Supplemental Information

1. Supplemental figure legends

Supplemental Figure S1. The on-target capture ratio of Exome-Seq. The ratio of uniquely mapped sequence reads that fell within the targeted regions. The mean values and standard deviation (SD) are shown. On average, 68.4% of the unique reads that mapped in consistent pairs fell within the targeted regions. The values varied quite slightly among chromosomes, except the Y chromosome, which is absent in cells from female patients and is frequently deleted in tumor cells from male patients.

Supplemental Figure S2. The fold enrichment for each target exome. The enrichment level of the targeted regions was calculated for each sample. Only the unique reads that mapped in consistent read pairs and fell within the targeted regions were counted. On average, the targeted regions of each exome were enriched by > 56-fold.

Supplemental Figure S3. Sequence coverage of targeted bases. The fraction of the targeted bases that covered by sequence reads is shown. The error bars represent the standard deviation of the mean read depth among 15 exomes. On average per exome, nearly 97% of all targeted bases have at least a 1x read depth and over 83% of all targeted bases have at least a 10x read depth.

Supplemental Figure S4. Length distribution of somatic indels. The total number of somatic small indels called in targeted regions in 15 tumors is shown. The lengths of the small indels varied from 1 to 29 base pairs (bp). 78% of the indels were 1-3 bp in length.

Supplemental Figure S5. The most frequent targets of homozygous deletions. The DNA copy number alterations were analyzed using the GIM algorithm. In each graph, the *y* axis indicates the adjusted log2 ratios of signal intensities between the tumor sample and its matched normal sample for the perfect match probes. The red line represents the allele with a higher copy number and the blue line represents the allele with a lower copy number. The log2 ratio of -1 and 0 theoretically corresponds to 0 and 1 copy, respectively. The *CDKN2A* locus at *9p21.3* and *SMAD4* locus at *18q21.2* were frequently deleted in tumor cell lines analyzed.

Supplemental Figure S6. The frequently altered signaling pathways. (A) The 9 core signaling pathways. The pathway analysis was performed using Gene Ontology (GO) database (http://www.geneontology.org/). Only the pathways that mutated in ≥ 40% fraction of tumor cell lines and with ≥ 2 somatic mutations identified in at least 2 of their gene members were listed. The somatic mutations mainly clustered in 9 pathways, whereas the pathway components that were altered in individual tumor cell line varied widely. **(B)** The normalized mutation rates of each pathway for cell lines in 3 subgroups. The *MLH1*-HD cell line showed a much higher prevalence of mutations involved in all 9 pathways (p = 0.0007). The *MLH1*-LOH cell lines showed slightly but significantly increased mutation rates in 7 of the 9 pathways (p = 0.037).

Supplemental Figure S7. Validation of the truncating mutations which identified by Exome-Seq but not confirmed by mRNA-Seq. All 13 mutations were heterozygous and introduced premature termination codons. The gDNA and cDNA of the tumor cell lines as well as the gDNA of their matched normal samples were sequenced using Sanger method. The sequence electropherograms were shown and the positions of mutations were indicated by arrows. The information for each mutation including the gene symbol, the wild-type allele, the mutant allele and the mutant allele ratio was shown at the top of each panel. Forward, the PCR product was sequenced using the forward primer; reverse, the PCR product was sequenced using the reverse primer. By Sanger sequencing, 12 of the 13 truncating mutations were successfully validated in the gDNA of the tumor cell lines rather than their cDNA. One mutation in PTPRS was not validated (gray arrow).

Supplemental Figure S8. Comparison of the recurrently mutated genes identified in this study with that of an earlier study performed by Jones and colleagues. In this study, only the genes identified with ≥ 2 deleterious mutations (the nonsense, missense substitutions, frame-shift indels and focal homozygous deletions) in ≥ 2 tumor cell lines were counted. A total of 7 recurrently mutated genes were common in both studies.

Supplemental Figure S9. The expression level of DNA MMR genes. The expression level was examined by mRNA-Seq. All DNA MMR genes except for *MLH1* were shown. The reads per kilobase per million mapped reads (RPKM) value was calculate for each transcript. No significant difference was observed in the expression of these MMR genes among subgroups.

Supplemental Figure S10. *MLH1* promoter methylation analysis. The *MLH1* promoter methylation status was quantitatively determined using MassARRAY. The amplified DNA which was not methylated at all in any CpG sites was used as unmethylated (0%) control. The amplified DNA, methylated by S*ss*I methylase, was used as a fully methylated (100%) control. None of the tumor cell line showed promoter hypermethylation of *MLH1*.

Supplemental Figure S11. The truncating indels and *TP53* expression. (**A**) The expression level of *TP53*. The expression level was examined by mRNA-Seq. The reads per kilobase per million mapped reads (RPKM) value was calculated. The *TP53* copy number status, the presence or absence of somatic indels, missense or nonsense substitutions or RNA editings were annotated for each tumor cell line. (**B**) The truncating indels identified in two *MLH1*-LOH cell lines. The up-right panel shows a 1-bp deletion identified in a *MLH1*-LOH cell line PA055. The upper graph shows the Sanger sequence electropherogram and the lower graph shows the copy number alterations. The bottom-right panel shows a 4-bp insertion detected in another *MLH1*-LOH cell line PA202. In both cell lines, one allele of *TP35* was lost by DNA copy number deletion and the remaining allele was inactivated by truncating indels.

Supplemental Figure S12. MSI analysis using the conventional assay. (**A**) The results of MSI analysis. "-" indicates negative, "+" indicates positive, "1L, 1R, 2L" indicates the size peaks of amplified DNA fragments from tumor cell line shifted to the left or right by one or two base pairs compared to that of its matched normal sample. MSS, microsatellite stable; MSI-H, high-frequency MSI. (**B**) Representative results of a MSI-positive case. For each marker, the top and bottom graphs show size peaks of the amplified DNA fragments from the tumor cell line and its matched normal sample, respectively. The extra peaks and shorter peaks observed in the tumor cell line were marked by arrows. The *x* axis is size in bases and the *y* axis is fluorescence intensity. Red peaks are internal size standards. The assay revealed that all 7 markers were stable in *MLH1*-ROH and *MLH1*-LOH cell lines and 2 markers, D17S250 and D2S123, were unstable in the *MLH1*-HD cell line. Using the conventional MSI assay, *MLH1*-LOH cell lines were indistinguishable from *MLH1*-ROH tumors.

Supplemental Figure S13. The somatic indels identified in primary RCCs using Exome-Seq. Target enrichment was done using the Agilent Human All Exon 50Mb Kit and Exome-Seq was performed using the HiSeq 2000 sequencing system. Among the 15 RCCs analyzed, 13 cases showed LOH at the *MLH1* locus on chromosome *3p* and 2 cases showed ROH. The number of somatic small indels identified by Exome-Seq is shown for each tumor. The *MLH1*-LOH tumors showed a significantly increased mutation rate of somatic indels (*p* = 0.0008). The majority of the indels identified in *MLH1*-LOH tumors were frame-shift.

Supplemental Figure S14. Characterization of the targeted regions of Exome-Seq. (A) Comparison of the targeted genes sequenced in this study with that of an earlier study performed by Jones and colleagues using Sanger sequencing method. **(B)** The size distribution of the targeted exons sequenced in this study.

Supplemental Figure S15. The flowchart of mutation detection by Exome-Seq. For each cell line and its matched normal sample, the passing filter reads were mapped to human reference genome hg18 using BWA aligner. After mapping, the ambiguously mapped reads, inconsistent read pairs, singletons and duplicate reads were excluded from further analysis. The substitutions were called using SAMtools algorithm and the indels were called using both SAMtools and Pindel algorithms. The low quality variants were removed using the “quality filters” and the somatic variants were selected using the “somatic filters”. The potential false-positive events were further removed using BLAT algorithm and IGV viewer.

Supplemental Figure S16. Evaluation of the concordance for common SNPs between Exome-Seq and genome-wide SNP array. For each of the common SNPs, the genotype that inferred by Exome-Seq was compared with that called by SNP array. In the upper panel, the *x* and *y* axes show the read depth and the non-reference allele ratio of each SNP detected by Exome-Seq, respectively. According to the genotypes data called by SNP array, the heterozygous SNPs were colored in red and the homozygous SNPs were colored in green if they were same as the reference bases and in blue if they were different. The lower panel shows the summary statistics. There are a total of 7,711 SNPs shared between two analyses. The numbers highlighted in green indicates the number of concordant SNPs. The concordance rate was 99.84%.

2. Supplemental Tables

Supplemental Table S4. Additional microsatellite markers used in this study.

Marker	MS repeat	Chromosomal location	Location of MS repeat	Strand	Primer sequences	Size of PCR product (bp)
LIAS	(A)8	*4p14*	exon 2	+	F 5'-ATATTTTTGCAGCCCAGTCA-3'	160
					R 5'-TTCTCCTTTCTGGCGTTTTA-3'	
WHSC1	(C)8	*4p16.3*	exon 22	+	F 5'-TCCTACTGCTGTGAGCATGA-3'	139
					R 5'-GCTATTTGCCCTCTGTGACT-3'	
CYLD	(CAAA)4	*16q12.1*	exon 4	+	F 5'-CTACTGGGAAGAGCGGATTT-3'	165
					R 5'-TGATTTTTCTTGCCTTTTGC-3'	