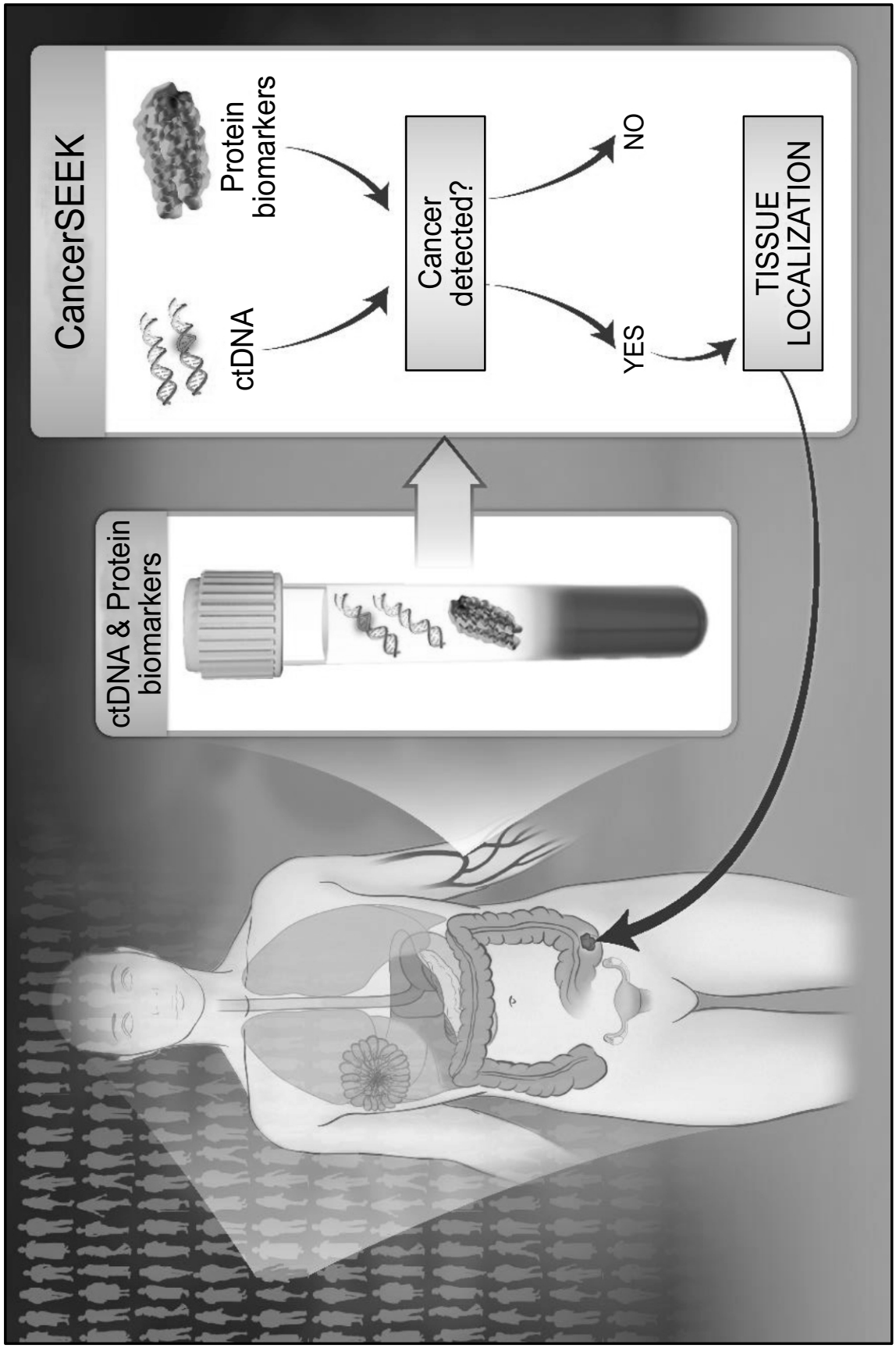

CancerSEEK
Protein biomarkers
ctDNA
Cancer detected?
NO
YES
TISSUE LOCALIZATION
ctDNA & Protein biomarkers

FIG. 1

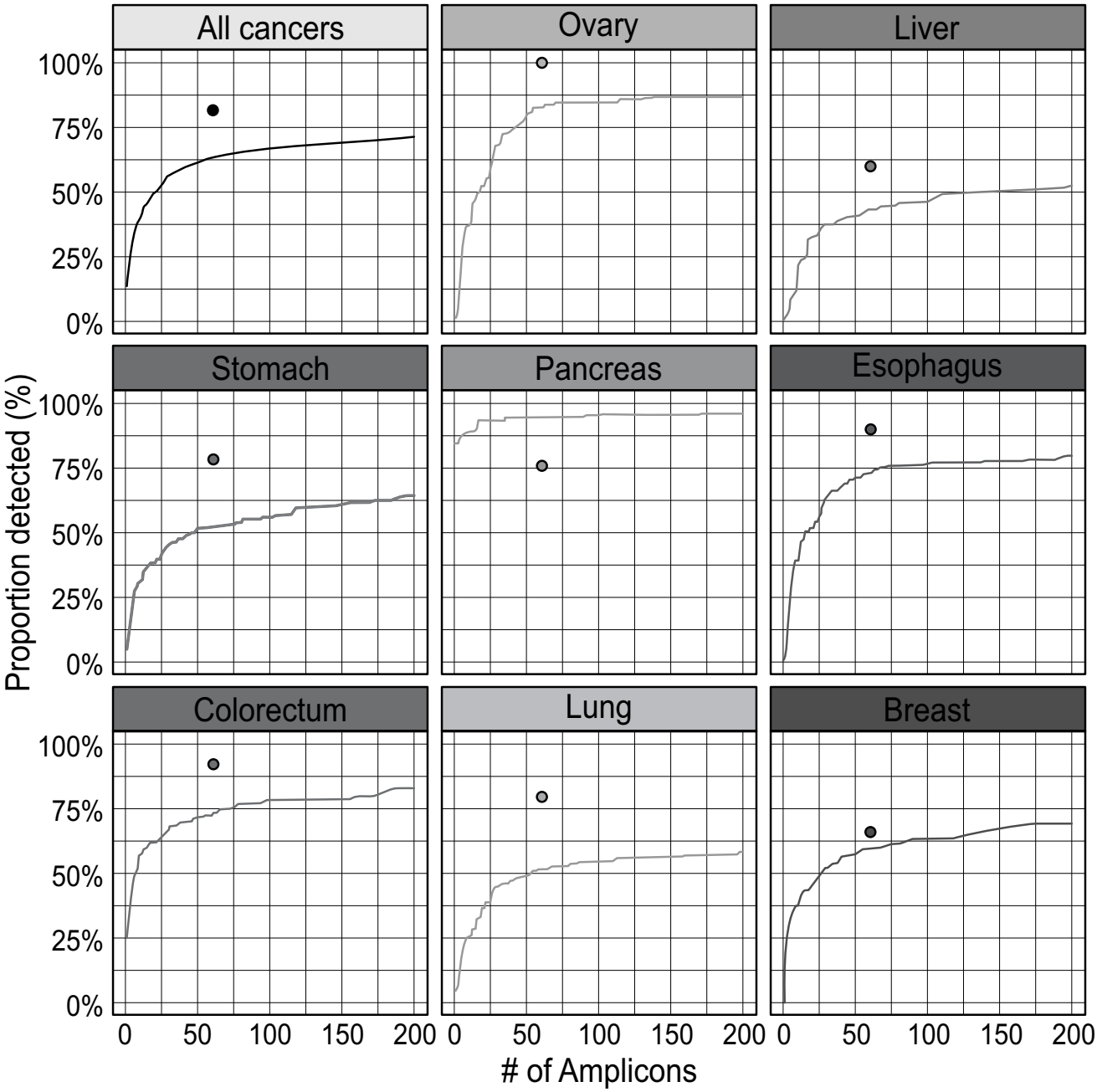

All cancers
Ovary
Liver
Stomach
Pancreas
Esophagus
Colorectum
Lung
Breast
Proportion detected (%)
100%
75%
50%
25%
0%
0
50
100
150
200
# of Amplicons

FIG. 2

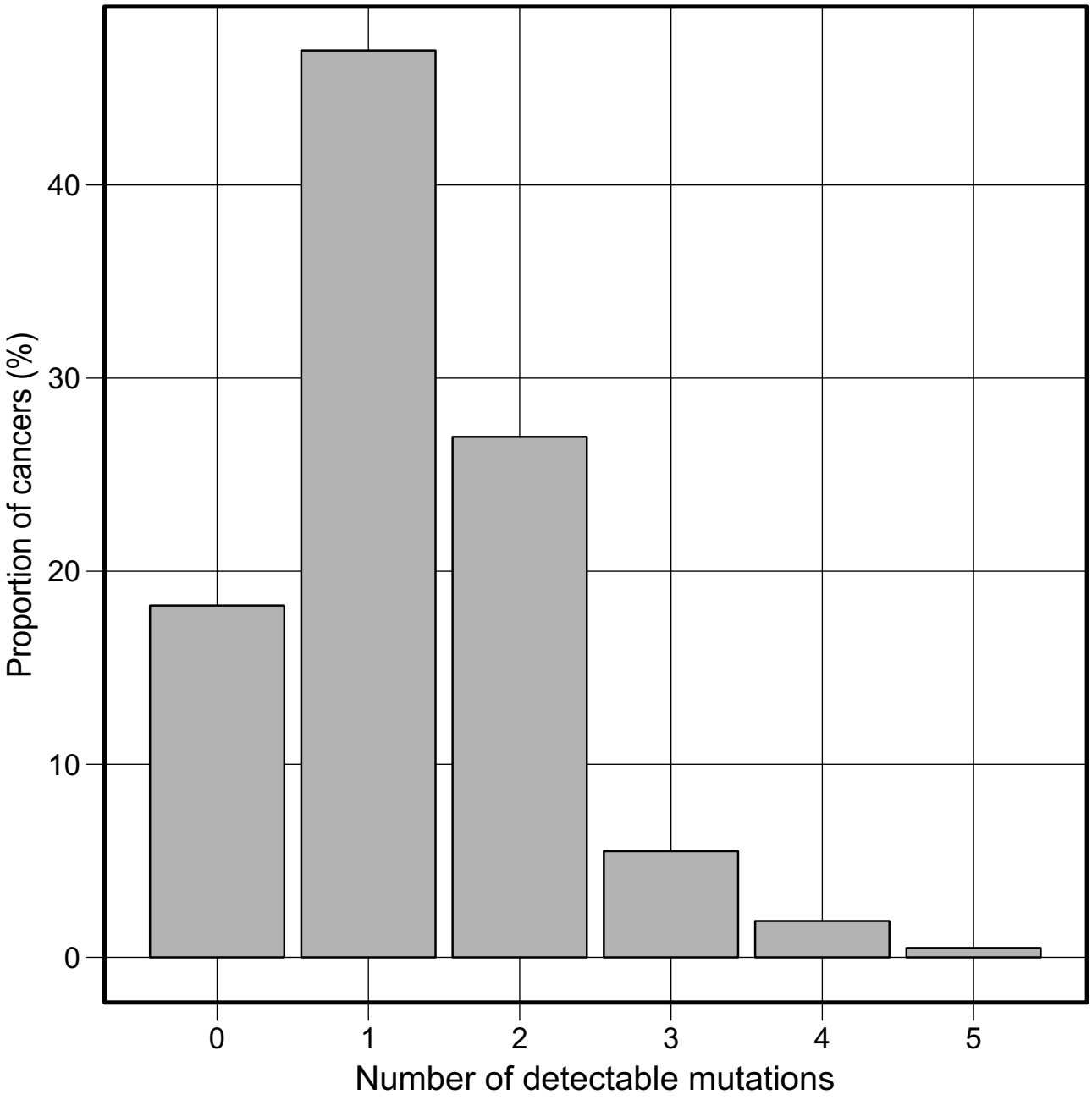

Proportion of cancers (%)
0
10
20
30
40
0
1
2
3
4
5
Number of detectable mutations

FIG. 3

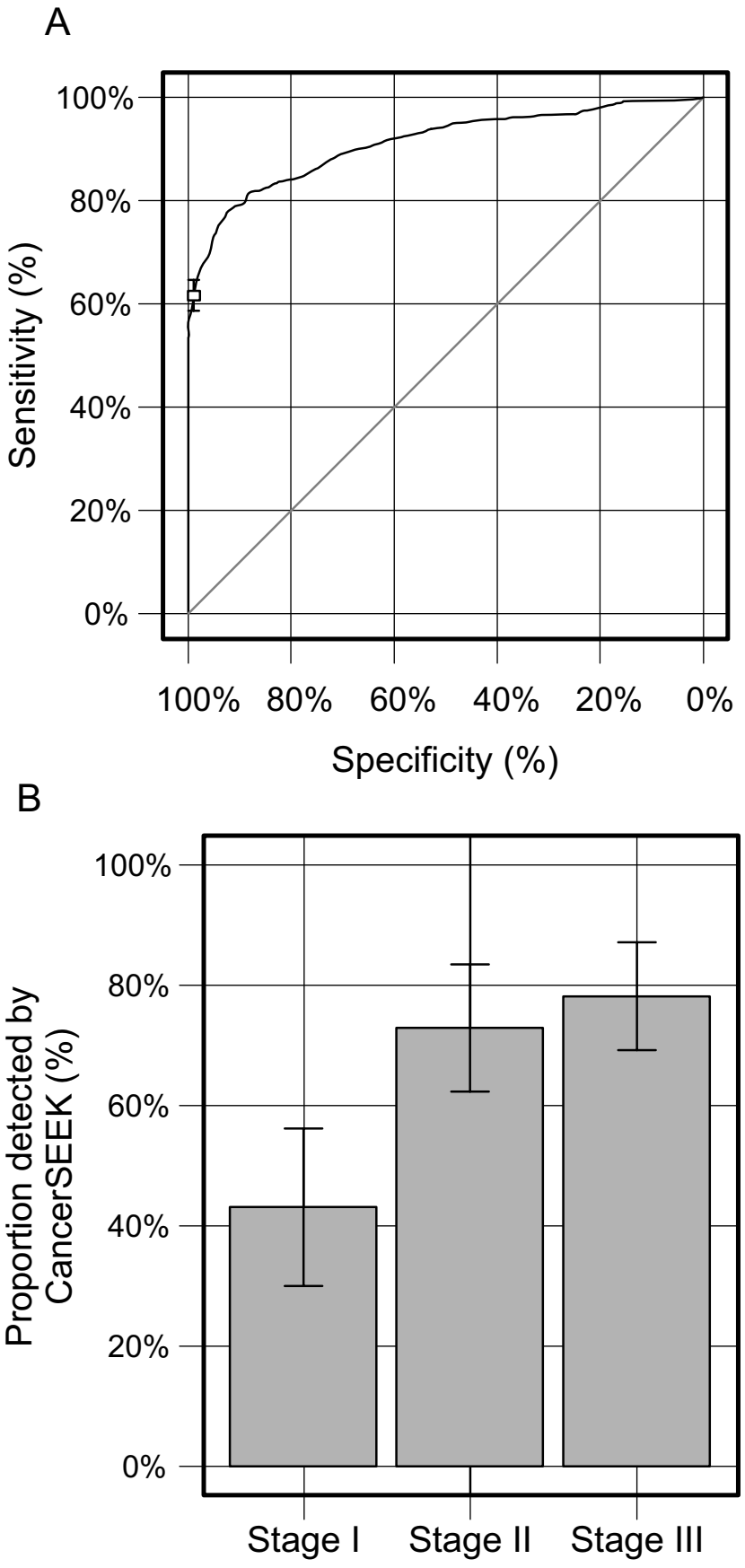

A
100%
80%
60%
40%
20%
0%
Sensitivity (%)
100%
80%
60%
40%
20%
0%
Specificity (%)
B
100%
80%
60%
40%
20%
0%
Proportion detected by CancerSEEK (%)
Stage I
Stage II
Stage III

FIG. 4

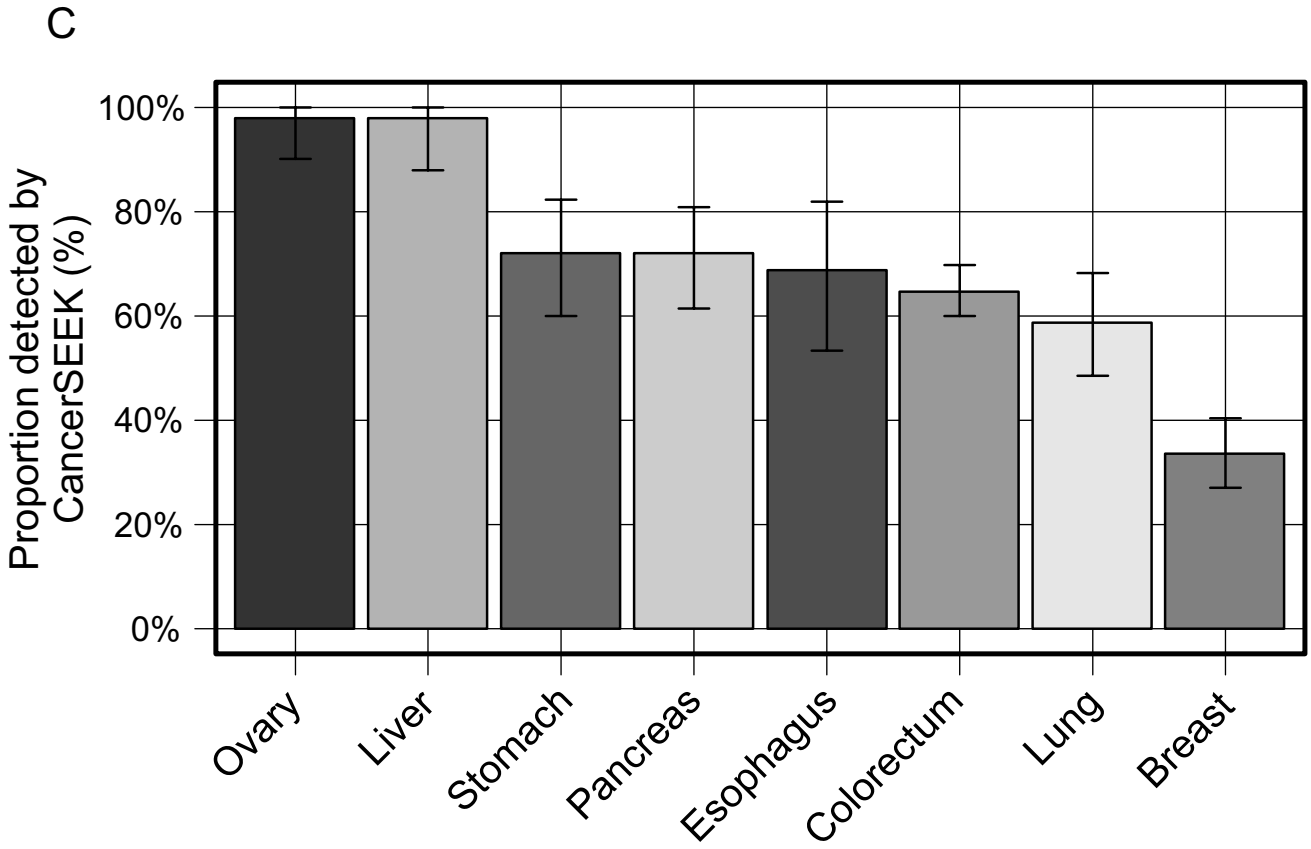

C
Proportion detected by CancerSEEK (%)
100%
80%
60%
40%
20%
0%
Ovary
Liver
Stomach
Pancreas
Esophagus
Colorectum
Lung
Breast

FIG. 4 (Cont.)

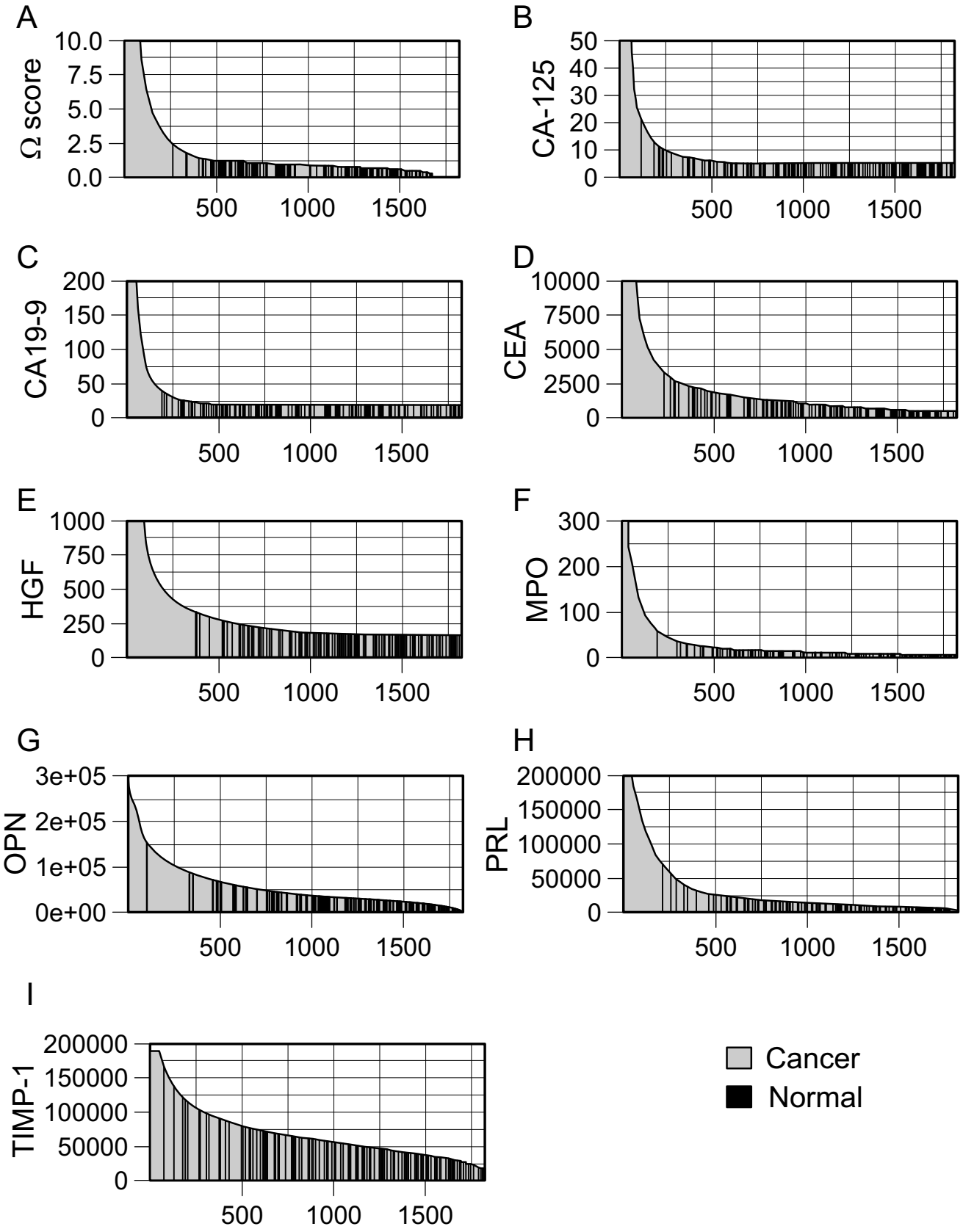

A
Ω score
10.0
7.5
5.0
2.5
0.0
500
1000
1500
B
CA-125
50
40
30
20
10
0
C
CA19-9
200
150
100
50
D
CEA
10000
7500
5000
2500
E
HGF
1000
750
500
250
F
MPO
300
200
100
G
OPN
3e+05
2e+05
1e+05
0e+00
H
PRL
200000
150000
100000
50000
I
TIMP-1
Cancer
Normal

FIG. 5

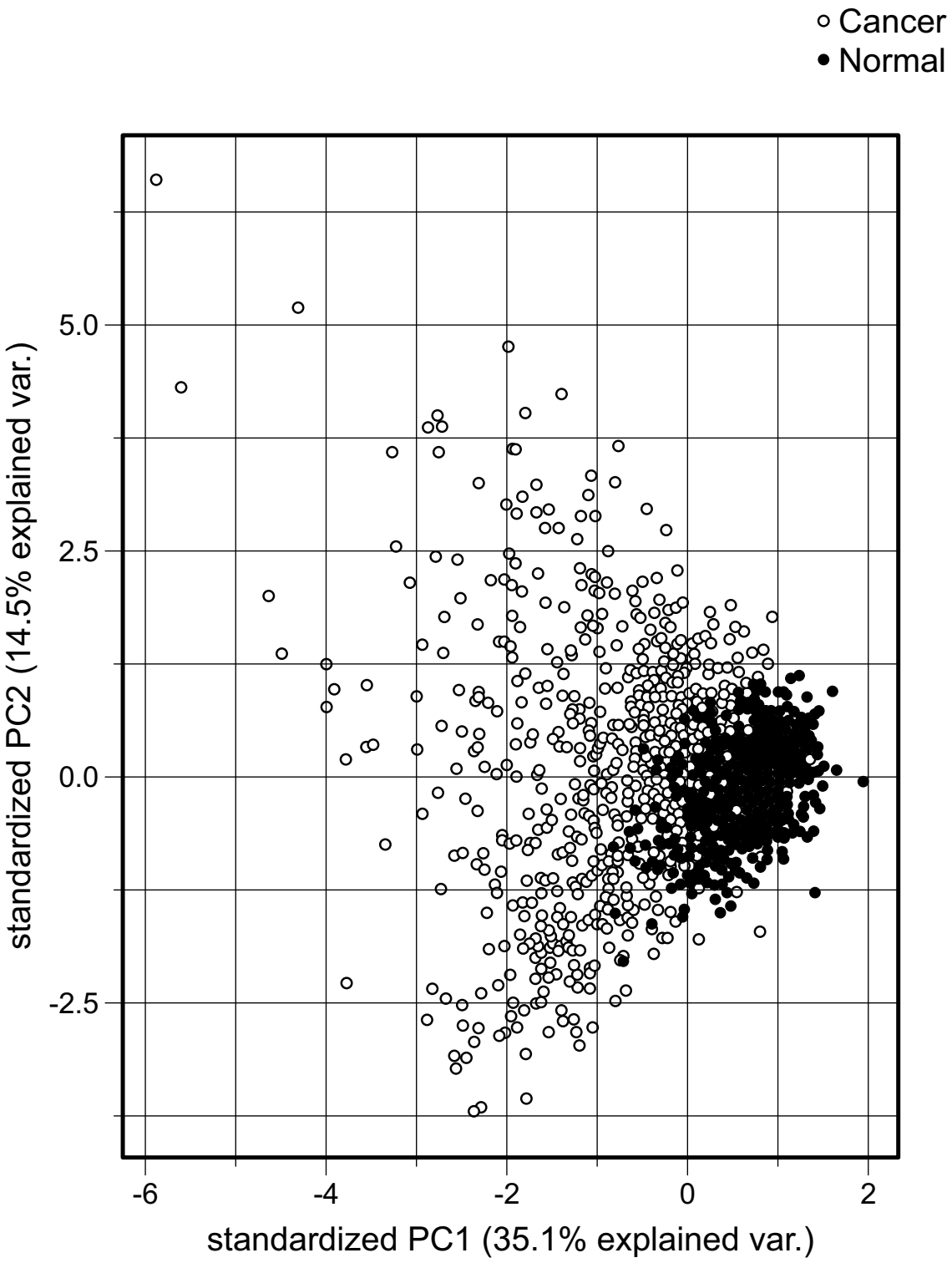

Cancer
Normal
5.0
2.5
0.0
-2.5
-6
-4
-2
0
2
standardized PC2 (14.5% explained var.)
standardized PC1 (35.1% explained var.)

FIG. 6

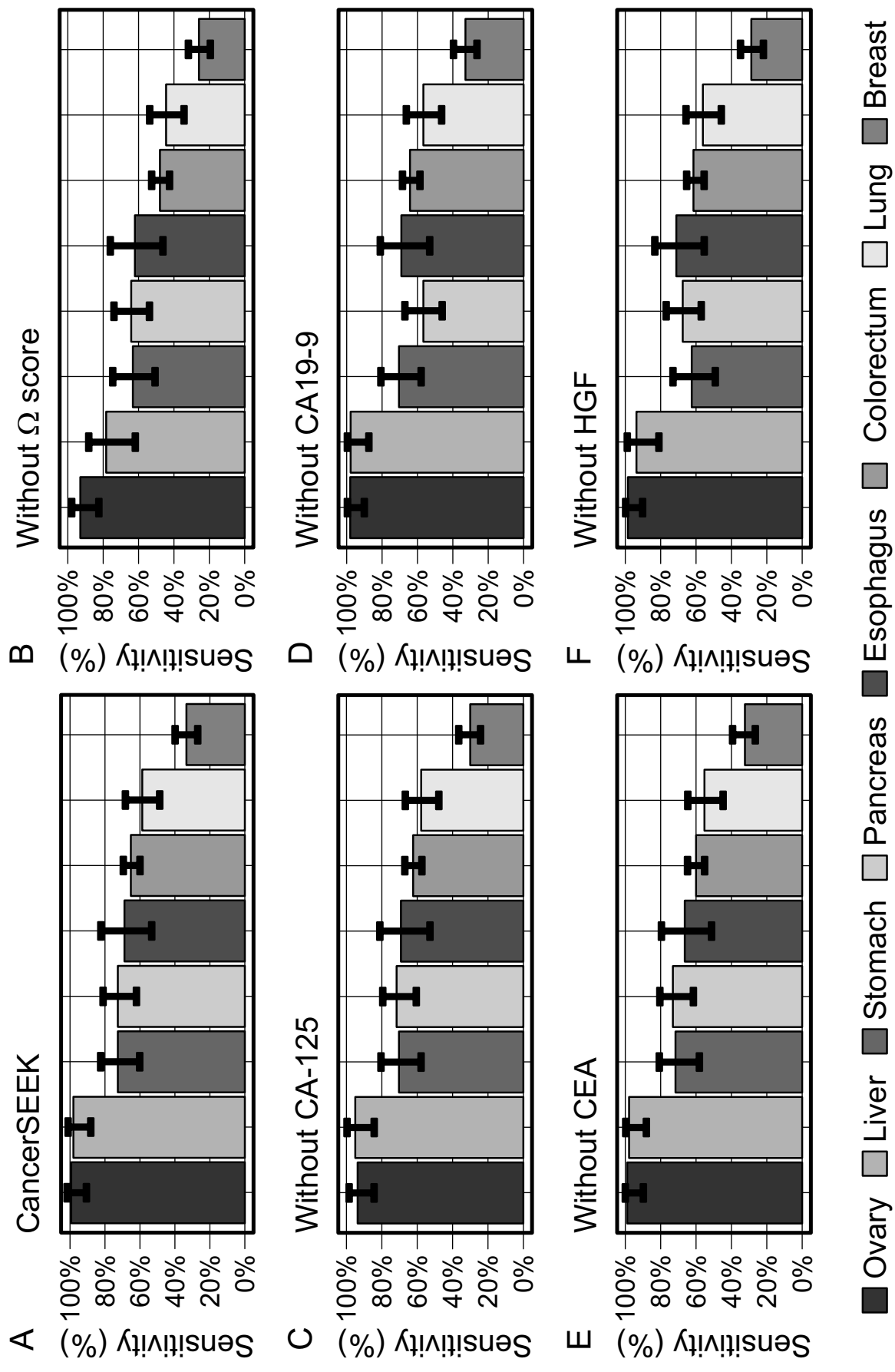

A
CancerSEEK
B
Without Ω score
C
Without CA-125
D
Without CA19-9
E
Without CEA
F
Without HGF
Sensitivity (%)
100%
80%
60%
40%
20%
0%
Ovary
Liver
Stomach
Pancreas
Esophagus
Colorectum
Lung
Breast

FIG. 7

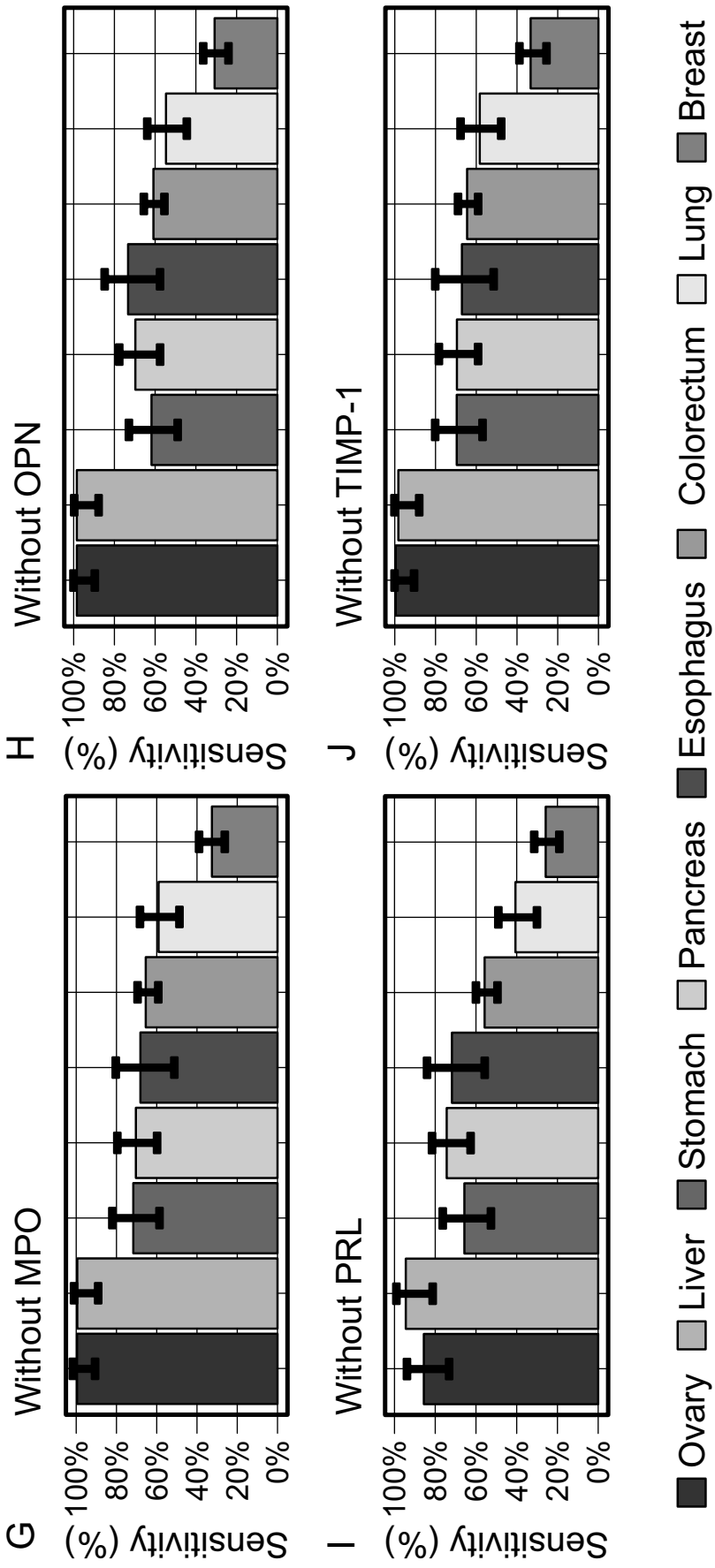

G
Without MPO
Sensitivity (%)
100%
80%
60%
40%
20%
0%
H
Without OPN
I
Without PRL
J
Without TIMP-1
Ovary
Liver
Stomach
Pancreas
Esophagus
Colorectum
Lung
Breast

FIG. 7 (Cont.)

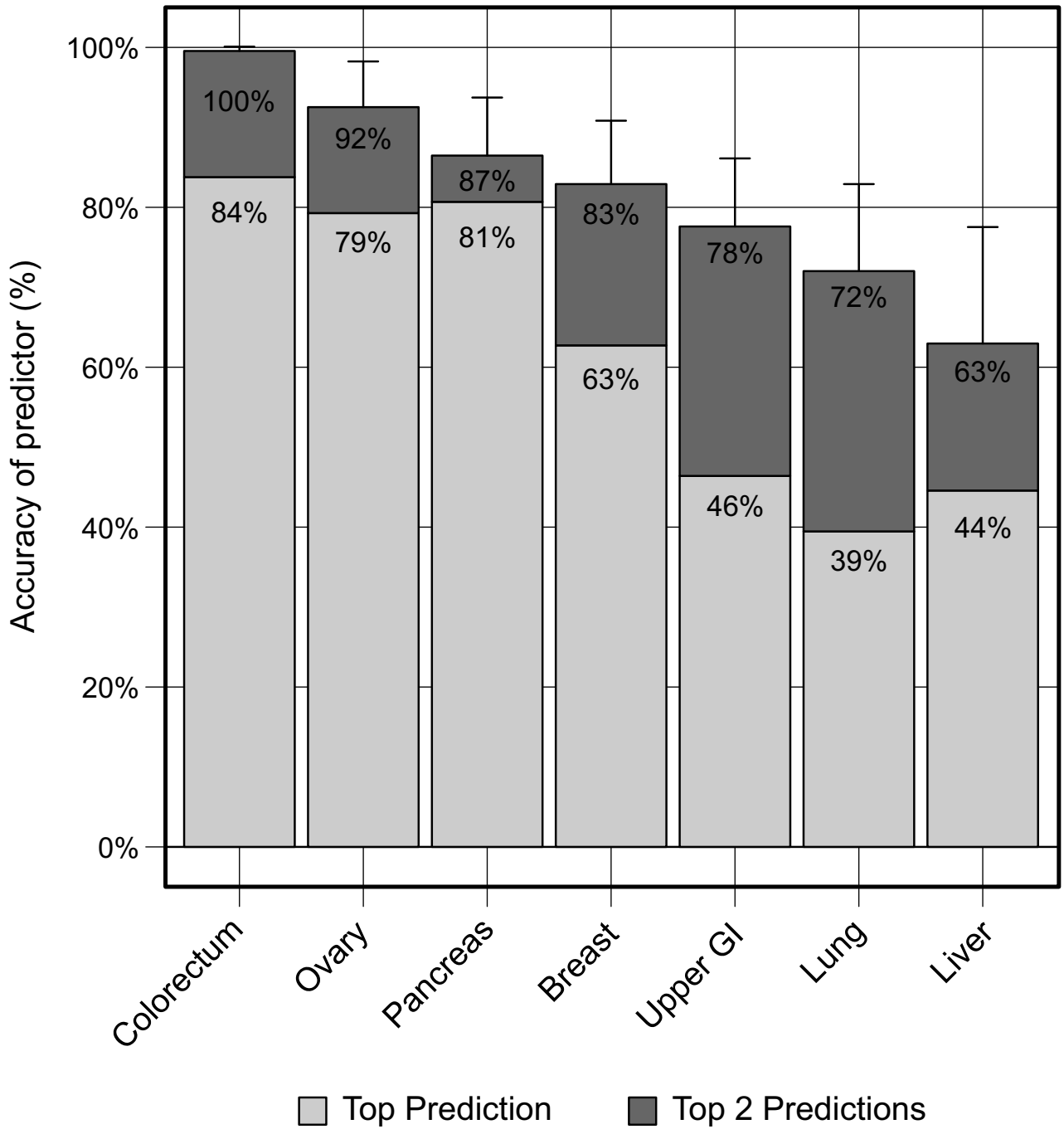

100%
80%
60%
40%
20%
0%
Accuracy of predictor (%)
100%
84%
92%
79%
87%
81%
83%
63%
78%
46%
72%
39%
63%
44%
Colorectum
Ovary
Pancreas
Breast
Upper GI
Lung
Liver
Top Prediction
Top 2 Predictions

FIG. 8

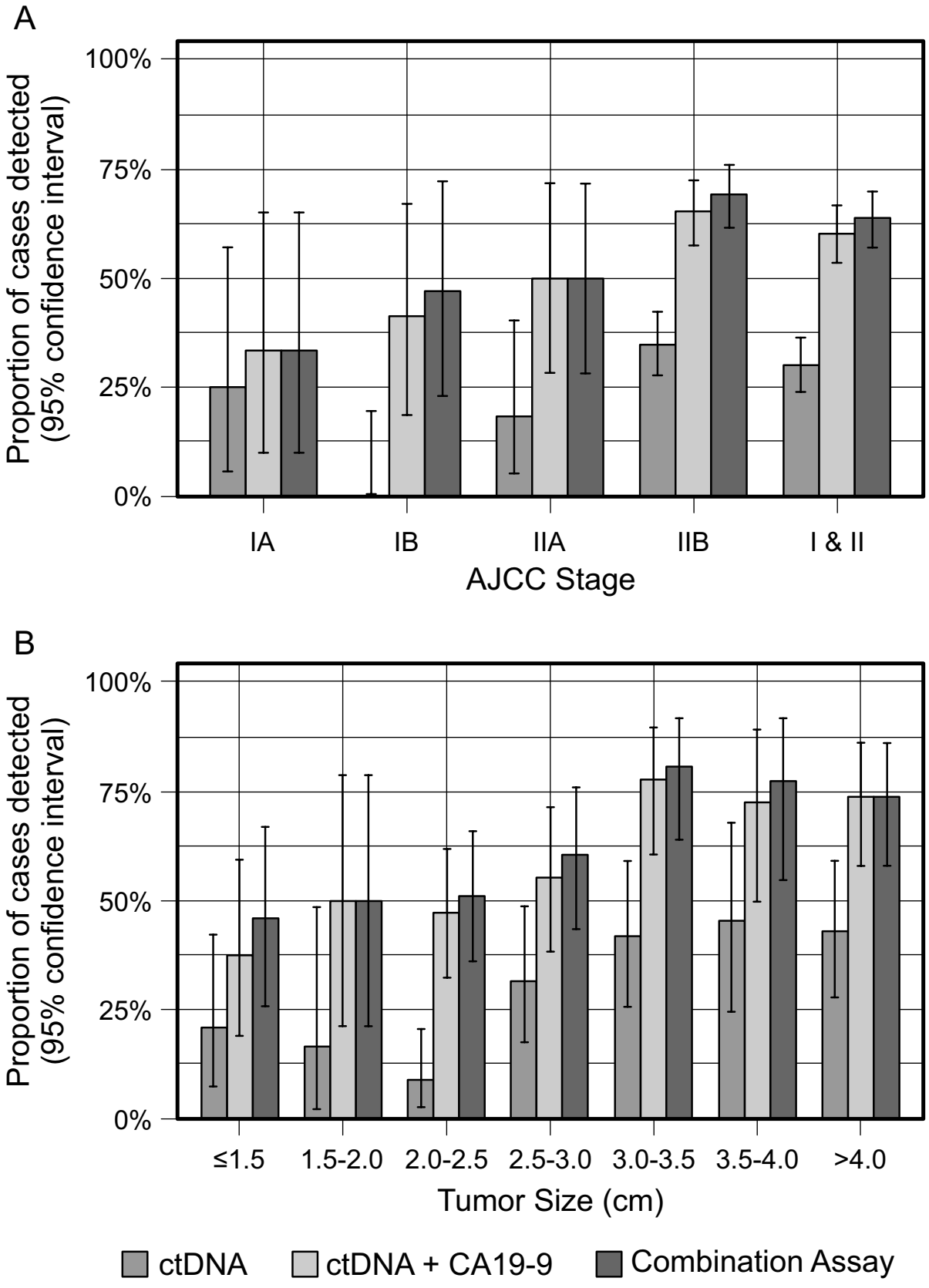

A
Proportion of cases detected (95% confidence interval)
100%
75%
50%
25%
0%
IA
IB
IIA
IIB
I & II
AJCC Stage
B
Proportion of cases detected (95% confidence interval)
100%
75%
50%
25%
0%
≤1.5
1.5-2.0
2.0-2.5
2.5-3.0
3.0-3.5
3.5-4.0
>4.0
Tumor Size (cm)
ctDNA
ctDNA + CA19-9
Combination Assay

FIG. 9

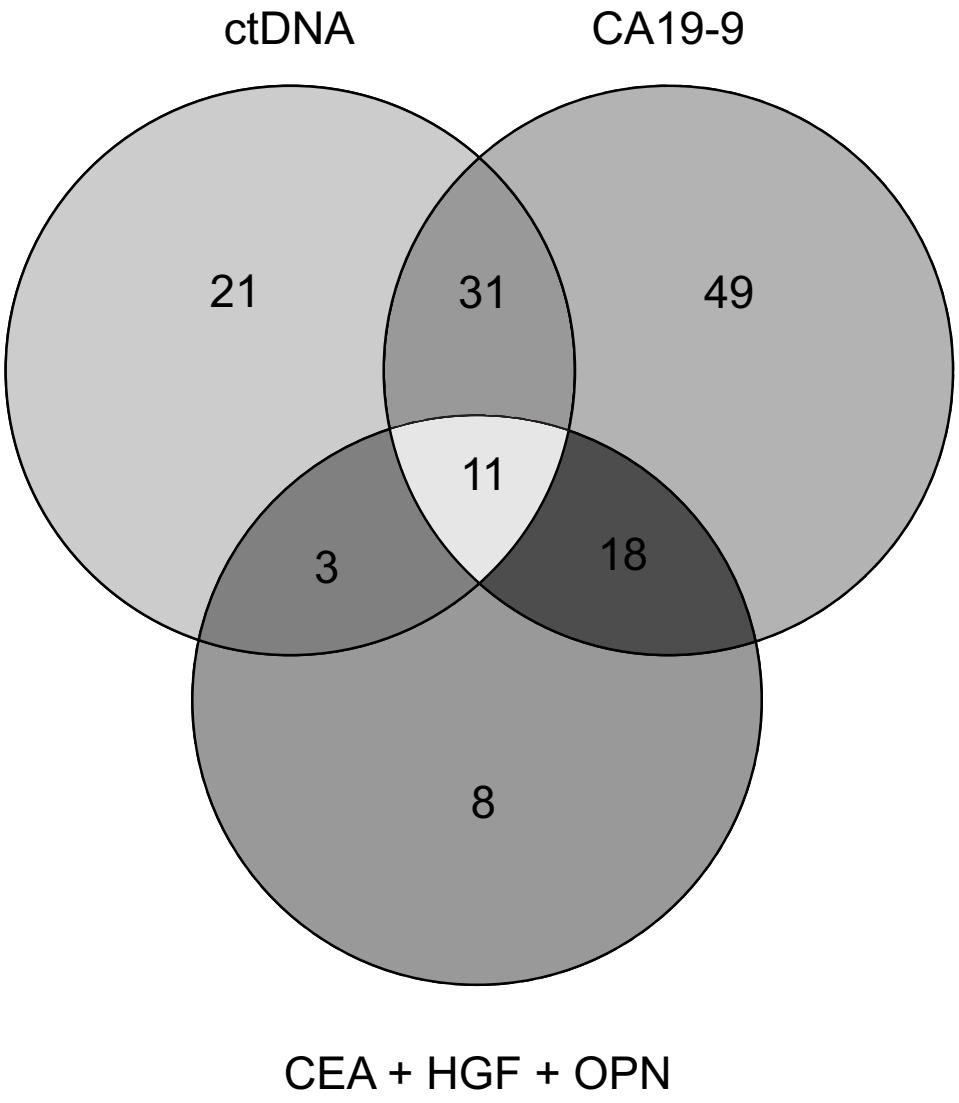


FIG. 10

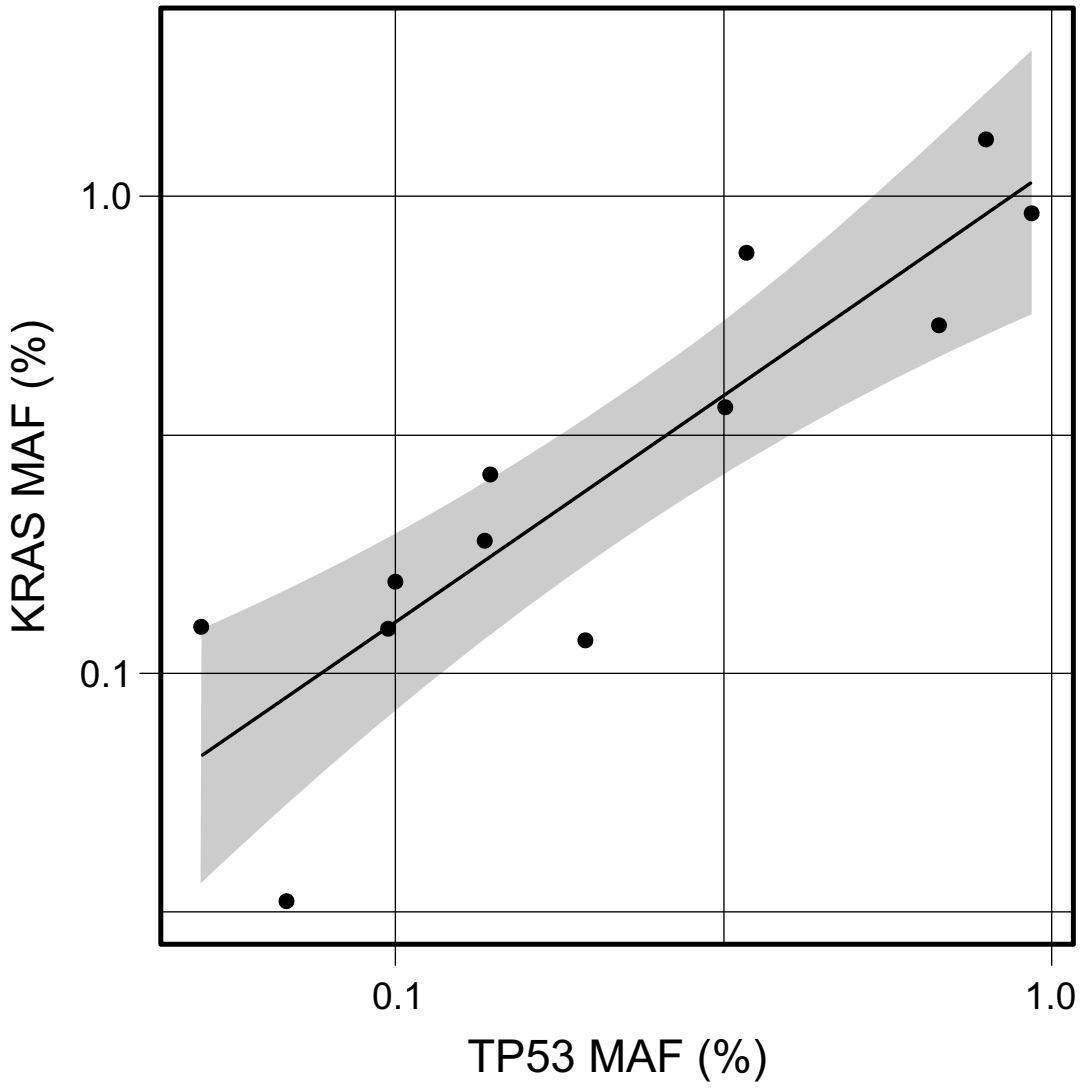


FIG. 11

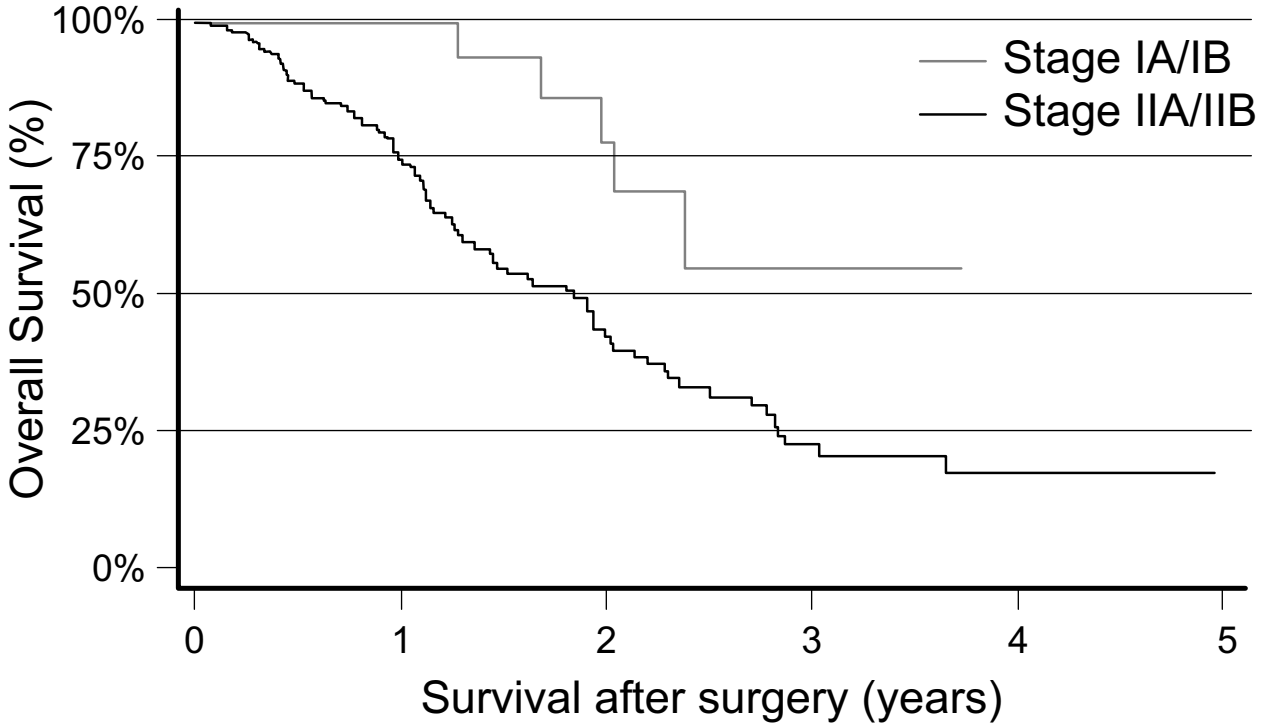

100%
75%
50%
25%
0%
Overall Survival (%)
0
1
2
3
4
5
Survival after surgery (years)
Stage IA/IB
Stage IIA/IIB

FIG. 12

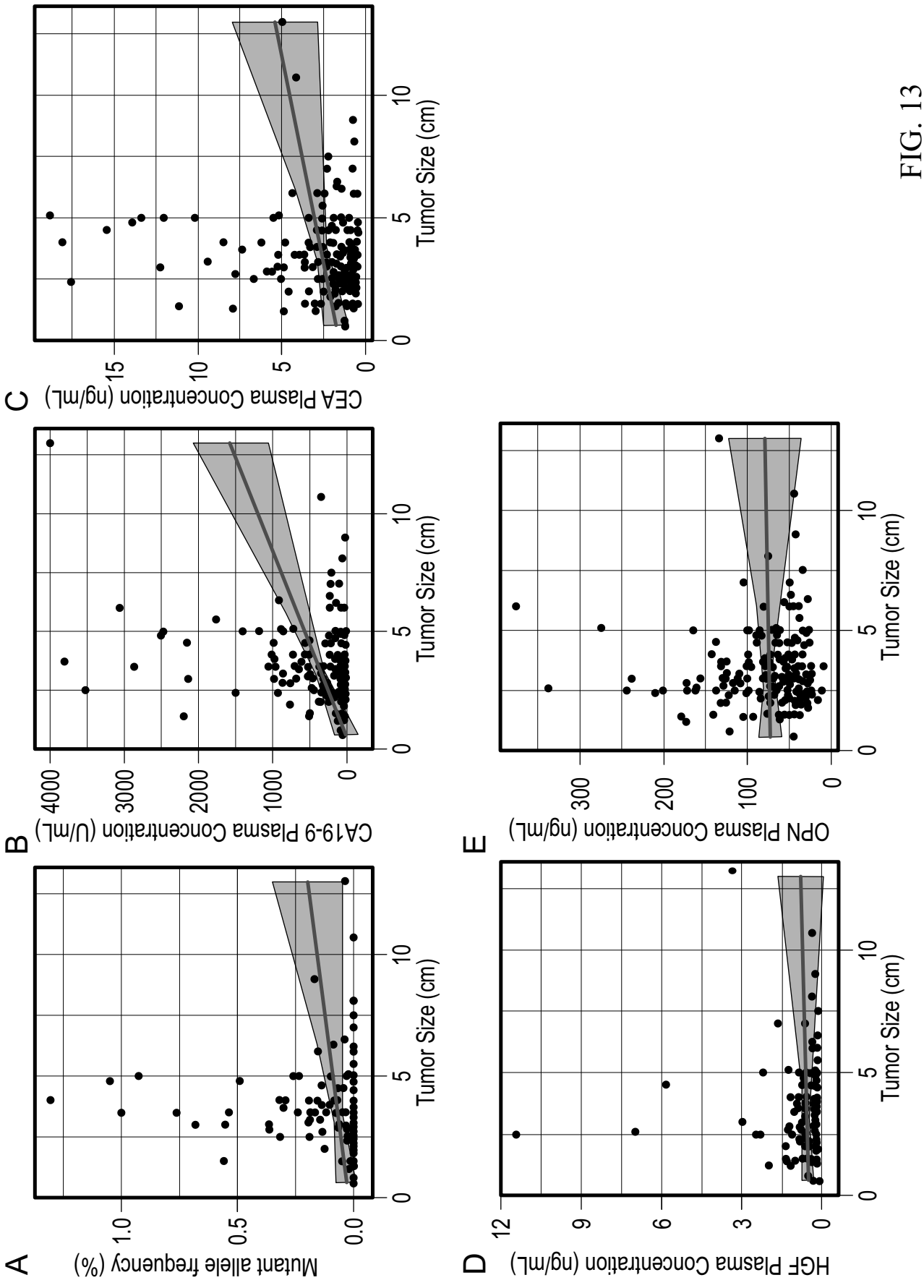

A
Mutant allele frequency (%)
0.0
0.5
1.0
0
5
10
Tumor Size (cm)
B
CA19-9 Plasma Concentration (U/mL)
0
1000
2000
3000
4000
0
5
10
Tumor Size (cm)
C
CEA Plasma Concentration (ng/mL)
0
5
10
15
0
5
10
Tumor Size (cm)
D
HGF Plasma Concentration (ng/mL)
0
3
6
9
12
0
5
10
Tumor Size (cm)
E
OPN Plasma Concentration (ng/mL)
0
100
200
300
0
5
10
Tumor Size (cm)

FIG. 13

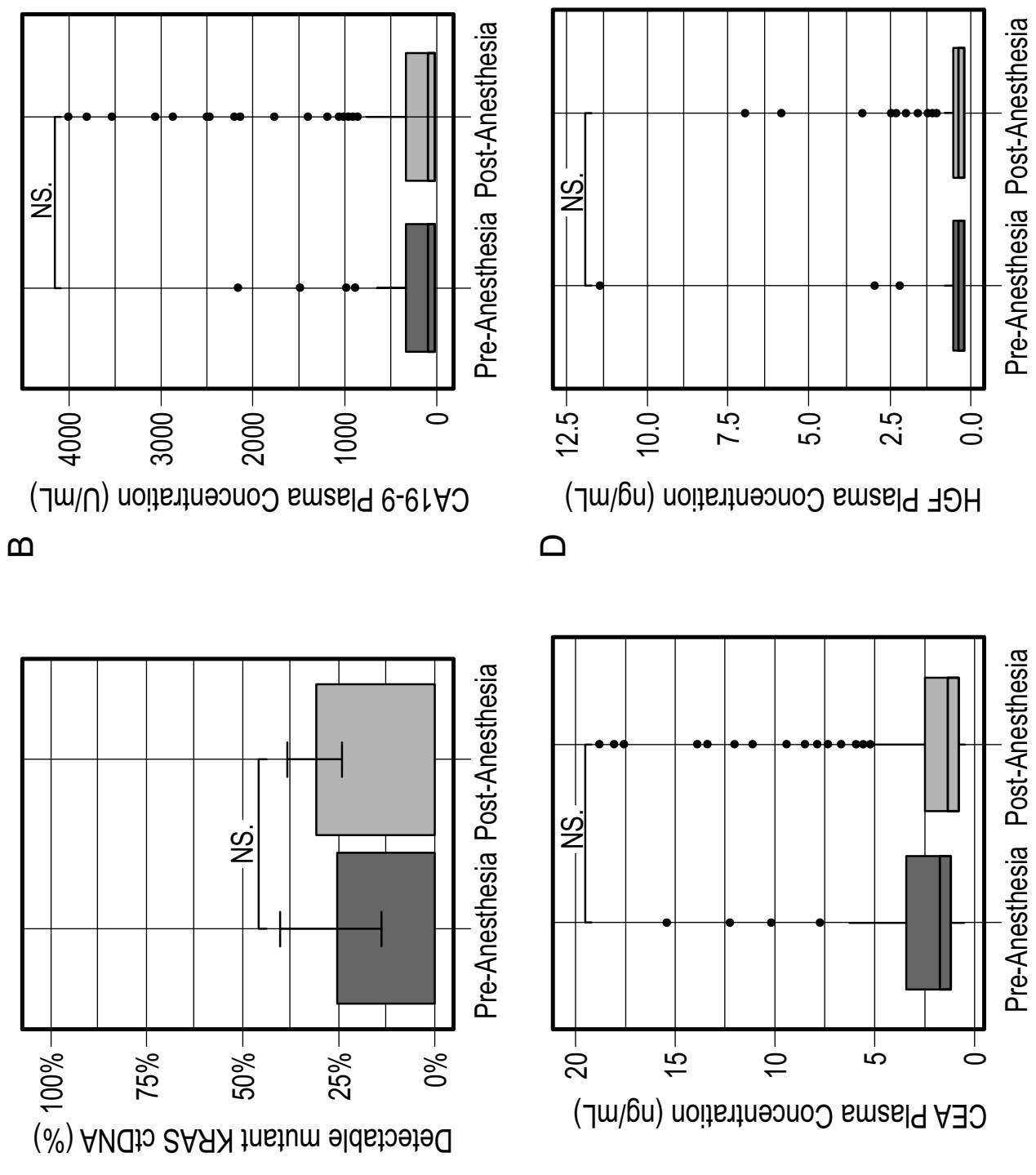
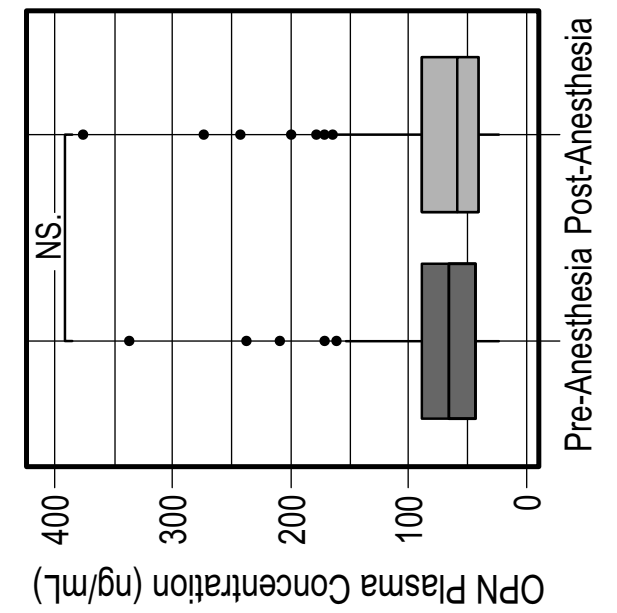
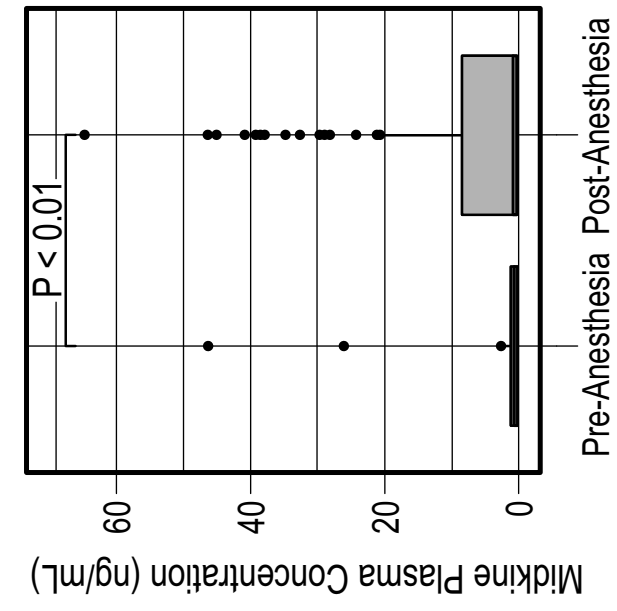

A
Detectable mutant KRAS ctDNA (%)
100%
75%
50%
25%
0%
NS.
Pre-Anesthesia
Post-Anesthesia
B
CA19-9 Plasma Concentration (U/mL)
4000
3000
2000
1000
0
NS.
Pre-Anesthesia
Post-Anesthesia
C
CEA Plasma Concentration (ng/mL)
20
15
10
5
0
NS.
Pre-Anesthesia
Post-Anesthesia
D
HGF Plasma Concentration (ng/mL)
12.5
10.0
7.5
5.0
2.5
0.0
NS.
Pre-Anesthesia
Post-Anesthesia

FIG. 14



F



E

G

FIG. 14 (Cont.)

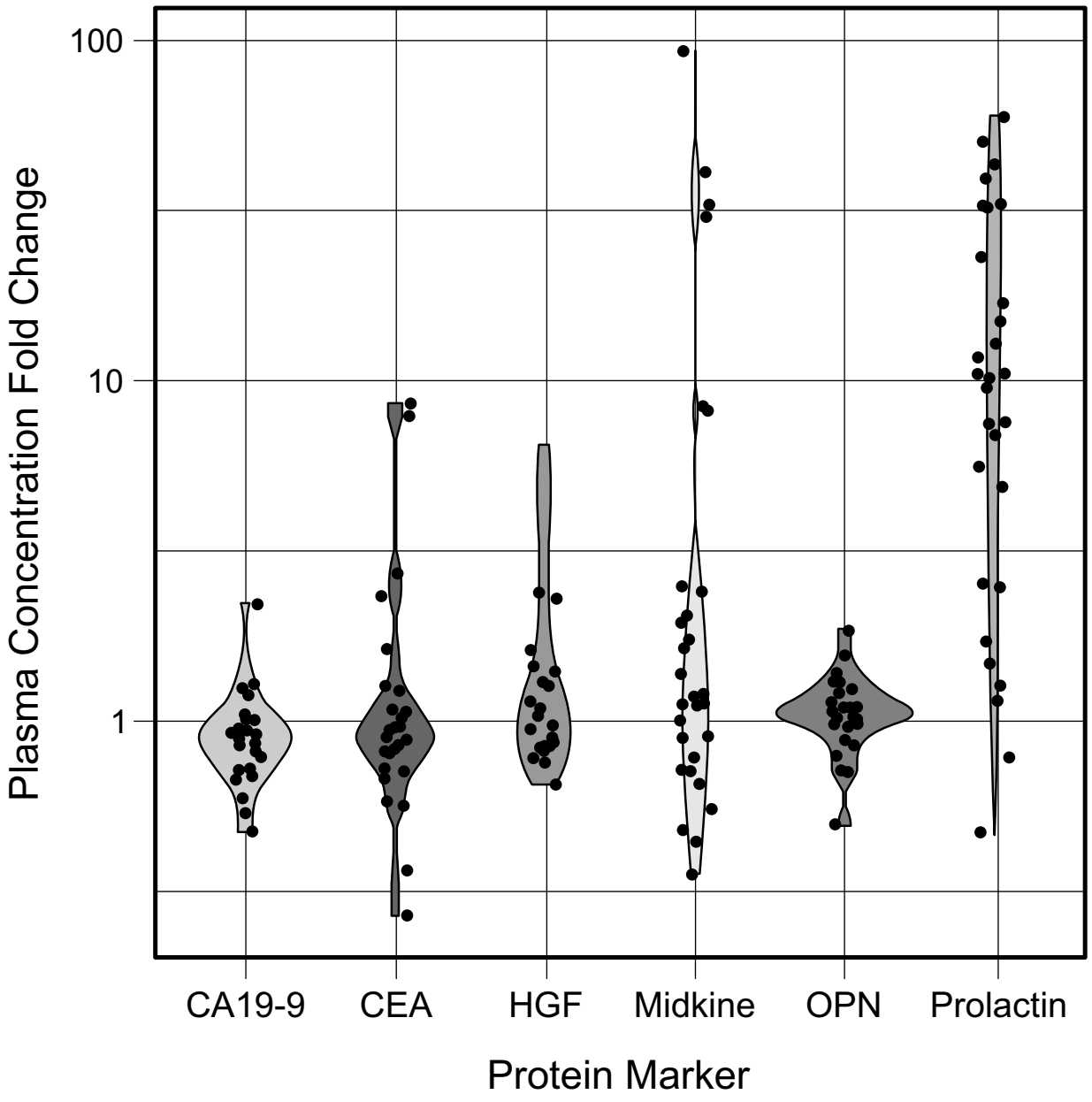

100
10
1
Plasma Concentration Fold Change
CA19-9
CEA
HGF
Midkine
OPN
Prolactin
Protein Marker

FIG. 15

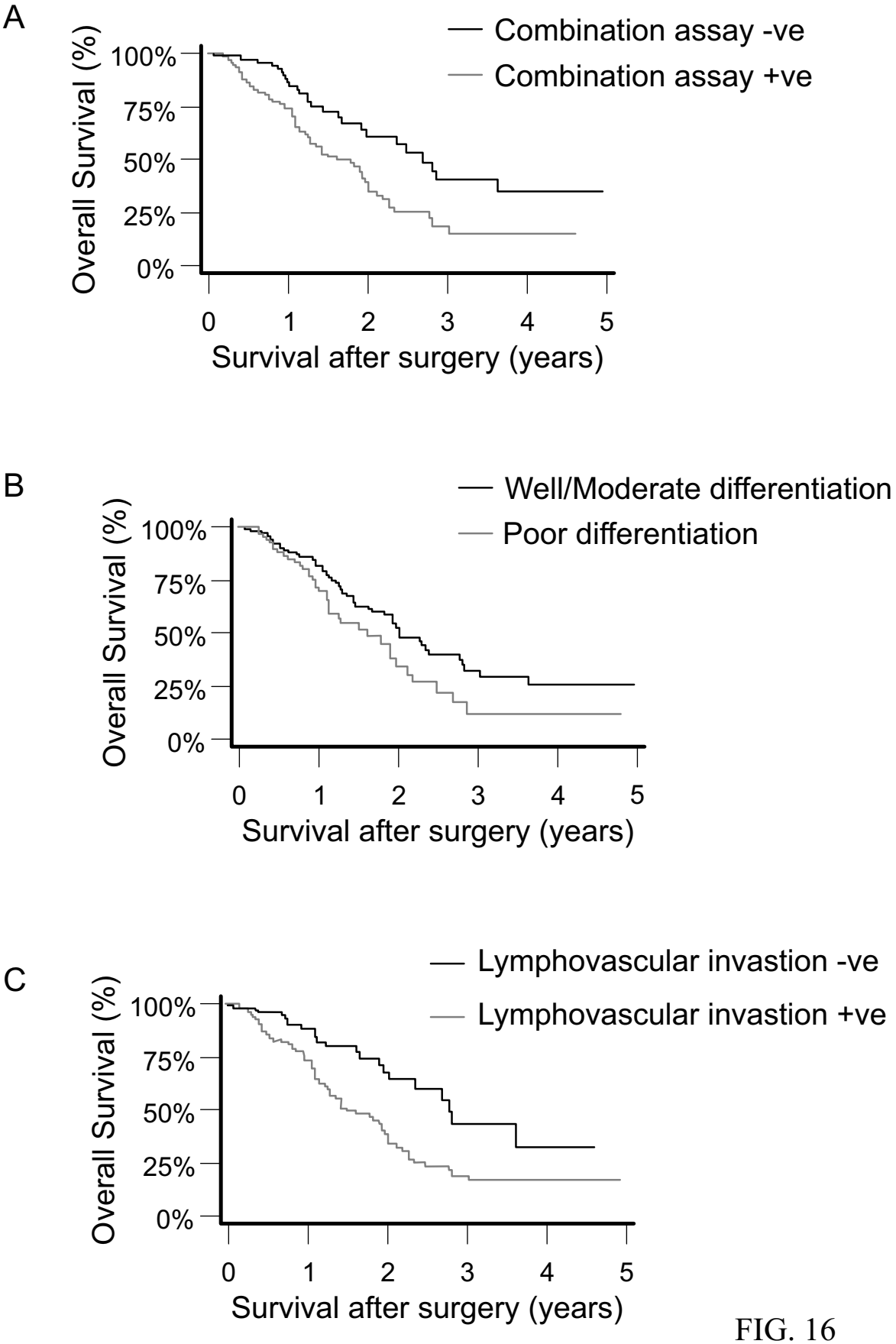
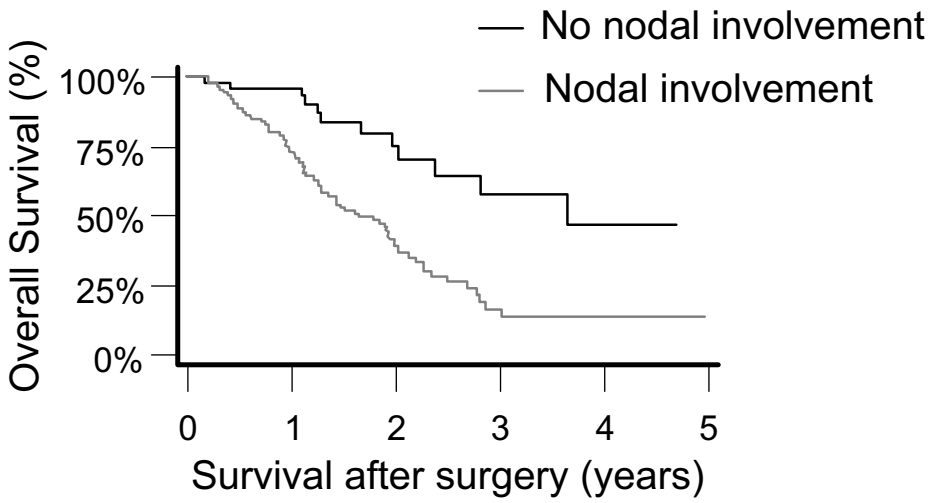

A
Combination assay -ve
Combination assay +ve
Overall Survival (%)
100%
75%
50%
25%
0%
0
1
2
3
4
5
Survival after surgery (years)
B
Well/Moderate differentiation
Poor differentiation
Overall Survival (%)
100%
75%
50%
25%
0%
0
1
2
3
4
5
Survival after surgery (years)
C
Lymphovascular invastion -ve
Lymphovascular invastion +ve
Overall Survival (%)
100%
75%
50%
25%
0%
0
1
2
3
4
5
Survival after surgery (years)

FIG. 16

D



E

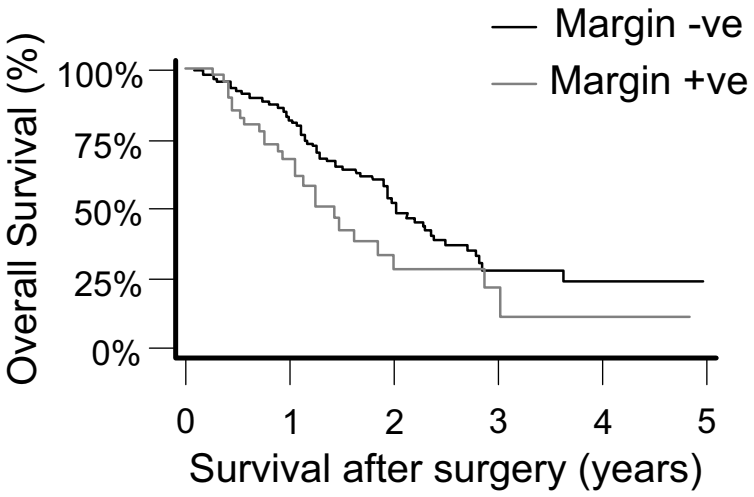


FIG. 16 (Cont.)

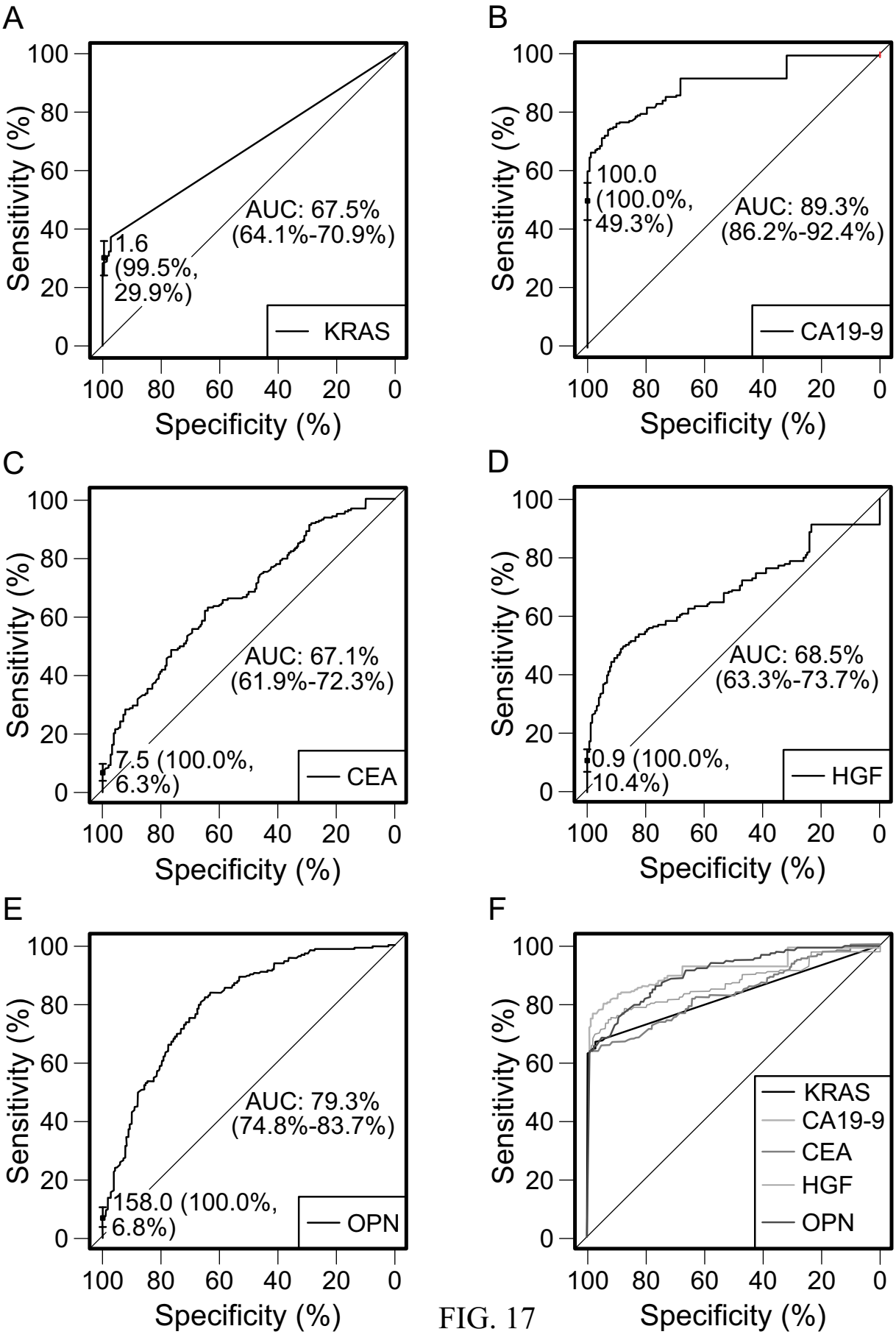

A
Sensitivity (%)
Specificity (%)
AUC: 67.5%
(64.1%-70.9%)
1.6
(99.5%,
29.9%)
KRAS
B
Sensitivity (%)
Specificity (%)
100.0
(100.0%,
49.3%)
AUC: 89.3%
(86.2%-92.4%)
CA19-9
C
Sensitivity (%)
Specificity (%)
AUC: 67.1%
(61.9%-72.3%)
7.5 (100.0%,
6.3%)
CEA
D
Sensitivity (%)
Specificity (%)
AUC: 68.5%
(63.3%-73.7%)
0.9 (100.0%,
10.4%)
HGF
E
Sensitivity (%)
Specificity (%)
AUC: 79.3%
(74.8%-83.7%)
158.0 (100.0%,
6.8%)
OPN
F
Sensitivity (%)
Specificity (%)
KRAS
CA19-9
CEA
HGF
OPN

FIG. 17

A

Tumor Type	SEER Incidence	Total #	# ctDNA Positive	# Protein Positive	# ctDNA + Protein Positive	% ctDNA Positive	% Protein Positive	% ctDNA + Protein Positive
Breast	43.8	150	24	30	47	16%	20%	31%
CRC	10	322	130	168	231	40%	52%	72%
Esophagus	0.63	43	21	28	34	49%	65%	79%
Gastric	1.34	65	20	38	39	31%	58%	60%
Liver	1.5	53	39	43	50	74%	81%	94%
Lung	3.31	109	43	30	61	39%	28%	56%
Ovarian	3.83	86	43	45	59	50%	52%	69%
Pancreas	3.8	412	91	248	271	22%	60%	66%
All Cancers	68.21	1240	411	630	792	33%	51%	64%
All Cancers (Incidence Weighted)	68.21	12239	3102.55	4341.11	5974.96	25%	35%	49%

FIG. 18

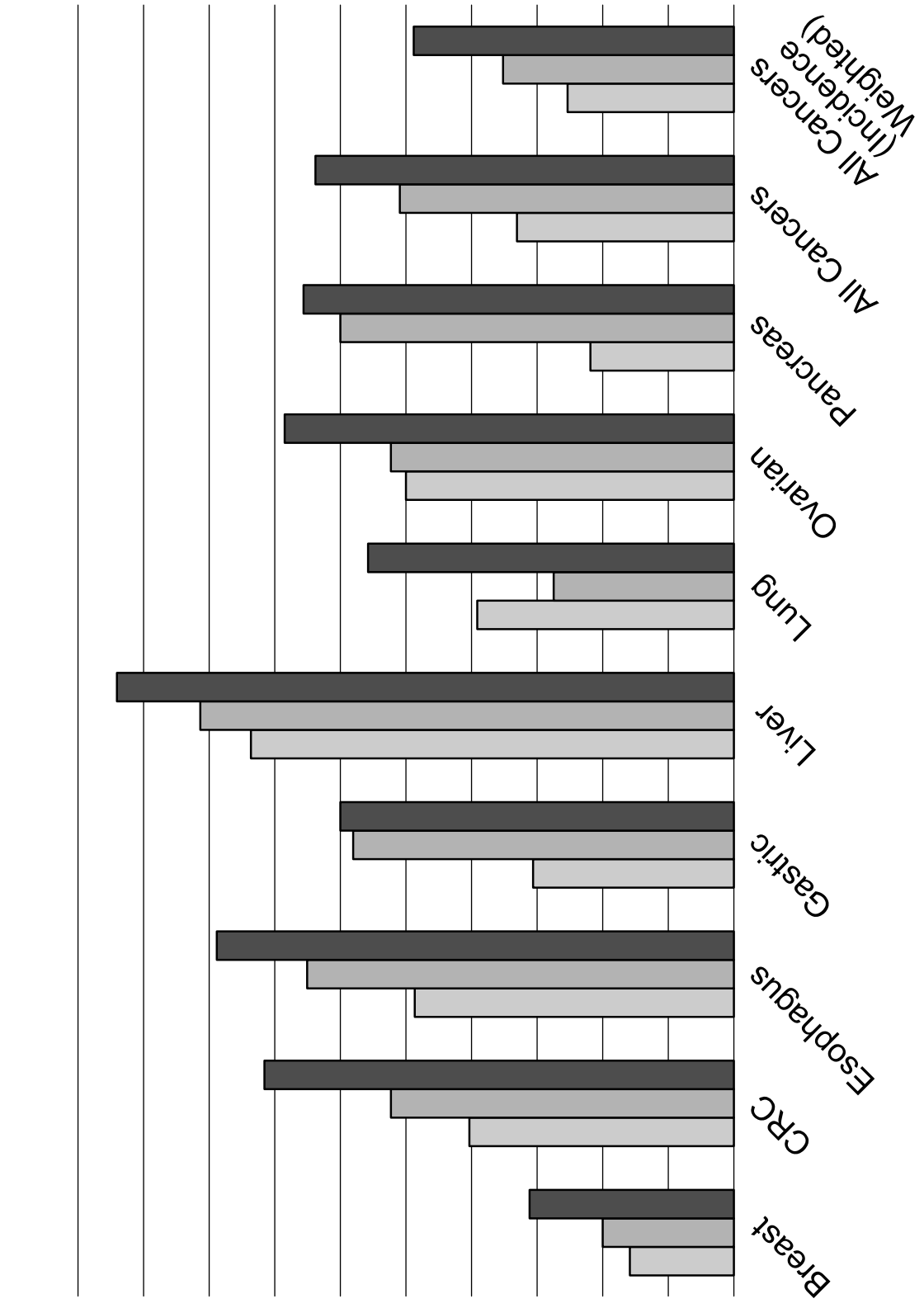

B
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Breast
CRC
Esophagus
Gastric
Liver
Lung
Ovarian
Pancreas
All Cancers
All Cancers (Incidence Weighted)
ctDNA
Protein
ctDNA + Protein

FIG. 18 (Cont.)

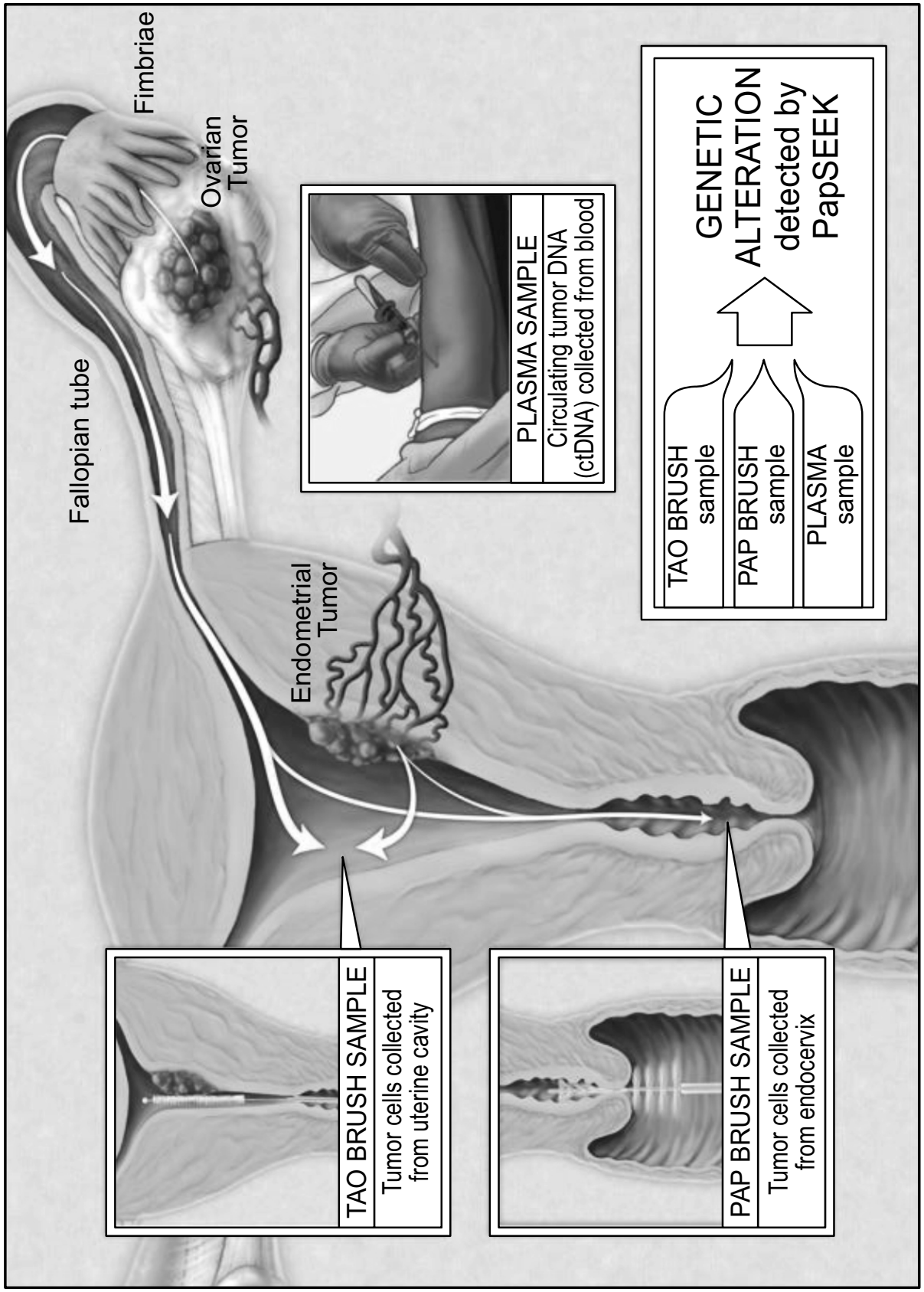

Fimbriae
Ovarian Tumor
Fallopian tube
Endometrial Tumor
TAO BRUSH SAMPLE
Tumor cells collected from uterine cavity
PAP BRUSH SAMPLE
Tumor cells collected from endocervix
PLASMA SAMPLE
Circulating tumor DNA (ctDNA) collected from blood
TAO BRUSH sample
PAP BRUSH sample
PLASMA sample
GENETIC ALTERATION detected by PapSEEK

FIG. 19

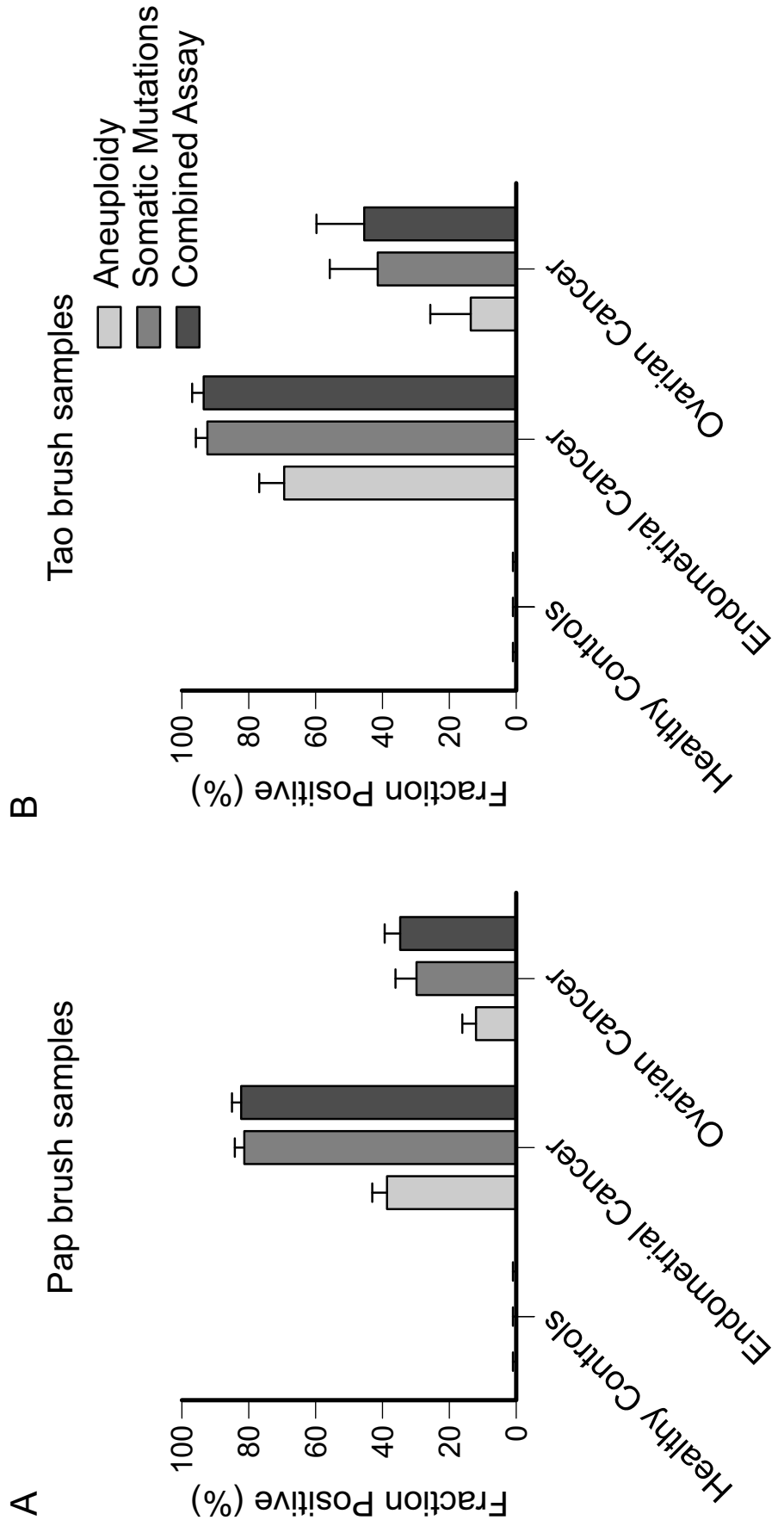
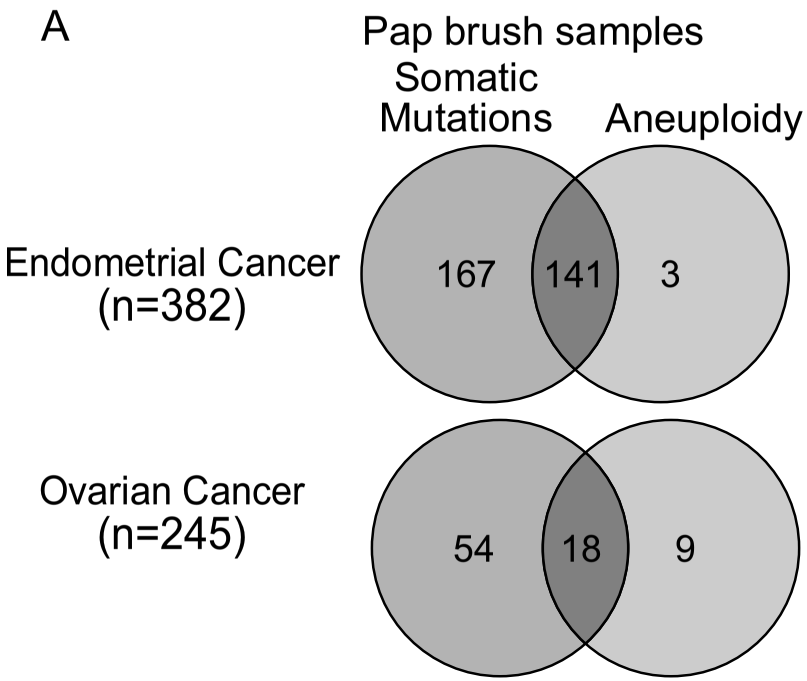
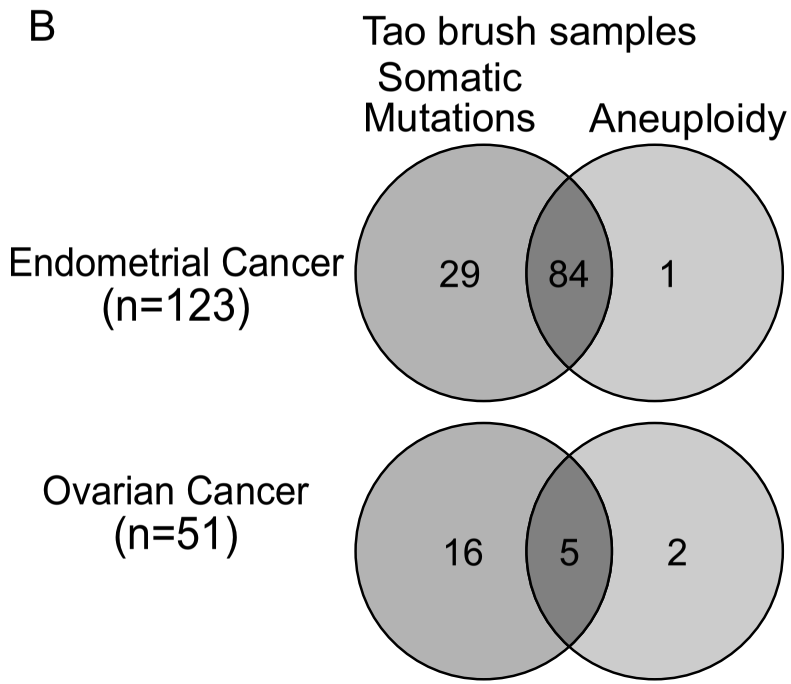

A
Pap brush samples
Fraction Positive (%)
0
20
40
60
80
100
Healthy Controls
Endometrial Cancer
Ovarian Cancer
B
Tao brush samples
Aneuploidy
Somatic Mutations
Combined Assay
Fraction Positive (%)
0
20
40
60
80
100
Healthy Controls
Endometrial Cancer
Ovarian Cancer

FIG. 20


A
Pap brush samples
Somatic Mutations
Aneuploidy
Endometrial Cancer (n=382)
167
141
3
Ovarian Cancer (n=245)
54
18
9


B
Tao brush samples
Somatic Mutations
Aneuploidy
Endometrial Cancer (n=123)
29
84
1
Ovarian Cancer (n=51)
16
5
2

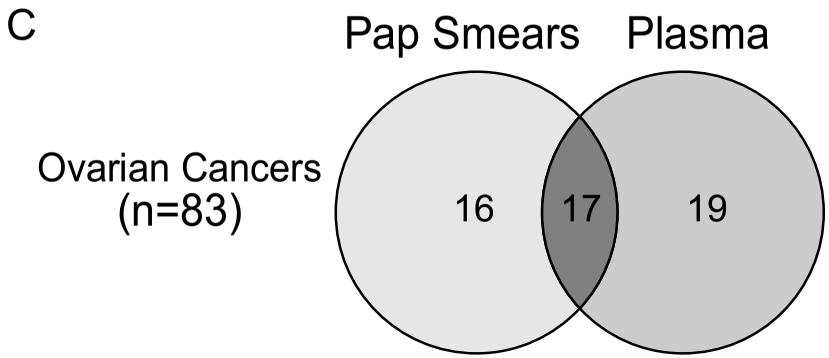

C
Pap Smears
Plasma
Ovarian Cancers (n=83)
16
17
19

FIG. 21

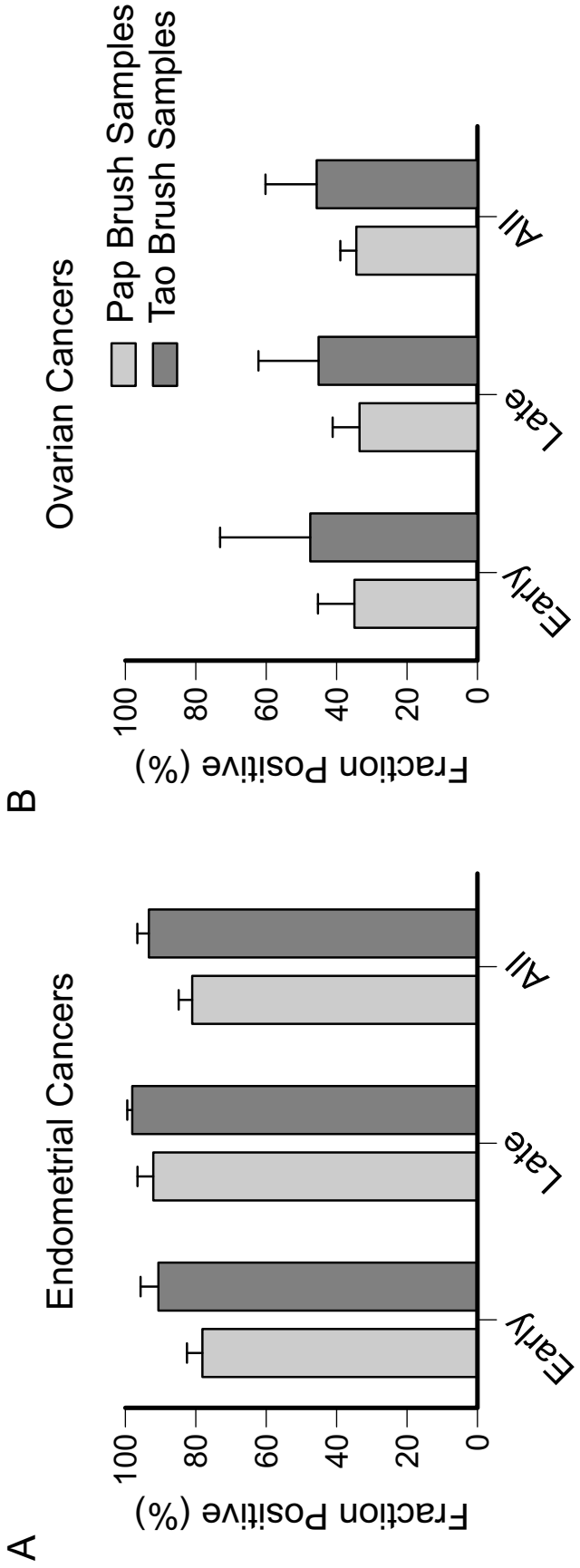

A
Endometrial Cancers
Fraction Positive (%)
100
80
60
40
20
0
Early
Late
All
B
Ovarian Cancers
Pap Brush Samples
Tao Brush Samples
Fraction Positive (%)
100
80
60
40
20
0
Early
Late
All

FIG. 22

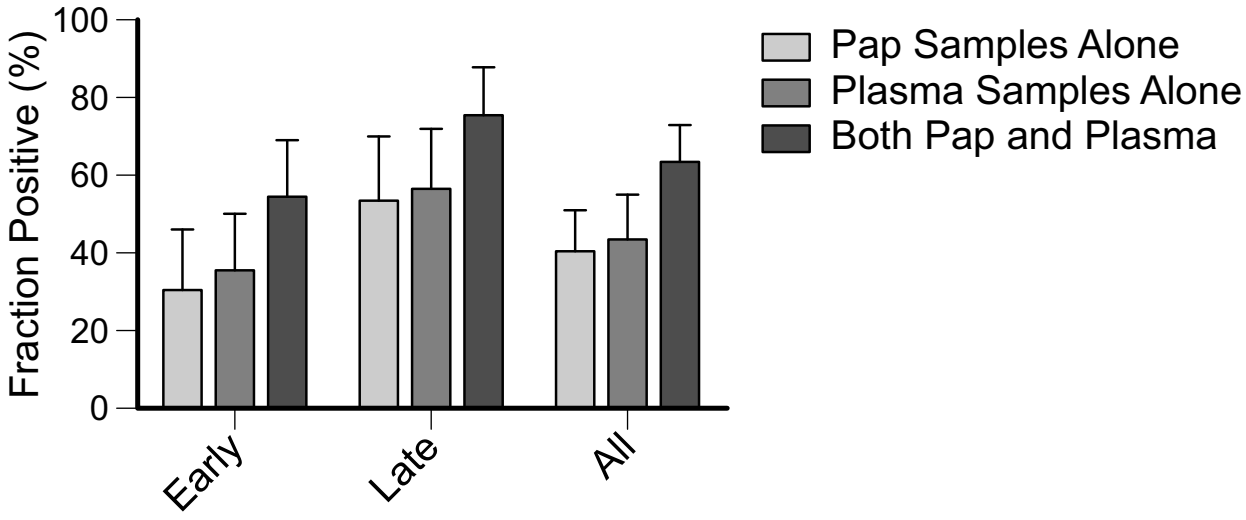

Fraction Positive (%)
100
80
60
40
20
0
Early
Late
All
Pap Samples Alone
Plasma Samples Alone
Both Pap and Plasma

FIG. 23

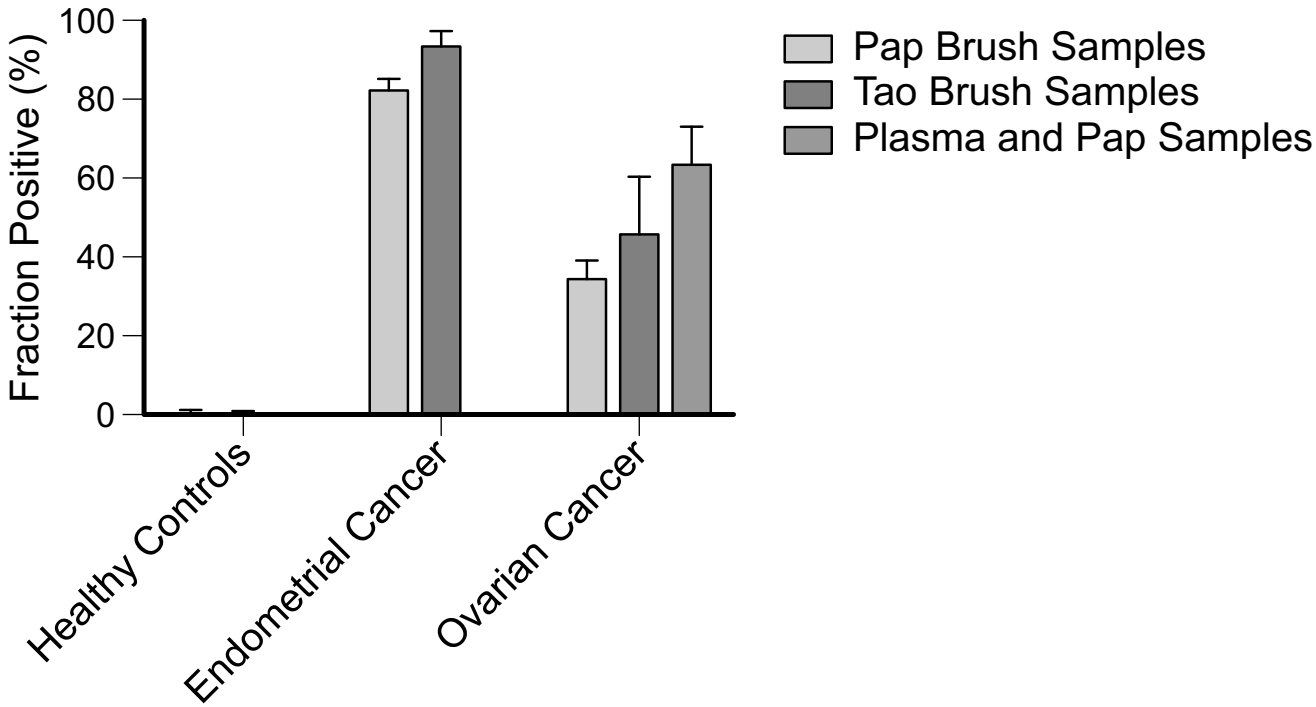

Fraction Positive (%)
100
80
60
40
20
0
Healthy Controls
Endometrial Cancer
Ovarian Cancer
Pap Brush Samples
Tao Brush Samples
Plasma and Pap Samples

FIG. 24

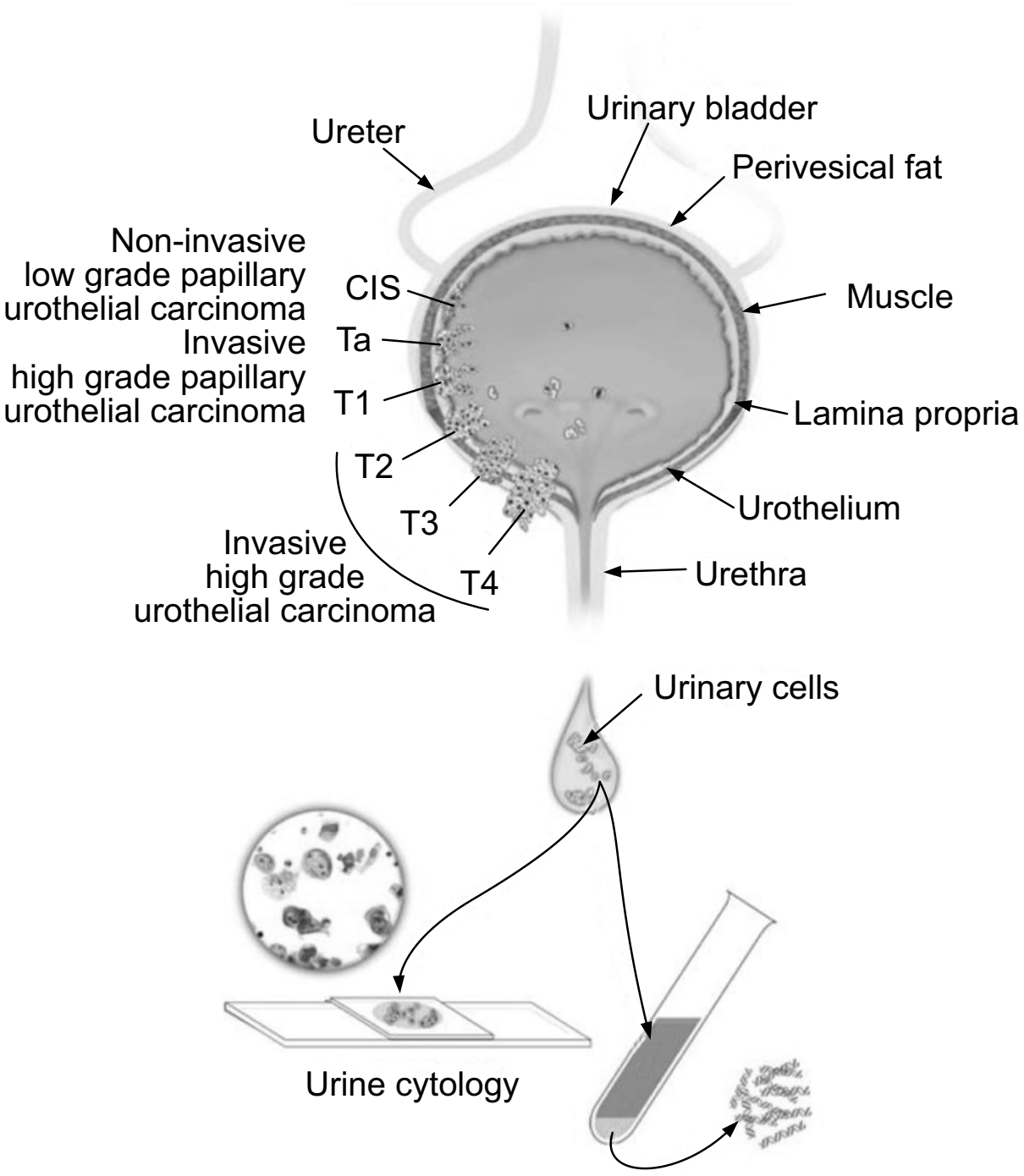

Ureter
Urinary bladder
Perivesical fat
Non-invasive
low grade papillary
urothelial carcinoma
CIS
Ta
Invasive
high grade papillary
urothelial carcinoma
T1
T2
T3
T4
Invasive
high grade
urothelial carcinoma
Muscle
Lamina propria
Urothelium
Urethra
Urinary cells
Urine cytology
UroSEEK:
Test for intragenetic mutations in 10 genes
Test for mutations in TERT promoter
Test for aneuploidy

FIG. 25

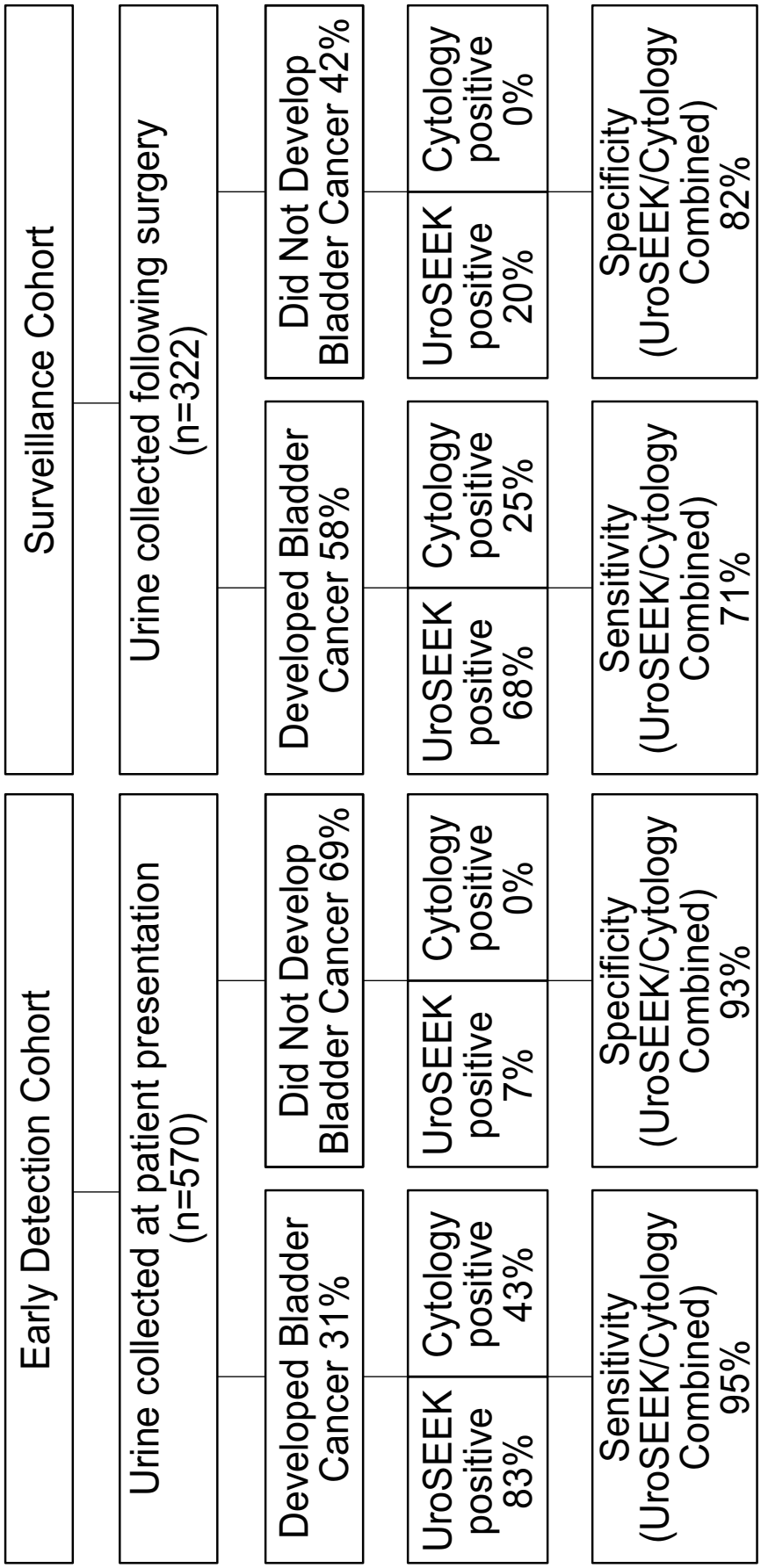

Early Detection Cohort
Urine collected at patient presentation (n=570)
Developed Bladder Cancer 31%
Did Not Develop Bladder Cancer 69%
UroSEEK positive 83%
Cytology positive 43%
UroSEEK positive 7%
Cytology positive 0%
Sensitivity (UroSEEK/Cytology Combined) 95%
Specificity (UroSEEK/Cytology Combined) 93%
Surveillance Cohort
Urine collected following surgery (n=322)
Developed Bladder Cancer 58%
Did Not Develop Bladder Cancer 42%
UroSEEK positive 68%
Cytology positive 25%
UroSEEK positive 20%
Cytology positive 0%
Sensitivity (UroSEEK/Cytology Combined) 71%
Specificity (UroSEEK/Cytology Combined) 82%

FIG. 26

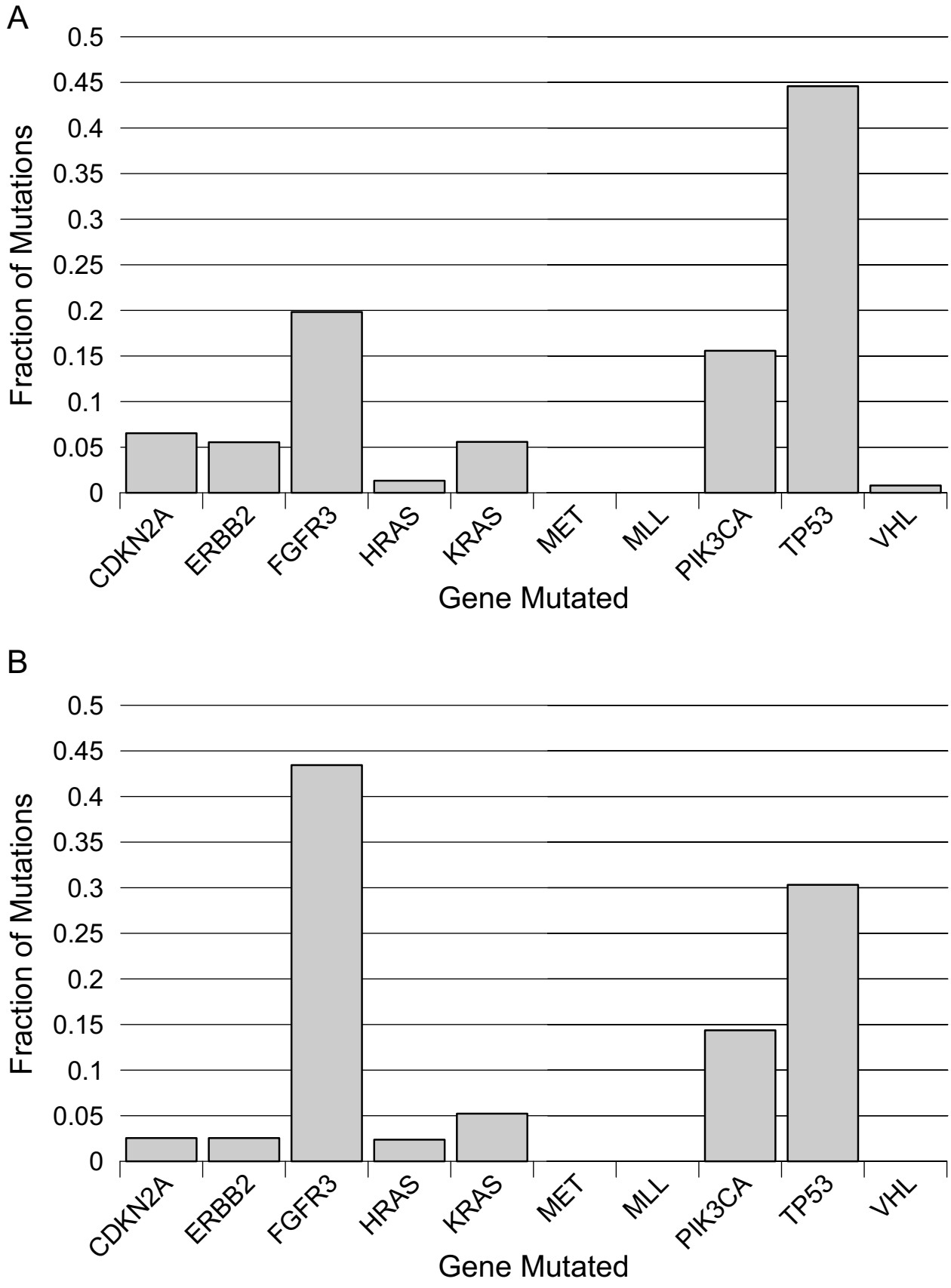

A
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Fraction of Mutations
CDKN2A
ERBB2
FGFR3
HRAS
KRAS
MET
MLL
PIK3CA
TP53
VHL
Gene Mutated
B
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Fraction of Mutations
CDKN2A
ERBB2
FGFR3
HRAS
KRAS
MET
MLL
PIK3CA
TP53
VHL
Gene Mutated

FIG. 27

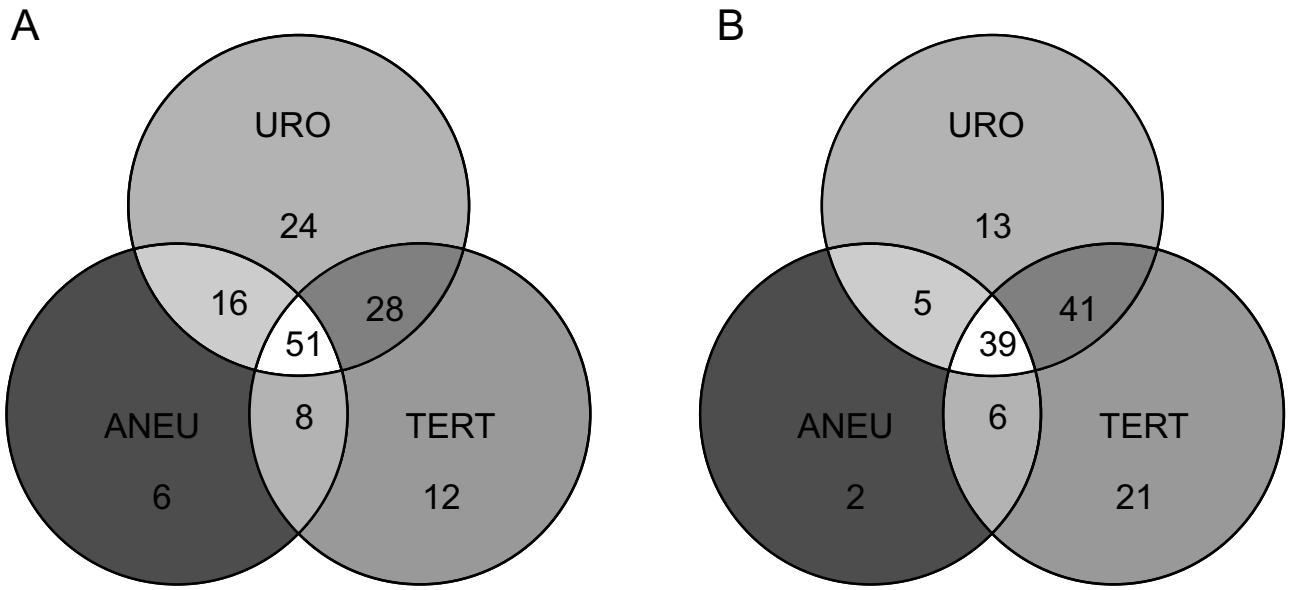


FIG. 28

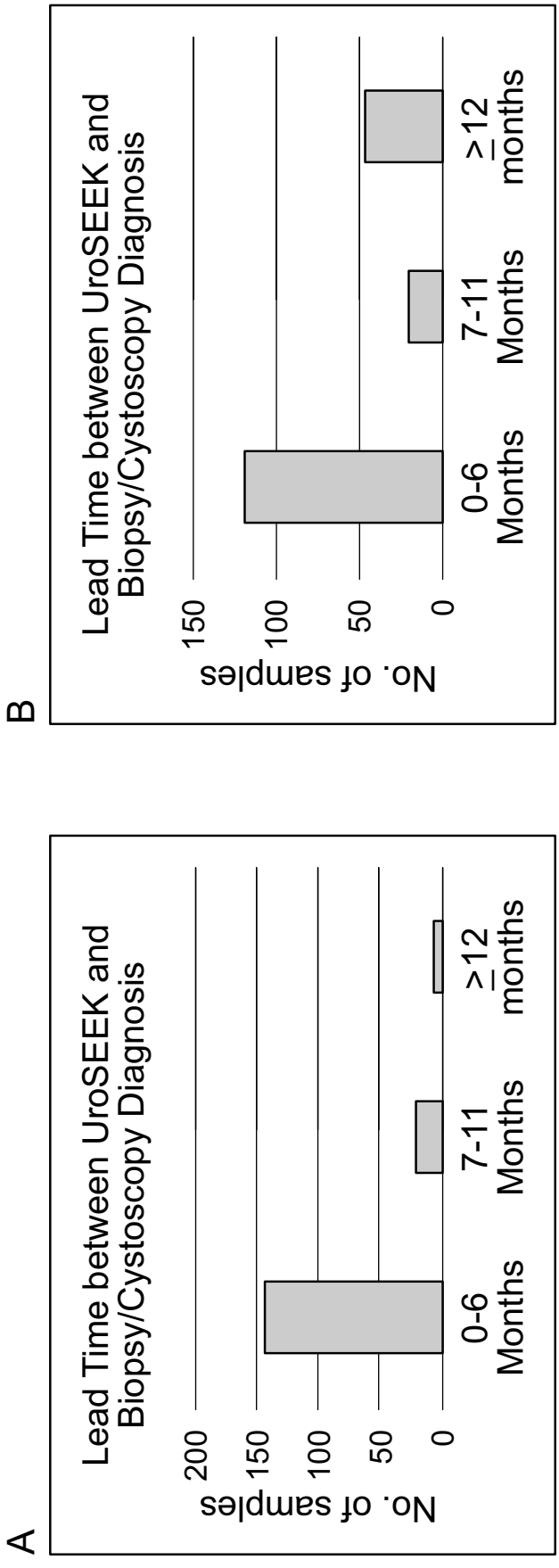

A
Lead Time between UroSEEK and Biopsy/Cystoscopy Diagnosis
No. of samples
200
150
100
50
0
0-6 Months
7-11 Months
≥12 months
B
Lead Time between UroSEEK and Biopsy/Cystoscopy Diagnosis
No. of samples
150
100
50
0
0-6 Months
7-11 Months
≥12 months

FIG. 29

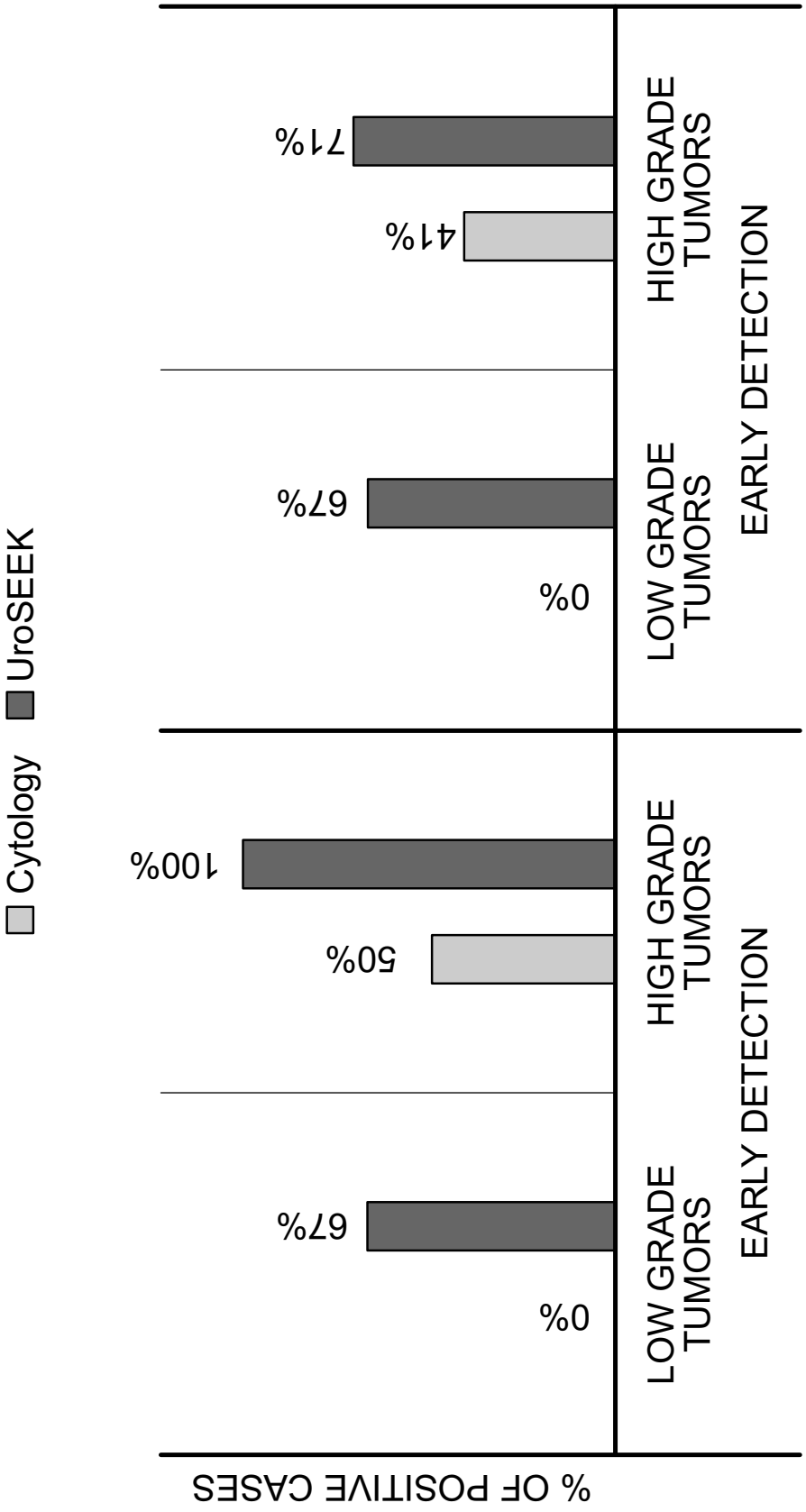
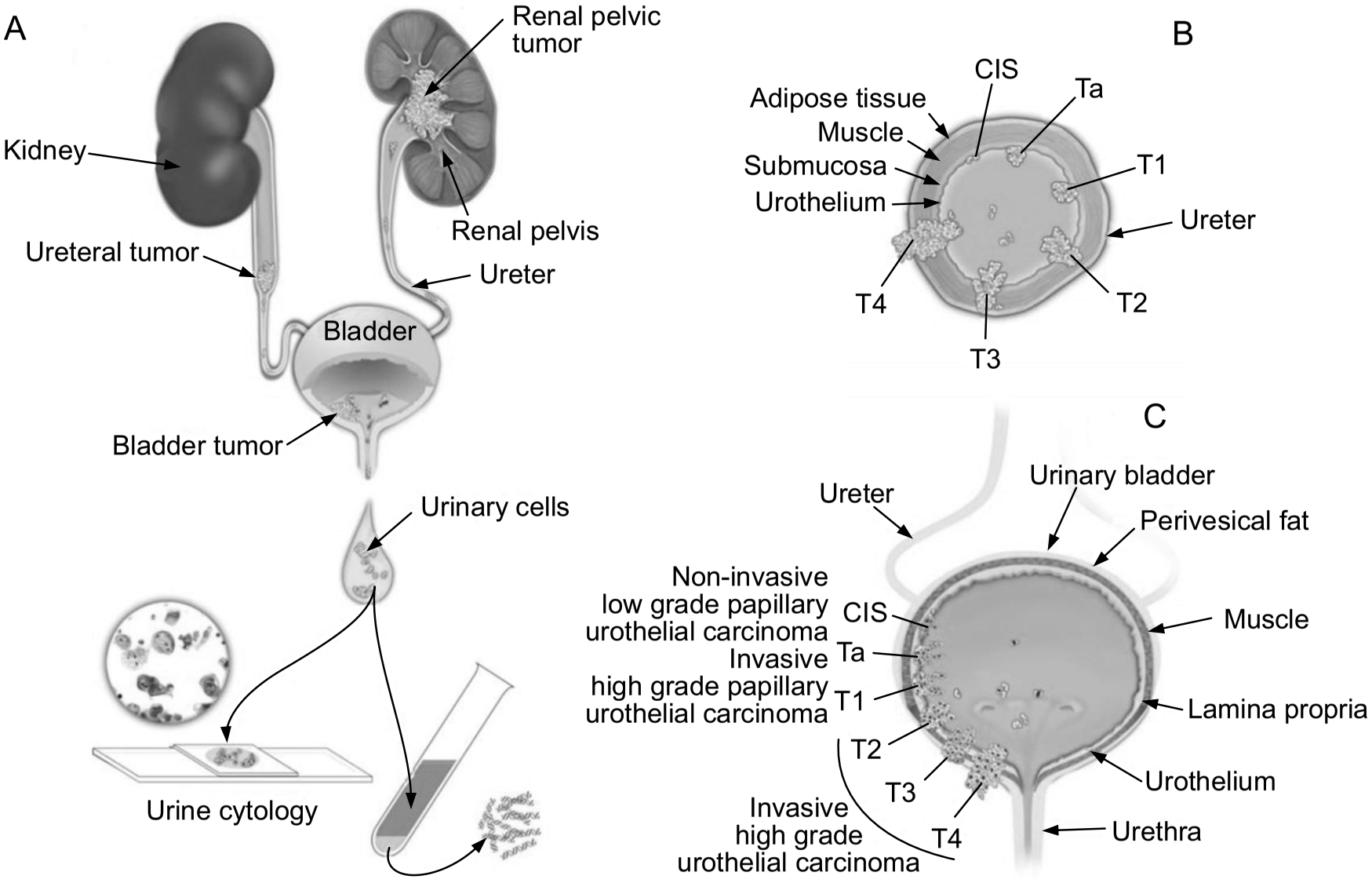

% OF POSITIVE CASES
LOW GRADE TUMORS
HIGH GRADE TUMORS
EARLY DETECTION
0%
67%
50%
100%
LOW GRADE TUMORS
HIGH GRADE TUMORS
EARLY DETECTION
0%
67%
41%
71%
Cytology
UroSEEK

FIG. 30



UroSEEK:
Test for intragenetic mutations in 10 genes
Test for mutations in TERT promoter
Test for aneuploidy

FIG. 31

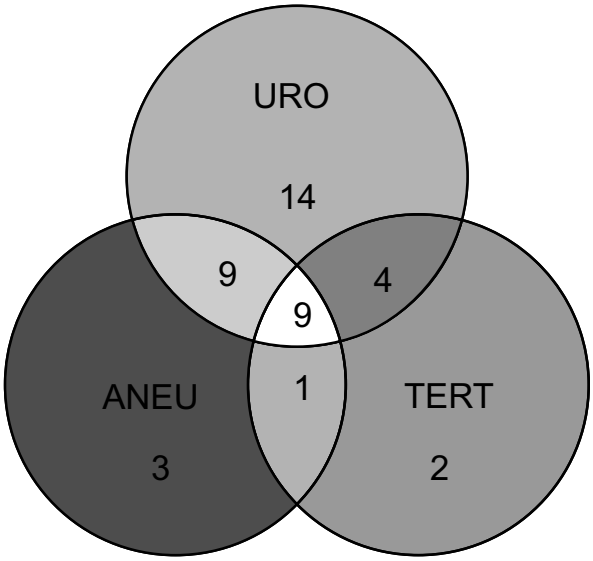


FIG. 32

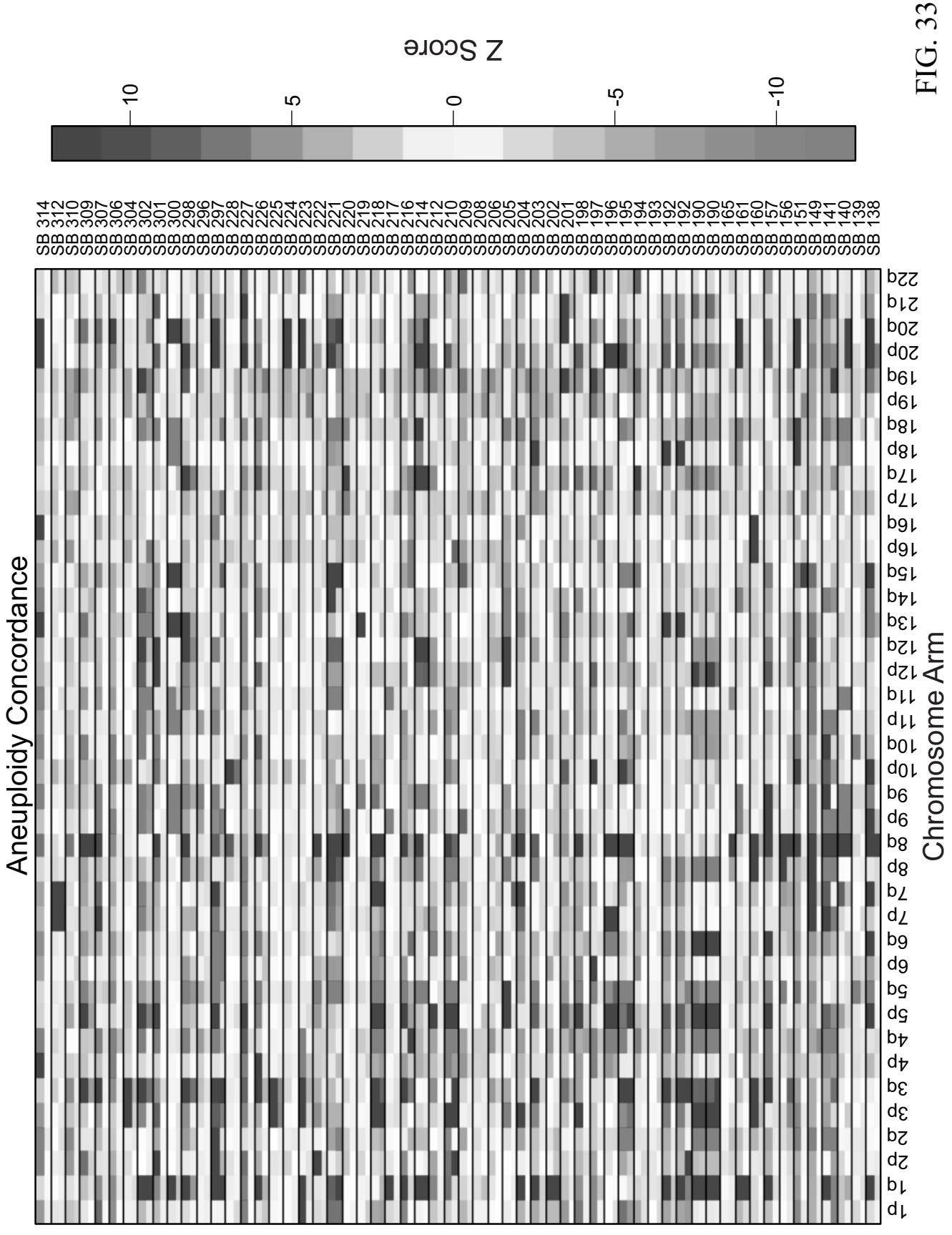

Aneuploidy Concordance
Chromosome Arm
1p
1q
2p
2q
3p
3q
4p
4q
5p
5q
6p
6q
7p
7q
8p
8q
9p
9q
10p
10q
11p
11q
12p
12q
13q
14q
15q
16p
16q
17p
17q
18p
18q
19p
19q
20p
20q
21q
22q
SB 314
SB 312
SB 310
SB 309
SB 307
SB 306
SB 304
SB 302
SB 301
SB 300
SB 298
SB 296
SB 297
SB 228
SB 227
SB 226
SB 225
SB 224
SB 223
SB 222
SB 221
SB 220
SB 219
SB 218
SB 217
SB 216
SB 214
SB 212
SB 210
SB 209
SB 208
SB 206
SB 205
SB 204
SB 203
SB 202
SB 201
SB 198
SB 197
SB 196
SB 195
SB 194
SB 193
SB 192
SB 190
SB 190
SB 165
SB 161
SB 160
SB 157
SB 156
SB 151
SB 149
SB 141
SB 140
SB 139
SB 138
Z Score
10
5
0
-5
-10

FIG. 33

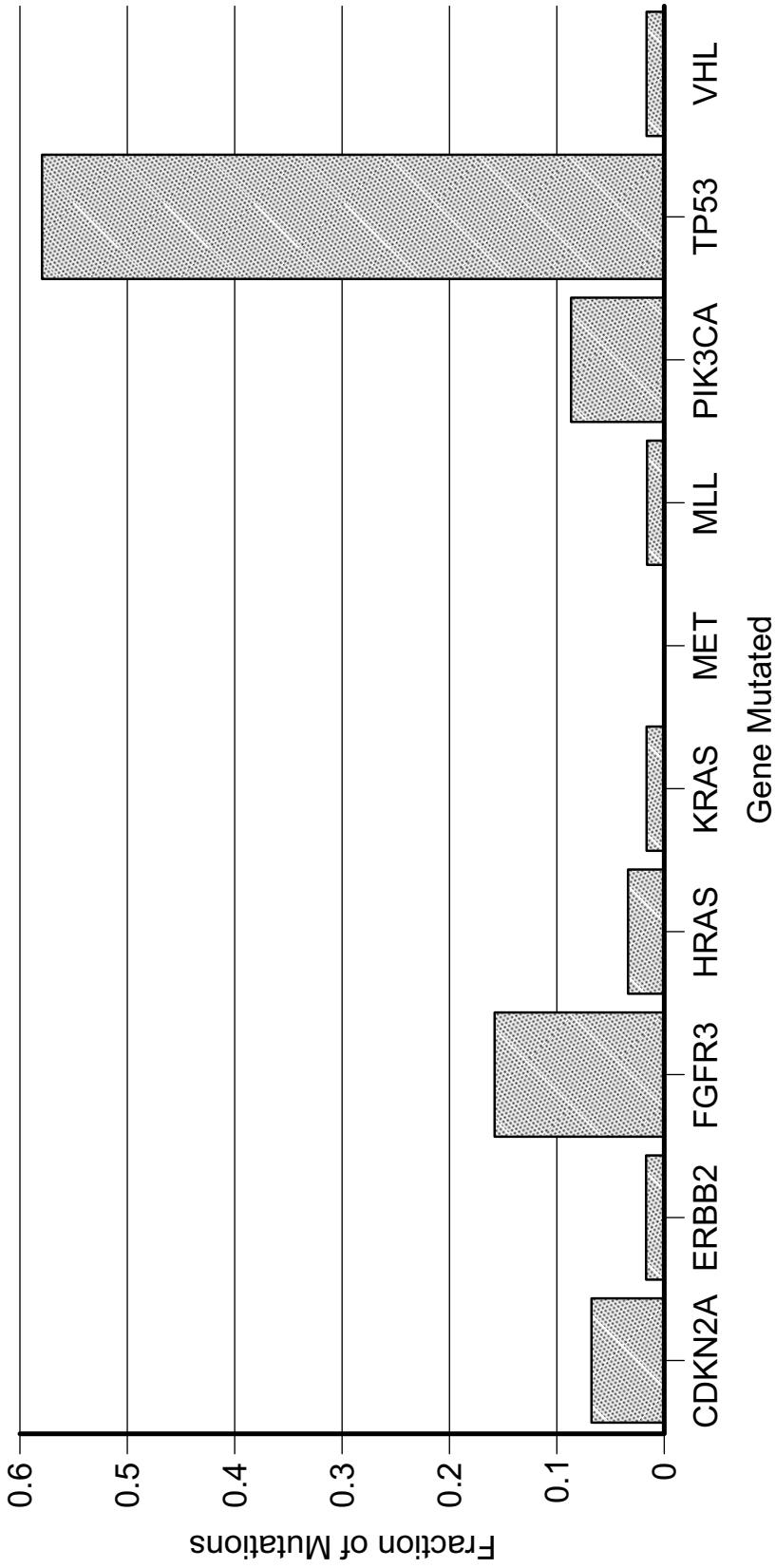

Fraction of Mutations
0.6
0.5
0.4
0.3
0.2
0.1
0
CDKN2A
ERBB2
FGFR3
HRAS
KRAS
MET
MLL
PIK3CA
TP53
VHL
Gene Mutated

FIG. 34

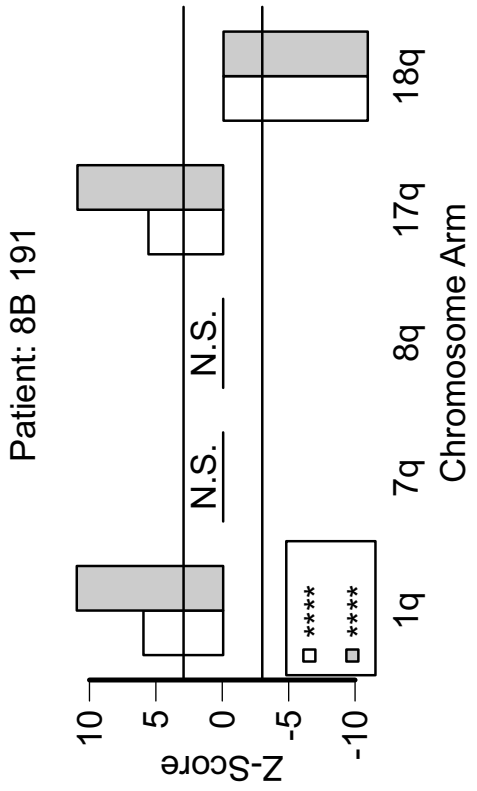


FIG. 35A

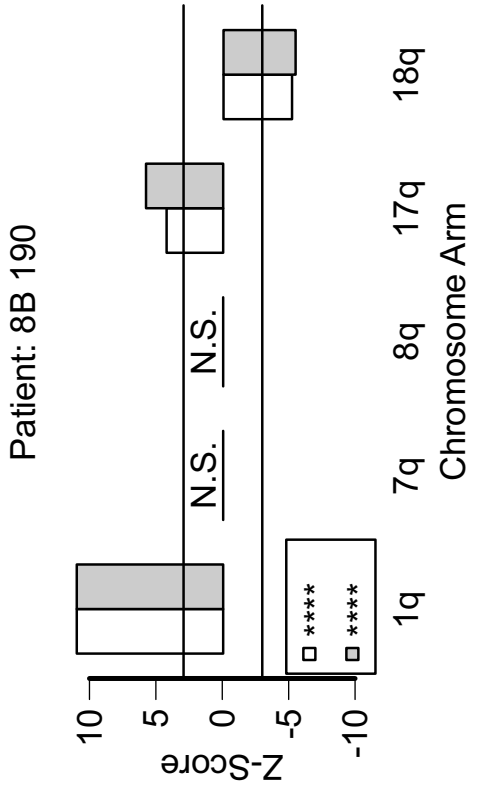


FIG. 35B

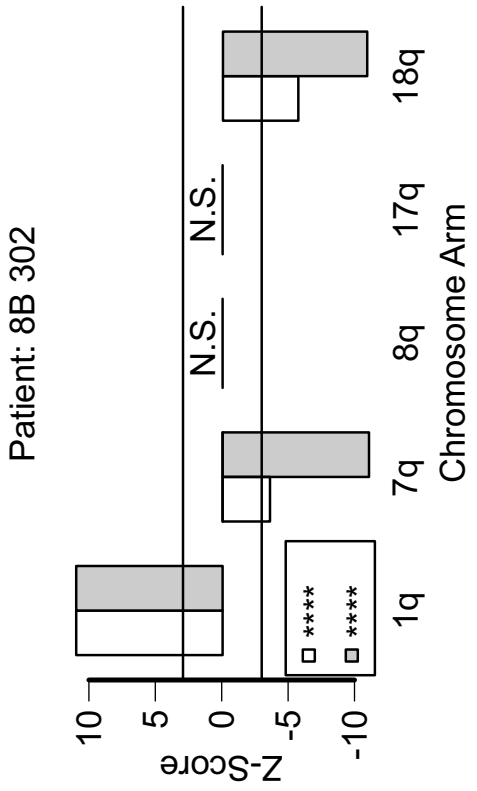


FIG. 35C

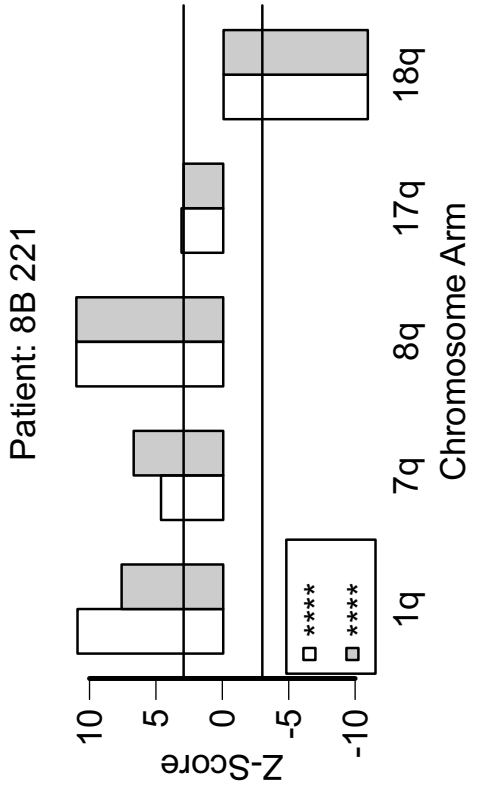


FIG. 35D

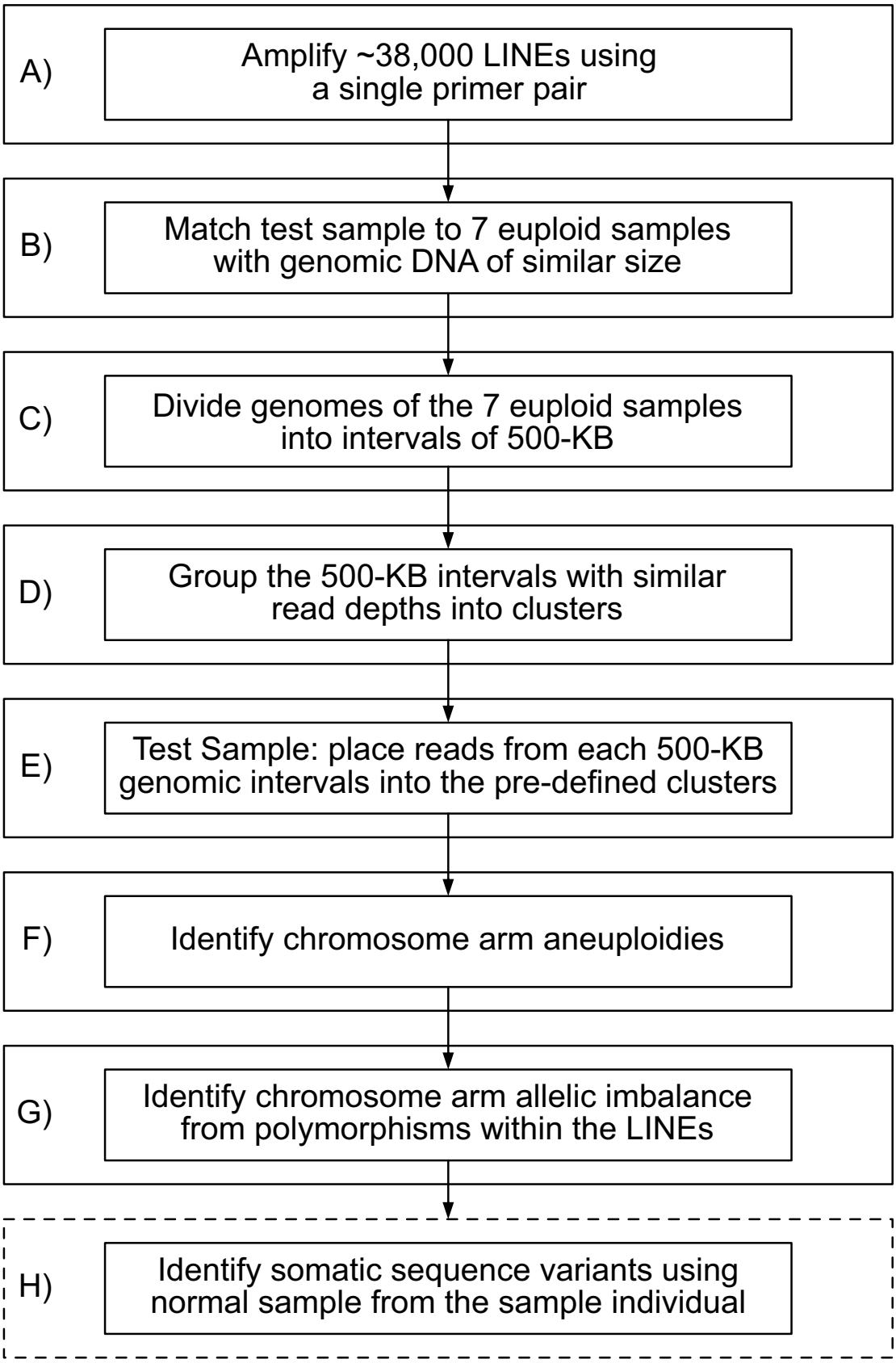

A)
Amplify ~38,000 LINEs using a single primer pair
B)
Match test sample to 7 euploid samples with genomic DNA of similar size
C)
Divide genomes of the 7 euploid samples into intervals of 500-KB
D)
Group the 500-KB intervals with similar read depths into clusters
E)
Test Sample: place reads from each 500-KB genomic intervals into the pre-defined clusters
F)
Identify chromosome arm aneuploidies
G)
Identify chromosome arm allelic imbalance from polymorphisms within the LINEs
H)
Identify somatic sequence variants using normal sample from the sample individual

FIG. 36

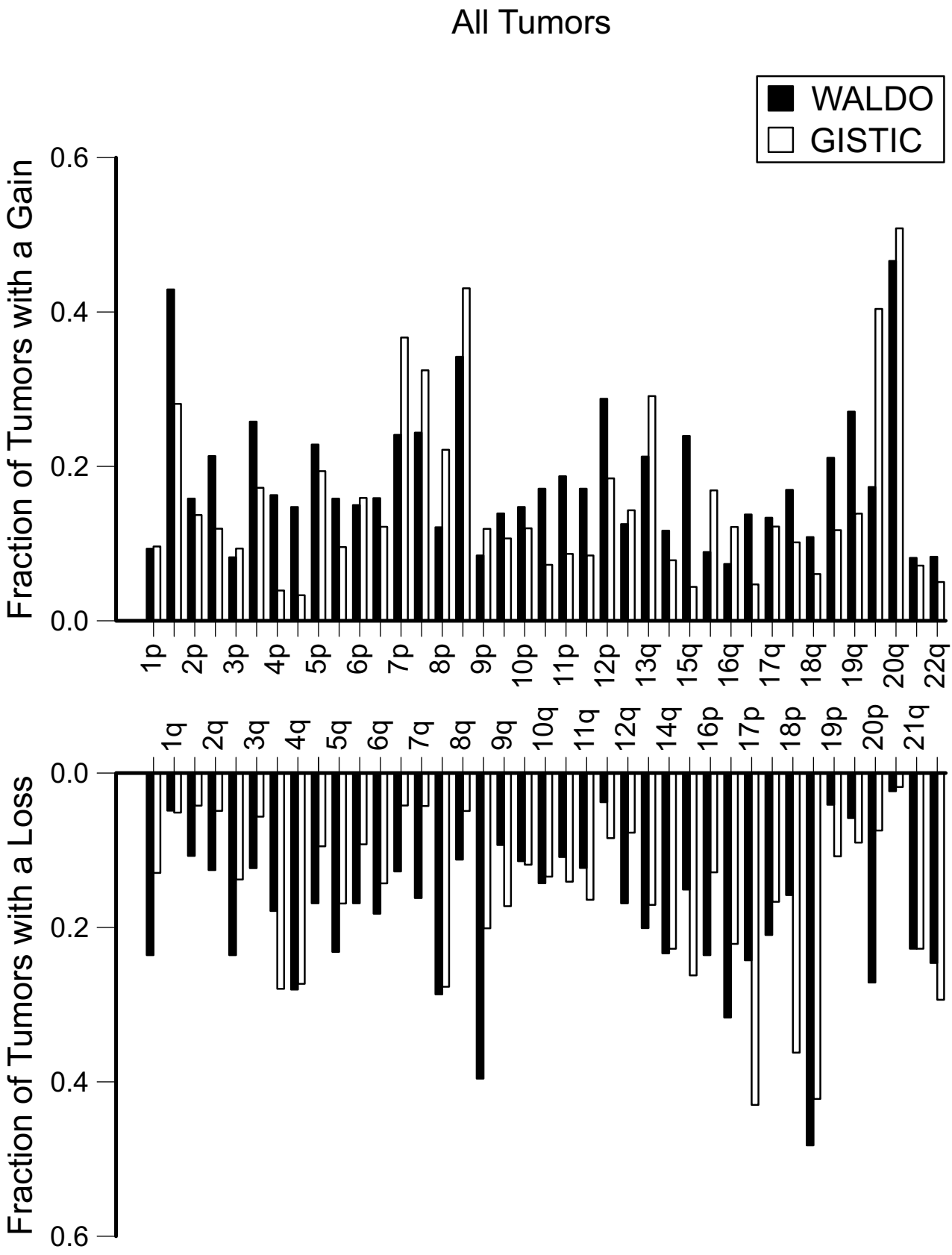

All Tumors
WALDO
GISTIC
Fraction of Tumors with a Gain
0.6
0.4
0.2
0.0
1p
2p
3p
4p
5p
6p
7p
8p
9p
10p
11p
12p
13q
15q
16q
17q
18q
19q
20q
22q
Fraction of Tumors with a Loss
0.0
0.2
0.4
0.6
1q
2q
3q
4q
5q
6q
7q
8q
9q
10q
11q
12q
14q
16p
17p
18p
19p
20p
21q

FIG. 37

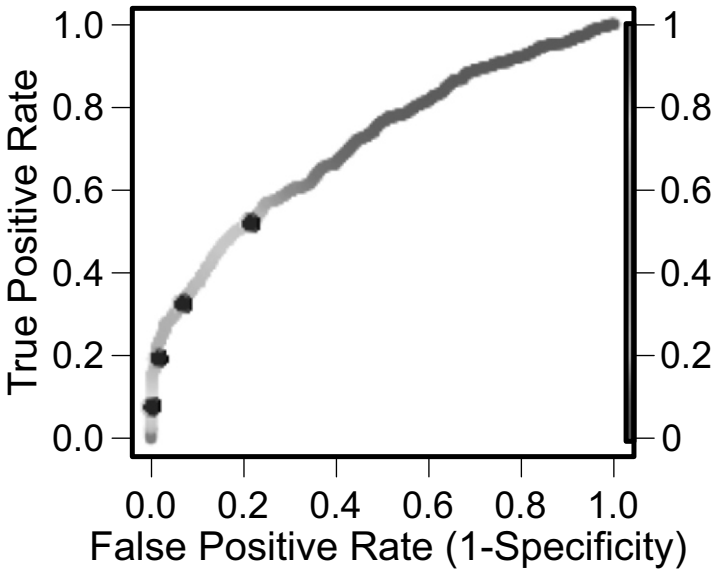


FIG. 38A

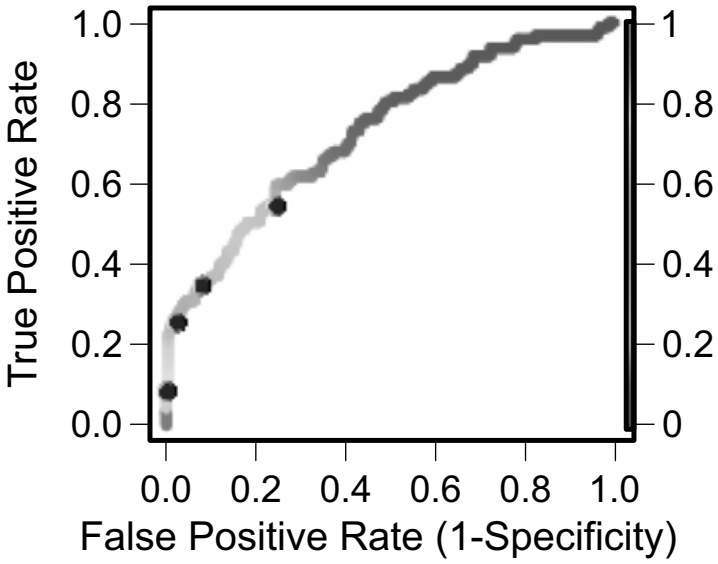


FIG. 38B

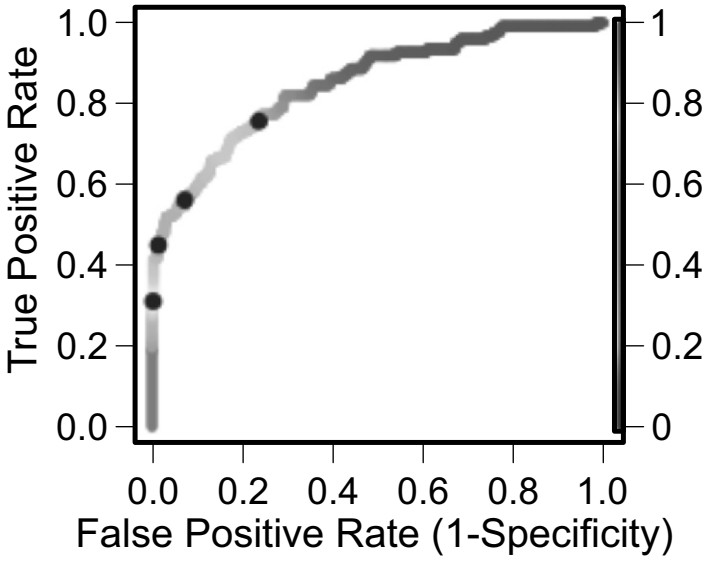


FIG. 38C

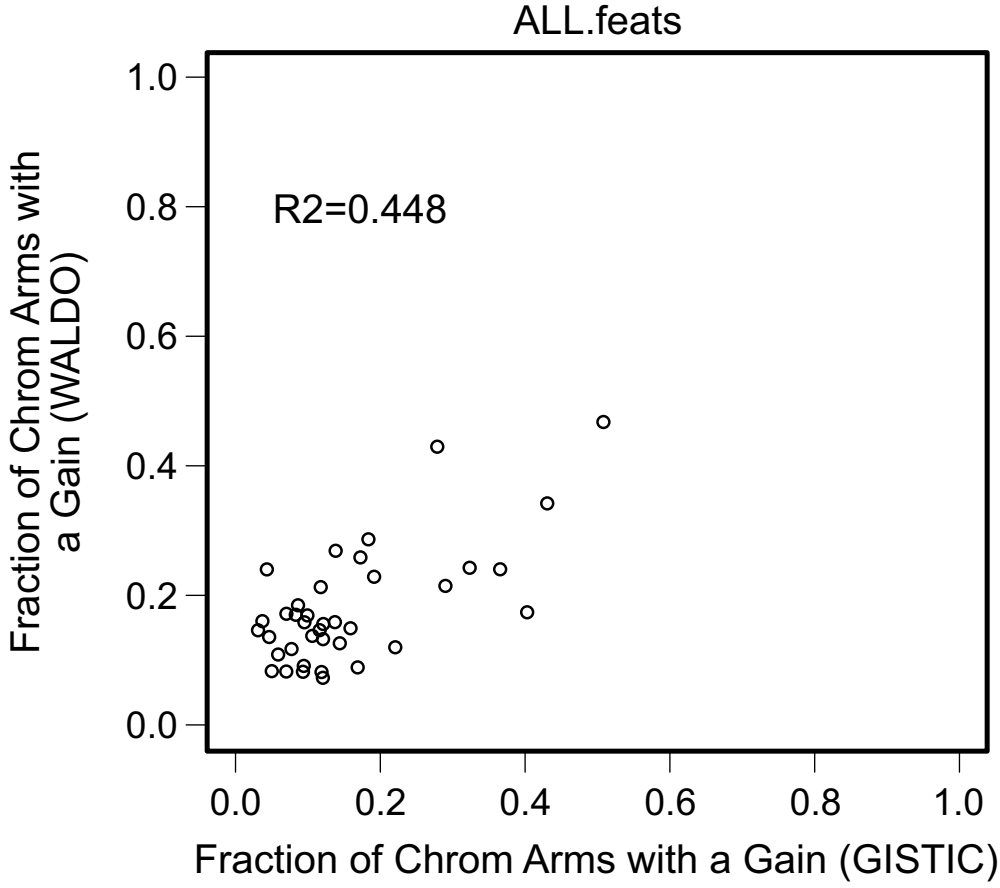


FIG. 39A

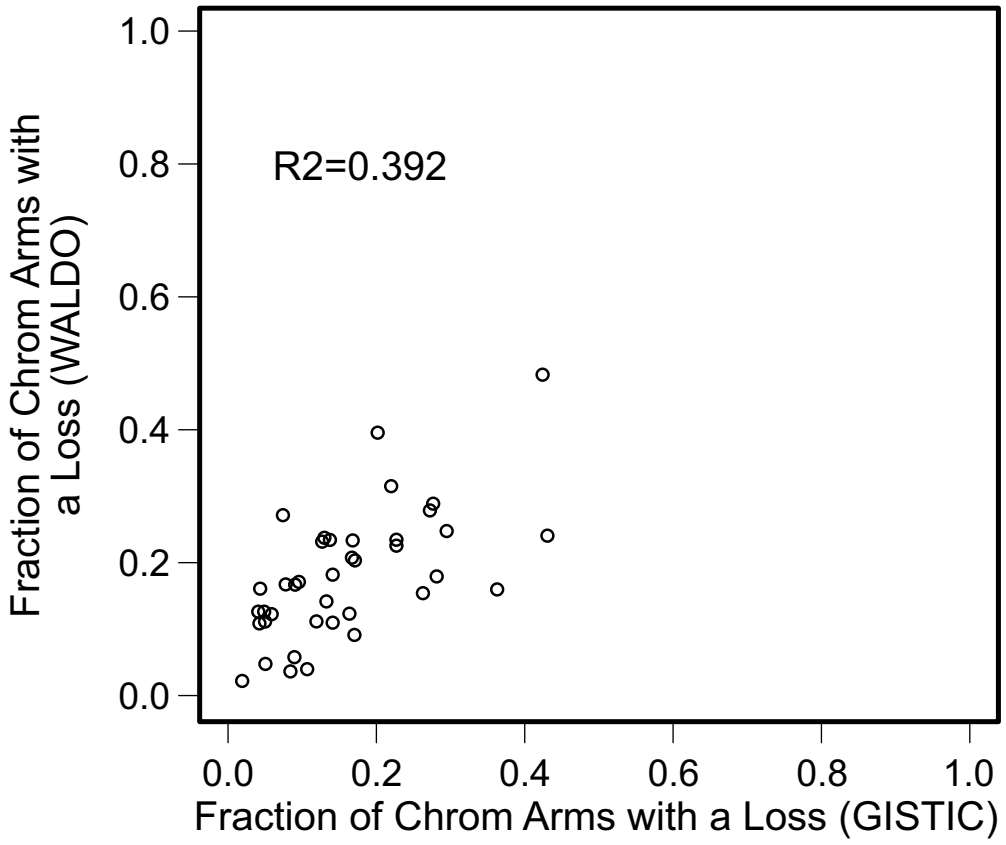


FIG. 39B

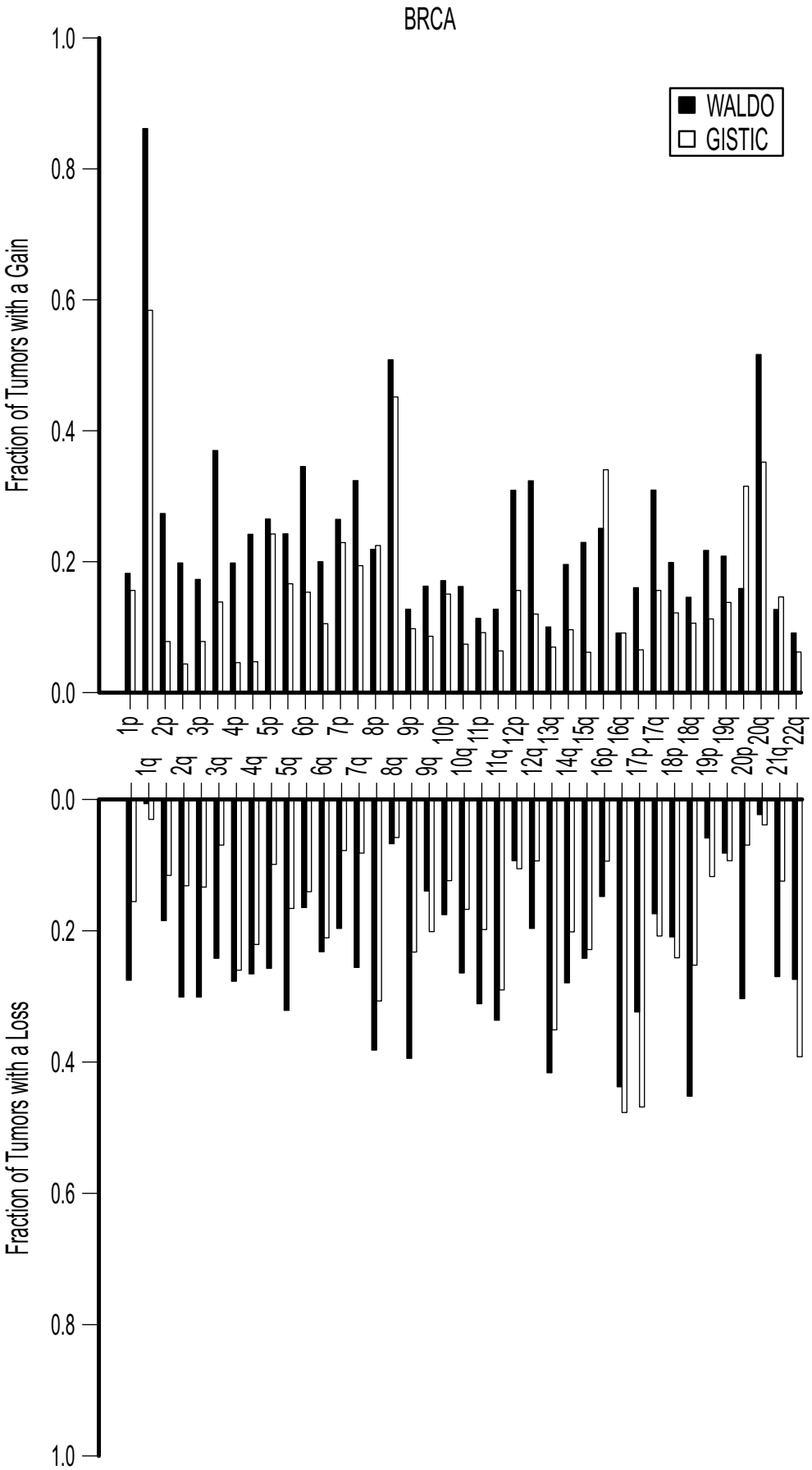
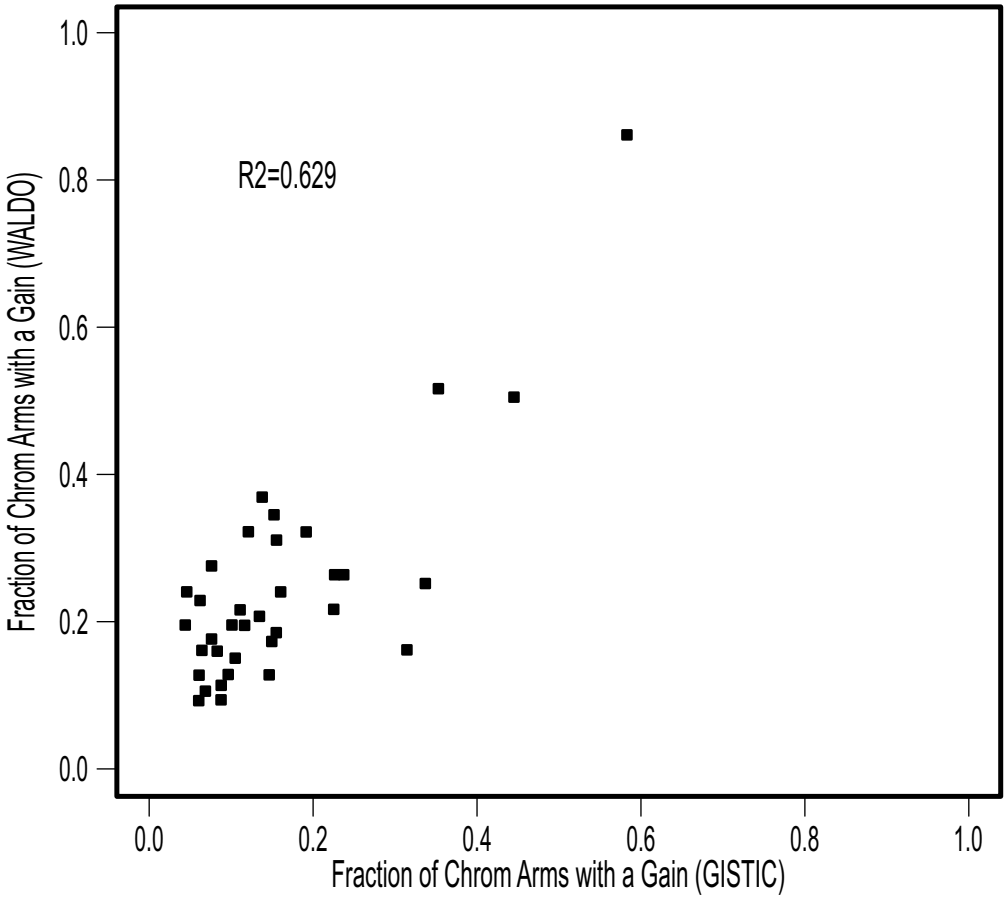

BRCA
WALDO
GISTIC
Fraction of Tumors with a Gain
Fraction of Tumors with a Loss
1.0
0.8
0.6
0.4
0.2
0.0
1p 2p 3p 4p 5p 6p 7p 8p 9p 10p 11p 12p 13q 15q 16q 17q 18q 19q 20q 22q
1q 2q 3q 4q 5q 6q 7q 8q 9q 10q 11q 12q 14q 16p 17p 18p 19p 20p 21q

FIG. 40A


R2=0.629
Fraction of Chrom Arms with a Gain (WALDO)
Fraction of Chrom Arms with a Gain (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

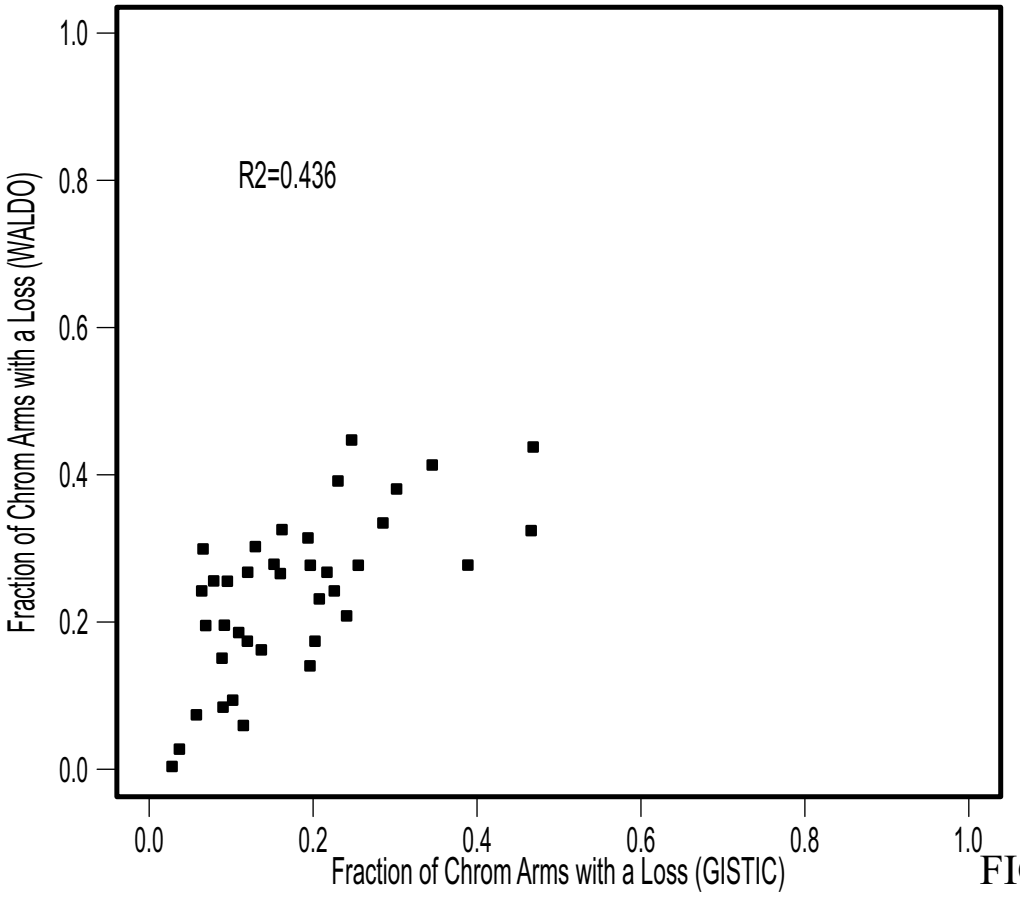

R2=0.436
Fraction of Chrom Arms with a Loss (WALDO)
Fraction of Chrom Arms with a Loss (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 40A (Cont.)

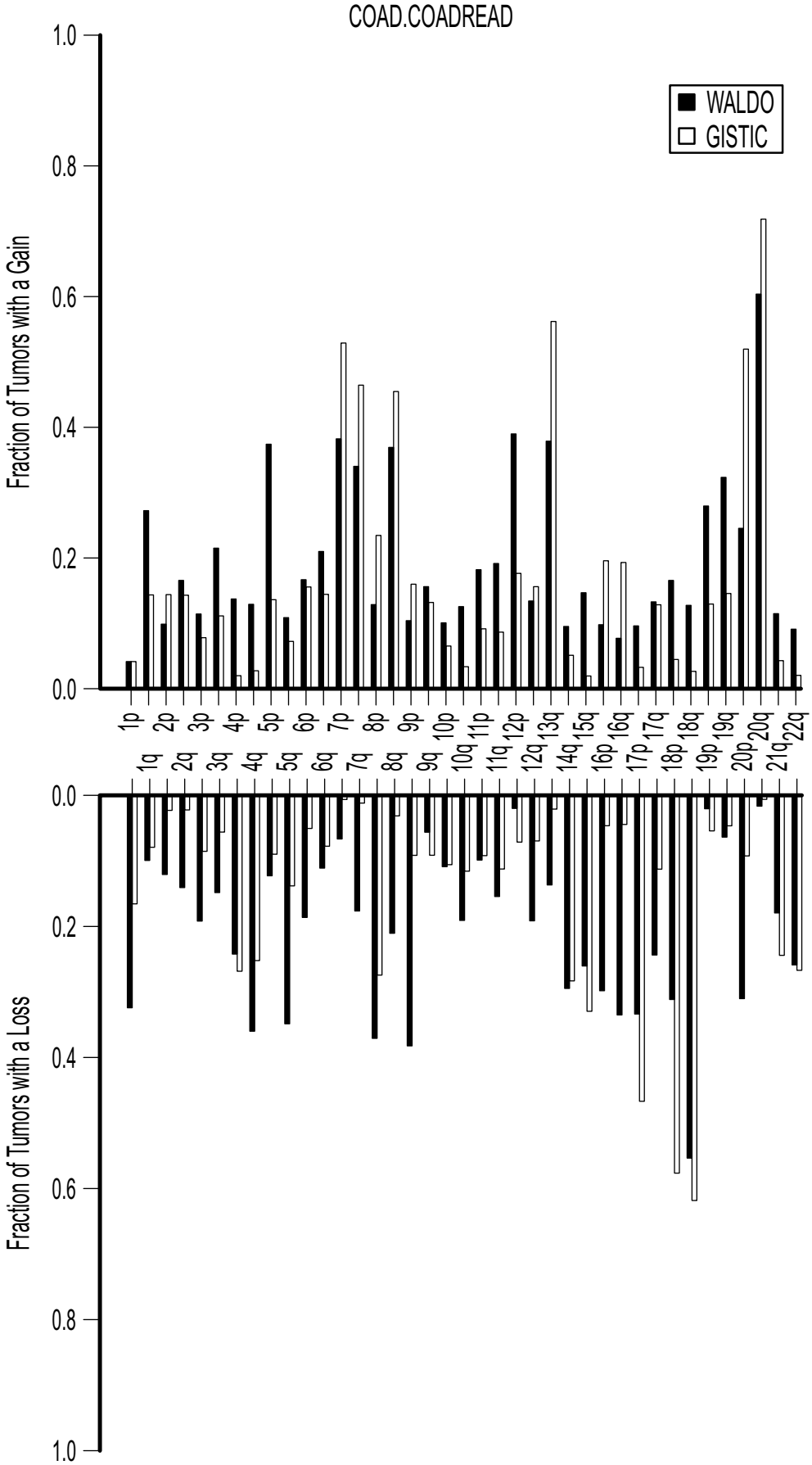
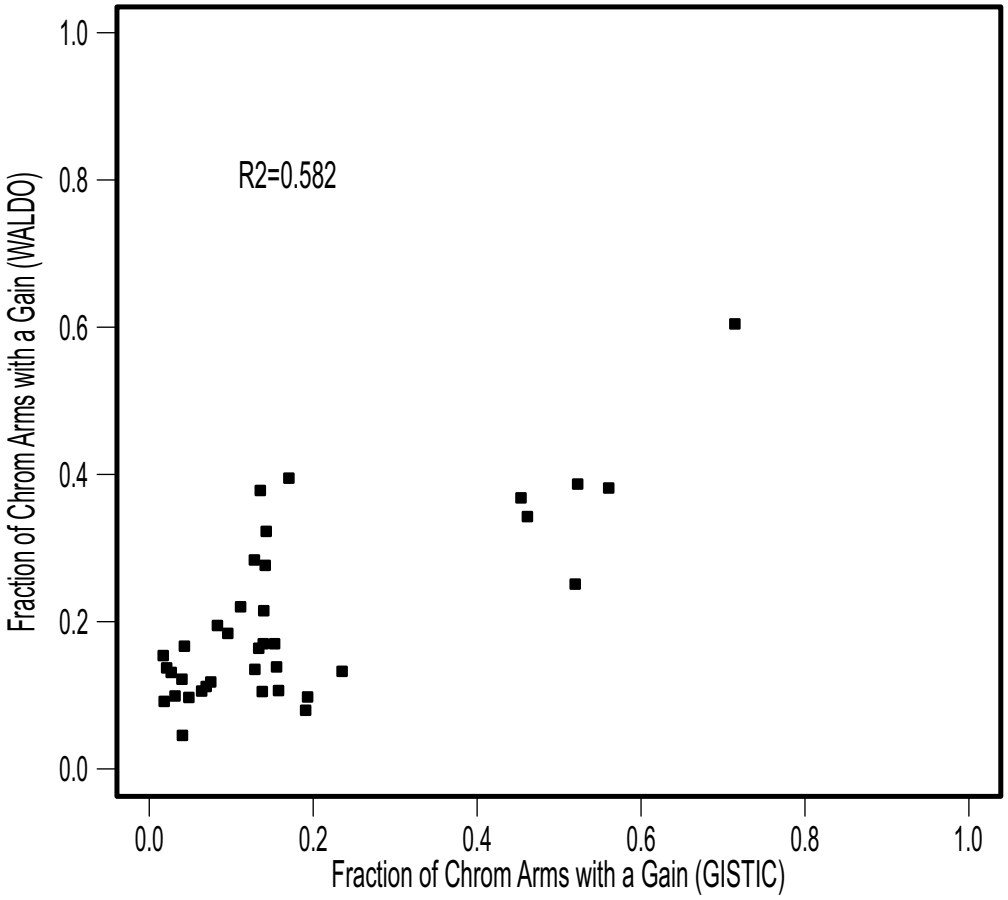

COAD.COADREAD
WALDO
GISTIC
Fraction of Tumors with a Gain
Fraction of Tumors with a Loss
1.0
0.8
0.6
0.4
0.2
0.0
1p 2p 3p 4p 5p 6p 7p 8p 9p 10p 11p 12p 13q 15q 16q 17q 18q 19q 20q 22q
1q 2q 3q 4q 5q 6q 7q 8q 9q 10q 11q 12q 14q 16p 17p 18p 19p 20p 21q

FIG. 40B


R2=0.582
Fraction of Chrom Arms with a Gain (WALDO)
Fraction of Chrom Arms with a Gain (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

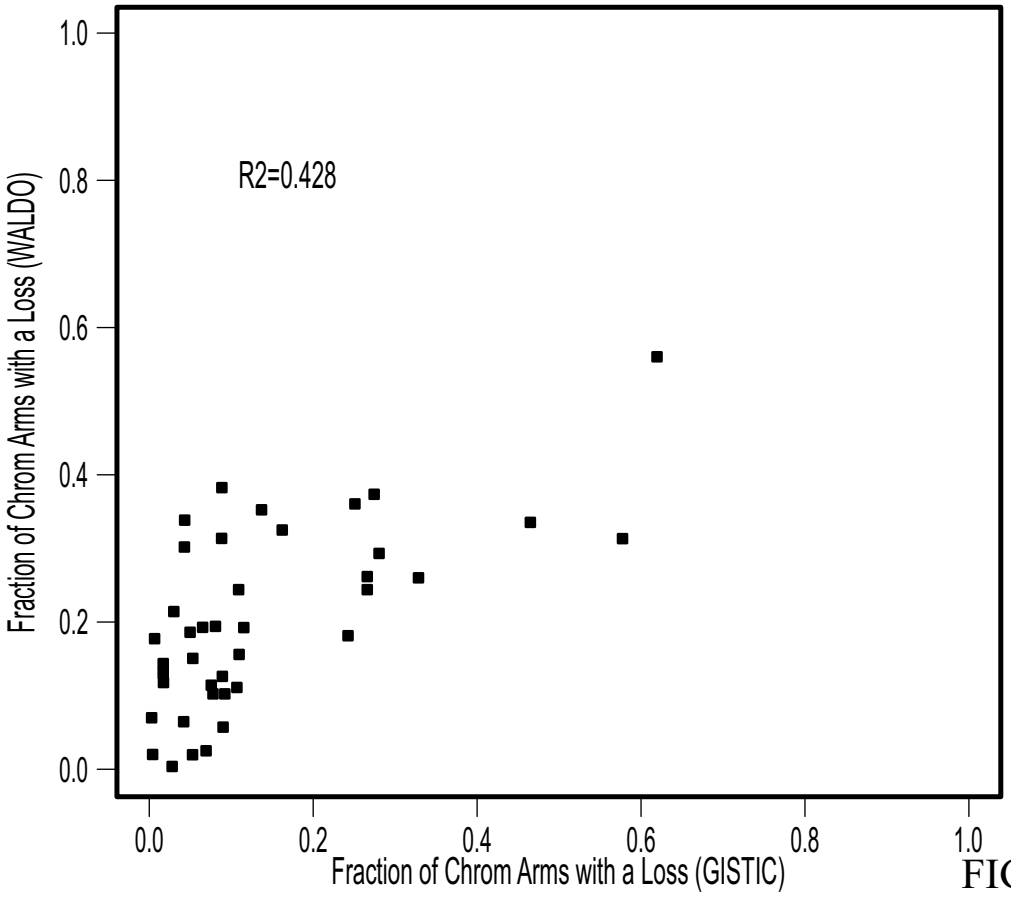

R2=0.428
Fraction of Chrom Arms with a Loss (WALDO)
Fraction of Chrom Arms with a Loss (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 40B (Cont.)

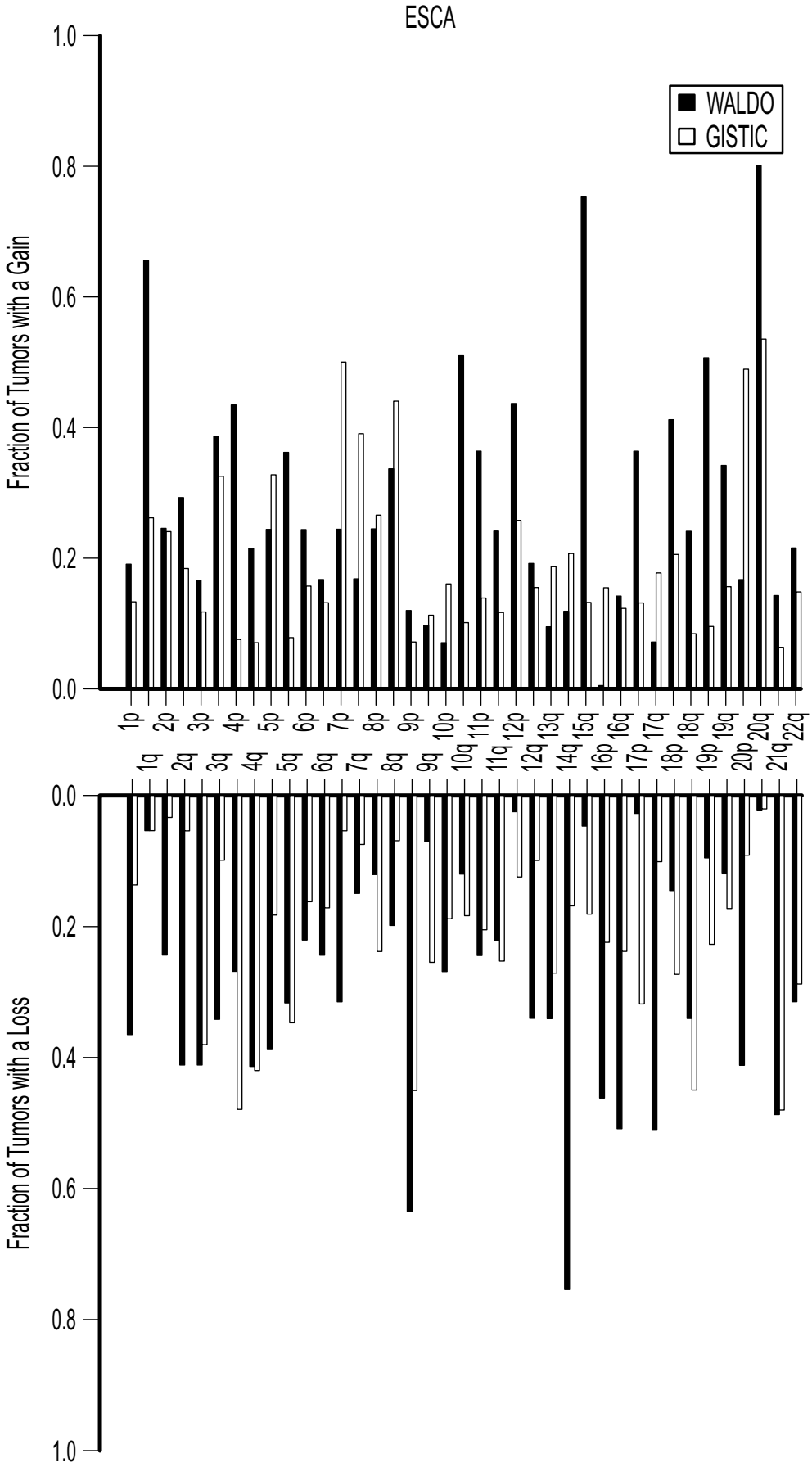
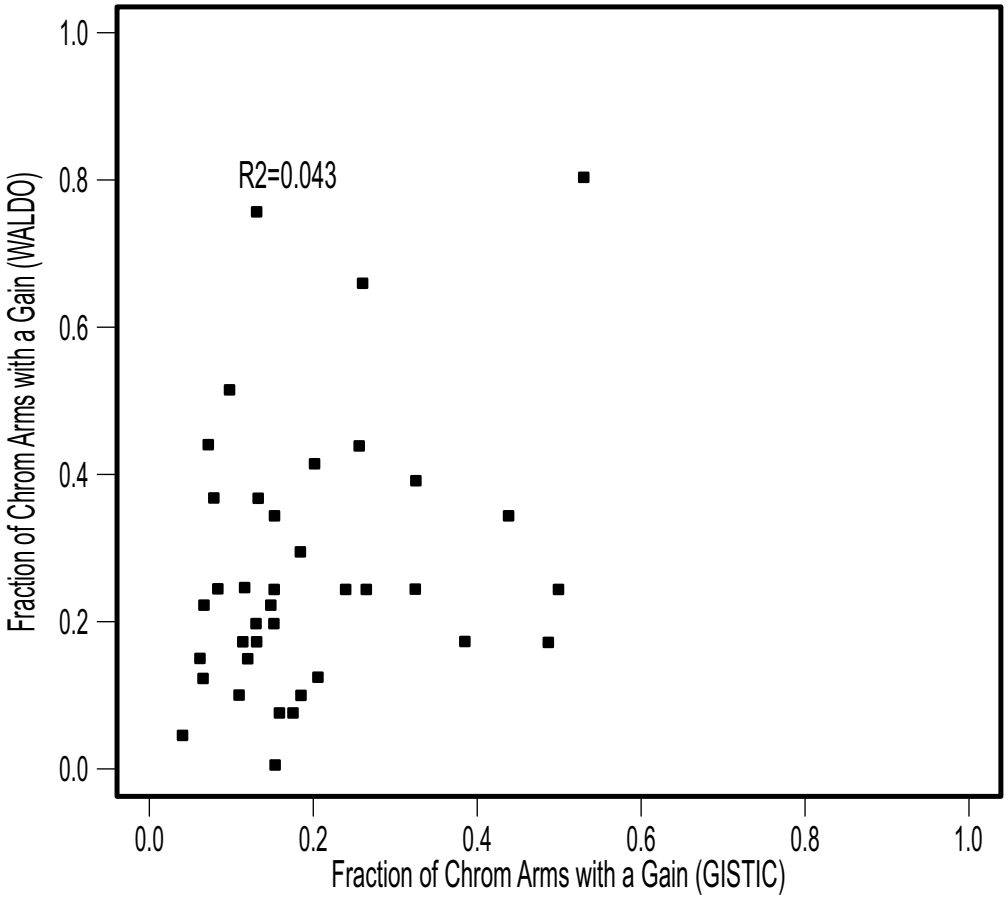

ESCA
WALDO
GISTIC
Fraction of Tumors with a Gain
Fraction of Tumors with a Loss
1.0
0.8
0.6
0.4
0.2
0.0
1p 2p 3p 4p 5p 6p 7p 8p 9p 10p 11p 12p 13q 15q 16p 16q 17p 17q 18p 18q 19p 19q 20p 20q 21q 22q
1q 2q 3q 4q 5q 6q 7q 8q 9q 10q 11q 12q 14q

FIG. 40C


R2=0.043
Fraction of Chrom Arms with a Gain (WALDO)
Fraction of Chrom Arms with a Gain (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

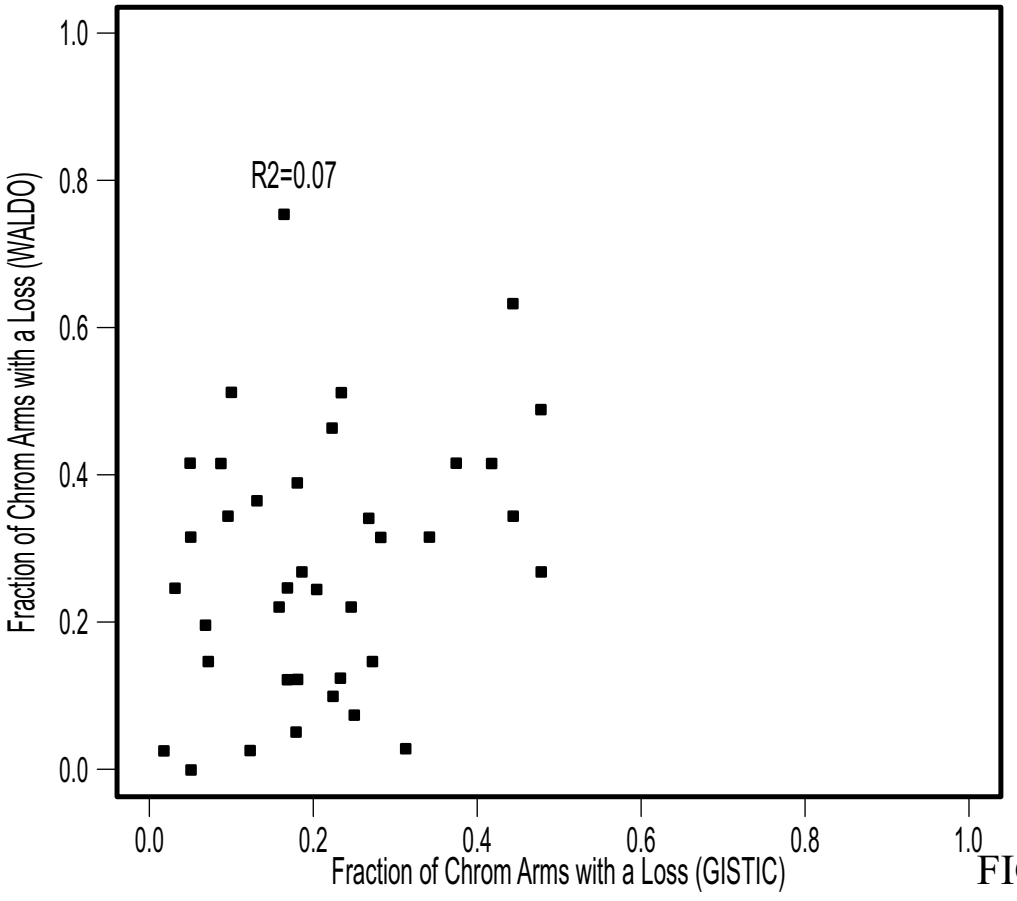

R2=0.07
Fraction of Chrom Arms with a Loss (WALDO)
Fraction of Chrom Arms with a Loss (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 40C (Cont.)

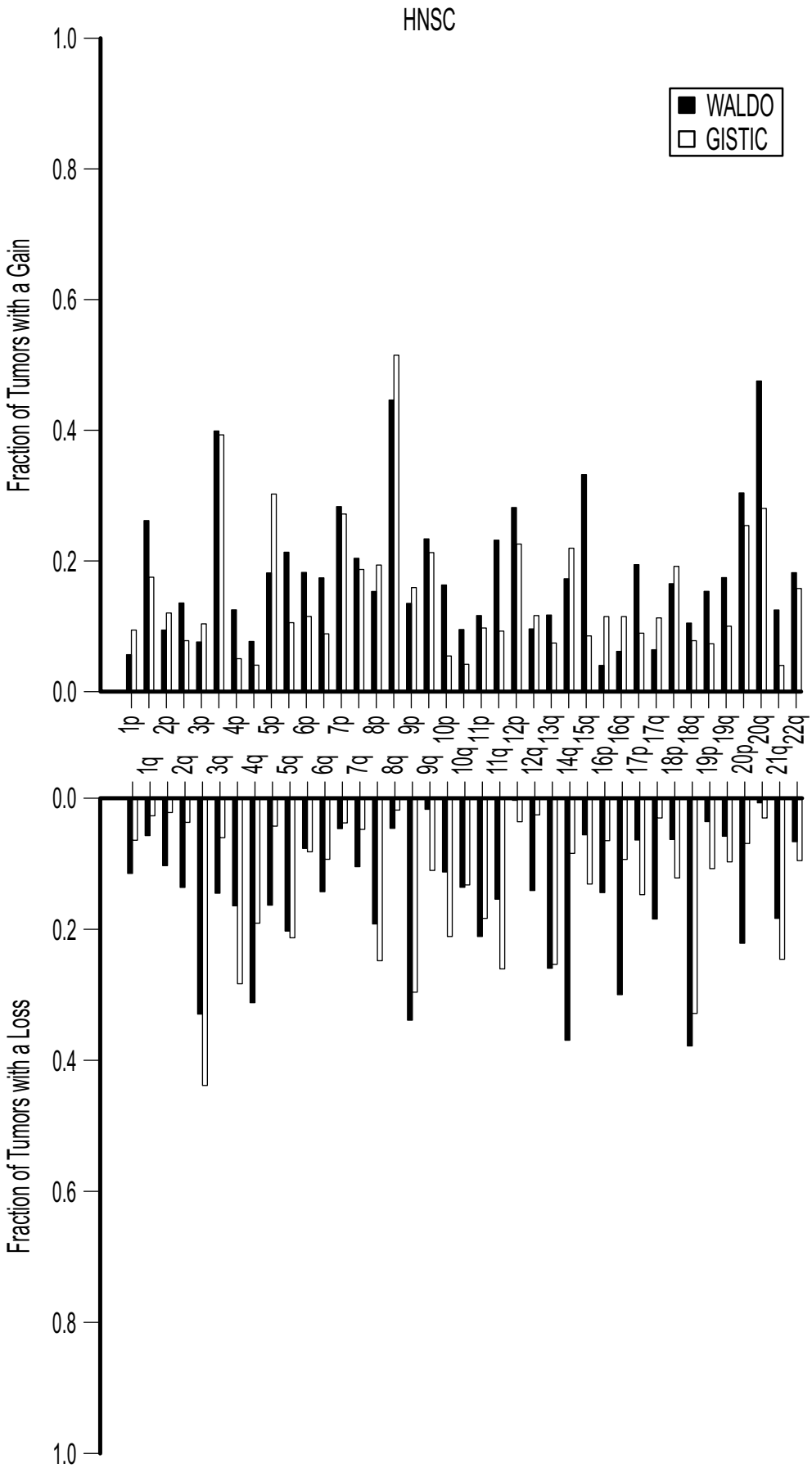
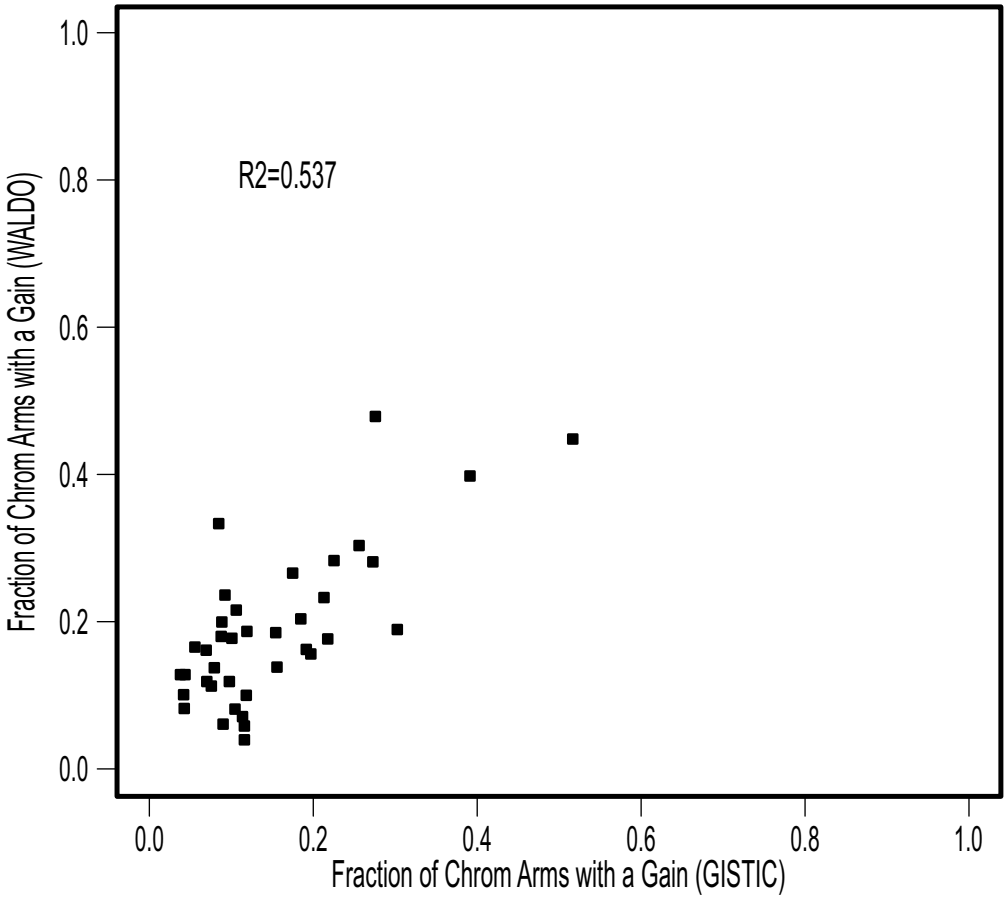

HNSC
WALDO
GISTIC
Fraction of Tumors with a Gain
Fraction of Tumors with a Loss
1.0
0.8
0.6
0.4
0.2
0.0
1p
2p
3p
4p
5p
6p
7p
8p
9p
10p
11p
12p
13q
15q
16p
16q
17p
17q
18p
18q
19p
19q
20p
20q
21q
22q
1q
2q
3q
4q
5q
6q
7q
8q
9q
10q
11q
12p

FIG. 40D


R2=0.537
Fraction of Chrom Arms with a Gain (WALDO)
Fraction of Chrom Arms with a Gain (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

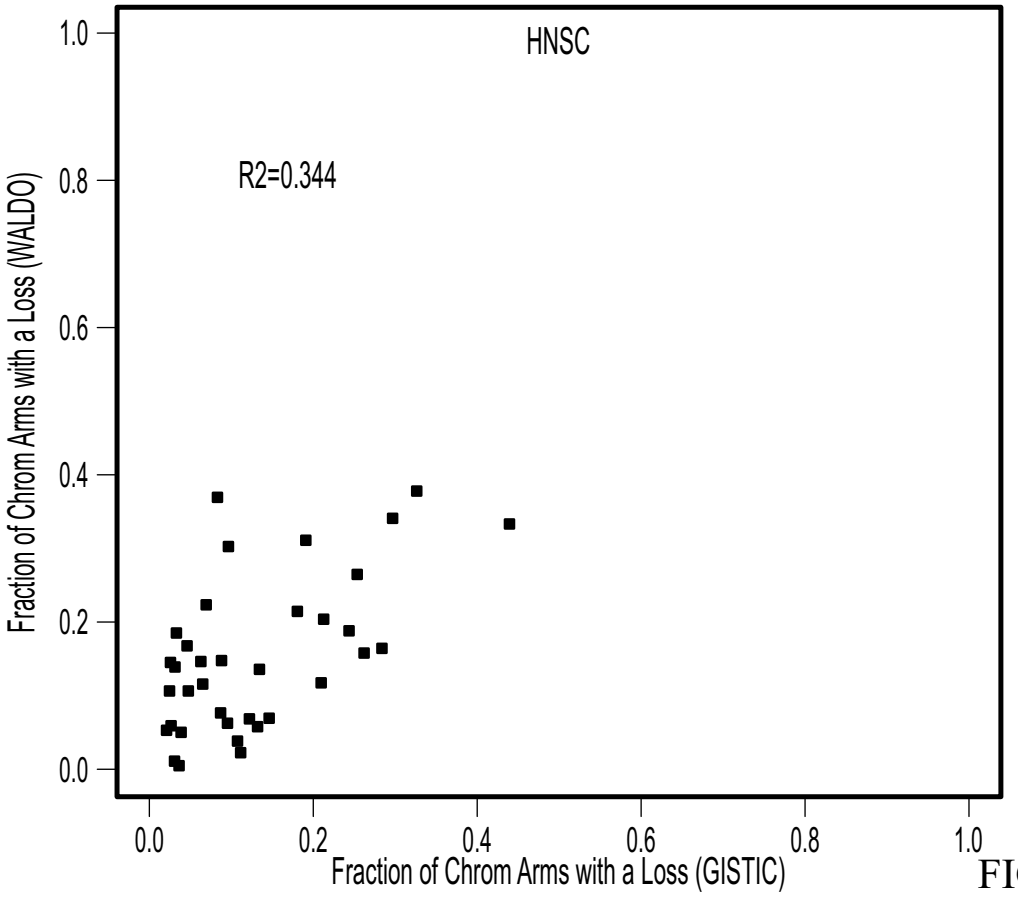

HNSC
R2=0.344
Fraction of Chrom Arms with a Loss (WALDO)
Fraction of Chrom Arms with a Loss (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 40D (Cont.)

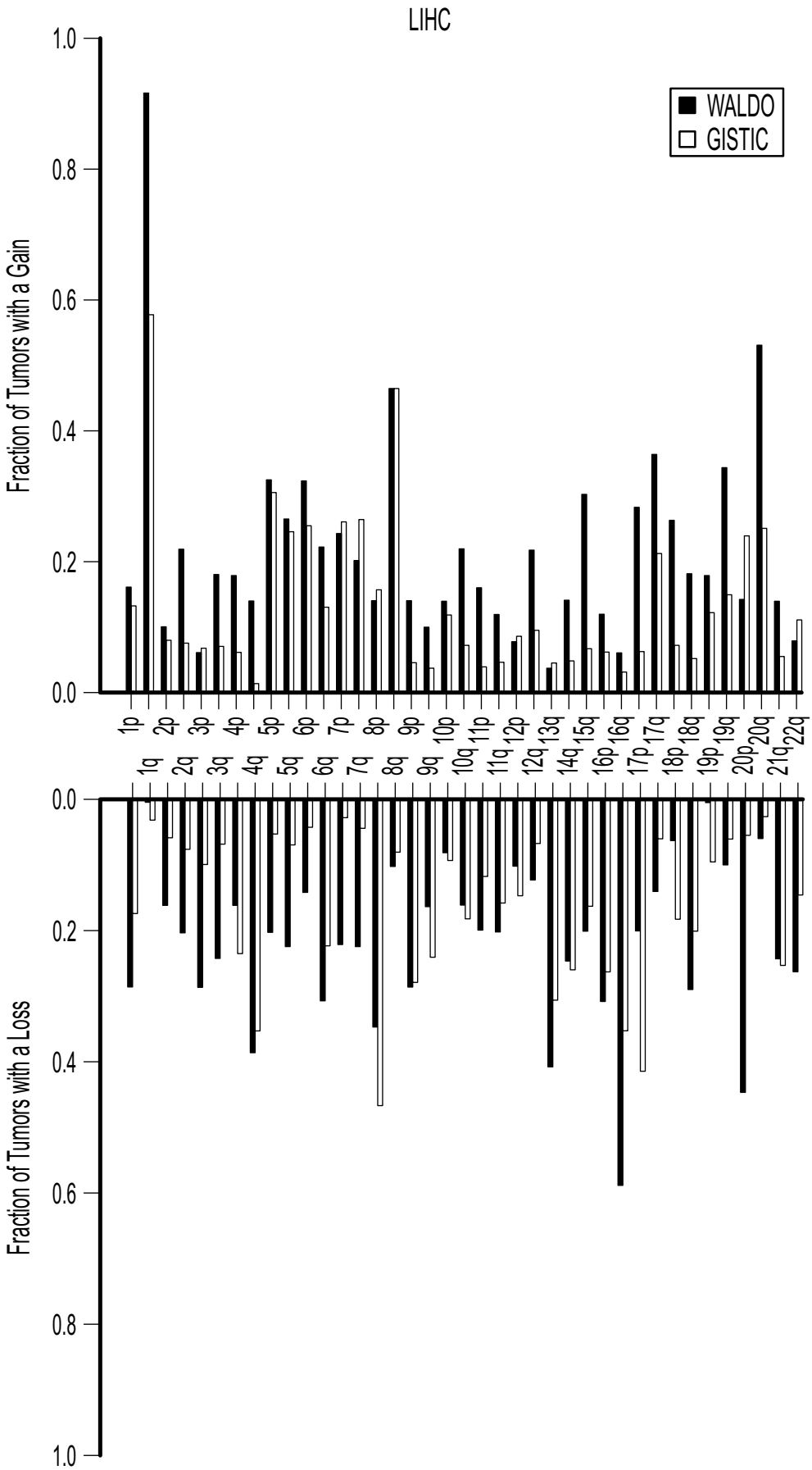
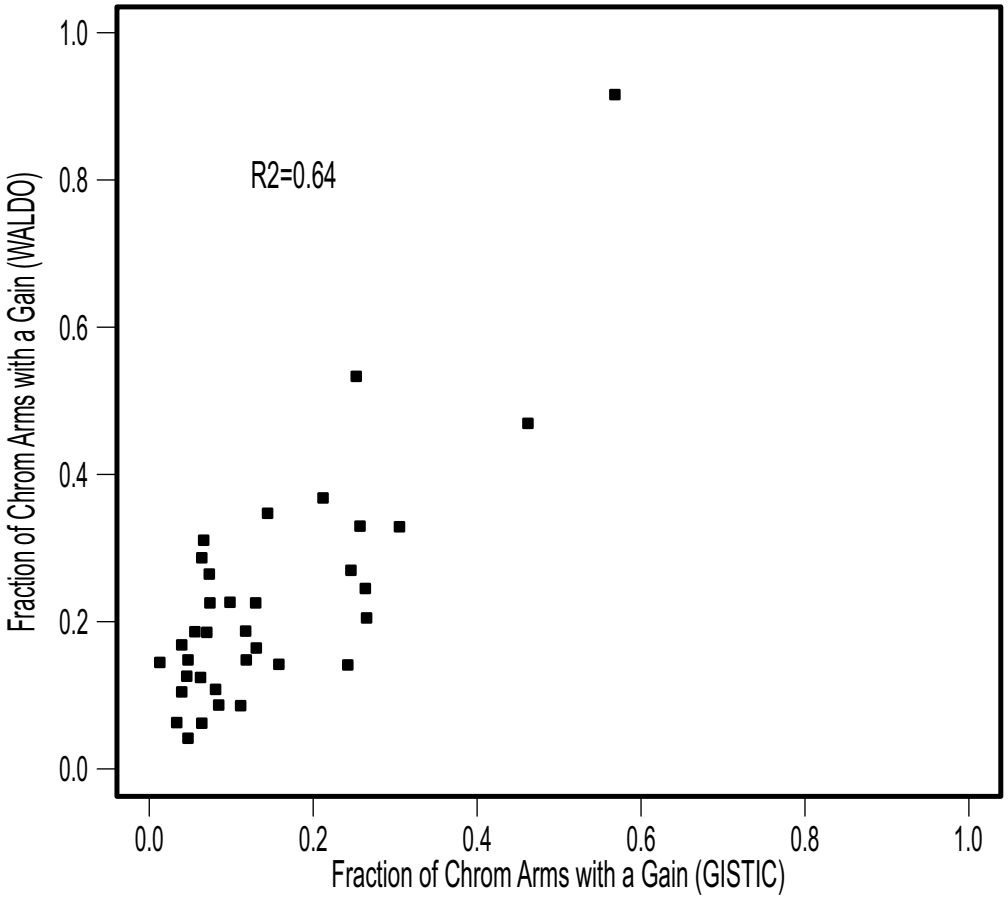

LIHC
WALDO
GISTIC
Fraction of Tumors with a Gain
Fraction of Tumors with a Loss
1.0
0.8
0.6
0.4
0.2
0.0
1p 2p 3p 4p 5p 6p 7p 8p 9p 10p 11p 12p 13q 14q 15q 16p 16q 17p 17q 18p 18q 19p 19q 20p 20q 21q 22q
1q 2q 3q 4q 5q 6q 7q 8q 9q 10q 11q 12q

FIG. 40E


R2=0.64
Fraction of Chrom Arms with a Gain (WALDO)
Fraction of Chrom Arms with a Gain (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

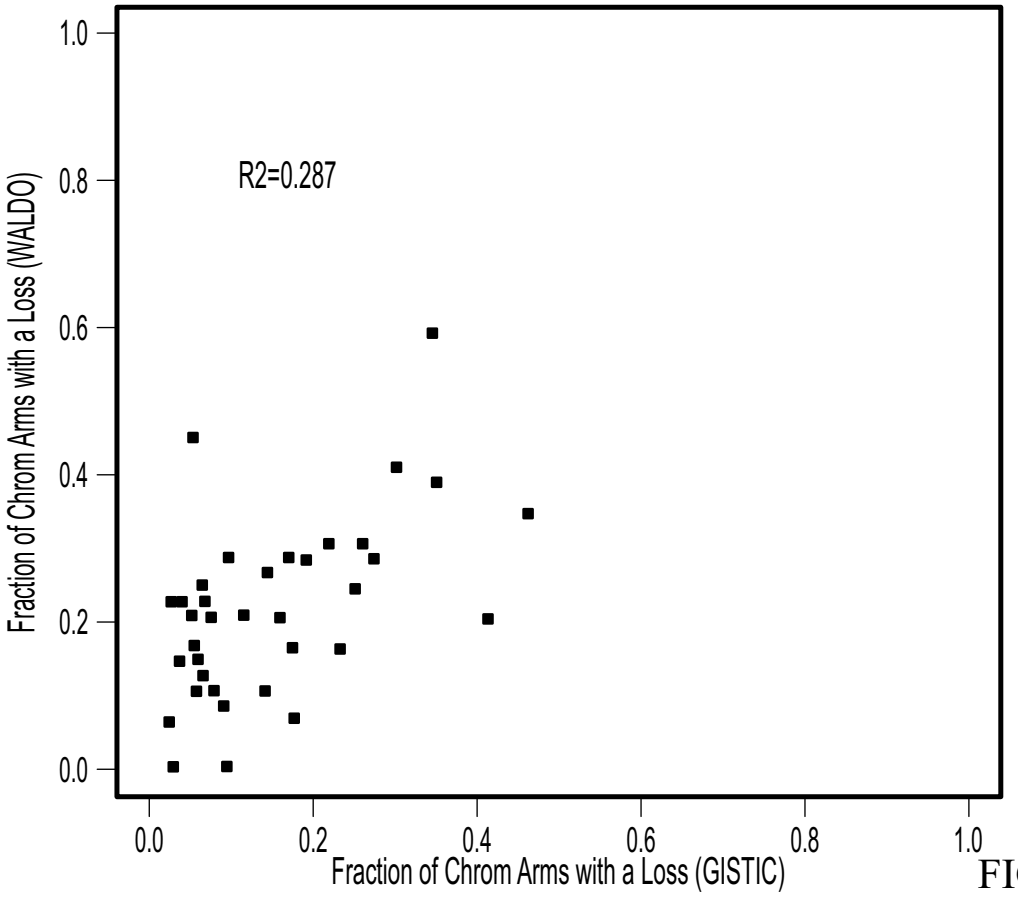

R2=0.287
Fraction of Chrom Arms with a Loss (WALDO)
Fraction of Chrom Arms with a Loss (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 40E (Cont.)

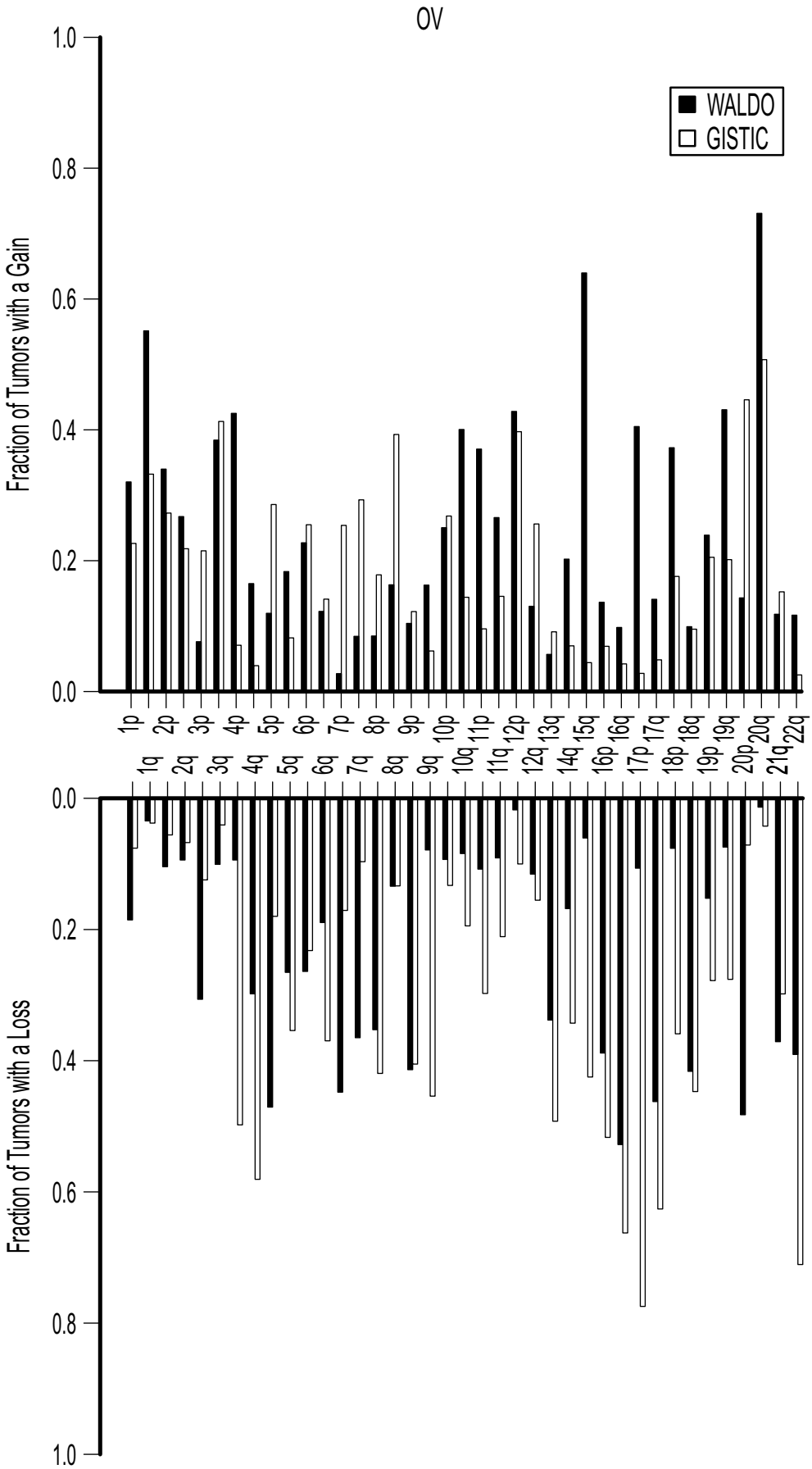
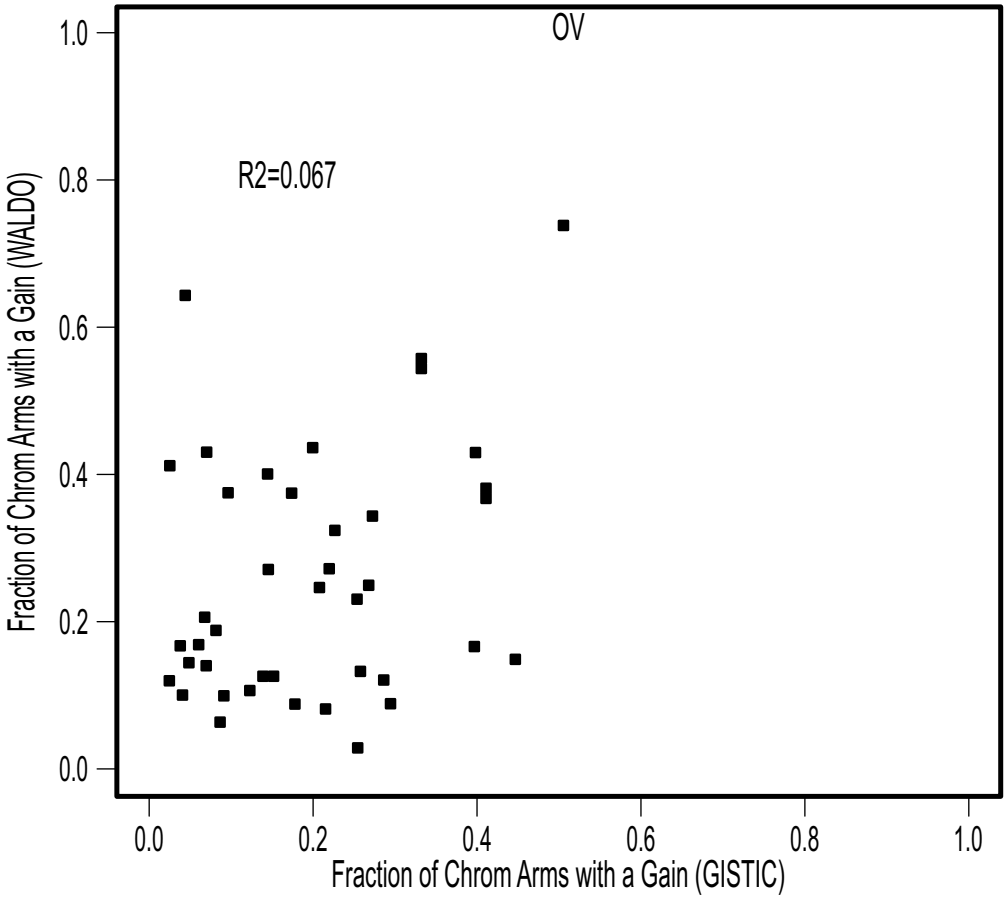

OV
WALDO
GISTIC
Fraction of Tumors with a Gain
Fraction of Tumors with a Loss
1.0
0.8
0.6
0.4
0.2
0.0
1p
2p
3p
4p
5p
6p
7p
8p
9p
10p
11p
12p
13q
15q
16p
16q
17p
17q
18p
18q
19p
19q
20p
20q
21q
22q
1q
2q
3q
4q
5q
6q
7q
8q
9q
10q
11q
12q
14q

FIG. 40F


OV
R2=0.067
Fraction of Chrom Arms with a Gain (WALDO)
Fraction of Chrom Arms with a Gain (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

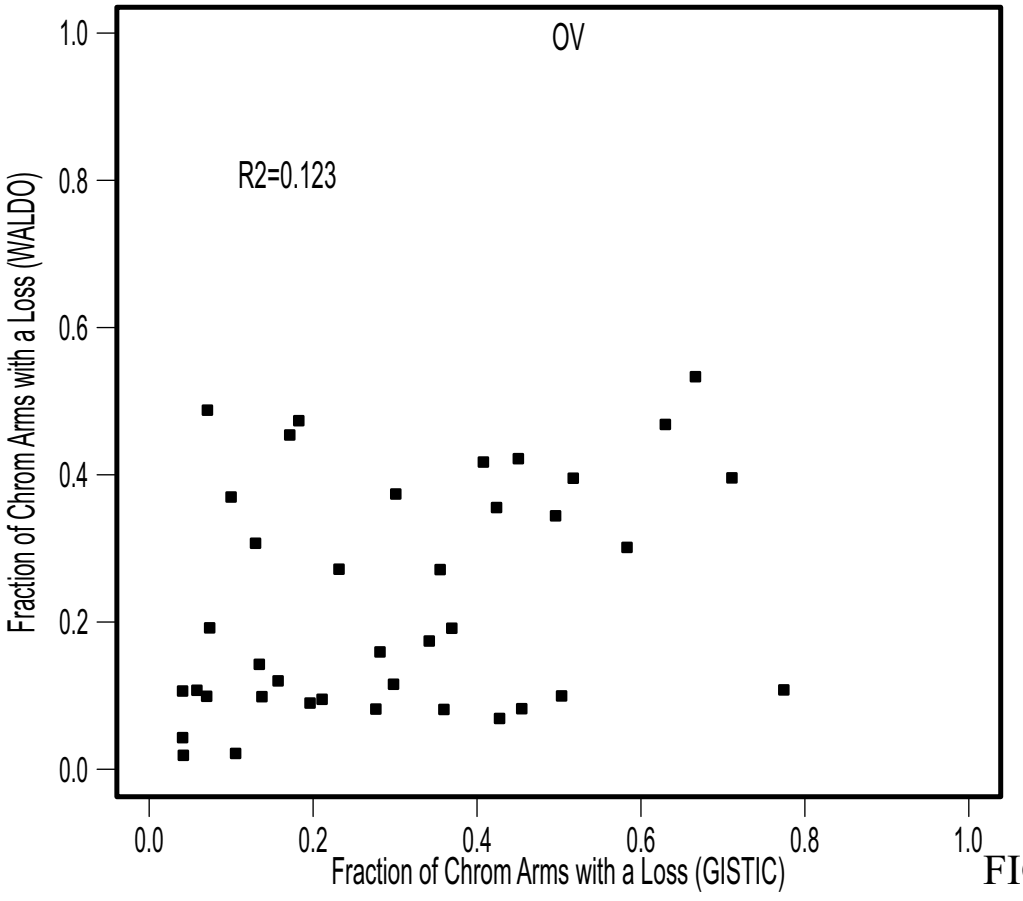

OV
R2=0.123
Fraction of Chrom Arms with a Loss (WALDO)
Fraction of Chrom Arms with a Loss (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 40F (Cont.)

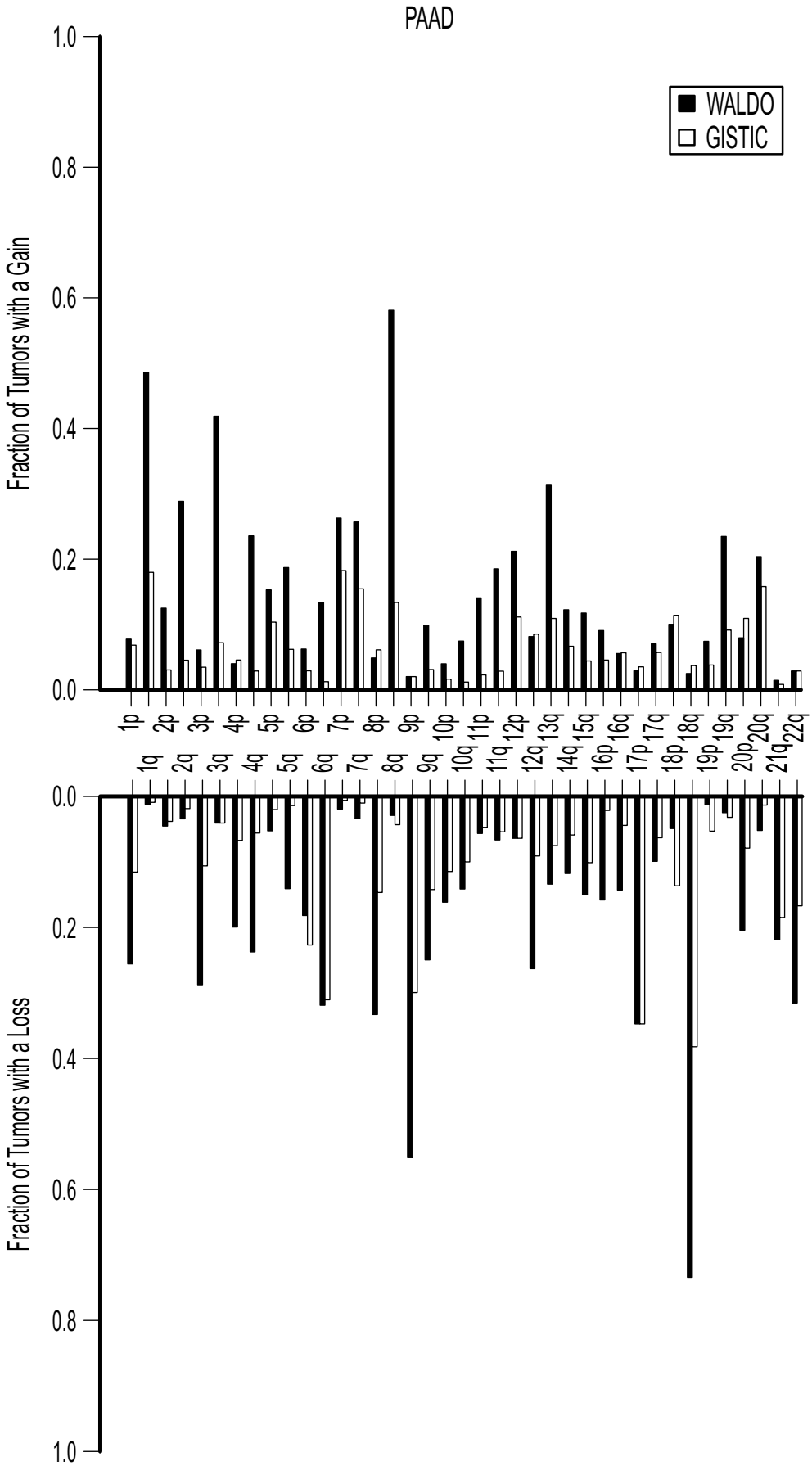
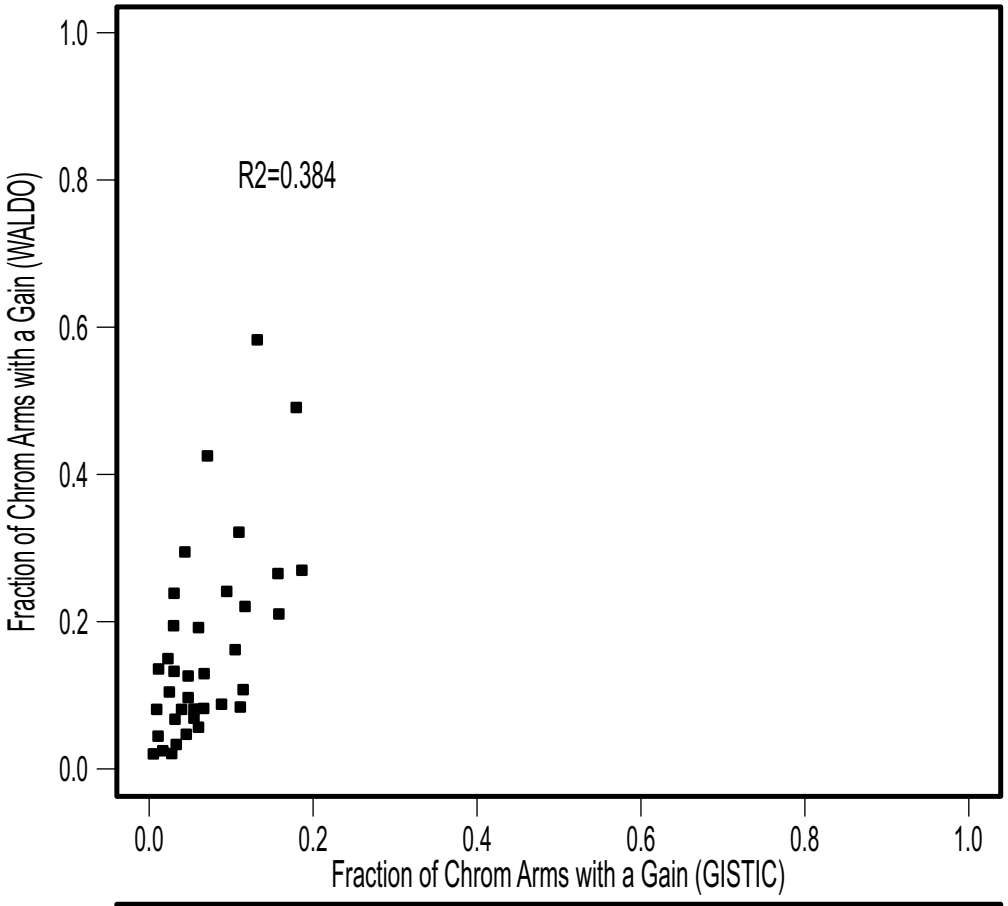

PAAD
WALDO
GISTIC
Fraction of Tumors with a Gain
Fraction of Tumors with a Loss
1.0
0.8
0.6
0.4
0.2
0.0
1p 2p 3p 4p 5p 6p 7p 8p 9p 10p 11p 12p 13q 15q 16q 17q 18q 19q 20q 22q
1q 2q 3q 4q 5q 6q 7q 8q 9q 10q 11q 12q 13q 14q 15q 16p 17p 18p 19p 20p 21q

FIG. 40G


R2=0.384
Fraction of Chrom Arms with a Gain (WALDO)
Fraction of Chrom Arms with a Gain (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

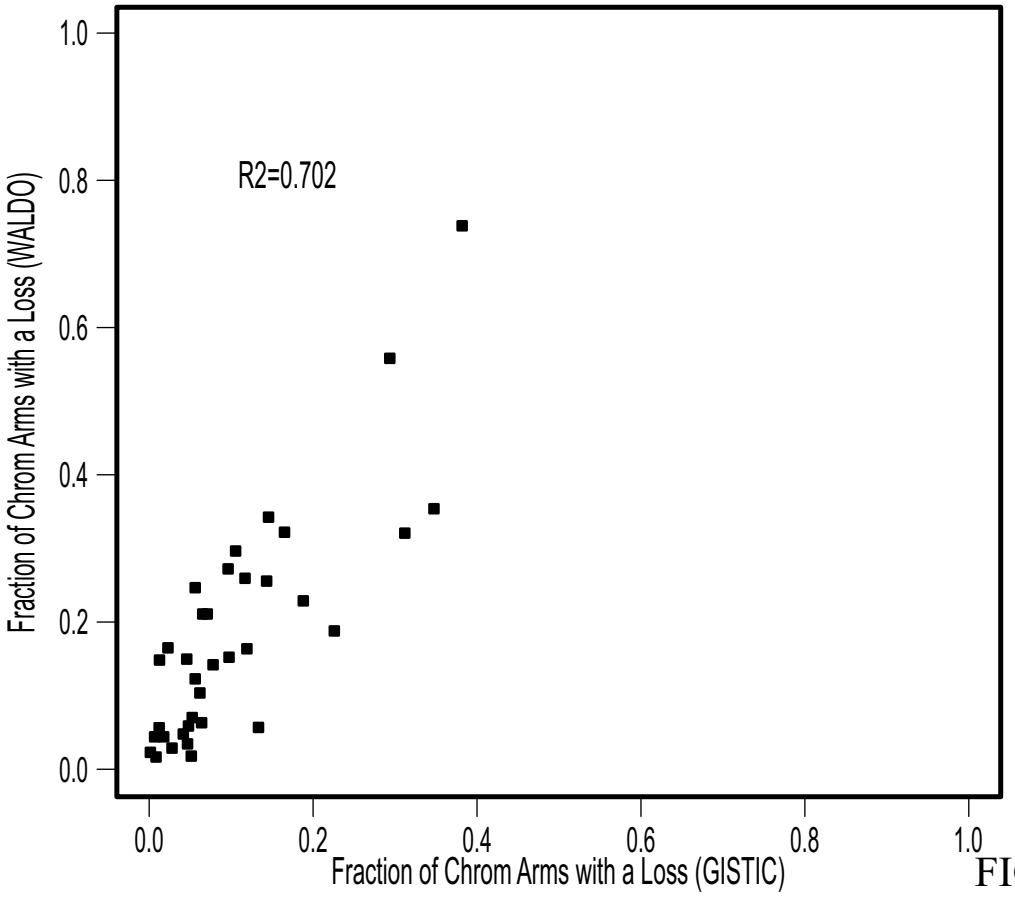

R2=0.702
Fraction of Chrom Arms with a Loss (WALDO)
Fraction of Chrom Arms with a Loss (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 40G (Cont.)

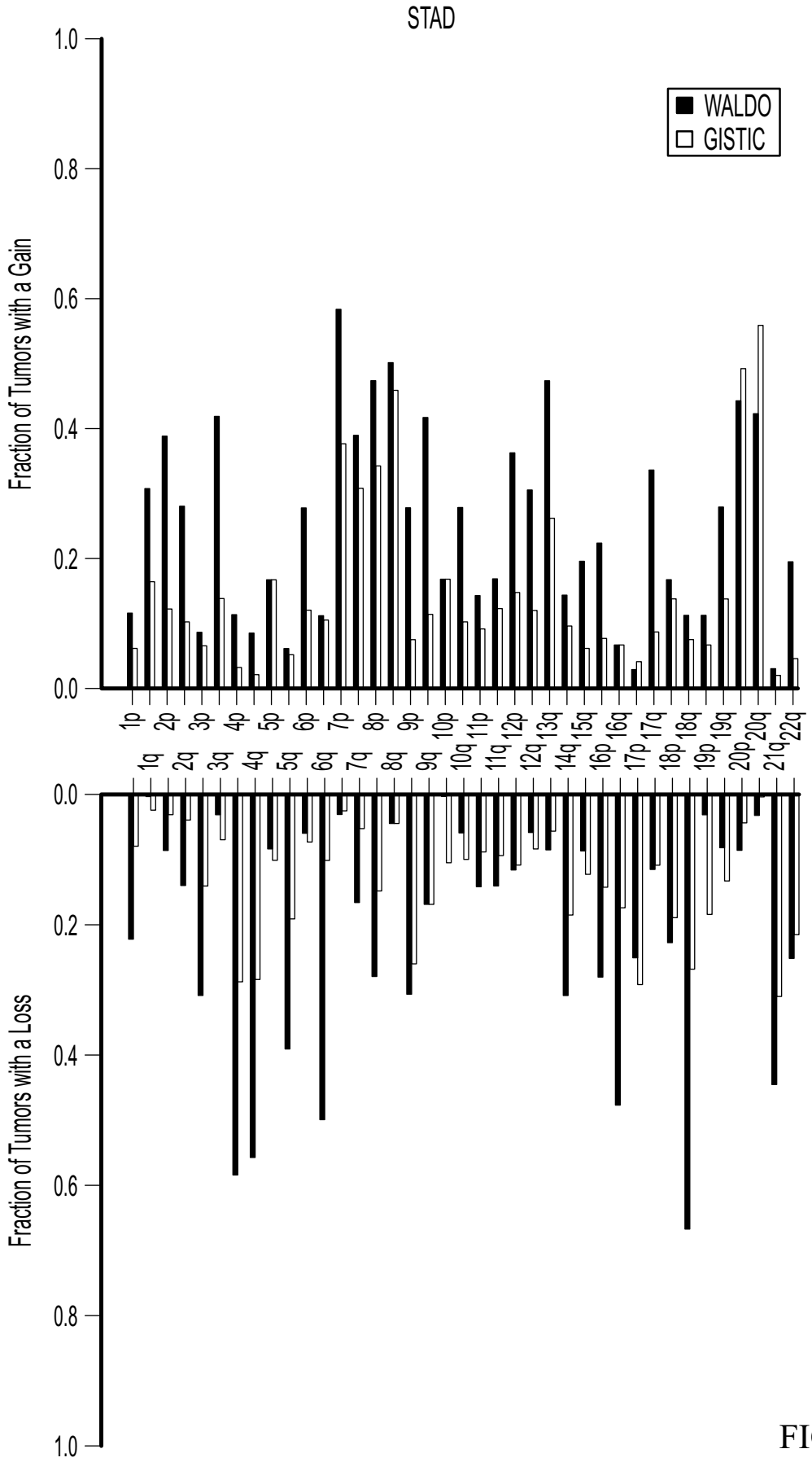
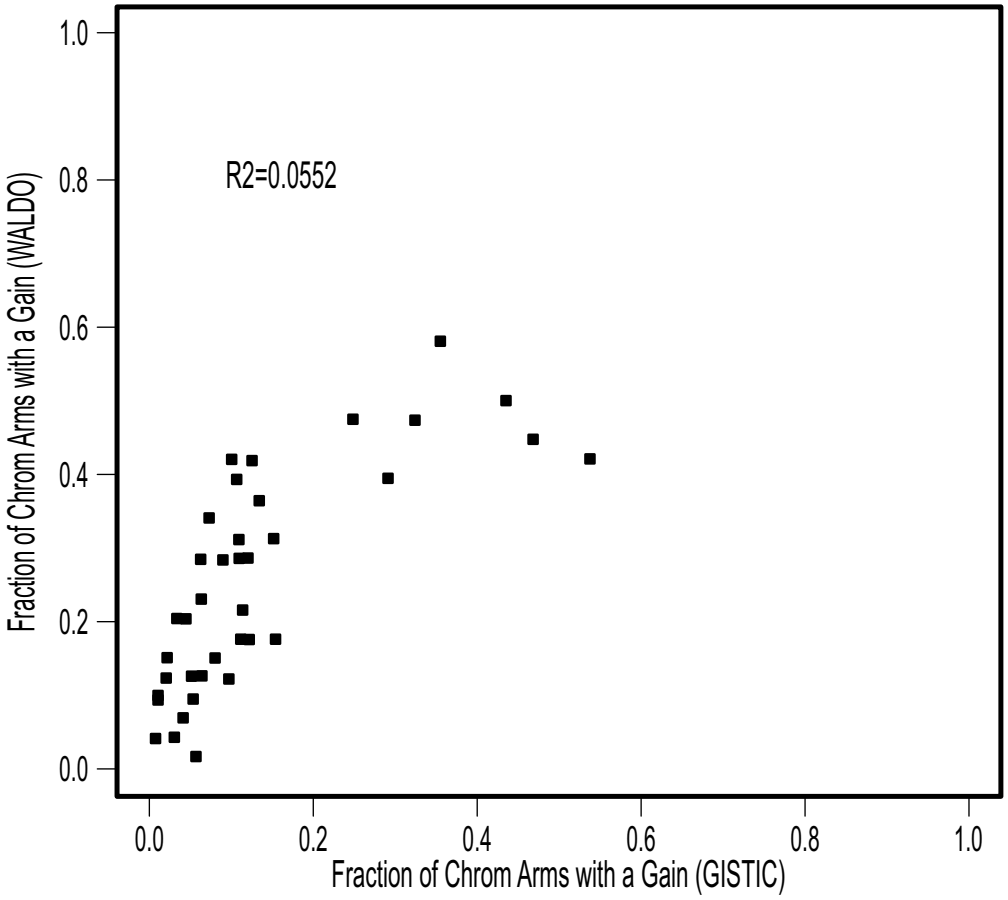

STAD
WALDO
GISTIC
Fraction of Tumors with a Gain
Fraction of Tumors with a Loss
1.0
0.8
0.6
0.4
0.2
0.0
1p 2p 3p 4p 5p 6p 7p 8p 9p 10p 11p 12p 13q 15q 16q 17q 18q 19q 20q 22q
1q 2q 3q 4q 5q 6q 7q 8q 9q 10q 11q 12q 14q 16p 17p 18p 19p 20p 21q

FIG. 40H


R2=0.0552
Fraction of Chrom Arms with a Gain (WALDO)
Fraction of Chrom Arms with a Gain (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

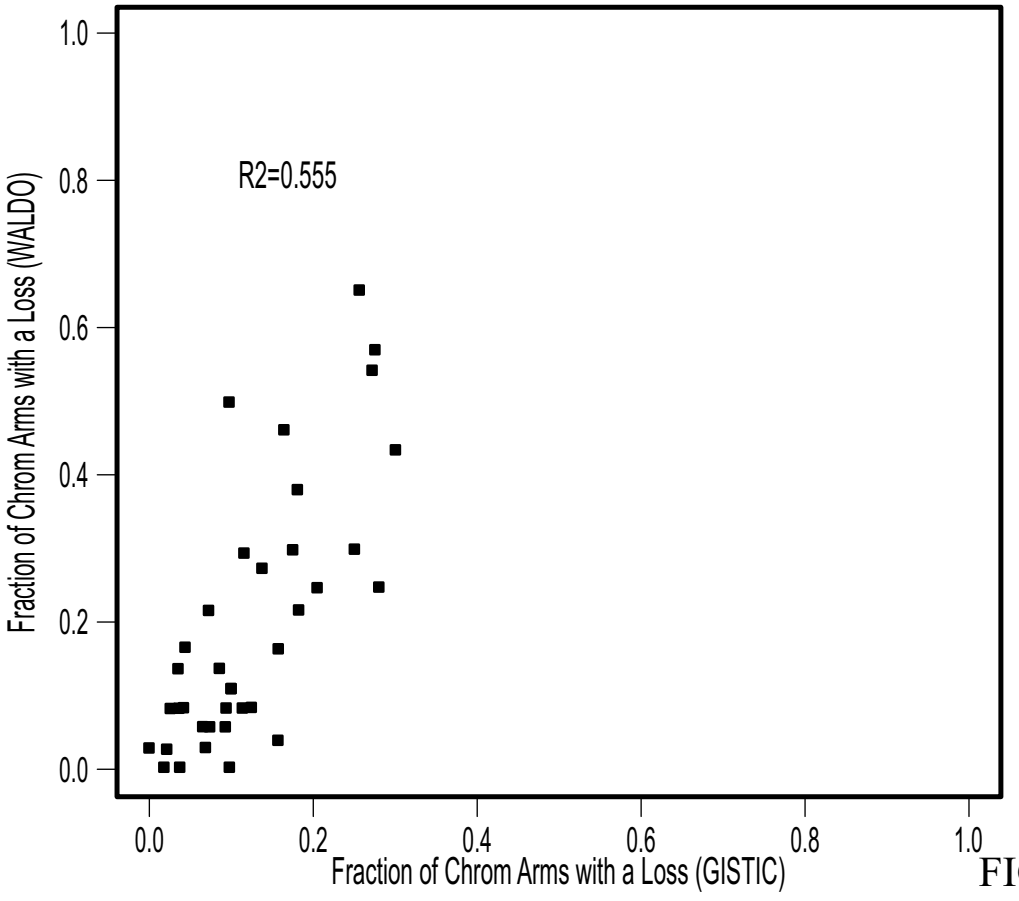

R2=0.555
Fraction of Chrom Arms with a Loss (WALDO)
Fraction of Chrom Arms with a Loss (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 40H (Cont.)

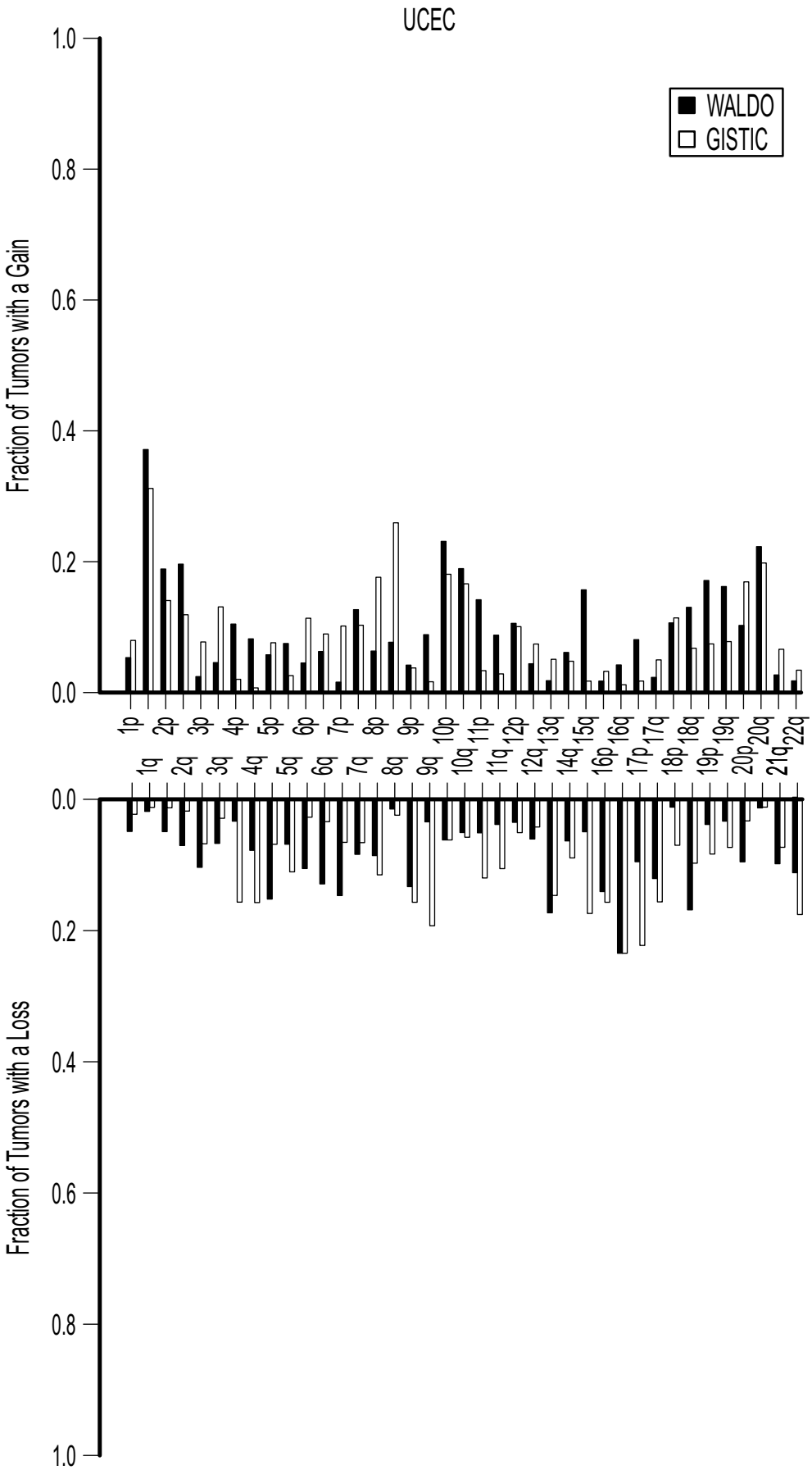
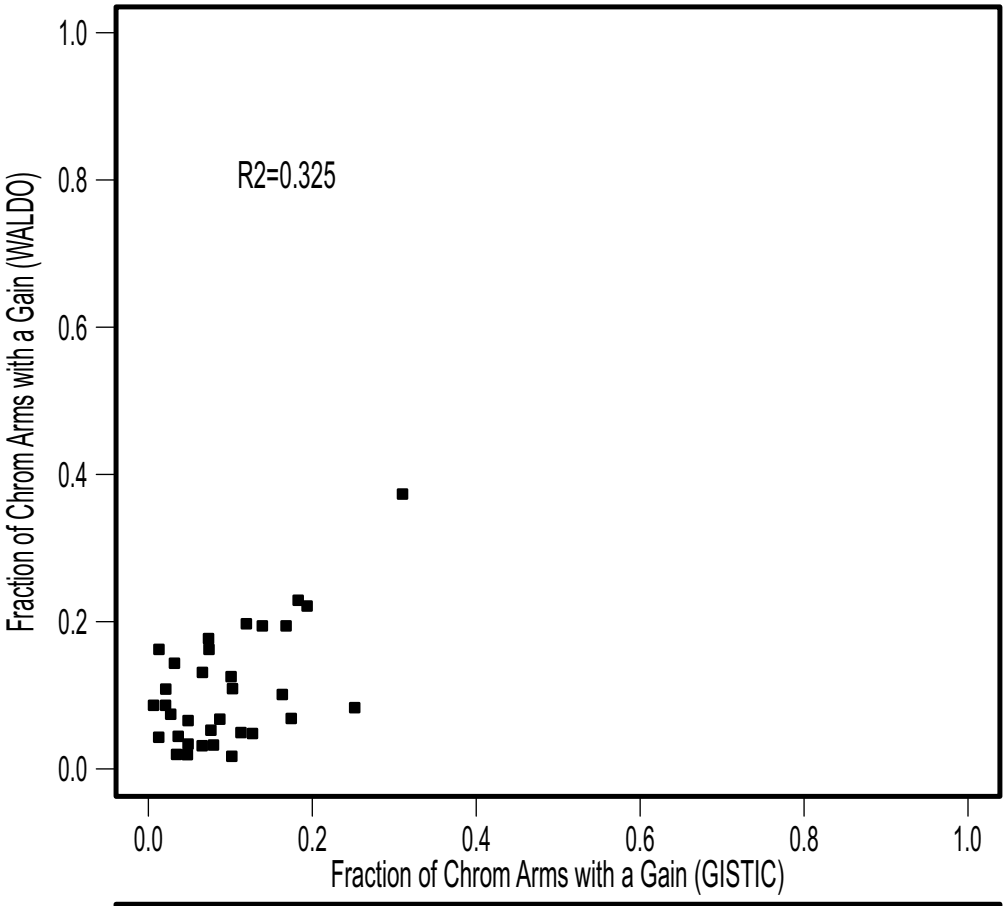

UCEC
WALDO
GISTIC
Fraction of Tumors with a Gain
Fraction of Tumors with a Loss
1.0
0.8
0.6
0.4
0.2
0.0
1p 2p 3p 4p 5p 6p 7p 8p 9p 10p 11p 12p 13q 15q 16q 17q 18q 19q 20q 22q
1q 2q 3q 4q 5q 6q 7q 8q 9q 10q 11q 12q 14q 16p 17p 18p 19p 20p 21q

FIG. 40I


R2=0.325
Fraction of Chrom Arms with a Gain (WALDO)
Fraction of Chrom Arms with a Gain (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

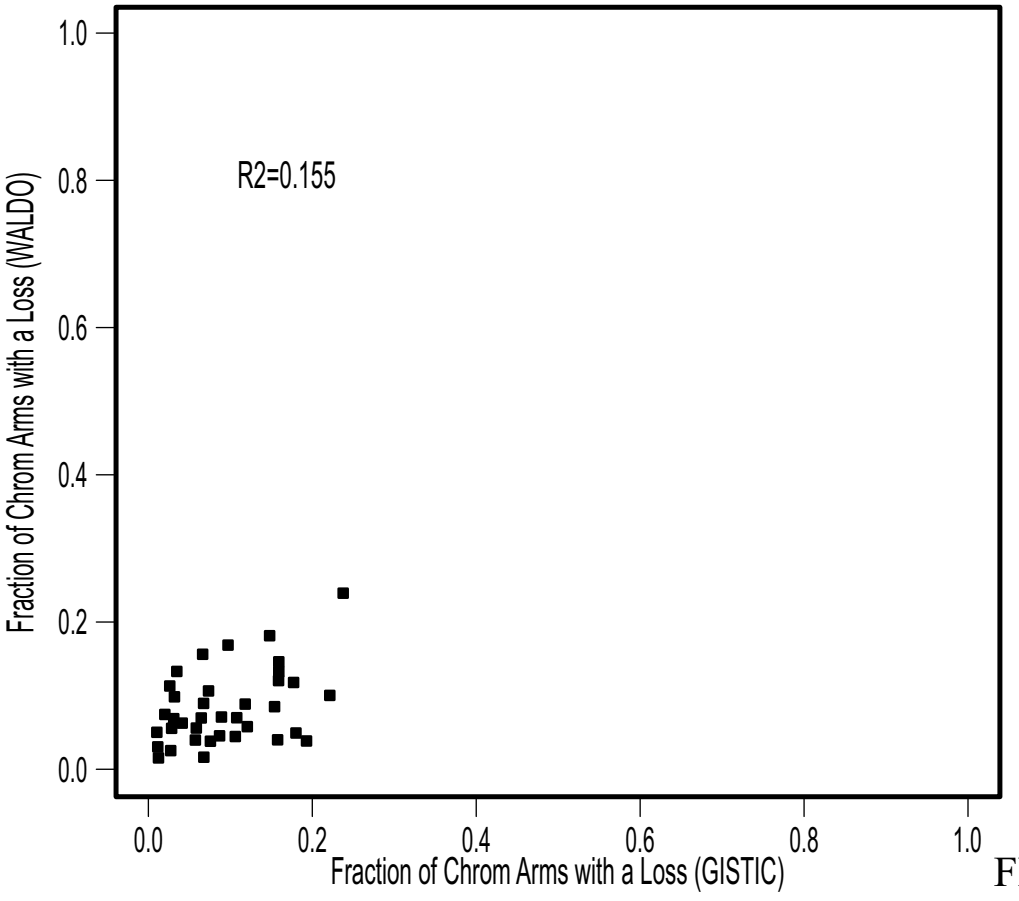

R2=0.155
Fraction of Chrom Arms with a Loss (WALDO)
Fraction of Chrom Arms with a Loss (GISTIC)
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 40I (Cont.)

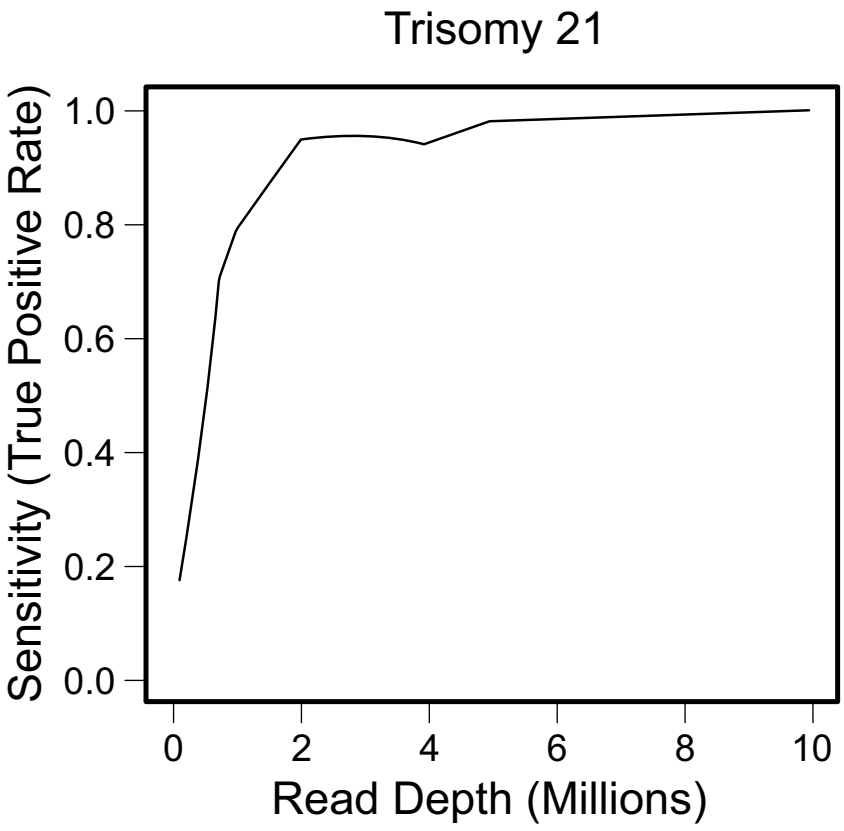


FIG. 41A

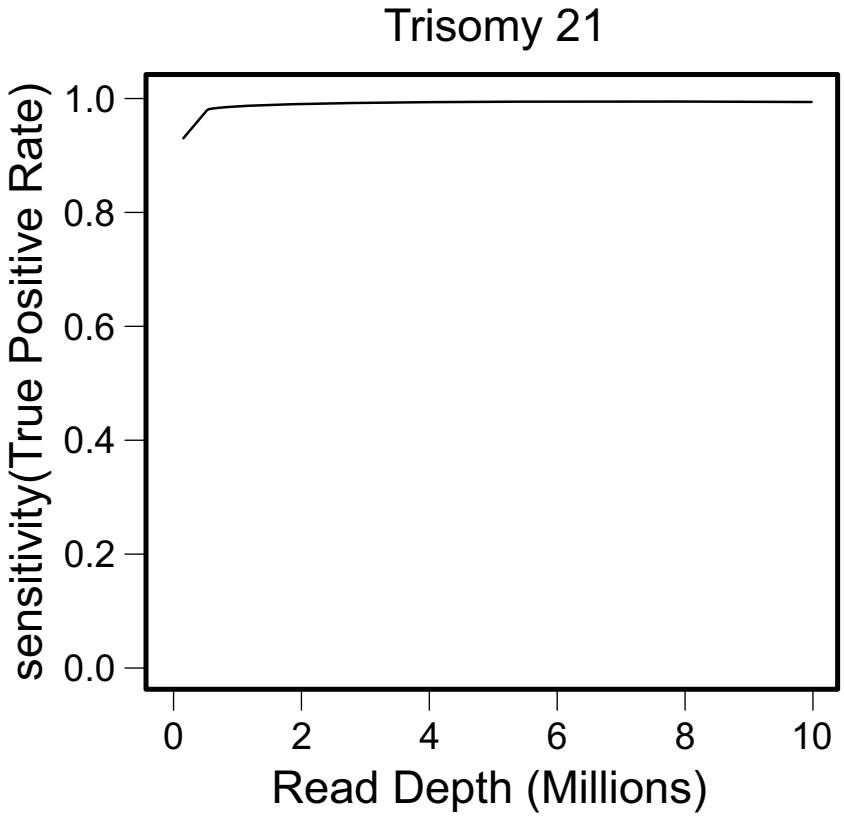


FIG. 41B

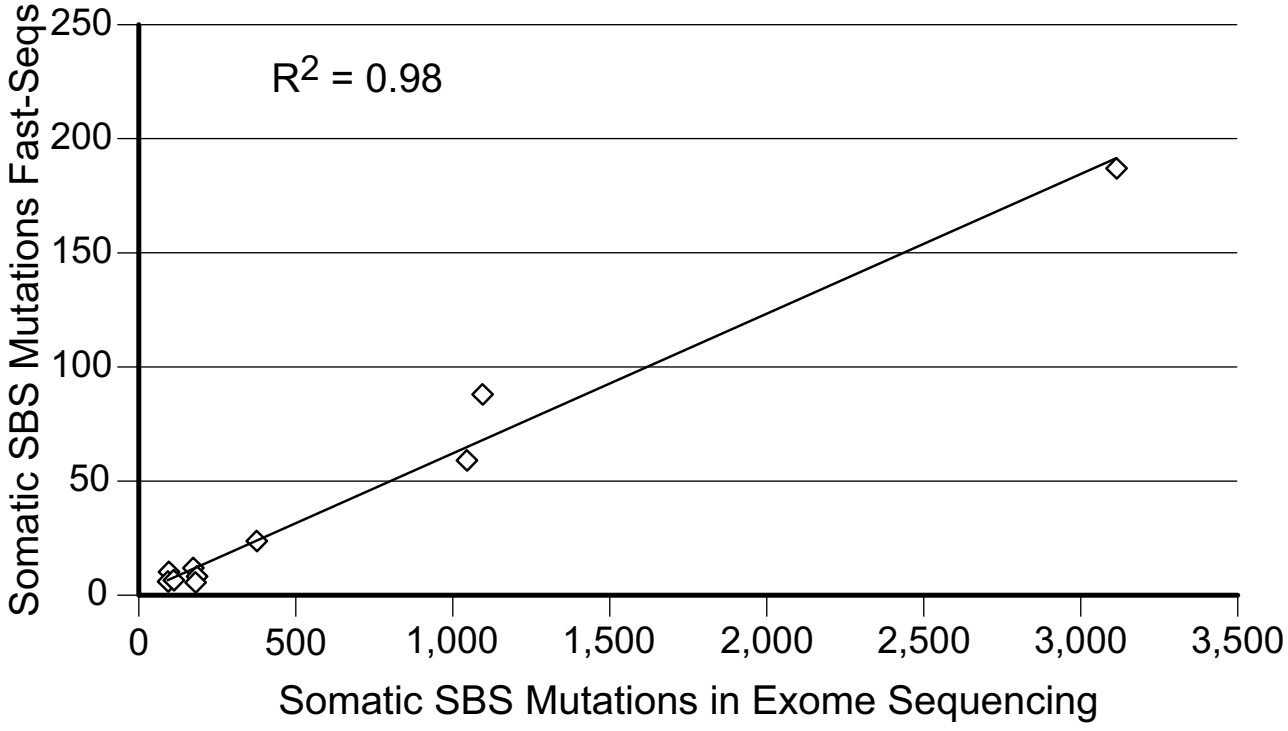

Somatic SBS Mutations Fast-Seqs
250
200
150
100
50
0
R^2 = 0.98
0
500
1,000
1,500
2,000
2,500
3,000
3,500
Somatic SBS Mutations in Exome Sequencing

FIG. 42

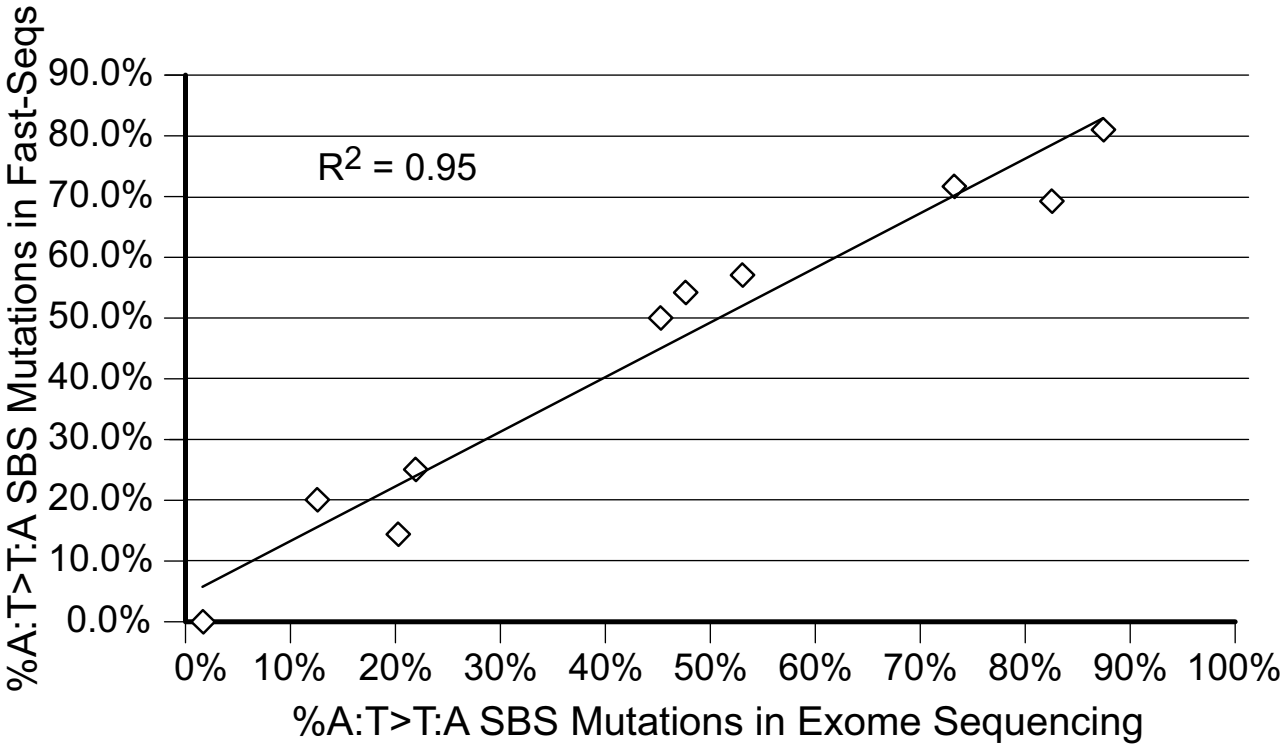

%A:T>T:A SBS Mutations in Fast-Seqs
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
R2 = 0.95
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
%A:T>T:A SBS Mutations in Exome Sequencing

FIG. 43

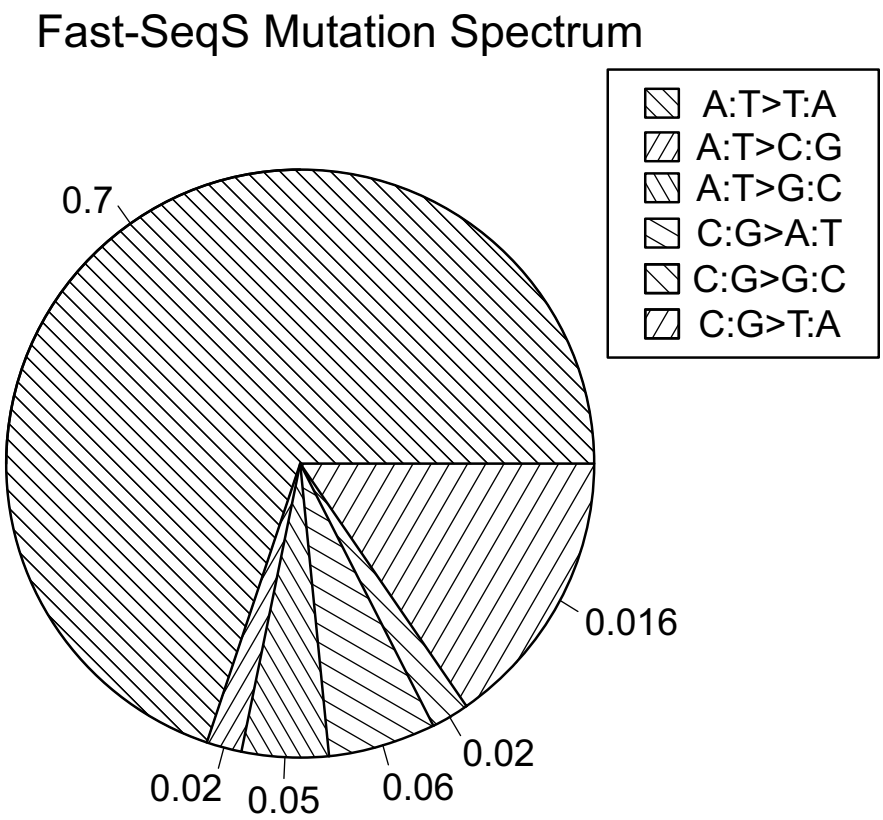


FIG. 44A

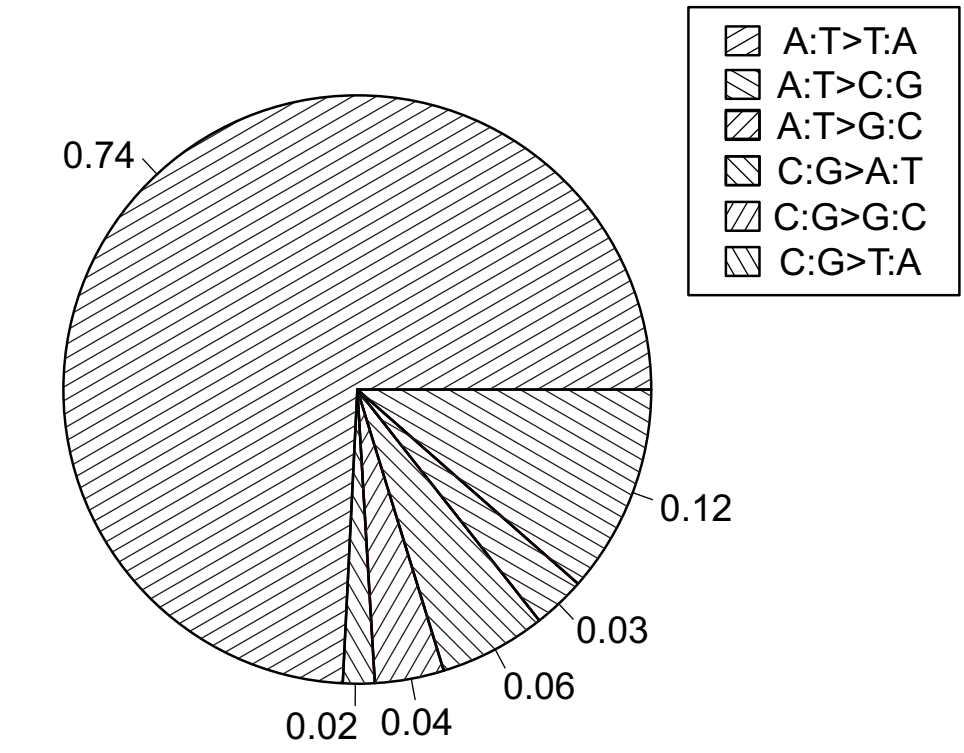


FIG. 44B

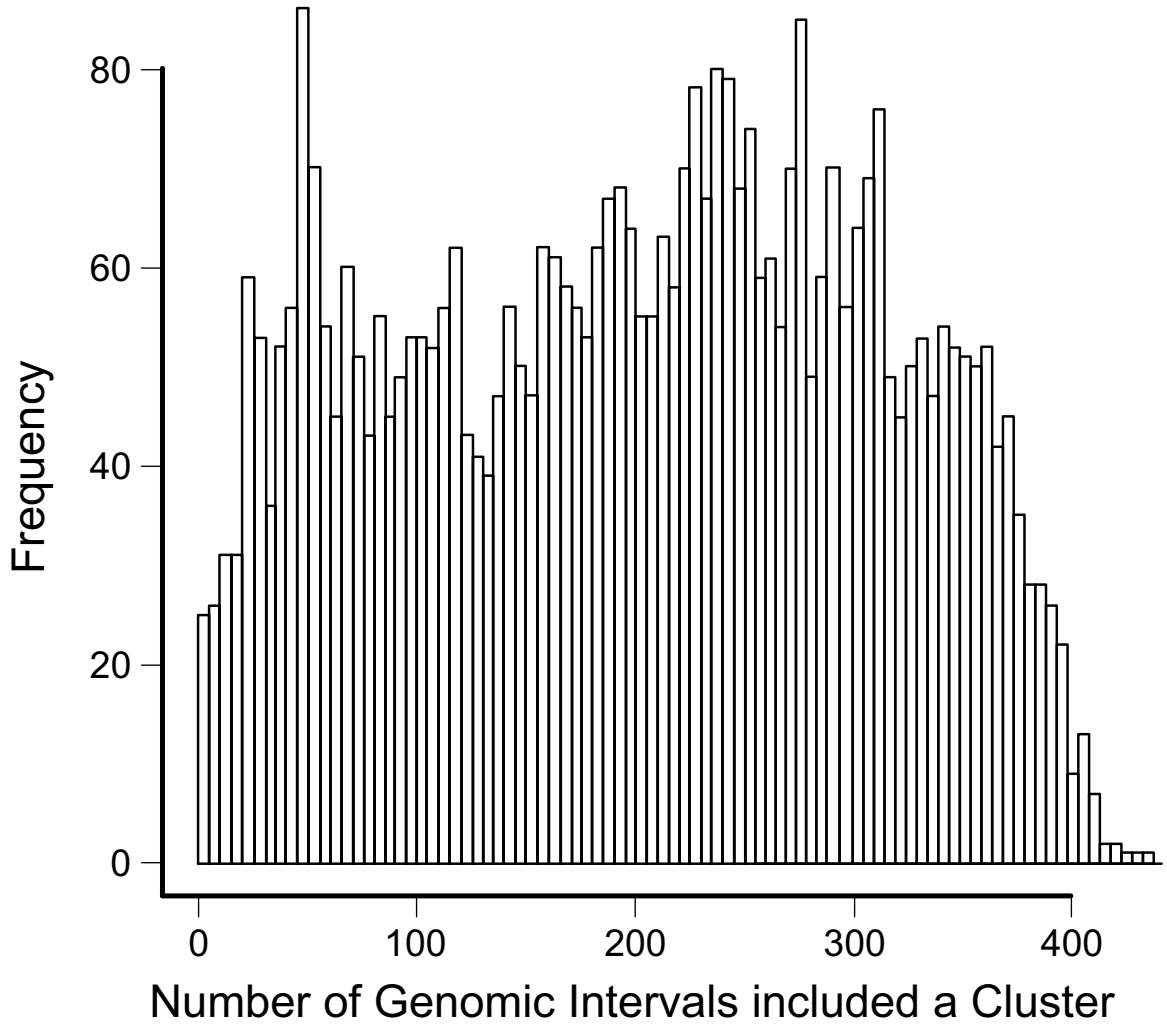

Frequency
80
60
40
20
0
0
100
200
300
400
Number of Genomic Intervals included a Cluster

FIG. 45

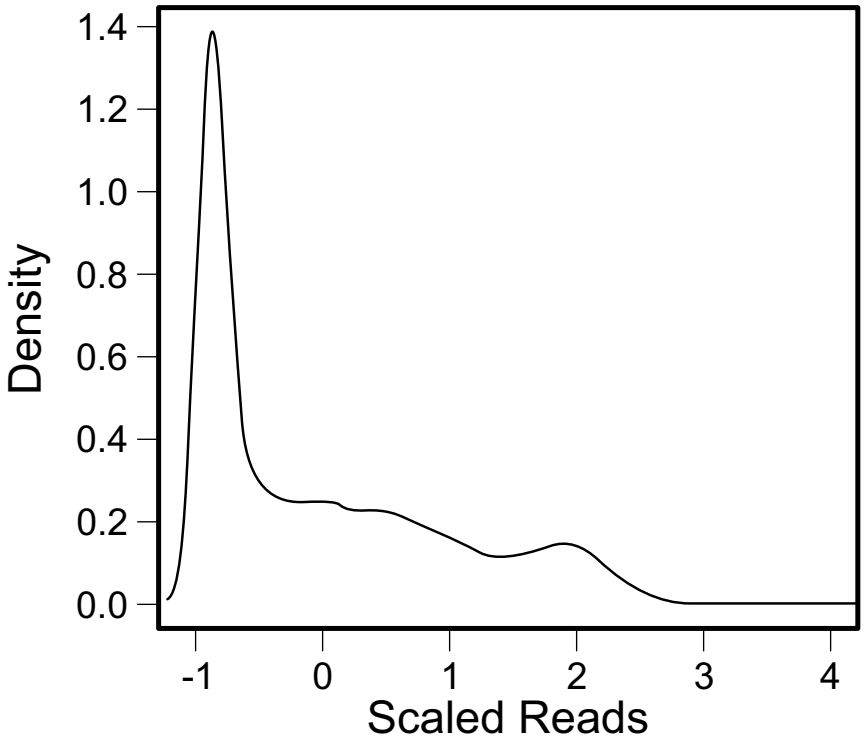


FIG. 46A

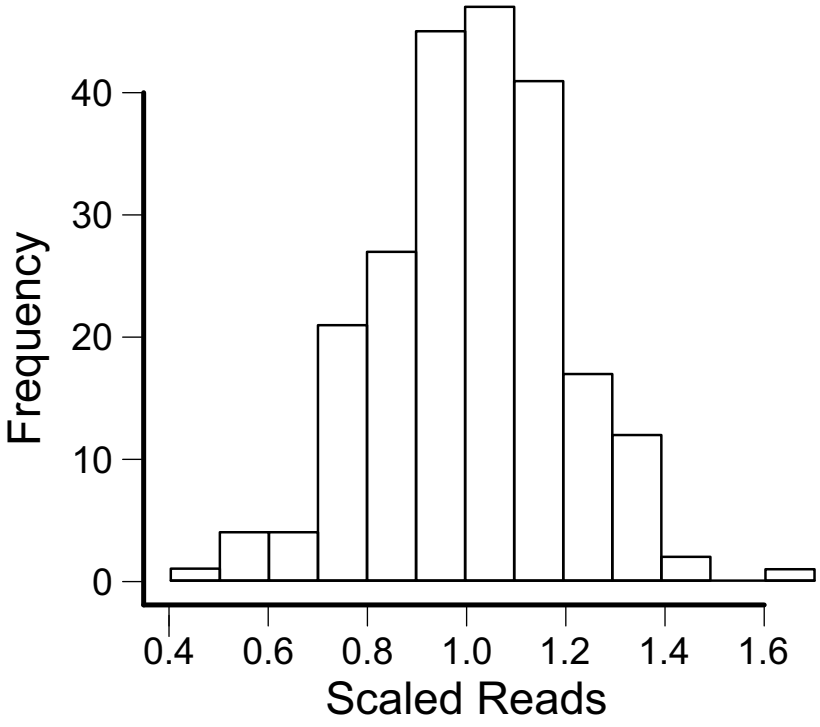


FIG. 46B

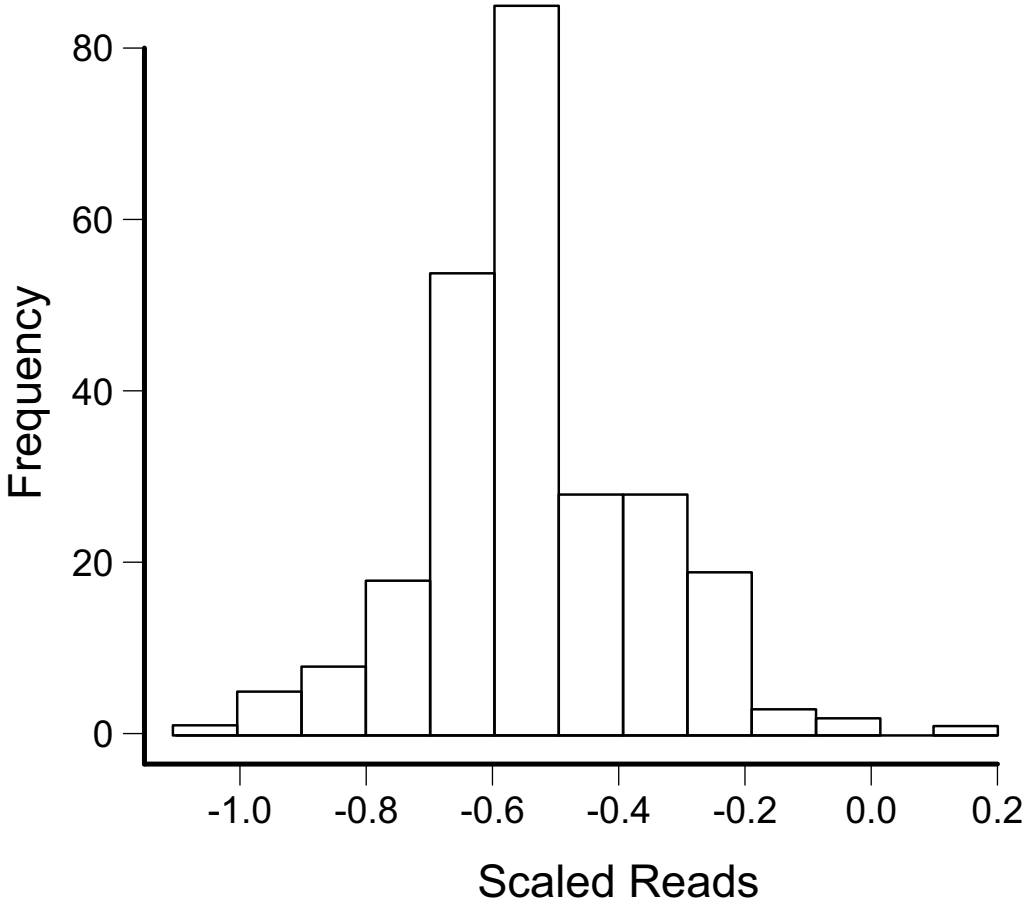
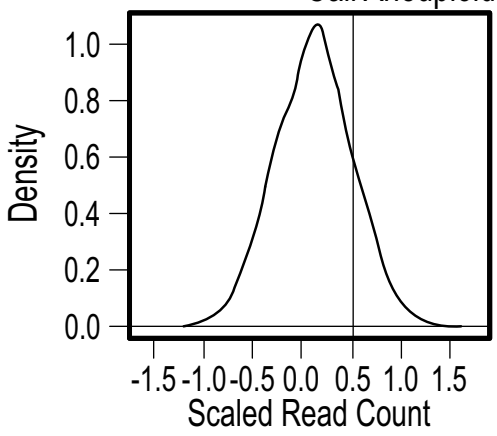
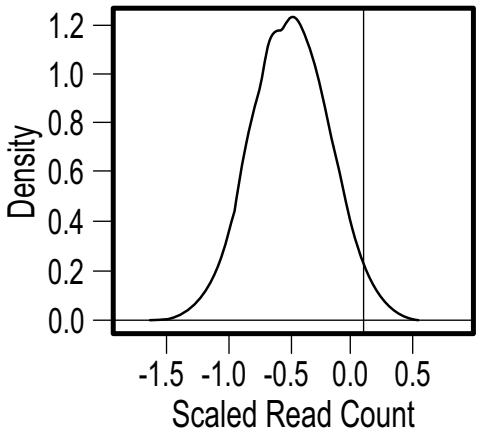


FIG. 46C

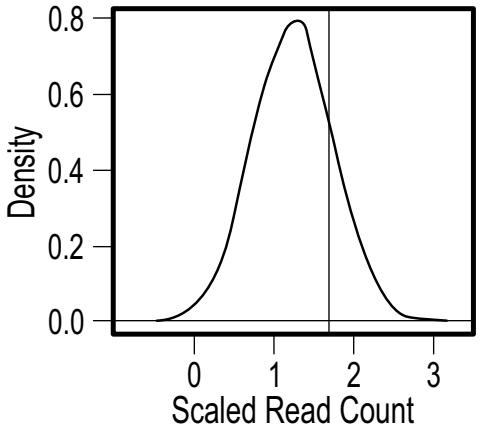
Call Aneuploidy on a Chromosome Arm



500kb Genomic Interval 1
Estimate parameters
$R_i \sim N(\mu_i, \sigma_i^2)$
$R_1 \sim N(\mu_1, \sigma_1^2)$
$R_1 = 0.5$
$\mu_1 = 0.1$
$\sigma_i^2 = 0.15$
$2*(1\text{-}CDF(\mu_1 = 0.15, \sigma_1^2 = 0.15 \mid R_1 = 0.5) = 0.301 >$
$\alpha \rightarrow$ not an outlier



500kb Genomic Interval 2
Estimate parameters
$R_i \sim N(\mu_i, \sigma_i^2)$
$R_2 \sim N(\mu_2, \sigma_2^2)$
$R_2 = 0.1$
$\mu_2 = -0.5$
$\sigma_i^2 = 0.1$
$2*(1\text{-}CDF(\mu_2 = -0.5, \sigma_2^2 = 0.1 \mid R_2 = 0.1) = 0.058 >$
$\alpha \rightarrow$ not an outlier



500kb Genomic Interval 3
Estimate parameters
$R_3 \sim N(\mu_3, \sigma_3^2)$
$R_3 = 1.7$
$\mu_3 = 1.25$
$\sigma_3^2 = 0.25$
$2*(1\text{-}CDF(\mu_3 = 1.25, \sigma_3^2 = 0.25 \mid R_3 = 1.7) = 0.37 >$
$\alpha \rightarrow$ not an outlier

Chromosome Arm

$$\sum_1^i R_i \sim N\left(\sum_1^i \mu_i, \sum_1^i \sigma_i^2\right)$$

$$\sum_1^i R_i = (0.5 + 0.1 + 1.7) = 2.3$$

$$\sum_1^i \mu_i = 0.1 + (-0.5) + 1.25 = 0.85$$

$$\sum_1^i \sigma_i^2 = 0.15 + 0.1 + 0.25 = 0.5$$

$1\text{-}CDF(\Sigma_1^i \mu_i = 0.85, \Sigma_1^i \sigma_i^2 = 0.5 \mid \Sigma_1^i R_i = 2.3) = 0.0201 < \alpha$
$Z = 2.05$
Chromosome Arm is gained!

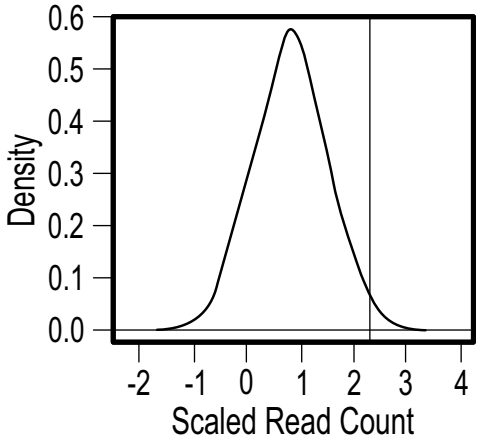


FIG. 47

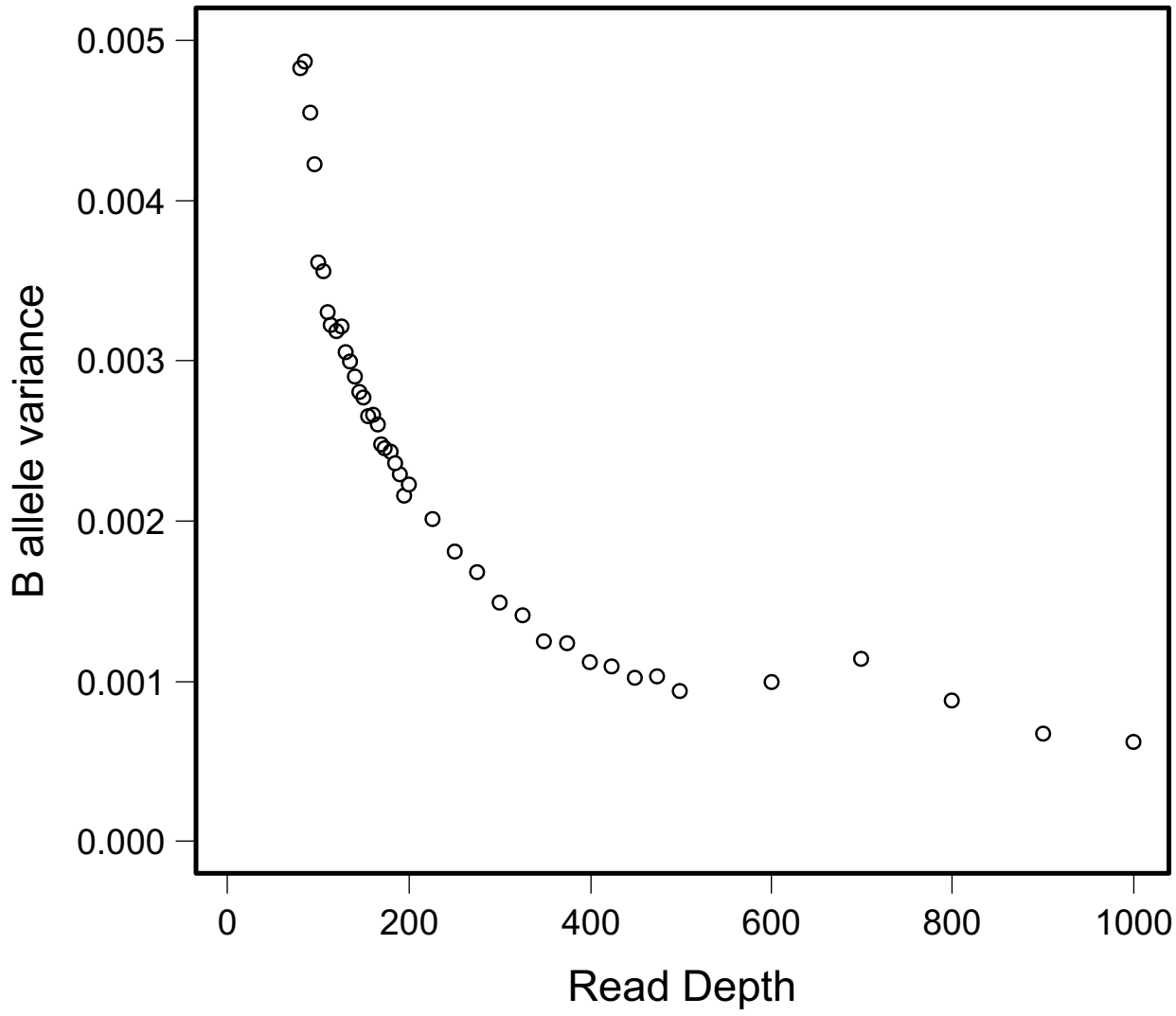

0.005
0.004
0.003
0.002
0.001
0.000
B allele variance
0
200
400
600
800
1000
Read Depth

FIG. 48

Synthetics with one arm alterations

```
N <- number of normal WBC samples used
f <- desired neoplastic cell fraction
r <- chromosome arm (1p..22q)
r_type <- alteration type (gain or loss)
p(r) <- probability of arm gain (0.5 default)
rho <- desired unique read depth of synthetic sample (9M default)
j<-desired number of repeats


For j=1:5
        For i=1:N
                s < - sample[i]
                U < - get UIDs from s
                For f in 1:F
                        For r in 1p:22q  #select alteration types and spike-in to t
                                For r_type in gain loss
                                        h <- new synthetic sample
                                        t <- copy of s
                                        u_ref <- get all reference allele UIDs that that map
                                                to chr arm r
                                        if r_type== gain
                                                t <- add duplicate u_ref reads
                                        else (r_type == loss)
                                                t <- subtract u_ref reads
                                        Endif
                                        h_normal<-random select (1-f)*rho UIDs from s
                                        h_aneuploid<-random select f*rho UIDs from t
                                        h<-h_normal+h_aneuploid
                                End
                        End
                End
```

FIG. 49

```
N <- number of normal WBC samples used
d <- desired degree of aneuploidy (number arms altered)
f <- desired neoplastic cell fraction
r <- chromosome arm (1p..22q)
r_type<-arm alteration type (gain or loss)
p(r) <- probability of arm gain (0.5 default)
rho <- desired unique read depth of synthetic sample (9M default)
j<- desired number of repeats


For j=1:5
        For i=1:N
                s <- sample[i]
                U <- get UIDs from s
                For f in 1:F
                        h <- new synthetic sample
                        For d in (5,10,15,20,25) #desired numbers of altered arms
                                t <- copy of s
                                For a in 1:d  #select alteration types and spike-in to t
                                        r <- random select arm
                                        r_type <- alteration type with p(r)
                                        u_ref <- get all reference allele UIDs that that map
                                                to chr arm r
                                        if r_type== gain
                                                t <- add duplicate u_ref reads
                                        else (r_type == loss)
                                                t <- subtract u_ref reads
                                        Endif
                                        h_normal<-random select (1-f)*rho UIDs from s
                                        h_aneuploid<-random select f*rho UIDs from t
                                        h<-h_normal+h_aneuploid
                                End
                        End
                End
        End
End
```

FIG. 50

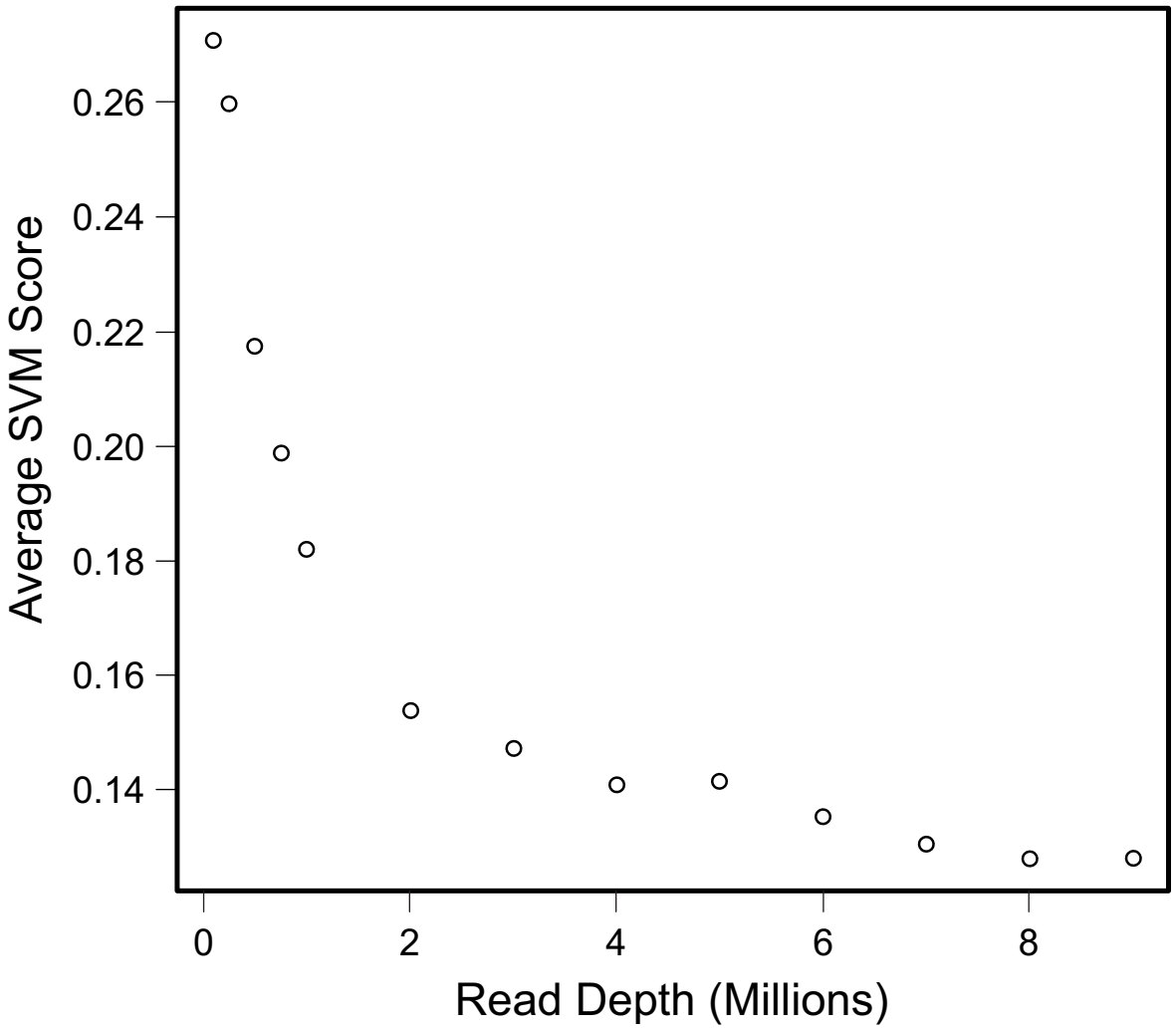

Average SVM Score
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0
2
4
6
8
Read Depth (Millions)

FIG. 51

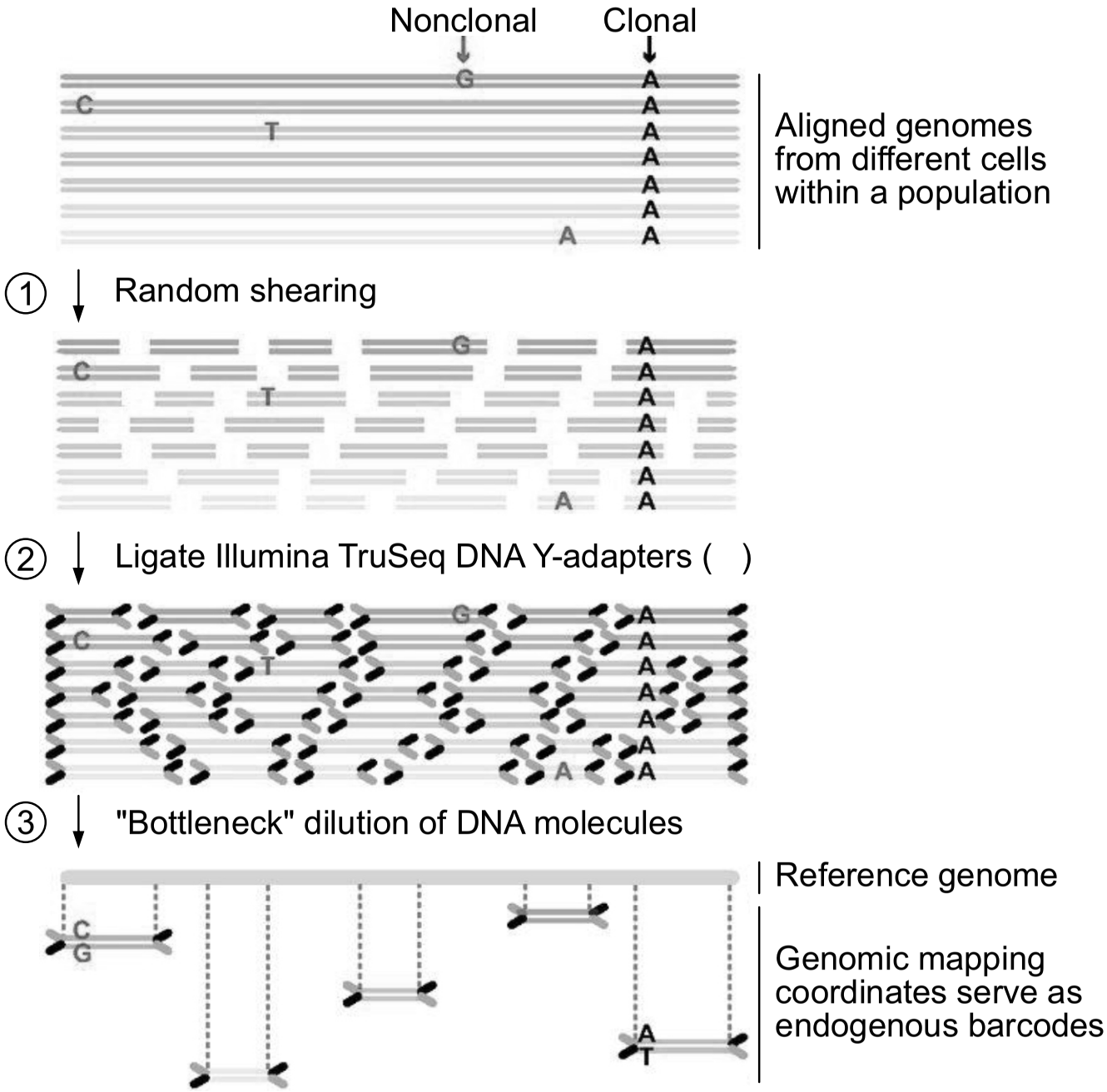

Nonclonal
Clonal
Aligned genomes from different cells within a population
① Random shearing
② Ligate Illumina TruSeq DNA Y-adapters ( )
③ "Bottleneck" dilution of DNA molecules
Reference genome
Genomic mapping coordinates serve as endogenous barcodes

FIG. 52

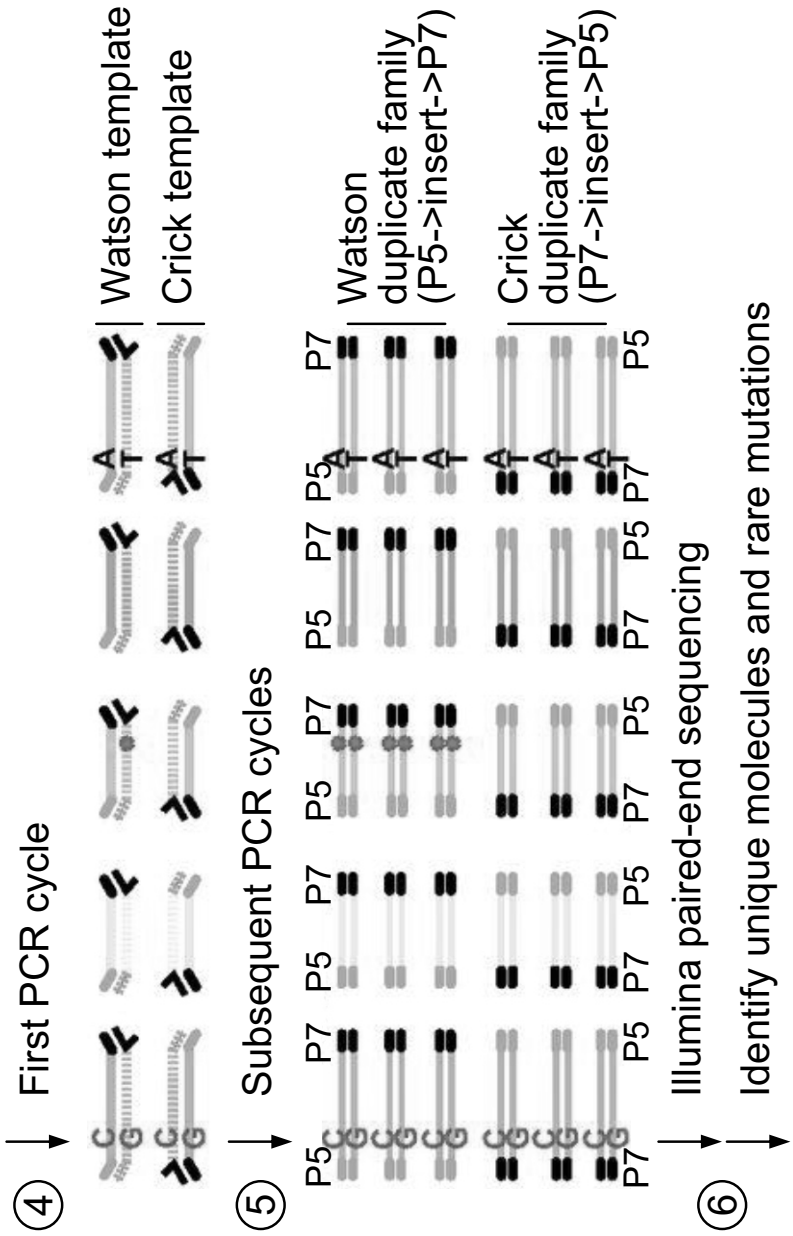

④ First PCR cycle
Watson template
Crick template
⑤ Subsequent PCR cycles
P5
P7
Watson duplicate family (P5->insert->P7)
Crick duplicate family (P7->insert->P5)
Illumina paired-end sequencing
⑥ Identify unique molecules and rare mutations
Mutation
Wild-type
Artifact
Wild-type
Clonal

FIG. 52 (Cont.)

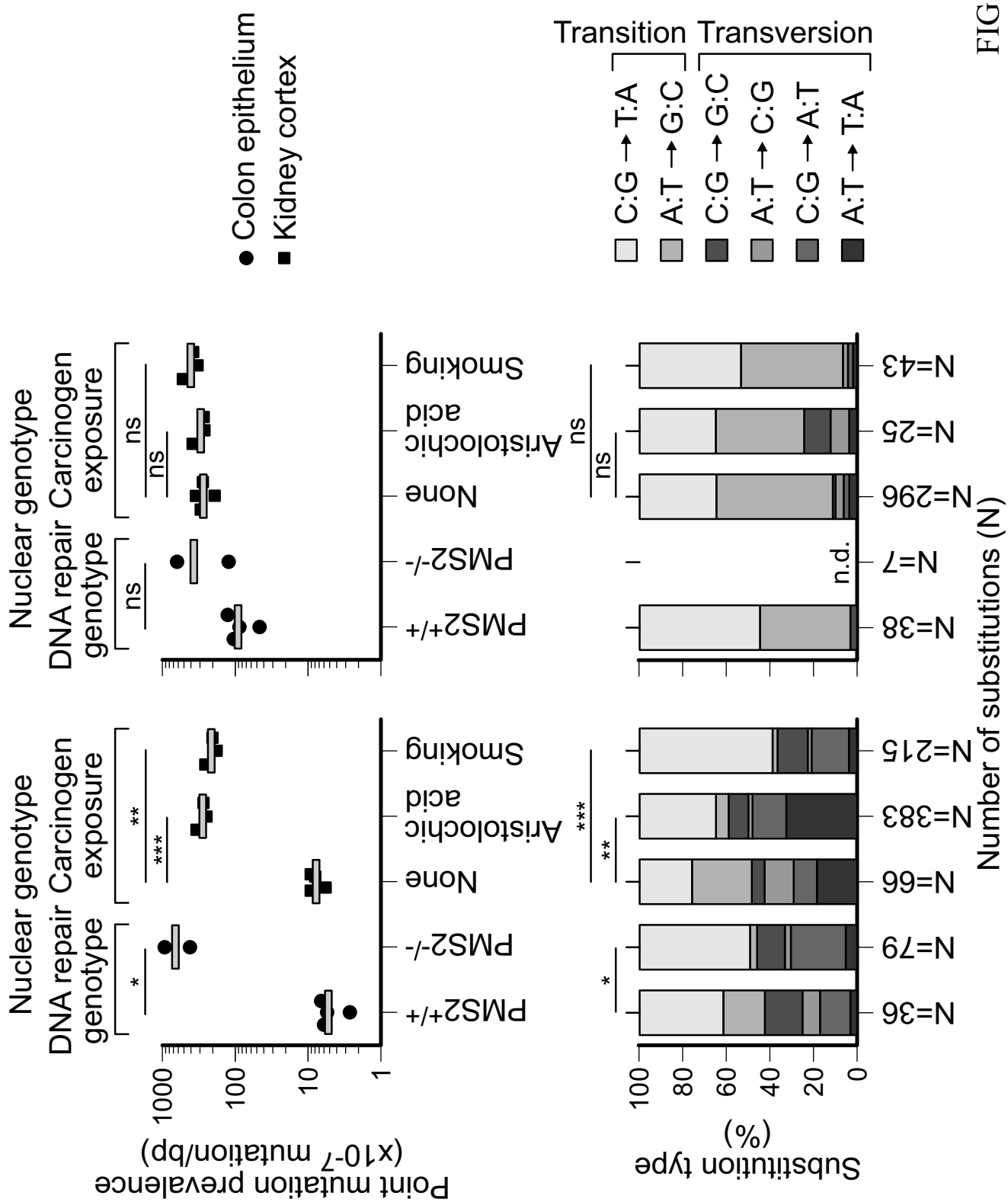

A
B
Point mutation prevalence
(x10-7 mutation/bp)
1000
100
10
1
Nuclear genotype
DNA repair genotype
Carcinogen exposure
*
***
**
ns
ns
ns
PMS2+/+
PMS2-/-
None
Aristolochic acid
Smoking
Colon epithelium
Kidney cortex
Substitution type
(%)
100
80
60
40
20
0
Transition
Transversion
C:G → T:A
A:T → G:C
C:G → G:C
A:T → C:G
C:G → A:T
A:T → T:A
n.d.
N=36
N=79
N=66
N=383
N=215
N=38
N=7
N=296
N=25
N=43
Number of substitutions (N)

FIG. 53

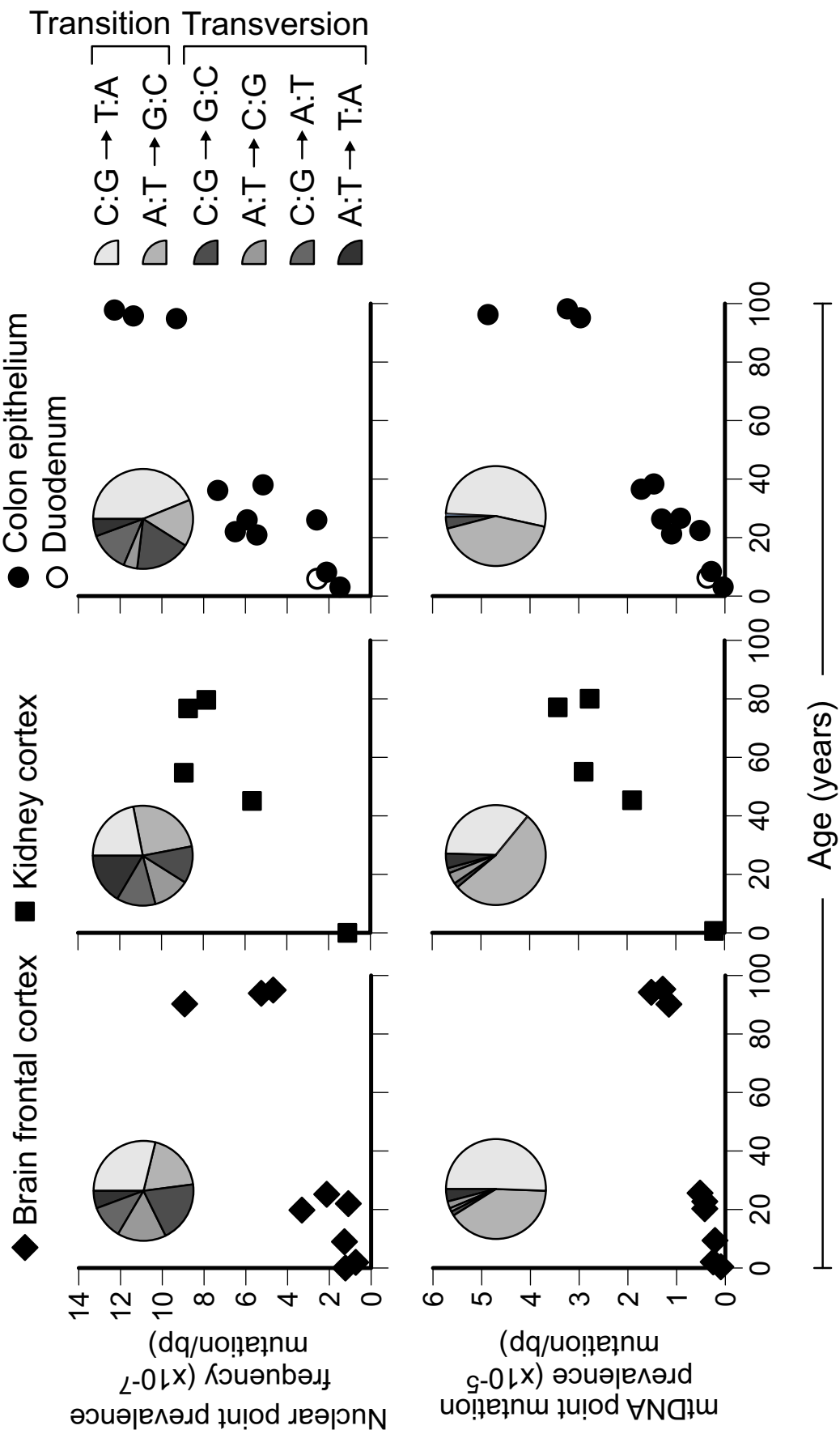

Transition
Transversion
C:G → T:A
A:T → G:C
C:G → G:C
A:T → C:G
C:G → A:T
A:T → T:A
Colon epithelium
Duodenum
Brain frontal cortex
Kidney cortex
Nuclear point prevalence frequency (x10-7 mutation/bp)
mtDNA point mutation prevalence (x10-5 mutation/bp)
Age (years)

FIG. 54

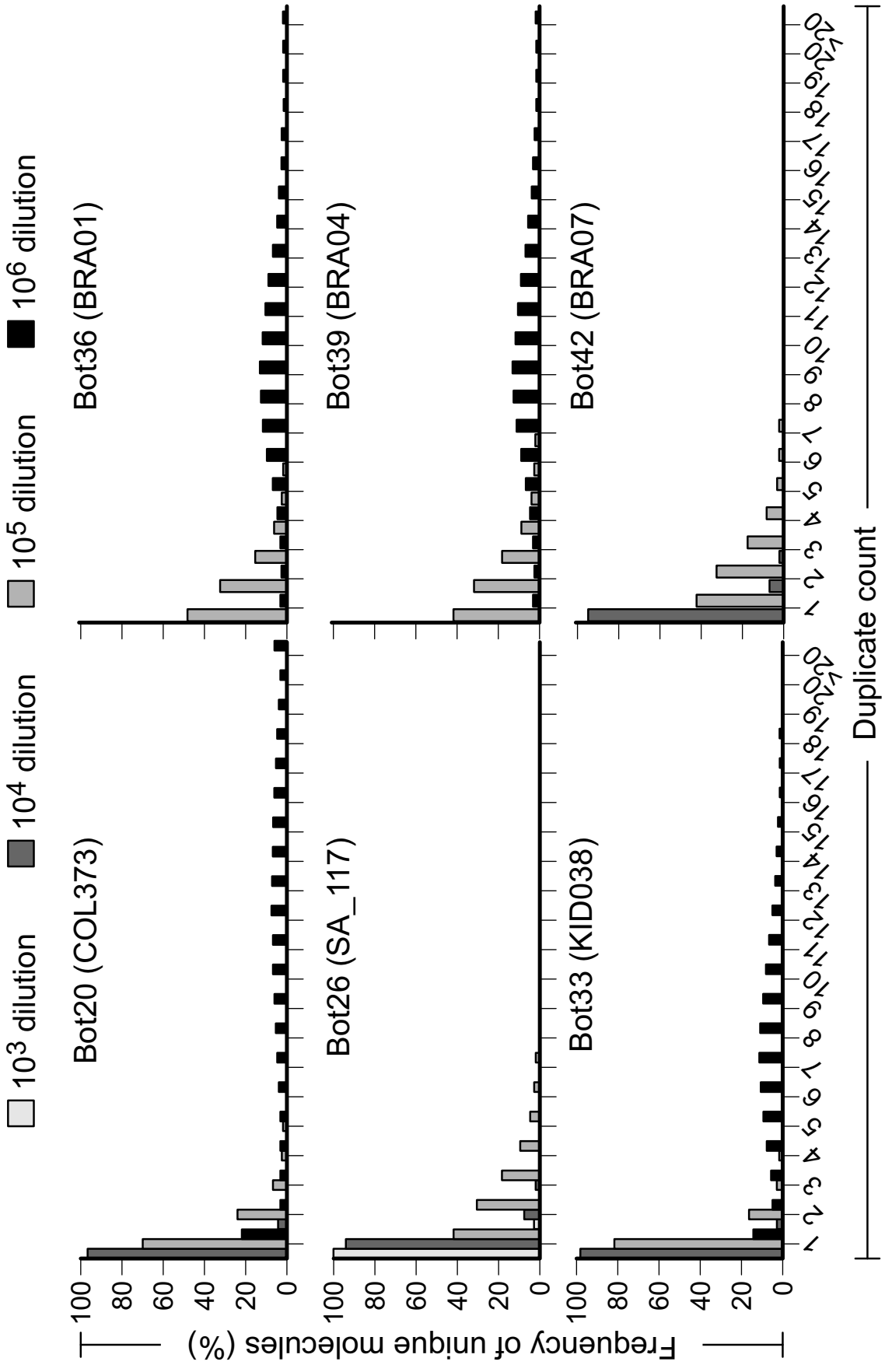

10^3 dilution
10^4 dilution
10^5 dilution
10^6 dilution
Bot20 (COL373)
Bot36 (BRA01)
Bot26 (SA_117)
Bot39 (BRA04)
Bot33 (KID038)
Bot42 (BRA07)
Frequency of unique molecules (%)
Duplicate count

FIG. 55

Bottleneck sequencing library
(sample name / library ID)
BRA09/Bo**
BRA08/Bo**
BRA07/Bo**
BRA06/Bo**
BRA05/Bo**
BRA04/Bo**
BRA03/Bo**
BRA02/Bo**
BRA01/Bo**
KID038/Bo**
KID038/Bo**
KID038/Bo**
KID037/Bo**
KID036/Bo**
KID035/Bo**
KID034/Bo**
SA_119/Bo**
SA_118/Bo**
SA_117/Bo**
AA_126/Bo**
AA_124/Bo**
AA_105/Bo**
COL375/Bo**
COL374/Bo**
COL373/Bo**
COL236/Bo**
COL235/Bo**
SIN230/Bo**
CCL239/Bo**
CCL238/Bo**
CCL237/Bo**
CCL236/Bo**
CCL235/Bo**
CCL234/Bo**
CCL233/Bo**
CCL232/Bo**
CCL231/Bo**
CCL229/Bo**
SA_119/Bo**
SA_118/Bo**
SA_117/Bo**
AA_126/Bo**
AA_124/Bo**
AA_105/Bo**
0
100
200
300
400
500
600
Duplicate count

FIG. 56

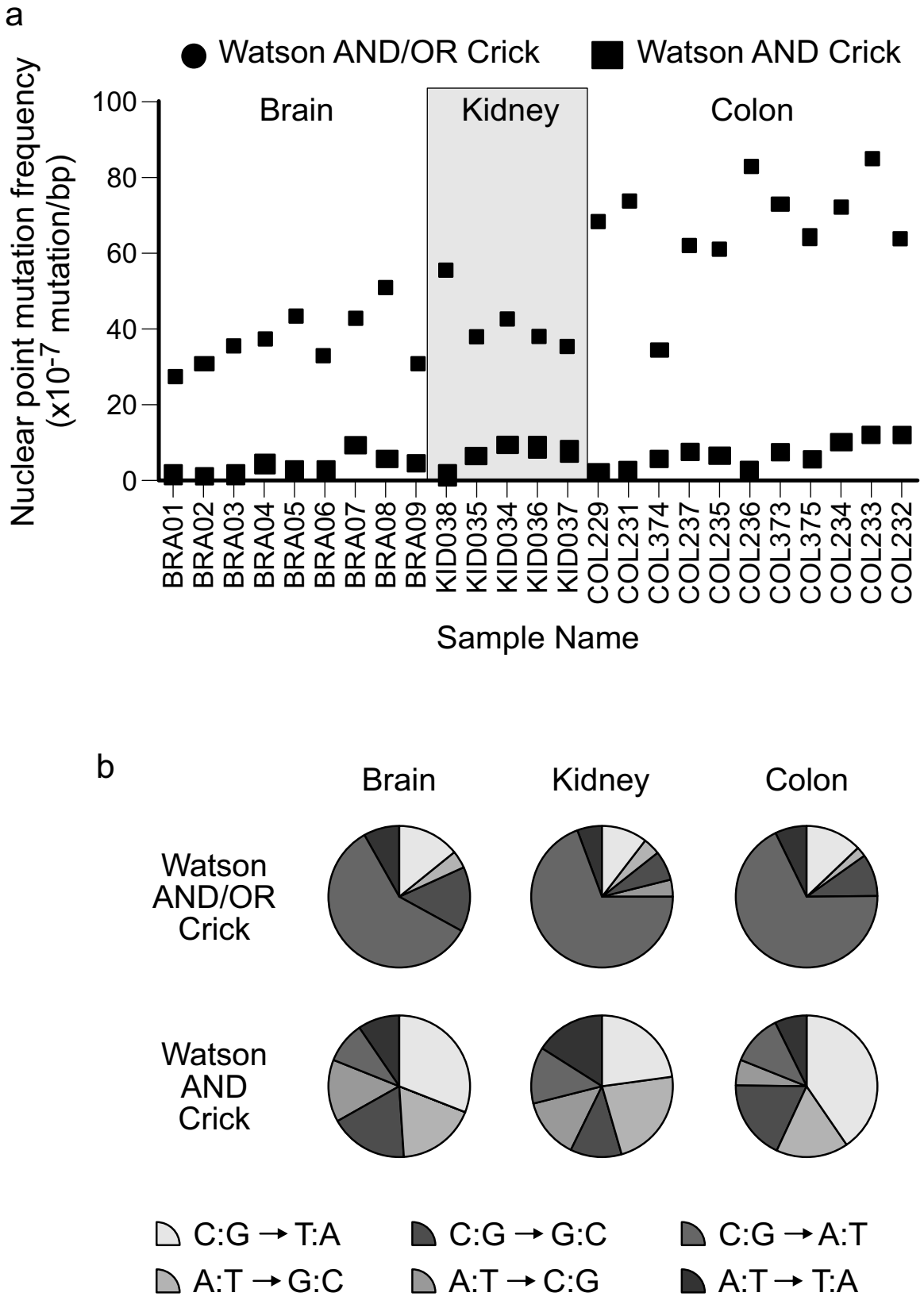

a
Watson AND/OR Crick
Watson AND Crick
Brain
Kidney
Colon
Nuclear point mutation frequency (x10-7 mutation/bp)
100
80
60
40
20
0
BRA01
BRA02
BRA03
BRA04
BRA05
BRA06
BRA07
BRA08
BRA09
KID038
KID035
KID034
KID036
KID037
COL229
COL231
COL374
COL237
COL235
COL236
COL373
COL375
COL234
COL233
COL232
Sample Name
b
Brain
Kidney
Colon
Watson AND/OR Crick
Watson AND Crick
C:G → T:A
C:G → G:C
C:G → A:T
A:T → G:C
A:T → C:G
A:T → T:A

FIG. 57

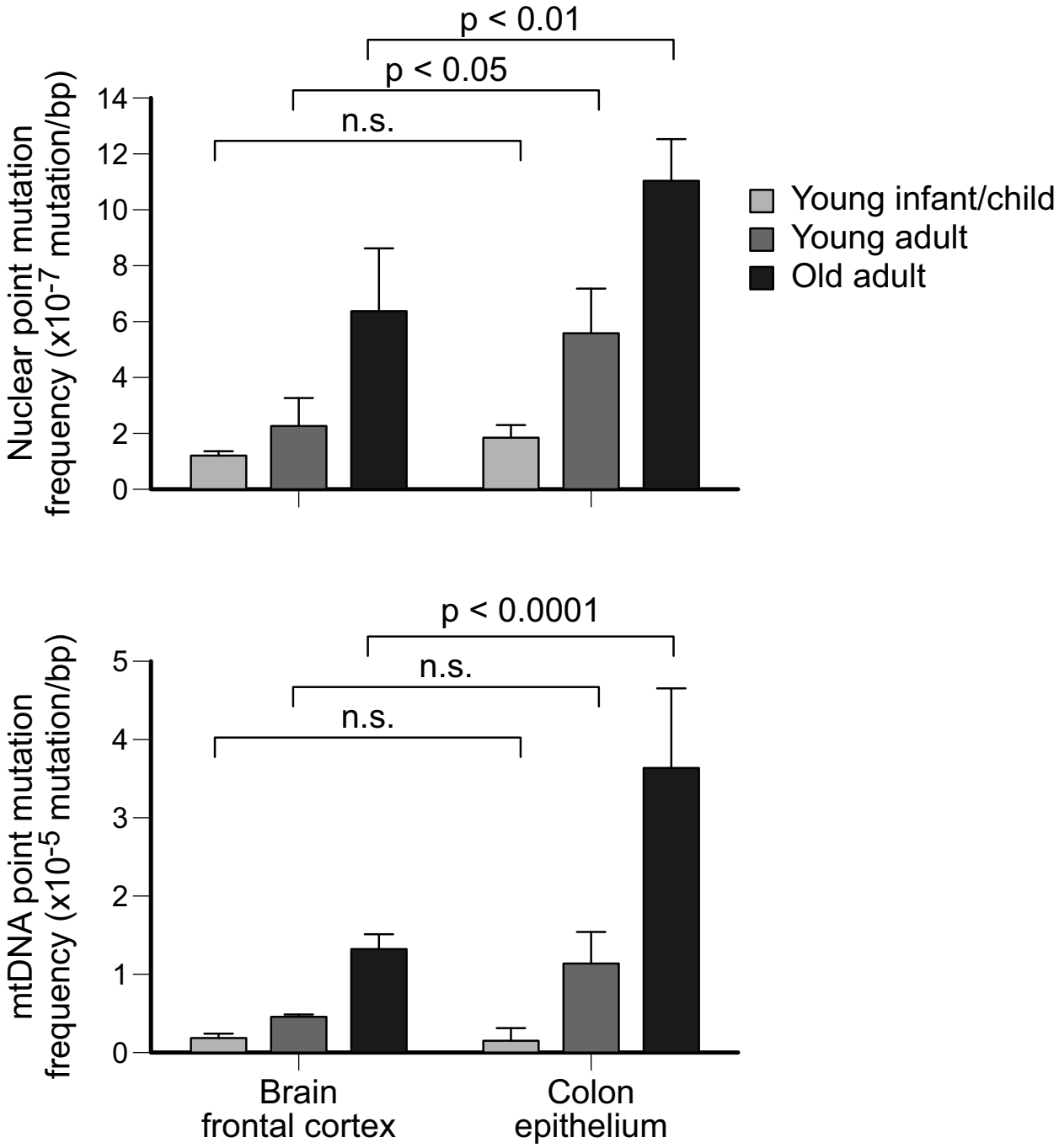

p < 0.01
p < 0.05
n.s.
Nuclear point mutation frequency (x10-7 mutation/bp)
14
12
10
8
6
4
2
0
Young infant/child
Young adult
Old adult
p < 0.0001
n.s.
n.s.
mtDNA point mutation frequency (x10-5 mutation/bp)
5
4
3
2
1
0
Brain frontal cortex
Colon epithelium

FIG. 58

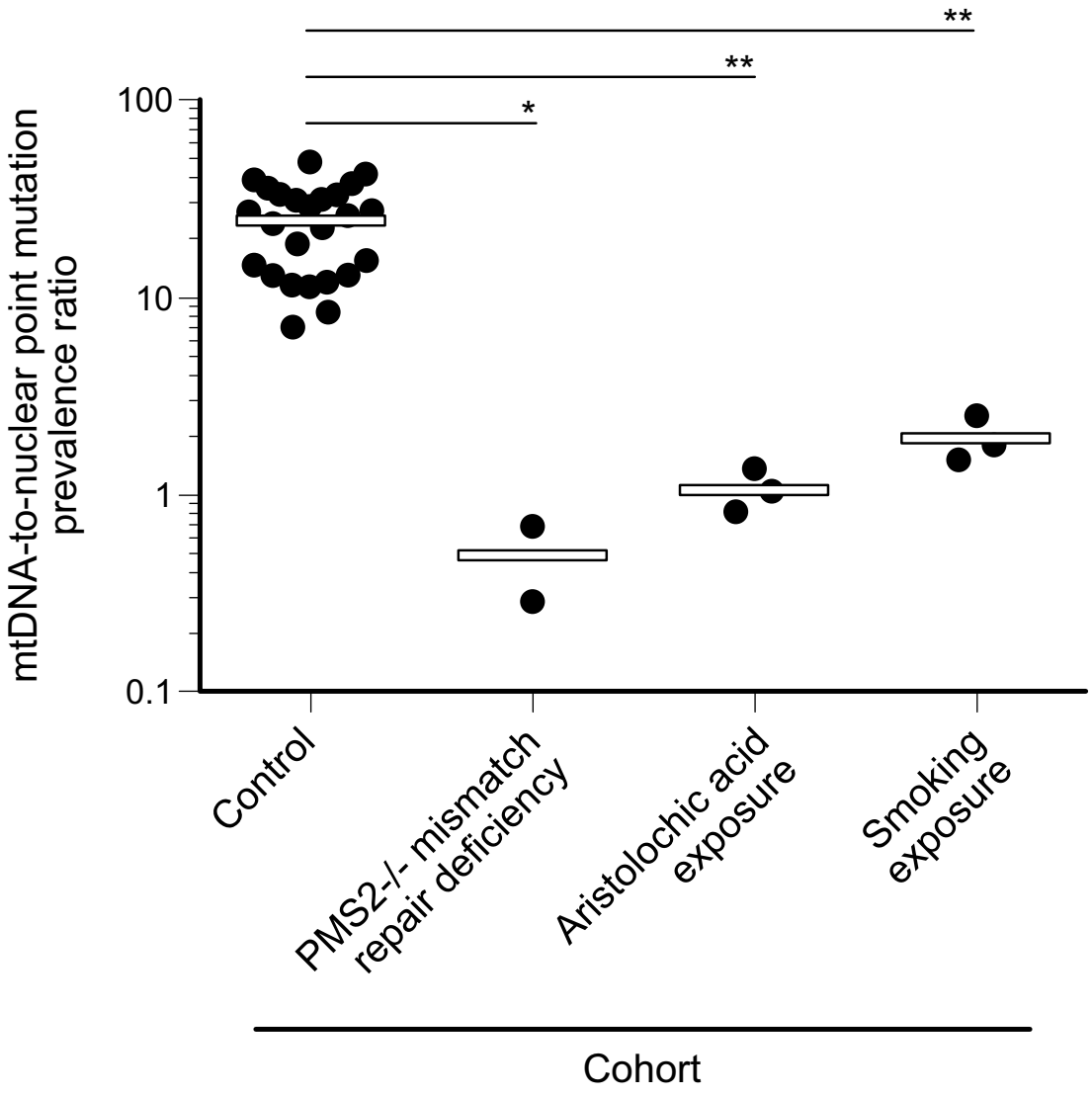

**
**
*
100
10
1
0.1
mtDNA-to-nuclear point mutation prevalence ratio
Control
PMS2-/- mismatch repair deficiency
Aristolochic acid exposure
Smoking exposure
Cohort

FIG. 59

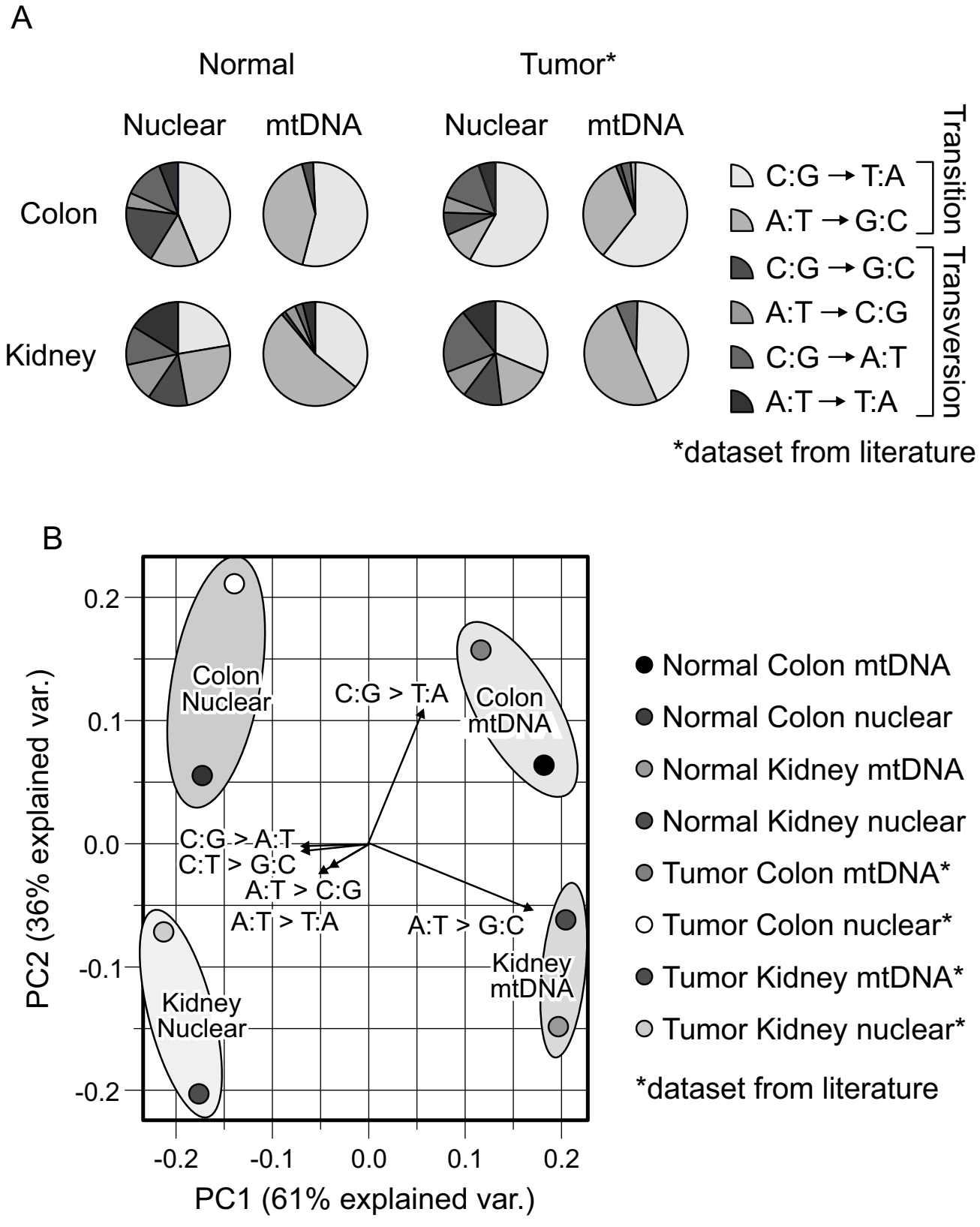

A
Normal
Tumor*
Nuclear
mtDNA
Nuclear
mtDNA
Colon
Kidney
C:G → T:A
A:T → G:C
C:G → G:C
A:T → C:G
C:G → A:T
A:T → T:A
Transition
Transversion
*dataset from literature
B
PC2 (36% explained var.)
PC1 (61% explained var.)
0.2
0.1
0.0
-0.1
-0.2
-0.2
-0.1
0.0
0.1
0.2
Colon Nuclear
Colon mtDNA
Kidney Nuclear
Kidney mtDNA
C:G > T:A
C:G > A:T
C:T > G:C
A:T > C:G
A:T > T:A
A:T > G:C
Normal Colon mtDNA
Normal Colon nuclear
Normal Kidney mtDNA
Normal Kidney nuclear
Tumor Colon mtDNA*
Tumor Colon nuclear*
Tumor Kidney mtDNA*
Tumor Kidney nuclear*
*dataset from literature

FIG. 60

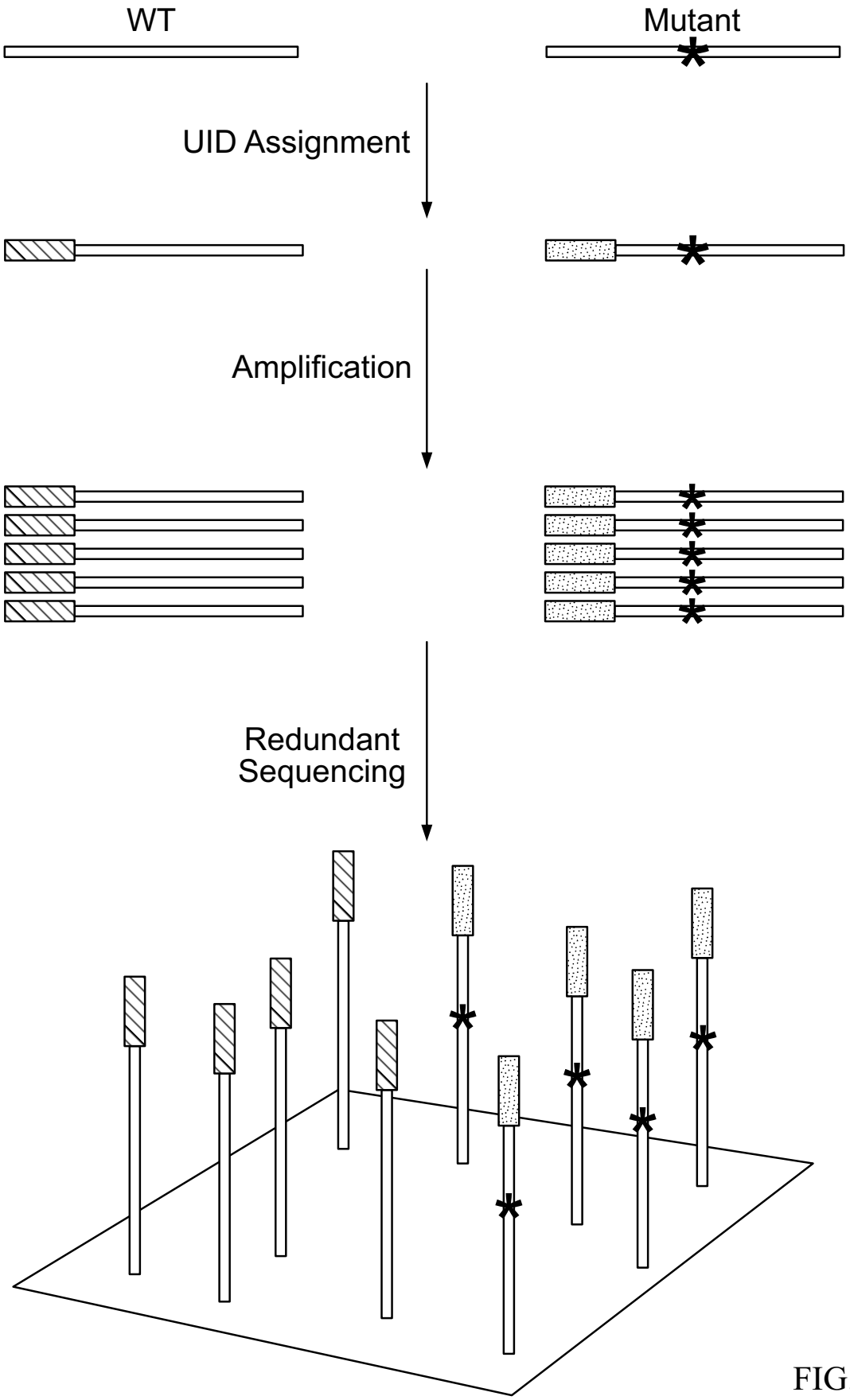

WT
Mutant
UID Assignment
Amplification
Redundant
Sequencing

FIG. 61

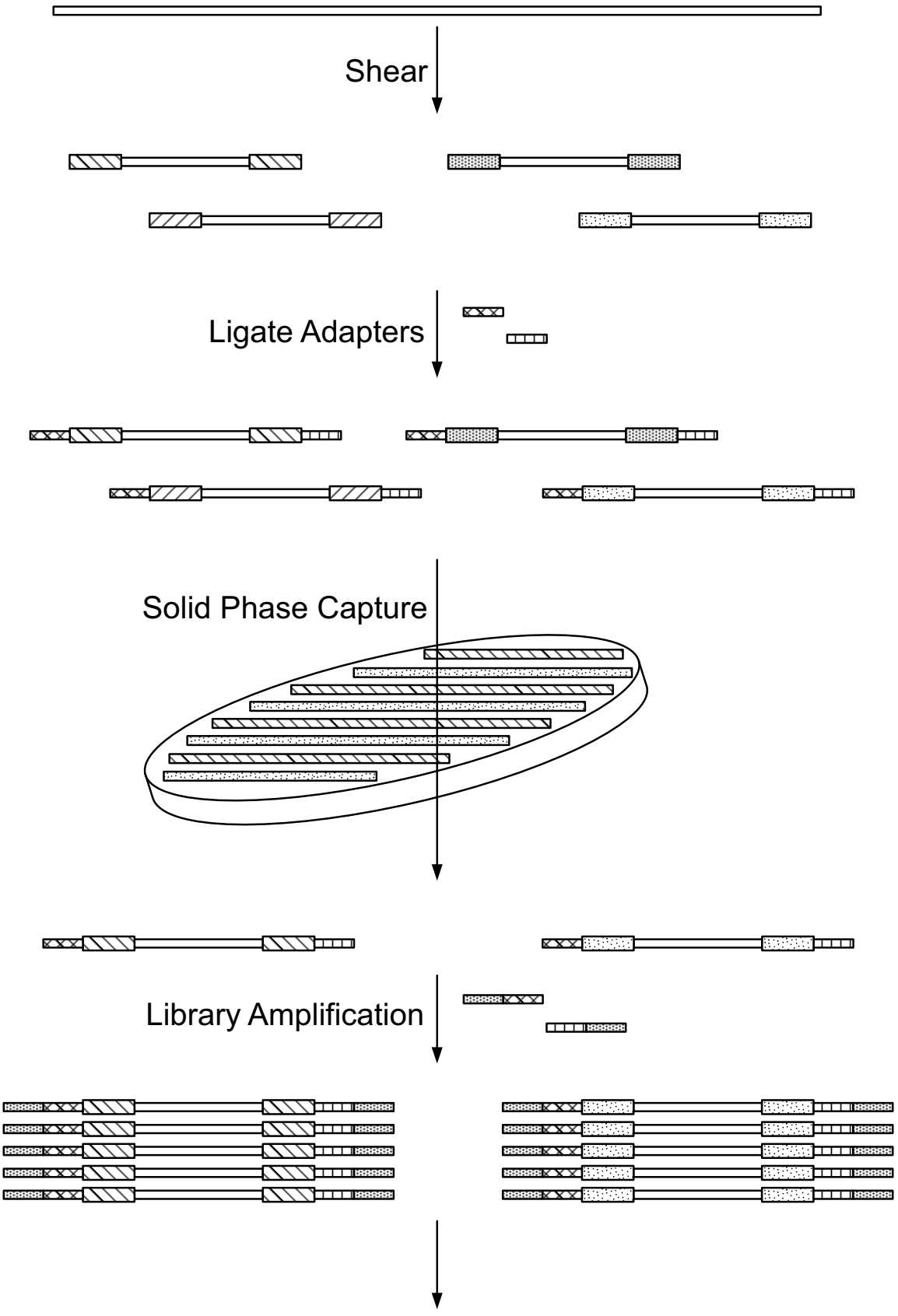

Shear
Ligate Adapters
Solid Phase Capture
Library Amplification
Redundant Sequencing

FIG. 62

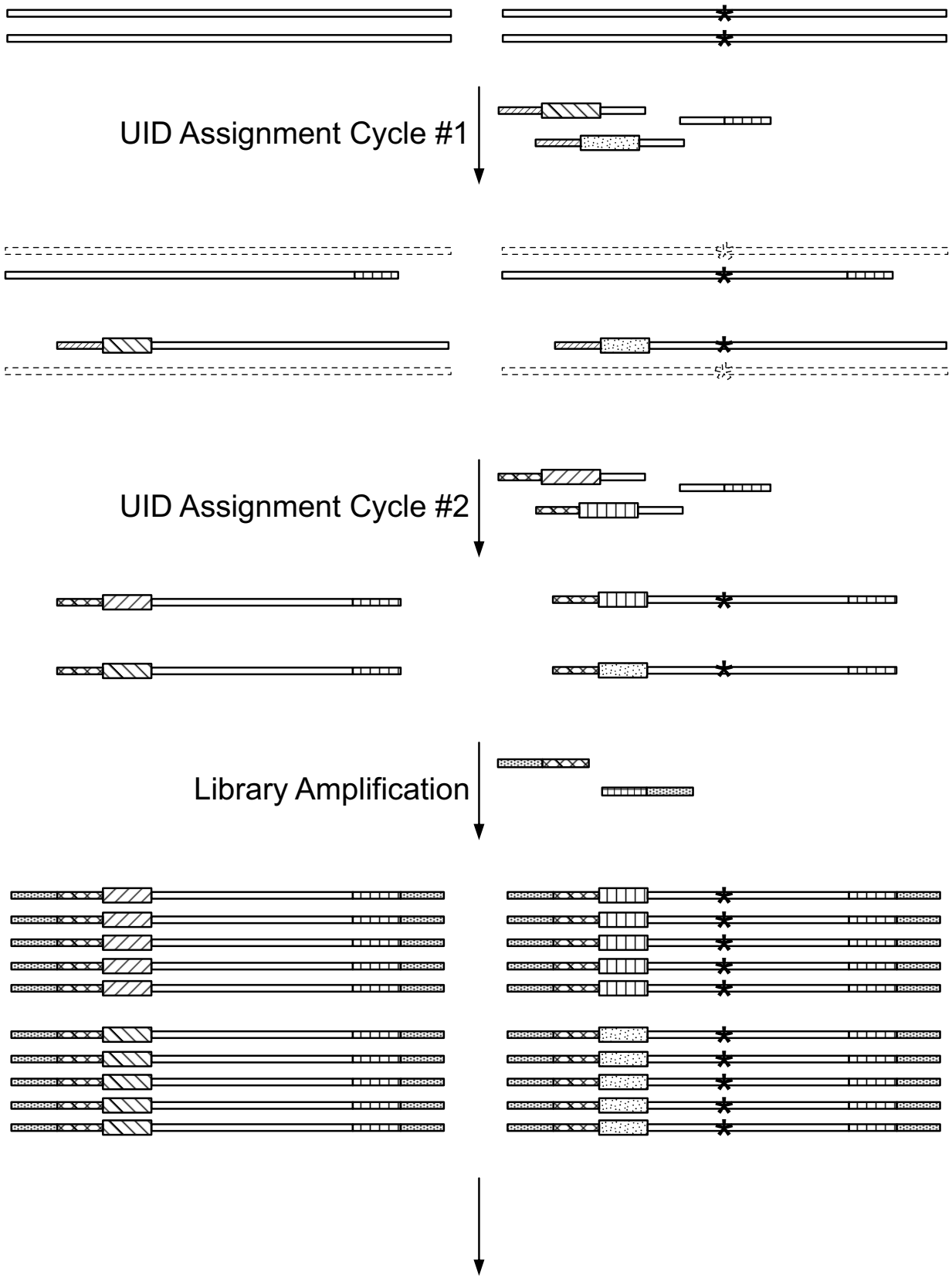

UID Assignment Cycle #1
UID Assignment Cycle #2
Library Amplification
Redundant Sequencing

FIG. 63

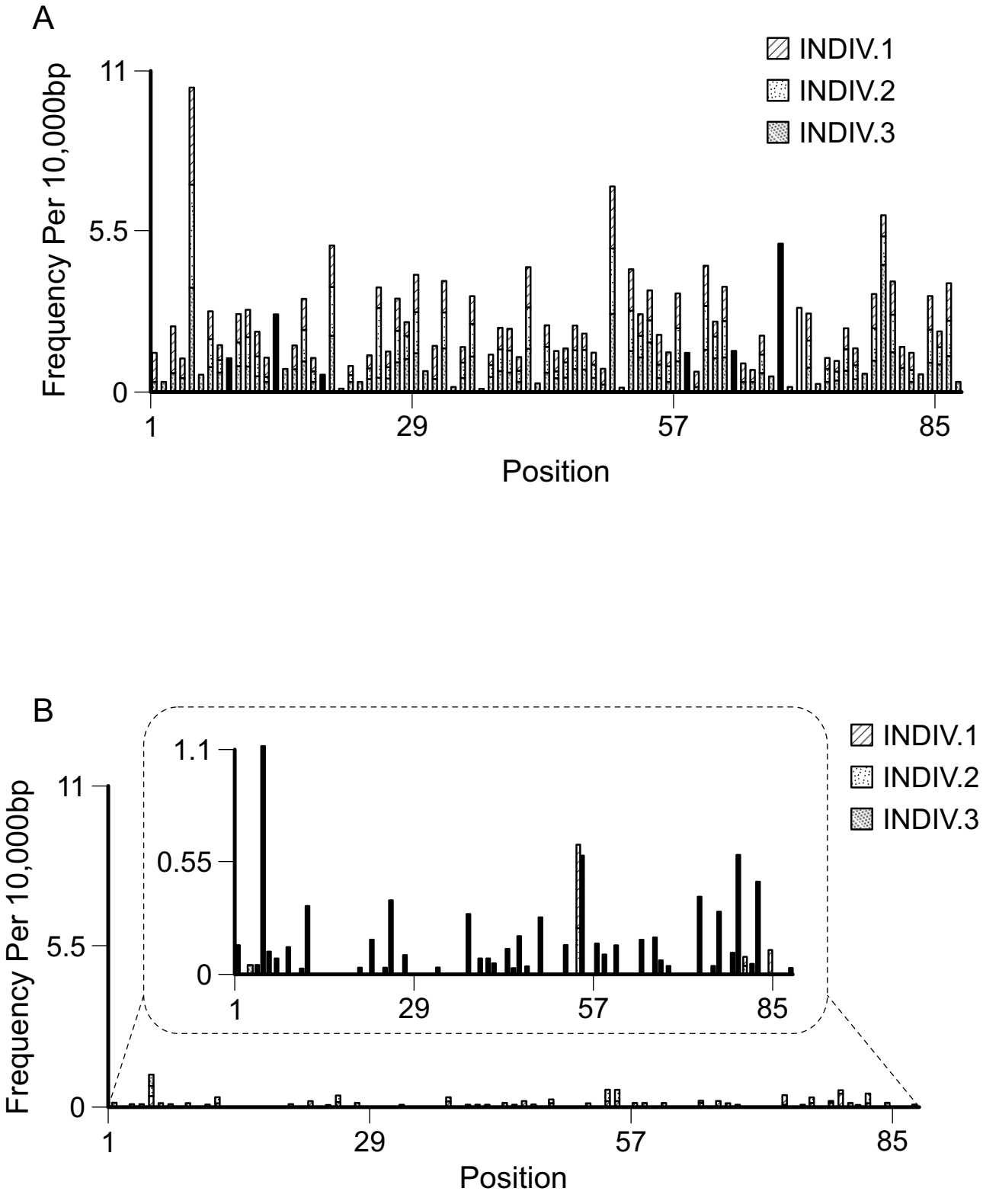

A
INDIV.1
INDIV.2
INDIV.3
Frequency Per 10,000bp
11
5.5
0
1
29
57
85
Position
B
INDIV.1
INDIV.2
INDIV.3
Frequency Per 10,000bp
11
5.5
0
1.1
0.55
0
1
29
57
85
1
29
57
85
Position

FIG. 64

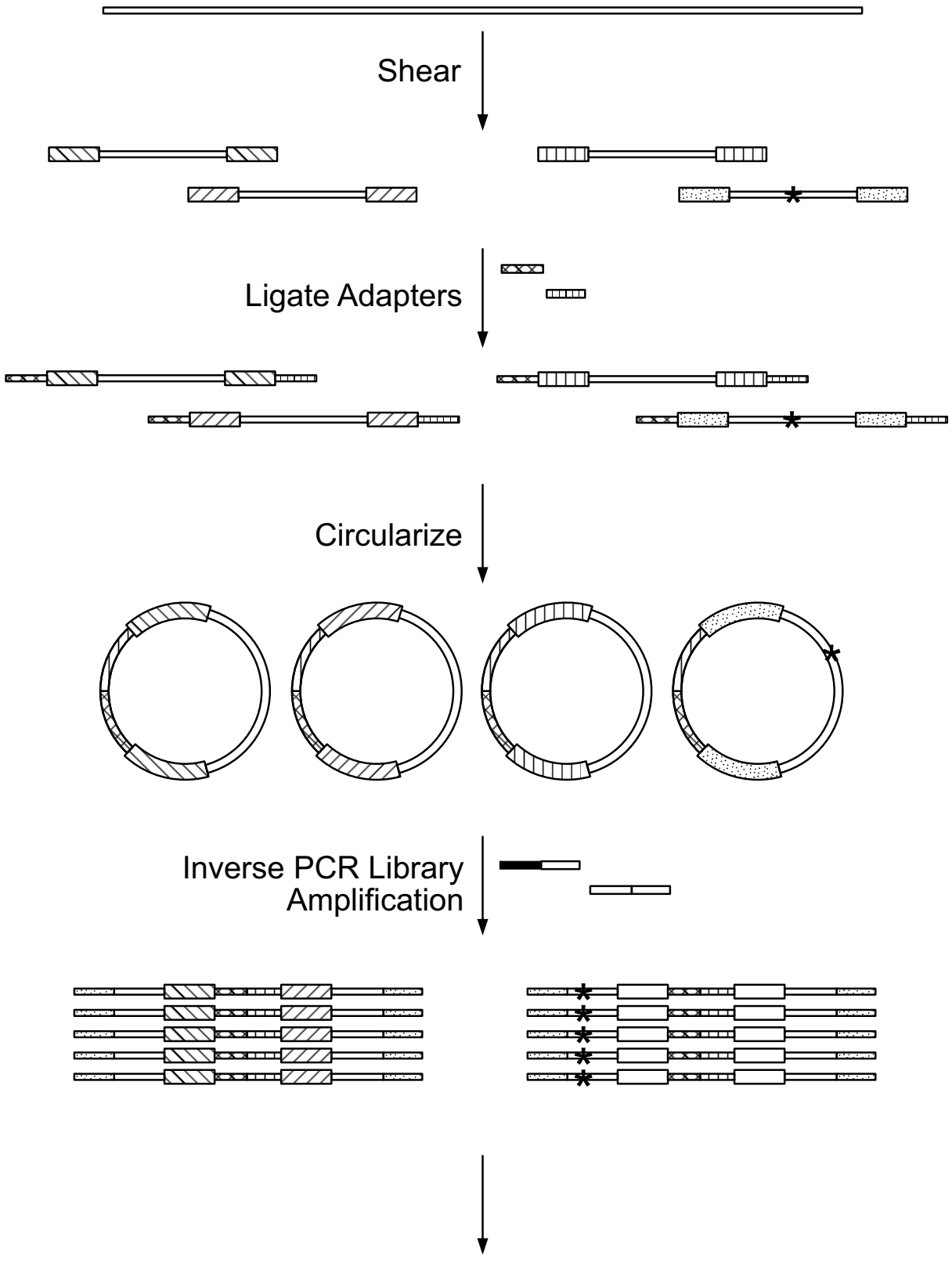

Shear
Ligate Adapters
Circularize
Inverse PCR Library
Amplification
Redundant Sequencing

FIG. 65

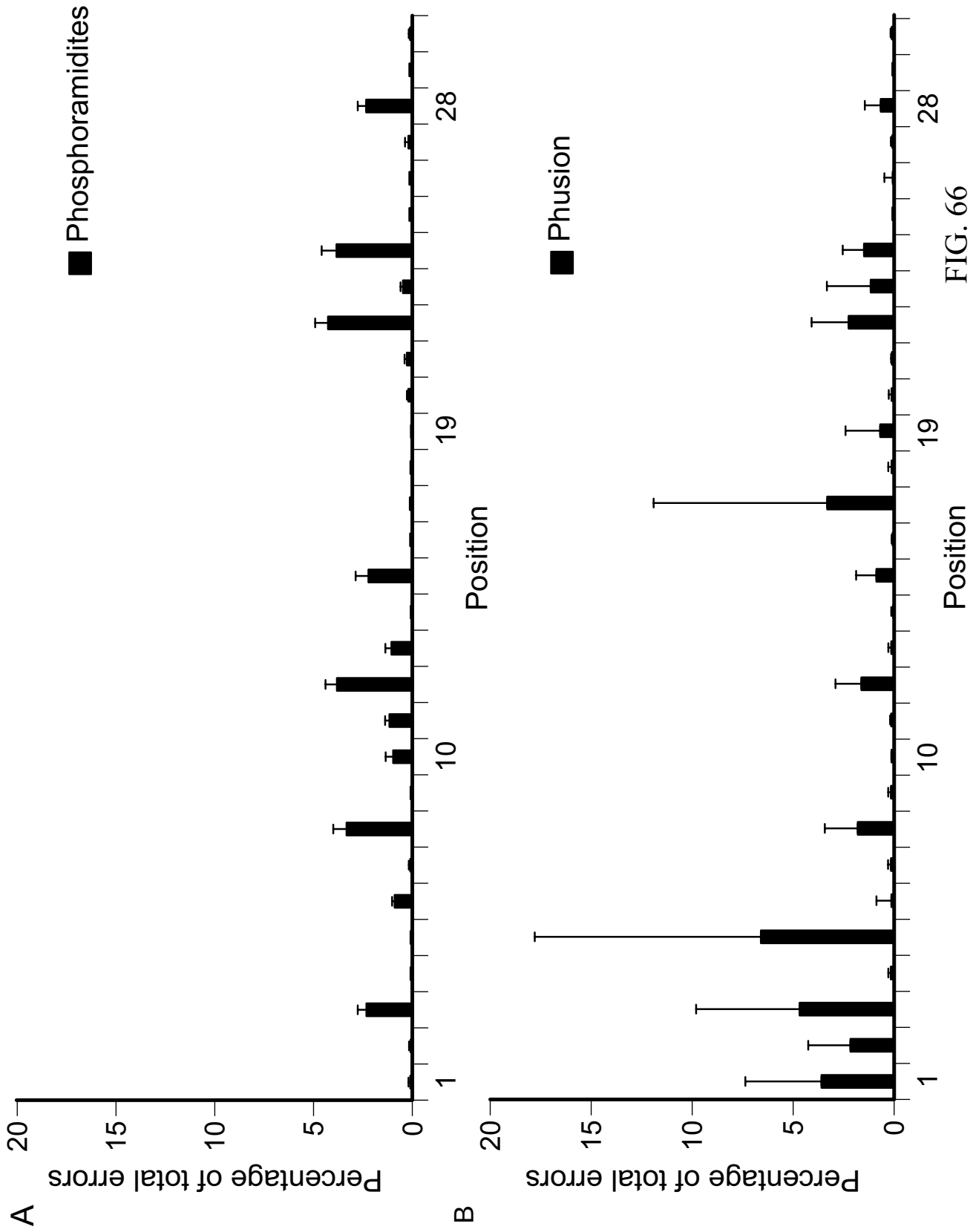

A
Phosphoramidites
Percentage of total errors
0
5
10
15
20
1
10
19
28
Position
B
Phusion
Percentage of total errors
0
5
10
15
20
1
10
19
28
Position

FIG. 66

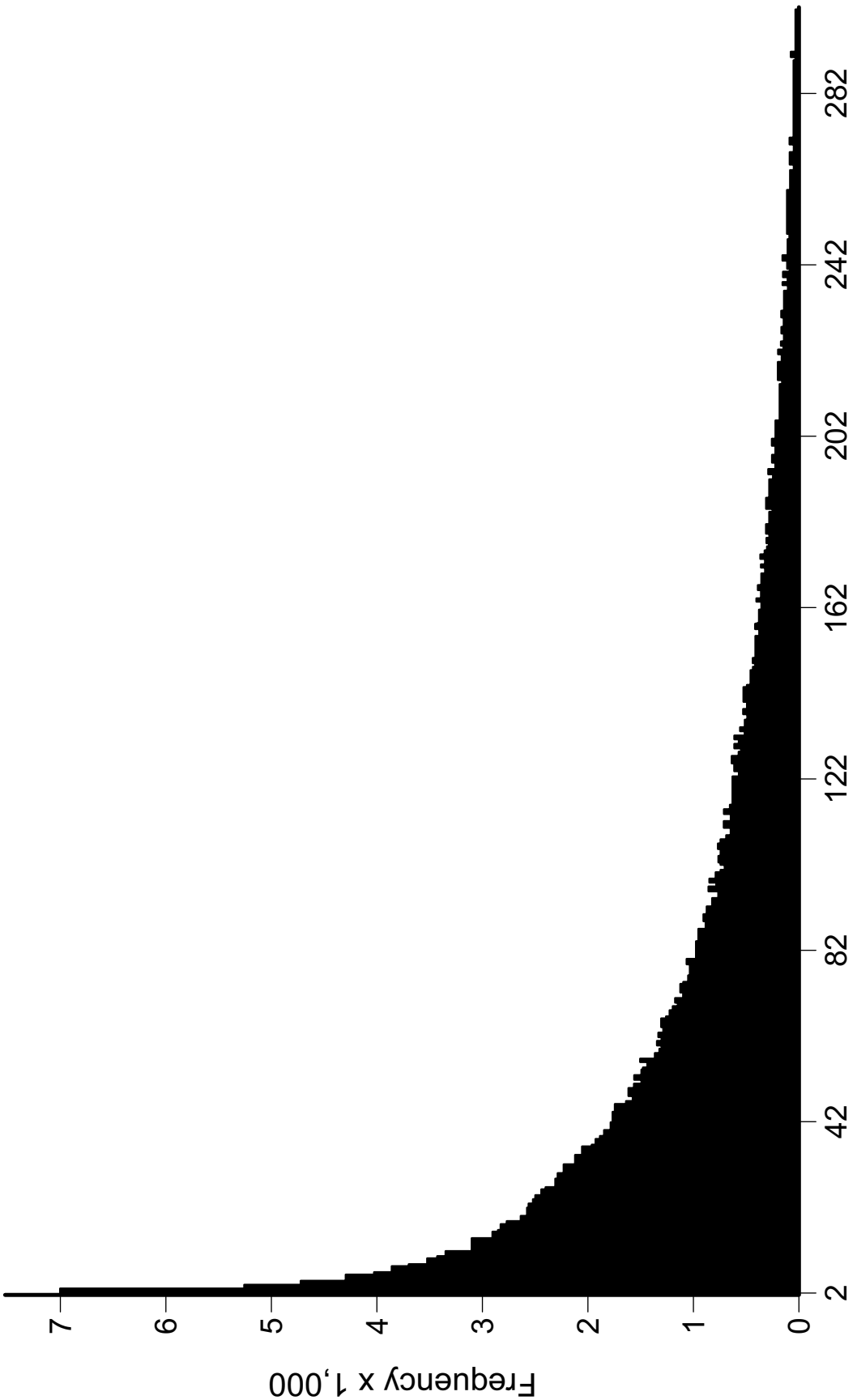

Frequency x 1,000
0
1
2
3
4
5
6
7
2
42
82
122
162
202
242
282
UID-Family Members

FIG. 67

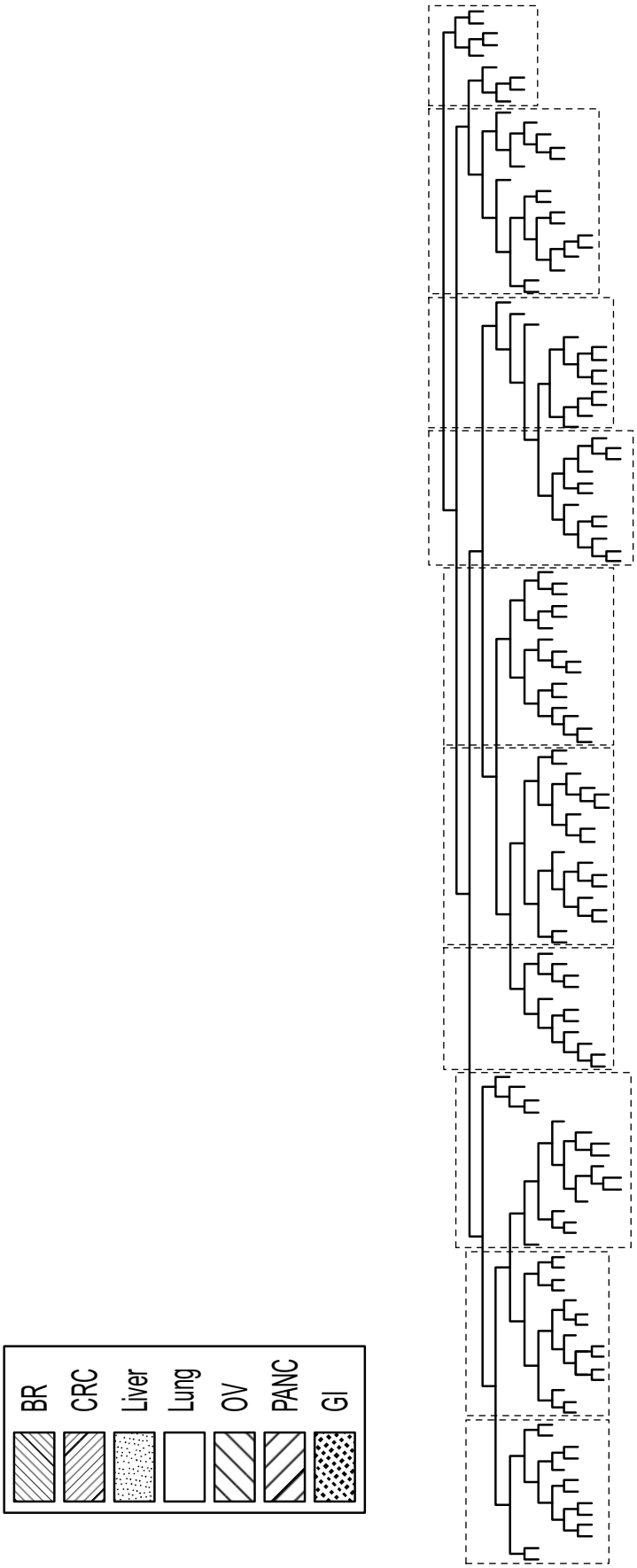


FIG. 68A

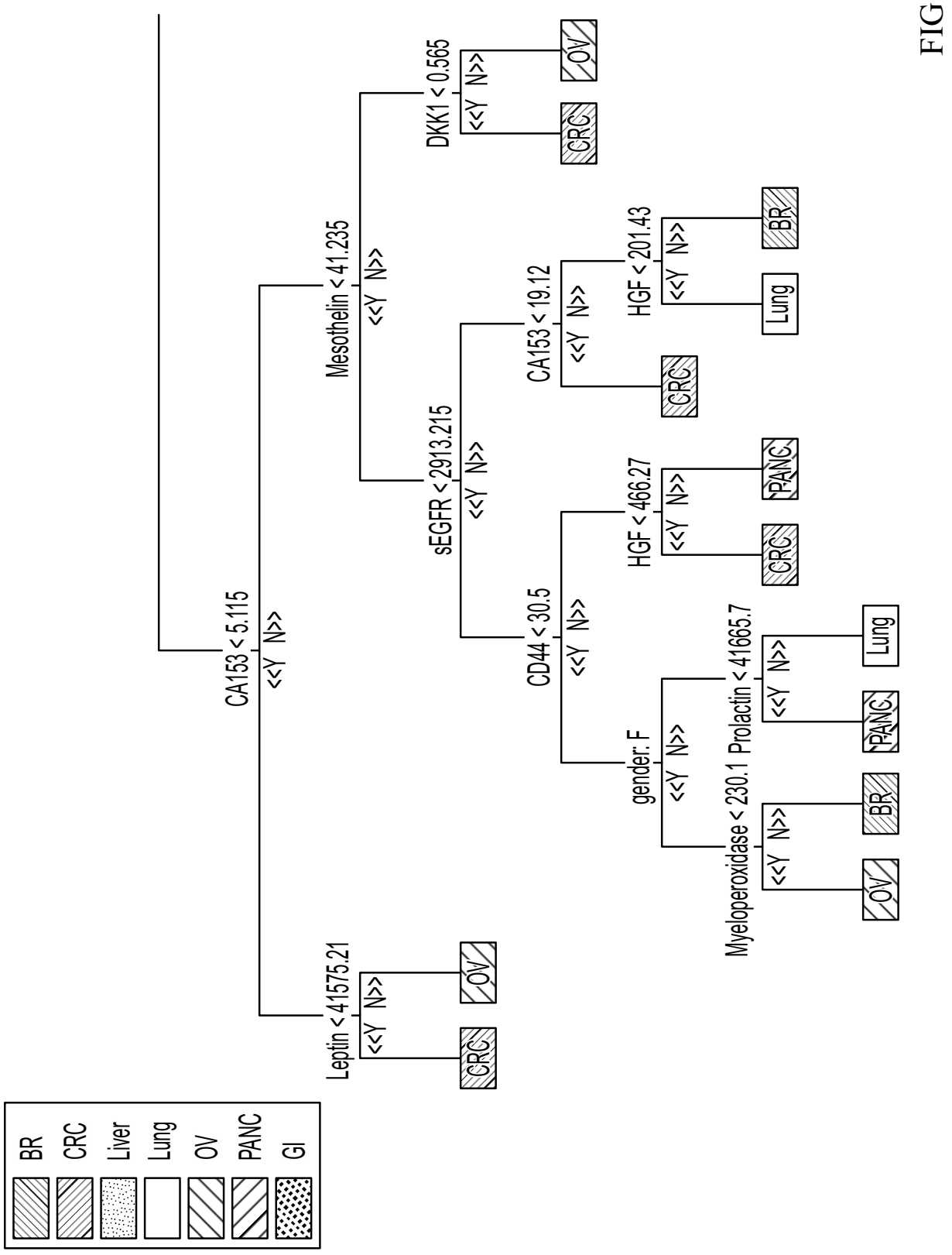

BR
CRC
Liver
Lung
OV
PANC
GI
CA153 < 5.115
<<Y N>>
Leptin < 41575.21
<<Y N>>
CRC
OV
Mesothelin < 41.235
<<Y N>>
DKK1 < 0.565
<<Y N>>
CRC
OV
sEGFR < 2913.215
<<Y N>>
CD44 < 30.5
<<Y N>>
gender: F
<<Y N>>
Myeloperoxidase < 230.1
<<Y N>>
OV
BR
Prolactin < 41665.7
<<Y N>>
PANC
Lung
HGF < 466.27
<<Y N>>
CRC
PANC
CA153 < 19.12
<<Y N>>
CRC
HGF < 201.43
<<Y N>>
Lung
BR

FIG. 68B

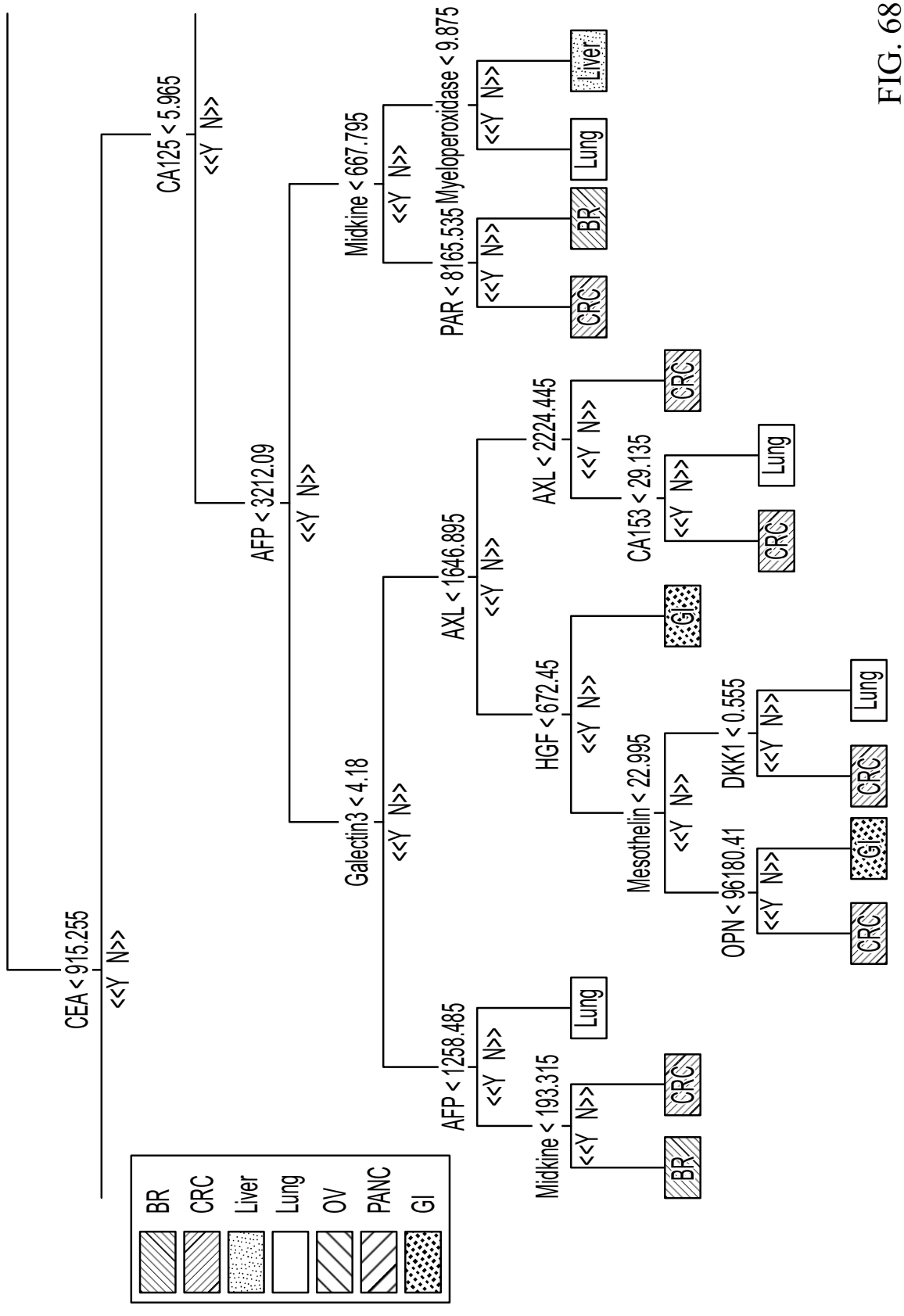

BR
CRC
Liver
Lung
OV
PANC
GI
CEA < 915.255
<<Y N>>
CA125 < 5.965
<<Y N>>
AFP < 3212.09
<<Y N>>
Galectin3 < 4.18
<<Y N>>
AFP < 1258.485
<<Y N>>
Lung
Midkine < 193.315
<<Y N>>
CRC
BR
AXL < 1646.895
<<Y N>>
HGF < 672.45
<<Y N>>
Mesothelin < 22.995
<<Y N>>
GI
OPN < 96180.41
<<Y N>>
CRC
GI
DKK1 < 0.555
<<Y N>>
CRC
Lung
AXL < 2224.445
<<Y N>>
CA153 < 29.135
<<Y N>>
CRC
Lung
CRC
Midkine < 667.795
<<Y N>>
PAR < 8165.535
<<Y N>>
CRC
Myeloperoxidase < 9.875
<<Y N>>
BR
Lung
Liver

FIG. 68C

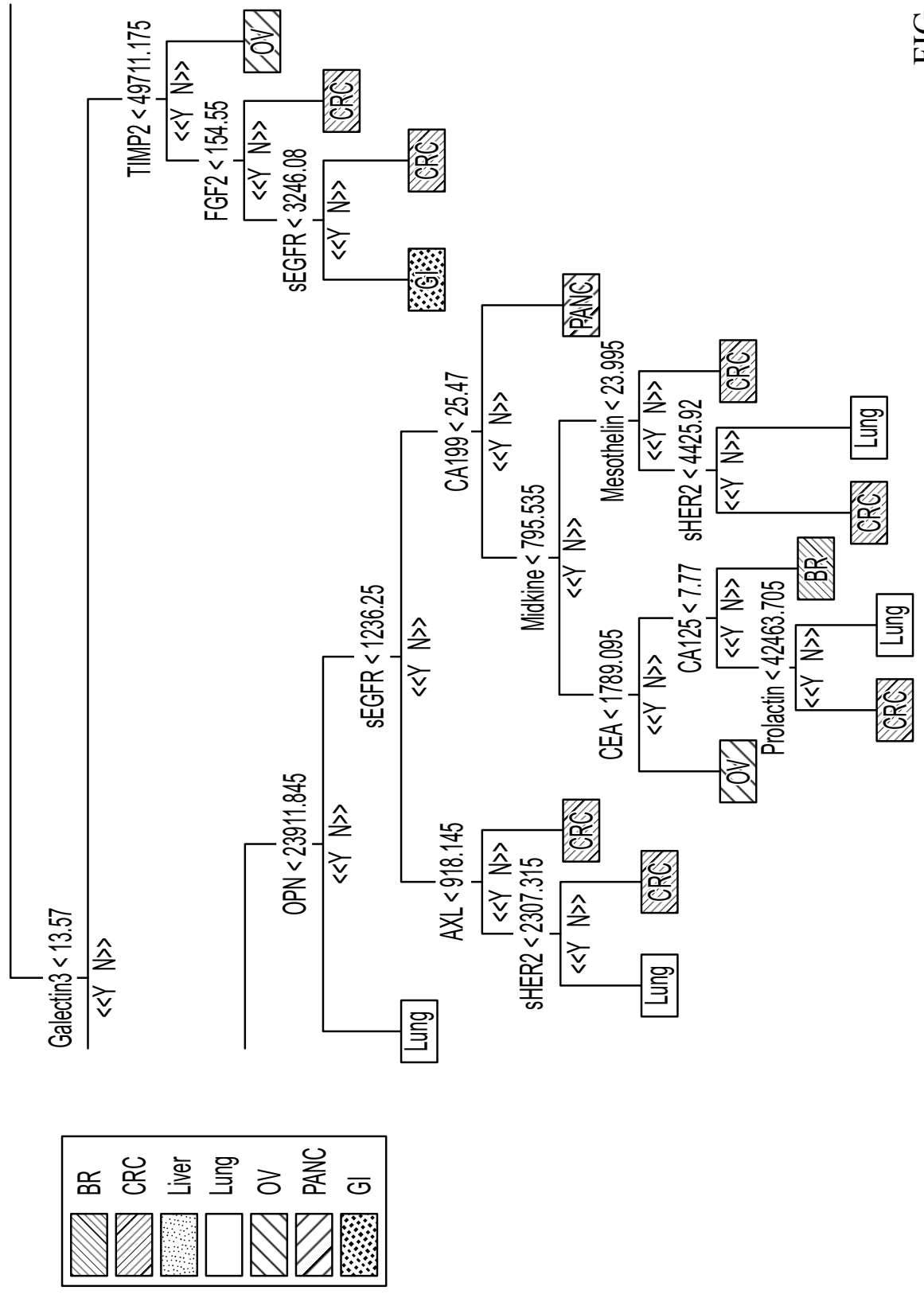

Galectin3 < 13.57
<<Y N>>
Lung
TIMP2 < 49711.175
<<Y N>>
OPN < 23911.845
<<Y N>>
FGF2 < 154.55
<<Y N>>
OV
sEGFR < 1236.25
<<Y N>>
sEGFR < 3246.08
<<Y N>>
CRC
AXL < 918.145
<<Y N>>
GI
CRC
CA199 < 25.47
<<Y N>>
sHER2 < 2307.315
<<Y N>>
PANC
Midkine < 795.535
<<Y N>>
CRC
Lung
CEA < 1789.095
<<Y N>>
Mesothelin < 23.995
<<Y N>>
CRC
CA125 < 7.77
<<Y N>>
sHER2 < 4425.92
<<Y N>>
OV
Prolactin < 42463.705
<<Y N>>
BR
CRC
Lung
CRC
Lung
BR
CRC
Liver
Lung
OV
PANC
GI

FIG. 68D

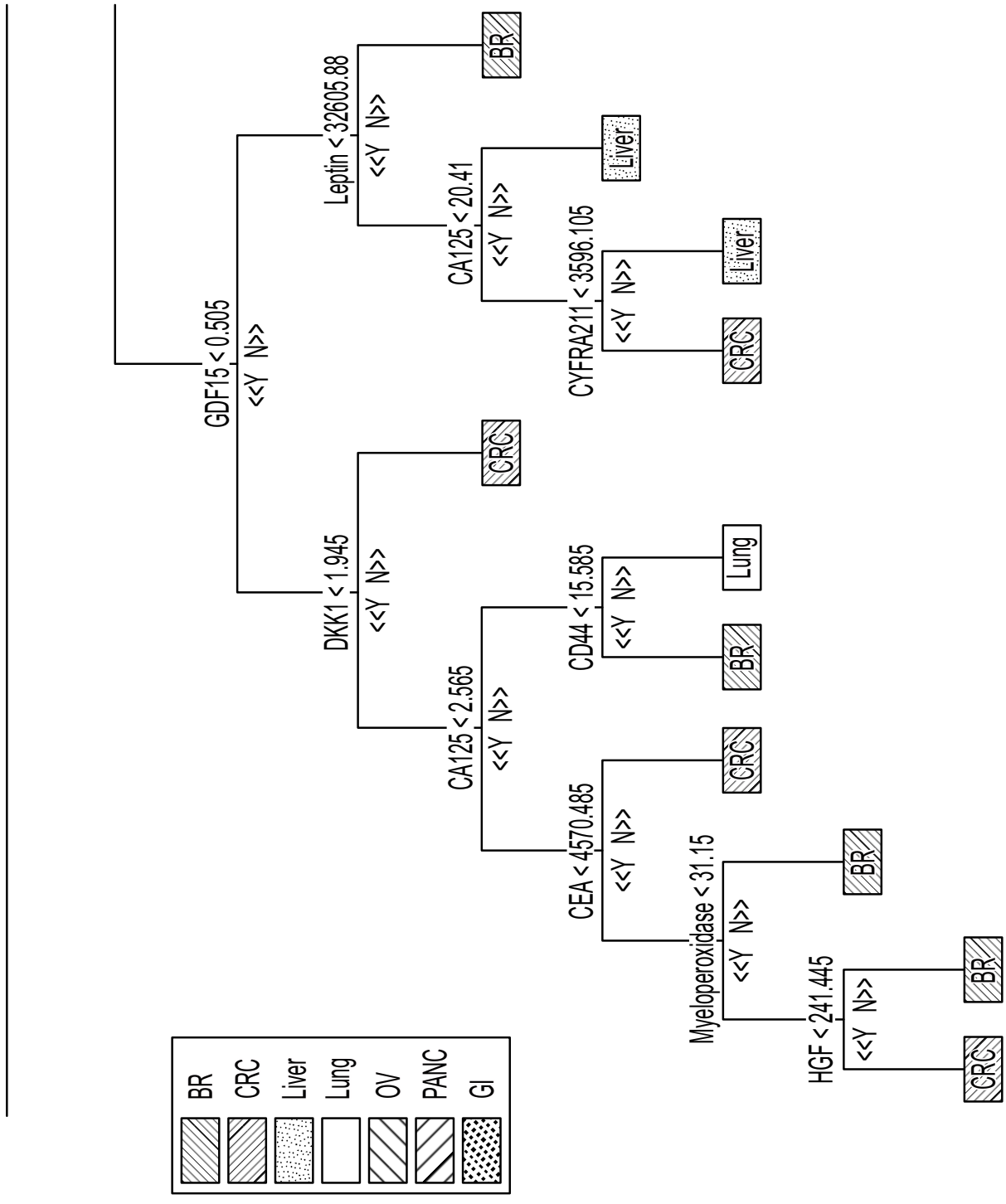

BR
CRC
Liver
Lung
OV
PANC
GI
GDF15 < 0.505
<<Y N>>
Leptin < 32605.88
<<Y N>>
BR
CA125 < 20.41
<<Y N>>
Liver
CYFRA211 < 3596.105
<<Y N>>
Liver
CRC
DKK1 < 1.945
<<Y N>>
CRC
CA125 < 2.565
<<Y N>>
CD44 < 15.585
<<Y N>>
Lung
BR
CEA < 4570.485
<<Y N>>
CRC
Myeloperoxidase < 31.15
<<Y N>>
BR
HGF < 241.445
<<Y N>>
BR
CRC

FIG. 68E

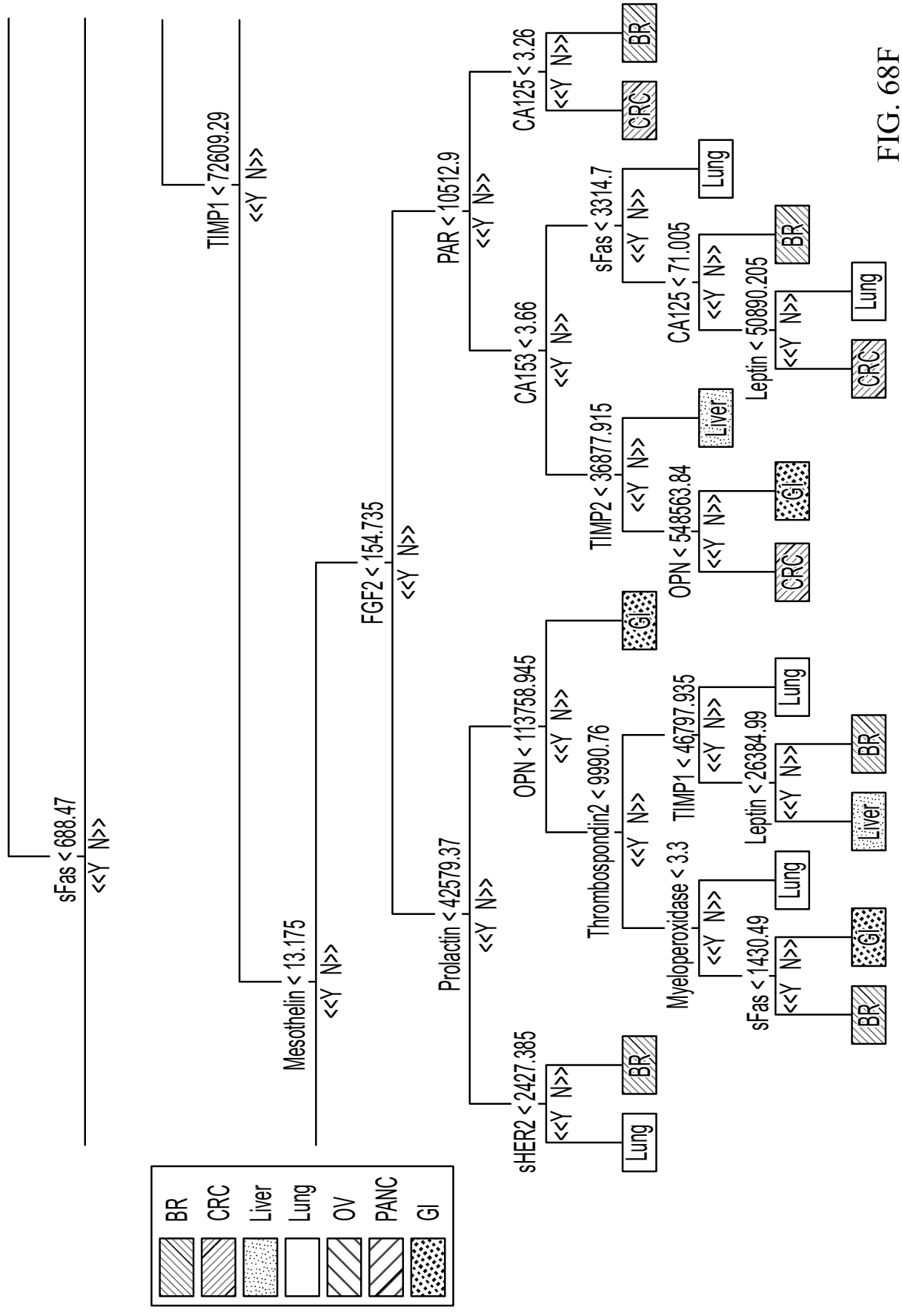

sFas < 688.47
<<Y N>>
TIMP1 < 72609.29
<<Y N>>
Mesothelin < 13.175
<<Y N>>
FGF2 < 154.735
<<Y N>>
Prolactin < 42579.37
<<Y N>>
sHER2 < 2427.385
<<Y N>>
Lung
BR
OPN < 113758.945
<<Y N>>
Thrombospondin2 < 9990.76
<<Y N>>
Myeloperoxidase < 3.3
<<Y N>>
sFas < 1430.49
<<Y N>>
BR
GI
Lung
Leptin < 26384.99
<<Y N>>
Liver
TIMP1 < 46797.935
<<Y N>>
BR
Lung
GI
PAR < 10512.9
<<Y N>>
CA153 < 3.66
<<Y N>>
TIMP2 < 36877.915
<<Y N>>
OPN < 548563.84
<<Y N>>
CRC
GI
Liver
CA125 < 3.26
<<Y N>>
CRC
BR
sFas < 3314.7
<<Y N>>
CA125 < 71.005
<<Y N>>
Leptin < 50890.205
<<Y N>>
CRC
Lung
BR
Lung
BR
CRC
Liver
GI
Lung
OV
PANC
GI

FIG. 68F

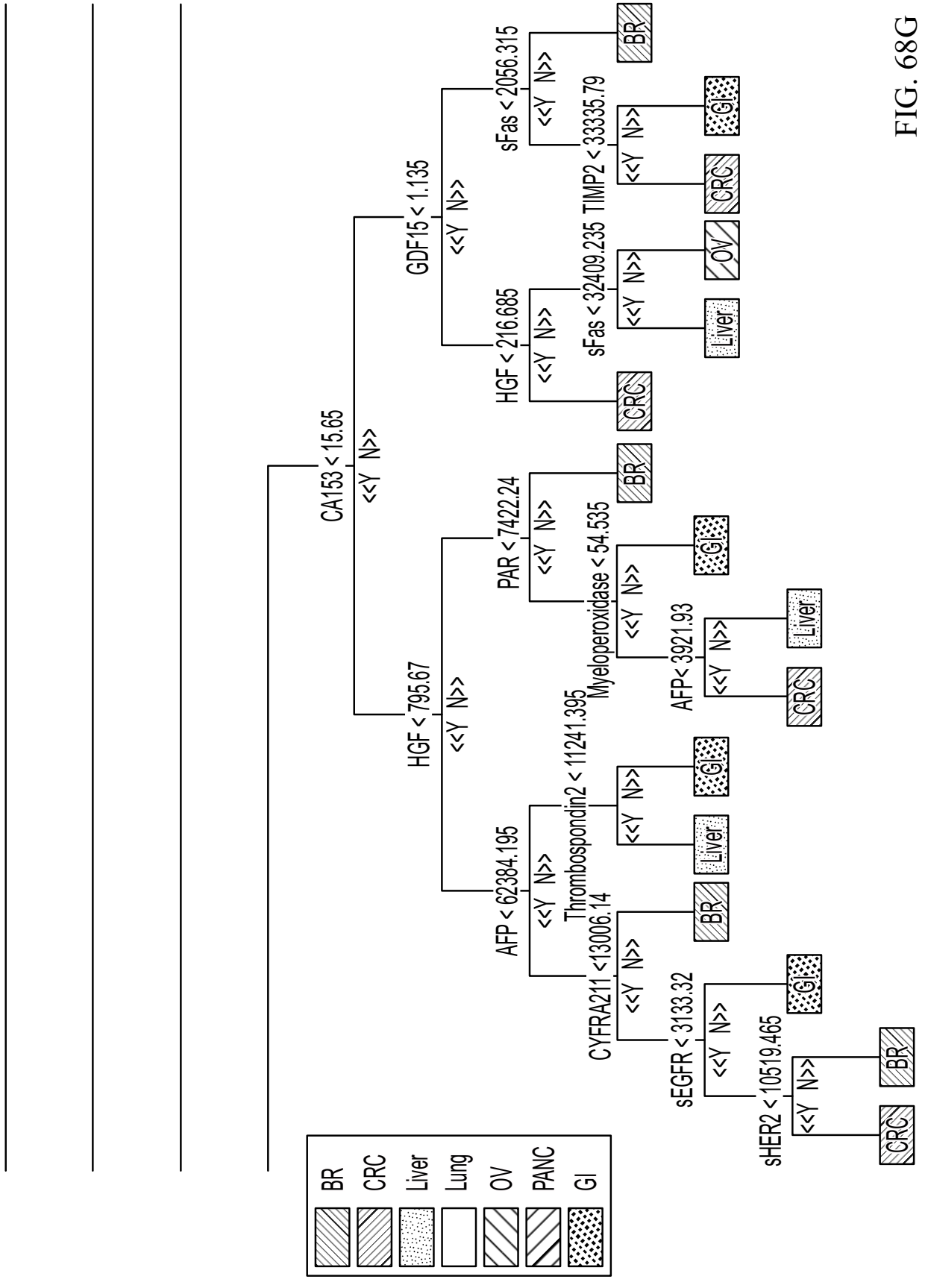

BR
CRC
Liver
Lung
OV
PANC
GI
CA153 < 15.65
<<Y N>>
HGF < 795.67
<<Y N>>
AFP < 62384.195
<<Y N>>
CYFRA211 <13006.14
<<Y N>>
sEGFR < 3133.32
<<Y N>>
sHER2 < 10519.465
<<Y N>>
CRC
BR
GI
Thrombospondin2 < 11241.395
<<Y N>>
BR
Liver
GI
PAR < 7422.24
<<Y N>>
Myeloperoxidase < 54.535
<<Y N>>
AFP< 3921.93
<<Y N>>
CRC
Liver
GI
GDF15 < 1.135
<<Y N>>
HGF < 216.685
<<Y N>>
CRC
sFas < 32409.235
<<Y N>>
Liver
OV
sFas < 2056.315
<<Y N>>
BR
TIMP2 < 33335.79
<<Y N>>
CRC
GI

FIG. 68G

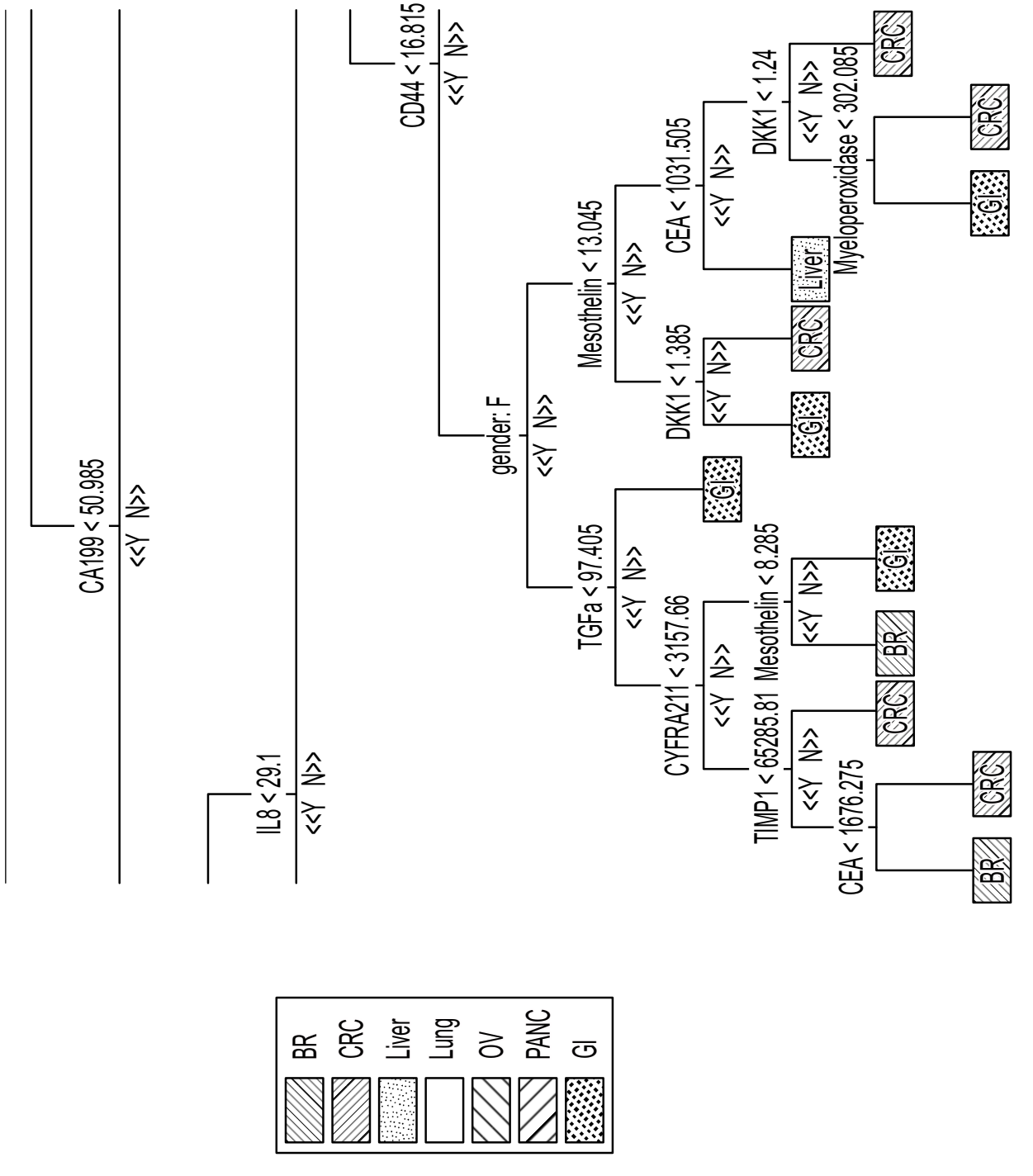

BR
CRC
Liver
Lung
OV
PANC
GI
CA199 < 50.985
<<Y N>>
IL8 < 29.1
<<Y N>>
CD44 < 16.815
<<Y N>>
gender: F
<<Y N>>
TGFa < 97.405
<<Y N>>
CYFRA211 < 3157.66
<<Y N>>
TIMP1 < 65285.81
<<Y N>>
CEA < 1676.275
BR
CRC
CRC
BR
Mesothelin < 8.285
<<Y N>>
GI
GI
Mesothelin < 13.045
<<Y N>>
DKK1 < 1.385
<<Y N>>
CRC
GI
CEA < 1031.505
<<Y N>>
Liver
DKK1 < 1.24
<<Y N>>
Myeloperoxidase < 302.085
CRC
GI
CRC

FIG. 68H

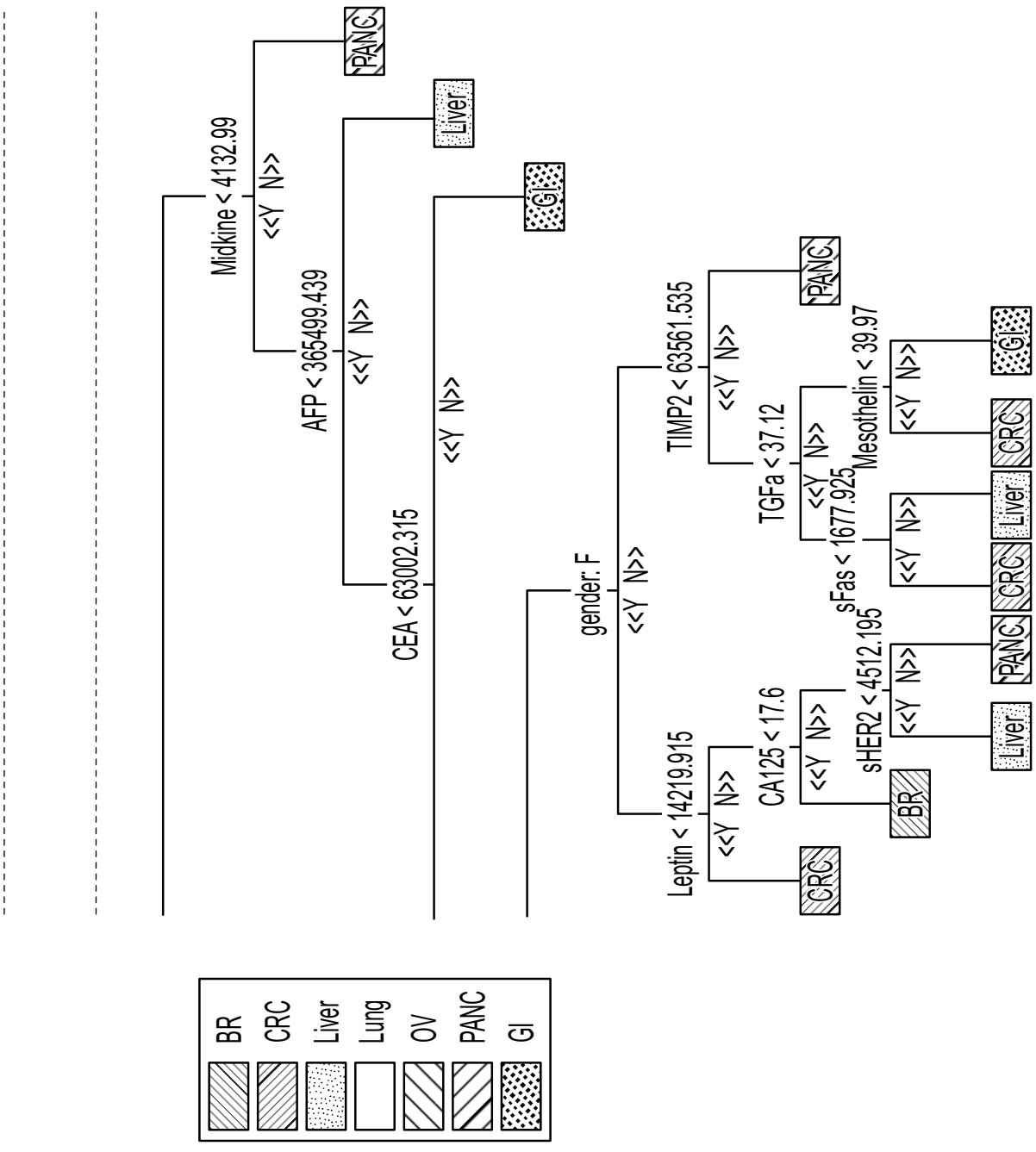

BR
CRC
Liver
Lung
OV
PANC
GI
Midkine < 4132.99
<<Y N>>
PANC
AFP < 365499.439
<<Y N>>
Liver
CEA < 63002.315
<<Y N>>
GI
gender: F
<<Y N>>
Leptin < 14219.915
<<Y N>>
CRC
CA125 < 17.6
<<Y N>>
BR
sHER2 < 4512.195
<<Y N>>
Liver
PANC
TIMP2 < 63561.535
<<Y N>>
PANC
TGFa < 37.12
<<Y N>>
sFas < 1677.925
<<Y N>>
CRC
Liver
Mesothelin < 39.97
<<Y N>>
CRC
GI

FIG. 68I

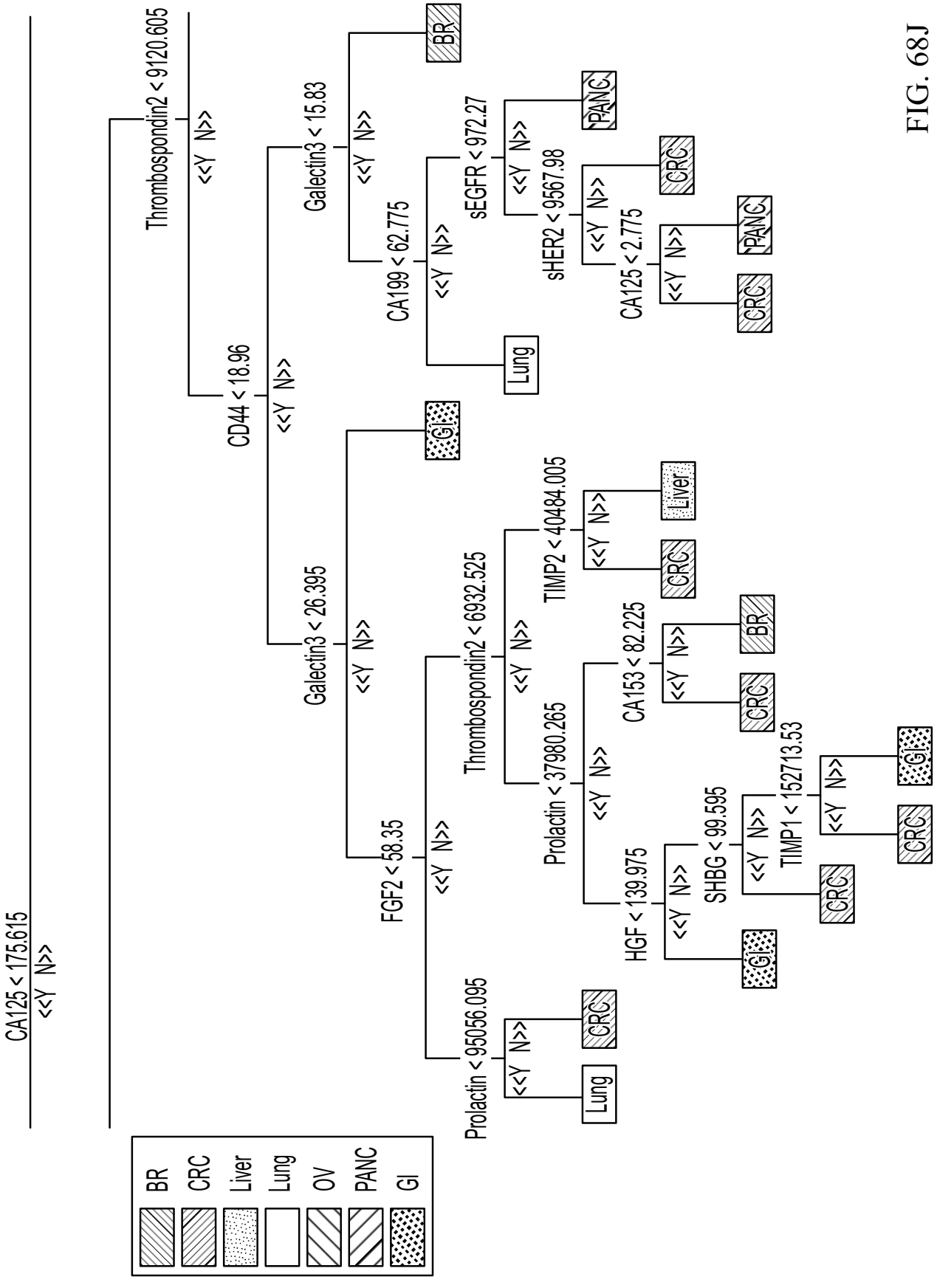

CA125 < 175.615
<<Y N>>
Thrombospondin2 < 9120.605
<<Y N>>
CD44 < 18.96
<<Y N>>
Galectin3 < 26.395
<<Y N>>
FGF2 < 58.35
<<Y N>>
Prolactin < 95056.095
<<Y N>>
Lung
CRC
Thrombospondin2 < 6932.525
<<Y N>>
Prolactin < 37980.265
<<Y N>>
HGF < 139.975
<<Y N>>
GI
SHBG < 99.595
<<Y N>>
CRC
TIMP1 < 152713.53
<<Y N>>
CRC
GI
CA153 < 82.225
<<Y N>>
CRC
BR
TIMP2 < 40484.005
<<Y N>>
CRC
Liver
GI
Galectin3 < 15.83
<<Y N>>
CA199 < 62.775
<<Y N>>
Lung
sEGFR < 972.27
<<Y N>>
sHER2 < 9567.98
<<Y N>>
CA125 < 2.775
<<Y N>>
CRC
PANC
CRC
PANC
PANC
BR
BR
CRC
Liver
Lung
OV
PANC
GI

FIG. 68J

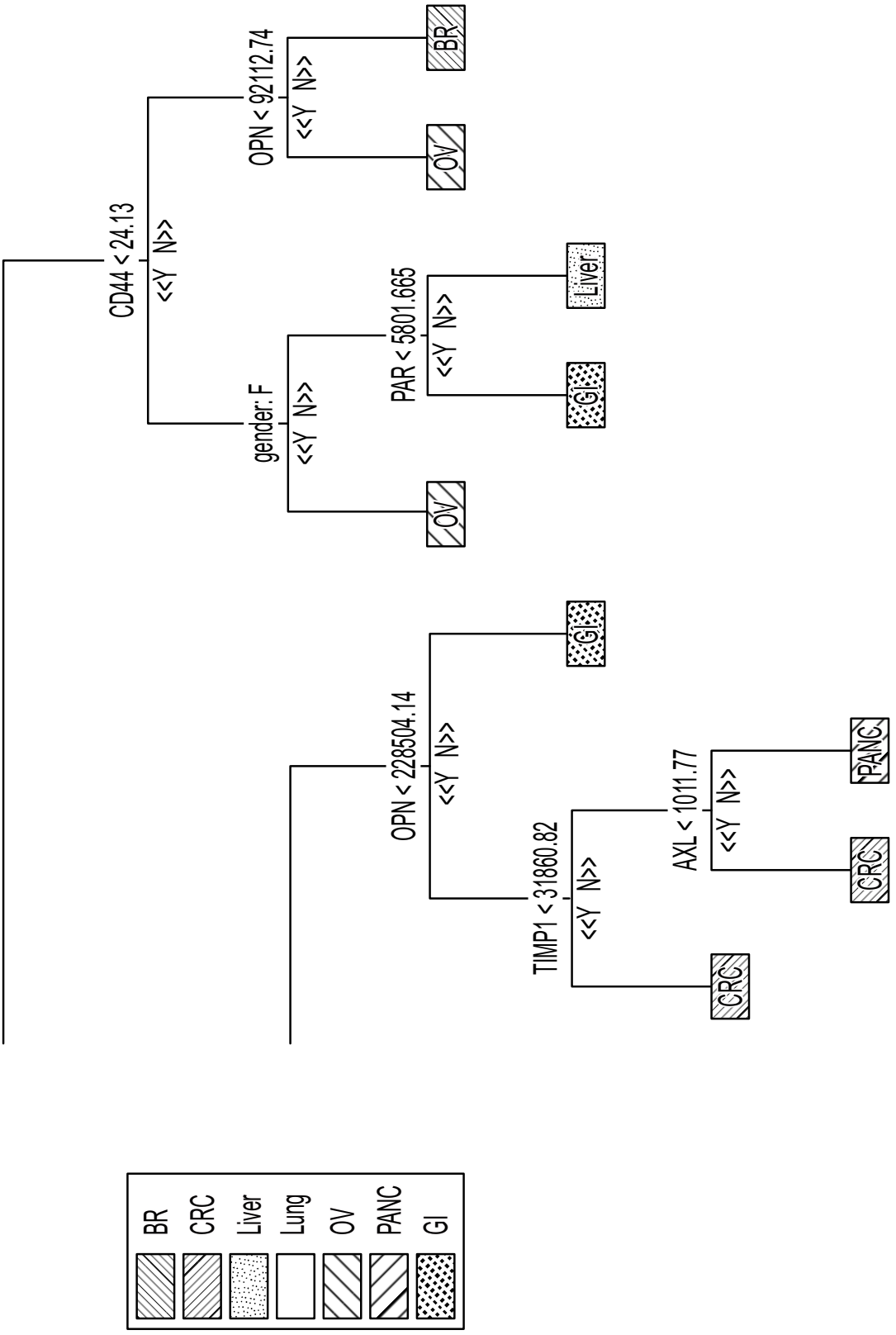

CD44 < 24.13
<<Y N>>
OPN < 92112.74
<<Y N>>
BR
OV
gender: F
<<Y N>>
PAR < 5801.665
<<Y N>>
Liver
GI
OV
OPN < 228504.14
<<Y N>>
GI
TIMP1 < 31860.82
<<Y N>>
AXL < 1011.77
<<Y N>>
PANC
CRC
CRC
BR
CRC
Liver
Lung
OV
PANC
GI

FIG. 68K

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States. 1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] caret_6.0-80          ggplot2_3.0.0         lattice_0.20-35
## [4] inTrees_1.2           DT_0.4                randomForest_4.6-14
##
## loaded via a namespace (and not attached):
## [1] magic_1.5-8           ddalpha_1.3.4         tidyr_0.8.1
## [4] sfsmisc_1.1-2         splines_3.5.0         foreach_1.4.4
## [7] prodlim_2018.04.18    assertthat_0.2.0      stats4_3.5.0
## [10] arules_1.6-1         DRR_0.0.3             yaml_2.1.19
## [13] robustbase_0.93-1    ipred_0.9-6           pillar_1.2.3
## [16] backports_1.1.2      glue_1.2.0            digest_0.6.15
## [19] RRF_1.8              colorspace_1.3-2      recipes_0.1.3
## [22] gbm_2.1.3            htmltools_0.3.6       Matrix_1.2-14
## [25] plyr_1.8.4           psych_1.8.4           timeDate_3043.102
## [28] pkgconfig_2.0.1      CVST_0.2-2            broom_0.4.5
## [31] purrr_0.2.5          xtable_1.8-2          scales_0.5.0
## [34] gower_0.1.2          lava_1.6.2            tibble_1.4.2
## [37] xgboost_0.71.2       withr_2.1.2           nnet_7.3-12
## [40] lazyeval_0.2.1       mnormt_1.5-5          survival_2.42-4
## [43] magrittr_1.5         evaluate_0.11         nlme_3.1-137
## [46] MASS_7.3-50          dimRed_0.1.0          foreign_0.8-70
## [49] class_7.3-14         tools_3.5.0           data.table_1.11.4
## [52] stringr_1.3.1        kernlab_0.9-26        munsell_0.5.0
## [55] bindrcpp_0.2.2       pls_2.6-0             compiler_3.5.0
## [58] RappRoll_0.3.0       rlang_0.2.1           grid_3.5.0
## [61] iterators_1.0.9      htmlwidgets_1.2       rmarkdown_1.10
## [64] geometry_0.3-6       gtable_0.2.0          ModelMetrics_1.1.0
## [67] codetools_0.2-15     abind_1.4-5           reshape2_1.4.3
## [70] R6_2.2.2             lubridate_1.7.4       knitr_1.20
## [73] dplyr_0.7.6          bindr_0.1.1           rprojroot_1.3-2
## [76] string_1.2.3         parallel_3.5.0        Rcpp_0.12.17
## [79] rpart_4.1-13         DEoptimR_1.0-8        tidyselect_0.2.4
```

FIG. 69

```
## [1] "Tumor samples: n = 626"
```

```
## [1] "Proteins: AFP, CA125, CA153, CA199, CEA, CYFRA211, FGF2, HE4, HGF,
IL6, IL8, Leptin, OPN, Prolactin, sFas, TGFa, TIMP1, TIMP2, Galectin3,
Myeloperoxidase, SHBG, DKK1, GDF15, NSE, OPG, CD44, Mesothelin, Midkine,
AXL, PAR, sEGFR, sHER2, Thrombospondin2"
```

```
## [1] "Predictions from randomForest:"
```

```
##
##      Breast         CRC       Liver        Lung     Ovarian    Pancreas      st_eso
##          48         392          17          25          41          62          41
```

```
## 18635 rules (length<=6) were extracted from the first 500 trees.
```

```
## [1] "Predictions from inTrees classifier:"
```

```
## predictions
##      Breast         CRC       Liver        Lung     Ovarian    Pancreas      st_eso
##          57         352          19          36          45          50          67
```

```
## [1] "Accuracies from the paper:"
```

```
##      Breast         CRC       Liver        Lung     Ovarian    Pancreas      st_eso
##        0.63        0.84        0.44        0.39        0.79        0.81        0.46
```

```
## [1] "Accuracies from inTrees classifier:"
```

```
##      Breast         CRC       Liver        Lung     Ovarian    Pancreas      st_eso
## 0.6376812   0.8893281   0.3863636   0.4098361   0.8113208   0.7058824   0.6025641
```

FIG. 69 (Cont.)