Online Data Supplement

Figure E1

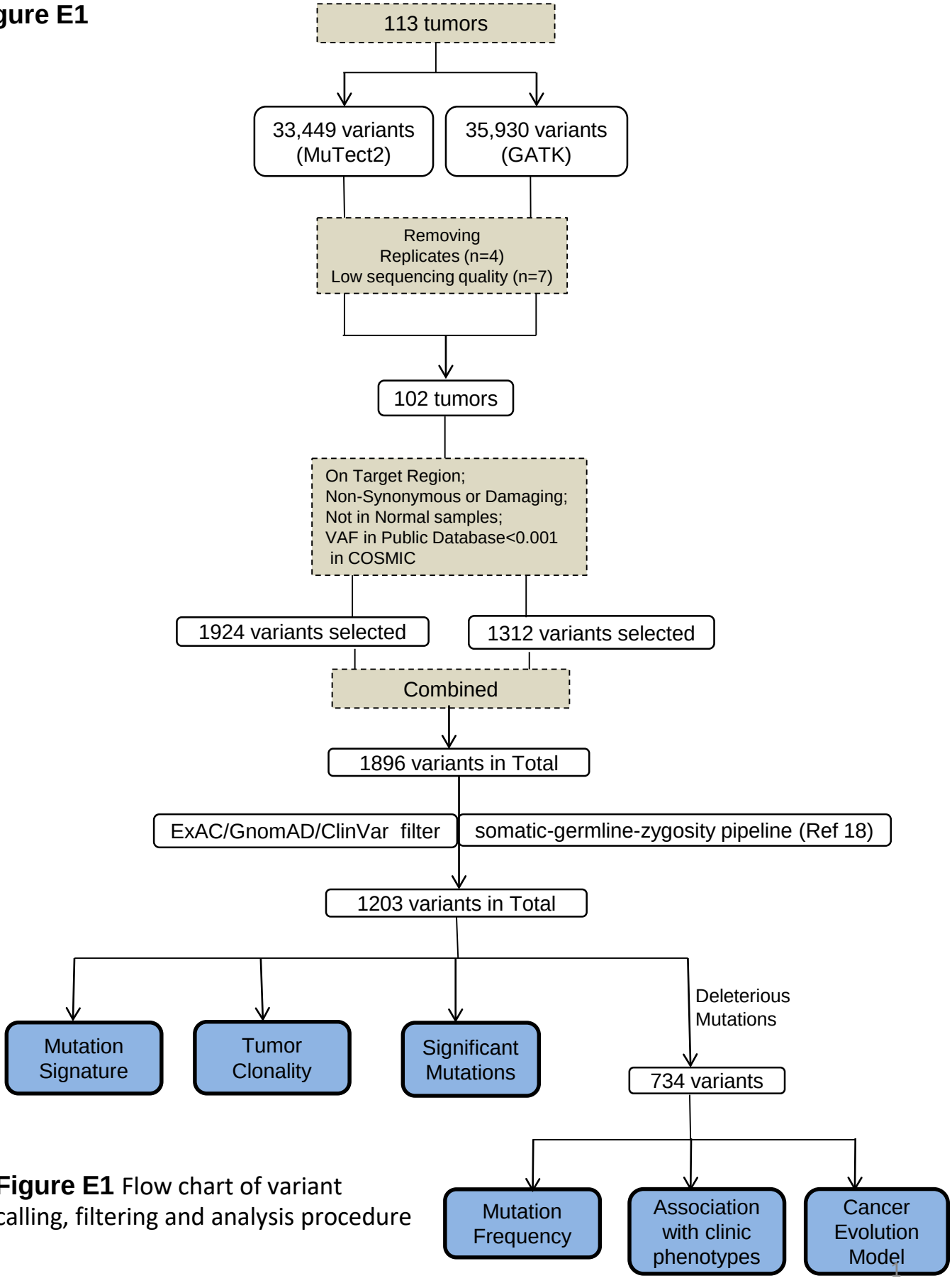


Figure E1 Flow chart of variant calling, filtering and analysis procedure

Figure E2. Somatic mutations and copy number in AIS, MIA and ADC. (A) The proportion of all mutation types in our cohort. (B) Nine genes harboring significant hot spot mutations were identified using the oncodriveCluster approach. (C) Frequencies of significantly amplified and deleted regions or genes identified using RUBIC tool. (D) Overview of top gene mutations and copy number alterations.

A

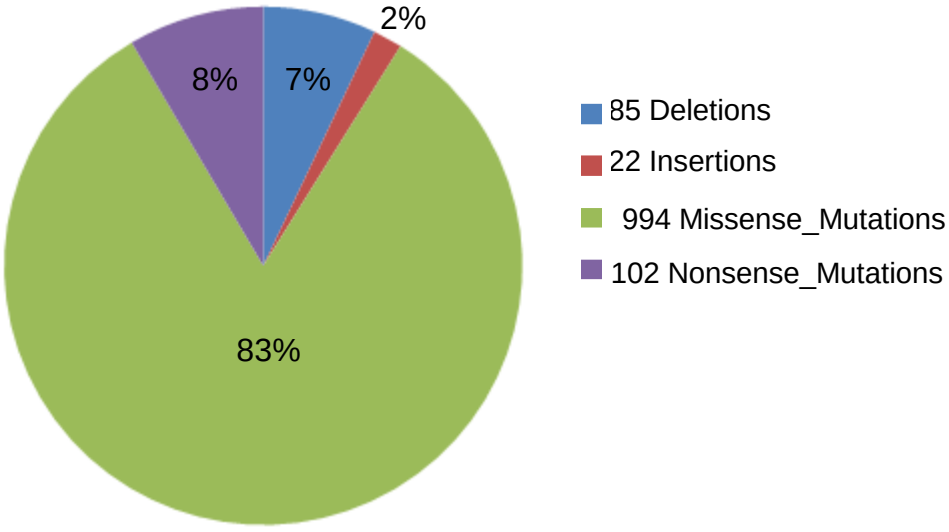
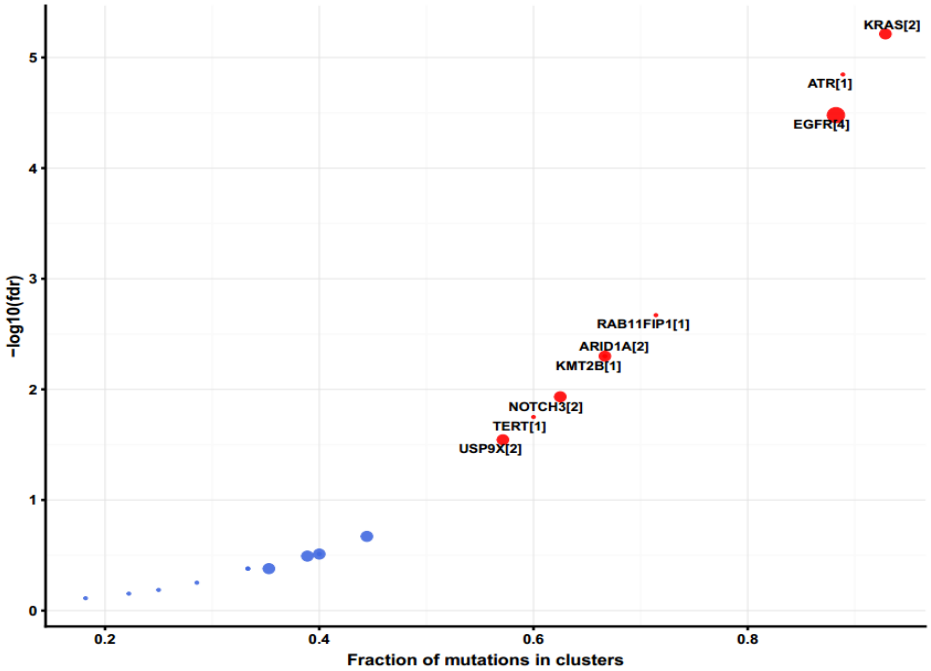


Figure E2

B


KRAS[2]
ATR[1]
EGFR[4]
RAB11FIP1[1]
ARID1A[2]
KMT2B[1]
NOTCH3[2]
TERT[1]
USP9X[2]
−log10(fdr)
Fraction of mutations in clusters
0
1
2
3
4
5
0.2
0.4
0.6
0.8

C

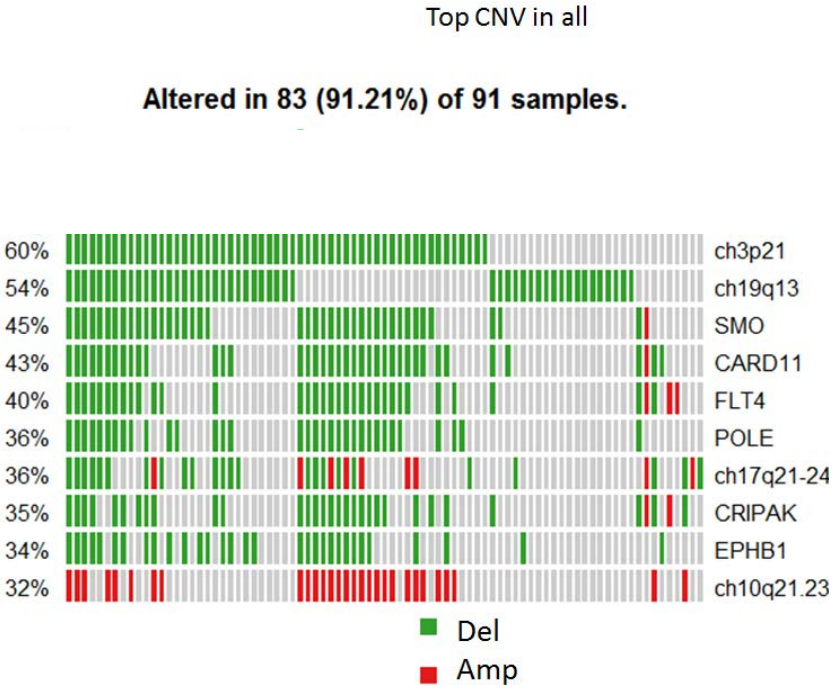

Top CNV in all
Altered in 83 (91.21%) of 91 samples.
60% ch3p21
54% ch19q13
45% SMO
43% CARD11
40% FLT4
36% POLE
36% ch17q21-24
35% CRIPAK
34% EPHB1
32% ch10q21.23
Del
Amp

Figure E2

D

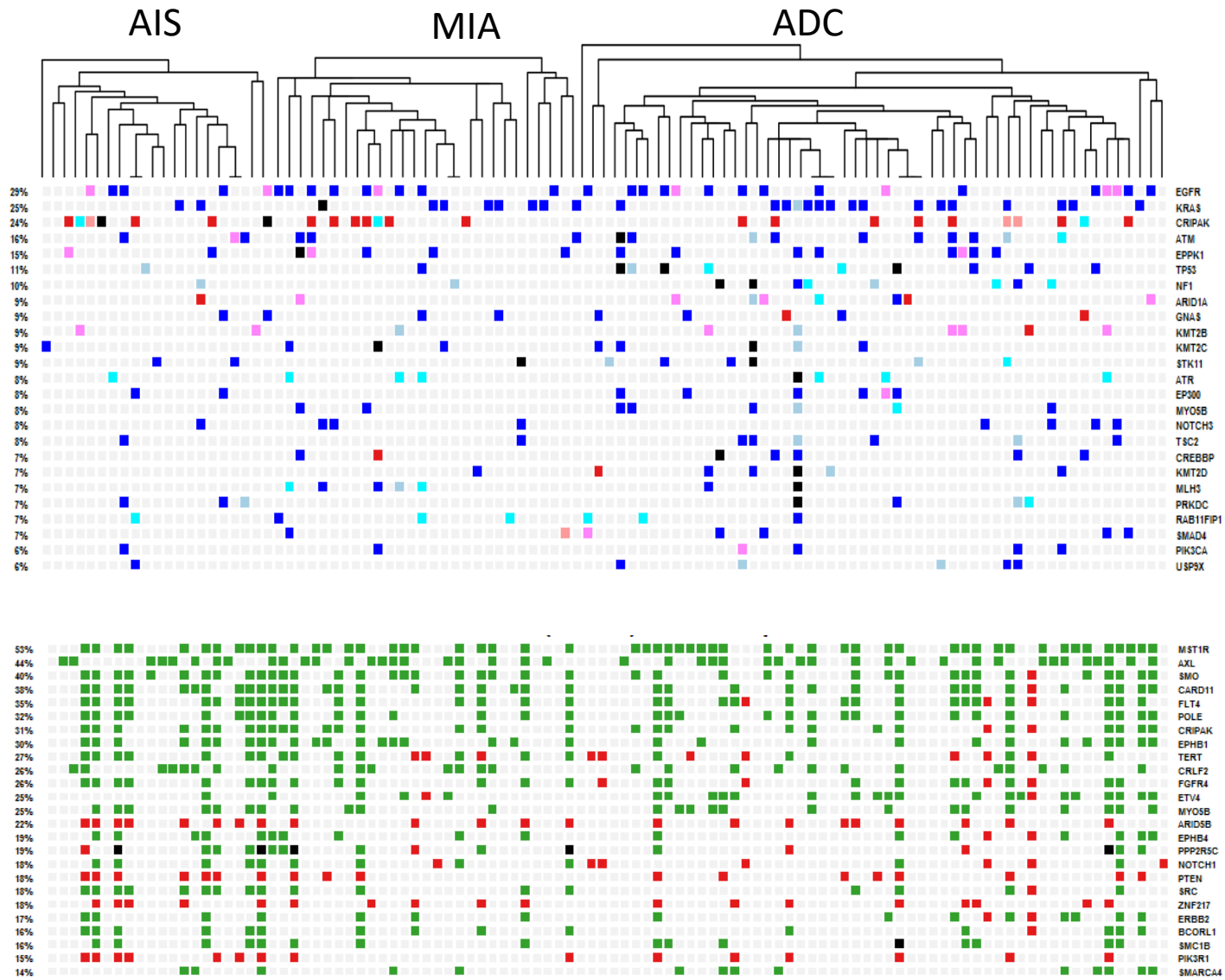

AIS
MIA
ADC
29% EGFR
25% KRAS
24% CRIPAK
16% ATM
15% EPPK1
11% TP53
10% NF1
9% ARID1A
9% GNAS
9% KMT2B
9% KMT2C
9% STK11
8% ATR
8% EP300
8% MYO5B
8% NOTCH3
8% TSC2
7% CREBBP
7% KMT2D
7% MLH3
7% PRKDC
7% RAB11FIP1
7% SMAD4
6% PIK3CA
6% USP9X
53% MST1R
44% AXL
40% SMO
38% CARD11
35% FLT4
32% POLE
31% CRIPAK
30% EPHB1
27% TERT
26% CRLF2
26% FGFR4
25% ETV4
25% MYO5B
22% ARID5B
19% EPHB4
19% PPP2R5C
18% NOTCH1
18% PTEN
18% SRC
18% ZNF217
17% ERBB2
16% BCORL1
16% SMC1B
15% PIK3R1
14% SMARCA4

Figure E3. Mutational signature and tumor heterogeneity analysis. (A) Five mutation signatures identified by non-negative matrix factorization (NMF) approach on 96 trinucleotide mutation types across all tumors. Five mutation signatures include smoke exposure (signature 4), APOBEC activation (signature 2), aging (signature 1), mismatch repair (MMR) (signature 6) and unknown (signature 18). Each bar corresponds to the probability of observing a particular mutation in a trinucleotide context within each signature. **(B)** Relative contribution of each mutation signature in each tumor. **(C)** Mutation burden was significantly higher in tumors enriched in APOBEC and Sig18.unknown when compared to MMR signature enriched tumors. P values were calculated using negative binomial regression model. **(D)** Mutant-allele tumor heterogeneity (MATH) score tended to be positively associated with mutation burden (r=0.19, P=0.07) and was not associated with smoking status, age , stage, gender or tumor size. **(E)** Two representative cancer cell fraction (CCF) density plot using Sciclone showed one mulita-clonal tumor (T45) and one mono-clonal tumor (T40). **(F).** The proportions of clonal/subclonal mutations were not significantly associated with the histology. **(G)** Tumors harboring any EGFR mutations or only subclonal EGFR mutations were not associated with overall survival. P values shown were calculated using Kaplan-Meier estimate. **(H)** VAF plot for EGFR mutations. The majority of subclonal mutations of EGFR were p.L858R.

Figure E3

A

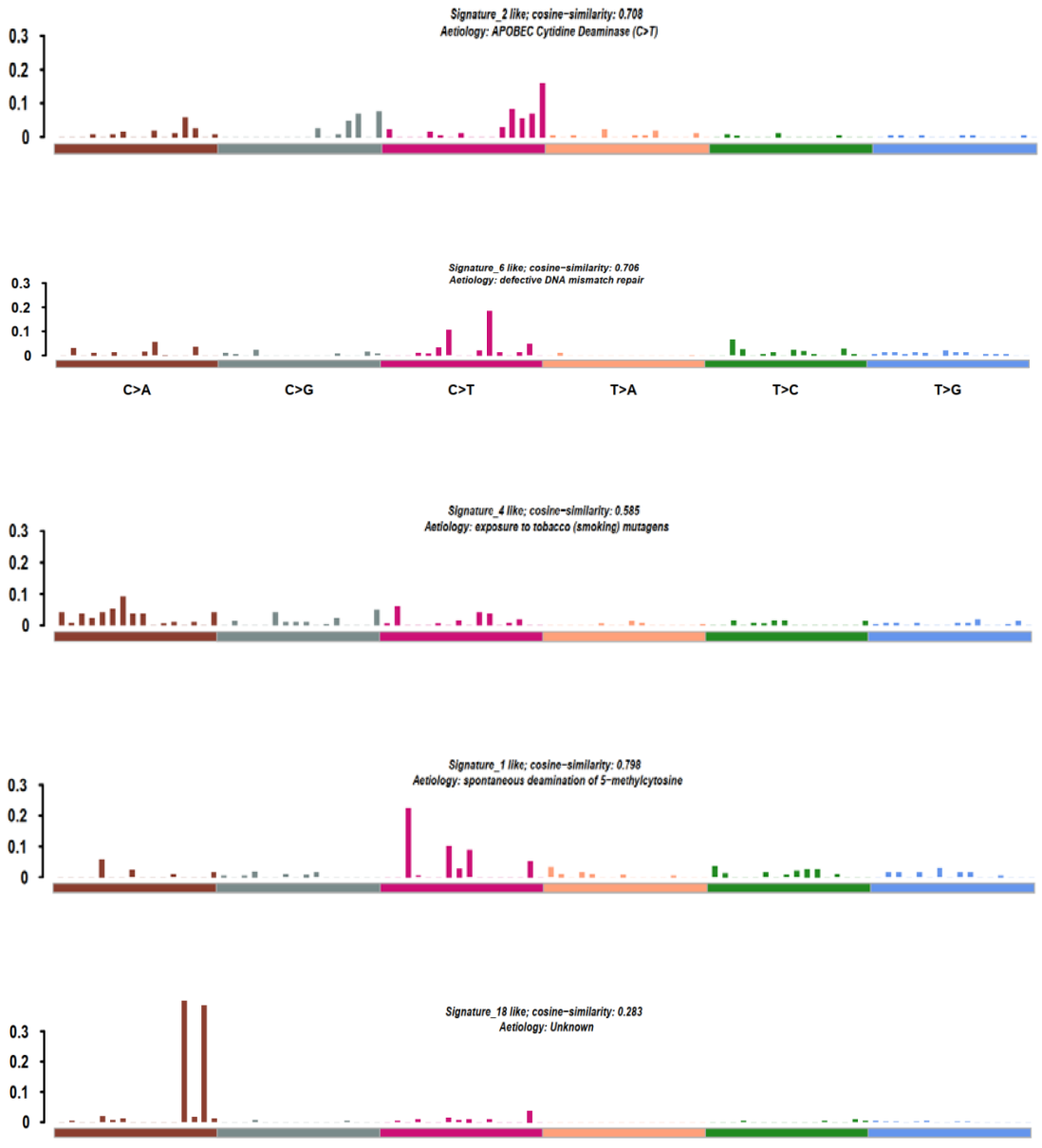
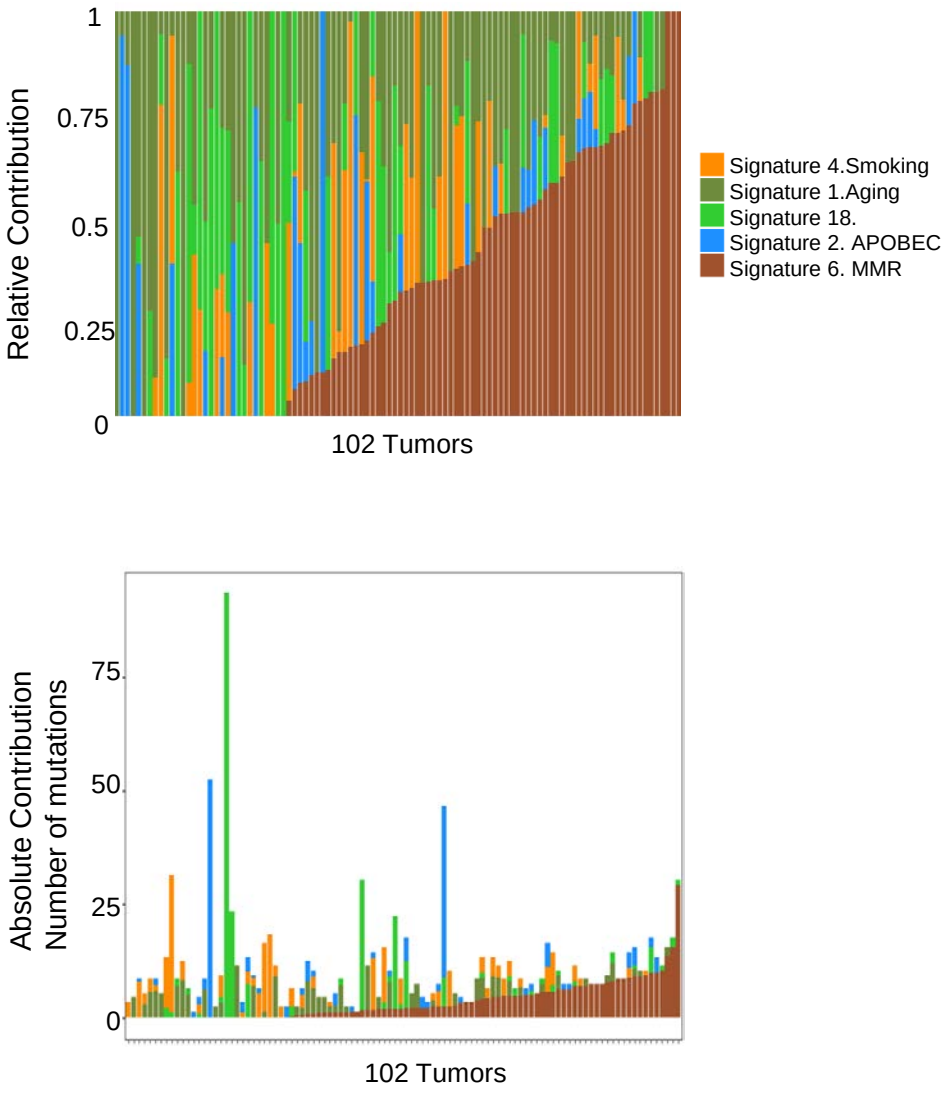

Signature_2 like; cosine−similarity: 0.708
Aetiology: APOBEC Cytidine Deaminase (C>T)
Signature_6 like; cosine−similarity: 0.706
Aetiology: defective DNA mismatch repair
Signature_4 like; cosine−similarity: 0.585
Aetiology: exposure to tobacco (smoking) mutagens
Signature_1 like; cosine−similarity: 0.798
Aetiology: spontaneous deamination of 5−methylcytosine
Signature_18 like; cosine−similarity: 0.283
Aetiology: Unknown
0.3
0.2
0.1
0
C>A
C>G
C>T
T>A
T>C
T>G

Figure E3

B


Relative Contribution
1
0.75
0.5
0.25
0
102 Tumors
Signature 4.Smoking
Signature 1.Aging
Signature 18.
Signature 2. APOBEC
Signature 6. MMR
Absolute Contribution
Number of mutations
75
50
25
0
102 Tumors

C

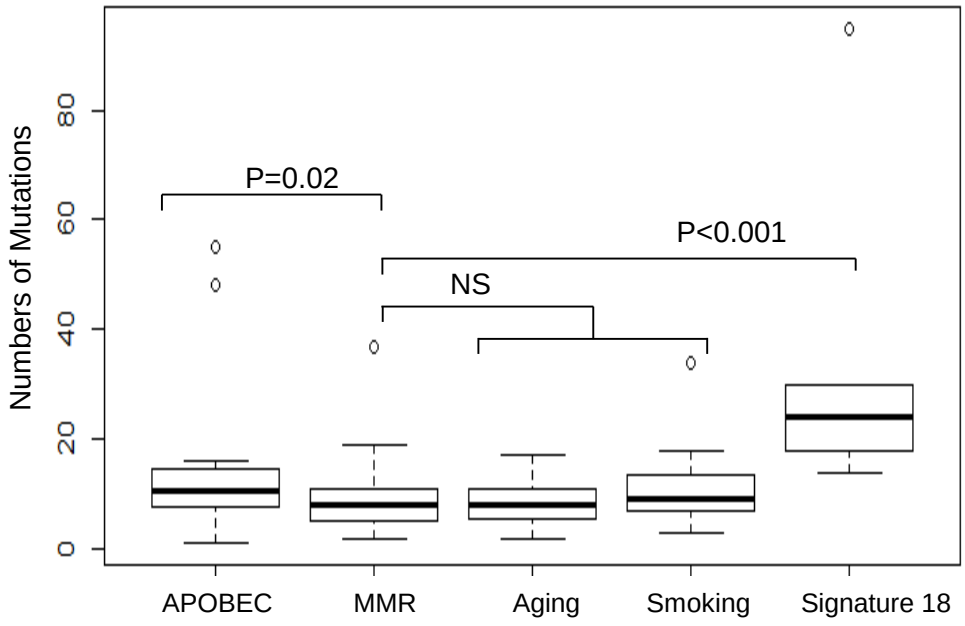
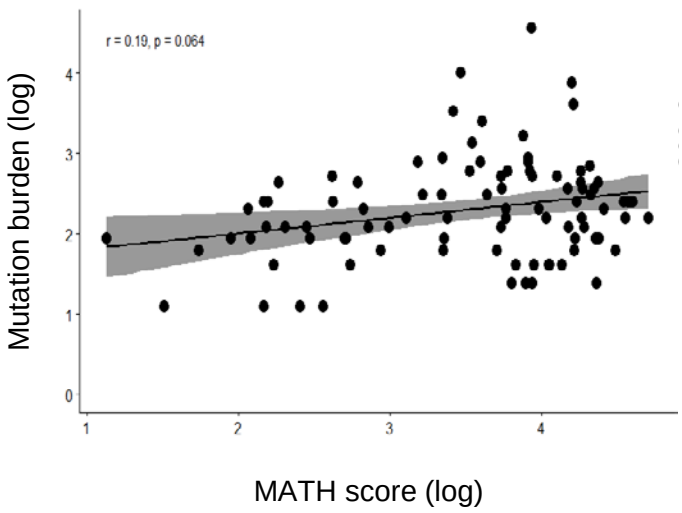
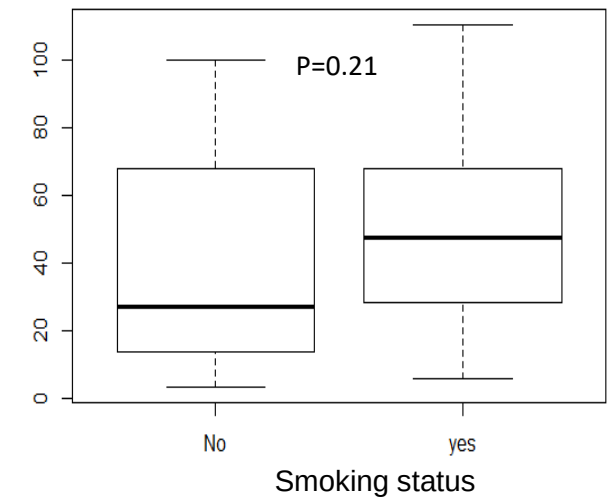
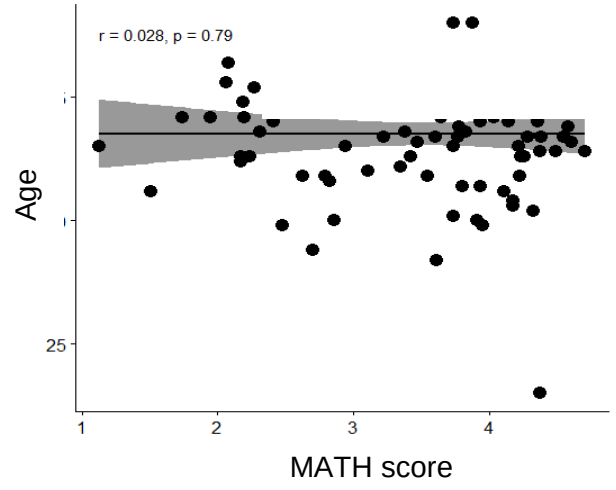

Numbers of Mutations
0
20
40
60
80
P=0.02
P<0.001
NS
APOBEC
MMR
Aging
Smoking
Signature 18

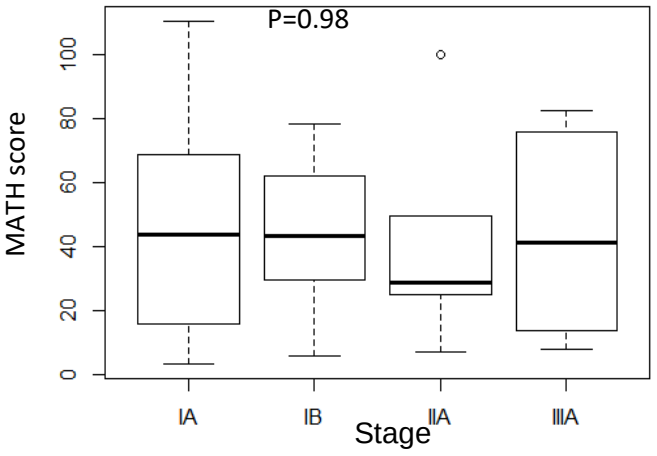
Figure E3

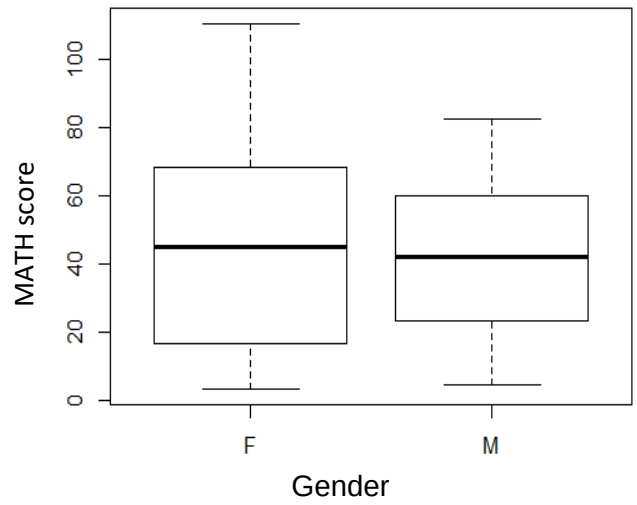
D


r = 0.19, p = 0.064
Mutation burden (log)
MATH score (log)


P=0.21
No
yes
Smoking status


r = 0.028, p = 0.79
Age
MATH score


P=0.98
MATH score
IA
IB
IIA
IIIA
Stage


MATH score
F
M
Gender

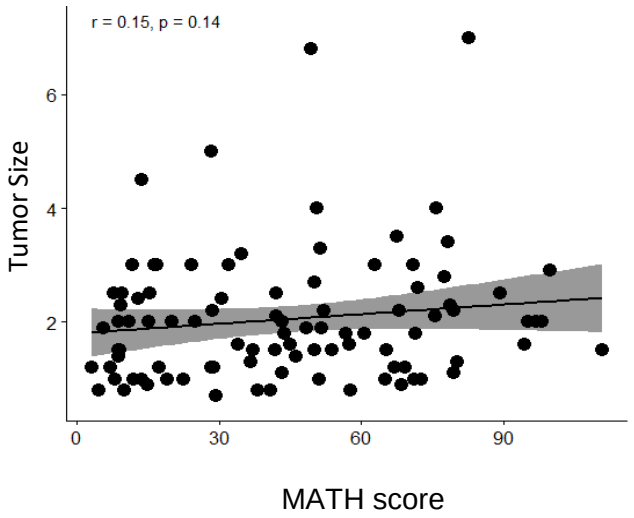
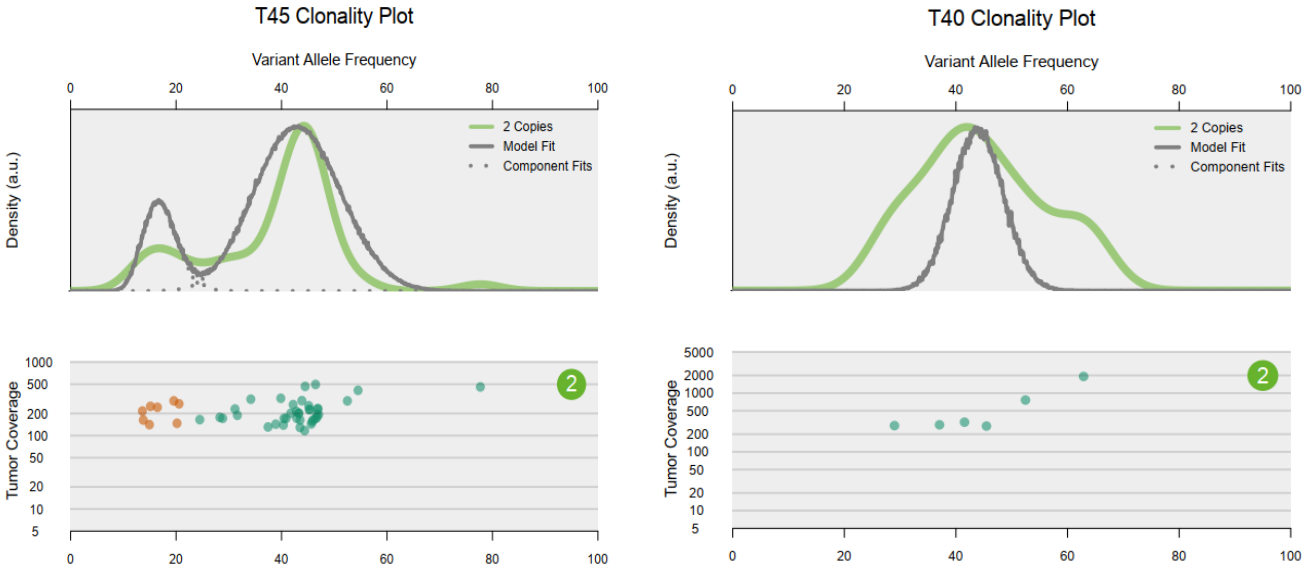

r = 0.15, p = 0.14
Tumor Size
MATH score

Figure E3

E


T45 Clonality Plot
Variant Allele Frequency
Density (a.u.)
2 Copies
Model Fit
Component Fits
Tumor Coverage
T40 Clonality Plot
Variant Allele Frequency
Density (a.u.)
2 Copies
Model Fit
Component Fits
Tumor Coverage

F

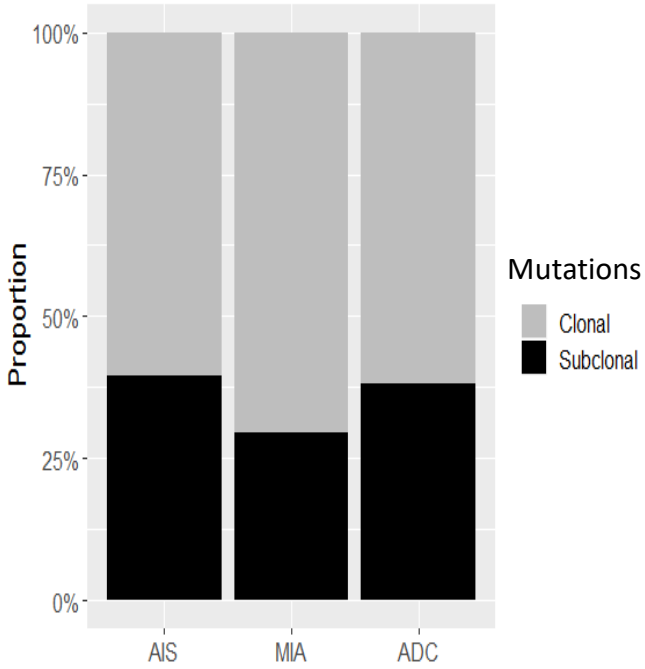
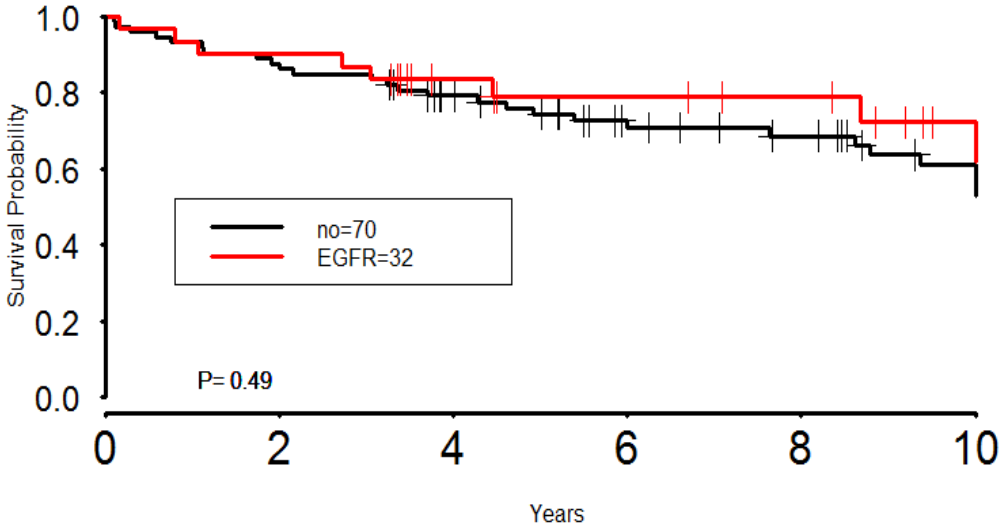

Proportion
100%
75%
50%
25%
0%
AIS
MIA
ADC
Mutations
Clonal
Subclonal

Figure E3

G


1.0
0.8
0.6
0.4
0.2
0.0
Survival Probability
no=70
EGFR=32
P= 0.49
0
2
4
6
8
10
Years

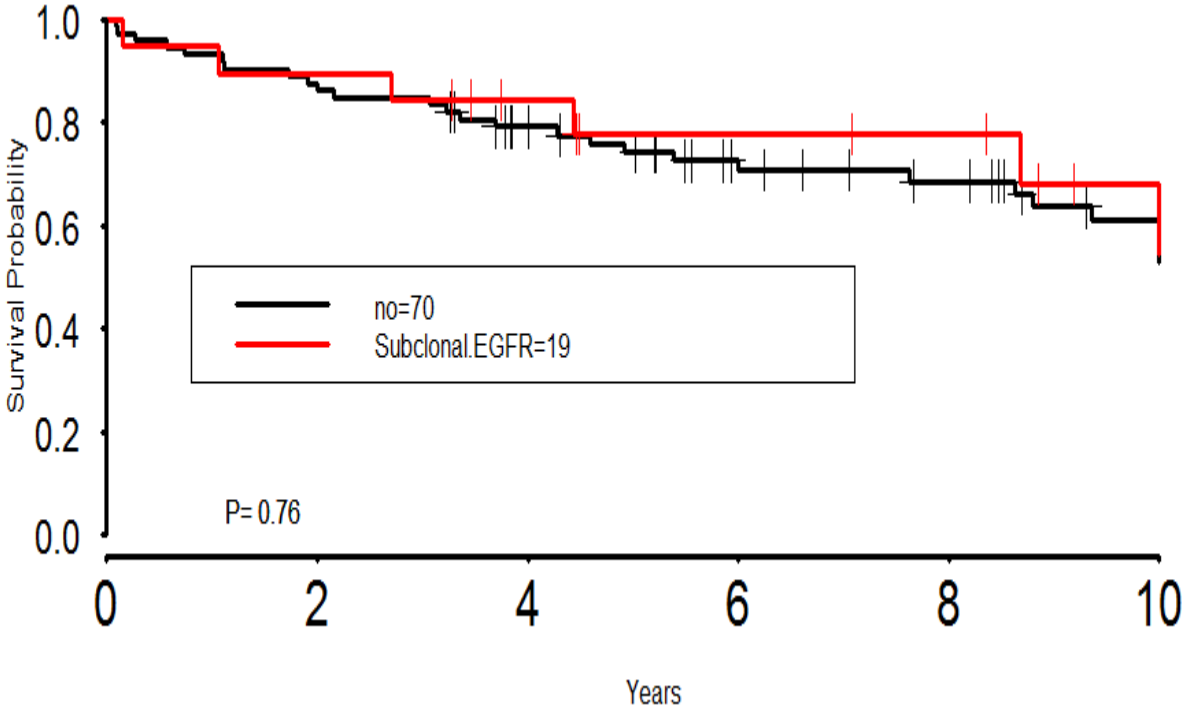

1.0
0.8
0.6
0.4
0.2
0.0
Survival Probability
no=70
Subclonal.EGFR=19
P= 0.76
0
2
4
6
8
10
Years

Figure E3

H

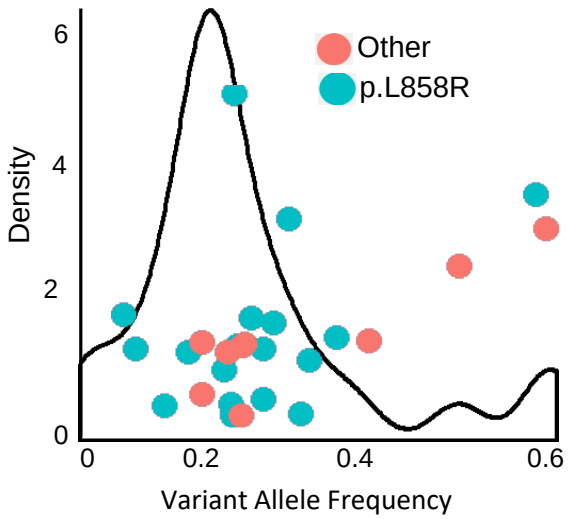
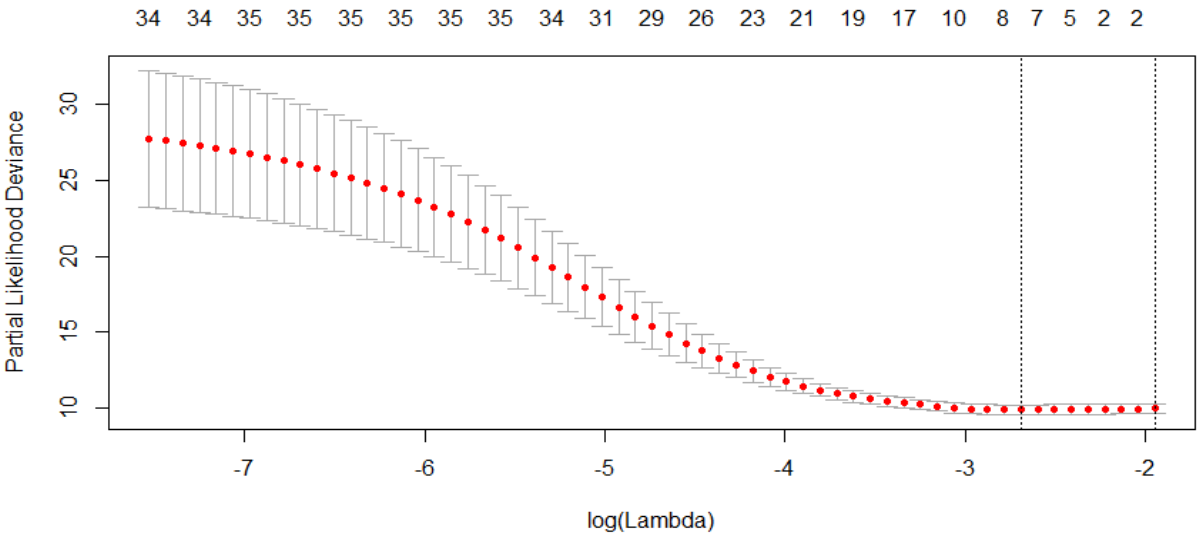
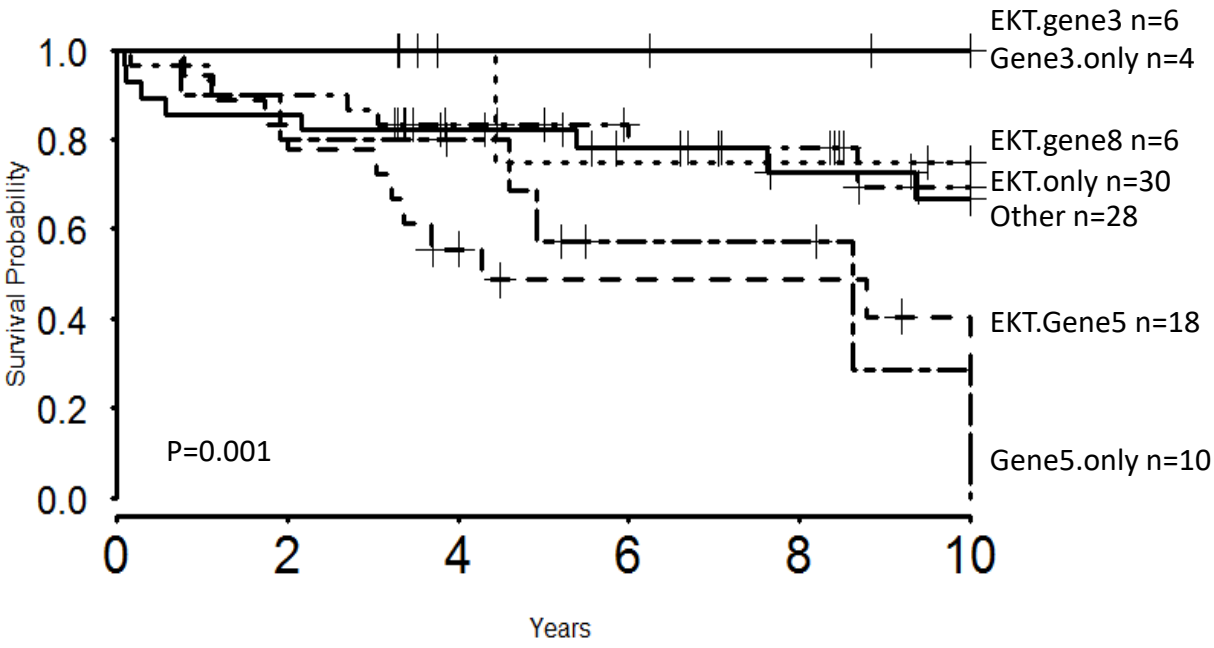

Other
p.L858R
Density
6
4
2
0
0
0.2
0.4
0.6
Variant Allele Frequency

Figure E4. Association between gene mutations and 10-year overall survival. **(A)** Eight mutated genes selected by using an elastic-net penalized regression modeling (Glmnet R package). **(B)** Kaplan-Meier (KM) plot showed three patient groups had distinct survival outcomes. Group 1 includes patients having 5 gene (Gene5) mutation set and patients having concurrent EKT mutations (EGFR, KRAS and TP53); group2 includes patients having any of EKT mutations, patients having concurrent EKT with Gene8 mutations and patients having none of EKT or Gene8 mutations; group3 includes those patients having any of Gene3 mutations only or concurrent EKT with Gene3 mutations. **(C)** Group1 tumors harboring Gene5 only mutations or harboring concurrent EKT with Gene5 mutations were significantly associated with worse overall survival, after adjusting for age, gender, LVI and histology. **(D)** Tumors harboring any of Gene5 mutations, but not having any of EKT or Gene3 mutations were significantly associated with worse overall survival, after adjusting for age, gender, LVI or histology.

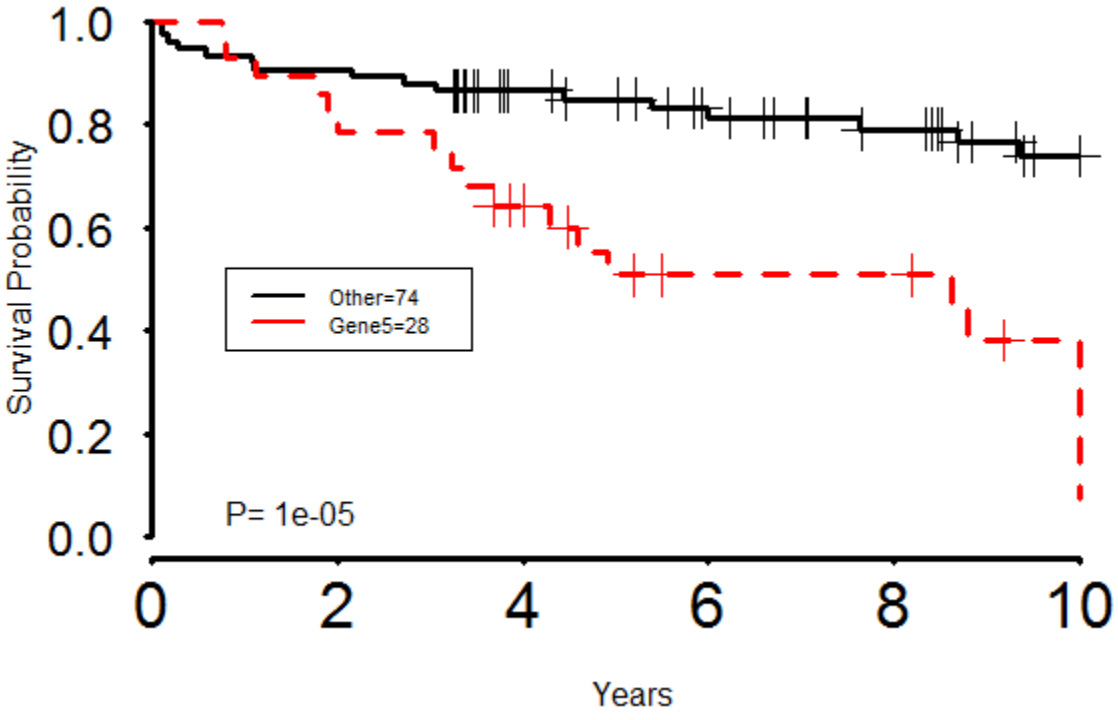
A.



B


EKT.gene3 n=6
Gene3.only n=4
EKT.gene8 n=6
EKT.only n=30
Other n=28
EKT.Gene5 n=18
Gene5.only n=10
P=0.001
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
Years

C


Other=74
Gene5=28
P= 1e-05
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
Years

D

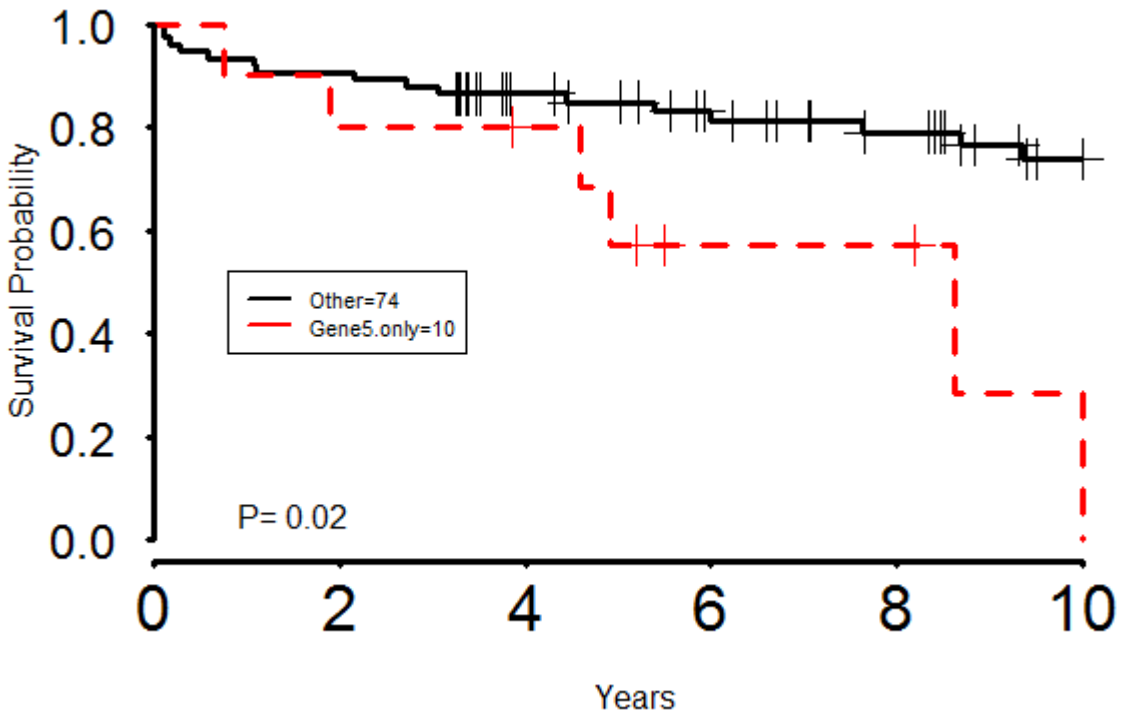

1.0
0.8
0.6
0.4
0.2
0.0
Survival Probability
Other=74
Gene5.only=10
P= 0.02
0
2
4
6
8
10
Years

Figure E5. Cancer evolution model in early ADC detected by CAPRI algorithm. **(A)** Figure E5A has all tumors harboring a total of 59 significantly mutated genes plus four significantly altered CNVs with confidence shown as edge labels. The first, the second and third are p-values for temporal priority, probability raising and hypergeometric test. The last represents the relation confidence estimated with 100 non-parametric bootstrap (npb score) iterations. Red p-values are above the minimum significance threshold of 0.05. See Methods in the online supplement and Figure 4A in the Main Text for an interpretation of this model. **(B)** and **(C)** showed cancer evolution model in AIS/MIA combined and ADC using CAPRI algorithm. Only nonparametric bootstrap scores were shown. **(D)** Significantly mutually exclusive mutations between EGFR and KRAS (P=0.0005), EGFR and NF1 (P=0.04), and co-occurring mutations between EGFR and ATR (P=0.0007), EGFR and SMAD4 (P=0.02), STK11 and KEAP1 (P=0.0001). P values were calculated using Fisher exact test.

Figure E5A

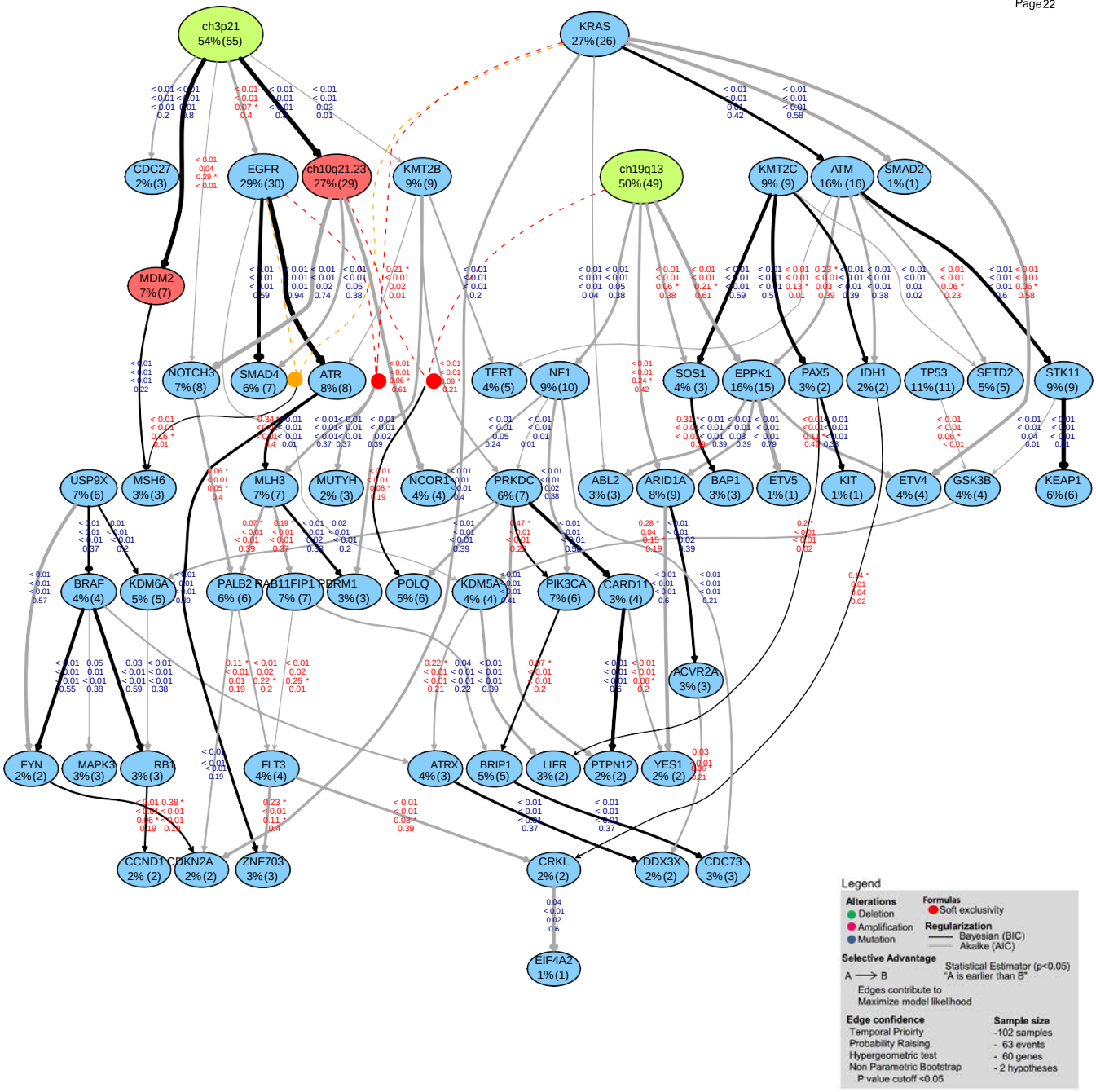

ch3p21
54% (55)
KRAS
27% (26)
CDC27
2% (3)
EGFR
29% (30)
ch10q21.23
27% (29)
KMT2B
9% (9)
ch19q13
50% (49)
KMT2C
9% (9)
ATM
16% (16)
SMAD2
1% (1)
MDM2
7% (7)
NOTCH3
7% (8)
SMAD4
6% (7)
ATR
8% (8)
TERT
4% (5)
NF1
9% (10)
SOS1
4% (3)
EPPK1
16% (15)
PAX5
3% (2)
IDH1
2% (2)
TP53
11% (11)
SETD2
5% (5)
STK11
9% (9)
USP9X
7% (6)
MSH6
3% (3)
MLH3
7% (7)
MUTYH
2% (3)
NCOR1
4% (4)
PRKDC
6% (7)
ABL2
3% (3)
ARID1A
8% (9)
BAP1
3% (3)
ETV5
1% (1)
KIT
1% (1)
ETV4
4% (4)
GSK3B
4% (4)
KEAP1
6% (6)
BRAF
4% (4)
KDM6A
5% (5)
PALB2
6% (6)
RAB11FIP1
7% (7)
PRRM1
3% (3)
POLQ
5% (6)
KDM5A
4% (4)
PIK3CA
7% (6)
CARD11
3% (4)
ACVR2A
3% (3)
FYN
2% (2)
MAPK3
3% (3)
RB1
3% (3)
FLT3
4% (4)
ATRX
4% (3)
BRIP1
5% (5)
LIFR
3% (2)
PTPN12
2% (2)
YES1
2% (2)
CCND1
2% (2)
CDKN2A
2% (2)
ZNF703
3% (3)
CRKL
2% (2)
DDX3X
2% (2)
CDC73
3% (3)
EIF4A2
1% (1)
Legend
Alterations
Deletion
Amplification
Mutation
Formulas
Soft exclusivity
Regularization
Bayesian (BIC)
Akaike (AIC)
Selective Advantage
A ⟶ B
Edges contribute to
Maximize model likelihood
Statistical Estimator (p<0.05)
"A is earlier than B"
Edge confidence
Temporal Priority
Probability Raising
Hypergeometric test
Non Parametric Bootstrap
P value cutoff <0.05
Sample size
-102 samples
- 63 events
- 60 genes
- 2 hypotheses

Figure E5B

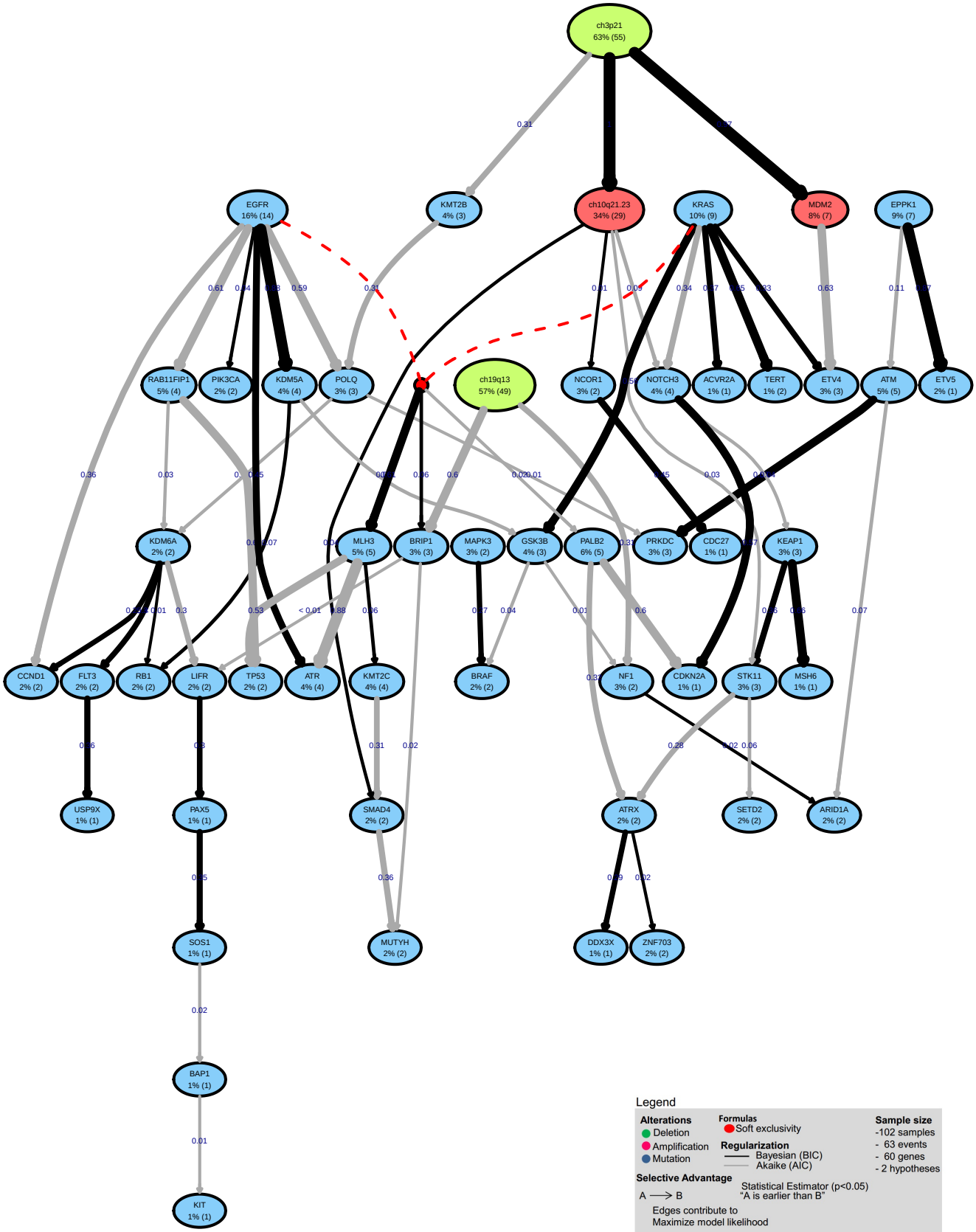
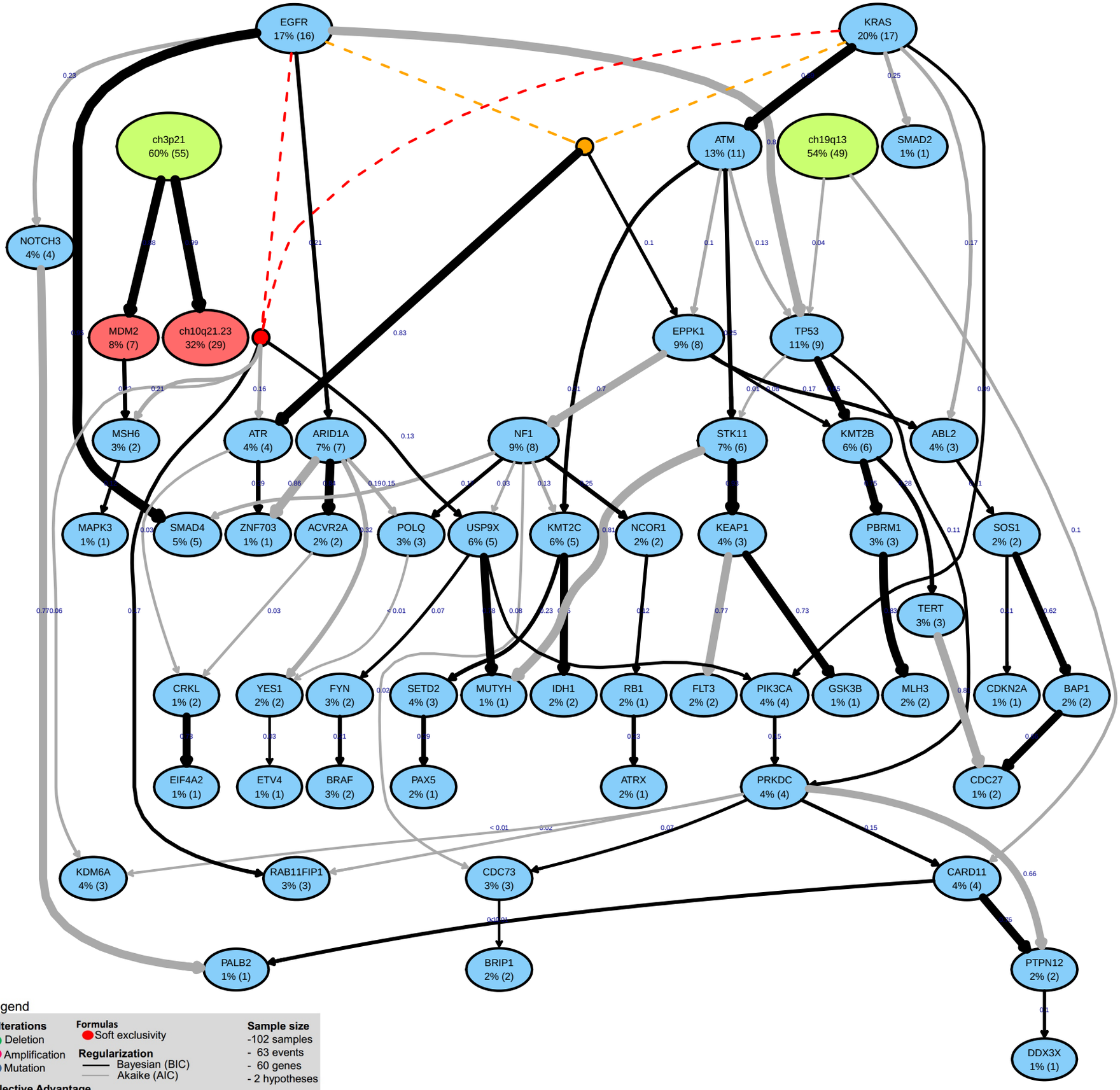

ch3p21
63% (55)
EGFR
16% (14)
KMT2B
4% (3)
ch10q21.23
34% (29)
KRAS
10% (9)
MDM2
8% (7)
EPPK1
9% (7)
RAB11FIP1
5% (4)
PIK3CA
2% (2)
KDM5A
4% (4)
POLQ
3% (3)
ch19q13
57% (49)
NCOR1
3% (2)
NOTCH3
4% (4)
ACVR2A
1% (1)
TERT
1% (2)
ETV4
3% (3)
ATM
5% (5)
ETV5
2% (1)
KDM6A
2% (2)
MLH3
5% (5)
BRIP1
3% (3)
MAPK3
3% (2)
GSK3B
4% (3)
PALB2
6% (5)
PRKDC
3% (3)
CDC27
1% (1)
KEAP1
3% (3)
CCND1
2% (2)
FLT3
2% (2)
RB1
2% (2)
LIFR
2% (2)
TP53
2% (2)
ATR
4% (4)
KMT2C
4% (4)
BRAF
2% (2)
NF1
3% (2)
CDKN2A
1% (1)
STK11
3% (3)
MSH6
1% (1)
USP9X
1% (1)
PAX5
1% (1)
SMAD4
2% (2)
ATRX
2% (2)
SETD2
2% (2)
ARID1A
2% (2)
SOS1
1% (1)
MUTYH
2% (2)
DDX3X
1% (1)
ZNF703
2% (2)
BAP1
1% (1)
KIT
1% (1)
Legend
Alterations
Deletion
Amplification
Mutation
Formulas
Soft exclusivity
Regularization
Bayesian (BIC)
Akaike (AIC)
Sample size
-102 samples
- 63 events
- 60 genes
- 2 hypotheses
Selective Advantage
A ⟶ B
Statistical Estimator (p<0.05)
"A is earlier than B"
Edges contribute to
Maximize model likelihood

Figure E5C



Legend

Alterations
- Deletion
- Amplification
- Mutation

Formulas
- Soft exclusivity

Regularization
- Bayesian (BIC)
- Akaike (AIC)

Sample size
-102 samples
- 63 events
- 60 genes
- 2 hypotheses

Selective Advantage

A ⟶ B

Statistical Estimator (p<0.05)
"A is earlier than B"

Edges contribute to
Maximize model likelihood

Figure E5

D

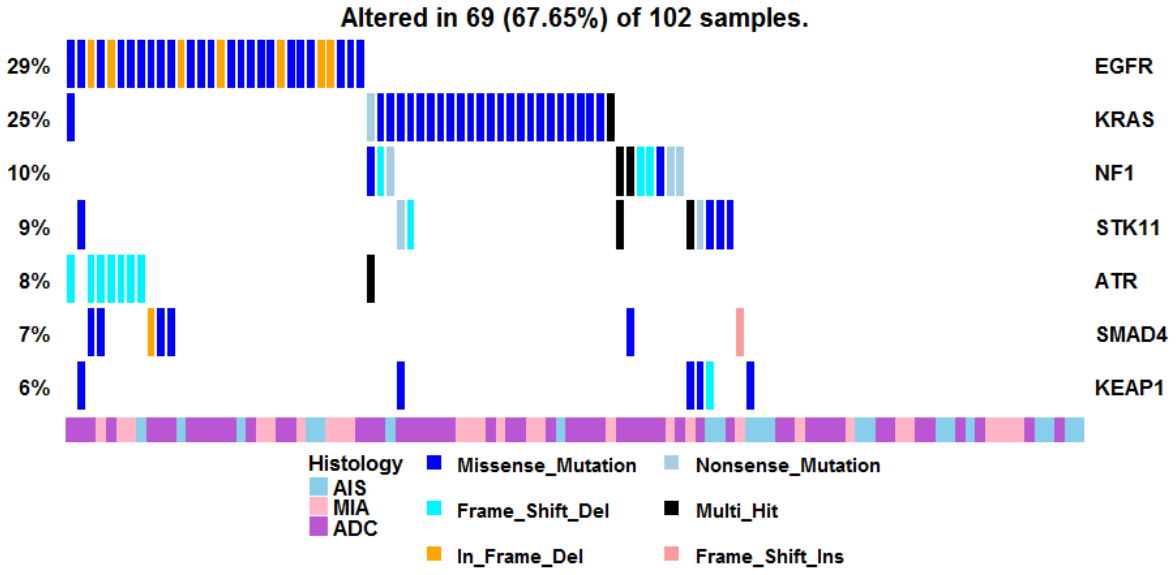

Altered in 69 (67.65%) of 102 samples.
29%
25%
10%
9%
8%
7%
6%
EGFR
KRAS
NF1
STK11
ATR
SMAD4
KEAP1
Histology
AIS
MIA
ADC
Missense_Mutation
Frame_Shift_Del
In_Frame_Del
Nonsense_Mutation
Multi_Hit
Frame_Shift_Ins

Figure E6

Figure E6. Non-supervised cluster analysis using genes that were mutated in greater than 5% tumors

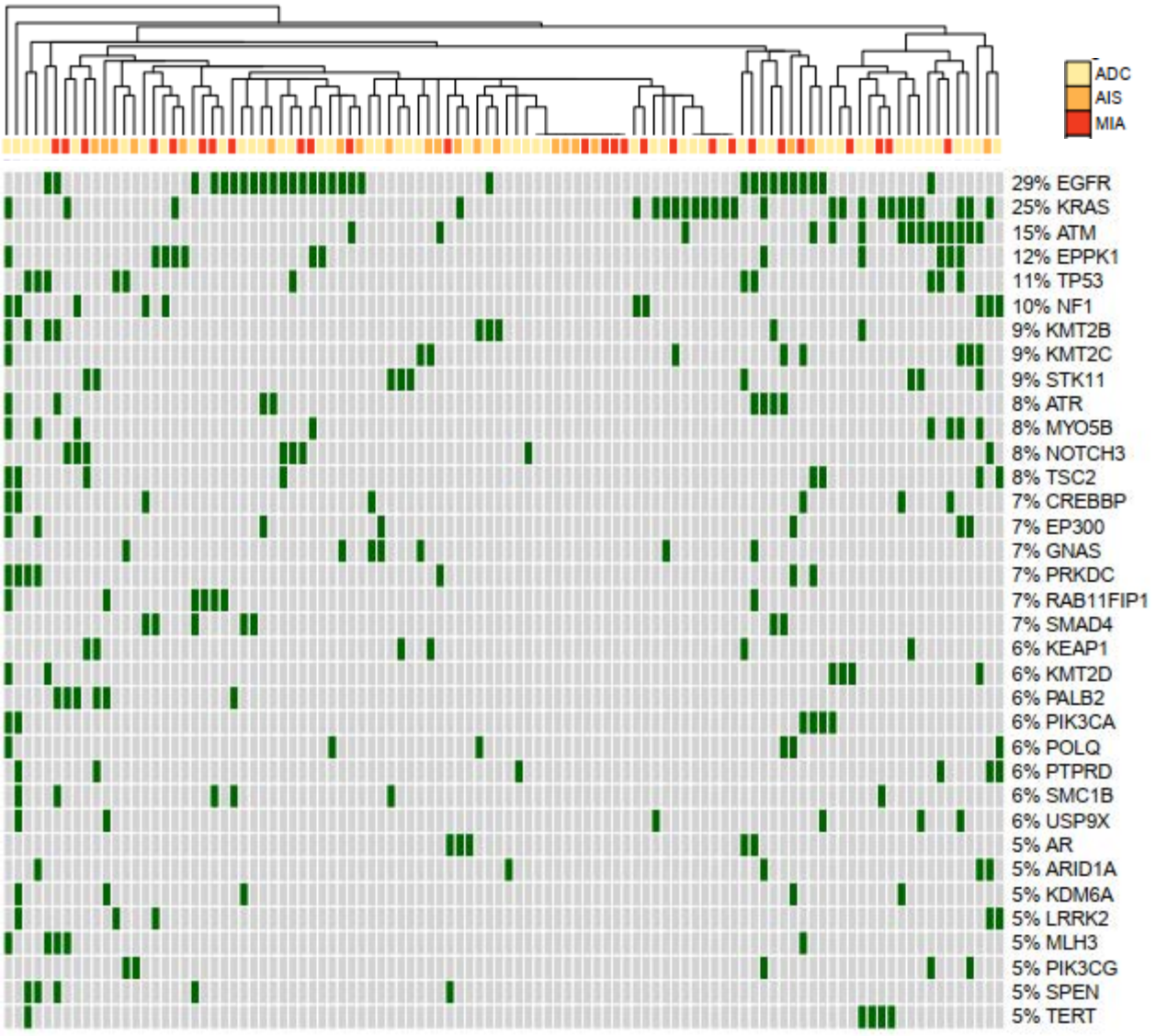


Figure E7

Figure E7. Clonality diversity is not associated with mutation signature