonstrate that the first component of the PCA is highly related to PCA (p-value of r-squared test). **B**, UMAP run on the complete dataset annotated by asP. It is clear that the three groups of points are induced by the difference in terms of asP **C**, UMAP non-linear dimensionality reduction of the gene level allele-specific CN data run on the PCA of the data where the first component (related to asP) was removed. Samples are color coded by tumor type. Boundaries of clusters identified by DBSCAN are shown. All genes are used and interpolation using flanking segments (weights based on log2 values similarities) is applied as necessary. **D**, Pie chart matrix of main tumor aberrations (rows) in panel **D** DBSCAN clusters (columns). Each pie reports the number of aberrant samples and the proportions per tumor type. The triangle indicates that two or more tumor types are enriched for the aberration in the cluster. Rows are annotated with the tumor type distribution of each aberration. Columns are annotated with the tumor type distribution in each DBSCAN cluster. Only data for clusters with at least one enriched aberration are shown.

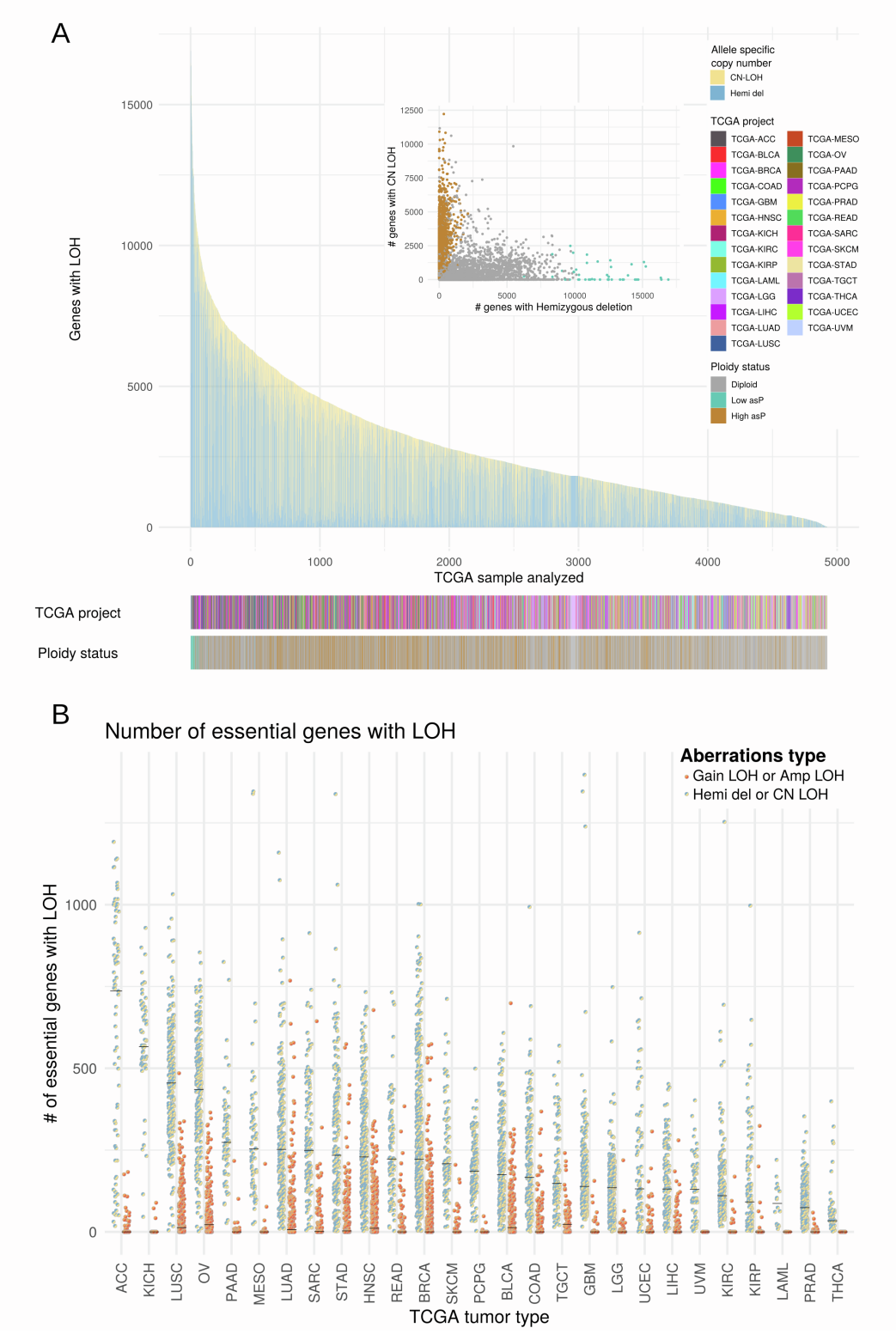

A
significance -log(p)
0
8
cluster 1
cluster 2
Proliferation signatures
Proteasome signatures
BRCA
LUSC
LIHC
BLCA
LGG
PAAD
LUAD
PRAD
OV
TGCT
SARC
KIRP
STAD
UCEC
READ
ACC
GBM
KIRC
KICH
HNSC
UVM
COAD
CCP
TRI70 neg
PCNA25
MKI67
TRI70 pos
HET70
proteasome
proteasome inhibition
proteasome destabilization
B
TP53 status
1.00
0.75
0.50
0.25
0.00
1
2
cluster
726
403
TP53 status
wt/wt
mut/wt
mut/mut
asP
diploid
high asP
C
proteasome
cluster 1
cluster 2
log2(scaled counts + 1)
E2F7
wt/wt
mut/wt
mut/mut
D
TP53
log2(scaled counts + 1)
TP53 status
E
E2F1
F
KI67
G
CDK6

Supplementary Fig. 10 | Effect of concomitant SNV and loss of wt copy of TP53 on target genes. Related to Figure 2. A, Dendrogram and corresponding heatmap of significance (-log(P-value)) of differential expression of proliferative signatures (left) and proteasome signatures (right) when comparing high asP against diploid samples in tumor type. Data is clustered in two groups: cluster #1, characterized by the activation of proliferative signatures in high asP samples; and cluster #2 by repression of proliferation in high asP samples. Five tumor types excluded due to the low number or absence of high asP samples. Activation of the proteasome signature is observed almost exclusively in high asP samples of tumor types included in cluster 1, while inhibition of the proteasome (Levin et al., 2018; Wang et al., 2017) was mainly associated to cluster 2. **B**, Fraction of samples in the two clusters, stratified by asCN and concomitant presence of *TP53* SNV and wild-type copy number. We observe significant enrichment for TP53 SNVs in cluster1 (Chi-squared test, fdr=3.95e-05). Cluster 1 is also enriched for *TP53* LOH events with respect to cluster 2 (top panel, 59% and 47%, respectively, Chi-squared test, p=1.33e-12), and number of samples with SNVs and concomitant loss of wild-type TP53 (mut/mut) (bottom panel, 92% and 86%, respectively, Chi-squared test, p<0.01). We also detected enrichment of copy gain focal events in cluster 1 high asP samples (**Table S19**). **C**, Expression levels of the proteasome signature and E2F7 stratified by cluster, *TP53* status, and ploidy. **D**, Expression level of *TP53* in samples with *TP53* SNV, stratified by the presence of a wt *TP53* copy (**E**, **F** and **G**). Expression levels of *E2F1*, *KI67* and *CDK6* in samples with *TP53* SNV stratified by cluster, wt allele presence, and asP. These analyses include all deleterious mutations.



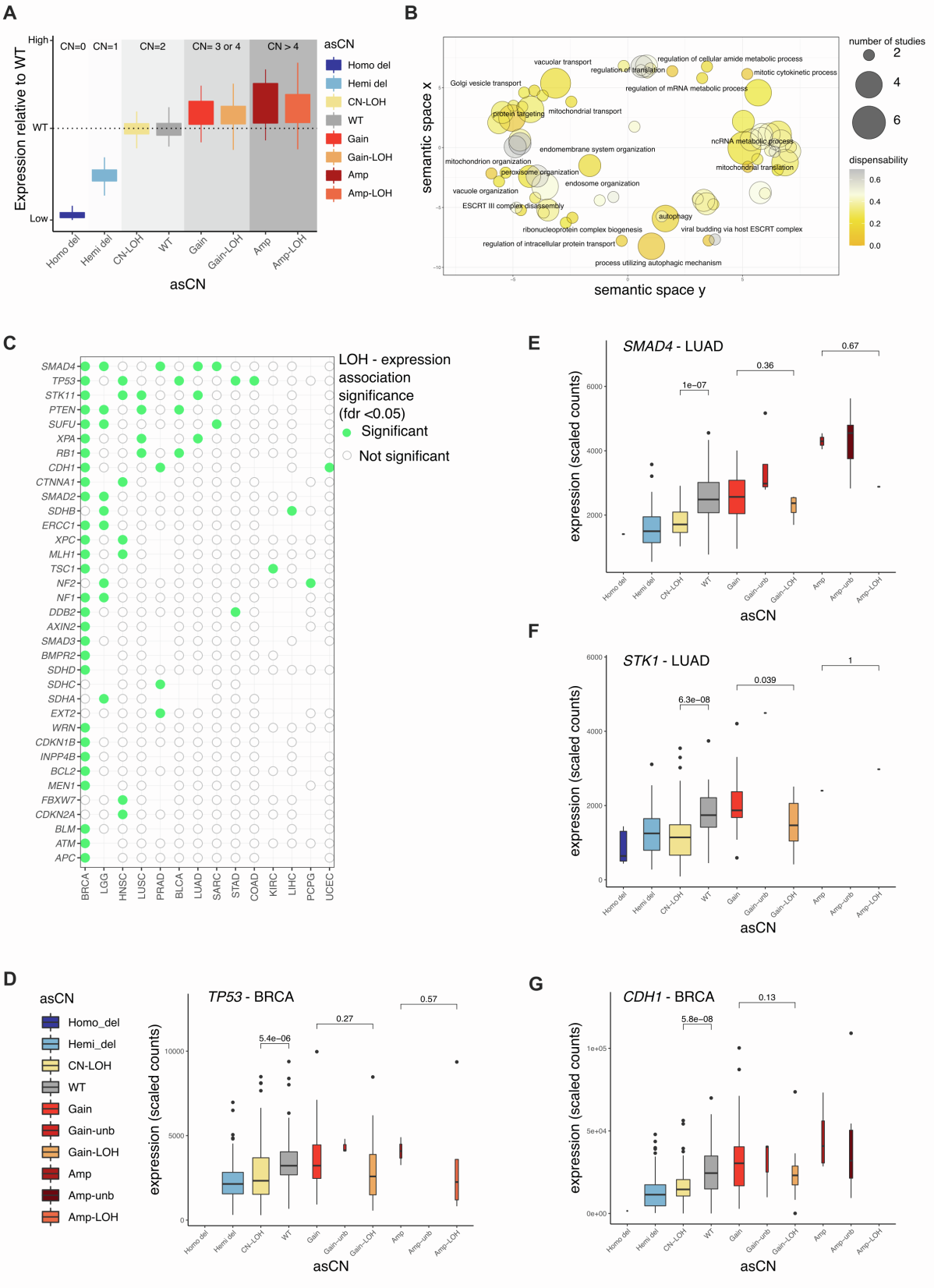
Supplementary Fig. 11 | Loss of Heterozygosity burden across TCGA studies. Related to Figure 3. **A**, Boxplot of copy number neutral loss (CN-LOH) burden against ploidy status. Significant levels of Wilcoxon rank-sum test are reported (*** indicate p-val < 0.001, ns = non-significant). **B**, Percentage of TCGA samples (y-axis) with CN-LOH ≥ x, for each value x of the CN-LOH burden (x-axis), stratified by ploidy status. **C**, For each TCGA study, the plot reports the distribution of CN-LOH burden on a sample basis. Samples from the same TCGA study are stratified by ploidy status. Statistics are reported in **Table S9**. **D**, Boxplot of the sum of copy gain loss of heterozygosity (Gain-LOH) and copy amplification loss of

heterozygosity (Amp-LOH) burden against ploidy status. Significant levels of Wilcoxon rank-sum test are reported (*** indicate p-val < 0.001, ns = non-significant). **E**, Dot plots and dendrogram reporting the fractions of samples showing LOH burden in each interval.



Supplementary Fig. 12 | LOH events across samples and tumor types. Related to Figure 3. A. Visualization of the number of genes that have undergone loss of heterozygosity (either hemizygous deletion or copy neutral LOH) within the TCGA cohort ordered by the total number of aberrations in each

sample. Under the barplot are reported the annotations from the TCGA project and the ploidy status for each sample. The inset compares the number genes with a hemizygous deletion (on the x axis) to the number of genes with CN-LOH (on the y axis). The plot shows that samples have either a high number of hemizygous deletion or a high number of CN-LOH. As evident from the colors of the points this phenomenon is related to the asP of the sample. **B**. Number of aberrations in essential genes (tot: 1478) stratified by aberration class and TCGA study. The black line in each group of points represents the median of the group.


A
High
WT
Low
Expression relative to WT
CN=0
CN=1
CN=2
CN= 3 or 4
CN > 4
asCN
Homo del
Hemi del
CN-LOH
WT
Gain
Gain-LOH
Amp
Amp-LOH
B
semantic space x
semantic space y
vacuolar transport
regulation of cellular amide metabolic process
regulation of translation
mitotic cytokinetic process
Golgi vesicle transport
regulation of mRNA metabolic process
protein targeting
mitochondrial transport
endomembrane system organization
ncRNA metabolic process
mitochondrion organization
peroxisome organization
mitochondrial translation
endosome organization
vacuole organization
ESCRT III complex disassembly
autophagy
ribonucleoprotein complex biogenesis
viral budding via host ESCRT complex
regulation of intracellular protein transport
process utilizing autophagic mechanism
number of studies
2
4
6
dispensability
0.6
0.4
0.2
0.0
C
LOH - expression association significance (fdr <0.05)
Significant
Not significant
SMAD4
TP53
STK11
PTEN
SUFU
XPA
RB1
CDH1
CTNNA1
SMAD2
SDHB
ERCC1
XPC
MLH1
TSC1
NF2
NF1
DDB2
AXIN2
SMAD3
BMPR2
SDHD
SDHC
SDHA
EXT2
WRN
CDKN1B
INPP4B
BCL2
MEN1
FBXW7
CDKN2A
BLM
ATM
APC
BRCA
LGG
HNSC
LUSC
PRAD
BLCA
LUAD
SARC
STAD
COAD
KIRC
LIHC
PCPG
UCEC
D
asCN
Homo_del
Hemi_del
CN-LOH
WT
Gain
Gain-unb
Gain-LOH
Amp
Amp-unb
Amp-LOH
TP53 - BRCA
5.4e−06
0.27
0.57
expression (scaled counts)
asCN
E
SMAD4 - LUAD
1e−07
0.36
0.67
expression (scaled counts)
asCN
F
STK1 - LUAD
6.3e−08
0.039
1
expression (scaled counts)
asCN
G
CDH1 - BRCA
5.8e−08
0.13
expression (scaled counts)
asCN

Supplementary Fig. 13 | Loss of heterozygosity events and their impact on gene expression. Related to Figure 4. A, synthetic data showing the expression levels, stratified by asCN, of a gene for which LOH is not associated on expression and CN is positively associated to expression. **B,** semantic clustering of GO terms associated to genes with LOH status negatively associated to expression. The term “ncRNA metabolic process” is enriched in 6 studies, suggesting a shared mechanism linked to regulation through ncRNAs that is fine-tuned by the number of alleles. **C**, dotmap showing TSG genes with decrease of gene expression upon LOH in each TCGA study. Empty dots indicate “no significance” while the absence of a dot indicates that the test has not been performed because of the low number of events (<10). **D, E, F, G,** examples showing the level of expression of TSG genes stratified by asCN. Width of boxplots is proportional to the number of events. All p-values of Mann-Whitney tests.