Weaver *et al.*,

Supplementary Information

Supplementary Note

Analysis of mutational context

As bases are paired across strands - A:T and G:C - the 12 possible mutation classes (A>T, A>C, A>G, T>C, T>G, T>A, G>A, G>T, G>C, C>A, C>T, C>G) can be collapsed to 6 classes where the mutated base is always considered as the pyrimidine i.e. A>C is equivalent to T>G. The bases immediately 5' and 3' of the mutated base are known to alter the rate at which differing mutational classes may occur at the base[1,2]. Each mutation class can therefore occur at one of 12 trinucleotide contexts, giving a total of 96 possible class-context combinations.

For each sample in the discovery cohort we extracted the trinucleotide context in which all somatic mutations occurred using the hg19 reference genome (https://www.genome.ucsc.edu/) and a custom perl script. The number of occurrences of each possible mutation-class-context combination was assessed. The prevalence of a specific mutation class at a given trinucleotide is determined both by the mutational processes active in a tumor and by the prevalence of the trinucleotide in the reference genome. To determine the fold enrichment due specifically to mutational processes affecting the tumor, we corrected for the relative prevalence of each trinucleotide within the mappable hg19 reference genome. To assess for novel mutational signatures heat maps were created and visually inspected following the protocol established in Nik-Zainal*et al*[1].

Selection of recurrently mutated target genes

Using SNV calls generated by STRELKA[3] (Illumina), genes were ordered by an estimated probability of frequency of mutation above a baseline, non-silent calling rate. Frequently mutated genes with a p-value $<4x10^{-5}$ (n=26 genes) were selected. We applied stringent filtering criteria to this cohort, removing those genes for which a mutation fell in a poorly mapping region (n=7, Supplementary Table 13) and those classified as uncharacterized (n=1), to enrich for functionally relevant mutation targets. We also removed a further two genes as members of large families we suspected were more likely to be passengers (OR10R2, C10orf71). A further two genes were removed as no mutations were identified in either in the discovery cohort under the adapted SNV filtering criteria (*PCDHGA11*, *HMX2*).

Additional genes were selected for validation based on a known association with carcinogenesis (*ABCB1, SMARCA4, UNC13C, CNTNAP5, MYO18B, MMP16*) and for their relevance to the NFκB pathway, known to be associated with the development of EAC (*TLR1, TLR4, TLR7, TLR9, MYD88, TRAF3, TRAF6*)[4]. In total 27 genes were taken forward to the primer design stage.

Immunohistochemsitry for ARID1A

Immunohistochemistry was performed on tissue microarrays containing tissue cores from 298 EACs. The ARID1A antibody - sigma, HPA005456 - was used at a dilution of 1:200. Staining was performed using a BONDMax autostainer (Leica, Milton Keynes, UK). Cores were scored as 0 (loss of staining), 1 (weak intensity staining), 2 (moderate intensity staining) or 3 (strong intensity staining).

Clonal analysis of 15 recurrently mutated genes in EAC

1) Germline heterozygous sites in non-coding regions were identified in the following manner. A GATK walker identified all sites with Q30 base-quality coverage of the normal sample between 30 and 150 at least 12 reads supporting a variant and a B-allele frequency of at least 0.35.

2) For each chromosomal arm, FREEC counts for 10000 base windows were iteratively segmented (using fastseg) and GC corrected.

3) Those segments were then themselves segmented by the B-allele frequency of sites identified as germline-heterozygous.

4) All segments of length >1000 SNPs are plotted on a depth vs BAF plot and regions sought that will positively identify the coverage/copy number relationship.

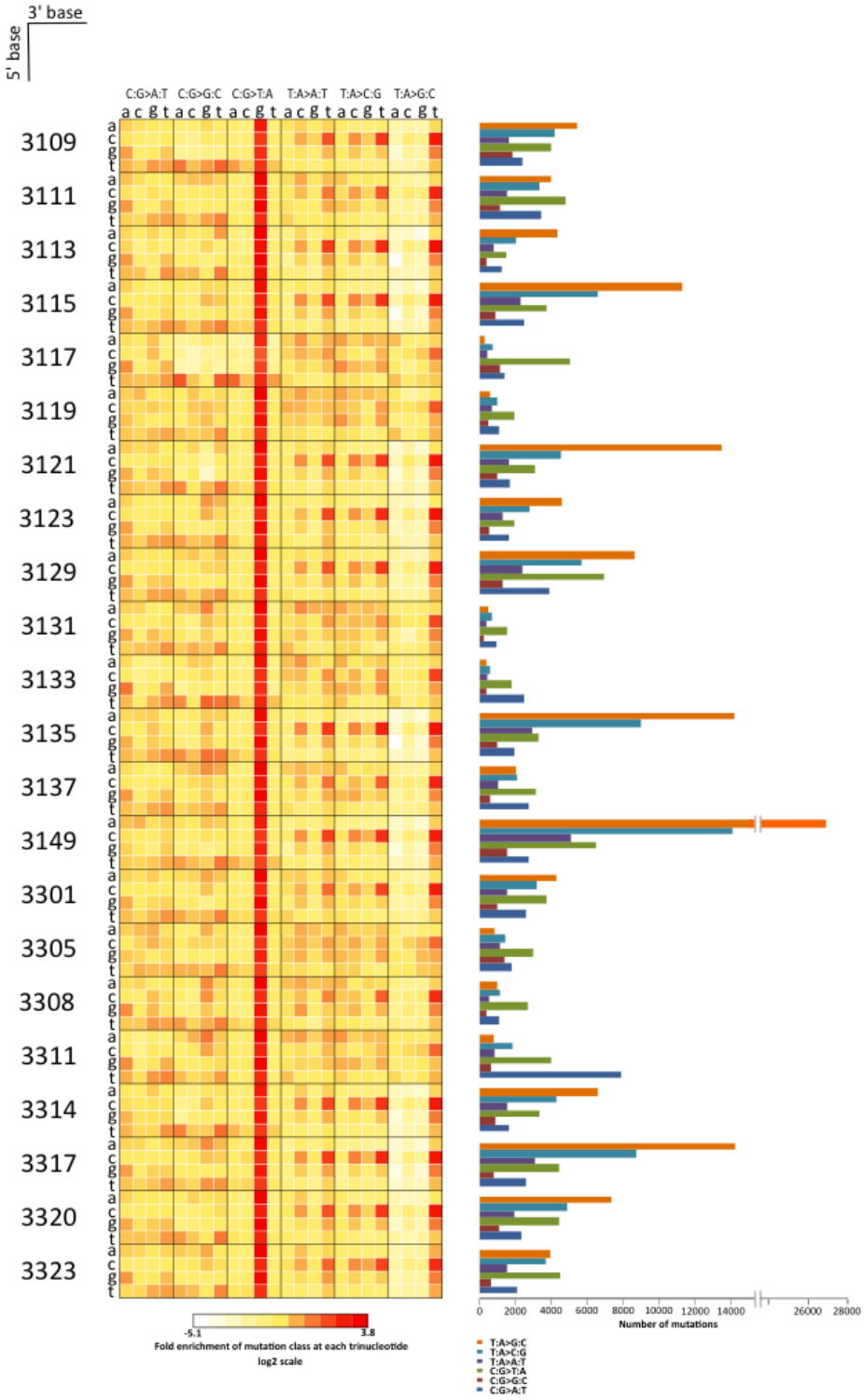
5) For the segment containing a gene of interest, all models consisting of a single copy-number state mixed with normal diploid (AB) tissue are considered. The distribution of depth in the segment, distribution of B-allele frequencies for germline heterozygous SNPs for the segment, and distribution of allele frequency for somatic variants are all considered to determine if the copy-number state is feasible, and if so what proportion of cells carry the copy number mutation (cellularity). If no model provides a good fit, the tumour is presumed to be a mixture of subclonal copy-number mutations. This step identifies whether LOH has taken place at the location of the gene.

6) All major clusters of regions in the depth v BAF plot are considered in this manner to determine the maximum cellularity of any region in the sample: presumed to be the proportion of cells that are malignant. The cellularity of the copy number change of the gene of interest is then compared to judge whether it might be present in all tumour cells. Where this is not clear (as is often the case with lower cellularity tumours, older copy number changes, or changes with no allelic imbalance), the 'benefit of the doubt' is given to the mutation and it is called 'clonal'

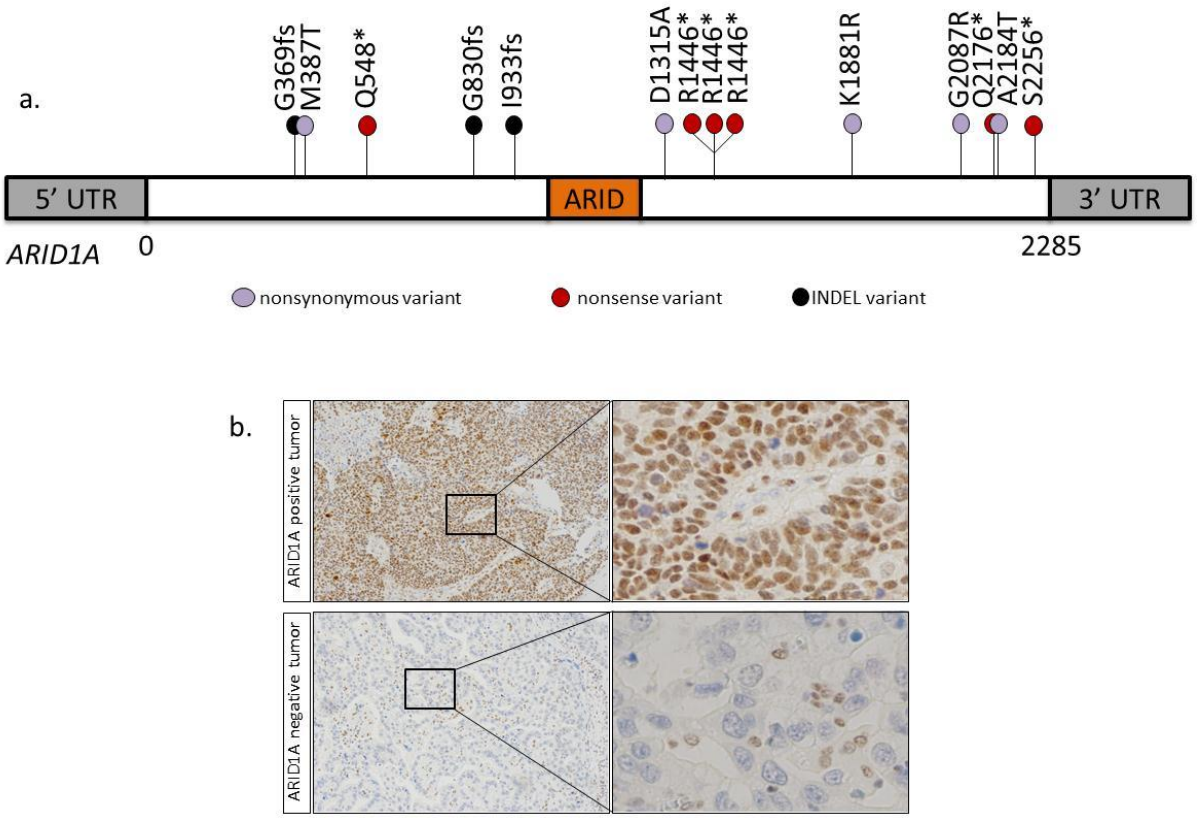
7) Finally the mutation is considered relative to the copy number state and assessed as being clonal (or not)

References to Supplementary Note

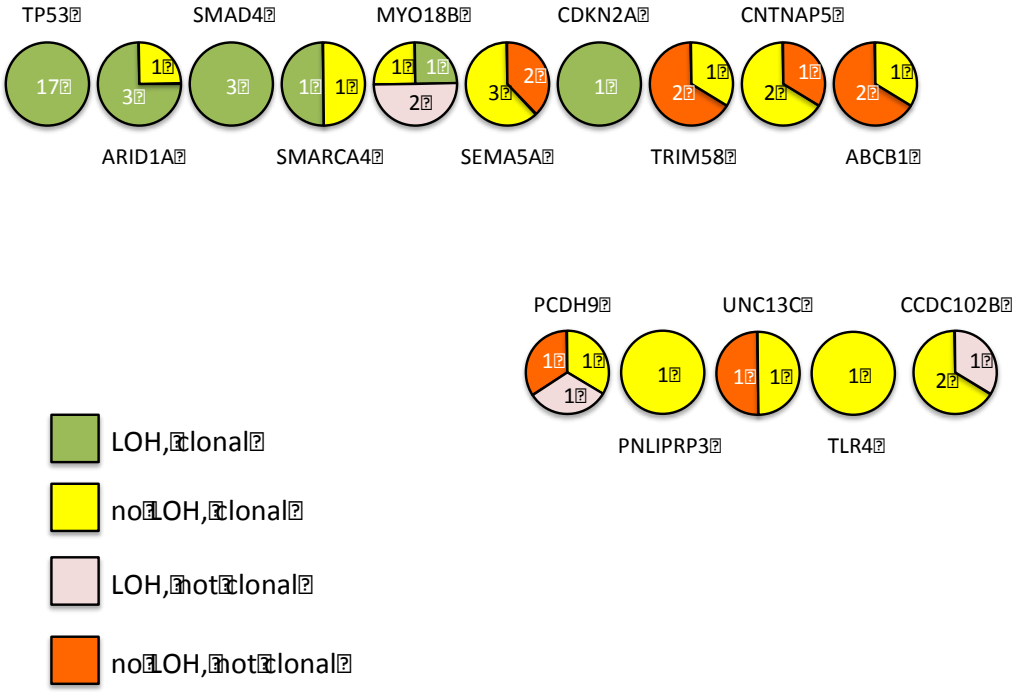
1. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).
2. Dulak, A.M. et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet* **45**, 478-86 (2013).
3. Saunders, C.T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-7 (2012).
4. Abdel-Latif, M.M. et al. NF-kappaB activation in esophageal adenocarcinoma: relationship to Barrett's metaplasia, survival, and response to neoadjuvant chemoradiotherapy. *Ann Surg* **239**, 491-500 (2004).



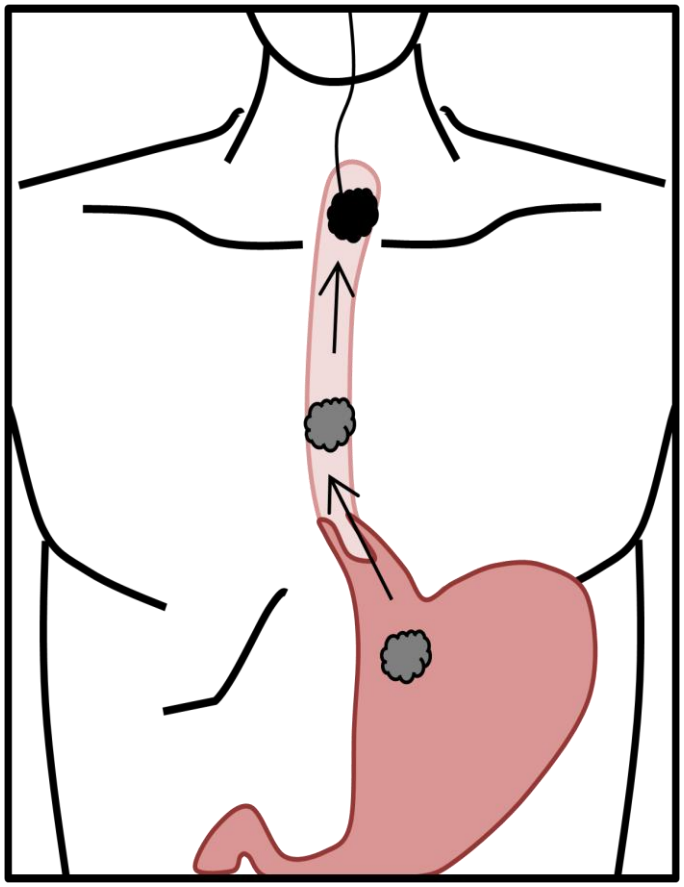
Supplementary Figure 1: Mutational patterns in the 22 discovery cohort EACs. The bar graph represents the total number of each possible mutation class change. The heat map displays the enrichment of each mutation class at any given trinucleotide context. For example the strong red strip running down the centre of the figure represents enrichment for C:G>T:A mutations at the XpCpG trinucleotide. Enrichment at the CTT trinucleotide can be seen for all 3 classes of thymidine mutation in the majority of samples.



Supplementary figure 2: ARID1A DNA mutations and protein expression. a. Schematic showing ARID1A mutations identified in the discovery and validation cohort (112 EAC cases). b. Examples of two EAC tumors, one staining positively for ARID1A protein expression and one staining negatively for ARID1A protein expression.



Supplementary figure 3: Clonal analysis of 15 recurrently-mutated genes in EAC



Supplementary figure 4: Schematic demonstrating Cytosponge™ sampling of cells from the top of the stomach, full length of the esophagus and oropharynx.

Supplementary Tables

Supplementary Table 1: Demographic data of the 22 patients in the discovery cohort

Case ID	Gender	Age (yrs)	Chemo treated	Differentiation	T	N	M	Stage	Alive/ Deceased	Survival (months)	Normal sequenced
3109	Female	80	No	moderate	3	1	X	IIIA	deceased	9*	Blood
3111	Male	70	No	poor	1	0	X	IB	alive	30	Blood
3113	Male	63	No	poor	3	1	X	IIIA	alive	32	Blood
3115	Male	82	No	moderate	2	1	X	IIB	alive	33	Blood
3117	Male	78	No	moderate	3	2	X	IIIB	alive	43	Normal esophagus
3119	Male	76	No	poor	3	2	X	IIIB	deceased	58*	Normal esophagus
3121	Male	70	No	poor	3	1	X	IIIA	deceased	22	Blood
3125	Male	68	Yes	moderate	4	3	X	IIIC	deceased	27	Blood
3129	Male	72	Yes	poor	3	3	X	IIIC	alive	38	Blood
3131	Male	57	Yes	moderate	1	0	X	IB	alive	24	Blood
3133	Male	77	Yes	poor	3	2	1	IV	alive	47	Blood
3135	Male	53	Yes	poor	3	0	X	IIB	deceased	15	Blood
3137	Male	54	Yes	moderate	3	1	0	IIIA	deceased	34*	Blood
3149	Female	63	No	moderate	3	0	X	IIB	deceased	26*	Blood
3302	Male	53	Yes	moderate	1	0	X	IB	alive	27	Normal esophagus
3305	Male	75	Yes	moderate	2	1	X	IIB	alive	5	Normal esophagus
3308	Male	59	Yes	mod/poor	2	0	X	IB	deceased	23*	Normal esophagus
3311	Female	75	Yes	poor	1	1	X	IIB	alive	33	Normal esophagus
3314	Female	76	Yes	poor	3	2	X	IIIB	deceased	31*	Normal esophagus
3317	Male	60	Yes	mod/poor	3	3	X	IIIC	deceased	9*	Normal esophagus
3320	Male	59	Yes	mod/poor	3	2	X	IIIB	deceased	20*	Normal esophagus
3323	Male	65	Yes	poor	2	0	X	IIA	alive	32	Normal esophagus

* Death due to cancer recurrence

Supplementary Table 2: Coverage data for the discovery cohort

Metric	Normal Reference	Tumor
Mapped sequence (Gb)	180 (±15)	170 (±17)
Mean read depth	67 (±6)	63 (±6)
Assembled genome covered (%)	99.9 (±0.03)	99.9 (±0.02)
10x or greater genome covered (%)	99.7 (±0.08)	99.7 (±0.07)
30x or greater genome covered (%)	98.9 (±0.48)	98.3 (±0.96)
50x or greater genome covered (%)	91.5 (±7.21)	79.8 (±10.51)

*Data presented as median (interquartile range)

Supplementary Table 3: Demographics of the 90 EAC patients in the validation cohort

Case ID	Gender	Age (yrs)	Chemo treated	Differentiation	T	N	M	Stage	Alive/ Deceased	Survival (months)
EAC001	Male	33	No	mod/poor	2	0	X	IB	Alive	58
EAC002	Male	67	No	mod/poor	4	3	0	IIIC	Deceased	14*
EAC003	Male	54	No	well/mod	4	3	1	IV	Alive	59
EAC004	Female	72	No	mod/poor	4	1	0	IIIC	Deceased	23*
EAC005	Male	62	No	moderate	3	3	X	IIIC	Deceased	6*
EAC006	Male	57	No	poor	3	1	X	IIIA	Deceased	38*
EAC007	Male	66	No	moderate	3	1	X	IIIA	Deceased	6*
EAC008	Male	77	No	N/A	2	0	X	N/A	Deceased	11
EAC009	Male	73	No	poor	3	0	X	IIB	Deceased	17*
EAC010	Female	62	No	N/A	2	0	0	N/A	Alive	45
EAC011	Male	60	No	mod/poor	3	0	X	IIB	Deceased	24*
EAC012	Male	64	No	poor	4	1	X	IIIC	Deceased	7*
EAC013	Male	56	No	poor	3	1	1	IV	Alive	2
EAC014	Male	74	No	moderate	4	0	0	IIIC	Deceased	12*
EAC015	Male	41	No	moderate	3	3	0	IIIC	Deceased	21*
EAC016	Male	61	No	N/A	X	1	1	IV	Deceased	16*
EAC017	Male	63	Yes	mod/poor	3	0	X	IIB	Deceased	10*
EAC018	Male	66	Yes	mod/poor	3	2	X	IIIB	Deceased	29*
EAC019	Male	54	Yes	poor	3	2	X	IIIB	Deceased	12*
EAC020	Male	77	No	mod/poor	3	2	0	IIIB	Deceased	20*
EAC021	Male	32	Yes	N/A	3	0	X	IIB	Alive	25
EAC022	Male	55	Yes	poor	3	0	X	IIB	Deceased	12*
EAC023	Male	79	No	moderate	2	1	0	IIB	Deceased	2
EAC024	Male	74	No	moderate	3	0	0	IIB	Alive	21
EAC025	Male	58	No	moderate	3	3	0	IIIC	Deceased	18
EAC026	Male	66	Yes	poor	3	2	0	IIIB	Alive	24
EAC027	Male	82	No	well/mod	3	0	0	IIB	Alive	19
EAC028	Male	67	Yes	moderate	3	2	X	IIIB	Alive	48
EAC029	Male	78	Yes	moderate	2	0	0	IB	Alive	20
EAC030	Female	75	No	moderate	1b	0	X	IA	Alive	15
EAC031	Male	81	No	well/mod	3	1	0	IIIA	Alive	20
EAC032	Male	70	Yes	well/mod	1	0	0	IB	Alive	17
EAC033	Male	64	Yes	moderate	1	0	X	IA	Deceased	1*
EAC034	Male	76	No	poor	3	1	0	IIIA	Deceased	11*
EAC035	Male	69	No	poor	X	1	0	N/A	Alive	9
EAC036	Male	67	No	poor	X	2	1	IV	Deceased	5*
EAC037	Male	62	Yes	moderate	3	3	X	IIIC	Alive	23
EAC038	Male	73	Yes	poor	2	0	X	IB	Alive	22
EAC039	Male	66	Yes	moderate	2	0	0	IB	Alive	22
EAC040	Male	62	Yes	poor	3	2	0	IIIB	Alive	20
EAC041	Male	62	Yes	poor	3	1	0	IIIA	Alive	14
EAC042	Male	69	Yes	poor	3	3	0	IIIC	Alive	24
EAC043	Female	67	No	poor	2	0	0	IB	Alive	134
EAC044	Male	74	No	poor	3	2	0	IIIB	Deceased	53
EAC045	Male	N/A	No	poor	N/A	N/A	N/A	N/A	N/A	N/A
EAC046	Female	69	No	poor	2	2	0	IIIA	Deceased	5
EAC047	Male	63	No	poor	3	1	0	IIIA	Deceased	13*
EAC048	Male	52	No	poor	3	2	0	IIIB	Deceased	5*
EAC049	Male	71	No	poor	2	1	0	IIB	Alive	73

EAC050	Male	43	No	poor	3	3	0	IIIC	Deceased	35*
EAC051	Female	N/A	No	poor	N/A	N/A	N/A	N/A	N/A	N/A
EAC052	Female	74	No	poor	3	2	0	IIIB	Deceased	27*
EAC053	Female	N/A	No	moderate	N/A	N/A	N/A	N/A	N/A	N/A
EAC054	Male	56	No	moderate	3	1	0	IIIA	Deceased	6*
EAC055	Male	56	No	moderate	1	0	0	IA	Alive	77
EAC056	Female	56	No	mod/poor	3	3	0	IIIC	Deceased	8*
EAC057	N/A	N/A	N/A	mod/poor	N/A	N/A	N/A	N/A	N/A	N/A
EAC058	Male	50	No	moderate	3	2	0	IIIB	Deceased	1*
EAC059	N/A	N/A	N/A	mod/poor	N/A	N/A	N/A	N/A	N/A	N/A
EAC060	Male	79	No	well	2	1	0	IIB	Deceased	38*
EAC061	Male	81	No	poor	3	3	0	IIIC	Deceased	15
EAC062	Female	44	No	poor	3	0	0	IIB	Deceased	92
EAC063	Male	57	Yes	poor	3	3	0	IIIC	Deceased	2*
EAC064	Male	76	No	poor	3	2	X	IIIB	N/A	N/A
EAC065	Male	75	Yes	poor	3	3	X	IIIC	Deceased	2*
EAC066	Female	69	No	poor	3	2	X	IIIB	Deceased	12*
EAC067	Male	70	Yes	moderate	3	2	0	IIIB	Deceased	38*
EAC068	Male	55	Yes	poor	3	1	X	IIIA	Deceased	21*
EAC069	Male	67	Yes	N/A	2b	0	X	N/A	Alive	50
EAC070	Male	52	Yes	poor	3	3	X	IIIC	Deceased	7*
EAC071	Male	71	Yes	poor	3	0	X	IIB	Deceased	21*
EAC072	Male	53	Yes	moderate	2	1	0	IIB	Deceased	12*
EAC073	Male	73	Yes	moderate	3	1	X	IIIA	Deceased	23
EAC074	Male	70	Yes	poor	3	0	x	IIB	Deceased	13*
EAC075	Male	83	Yes	moderate	3	2	X	IIIB	Deceased	21*
EAC076	Female	66	Yes	moderate	3	1	X	IIIA	Deceased	15*
EAC077	Male	66	Yes	moderate	3	2	X	IIIB	Alive	34
EAC078	Female	70	No	poor	3	2	X	IIIB	Alive	27
EAC079	Male	75	No	well/mod	1	0	X	IA	Alive	27
EAC080	Male	76	Yes	poor	3	1	X	IIIA	Deceased	1
EAC081	Male	63	Yes	moderate	3	1	X	IIIA	Alive	24
EAC082	Male	67	No	poor	1	0	X	IA	Alive	24
EAC083	Male	74	Yes	poor	3	3	X	IIIC	Deceased	12
EAC084	Male	56	No	moderate	1b	0	X	IA	Alive	18
EAC085	Male	75	Yes	moderate	3	1	0	IIIA	Deceased	11*
EAC086	Male	79	Yes	poor	3	3	0	IIIC	Alive	23
EAC087	Male	65	Yes	moderate	1	0	X	IA	Deceased	9*
EAC088	Female	80	No	moderate	3	3	X	IIIC	Alive	14
EAC089	Male	70	No	moderate	2	2	X	IIIA	Deceased	6*
EAC090	Male	75	No	N/A	2	0	0	IB	Alive	4

* Cause of death related to cancer

Supplementary Table 4: Validation using external data. The published mutations of Dulak et al. (supplementary table 6) were interrogated for the fifteen genes that were mutated in four or more samples of our data (see Figure 2). The percentage of samples carrying a mutation in that data set is compared to the percentage carrying a mutation in ours. For each sample in the Dulak *et al.* data set, the allele frequencies of the observed mutations were ranked, and the percentile associated with the genes of interest noted. The median such percentile for all samples carrying that gene is noted (e.g. a percentile of 100 would mean that the mutation in the gene always had the highest allele frequency in a given sample). The table is ranked by this statistic.

	Dulak et al.		Weaver *et al*
Gene	**Allele Frequency Percentile**	**Mutation Percentage**	**Mutation Percentage**
TP53	92.5	73.1	68.8
SMAD4	84.7	9.0	11.6
CCDC102B	83.2	2.1	3.6
CDKN2A	83.1	13.8	8.0
ARID1A	78.2	9.7	11.6
SMARCA4	72.5	6.9	6.3
SEMA5A	67.2	8.3	8.0
TRIM58	62.6	2.8	6.3
MYO18B	57.1	3.4	11.6
TLR4	56.8	6.2	3.6
CNTNAP5	49.7	11.7	6.3
UNC13C	48.3	8.3	4.5
PCDH9	48.1	11.0	6.3
ABCB1	28.7	5.5	6.3
PNLIPRP3	28.4	4.8	4.5

Supplementary Table 5: Point mutations identified in the never-dysplastic Barrett's esophagus and high-grade dysplasia (HGD) samples.

Name	Diagnosis	Gene	Chromosome	Position	Ref	Alt	Mutation type
HGD_01	HGD	TP53	chr17	7574003	G	A	nonsense
HGD_02	HGD	TLR4	chr9	120476906	C	A	missense
HGD_02	HGD	TP53	chr17	7579316	-	A	Frame_shift_INS
HGD_03	HGD	CDKN2A	chr9	21971111	G	A	missense
HGD_03	HGD	TP53	chr17	7578406	C	T	missense
HGD_04	HGD	TP53	chr17	7577551	C	T	missense
HGD_05	HGD	ARID1A	chr1	27105550	C	T	nonsense
HGD_05	HGD	MYO18B	chr22	26157079	T	G	missense
HGD_05	HGD	TP53	chr17	7577547	C	G	missense
HGD_06	HGD	CNTNAP5	chr2	125192118	A	G	missense
HGD_07	HGD	ARID1A	chr1	27106025	G	A	missense
HGD_08	HGD	HMX2	chr10	124908019	C	T	missense
HGD_08	HGD	MYO18B	chr22	26423117	G	A	missense
HGD_08	HGD	TP53	chr17	7577120	C	T	missense
HGD_09	HGD	ARID1A	chr1	27107084	G	A	missense
HGD_09	HGD	CDKN2A	chr9	21974695	-	T	Frame_shift_INS
HGD_10	HGD	TLR4	chr9	120475722	A	G	missense
HGD_10	HGD	TP53	chr17	7577536	T	C	missense
HGD_10	HGD	UNC13C	chr15	54685369	G	T	missense
HGD_11	HGD	TP53	chr17	7578526	C	A	missense
HGD_12	HGD	ARID1A	chr1	27087503	C	T	nonsense
HGD_12	HGD	TP53	chr17	7578205	C	G	missense
HGD_13	HGD	PCDHGA11	chr5	140803181	A	C	missense
HGD_13	HGD	SMARCA4	chr19	11134230	C	T	missense
HGD_13	HGD	TP53	chr17	7577538	C	T	missense
HGD_13	HGD	ARID1A	chr1	27056160	-	A	Frame_shift_INS
HGD_13	HGD	CDKN2A	chr9	21974695	-	T	Frame_shift_INS
HGD_14	HGD	ABCB1	chr7	87183086	T	G	missense
HGD_15	HGD	ARID1A	chr1	27105946	G	T	nonsense
HGD_15	HGD	TLR4	chr9	120475115	T	G	missense
HGD_15	HGD	TP53	chr17	7578479	G	A	missense
HGD_16	HGD	HMX2	chr10	124908015	C	T	missense
HGD_16	HGD	TP53	chr17	7577514	TGA	-	Inframe_DEL
HGD_17	HGD	TP53	chr17	7574003	G	A	nonsense
HGD_18	HGD	TP53	chr17	7577547	C	T	missense
HGD_19	HGD	CNTNAP5	chr2	125521646	T	A	missense
HGD_20	HGD	CNTNAP5	chr2	125530484	A	G	missense
HGD_20	HGD	TP53	chr17	7579373	C	G	missense
HGD_21	HGD	ABCB1	chr7	87196200	A	G	missense
HGD_21	HGD	MYO18B	chr22	26157069	T	C	missense
HGD_21	HGD	MYO18B	chr22	26423567	C	G	missense
HGD_22	HGD	UNC13C	chr15	54614234	A	T	missense
HGD_23	HGD	ARID1A	chr1	27089462	A	C	splice_site
HGD_23	HGD	TP53	chr17	7577121	G	A	missense
HGD_23	HGD	UNC13C	chr15	54919152	G	T	missense
HGD_24	HGD	TLR4	chr9	120476598	G	A	missense
HGD_24	HGD	TP53	chr17	7578406	C	T	missense

HGD_25	HGD	TP53	chr17	7579317	AGTC	-	Frame_shift_DEL
HGD_26	HGD	TP53	chr17	7577120	C	T	missense
HGD_27	HGD	MYO18B	chr22	26247524	C	T	missense
HGD_27	HGD	TP53	chr17	7578263	G	A	nonsense
HGD_28	HGD	ARID1A	chr1	27057685	C	T	nonsense
HGD_28	HGD	CDKN2A	chr9	21971035	T	C	missense
HGD_28	HGD	TLR4	chr9	120475402	G	T	missense
HGD_29	HGD	TP53	chr17	7578280	G	A	missense
HGD_30	HGD	SMARCA4	chr19	11152161	A	T	missense
HGD_30	HGD	TP53	chr17	7578443	A	G	missense
HGD_30	HGD	TP53	chr17	7577099	C	T	missense
HGD_30	HGD	ARID1A	chr1	27056192	G	-	Frame_shift_DEL
HGD_31	HGD	PCDH9	chr13	67801846	T	C	missense
HGD_31	HGD	TP53	chr17	7574018	G	A	missense
HGD_32	HGD	TP53	chr17	7577114	C	A	missense
HGD_33	HGD	TP53	chr17	7574003	G	A	nonsense
HGD_34	HGD	TP53	chr17	7577538	C	T	missense
HGD_35	HGD	CDKN2A	chr9	21971029	C	T	nonsense
HGD_35	HGD	SMARCA4	chr19	11132428	G	A	missense
HGD_35	HGD	TP53	chr17	7578406	C	T	missense
HGD_36	HGD	CNTNAP5	chr2	124979372	T	C	missense
HGD_36	HGD	TP53	chr17	7578403	C	A	missense
HGD_37	HGD	CNTNAP5	chr2	124979372	T	G	missense
HGD_37	HGD	CNTNAP5	chr2	125204408	C	T	missense
HGD_37	HGD	TP53	chr17	7578190	T	C	missense
HGD_37	HGD	TP53	chr17	7574003	G	A	nonsense
HGD_38	HGD	TP53	chr17	7577538	C	T	missense
HGD_39	HGD	CDKN2A	chr9	21971186	G	A	nonsense
HGD_39	HGD	MYF6	chr12	81101952	G	A	missense
HGD_39	HGD	TP53	chr17	7577511	A	C	missense
NDBE_01	Never-dysplastic BE	UNC13C	chr15	54306592	AGC	-	Inframe_Del
NDBE_02	Never-dysplastic BE	CNTNAP5	chr2	125555838	T	G	missense
NDBE_02	Never-dysplastic BE	UNC13C	chr15	54614286	A	C	missense
NDBE_03	Never-dysplastic BE	CNTNAP5	chr2	124999875	A	G	missense
NDBE_04	Never-dysplastic BE	TP53	chr17	7577120	C	T	missense
NDBE_05	Never-dysplastic BE	UNC13C	chr15	54306592	AGC	-	Inframe_Del
NDBE_06	Never-dysplastic BE	ARID1A	chr1	27106373	C	T	missense
NDBE_06	Never-dysplastic BE	ARID1A	chr1	27023239	-	A	Frame_shift_INS
NDBE_07	Never-dysplastic BE	CNTNAP5	chr2	125627319	T	A	missense
NDBE_07	Never-dysplastic BE	MYO18B	chr22	26423082	G	A	missense
NDBE_07	Never-dysplastic BE	ARID1A	chr1	27100943	-	CG	Frame_shift_INS
NDBE_08	Never-dysplastic BE	ABCB1	chr7	87145865	G	T	missense
NDBE_08	Never-dysplastic BE	SMARCA4	chr19	11141499	G	A	missense
NDBE_09	Never-dysplastic BE	SEMA5A	chr5	9197365	G	A	missense
NDBE_10	Never-dysplastic BE	CDKN2A	chr9	21970971	G	T	Nonsense
NDBE_10	Never-dysplastic BE	MYO18B	chr22	26291140	C	T	missense
NDBE_10	Never-dysplastic BE	MYF6	chr12	81101695	-	G	Frame_shift_INS
NDBE_11	Never-dysplastic BE	CDKN2A	chr9	21971186	G	A	Nonsense
NDBE_12	Never-dysplastic BE	MYF6	chr12	81102313	C	T	missense
NDBE_12	Never-dysplastic BE	TRAF3	chr14	103371607	G	A	missense
NDBE_12	Never-dysplastic BE	CDKN2A	chr9	21994319	C	-	Frame_shift_Del
NDBE_13	Never-dysplastic BE	CCDC102B	chr18	66504227	G	A	missense
NDBE_13	Never-dysplastic BE	CDKN2A	chr9	21974676	C	T	missense
NDBE_13	Never-dysplastic BE	SMARCA4	chr19	11134252	G	A	missense

NDBE_14	Never-dysplastic BE	CDKN2A	chr9	21971111	G	A	missense
NDBE_14	Never-dysplastic BE	TRIM58	chr1	248039463	A	G	missense
NDBE_15	Never-dysplastic BE	MYF6	chr12	81101649	G	A	missense
NDBE_15	Never-dysplastic BE	SSTR4	chr20	23016691	C	T	missense
NDBE_15	Never-dysplastic BE	UNC13C	chr15	54529829	C	T	missense
NDBE_16	Never-dysplastic BE	SMARCA4	chr19	11144146	C	T	missense
NDBE_17	Never-dysplastic BE	PNLIPRP3	chr10	118236283	A	C	missense
NDBE_18	Never-dysplastic BE	MMP16	chr8	89068483	C	T	missense
NDBE_18	Never-dysplastic BE	PCDH9	chr13	67801669	G	C	missense
NDBE_19	Never-dysplastic BE	PCDH9	chr13	67801805	T	G	missense
NDBE_20	Never-dysplastic BE	CDKN2A	chr9	21971120	G	A	Nonsense
NDBE_21	Never-dysplastic BE	ARID1A	chr1	27100943	-	CG	Frame_shift_INS

Supplementary Table 6: Number of mutations identified in EAC samples as well as Barrett's esophagus samples with no dysplasia (BE) and high grade dysplasia (HGD). The Fisher's Exact p value is shown (p-value) as well as the Benjamini-Hochberg adjusted p-value (BH_adjusted p-value).

Gene_ID	Tumor_WT	Tumor_mutant	HGD_WT	HGD_mutant	BE_WT	BE_mutant	p-value	BH_adjusted_p-value	significant	TvHGD	TvBE	HGDvBE
TP53	35	77	12	31	65	1	<0.0001	<0.0001	YES	0.8455	<0.0001	<0.0001
SMAD4	99	13	43	0	66	0	0.0005	0.0061	YES	0.0201	0.0022	1
MYO18B	99	13	39	4	64	2	0.1225	0.5473	NO			
ARID1A	99	13	35	8	62	4	0.1428	0.5473	NO			
SEMA5A	103	9	43	0	65	1	0.0406	0.2333	NO			
CDKN2A	103	9	37	6	60	6	0.5156	0.8479	NO			
TRIM58	105	7	43	0	65	1	0.1761	0.5787	NO			
CNTNAP5	105	7	38	5	63	3	0.3566	0.7456	NO			
ABCB1	105	7	41	2	65	1	0.4241	0.8129	NO			
PCDH9	105	7	42	1	64	2	0.5161	0.8479	NO			
SMARCA4	105	7	40	3	63	3	0.8666	1	NO			
PNLIPRP3	107	5	43	0	65	1	0.3365	0.7456	NO			
UNC13C	107	5	40	3	62	4	0.7893	1	NO			
TLR4	108	4	38	5	66	0	0.0122	0.0932	NO			
CCDC102B	108	4	43	0	65	1	0.5928	0.9090	NO			
FGF10	109	3	43	0	66	0	0.3037	0.7456	NO			
TRAF6	109	3	43	0	66	0	0.3037	0.7456	NO			
MYF6	109	3	42	1	63	3	0.8766	1	NO			
MMP16	110	2	43	0	65	1	1	1	NO			
SSTR4	110	2	43	0	65	1	1	1	NO			
MYD88	111	1	43	0	66	0	1	1	NO			
CCDC153	111	1	43	0	66	0	1	1	NO			
TRAF3	111	1	43	0	65	1	1	1	NO			

Supplementary table 7: Expanded cohort for MYO18B and ARID1A. To better characterize the percentage of pre malignant lesions harboring mutations we resequenced the top two non-significantly mutated genes in our EAC cohort in a further 25 NDBE samples and 11 HGD samples giving a total of 91 NDBE and 54 HGD samples. No significant difference in mutation frequencies was observed. The Fisher's exact test p-value is shown (p-value).

Gene_ID	Tumour_WT	Tumour_mutant	HGD_WT	HGD_mutant	BE_WT	BE_mutant	p value
MYO18B	99	13	49	5	88	3	0.0698
ARID1A	99	13	43	11	82	9	0.1808

Supplementary table 8: Table showing all 22 high grade BE patients for which TP53 sequencing was performed on their Cytosponge samples. For each patient that a TP53 mutation was identified, the exact base change as well as the exact location is noted, together with the allele fraction for each replicate (Freq1 and Freq2), the depth of sequencing obtained for the base in question, the expected variant allele fraction for the two replicates as well as the observed allele fractions. NS = nonsynonymous, HGD = High grade dysplasia.

	Diagnosis	Exonic	Type	Chr	Position	Ref	Mut	Freq1	Freq2	Ave Freq	medianFreq	Depth1	Depth2	Expected1	Expected2	Observed1	Observed2
Cyto_HGD_1	HGD	exonic	NS	17	7578403	C	A	0.113	0.077	0.095	0	373	416	0.043882	0.048941	42	32
Cyto_HGD_2	HGD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cyto_HGD_3	HGD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cyto_HGD_4	HGD	exonic	NS	17	7577121	G	A	0.167	0.155	0.161	0	935	865	4.3287	4.0046	156	134
Cyto_HGD_5	HGD	exonic	INDEL, FS	17	7578458												
Cyto_HGD_6	HGD	exonic	stopgain	17	7578524	G	A	0.018	0.009	0.0135	0	9792	2533	1.3663	0.35344	176	23
Cyto_HGD_7	HGD	exonic	INDEL, FS	17	7573993												
Cyto_HGD_8	HGD	exonic	NS	17	7577580	T	A	0.019	0.036	0.0275	0	946	220	1.247	0.29	18	8
Cyto_HGD_9	HGD	exonic	NS	17	7577120	C	T	0.175	0.178	0.1765	0	1270	1579	1.7639	2.1931	222	281
Cyto_HGD_10	HGD	exonic	NS	17	7577539	G	A	0.038	0.022	0.03	0	1804	504	0.74282	0.20753	69	11
Cyto_HGD_11	HGD	exonic	stopgain	17	7574003	G	A	0.156	0.166	0.161	0	1218	1104	0.27405	0.2484	190	183
Cyto_HGD_12	HGD	exonic	NS	17	7578265	A	G	0.153	0.152	0.1525	0	432	618	0	0	66	94
Cyto_HGD_13	HGD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cyto_HGD_14	HGD	exonic	NS	17	7577094	G	A	0.017	0.015	0.016	0	2940	1831	2.8224	1.7578	50	27
Cyto_HGD_15	HGD	exonic	stopgain	17	7578263	G	A	0.01	0.012	0.011	0	3632	1035	2.1291	0.60672	36	12
Cyto_HGD_16	HGD	exonic	NS	17	7577551	C	T	0.322	0.391	0.3565	0	90	23	0	0	29	9
Cyto_HGD_17	HGD	exonic	NS	17	7578190	T	G	0.247	0.246	0.2465	0	1154	878	0	0	285	216
Cyto_HGD_18	HGD	exonic	NS	17	7577046	C	T	0.006	0.005	0.0055	0	2979	4171	1.3983	1.9578	18	21
Cyto_HGD_19	HGD	exonic	NS	17	7577085	C	T	0.205	0.388	0.2965	0	39	49	0.020893	0.02625	8	19
Cyto_HGD_20	HGD	exonic	NS	17	7578272	G	A	0.026	0.051	0.0385	0	1676	1403	0.14366	0.12026	44	72
Cyto_HGD_21	HGD	exonic	NS	17	7577121	G	A	0.047	0.068	0.0575	0	236	370	0.25567	0.40083	11	25
Cyto_HGD_22	HGD	exonic	NS	17	7577556	C	T	0.021	0.019	0.02	0	977	2120	0.11273	0.24462	21	40

Supplementary Table 9: Sequencing metrics for the discovery cohort. For each of the 44 samples sent for sequencing, we report the number of read pairs produced, their alignment rate with BWA and the base error rate reported by the aligner. IQR: interquartile range (from 0.025 quantile, i) 97.5 quantile. The duplication rate, optical duplicate rate and the estimated library size were obtained from PICARD.Whether our reads look like a random sample from the genome we can measure via relative entropy. A perfectly random sample with no biases or structure would give a value of zero. Since structure is more common at the start of a read, we report this measure for the first 5 bases of the read (start), and bases 31-35 (mid). We define the mappable genome as that excluding the gaps in assembly detailed in the BioConductor package Bsgenome.Hsapiens.UCSC.hg19_1.3.17. Our coverage statistics exclude the sex chromosomes, and we give the mean depth in these regions, the proportion of these regions that are covered to at least 1x, 10x, 30x and 50x. There is naturally a particular interest in mutations within the protein coding regions, and these statistics are reported solely for such regions (as defined from Ensembl).

Case ID	Tumor / Normal	Lanes	Read pairs ($x10^9$)	Rate (%)		Insert size median (IQR	Duplication rate (%)		Library size ($x10^9$)	Relative Entropy		% Q30 mapped bases	Assembled regions					Protein coding					
				Alignment	Error		Pair	Optical		Start	Mid		Mapped sequence (Gb)	Mean read depth (Fold coverage)	1x	10x	30x	50x	Mapped sequence (Gb)	Mean read depth (fold coverage)	1x	10x	30x	50x
3108	N	6	1.24	93.08	0.84	304 (234-384)	6.15	0.01	9.02	0.00551	0.00541	86.89	204.8	77.1	99.9	99.7	98.9	93.9	2.83	84.3	99.7	99.3	98.5	95.4
3109	T	5	1.02	95.91	0.51	296 (211-381)	11.17	0.02	4.07	0.00334	0.00208	90.86	166.5	62.1	99.9	99.7	98.6	83.6	2.00	59.5	99.7	99.3	97.3	80.3
3110	N	6	1.25	92.57	0.85	309 (242-382)	7.86	0.01	6.99	0.00661	0.00622	86.79	206.4	77.4	99.9	99.7	99.0	95.6	2.72	81.1	99.6	99.1	98.2	95.3
3111	T	5	1.03	95.85	0.51	295 (215-380)	10.6	0.02	4.33	0.00425	0.00285	90.45	170.0	63.0	99.9	99.7	98.5	84.3	2.00	59.0	99.7	99.2	97.0	78.1
3112	N	7	1.37	93.79	0.71	306 (231-391)	6.94	0.01	8.85	0.00823	0.0081	87.56	230.9	86.8	99.9	99.7	99.1	96.	3.13	93.5	99.5	99.0	98.3	96.3
3113	T	5	0.94	94.34	0.48	283 (208-358)	5.99	0.02	7.11	0.00646	0.00609	91.2	160.3	60.5	99.9	99.6	98.2	78.4	2.14	63.8	99.6	99.1	97.5	85.7
3114	N	5	0.94	93.79	0.82	315 (235-402)	5.05	0.01	8.45	0.00859	0.00876	86.65	160.1	60.5	99.9	99.6	97.3	71.8	2.34	69.6	99.7	99.3	97.8	84.1
3115	T	5	0.98	95.04	0.55	293 (213-383)	13.41	0.01	3.17	0.0054	0.00299	90.2	157.0	58.1	99.9	99.7	97.0	67.6	1.93	57.5	99.8	99.3	96.1	66.1
3116	N	5	0.99	94.16	0.79	301 (231-381)	11.92	0.01	3.59	0.00581	0.00385	88.32	156.3	58.3	99.9	99.7	98.5	81.2	1.89	56.3	99.6	99.0	96.8	77.7
3117	T	5	0.97	94.61	0.57	308 (223-405)	6.99	0.02	6.30	0.00516	0.00424	90.32	164.7	62.6	99.9	99.7	98.3	80.4	2.13	64.4	99.7	99.1	97.4	85.2
3118	N	6	1.24	92.96	0.9	282 (212-357)	14.49	0.01	3.61	0.00561	0.00352	85.96	191.3	71.2	99.9	99.7	99.1	96.1	2.27	67.7	99.7	99.3	98.0	93.2
3119	T	5	0.99	95.22	0.5	307 (216-411)	3.98	0.02	11.52	0.00597	0.00624	90.34	173.3	65.0	99.8	99.6	96.4	70.0	2.53	74.8	99.3	98.6	96.3	80.9
3120	N	5	1.18	94.35	0.53	325 (230-435)	4.01	0.01	13.53	0.01070	0.00985	89.02	205.1	77.7	99.9	99.7	98.8	92.6	3.06	91.2	99.5	99.0	98.2	94.8
3121	T	5	1.16	94.21	0.62	330 (250-410)	4.73	0.02	11.22	0.01120	0.00926	88.37	200.2	75.4	99.9	99.6	98.6	89.1	2.97	88.0	99.5	98.9	98.0	93.2
3124	N	5	1.04	94.87	0.55	304 (235-379)	8.07	0.01	5.80	0.00641	0.00549	89.43	174.9	65.6	99.9	99.7	98.5	88.4	2.29	68.2	99.8	99.3	98.1	91.4
3125	T	5	1.00	95.1	0.51	305 (225-395)	6.81	0.03	6.69	0.00516	0.00471	90.92	170.4	64.0	99.9	99.7	98.5	84.9	2.25	67.2	99.7	99.2	97.9	89.5
3128	N	5	1.02	94.47	0.8	315 (245-390)	4.8	0.01	9.78	0.01080	0.01030	82.22	174.5	65.8	99.9	99.6	98.2	83.45	2.49	74.1	99.6	99.3	98.0	90.0
3129	T	5	1.03	95.21	0.55	277 (197-371)	8.67	0.02	5.32	0.00421	0.00298	90.14	173.6	61.5	99.9	99.7	98.2	79.05	2.07	59.6	99.6	99.0	96.6	76.5
3130	N	5	1.08	95.14	0.53	281 (206-364)	8.34	0.01	5.83	0.00569	0.00452	89.87	181.1	67.7	99.9	99.7	98.8	92.9	2.24	66.7	99.7	99.2	97.7	92.3
3131	T	5	0.98	94.93	0.49	314 (238-404)	4.94	0.03	9.15	0.00758	0.00773	91.93	170.3	64.3	99.9	99.6	97.6	76.0	2.45	73.6	99.7	99.2	97.7	86.3
3132	N	5	1.07	94.82	0.54	308 (228-398)	6.03	0.02	8.13	0.00647	0.00619	89.57	184.1	69.3	99.9	99.7	98.7	89.8	2.49	74.4	99.6	99.2	98.2	92.8
3133	T	5	1.08	94.97	0.5	297 (232-372)	7.48	0.02	6.53	0.00912	0.00416	91.16	182.9	68.51	99.9	99.7	98.8	91.2	2.32	69.4	99.8	99.3	98.0	92.2
3134	N	5	1.16	94.55	0.6	321 (241-415)	4.38	0.02	12.18	0.00951	0.00964	88.15	201.4	76.0	99.9	99.7	98.7	91.9	2.95	88.9	99.7	99.4	98.6	95.0
3135	T	5	1.12	94.84	0.57	326 (246-410)	6.42	0.02	7.91	0.00816	0.00721	89.45	190.1	71.1	99.9	99.6	98.1	85.5	2.63	77.	99.6	99.0	97.5	89.2
3136	N	5	1.05	94.67	0.63	308 (218-393)	5.76	0.01	8.29	0.00647	0.00597	86.84	179.7	67.2	99.9	99.7	98.5	90.5	2.29	68.4	99.8	99.4	98.1	92.1
3137	T	5	0.98	94.27	0.5	311 (201-433)	4.88	0.02	9.15	0.01050	0.00866	91.0	168.9	63.2	99.9	99.6	97.4	73.7	2.46	72.9	99.7	99.2	97.9	86.1
3148	N	5	1.09	94.49	0.68	328 (233-446)	3.73	0.02	13.49	0.00901	0.00857	87.31	186.2	70.3	99.9	99.7	98.3	86.0	2.74	81.7	99.7	99.3	98.2	91.6
3149	T	5	1.09	93.88	0.61	309 (214-409)	4.75	0.02	10.45	0.01170	0.01260	88.35	184.9	70.5	99.8	99.0	92.9	75.0	2.78	83.0	99.7	99.1	96.6	86.9
3301	N	5	0.92	96.23	0.45	351 (221-506)	5.92	0.03	7.25	0.00545	0.00296	93.34	160.5	59.6	99.9	99.7	98.8	83.8	1.89	56.4	99.9	99.6	97.6	76.2

3302	T	5	1.05	96.15	0.42	340 (190-530)	5.97	0.02	8.18	0.00708	0.00247	92.75	183.7	68.7	99.9	99.7	98.5	79.0	2.19	66.2	99.7	98.9	96.1	73.8
3304	N	5	1.06	95.6	0.45	327 (142-542)	8.37	0.02	5.73	0.00565	0.00411	92.83	178.1	66.1	99.9	99.8	99.2	93.5	2.06	61.5	99.9	99.6	98.0	86.3
3305	T	5	1.01	95.63	0.54	323 (178-495)	7.35	0.01	6.24	0.00693	0.00347	89.81	170.8	62.8	99.9	99.8	98.9	83.2	2.13	62.3	99.9	99.7	98.0	80.7
3307	N	5	1.08	96.16	0.49	332 (202-477)	6.85	0.02	7.26	0.00604	0.00332	91.42	186.1	69.1	99.9	99.7	99.0	94.7	2.19	65.4	99.9	99.5	98.1	90.8
3308	T	5	1.03	95.96	0.41	347 (182-572)	9.96	0.04	4.63	0.00943	0.00284	93.68	170.2	63.7	99.93	99.7	98.9	85.1	2.03	60.5	99.6	98.9	96.3	79.7
3310	N	6	1.08	95.54	0.55	364 (261-481)	10.71	0.01	4.46	0.00796	0.00347	88.61	172.8	64.2	99.9	99.8	99.0	90.9	2.04	61.0	99.9	99.7	98.3	85.47
3311	T	5	0.91	95.71	0.5	351 (256-493)	6.99	0.02	5.92	0.00471	0.00305	90.17	151.0	56.5	99.9	99.7	97.6	64.0	1.85	55.7	99.9	99.7	96.8	60.5
3313	N	5	1.04	96.49	0.4	342 (212-492)	5.22	0.02	9.29	0.00651	0.00295	93.33	178.5	66.3	99.9	99.8	99.1	93.2	2.08	62.2	99.9	99.5	97.9	87.64
3314	T	5	1.03	96.49	0.44	353 (183-546)	5.21	0.03	9.25	0.00404	0.00246	93.06	180.9	67.5	99.9	99.7	97.0	81.8	2.33	68.	99.9	99.8	97.6	85.1
3316	N	5	1.01	95.88	0.39	331 (150-550)	5.99	0.05	7.82	0.01160	0.00498	94.16	174.7	64.7	99.9	99.7	98.9	89.6	1.92	57.4	99.7	99.0	95.8	73.8
3317	T	5	1.08	95.69	0.46	338 (233-458)	8.37	0.02	5.86	0.00618	0.00368	91.44	183.6	67.7	99.9	99.8	98.9	84.6	2.14	65.0	99.9	99.6	97.8	78.8
3319	N	5	1.09	95.72	0.48	362 (212-522)	8.24	0.02	6.00	0.01320	0.00515	91.86	184.3	68.5	99.9	99.8	99.1	94.7	2.22	66.3	99.9	99.6	98.4	92.03
3320	T	5	0.95	95.71	0.49	350 (203-520)	8.72	0.02	4.93	0.00609	0.00358	90.34	160.6	59.1	99.9	99.7	98.2	73.2	1.96	57.8	99.9	99.6	97.2	71.6
3322	N	5	0.97	95.88	0.42	355 (200-530)	11.34	0.03	3.79	0.00679	0.00367	93.49	158.2	58.8	99.4	99.8	98.8	81.6	1.85	55.4	99.9	99.5	97.4	73.12
3323	T	5	0.92	96.27	0.41	350 (220-495)	5.79	0.04	7.41	0.00596	0.00297	94.13	160.1	59.3	99.9	99.7	98.3	72.6	1.89	56.4	99.9	99.5	96.6	65.9

Supplementary Table 10: Tiles removed post-QC.

Case ID	Tumor/Normal	Flow cell	Lane	Tiles removed
3108	N	HS2000-920_87	2	1103
3110	N	HS2000-920_87	4	2208, 2308
3118	N	HS2000-920_87	5	1108
3118	N	HS2000-920_87	6	1208, 2108

Supplementary Table 11: Lanes with reads trimmed post-QC

Case ID	Flow cell	Lane	Read	Bases trimmed
3114	HS2000-1010_46	3	2	85 - 100
3116	HS2000-645_109	7	2	70 - 100
3128	HS2000-793_102	4	1	80 - 100

Supplementary Table 12: Per-sample coverage statistics for the callable genome. Our rules for calling somatic single nucleotide variants require a minimum of 10-fold coverage in both the tumor and germline sequencing. For each pair of samples from the discovery set, for the assembled regions in the auto chromosomal chromosomes, we report the number of bases covered to this depth in the tumor sample, the normal sample, and in both. This is also given as a percentage of the assembled region for the three cases. The percentages that would be expected if the variations in coverage of the two samples were independent are reported in the final column.

Normal ID	Tumor ID	≥10x in normal	≥10x in tumor	≥10x in both	Mappable genome size	percentage in normal	percentage in tumor	percentage in both	Percentage expected if independent
3108	3109	2677470246	2677568854	2676219799	2684578480	99.74	99.74	99.69	99.47
3110	3111	2677293751	2676993577	2675675217	2684578480	99.73	99.72	99.67	99.45
3112	3113	2677694607	2676175146	2675513115	2684578480	99.74	99.69	99.66	99.43
3114	3115	2675911765	2676895348	2674218323	2684578480	99.68	99.71	99.61	99.39
3116	3117	2676751763	2676394568	2675205966	2684578480	99.71	99.7	99.65	99.4
3118	3119	2678599326	2673847404	2673266814	2684578480	99.78	99.6	99.58	99.38
3120	3121	2677092193	2676360689	2675551353	2684578480	99.72	99.69	99.66	99.42
3124	3125	2676749324	2676939396	2675590851	2684578480	99.71	99.72	99.67	99.42
3128	3129	2676382727	2676682610	2674622478	2684578480	99.69	99.71	99.63	99.4
3130	3131	2677099380	2675246020	2674305225	2684578480	99.72	99.65	99.62	99.37
3132	3133	2677166211	2677772556	2676296861	2684578480	99.72	99.75	99.69	99.47
3134	3135	2677146070	2675670771	2674765029	2684578480	99.72	99.67	99.63	99.39
3136	3137	2676470255	2675737311	2674518298	2684578480	99.7	99.67	99.63	99.37
3148	3149	2676701946	2658479332	2657761547	2684578480	99.71	99.03	99	98.74
3301	3302	2678967522	2677701944	2676509908	2684578480	99.79	99.74	99.7	99.54
3304	3305	2680475788	2679564294	2678958903	2684578480	99.85	99.81	99.79	99.66
3307	3308	2678621469	2678544268	2676799164	2684578480	99.78	99.78	99.71	99.55
3310	3311	2679698213	2677434093	2676908392	2684578480	99.82	99.73	99.71	99.55
3313	3314	2679738979	2678361776	2677334630	2684578480	99.82	99.77	99.73	99.59
3316	3317	2678231813	2680001162	2677519640	2684578480	99.76	99.83	99.74	99.59
3319	3320	2679435093	2678990848	2678012748	2684578480	99.81	99.79	99.76	99.6
3322	3323	2679436003	2678101117	2677403328	2684578480	99.81	99.76	99.73	99.57

Supplementary Table 13: List of frequently mutated genes excluded from the validation

Gene	Sample count	Coding length	p-value	Exclusion criteria
KRTAP4-7	3	465	4.43E-08	Mutations in poorly mapping regions
CTD-2144E22.5	3	519	6.86E-08	Uncharacterised Protein
FRG1	3	814	4.08E-07	Mutations in poorly mapping regions
OR10R2	3	1005	9.38E-07	Large gene family
MAGEC2	3	1135	1.52E-06	Mutations in poorly mapping regions
DIRC3	2	413	4.16E-06	Failed primer design
CST2	2	431	4.72E-06	Mutations in poorly mapping regions
ZNF730	3	1529	4.88E-06	Mutation in poorly mapping region
IVL	3	1763	8.52E-06	Mutations in poorly mapping regions
REG3A	2	553	9.90E-06	Mutation in poorly mapping region
C10orf71	4	4406	1.54E-05	Large gene family

Supplementary Table 14: Coverage (≥100 fold) achieved for the 26genes screened by amplicon re-sequencing for mutations in the validation cohort.

Gene Name	Covered Bases	Targeted Region	Proportion of the gene covered (≥100 fold)
ABCB1	3909	4023	0.97
ARID1A	5728	6935	0.83
CCDC102B	1775	1849	0.96
CCDC153	658	658	1
CDKN2A	709	860	0.82
CNTNAP5	4010	4010	1
FGF10	632	640	0.99
MMP16	1782	1930	0.92
MYD88	967	967	1
MYF6	734	734	1
MYO18B	7659	7884	0.97
PCDH9	3705	3789	0.98
PNLIPRP3	1444	1445	1
SEMA5A	3146	3436	0.92
SMAD4	1693	1803	0.94
SMARCA4	4871	5406	0.90
SSTR4	1047	1164	0.90
TLR1	288	2398	0.12
TLR4	2423	2533	0.96
TLR7	744	3159	0.24
TLR9	949	3100	0.31
TP53	1254	1430	0.88
TRAF3	1650	1772	0.93
TRAF6	1594	1602	1
TRIM58	1303	1478	0.88
UNC13C	6485	6845	0.95

Supplementary Table 15: p53 primers used in the p53 multiplex assay. All forward primers contained the CS1 sequence (5′- ACACTGACGACATGGTTCTACA – 3′) and all the reverse primers contained the CS2 sequence (5′- TACGGTAGCAGAGACTTGGTCT – 3′) in order to allow the addition of a unique barcode as well as the Illumina adapter sequence in the second PCR.

Target Specific – Forward	Target Specific - Reverse
ACACTGACGACATGGTTCTACAGACCCAAAACCCAAAATGGC	TACGGTACGAGAGACTTGGTCTTCCCTGCTTCTGTCTCCTAC
ACACTGACGACATGGTTCTACACTGGTGTTGTTGGGCAGT	TACGGTACGAGAGACTTGGTCTATCTCCGCAAGAAAGGGGAG
ACACTGACGACATGGTTCTACATCCAAATACTCCACACGCAAA	TACGGTACGAGAGACTTGGTCTGCTGCCCCCACCATGAG
ACACTGACGACATGGTTCTACATGTGCTGTGACTGCTTGTAG	TACGGTACGAGAGACTTGGTCTTGCCCTGACTTTCAACTCTGT
ACACTGACGACATGGTTCTACAGGAAACCGTAGCTGCCCTG	TACGGTACGAGAGACTTGGTCTAAGACCCAGGTCCAGATGAA
ACACTGACGACATGGTTCTACAGGAATCCTATGGCTTTCCAACC	TACGGTACGAGAGACTTGGTCTCCCCCTCCTCTGTTGCTG
ACACTGACGACATGGTTCTACATCTGTATCAGGCAAAGTCATAGAA	TACGGTACGAGAGACTTGGTCTGCCTCAAAGACAATGGCTCC
ACACTGACGACATGGTTCTACAAGAAAACGGCATTTTGAGTGT	TACGGTACGAGAGACTTGGTCTAAGGGTGCAGTTATGCCTCA
ACACTGACGACATGGTTCTACATGTCCTGCTTGCTTACCTCG	TACGGTACGAGAGACTTGGTCTGCCTCTTGCTTCTCTTTTCCT
ACACTGACGACATGGTTCTACAGGGGTCAGAGGCAAGCAG	TACGGTACGAGAGACTTGGTCTCTTGGGCCTGTGTTATCTCC
ACACTGACGACATGGTTCTACAGAGAAAGCCCCCCTACTGC	TACGGTACGAGAGACTTGGTCTAGCATCTTATCCGAGTGGAAGG
ACACTGACGACATGGTTCTACAAGCTGCTCACCATCGCTA	TACGGTACGAGAGACTTGGTCTCCAACTGGCCAAGACCT
ACACTGACGACATGGTTCTACAATACGGCCAGGCATTGAAGT	TACGGTACGAGAGACTTGGTCTCCTCCTGGCCCCTGTC
ACACTGACGACATGGTTCTACACAGCCTCTGGCATTCTGG	TACGGTACGAGAGACTTGGTCTCCTGGTCCTCTGACTGCTCT

Supplementary Table 16: Genomic co-ordinates (hg19) that each of the p53 primers used in the p53 multiplex assay amplify

Gene	chr	amp_start	amp_end	Called	Pool
TP53	chr17	7572850	7573030	TP53_1	Pool 1
TP53	chr17	7576908	7577075	TP53_5	Pool 1
TP53	chr17	7578229	7578406	TP53_9	Pool 1
TP53	chr17	7578425	7578594	TP53_11	Pool 1
TP53	chr17	7579359	7579520	TP53_13	Pool 1
TP53	chr17	7573859	7574054	TP53_2	Pool 2
TP53	chr17	7576584	7576734	TP53_3	Pool 2
TP53	chr17	7576786	7576983	TP53_4	Pool 2
TP53	chr17	7577003	7577187	TP53_6	Pool 2
TP53	chr17	7577432	7577631	TP53_7	Pool 2
TP53	chr17	7578091	7578274	TP53_8	Pool 2
TP53	chr17	7578361	7578525	TP53_10	Pool 2
TP53	chr17	7579260	7579421	TP53_12	Pool 2
TP53	chr17	7579479	7579626	TP53_14	Pool 2

Supplementary Table 17: Sequences of the Fluidigm barcode primers containing PE1 in the forward primer and PE2 and the unique barcode (BC) in the reverse primer.

Primer	Sequence
PE1-CS1	5'-AATGATACGGCGACCACCGAGATCTACACTGACGACATGGTTCTACA-3'
PE2-BC-CS2	5'-CAAGCAGAAGACGGCATACGAGAT-[BC]-TACGGTAGCAGAGACTTGGTCT-3'