

Supplemental materials

Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers

Kwiyeom Yoon¹, Sunghoon Lee¹, Tae-Su Han¹, So Yeon Moon, Sun Mi Yun, Seong-Ho Kong, Sungwoong Jho, Jinny Choe, Jieun Yu, Hyuk-Joon Lee, Ji Hyun Park, Hak-Min Kim, So Yeun Lee, Jongsun Park, Woo-Ho Kim, Jong Bhak, Han-Kwang Yang, Seong-Jin Kim²

²Correspondence should be addressed to S-J.K. (kimsj@cha.ac.kr)

¹These authors equally contributed to this work.

This PDF file includes

Supplemental Figures S1 to S5

Supplemental Tables S1 to S9

Supplemental Tables S2, S5 and S7 are attached individually in Excel.

Supplemental Note

Other Supplemental materials for this manuscript include the following:

Supplemental Table S2. The list of length alterations with all deletions and insertions in gene regions from whole-genome sequencing and RNA sequencing.

Supplemental Table S5. The list of candidate driver genes in gastric cancer predicted from whole-genome sequencing of six gastric cancer cell lines.

Supplemental Table S7. The list of genes with MS mutations that correlate with dysregulated mRNA expression.

Index of supplemental Figures, Tables and Notes

Supplemental Figure Legends

Supplemental Figure S1. Overview of MS mutations identification. Several steps were involved in identifying MS mutations in 18 human gastric cancer cell lines (three MSI-H and 15 MSS cell lines) and sixteen matched pairs of primary human gastric cancer and normal tissues (eight MSI-H and eight MSS cancer tissues). Whole-genome sequencing was performed in six gastric cancer cell lines (three MSI-H and three MSS as a control). Two different algorithms were used for RNA sequencing and whole-genome sequencing data analyses. Detailed methods are provided in the **Methods**. MS mutations in gene regions (CDS and UTRs) were extracted from a list of MS-associated indels (length of 5 or more repetitive nucleotide units) from whole-genome or transcriptome analyses.

Supplemental Figure S2. (A) A Venn diagram depicting the distribution of MS mutations identified in whole-genome sequencing, including those in intergenic and intron regions, between three MSI-H and sum of three MSS gastric cancer cell lines. The number of novel genes is displayed in parentheses. (B) The mutation ratio of the MSs (A/T and G/C) across the whole-genome. Mutated MS and all MS indicate mutated repeat tracts and all existed repeat tracts both containing five or more nucleotides in length, respectively. (C) Length distribution of mononucleotide repeat instabilities of whole-genome found in one, two or all three cell lines.

Supplemental Figure S3. Total deletion counts of MS-associated repeats with a length of 5 or more nucleotide units and number of deletions in gene regions containing mononucleotide repeats identified via whole-genome sequencing (A) and RNA sequencing (B, C) of gastric cancers. The list of genes of (A-C) is summarized in **Supplemental Table S2**.

Supplemental Figure S4. Sequencing chromatograms of novel candidate genes from primary gastric

tissues. Candidates containing microsatellites in their coding regions (A, B) and in the 3' UTRs (C) compared to MSS tissues are displayed. Refseq denotes the reference sequence from Ensembl database. WT and MT indicate the wild-type and mutated (deletion) sequences, respectively.

Supplemental Figure S5. Expression of identified genes with down-regulated mRNA transcription by comparing the mutation status and transcriptional levels in three MSI-H cell lines. (A-D) mRNA transcriptional levels of common down-regulated genes with MSIs in their 3'UTR in MSI-H cell lines were performed by Q-PCR. The levels of mRNA expression by FPKM measured from RNA sequencing are shown. The list of primers is available in **Supplemental Table S8**.

Supplemental Table Legends

Supplemental Tables S2, S5 and S7 are attached individually in Excel.

Supplemental Table S1. Summary of the genome and transcriptome sequencing statistics for the identification of MS mutations in gastric cancer cell lines and primary patient tissues. The sensitivity of the sequencing was performed by comparing indels from whole-genome sequencing data to those from RNA sequencing data of corresponding cell lines. Detailed method is available in **Supplemental Notes**.

Supplemental Table S2. The list of length alterations with all deletions and insertions in gene regions from whole-genome sequencing and RNA sequencing. Genome_Cell lines, Transcriptome_Cell lines and Transcriptome_Tissues indicate whole-genome sequencing of gastric cancer cell lines, RNA sequencing of gastric cancer cell lines, and RNA sequencing of gastric cancer primary patient tissues in each tab of the file, respectively.

Supplemental Table S3. Summary of total indel counts (Total indel), MS-associated indel counts (IndelMS) and MSI target repeat sequence counts in gene regions (IndelMS of gene regions) in gastric cancer cell lines and tissues. MS-associated indel counts were analyzed in tracts with a length of 5 or more nucleotide units of mononucleotide or dinucleotide repeats. “Total indel” and “IndelMS” counts for whole-genome sequencing data include intergenic and intron regions. “Total indel” and “IndelMS” counts for RNA sequencing data include non-coding regions and non-annotated protein coding genes. “Gene regions” are both coding sequences (CDS) and untranslated regions (UTRs). Different indels (i.e., -A, -AA, +A or +AA) occurred at the same MS site were individually counted for “Total indel” but were counted as one deletion or insertion for “IndelMS” and “IndelMS of gene regions”. Dels indicates deletions.

Supplemental Table S4. Frequency of deletions and insertions according to each indel lengths of MSI target mononucleotide repeats within gene regions in human gastric cancer cell lines and primary gastric cancer tissues. “ $\geq \pm 6$ ” indicates deletions or insertions of 6 to 30 nucleotides.

Supplemental Table S5. The list of candidate driver genes in gastric cancer predicted from whole-genome sequencing of six gastric cancers. We evaluated whether the frequency of mutated MSs located in intron and CDS region was statistical significantly higher than expected by the frequency of all existed MSs of a gene.

Supplemental Table S6. The validation result of candidate genes with MS mutations in 22 MSI-H and 13 MSS primary gastric cancer patient tissues. MSI statuses of candidates were verified by PCR and Sanger sequencing with the primary gastric cancer genomic DNAs. Reference sequences are available from Ensembl database under the accession numbers.

Supplemental Table S7. The list of candidate genes with MS mutations that correlate with dysregulated mRNA expression. Up- or down-regulated genes with MS mutations were compared the mutation status and transcriptional levels of specific genes from whole-genome analysis and transcriptome analysis, respectively. The expression levels of down-regulated genes were low or undetectable in the transcriptomes of MSI-H gastric cancer cell lines. Differential gene expression is reported as the log₂ ratio of MSI-H and MSS. Up-regulated genes (grey) have a log₂ ratio of one or greater, while down-regulated genes have a negative ratio value. A log ratio of one represents a two-fold change. The candidate genes were analyzed with statistical significance ($P < 0.05$).

Supplemental Table S8. The list of Q-PCR primers for validation of the expression of genes with UTR MSI.

Supplemental Table S9. The clinical information of the gastric tumors in this study.

Supplemental Notes

Validation of indel accuracy

We checked the specificity and sensitivity of indels detected in six cell lines genomes. The specificity check was performed by Sanger sequencing of 123 indels, and the result showed that all the 123 (100%) were true indels. The sensitivity check was performed by comparing indels from whole-genome sequencing data to those from RNA sequencing data of corresponding cell lines. To acquire reliable indel set of RNA sequencing data, we used the intersection of indels predicted commonly by both methods, Tophat and BWA (Supplemental Fig. S1 and Methods). We found that the indel concordance between whole-genome sequencing and RNA sequencing of MSS cell lines were 64.6~78.3%, and those of MSI-H were 87.1~92.4% (Supplemental Table S1).

The higher concordance ratio of MSI-H cell line is thought to be the higher proportion of longer indels,

which can be more accurately detected by indel detection algorithm.

Calculation of driver gene score calculation

We calculated a driver gene score using a method suggested by Wang et al. with modification (Wang et al. 2011). We started by filtering MSI specific indels by excluding indels occurred in all MSS and MSI-H samples. We used the following criteria to define MSI specific indels that occurred within microsatellites. .

1. Microsatellites of average read mapping depth ≥ 5 ×.
2. Microsatellites composed of A, C, AC, AG, AT, and CG as their repeat units.

Then, we calculated observed number of MSI specific indels within microsatellites of the CDS region in a gene (N_o) and the expected number of MSI specific indels of the CDS region in a gene (N_b), which was estimated using the MSI specific indels in microsatellites in the intron region of a gene (N_{bg}) or genes in a chromosome (N_{bc}). Microsatellites of the intron region used to calculate N_b do not contain microsatellites of repeat number bigger than that of the CDS region. N_b , N_{bg} , and N_{bc} were calculated as

$$N_b = \max(N_{bc}, N_{bg})$$

$$N_{bg} = r_u(uN_{bg})$$

$$N_{bc} = r_u(uN_{bc})$$

where $u \in \{A, C, AC, AG, AT, CG\}$. A, C, AC, AG, AT, and CG are repeat units of microsatellites used for driver gene score calculation. uN_{bc} is the expected number of MSI specific indels of the CDS region in a gene that is calculated from the intron region of genes in a chromosome for a microsatellite repeat unit. uN_{bg} and uN_{bc} were calculated as

$$uN_{bg} = uN_{gcm} cm_u N_{gimi} / uN_{gim}$$

$$uN_{bc} = uN_{gcm} cm_u N_{cimi} / uN_{cim}$$

where, uN_{gcm} is the number of microsatellites in CDS region of a gene for a microsatellite repeat unit, uN_{gim} is the number of microsatellites in intron region of a gene for a microsatellite unit, uN_{cim} is the number of

microsatellites in intron region of genes in a chromosome for a microsatellite unit, ${}_uN_{gimi}$ is the number of MSI specific indels in microsatellites in intron region of a gene for a microsatellite unit, and ${}_uN_{cimi}$ is the number of MSI specific indels in microsatellites in intron region of genes in a chromosome for a microsatellite unit.

Given the N_o and N_b , we calculated P-value based on the Poisson distribution. Then the Q-value (Q_g) was calculated from the P-value of the gene using Q-value library in R (Storey and Tibshirani 2003). Driver gene score (DS_g) was calculated as

$$DS_g = -\log_{10}(Q_g)$$

Supplemental References

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 100(16): 9440-9445.

Wang K, Kan J, Yuen ST, Shi ST, Chu KM, Law S, Chan TL, Kan Z, Chan AS, Tsui WY et al. 2011. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nature genetics* 43(12): 1219-1223.

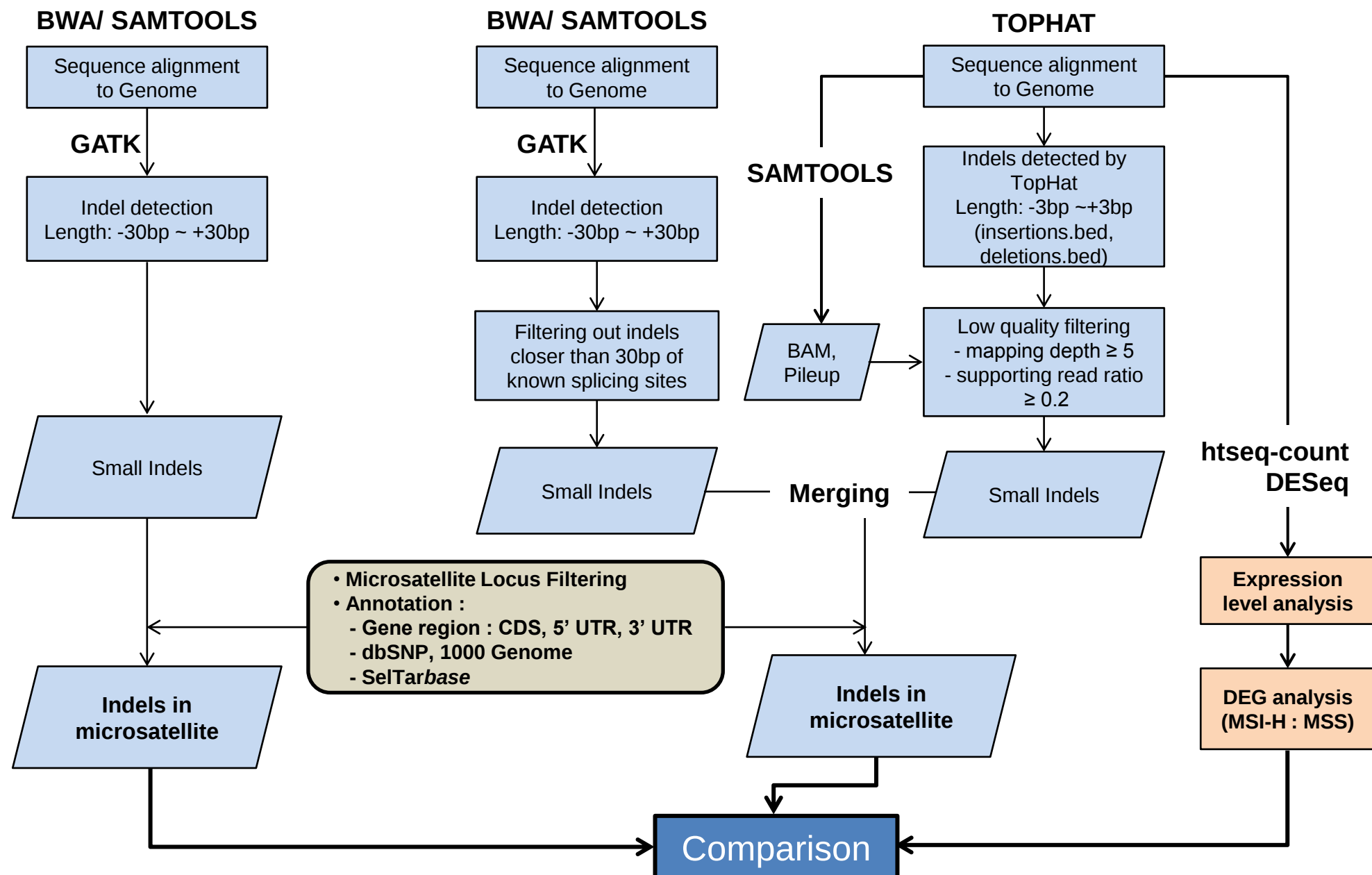
Supplemental Figure S1

Whole-genome Sequencing

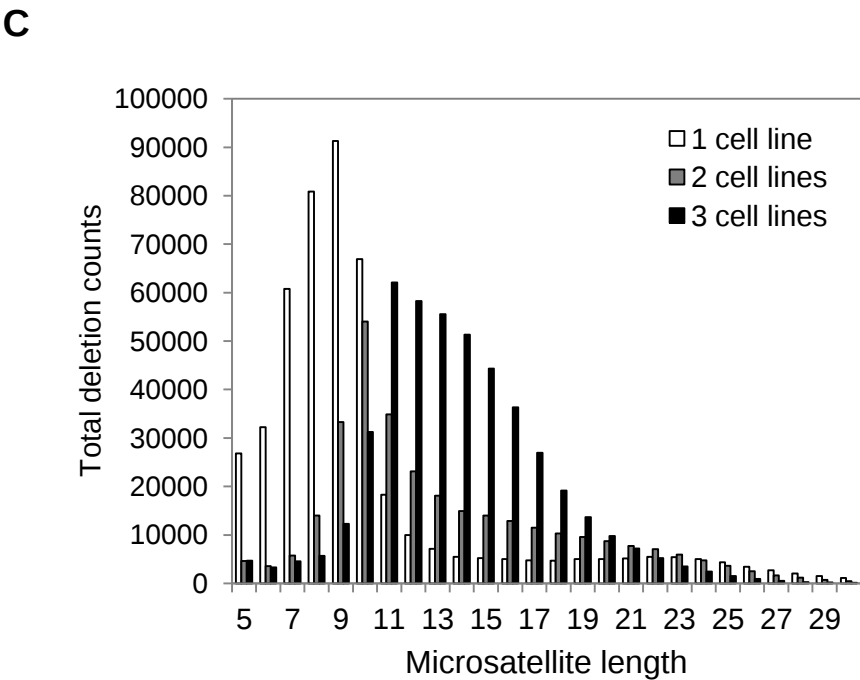
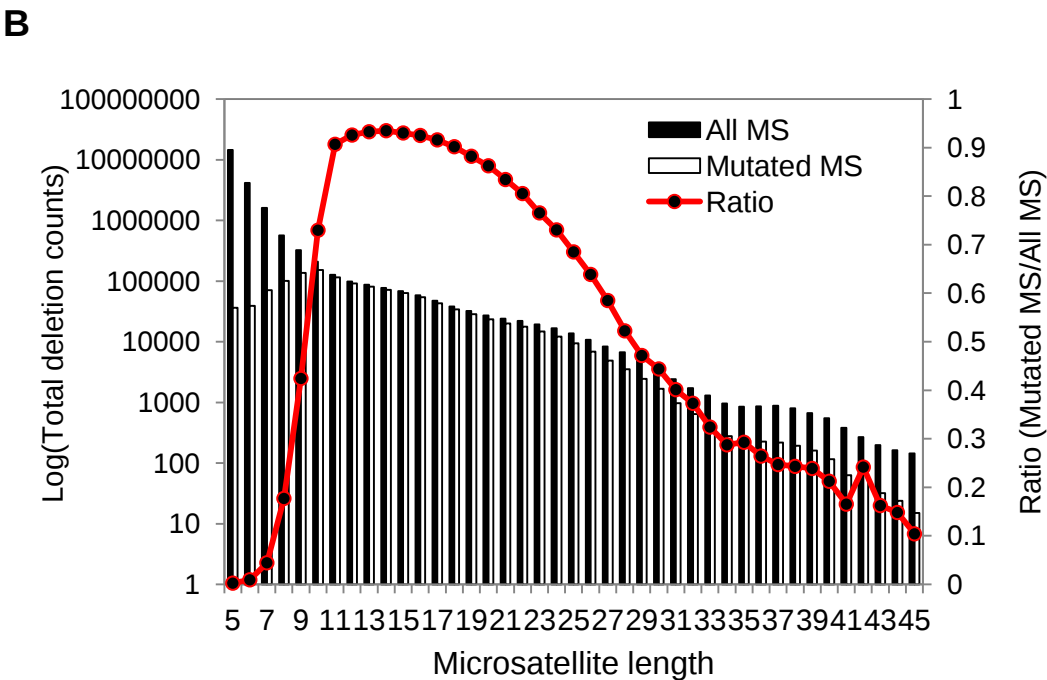
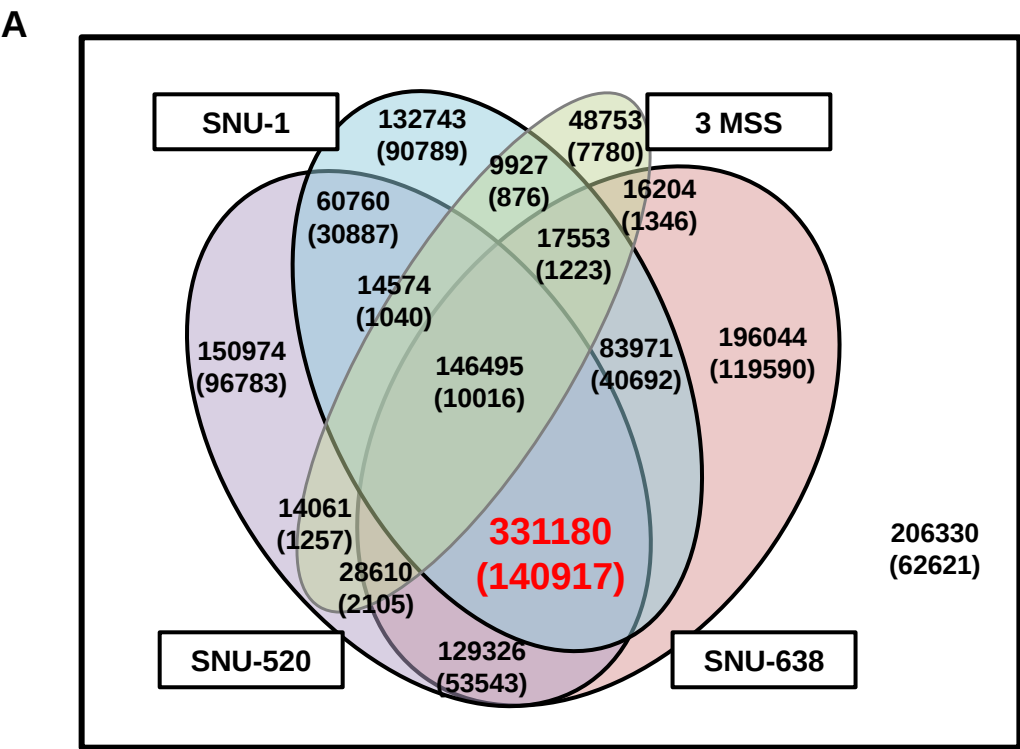
- 6 Cell lines : 3 MSI-H, 3 MSS

RNA Sequencing

- 18 Cell lines : 3 MSI-H, 15 MSS
- 16 Normal-Tumor tissue pairs : 8 MSI-H pair, 8 MSS pair

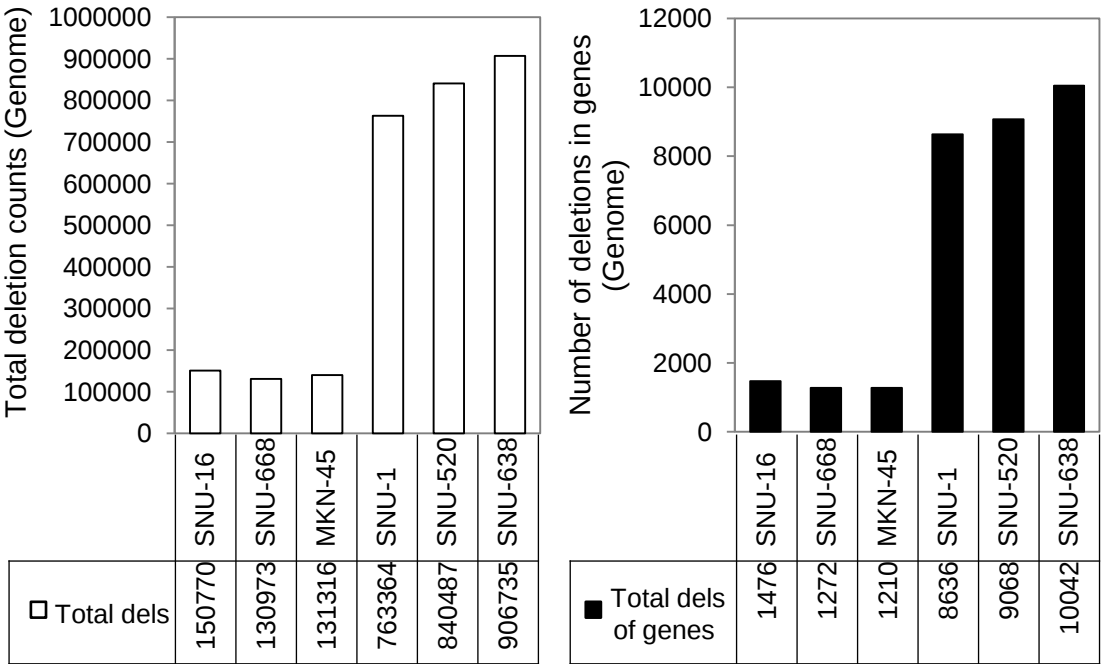


Supplemental Figure S2

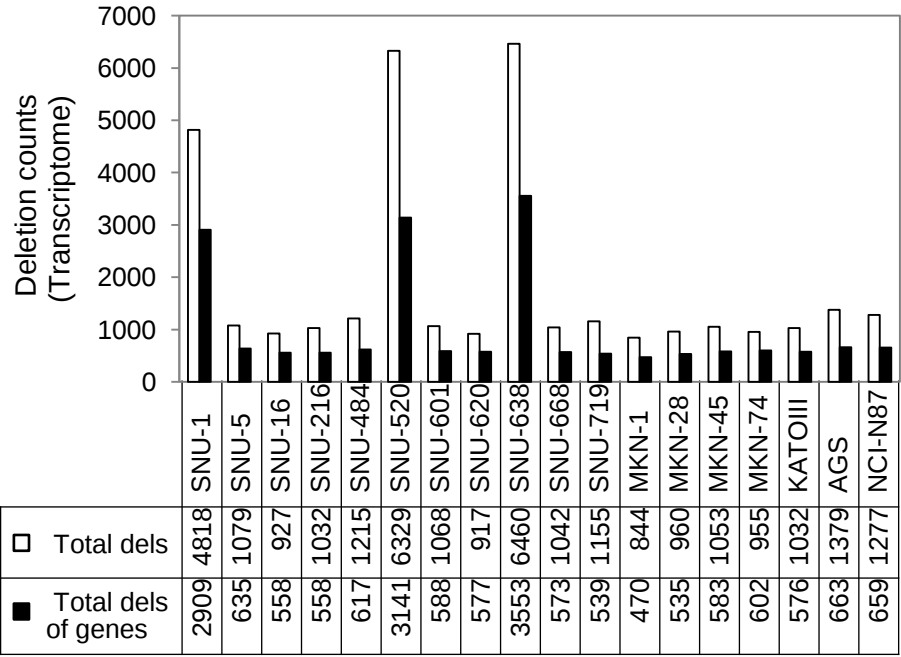


Supplemental Figure S3

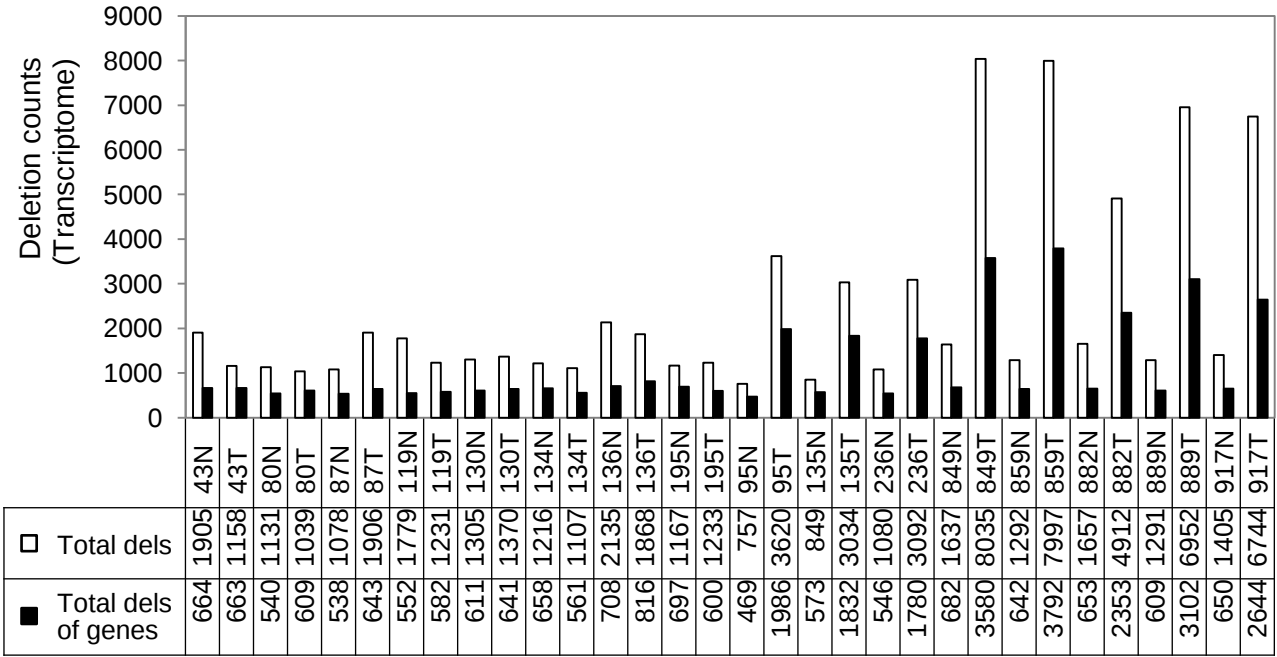
A



B

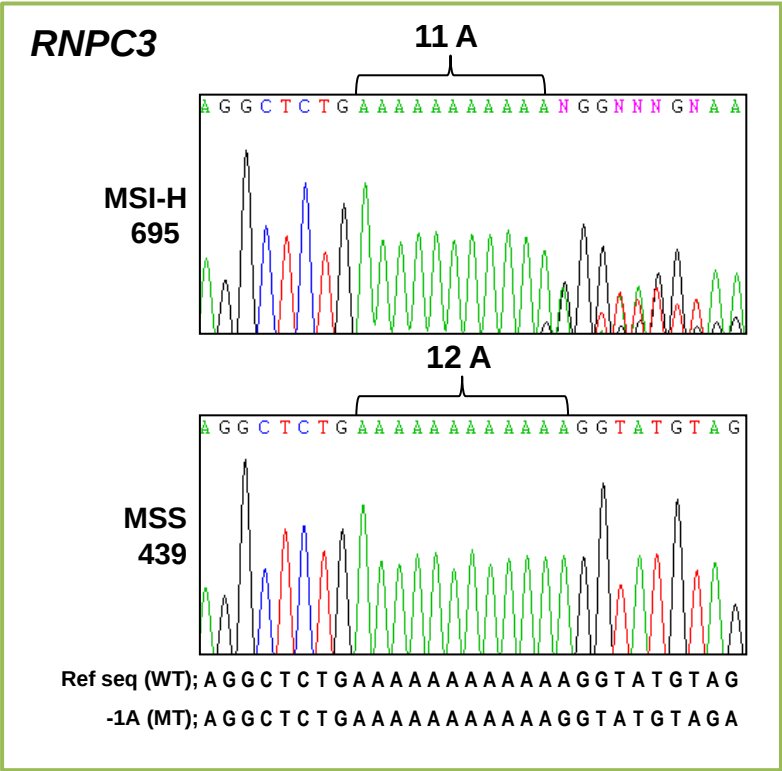


C

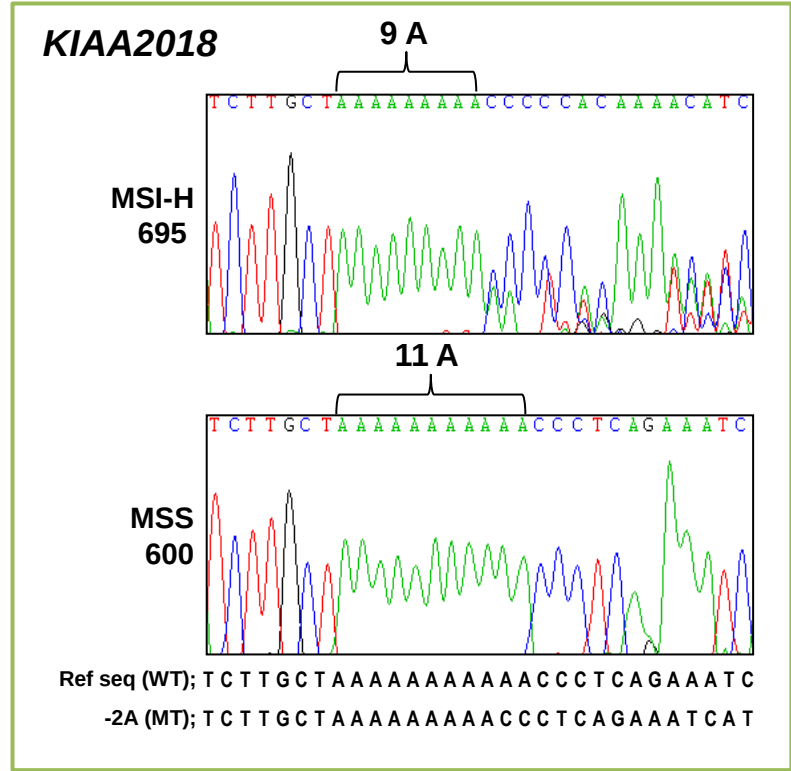


Supplemental Figure S4

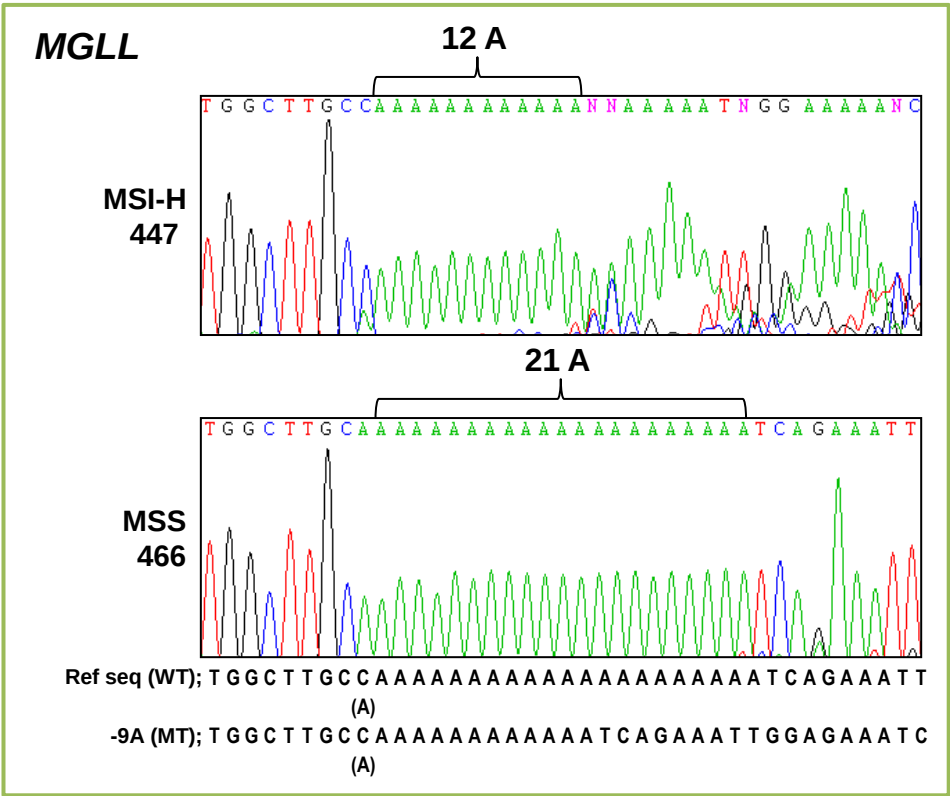
A



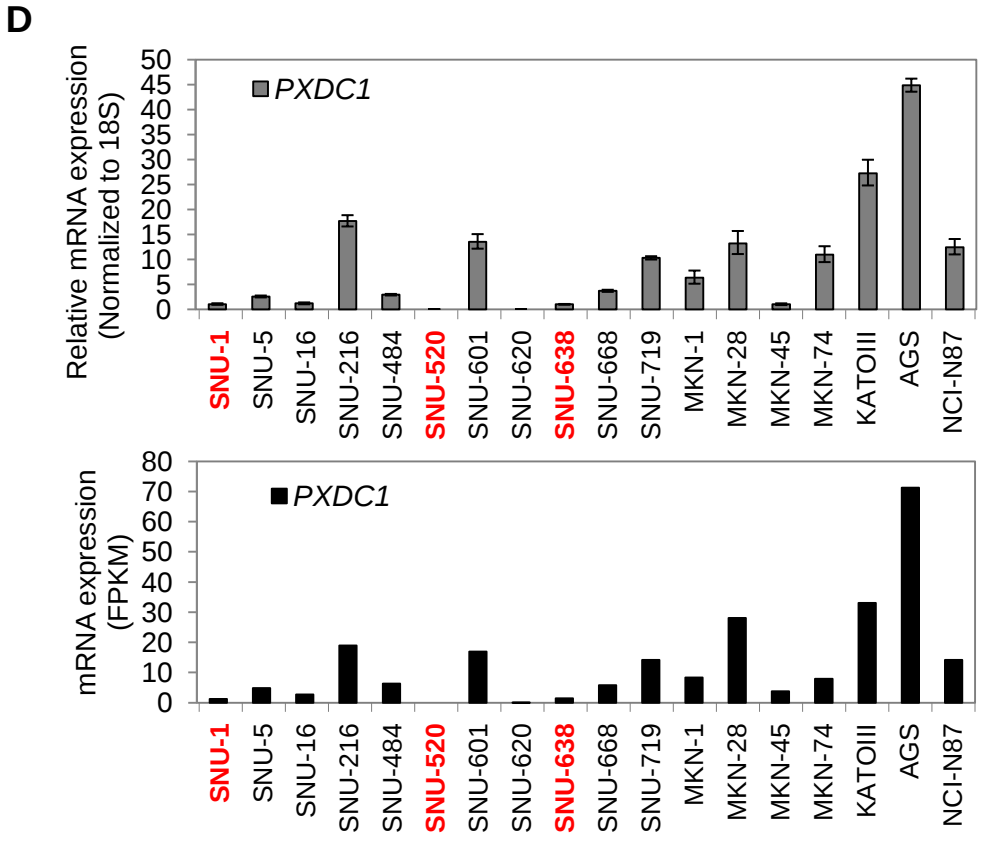
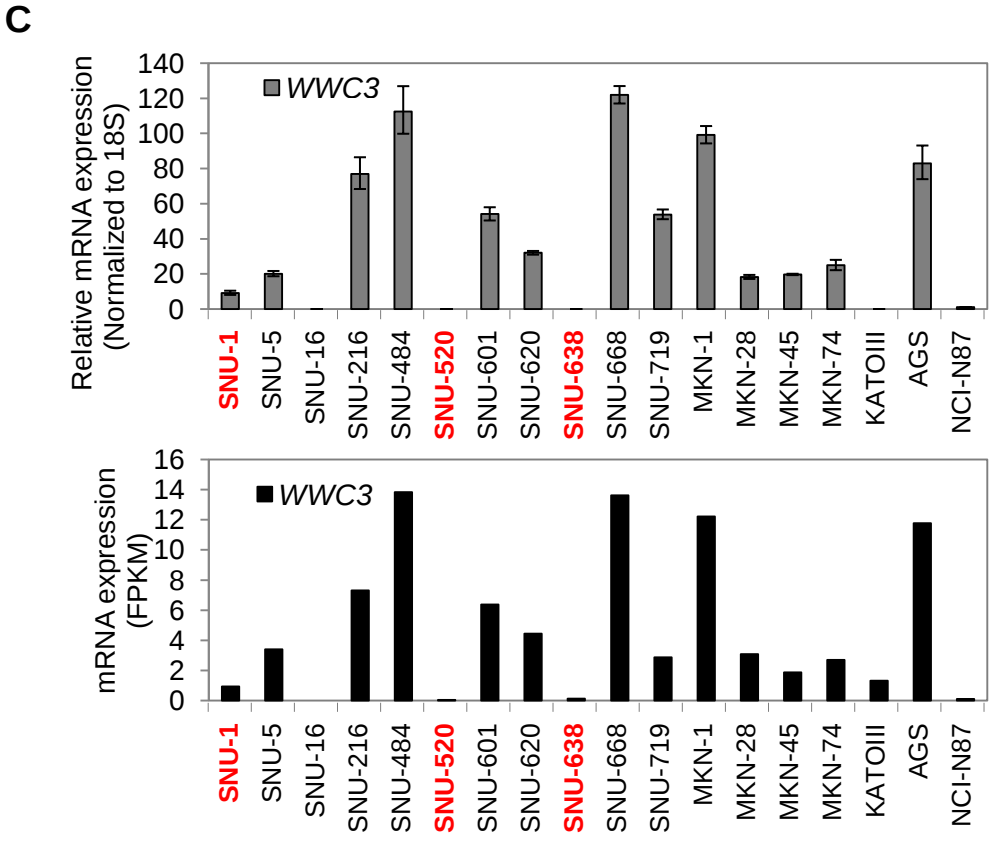
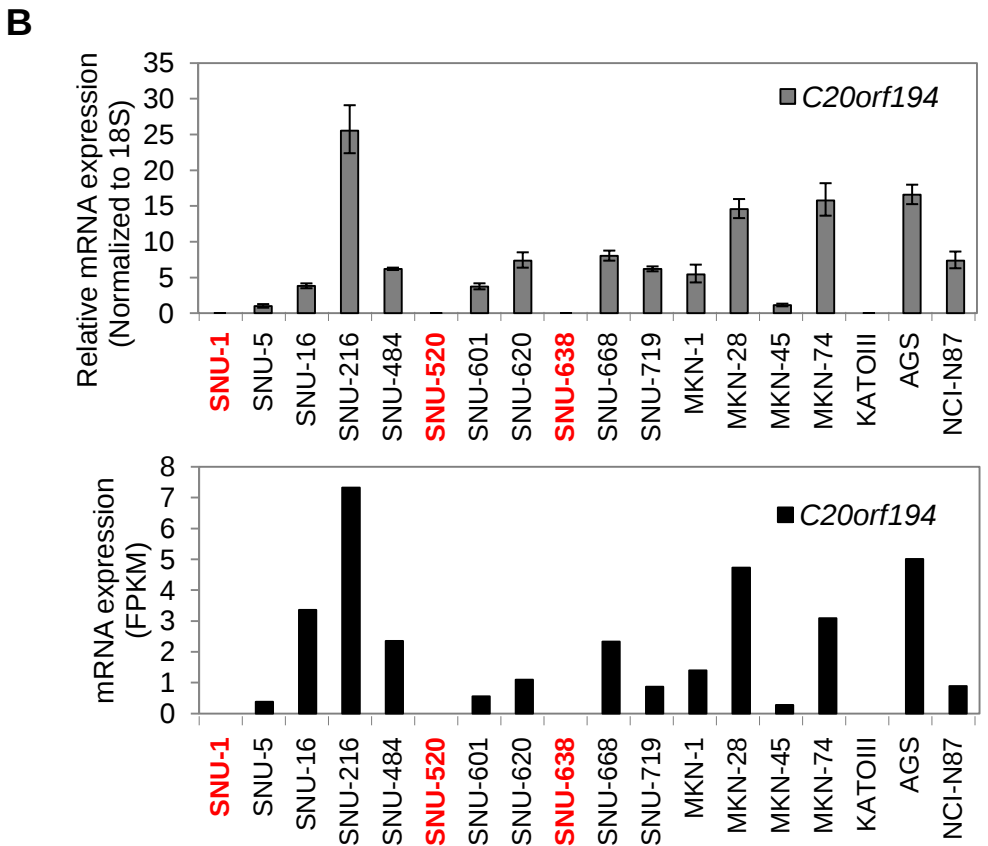
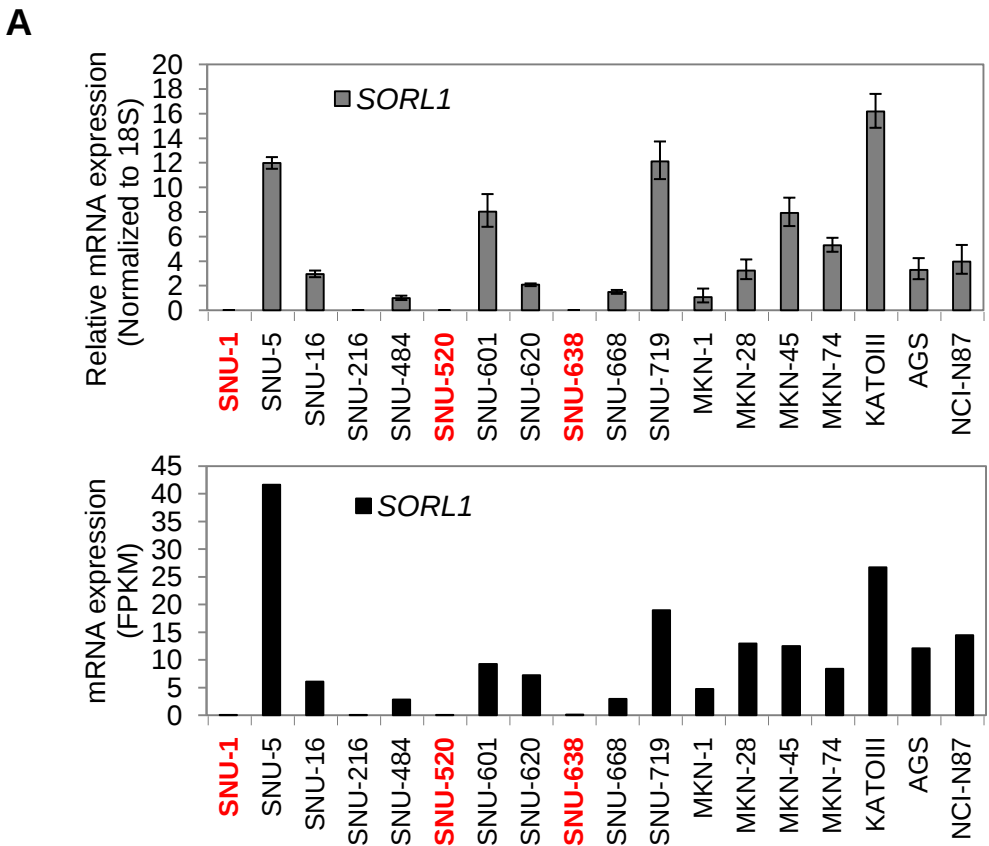
B



C



Supplemental Figure S5



Supplemental Table S1

Whole-genome sequencing of Gastric cancer cell lines												
Cell line	Type	Read length (bp)	Sequencing yield (Gbp)	Total sequenced -reads	Mapped reads	Mapping rate	Mapping depth	Coverage	MS Indels			
SNU-16	MSS	90	97.1	1,079,287,176	1,040,157,953	96.37%	32.7 x	98.90%	333,437			
SNU-668	MSS	100	106.6	1,065,870,022	1,018,144,289	95.52%	35.5 x	98.70%	332,657			
MKN-45	MSS	100	121.9	1,219,219,860	1,169,630,082	95.93%	40.8 x	98.40%	342,604			
SNU-1	MSI-H	90	92.7	1,029,484,456	961,137,822	93.36%	30.2 x	99.40%	915,006			
SNU-520	MSI-H	90	115.1	1,279,112,782	1,200,609,033	93.86%	37.7 x	98.80%	989,924			
SNU-638	MSI-H	90	97.4	1,082,543,620	1,029,112,841	95.06%	32.3 x	99.30%	1,067,370			
RNA sequencing of Gastric cancer cell lines												
Cell line	Type	Read length (bp)	Sequencing yield (Gbp)	Total sequenced -reads	Mapping algorithm	Mapped reads	Mapping rate	MS Indels ¹	MS Indels with co-occurrence in Tophat and BWA ²	Total number of Indels in MS region ³		
SNU-1	MSI-H	90	4.61	51,186,632	Tophat BWA	46,704,593 43,579,558	91.24% 85.14%	3832 (2902) 5335 (4338)	2664 (2359, 89.9%)	6503 (4881, 75.1%)		
SNU-5	MSS	90	4.52	50,218,140	Tophat BWA	46,346,720 43,307,380	92.29% 86.24%	1610 2252			939	2923
SNU-16	MSS	90	4.80	53,333,334	Tophat BWA	49,769,540 46,514,167	93.32% 87.21%	1745 (807) 2275 (1301)	910 (674, 74.1%)	3110 (1434, 46.1%)		
SNU-216	MSS	90	4.69	52,157,444	Tophat BWA	47,800,642 44,681,873	91.65% 85.67%	2001 2535			975	3561
SNU-484	MSS	90	4.69	52,091,840	Tophat BWA	48,244,556 44,910,361	92.61% 86.21%	2183 2848	1091	3940		
SNU-520	MSI-H	90	4.14	45,982,088	Tophat BWA	42,062,675 39,459,877	91.48% 85.82%	4932 (3545) 7111 (5914)			3300 (2874, 87.1%)	8743 (6585, 75.3%)
SNU-601	MSS	90	4.80	53,333,334	Tophat BWA	48,566,465 45,570,170	91.06% 85.44%	1985 2557	1001	3541		
SNU-620	MSS	90	4.62	51,336,766	Tophat BWA	46,840,267 43,822,966	91.24% 85.36%	1201 1833			718	2316
SNU-638	MSI-H	90	4.19	46,605,498	Tophat BWA	43,225,859 40,908,666	92.75% 87.78%	4492 (3713) 6888 (6105)	3282 (3033, 92.4%)	8098 (6785, 87.8%)		
SNU-668	MSS	90	4.07	45,176,716	Tophat BWA	42,110,822 39,257,841	93.21% 86.90%	1656 (834) 2248 (1381)			911 (713, 78.3%)	2993 (1502, 50.2%)
SNU-719	MSS	90	4.80	53,333,334	Tophat BWA	50,626,941 47,082,538	94.93% 88.28%	1742 2360	902	3200		
MKN-1	MSS	90	3.89	43,271,438	Tophat BWA	39,750,685 36,956,000	91.86% 85.41%	1612 2061			803	2870
MKN-28	MSS	90	4.19	46,573,372	Tophat BWA	43,075,296 40,334,728	92.49% 86.60%	1955 2489	989	3455		
MKN-45	MSS	90	4.47	49,654,424	Tophat BWA	45,570,755 42,570,484	91.78% 85.73%	2169 (847) 2710 (1416)			1109 (716, 64.6%)	3770 (1547, 41.0%)
MKN-74	MSS	90	4.80	53,333,334	Tophat BWA	48,490,043 45,690,282	90.92% 85.67%	1426 2060	809	2677		
KATOIII	MSS	90	4.80	53,333,334	Tophat BWA	49,163,398 45,942,880	92.18% 86.14%	1866 2432			954	3344
AGS	MSS	90	6.67	74,059,044	Tophat BWA	67,328,252 62,508,737	90.91% 84.40%	2397 3205	1244	4358		
NCI-N87	MSS	90	4.29	47,684,748	Tophat BWA	44,155,353 41,591,417	92.60% 87.22%	2598 3159			1290	4467
RNA sequencing of Gastric cancer tissues												
Patient	Type	Sample	Read length (bp)	Sequencing yield (Gbp)	Total sequenced -reads	Mapping algorithm	Mapped reads	Mapping rate	MS Indels	Merged MS Indels		
43	MSS	Normal	90	4.69	52,088,890	Tophat BWA	47,206,964 44,773,995	90.63% 85.96%	3621 6021	7967		
						Tumor	90	4.69	52,088,890	Tophat BWA	48,091,237 45,326,232	89.10% 87.02%
		Normal	90	3.89	43,269,450					Tophat BWA	42,892,474 36,359,801	87.90% 84.03%
						Tumor	90	3.87	42,996,076	Tophat BWA	40,789,610 38,108,455	92.30% 88.63%
80	MSS	Normal	90	4.00	44,471,834					Tophat BWA	39,500,055 37,392,955	85.10% 84.08%
						Tumor	90	4.48	49,814,912	Tophat BWA	47,574,144 45,094,950	91.70% 90.53%
		Normal	90	4.44	49,357,002					Tophat BWA	44,816,157 38,284,912	90.80% 77.57%
						Tumor	90	4.21	46,801,172	Tophat BWA	44,041,524 41,582,261	90.30% 88.85%
119	MSS	Normal	90	4.69	52,088,890					Tophat BWA	58,236,973 43,752,083	89.40% 84.00%
						Tumor	90	4.95	55,022,224	Tophat BWA	51,364,571 49,502,563	88.70% 89.97%
		130	MSS	Normal	90					4.19	46,515,690	Tophat BWA
						Tumor	90	4.43	49,181,346			Tophat BWA
134	MSS			Normal	90					4.72	52,400,004	Tophat BWA
						Tumor	90	4.95	55,022,224			Tophat BWA
		135	MSI-H	Normal	90					4.69	52,088,890	Tophat BWA
						Tumor	90	4.70	52,266,668			Tophat BWA
136	MSS			Normal	90					4.93	54,755,556	Tophat BWA
						Tumor	90	4.82	53,511,112			Tophat BWA
		195	MSS	Normal	90					4.82	53,511,112	Tophat BWA
						Tumor	90	4.69	52,088,890			Tophat BWA
236	MSI-H			Normal	90					3.71	41,204,902	Tophat BWA
						Tumor	90	3.96	43,956,792			Tophat BWA
		849	MSI-H	Normal	90					5.23	58,124,992	Tophat BWA
						Tumor	90	5.75	63,880,118			Tophat BWA
859	MSI-H			Normal	90					5.31	59,019,350	Tophat BWA
						Tumor	90	5.06	56,227,424			Tophat BWA
		882	MSI-H	Normal	90					5.72	63,549,430	Tophat BWA
						Tumor	90	5.32	59,102,148			Tophat BWA
889	MSI-H			Normal	90					4.00	44,498,444	Tophat BWA
						Tumor	90	5.37	59,655,378			Tophat BWA
		917	MSI-H	Normal	90					5.65	62,755,892	Tophat BWA
						Tumor	90	5.70	63,306,020			Tophat BWA
¹ MS Indels in each Tophat and BWA (The number of concordant Indels with whole-genome sequencing (WGS) Indels)												
² MS Indels with co-occurrence in Tophat and BWA (The number of concordant Indel with WGS Indels, Concordant ratio (%))												
³ Total number of Indels (Sum of indels of Tophat and BWA) in MS region (The number of concordant Indel with WGS Indels, Concordant ratio (%))												

Supplemental Table S3

Whole-genome sequencing				RNA sequencing											
				Cell lines						Primary tissues					
	Total indel	IndelMS	IndelMS of gene regions		Total indel	IndelMS	IndelMS of gene regions			Total indel	IndelMS	IndelMS of gene regions			
					Total	Dels	Total	Dels			Total	Dels	Total	Dels	
SNU-16	557922	Total; 333437	Total; 3278	SNU-1	10357	6570	4976	3896	3010	43N	13118	8092	2110	2980	764
		Del; 181851	Del; 1683	SNU-5	6476	2962	1213	1839	725	43T	8144	4063	1330	2475	771
SNU-668	546302	Total; 332657	Total; 3018	SNU-16	6685	3176	1042	1926	625	80N	8250	4784	1318	2424	639
		Del; 168413	Del; 1504	SNU-216	7361	3622	1184	2071	648	80T	7197	3875	1217	2350	722
MKN-45	560062	Total; 342604	Total; 3071	SNU-484	7965	4003	1380	2229	701	87N	8318	4603	1229	2440	631
		Del; 177209	Del; 1502	SNU-520	13006	8854	6566	4274	3258	87T	14282	8370	2182	2993	738
SNU-1	1142502	Total; 915006	Total; 10134	SNU-601	7581	3615	1233	2096	696	95N	5906	3270	886	2118	554
		Del; 801531	Del; 8893	SNU-620	5793	2359	1034	1527	661	95T	11041	6444	3877	3229	2114
SNU-520	1221768	Total; 989924	Total; 10571	SNU-638	11925	8183	6715	4417	3700	119N	11294	7344	1965	2350	621
		Del; 881725	Del; 9374	SNU-668	6421	3041	1197	1733	667	119T	8112	4238	1414	2094	673
SNU-638	1302482	Total; 1067370	Total; 11575	SNU-719	7217	3244	1317	1617	629	130N	8442	4906	1486	2441	705
		Del; 956735	Del; 10394	MKN-1	6393	2919	992	1772	568	130T	8905	4911	1583	2357	742
				MKN-28	6928	3519	1118	2024	624	134N	8446	4636	1398	2612	754
				MKN-45	7537	3839	1190	2240	668	134T	7818	3866	1270	2078	640
				MKN-74	6206	2728	1074	1782	686	135N	6306	3198	960	2267	653
				KATO III	7154	3398	1173	1948	663	135T	10397	5857	3248	3313	1958
				AGS	9169	4438	1583	2187	764	136N	12316	7780	2347	2773	800
				NCI-N87	8826	4550	1433	2486	756	136T	11930	6580	2104	3020	934
										195N	8260	4346	1312	2683	789
										195T	6345	2880	1398	1408	676
										236N	7557	4307	1236	2222	631
										236T	10702	6337	3340	3398	1917
										849N	17968	9066	1936	4387	818
										849T	22854	12930	8498	5747	3739
										859N	14719	7215	1518	4071	760
										859T	20817	11602	8409	5634	3988
										882N	19453	9990	1994	4637	811
										882T	22789	13615	5273	6340	2507
										889N	13522	6389	1522	3273	732
										889T	22207	13065	7344	5964	3275
										917N	14964	7050	1672	4038	793
										917T	28719	18122	7189	6863	2788

Supplemental Table S4

Gastric cancer cell lines																													
Whole-genome sequencing								RNA-sequencing																					
indel unit		SNU-16	SNU-668	MKN-45		SNU-1	SNU-520	SNU-638		SNU-1	SNU-5	SNU-16	SNU-216	SNU-484		SNU-520	SNU-601	SNU-620		SNU-638	SNU-668	SNU-719	MKN-1	MKN-28	MKN-45	MKN-74	KATOIII	AGS	NCI-N87
AorT	≥-6	49	54	53		518	806	1079		91	14	11	15	23		124	18	14		190	11	12	11	11	12	18	8	13	18
AorT	-5	17	15	13		373	408	495		41	9	3	4	4		53	7	4		91	3	1	5	4	2	1	4	3	5
AorT	-4	45	53	41		593	680	714		85	13	10	11	13		103	14	11		121	12	18	9	11	13	15	13	12	17
AorT	-3	67	55	49		996	979	1056		439	32	40	27	29		493	25	23		497	29	30	23	23	30	30	40	31	31
AorT	-2	164	139	152		1603	1600	1668		559	71	75	77	78		570	72	63		667	65	62	48	75	67	77	68	83	68
AorT	-1	867	771	787		3907	4011	4311		1502	428	354	359	403		1643	382	386		1781	405	352	308	357	393	400	388	449	447
AorT	+1	942	906	932		715	644	632		610	824	991	1136	1196		690	1121	629		424	792	727	932	1109	1255	824	998	1082	1367
AorT	+2	127	152	144		44	54	57		29	65	75	69	99		27	66	47		22	73	60	63	73	96	69	77	85	92
AorT	+3	58	55	64		22	16	27		36	34	30	33	29		38	31	31		35	24	22	22	33	50	19	35	38	43
AorT	+4	35	34	42		27	27	22		8	17	12	12	16		10	16	15		4	10	13	18	16	13	17	13	21	19
AorT	+5	15	15	19		12	13	13		7	12	16	12	17		6	17	7		10	11	11	10	15	12	15	9	18	22
AorT	≥+6	34	26	37		12	18	16		7	21	16	17	21		11	14	8		9	13	14	13	16	16	16	19	21	19
GorC	≥-6	8	4	6		23	17	15		2	4	3	6	4		5	4	6		5	3	3	2	3	4	5	6	2	1
GorC	-5	2	1	0		2	5	9		1	0	0	1	0		1	0	1		1	0	0	0	0	0	0	1	0	0
GorC	-4	2	0	1		11	9	13		3	1	2	2	0		5	2	1		2	1	2	0	1	2	1	1	2	2
GorC	-3	9	1	3		20	22	32		11	4	7	7	9		17	10	7		20	3	13	5	4	7	4	4	8	5
GorC	-2	18	12	13		44	50	48		16	9	7	7	4		14	8	12		23	5	6	7	7	6	8	4	5	10
GorC	-1	104	87	76		372	302	417		159	50	46	42	50		113	46	49		155	36	40	52	39	47	43	39	55	55
GorC	+1	109	62	70		161	164	163		77	61	62	53	61		79	61	57		88	53	63	52	57	46	54	60	61	61
GorC	+2	5	7	8		10	10	13		6	5	4	9	9		6	6	7		7	5	5	9	4	3	5	4	7	6
GorC	+3	4	5	4		3	4	3		2	2	2	2	2		2	0	0		6	2	1	4	3	2	1	2	3	2
GorC	+4	0	1	3		0	0	3		0	0	1	1	1		2	1	0		0	1	0	0	0	0	0	1	2	0
GorC	+5	1	2	1		0	0	0		0	0	2	0	0		0	0	1		0	1	0	0	0	1	0	0	1	0
GorC	≥+6	3	2	0		0	1	1		0	2	0	1	1		2	0	2		1	1	1	1	2	1	3	1	0	1

Gastric cancer Tissues (RNA-sequencing)																																	
indel unit		43N	43T	80N	80T	87N	87T	95N	95T	119N	119T	130N	130T	134N	134T	135N	135T	136N	136T	195N	195T	236N	236T	849N	849T	859N	859T	882N	882T	889N	889T	917N	917T
AorT	≥-6	21	17	11	10	16	15	4	72	13	12	13	8	16	7	15	64	15	26	20	19	12	68	19	172	24	190	13	60	17	199	14	80
AorT	-5	5	6	4	4	4	6	3	30	5	5	4	5	4	2	9	26	7	7	3	3	6	37	9	92	10	84	9	44	5	73	7	48
AorT	-4	11	11	6	9	13	10	11	57	12	10	14	10	11	8	17	58	17	21	12	6	16	72	14	142	15	146	17	60	21	134	23	96
AorT	-3	30	30	26	29	17	35	26	226	32	24	38	36	30	23	26	181	36	39	36	30	31	196	34	375	28	410	21	268	24	343	20	287
AorT	-2	77	71	70	86	71	81	59	355	60	60	79	75	78	68	68	321	80	93	85	59	69	295	83	594	60	587	68	393	64	537	77	441
AorT	-1	439	441	345	388	347	424	314	1113	358	387	403	424	438	381	377	1067	476	540	467	411	349	986	442	1875	421	2106	440	1334	397	1585	413	1496
AorT	+1	1768	1340	1485	1315	1458	1834	1302	776	1365	1101	1413	1268	1457	1107	1276	957	1553	1657	1500	480	1288	1097	3010	1533	2768	1195	3230	3245	2066	2281	2724	3448
AorT	+2	169	111	84	73	113	147	77	59	134	77	97	96	128	93	110	67	127	123	126	53	99	72	172	45	177	44	185	128	121	44	154	153
AorT	+3	61	36	32	30	43	62	36	69	40	36	31	29	41	39	35	84	63	62	44	29	35	71	53	103	50	81	61	151	61	88	56	137
AorT	+4	24	20	13	14	18	25	7	26	12	13	15	18	19	12	23	34	23	15	17	9	9	20	29	21	24	17	22	32	19	19	19	20
AorT	+5	16	16	13	13	15	23	7	8	18	15	13	20	15	26	13	17	18	17	16	13	13	18	46	16	40	12	51	20	34	10	25	14
AorT	≥+6	20	19	9	13	16	24	12	11	22	23	18	18	15	13	8	18	23	21	24	14	11	9	26	9	33	13	30	13	24	9	22	11
GorC	≥-6	5	4	3	4	7	4	2	2	3	5	3	4	7	4	1	3	3	3	5	2	2	5	6	7	6	9	6	8	3	6	7	10
GorC	-5	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	1	0	1	0	2	1	3
GorC	-4	2	2	2	2	1	1	1																									

[illegible]

Supplemental Table S8

Ensembl Accession	Gene Symbol	Primer sequences
ENSG00000074416	<i>MGLL</i>	<i>MGLL</i> F; CCCTCATCTTTGTGTCCCATG <i>MGLL</i> R; ACTACCATCCTCTCCCCTTC
ENSG00000137642	<i>SORL1</i>	<i>SORL1</i> F; TTGGCTTGGGAAATAGGAGTG <i>SORL1</i> R; CCTTGAAGAGTGTTGCAGAAG
ENSG00000088854	<i>C20orf194</i>	<i>C20orf194</i> F; GAAGGACATCTGGTACGAAGC <i>C20orf194</i> R; CCACCCAAGTAAGGAACATGAG
ENSG00000047644	<i>WWC3</i>	<i>WWC3</i> F; TGTTGCATGGGTATCGACTC <i>WWC3</i> R; AGATCTTCCGTGTTGGTTTCC
ENSG00000168994	<i>PXDC1</i>	<i>PXDC1</i> F; GTGGACCCAACAGAGCATT <i>PXDC1</i> R; AGAGGTTTCAGTCCCAAATGTC

Supplemental Table S9

Serial	Sex	Age	Location	Size (cm)	Histologic subtype (WHO classification)	Lauren	Stage	MSI Status
43	M	51	U	7.0	WD	Gastric	IIIb	MSS
80	M	56	L	5.0	PD	Diffuse	IIIb	MSS
87	M	53	U	8.0	MD	Diffuse	IIIb	MSS
95	M	65	L	6.0	WD	Intestinal	IIa	MSI-H
119	M	48	L	10.0	WD	Intestinal	IIa	MSS
130	M	57	L	6.5	MD	Intestinal	Ib	MSS
134	M	57	L	8.5	WD	Intestinal	IIIb	MSS
135	M	37	L	6.0	Mucinous	Intestinal	Ib	MSI-H
136	M	55	L	4.5	MD	Diffuse	IIa	MSS
195	M	67	U	6.0	Papillary	Intestinal	IIa	MSS
236	F	72	L	6.5	MD	Mixed	Ia	MSI-H
849	M	64	L	5.4	PD	Intestinal	Ib	MSI-H
859	F	60	L	5.0	MD	Intestinal	Ib	MSI-H
882	F	75	L	3.5	MD	Intestinal	Ia	MSI-H
889	F	66	L	8.7	MD	Intestinal	Ia	MSI-H
917	M	73	L	7.9	MD	Intestinal	Ia	MSI-H
432	F	66	M	12.5	Undifferentiated	Intestinal	Ib	MSI-H
466	F	86	U	6.0	MD	Diffuse	Ia	MSI-H
473	F	73	L	7.5	WD	Intestinal	Ib	MSI-H
474	M	67	L	6.2	MD	Intestinal	IIIb	MSI-H
477	M	82	M	12.7	Undifferentiated	Intestinal	IIc	MSI-H
483	M	75	M	6.0	WD	Intestinal	Ia	MSI-H
506	M	64	L	6.4	MD	Intestinal	Ia	MSI-H
508	M	72	L	7.5	WD	Intestinal	Ib	MSI-H
528	M	74	L	11.2	WD	Intestinal	IIa	MSI-H
529	F	57	L	6.5	Papillary	Intestinal	IIc	MSI-H
532	M	62	L	5.0	WD	Intestinal	Ib	MSI-H
627	F	80	M	10.0	WD	Intestinal	Ib	MSI-H
637	F	62	L	6.5	MD	Diffuse	Ib	MSI-H
641	M	63	M	9.5	Undifferentiated	Diffuse	IV	MSI-H
646	F	52	U	5.5	MD	Diffuse	Ib	MSI-H
662	M	55	L	3.5	MD	Diffuse	Ib	MSI-H
669	M	47	L	11.0	WD	Intestinal	Ia	MSI-H
678	F	51	L	8.0	WD	Intestinal	IIIb	MSI-H
692	F	68	L	5.8	WD	Intestinal	Ia	MSI-H
693	F	76	L	8.0	MD	Diffuse	Ib	MSI-H
695	F	57	L	4.5	MD	Diffuse	Ib	MSI-H
713	M	78	L	11.0	MD	Intestinal	Ia	MSI-H
433	F	59	L	5.5	Papillary	Intestinal	Ib	MSS
434	M	46	L	8.5	MD	Diffuse	IIc	MSS
436	M	24	M	8.0	MD	Diffuse	IV	MSS
437	F	43	U	4.5	WD	Intestinal	Ia	MSS
438	M	56	L	9.0	Papillary	Diffuse	IIa	MSS
439	M	48	U	5.8	WD	Intestinal	Ib	MSS
440	F	70	U	6.0	WD	Intestinal	Ia	MSS
441	M	78	U	5.2	WD	Intestinal	Ia	MSS
447	M	37	L	4.8	WD	Intestinal	IIa	MSS
449	M	64	L	6.2	Undifferentiated	Intestinal	IIIb	MSS
450	F	66	U	11.0	MD	Diffuse	IV	MSS
533	M	66	M	7.5	WD	Intestinal	Ia	MSS
600	F	40	M	3.5	MD	Diffuse	Ib	MSS

U, upper third; M, middle third; L, lower third
WD, tubularadenocarcinoma and well differentiated; MD, tubularadenocarcinoma and moderately differentiated;
PD, tubularadenocarcinoma and poorly differentiated
MSI-H, High-level microsatellite instability; MSS, microsatellite stable