Supplementary File 2

Single-cell RNA sequencing highlights the functional role of human endogenous retroviruses in gallbladder cancer
Supplementary Figure 1. Major cell type identification in GBC and adjacent normal tissues.

a tSNE projection of 28,301 single cells from tumor and adjacent normal tissues, color-coded by clusters (left), tissues (middle), and samples (right).

b tSNE plot displaying the expression of canonical marker genes for major cell types.

c Cell number (left, middle) and relative proportion (right) of major cell types in different tissue types and samples.

d tSNE plot displaying the expression of HERV families in tumor and adjacent normal tissues.
Supplementary Figure 2. Experimental verification of HERV-derived enhancer activity.

a Overall expression levels of HERVs in major cell types.
b Mean HERV expression level across all cells for a given cell type.
c Significantly upregulated HERV loci identified in GBC-derived T cells (left) and DCs (right).
Supplementary Figure 3. Epithelial cell subtypes.

a tSNE projection of epithelial cells, color-coded by clusters (left), tissues (middle), and samples (right).

b Heatmap showing large-scale CNVs for each pair of samples.

c tSNE plot displaying the CNV score for epithelial cells (left) and boxplot showing the CNV score for epithelial cell clusters (right).

d Violin plot displaying the expression of selected genes in two non-malignant subtypes.

e Cell number (left, middle) and relative proportion (right) of epithelial cell subtypes in different tissue types and samples.

f HERV families upregulated in Epi_EDN1(high) cells compared to Epi_EDN1(low) cells.

g Pie chart displaying the percentage of the upregulated HERVH loci.

h Differentiation trajectory of malignant and normal epithelial cells inferred by Monocle2, color-coded by pseudotime (left) and states (right).

i Top 10 GO terms for genes correlated with upregulated HERV families.

j Scatter plot showing correlation between the indicated gene and HERV (top) and expression of this gene depicted in violin plot (bottom).
Supplementary Figure 4. T cell subtypes.

a tSNE projection of T cells, color-coded by clusters (left), tissues (middle), and samples (right).
b tSNE plot displaying the expression of canonical marker genes for T cell subtypes. Clusters (cluster 10, 11, and 12) that did not express T cell marker genes (CD3D, CD3E, CD3G and CD2) were excluded from downstream analyses due to ambiguity in their identity. Cluster 0 expressed T cell gene markers but hardly expressed CD4 and CD8A and was thought to represent CD4-/CD8- T cells.

c Cell number (left, middle) and relative proportion (right) of T cell subtypes in different tissue types and samples.

d Top 10 GO terms for genes correlated with upregulated HERV families.
Supplementary Figure 5. B cell subtypes.

a tSNE projection of B cells, color-coded by clusters (left), tissues (middle), and samples (right).

b tSNE plot displaying the expression of canonical marker genes for B cell subtypes.

c Cell number (left, middle) and relative proportion (right) of B cell subtypes in different tissue types and samples.
Supplementary Figure 6. Myeloid cell subtypes.
a tSNE projection of DCs, color-coded by clusters (left), tissues (middle), and samples (right).
b tSNE plot displaying the expression of canonical marker genes for DC subtypes.
c Cell number (left, middle) and relative proportion (right) of DC subtypes in different tissue types and samples.
d Top 10 GO terms for genes correlated with upregulated HERV families.

e Scatter plot showing correlation between gene MMP7 and HERV LTR7Y in the indicated cell populations (left) and expression of gene MMP7 depicted in violin plot (right).
**Supplementary Figure 7. Fibroblast subtypes.**

a tSNE projection of fibroblasts, color-coded by clusters (left), tissues (middle), and samples (right).

b tSNE plot displaying the expression of canonical marker genes for fibroblast subtypes.

c Cell number (left, middle) and relative proportion (right) of fibroblast subtypes in different tissue types and samples.

d Top 10 GO terms for genes correlated with upregulated MER65-int.
Supplementary Table 1. QC metrics used to select cells.

| sample | nfeature.min | nfeature.max | mitochondrial contamination |
|--------|--------------|--------------|-----------------------------|
| N1     | 200          | 5000         | 30%                         |
| T1     | 200          | 6000         | 30%                         |
| N2     | 200          | 6000         | 30%                         |
| T2     | 200          | 6000         | 30%                         |
| N3     | 200          | 5000         | 30%                         |
| T3     | 200          | 6000         | 30%                         |
| N4     | 200          | 8000         | 30%                         |
| T4     | 200          | 8000         | 30%                         |
### Supplementary Table 2. Clinical information and single-cell RNA sequencing data.

| Patient name | Sample name | Gender | Age (year) | Tumor subtype | Sample collection | Number of cells | Total filtered reads | Reads mapped ambiguously (%) (gene | HERV) |
|--------------|-------------|--------|------------|---------------|-------------------|-----------------|--------------------|-----------------------|--------|
| P1           | N1          | female | 64         | adenocarcinoma | adjacent normal tissue | 6,491           | 265,093,671         | 2.29 | 6.07 |
| P1           | T1          | female | 64         | adenocarcinoma | tumor tissue       | 3,908           | 289,106,850         | 2.14 | 5.61 |
| P2           | N2          | male   | 56         | adenocarcinoma | adjacent normal tissue | 2,581           | 667,559,625         | 2.33 | 6.94 |
| P2           | T2          | male   | 56         | adenocarcinoma | tumor tissue       | 2,568           | 651,349,051         | 2.53 | 6.48 |
| P3           | N3          | female | 72         | adenocarcinoma | adjacent normal tissue | 5,000           | 465,915,248         | 2.39 | 6.52 |
| P3           | T3          | female | 72         | adenocarcinoma | tumor tissue       | 2,999           | 608,028,773         | 2.67 | 7.17 |
| P4           | N4          | male   | 54         | adenocarcinoma | adjacent normal tissue | 2,445           | 419,464,683         | 3.2  | 6.62 |
| P4           | T4          | male   | 54         | adenocarcinoma | tumor tissue       | 2,309           | 441,603,902         | 3.35 | 7.45 |
## Supplementary Table 3. Primer sequences for nested PCR.

| Primer Name | Primer sequence |
|-------------|-----------------|
| MER4A1-dup6-chr1-1-F | 5'- TTCAGGATACCTGAGCCCTGCTG -3' |
| MER4A1-dup6-chr1-1-R | 5'- TGCCCCACAGGTCTTTGTTATCATG -3' |
| MER4A1-dup6-chr1-2-F | 5'- CATTTCCTCTATCGATAGGTACTGTAACCCAAAATATCTC -3' |
| MER4A1-dup6-chr1-2-R | 5'- GCACCGCTAGAAAGGCTCGGTACTTCCTCTCTTCTCAC -3' |
| MER11B-dup23-chr5-1-F | 5'- TAAGAGGTGTAGTCTTGGTAGAAAGTGC -3' |
| MER11B-dup23-chr5-1-R | 5'- AAAGGCTATTTGCACCCATAGCTCAG -3' |
| MER11B-dup23-chr5-2-F | 5'- CATTTCCTATCGATAGGTAACGACCATG -3' |
| MER11B-dup23-chr5-2-R | 5'- GCACCGCTAGAAAGGCTCGGTACTTTTATG -3' |
| LTR13A-dup8-chr11-1-F | 5'- TCTCCACCTCCGCAATGGAAGCT -3' |
| LTR13A-dup8-chr11-1-R | 5'- TTCAGCATAACAGCAAGCTATAGTCTGAGTAC -3' |
| LTR13A-dup8-chr11-2-F | 5'- GAAGATCTCTCTCTCTTTTAAAATAGCTG -3' |
| LTR13A-dup8-chr11-2-R | 5'- CAAGGCTCGGTACTTCCTCTGTGAC -3' |
| LTR33-dup38-chr16-1-F | 5'- TGACCTCAAGTAGTAATTCCTCTTCTCCTTG -3' |
| LTR33-dup38-chr16-1-R | 5'- GGCTCTTCTCTCTGGTGTACCTTTTAC -3' |
| LTR33-dup38-chr16-2-F | 5'- CATTTCCTATCGATAGGTACTTTTTTCACTG -3' |
| LTR33-dup38-chr16-2-R | 5'- TAAGAGGTGTAGTCTTGGTAGAAAGTGC -3' |
| MLT1F-int-dup5-chr16-1-F | 5'- CATTTCCTATCGATAGGTACTCTCTCTGTCAAG -3' |
| MLT1F-int-dup5-chr16-1-R | 5'- CATTTCCTATCGATAGGTACTCTCTCTGTCAAG -3' |
| MLT1F-int-dup5-chr16-2-R | 5'- CATTTCCTATCGATAGGTACTCTCTCTGTCAAG -3' |
| LTR2-dup7-chr16-1-F | 5'- CTTCTACTATCGATAGGTACTCTCTCTGTCAAG -3' |
| LTR2-dup7-chr16-1-R | 5'- CTCTACTATCGATAGGTACTCTCTCTGTCAAG -3' |
| LTR10C-dup13-chr19-1-F | 5'- GTAAGGAAAGAAGTAGGATGCCTCTCTCATCGAG -3' |
| LTR10C-dup13-chr19-1-R | 5'- CAAGGCTAGACTGCTGTACCTATTATTATGTTATG -3' |
| LTR10C-dup13-chr19-2-F | 5'- GAAGGCTAGACTGCTGTACCTATTATTATGTTATG -3' |
| LTR10C-dup13-chr19-2-R | 5'- CATTTCCTATCGATAGGTACTCTCTCTGTCAAG -3' |
| MSTD-dup87-chr20-1-F | 5'- GCACCGCTAGAAAGGCTCGGTACTTCCTCTCTTTT -3' |
| MSTD-dup87-chr20-1-R | 5'- GCACCGCTAGAAAGGCTCGGTACTTCCTCTCTTTT -3' |
| MSTD-dup87-chr20-2-F | 5'- GCACCGCTAGAAAGGCTCGGTACTTCCTCTCTTTT -3' |
| MSTD-dup87-chr20-2-R | 5'- GCACCGCTAGAAAGGCTCGGTACTTCCTCTCTTTT -3' |
Supplementary Table 4. HERVs with enhancer activity and their neighboring DEGs.

| HERV with enhancer activity | Location | Neighboring DEGs |
|----------------------------|----------|------------------|
| MER4A1--dup6-chr1          | chr1: 22982843-22983338 | C1QA, C1QB, KDM1A |
| MER11B-dup23-chr5          | chr5: 151729381-151730458 | ATOX1, G3BP1 |
| LTR13A-dup8-chr11          | chr11: 65521819-65522784 | NEAT1 |
| LTR33-dup38-chr16          | chr16: 18936807-18937237 | ARL6IP1 |
| MLT1F-int-dup5-chr16       | chr16: 29915600-29916253 | ASPHD1, INO80E, KCTD13, MVP, TAOK2, TMEM219, YPEL3 |
| LTR2-dup7-chr16            | chr16: 30541069-30541510 | CD2BP2, DCTPP1, FBRS, ORAI3, PRR14, RNF40, SEPHS2, STX4, ZNF688, ZNF689 |
| LTR10C-dup13-chr19         | chr19: 15837284-15837348 | UCA1 |
| MSTD-dup87-chr20           | chr20: 49769262-49769601 | B4GALT5, RNF114 |