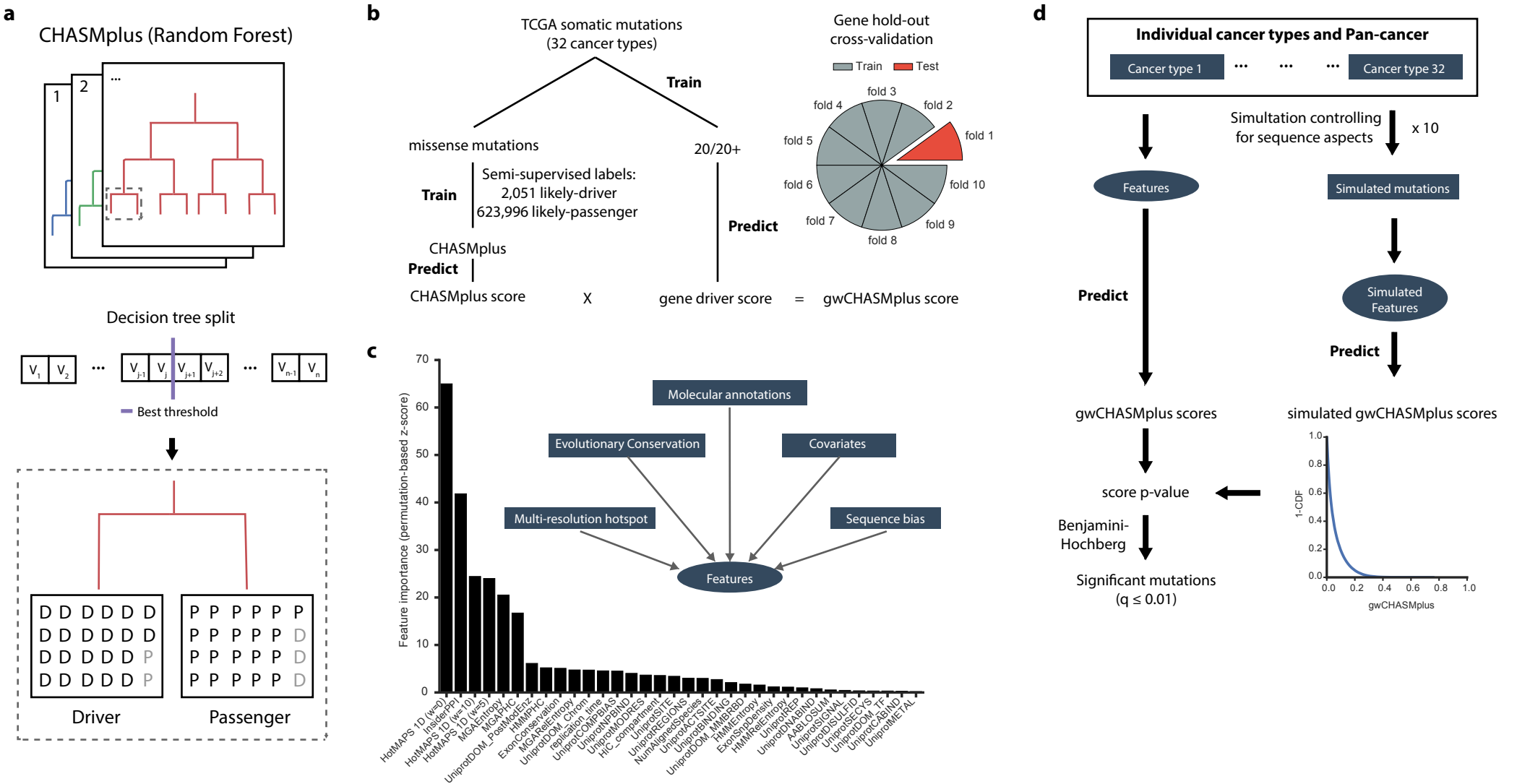
a
CHASMplus (Random Forest)
1
2
...
Decision tree split
Best threshold
D D D D D D
D D D D D D
D D D D D P
D D D D D P
P P P P P P
P P P P P D
P P P P P D
P P P P P D
Driver
Passenger
b
TCGA somatic mutations
(32 cancer types)
Train
missense mutations
Train
Semi-supervised labels:
2,051 likely-driver
623,996 likely-passenger
CHASMplus
Predict
CHASMplus score
20/20+
Predict
Gene hold-out
cross-validation
Train
Test
fold 1
fold 2
fold 3
fold 4
fold 5
fold 6
fold 7
fold 8
fold 9
fold 10
X
gene driver score
=
gwCHASMplus score
c
Feature importance (permutation-based z-score)
Molecular annotations
Evolutionary Conservation
Covariates
Multi-resolution hotspot
Sequence bias
Features
HotMAPS 1D (w=0)
InsideOutPPI
HotMAPS 1D (w=10)
HotMAPS 1D (w=5)
MGAEntropy
MGAPHC
UniprotDOM_PostModEnz
HMMPHC
ExonConservation
MGAReiEntropy
UniprotDOM_Chrom
replication_time
UniprotCOMPBIAS
UniprotNPBIND
UniprotMODRES
HiC_compartment
UniprotSITE
UniprotREGIONS
NumAlignedSpecies
UniprotACTSITE
UniprotBINDING
UniprotDOM_MMBRBD
HMMEntropy
ExonSnpDensity
HMMReiEntropy
UniprotREP
UniprotDNABIND
AABLOSUM
UniprotSIGNAL
UniprotDISULFID
UniprotSECYS
UniprotDOM_TF
UniprotCABIND
UniprotMETAL
d
Individual cancer types and Pan-cancer
Cancer type 1
Cancer type 32
Simultation controlling
for sequence aspects
x 10
Features
Simulated mutations
Predict
Simulated
Features
Predict
gwCHASMplus scores
simulated gwCHASMplus scores
score p-value
Benjamini-
Hochberg
Significant mutations
(q ≤ 0.01)
1-CDF
gwCHASMplus