Supplemental Methods

Whole-exome sequencing analysis

Libraries were made with the SureSelect XT Human All Exon V5+UTR kits (Agilent) and sequenced on HiSeqX (Illumina). We used FastQC v0.11.5[1] and multiQC v1.2[2] to check the quality of the raw FASTQ files and cutadapter v1.5 to remove adapter sequences.[3] The bwa-mem v0.7.15[4] with default parameters was used to align to the hg19 reference genome.[5] PCR-duplicated reads were removed using Picard tools v2.9.0,[6] then alignments were recalibrated using GenomeAnalysisToolkit (GATK) v3.7[7] with known variant databases.[8-10]

We used MuTect v1.1.7[11] and Strelka2 v2.8.2[12] to detect somatic variants for tumor and matched normal samples. For MuTect, alignment files were realigned using GATK, then were inputted into MuTect with the dbSNP[8] and COSMIC databases.[13] Strelka2 was executed in somatic configuration with default parameters. We selected somatic variants that were detected by either MuTect or Strelka2 and annotated by ANNOVAR.[14] We further filtered variants that were not present within the exome capture kit. The variants present in the intronic or intergenic regions were also excluded. In addition, the variants with low variant allele frequency (VAF) with low quality supporting reads were filtered out. Specifically, variants with VAF ≤ 0.2 were excluded if they had less than 3 high quality supporting reads. We defined high quality supporting read as a read containing the mutated nucleotide with a base quality score of 30 or higher at the mutation as well as alignment quality score of 50 or higher given by the aligner for the entire read.

We used MutsigCV 1.4.1[15] to identify statistically significant recurrent somatic variations. Full exome coverage data from MutsigCV website was used for this test. We used *p* value ≤ 0.05 as a cutoff for determining the significance of the recurrent mutations. We used GenVisR to generate a waterfall plot to visualize a pattern of recurrent variations in our cohort.[16]

Germline variations were identified with GATK Haplotypecaller and Strelka2. For GATK Haplotypecaller, we used recommended parameters and a variant recalibration stage with

known variant databases.[8-10,17] Because we focused on few specific cancer related genes, we used the union of the two germline mutation profiles to increase the sensitivity, then we manually inspected potential germline variations using IGV.[18] We also annotated the germline mutations using ANNOVAR, then selected a subset of mutations as potentially pathogenic by the following two criteria: 1) predicted as deleterious by SIFT[19] and PolyPhen2[20] and 2) having a lower population frequency than 0.01 in gnomAD v2.1.1 (https://gnomad.broadinstitute.org).

RNA sequencing and analysis

Libraries were made with TruSeq RNA Access Library kit (Illumina) and sequenced on HiSeq2500 (Illumina). We used FastQC, multiQC, and cutadapt for quality control and preprocessing. STAR v2.4.2a[21] was used to align to a GRCh38_P01 reference with Genecode v22 gene annotation, then HTSeq 0.6.0[22] was used to generate counts for gene expression quantification. The same parameters used in the TCGA STAD project[23] were used. DESeq2[24] was used to detect differentially expressed genes (DEG).

Consensus clustering using gene expression data was performed with ConsensusClusterPlus.[25] The raw read counts were normalized using voom function in the limma package. Various k (number of clusters) from 2 to 20 were tested and k = 5 was selected based on Cophenetic correlation coefficient. We performed DEG analysis between the identified cluster 1 and clusters 2+3+4+5 and cluster 4 and clusters 1+2+3+5 using the DESeq2 and used GSEA for gene set enrichment analysis.[26]

FusionCatcher[27] and STAR-Fusion[28] were used with default parameters to detect gene fusion events. Fusion events identified by both algorithms were used for further analysis. Non-clipped raw FASTQ data and Ensembl v89 database[29] were used as input and database, respectively. We visualized and manually inspected fused transcripts using supporting reads provided by the FusionCatcher and UCSC genome browser.[30]

Ancestry analysis

To identify each sample's inherited genetic characteristics, we performed an ancestry inference using Locating Ancestry from SEquence Reads (LASER) to analyze whole-exome sequencing data[31] with default parameters. LASER constructs a reference principal component (PC) space with a set of reference individuals and places test samples into the PC space. Ancestry of each sample can be inferred using distances in the PC space between the sample and the reference individuals. We downloaded and used a reference PC space data from the LASER website. The reference PC space was constructed with Human Genome Diversity Project[32] data that contained 938 reference individuals from various ethnic groups. Then we calculated the first 4 PCs for each normal Hispanic/Latino sample and mapped it to the PC space.

Metagenomics for Epstein-Barr virus and Helicobacter pylori

We used PathoScope 2.0[33] with parameters (-b very-sensitive-local -m hi -k 100 -t 50 -L 101 -s 0.95 --adjreflen --reuse) to identify EBV infections using whole-exome and RNA sequencing data. We used the target microbial database (PathoDB) available from PathoScope 2.0 release, which was built from NCBI nr (non-redundant) nucleotide database as of 2014.[34] To increase sensitivity, we performed the metagenomics analysis on both WES data and RNA-seq data.

Determining microsatellite instability

We used MSISensor[35] with default parameters to predict the MSI status by calculating and comparing length distributions of microsatellites between tumor and normal sample. The MSISensor calculated a score for each sample to determine the MSI status (e.g., higher scores indicate the sample is more likely to have MSI). If we have both of normal and blood samples for

a patient, we did two tests and averaged the scores. Then we chose a cutoff 10 based on a pan-cancer MSI assessment study using the MSISensor.[36]

Determining somatic copy number alterations

We used CNVkit[37] to perform copy number alteration (CNA) analysis for whole-exome sequencing data. Because our cohort was sequenced by two different vendors (DNA Link, Inc (San Diego) and Admera Health (New Jersey)) we divided the cohort into two batches based on the vendors and ran the CNVkit separately. For each batch, a pooled reference panel was built using normal samples, then somatic CNAs were called for each tumor sample. The CNA calling results were merged then GISTIC2[38] was used to identify recurrent CNA regions. CNA regions with |CNA value| < 0.1 were filtered out. We adopted a method from Ichikawa et al.[39] and (number of CNA regions > 41) was defined as a cutoff to stratify between genomically stable (GS) and chromosomal instability (CIN) subtypes.

References for Supplemental Methods

1. Andrews S. FastQC: A quality control tool for high throughput sequence data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed 2018].
2. Ewels P, Magnusson M, Lundin S, et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32:3047–3048.
3. Martin M. *Cutadapt removes adapter sequences from highthroughput sequencing reads. EMBnet J 17: 10–12*. 2011.
4. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589–595.
5. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–1774.
6. The Broad Institute. Picard tools. httpsbroadinstitute.github.iopicard. Available at: https://broadinstitute.github.io/picard/ [Accessed 2018].
7. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–1303.

8. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–311.

9. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature 2015;526:68–74.

10. Mills RE, Pittard WS, Mullaney JM, et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. 2011;21:830–839.

11. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213–219.

12. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods 2018;15:591–594.

13. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–D783.

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

15. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–218.

16. Skidmore ZL, Wagner AH, Lesurf R, et al. GenVisR: Genomic Visualizations in R. Bioinformatics 2016;32:3012–3014.

17. International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–796.

18. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol 2011;29:24–26.

19. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–1081.

20. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat. Methods 2010;7:248–249.

21. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.

22. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166–169.

23. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513:202–209.

24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 2014;15:550.

25. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–1573.

26. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. 2005;102:15545–15550.

27. Nicorici D, Satalan M, Edgren H, et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv 2014:011650.

28. Haas B, Dobin A, Stransky N, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. bioRxiv 2017:120295.

29. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. Nucleic Acids Res. 2018;46:D754–D761.

30. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.

31. Wang C, Zhan X, Liang L, et al. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am. J. Hum. Genet. 2015;96:926–937.

32. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 2008;319:1100–1104.

33. Hong C, Manimaran S, Shen Y, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome 2014;2:33.

34. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2014;42:D7–17.

35. Niu B, Ye K, Zhang Q, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 2014;30:1015–1016.

36. Middha S, Zhang L, Nafa K, et al. Reliable Pan-Cancer Microsatellite Instability Assessment by Using Targeted Next-Generation Sequencing Data. JCO Precision Oncology 2017:1–17.

37. Talevich E, Shain AH, Botton T, et al. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comp Biol 2016;12:e1004873.

38. Mermel CH, Schumacher SE, Hill B, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology 2011;12:R41.

39. Ichikawa H, Nagahashi M, Shimada Y, et al. Actionable gene-based classification toward precision medicine in gastric cancer. Genome Medicine 2017;9:93.