**Supplementary methods**

**EBV variant analysis**

All RNA-sequencing libraries were screened for quality reads and PCR duplicates (See integration analysis). The reads were mapped using STAR aligner (95) index that was created for EBV genome (NC\_007605) from a revised annotation file for EBV genes (downloaded from NCBI Nucleotide database). The “genomeSAindexbases” parameter of STAR was set to 8. After screening for quality and PCR duplicates, all the remaining sequencing reads were aligned to the EBV genome using STAR aligner with the above index. The high quality and properly paired aligned reads were obtained using “samtools view” with parameters, “-F 12 –q 254”, and then piled up against the EBV genome using “samtools mpileup” with the following parameters, “–B –d 8000 –L 8000 –t AD –I” (96). The variants calling and filtering was performed using “bcftools” with the filtering options set to “AD[1]>5 && AD[1]>AD[0] && QUAL>50” to ensure that at least 5 high quality reads support the variant presence. The spurious variants (e.g. variants near junctions or variants that align using blat with equal or better quality somewhere else in the genome) were filtered. All the remaining variants were annotated using ANNOVAR (97). The mutational type and spectrum analysis were done using “MutationalPatterns” package in R (98). The lolliplots were generated using the “trackViewer” package from “Bioconductor” in R.

**Human variant analysis**

All RNA-sequencing libraries were screened for quality reads and PCR duplicates (See integration analysis). The reads were mapped to STAR indexed genome that was created using GENECODE 19 annotations and default parameters (95). The high quality (–q 254) and properly-paired mapped reads (–F 12) were obtained using “samtools view” and piled up against the human genome (hg19) using “samtools mpileup” command with following parameters, “–B –d 8000 –L 8000 –t AD –I –q 10” (99). The variant calling and filtering was done using “bcftools filter” with “AD[1]>10 && AD[1]>AD[0] && QUAL>50” to ensure that at least 10 high quality reads support the variant presence. All common variants in the human genome was screened against dbSNP (avsnp147) using ANNOVAR (97). The SNPs that were mapped to chrM were discarded. The remaining SNPs were further filtered if they were present in more than one of our control Blood or Lymph node samples. The spurious variants (e.g. variants near junctions or variants that align using blat with equal or better quality somewhere else in the genome) were filtered. The list of 198 expert curated genes were obtained from COSMIC database (64) and augmented with six frequently mutated genes in lymphoma (MYC, ID3, DDX3X, TCF3, BCL2, TRAF3). All the exonic mutations that fall into this list were considered for analysis.

**EBV integration analysis**

A chimeric genome index was created from human (hg19) and EBV (NC\_007605) genomes. The low-quality sequencing reads and those originating from potential PCR clones were removed (100). All RNA-sequencing libraries were mapped to the chimeric genome using Burrows–Wheeler Aligner with default parameters (101). The whole genome sequencing, cDNA, and direct RNA sequencing libraries for GM12878 nanopore sequencing data (42) were mapped using the MEM mode optimized for long sequencing reads and parameters '-x ont2d' and '-L [1,1]'. The chimeric reads that were partly mapped to chrM were discarded. An in-house script (code will be available upon request) was used to discard the chimeric reads with <90% alignment score and/or those that align to more than one genomic locus (with the same score or better). To increase the sensitivity, the remainder of the sequencing reads were aggregated into 1000bp non-overlapping blocks of the human genome. In other word, we tiled the human genome to non-overlapping 1000bp blocks and counted the chimeric reads that align anywhere in each block from any EBV+ sample. The blocks with more than 20 mapped reads were considered as EBV integration loci and annotated using HOMMER annotatePeaks software (97). The consecutive blocks were merged. We then examined each individual sample to examine if it contains EBV integration to the identified blocked by having at least two mapped reads to the block. The integration plot was drawn using circular visualization software, Circos (102).

**RNA-seq data analysis**

All the sequencing reads were screened for quality and PCR clones before estimating the mRNA expression levels. The human mRNA expression levels in each cancer sample were estimated by RSEM software (103) using “rsem-calculate-expression” with the following parameters, “--bowtie-n 1 –bowtie-m 100 –seed-length 28” (**Table S5J**). The RSEM required bowtie index was created by “rsem-prepare-reference” on all RefSeq genes downloaded from UCSC table browser on April 2017. The differentially expressed genes were identified using edgeR package (104). The EBV gene expression levels were estimated by RSEM using the STAR indexed EBV genome on a revised EBV gene annotation file (**Table S1D**).

**RNA-seq data sets**

We obtained 1051 RNA-seq samples across 15 different cancer types, 57 lymphoblastoid, and 234 control samples from publicly available databases (8,32-39). Detailed information, GEO numbers and Pubmed IDs are all included in **Table S1A**. Samples were obtained from diverse anatomical locations according to the site of the primary tumor (**Table S1A**). Tumor samples were collected and processed at different geographical locations based on local prevalence of the tumor type. For example, NPC samples were from patients diagnosed in Singapore, BL samples were collected from participants in Uganda, and NKTCL samples were from Chinese participants.

**Super-enhancer analysis**

The list of super enhancers (SEs) and their corresponding genes for the selected immune cell types and cancer cells in **Fig. 1E** were obtained (44,47). *De novo* SEs from H3K27ac in Raji cells (105) were called by findPeaks program from HOMER using “super” option (106). For each cell type in **Fig. 1E**, the SE containing genes and the genes with recurrent EBV integration events were overlapped and the significance of the overlap was calculated using fisher exact test based on all human coding genes from RefSeq. All the SE locations were merged and overlapped with EBV integration loci using “bedtools intersect” program (**Table S1C**). For **Fig. S6K,** all the super enhancer containing genes cells were pooled together to obtain a list of unique super enhancer containing genes. This list was then overlapped with differentially expressed genes in each cancer type and the significance of the overlap was calculated using fisher exact test.

**Biological pathway analysis**

The pathway analyses in **Fig. 6A, Fig. S1B,** and **Fig. S6B-I** were performed using MetaScape (67). All differentially expressed genes (**Table S5A-G**) between EBV+ and EBV– samples in each cancer type, were provided individually (**Fig. S6B-I**), or as a “Multiple Gene List” (**Fig. 6A**). The list of genes with recurrent EBV integration events (**Table S1C**) in their proximity was used to generate **Fig. S1C**. All other parameters were kept at default. To derive Type I and Type II interferon activity scores for each individual sample, we have first sourced a list of Type I interferon response genes (GO:0034340) and Type II interferon response genes (GO:0071346) from Gene Ontology database. We discarded the genes that did not have any experimental evidence according to the INTERFEROME database (Rusinova et al. 2013) (http://www.interferome.org) and kept only genes for which the majority of experiments reported >2-fold induction in response to interferon treatment. This resulted in 51 and 85 genes in Type I and Type II interferon pathways, respectively (**Table S5H**). For each tumor type, the expression of each gene was converted to a z-score by subtracting the mean value and dividing it by standard deviation across samples. The sum of z-scores for genes in each pathway were then used to denote the activity score of that pathway (**Table S5I**).

**Supplementary Table Legends.**

**Supplementary Table S1. A,** detail information on all 1342 cancer samples used in this study. **B,** parts per million (P.P.M.) values for each cancer along with the number of reads that mapped to the EBV genome and the library size after removing low-quality reads and PCR duplicates. **C**, 56 EBV-integrated loci in the human genome. For each integration (rows), shown are the integration location, the nearest gene and its hg19 coordinates, the number of samples with the integration in each tissue type, and the coordinates of overlapping SEs from genesets shown in **Fig. 1E**. The total number of EBV+ samples analyzed in each tumor type is indicated in the header row. **D**, for each sample (rows), shown are the EBV integration location. **E**, expression (transcripts per million; T.P.M.) of all EBV genes (rows) in each sample (columns).

**Supplementary Table S2.** List of all primers used in the study (see methods).

**Supplementary Table S3. A-B,** synonymous (**A**) and non-synonymous (**B**) variants for each EBV gene (rows) in each sample (columns). Missense and nonsense variants are denoted by “1” and “2” respectively. **C**, catalog of all nonsynonymous substitutions for each sample and each EBV gene. 1 and 0 denote the presence or absence of each mutation, respectively. **D**, for each nonsynonymous variant in each EBV gene, shown are the substitutions and the number of samples with the given variant. **E,** mutational signatures of the EBV transcriptome in each EBV+ cancer sample. Shown are the contribution from two mutational signatures, A and B, derived by nonnegative matrix factorization (NMF) for each EBV+ cancer sample.

**Supplementary Table S4. A,** mutational signatures of host transcriptomes in each EBV+ and EBV– cancer sample. Shown are the nonnegative matrix factorization (NMF) scores for mutation signatures A and B for each sample. **B,** frequent mutations in cancer driver genes. Shown are the number of samples with mutation in that gene identified from RNA-seq data across 15 cancer types (columns) for each gene (rows). The number of EBV+ samples with mutations are shown in parenthesis. The total number of samples in each tumor type is indicated in the header row. **C**, missense and nonsense mutations in 8 cancer driver genes in each sample and whether the observed mutation has previously annotated in the COSMIC database (last column).

**Supplementary Table S5.** list of DEGs (fold change >1.5, FDR<0.05) between EBV+ and EBV– samples in each cancer (**A-G**). For each gene, shown are the mean expression value across EBV– and EBV+ cancer samples. The log-fold change, *p*-value and FDR values were obtained using differential expression analysis in edgeR. **H**, genes involved in Type I and Type II interferon activity (see methods). **I**, interferon activity of EBV+ tumor types. Shown in columns are cancer types, classification into IFN+ or IFN– groups for EBV+ samples or control group for EBV– samples, EBV status, Type I and Type II interferon signature score, and z-scores of genes related to Type I and Type II interferon signatures, for each sample (rows). **J**, expression values (transcripts per million; T.P.M.) of all genes (rows) in each sample (columns).

**Table S2.** List of all primers used in the study (see methods).

|  |  |  |
| --- | --- | --- |
| EBV Genes (qPCR) | BZLF1-F  | 5' - AGG CCA GCT AAC TGC CTA TC - 3' |
| BZLF1-R | 5' - TGA TTC TGG GTT ATG TCG GA - 3' |
| BRLF1-F | 5' - ACA CTC CCG GCT GTA AAT TC - 3' |
| BRLF1-R | 5' - TGG CTT GGA AGA CTT TCT GA - 3' |
| BLLF1-F | 5' - TGG CGA GTT TGC GTC CTC AG - 3' |
| BLLF1-R | 5' - CGT CCA GTG TCA CGA TTT CTT GG - 3' |
| BcLF1-F | 5' - CCT CCC TGA CCG TTC CCA G - 3' |
| BcLF1-R | 5' - GCA GTT TGA GAC CGC CAC ATC - 3' |
| BGLF5-F | 5' - TTC GGC CGC TAT TAG CTT AG - 3' |
| BGLF5-R | 5' - GAC GGG GGA ATA ATC AAC CT - 3' |
| BSLF1-F | 5' - GCG GGT CCT CTG GAT TAG ATA GTC - 3' |
| BSLF1-R | 5' - CAG GGG CGG TGG TCT TAG C - 3' |
| BDLF1-F | 5' - GCA CCT CCT CTG CTA TGG GC - 3' |
| BDLF1-R | 5' - TGA TAC TCA CCA AGA TTG TTC CAG G - 3' |
| BMRF1-F | 5' - CGT GCC AAT CTT GAG GTT TT - 3' |
| BMRF1-R | 5' - CGG AGG CGT GGT TAA ATA AA - 3' |
| EBNA1-F | 5' - CCA CAA TGT CGT CTT ACA CC - 3' |
| EBNA1-R | 5' - ATA ACA GAC AAT GGA CTC CCT - 3' |
| LMP1-F | 5'- GCCCTTTGTATACTCCTACTGATG -3' |
| LMP1-R | 5'- AGACAAGTAAGCACCCGAAGAT -3' |
| BBLF4-F | 5' - AAG CCT GCC TCA TCC TTG ACC - 3' |
| BBLF4-R | 5' - GAC GAG CCT CTC CTT CAC GG - 3' |
| Site-directed Mutagenesis | LMP1\_Q322X F  | 5' - AGG CCC TCC AtA ATT GAC GGA - 3' |
| LMP1\_Q322X R  | 5' - CCA TCA TTT CCA GCA GAG TC - 3' |
| LMP1\_G342X F  | 5' - GAT GAC AGA CtG AGG CGG CGG TC - 3' |
| LMP1\_G342X R  | 5' - AAA GGC GGG CCC TGG TCA - 3' |
| Human Genes (qPCR) | PD-L1 - F  | 5' - ATGCCTTGGTGTAGCACTGA - 3' |
| PD-L1 - R | 5' - TGCTGGATTACGTCTCCTCC - 3' |
| UBC – F  | 5' - ATTTGGGTCGCGGTTCTTG - 3' |
| UBC – R | 5' - TGCCTTGACATTCTCGATGGT - 3' |
| EBNA1 guide RNAs | EBNA1sg1-F | ACACGAATGGTGTAAGACGACATTGG |
| EBNA1sg1-R | AAAACCAATGTCGTCTTACACCATTC |
| EBNA1sg2-F | ACACCGGGTGGTTTGGAAAGCATCGG |
| EBNA1sg2-R | AAAACCGATGCTTTCCAAACCACCCG |