Supplementary Appendix

Supplementary Methods

Exome Capture, Library preparation and sequencing

The qualified genomic DNA samples were first fragmented into about 200-300bp using a covaris sonication system (Covaris S2). Adapters were then ligated to both ends of the obtaining fragments. The library was constructed including ligation-mediated PCR (LM-PCR), purification, hybridizing with the NimbleGenEZ 44M human exome array. Non-hybridized fragments were washed out. Both non-captured and captured LM-PCR products were subjected to quantitative PCR to estimate the magnitude of enrichment. Each captured library was loaded on Illumina Hiseq 2000 platform, and high-throughput sequencing were performed for each library independently to ensure that each sample meet the desired average fold-coverage. Raw image files were processed by Illumina base calling Software V1.7 for base calling, with default parameters and the sequences of each individual were generated as 90bp paired-end reads.

Reads mapping and mutations detection

After removing reads containing sequencing adaptors and low-quality reads with more than five unknown bases, the high quality reads were aligned to the NCBI human reference genome (hg19) using BWA (Li *et al*. 2009) with the default options. PCR duplicates were removed by Picard (Koboldt *et al*. 2009). The BWA raw results were realigned by the Genome Analysis Toolkit (GATK IndelRealigner (Li *et al*. 2009)) to improve alignment performance.

SNPs were detected using SOAPsnp with default parameters. Somatic substitutions were first predicted with Varscan2 (Mckenna *et al*. 2010) (http://varscan.sourceforge.net/Samtools mpileup parameter was –Q 0 and Varscan2 parameters were --min-coverage 10 --min-coverage-normal 10

--min-coverage-tumor 10 --min-var-freq 0.1 --min-avg-qual 0), and confirmed if the following criteria were met: (1) Adjacent somatic mutation distance ≥ 10; (2) Map Quality score not significantly lower than other alleles (Map Quality score cutoff is 30.Fisher's exact test, $p < 0.20$) (3) Base Quality score not significantly lower than other alleles (Base Quality score cutoff is 20.Fisher's exact test, $p < 0.05$); (4) Mutant allele frequency change between tumour and adjacent normal (Fisher's exact test $p < 0.05$); (5) Mutations were not in gap aligned reads(In neither 20bp flank region less than 10 gap flag); (6) Mutant allele not significantly enriched within 10 bps of 5' or 3' ends of reads (Fisher's exact test, $p < 0.05$); (7) Mutations were not in simpleRepeatRegion (Wang *et al.* 2010) (Repeat events less than 6). Somatic InDels were predicted with GATK SomaticIndelDetector and confirmed with the following conditions: (1) Reads depth (DP) > 5 in both tumor and normal samples; (2) An average mismatch rate < 0.5 in both reference and mutated alleles supported reads; (3) Average mapping quality > 20 in both reference and mutated alleles supported reads in tumor; (4) Median indel offsets from reads terminal > 5bp. Mutations were annotated with ANNOVAR (Website: SimpleRepeatRegion).

Real-time PCR

Total RNA was extracted from micro-dissected samples using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) and cDNA synthesis performed from 1 μg of RNA using a reverse transcription system (Promega, Madison, Wis). Quantitative PCR analysis was performed by SYBR green method using Applied Biosystems 7300. The primers used are provided in Supplementary Table 9. Results were analyzed by the $\Delta\Delta$CT method using GAPDH for data normalization (Supplementary Table 10).

References

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L & Wilson RK. 2012 VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22** 568-576.

Li H & Durbin R 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25** 1754-1760.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K & Wang J. 2009 SNP detection for massively parallele whole-genome resequencing. *Genome Res* **19** 1124-1132.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*. 2010 The Genome Analysis Toolkit: A mapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20** 1297-1303.

SimpleRepeatRegion

(http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/simpleRepeat.txt.gz)

Wang K, Li M & Hakonarson H. 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38** e164