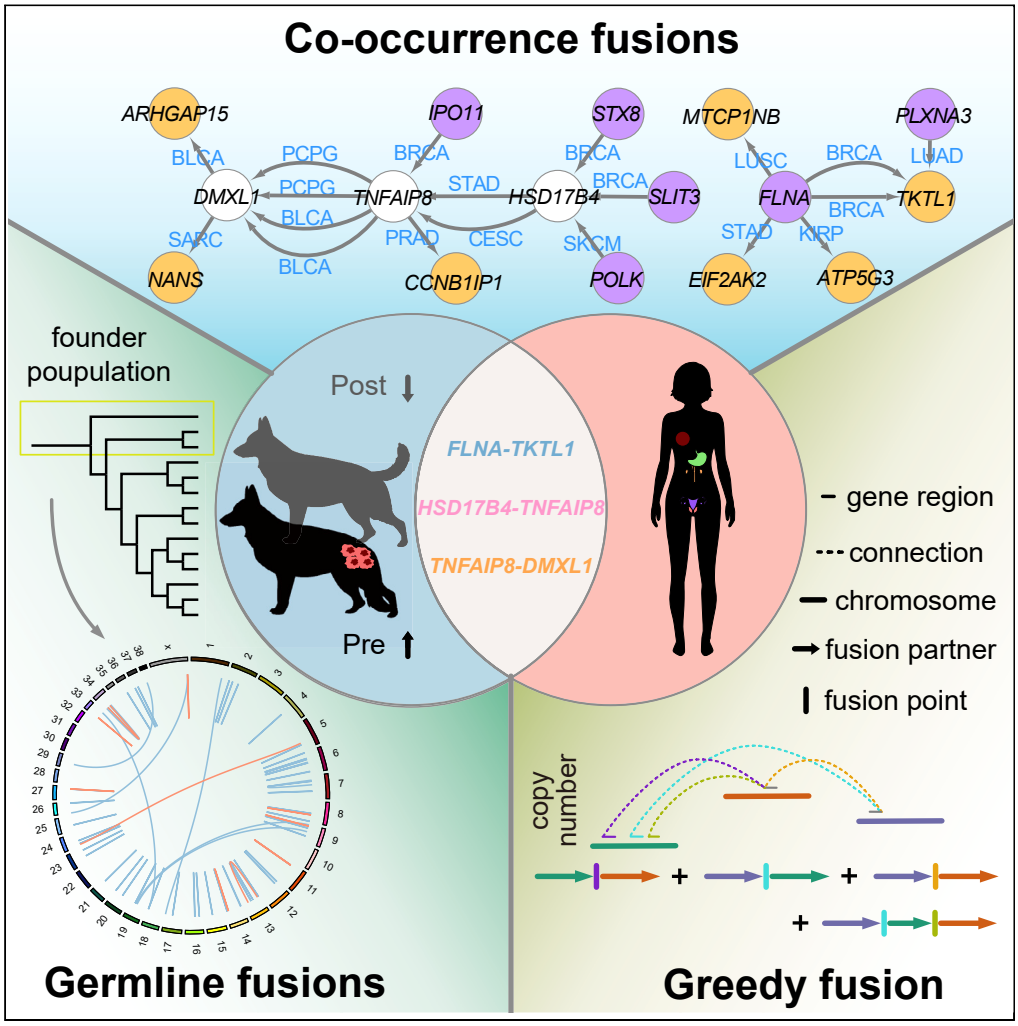
iScience

CellPress
OPEN ACCESS

Article

Germline gene fusions across species reveal the chromosomal instability regions and cancer susceptibility



Bo-Wen Zhou, Qing-Qin Wu, David H. Mauki, ..., Yan-Hu Liu, Guo-Dong Wang, Ya-Ping Zhang

liuyanhu@mail.kiz.ac.cn (Y.-H.L.)
wanggd@mail.kiz.ac.cn (G.-D.W.)
zhangyp@mail.kiz.ac.cn (Y.-P.Z.)

Highlights

The landscape of germline gene fusions in CTVT has been comprehensively identified

Fusion patterns are convergent across different tumor types

Greedy fusion is a new class of chromosomal structural alterations

Changes in germline gene fusions expression are associated with DNA methylation levels

Zhou et al., iScience 26, 108431
December 15, 2023
https://doi.org/10.1016/j.isci.2023.108431

iScience

CellPress
OPEN ACCESS

Article

Germline gene fusions across species reveal the chromosomal instability regions and cancer susceptibility

Bo-Wen Zhou,[1,2,7] Qing-Qin Wu,[3,7] David H. Mauki,[4] Xuan Wang,[1,2] Shu-Run Zhang,[1] Ting-Ting Yin,[1] Fang-Liang Chen,[5] Chao Li,[6] Yan-Hu Liu,[1,*] Guo-Dong Wang,[1,2,*] and Ya-Ping Zhang[1,2,8,*]

SUMMARY

The canine transmissible venereal tumor (CTVT) is a clonal cell-mediated cancer with a long evolutionary history and extensive karyotype rearrangements in its genome. However, little is known about its genetic similarity to human tumors. Here, using multi-omics data we identified 11 germline gene fusions (GGFs) in CTVT, which showed higher genetic susceptibility than others. Additionally, we illustrate a mechanism of a complex gene fusion of three gene segments (*HSD17B4-DMXL1-TNFAIP8*) that we refer to "greedy fusion". Our findings also provided evidence that expressions of GGFs are downregulated during the tumor regressive phase, which is associated with DNA methylation level. This study presents a comprehensive landscape of gene fusions (GFs) in CTVT, which offers a valuable genetic resource for exploring potential genetic mechanisms underlying the development of cancers in both dogs and humans.

INTRODUCTION

Over the past few decades, it has been discovered that a substantial fraction of cancers can be genetically inherited and their germline associated cancerous genes can be termed cancer predisposition genes (CPGs) since they possess genetic risk for cancer development.[1–3] It is estimated that around 3%–8% of all cancers are caused by CPGs,[2,4,5] but of course, these data are probably underestimated due to the following reasons. First, while long-term follow-up of familial cancer clusters is a tiring and daunting process,[3] it has however been discovered that the susceptibility or predisposition variants are present not only in children but also in adults.[4] Second, the definition of CPG is not well defined, and the more classical hypothesis currently accepted is the two-hit hypothesis, in which one allele is mutated in the germline and the other allele is mutated in the somatic.[6] Finally, screening of cancer therapeutic genes is currently focused on somatic mutations[7] and thus leaving prospective germline variant analysis enigmatic,[8–10] which our study intends to address.

Canine transmissible venereal tumor (CTVT) is an ancient cancer believed to have occurred thousands of years ago and is known to infect canines through allogeneic transfer of living tumor cells.[11,12] Due to its extraordinary evolutionary history, mutations originating from the "founder animal" are considered germline variants since they can be transferred from thousands of generations to the present day. Nonetheless, CTVT genome has been described as having a hybrid of canine (*Canis latrans*) introgression yet with stable genomic constitution.[13,14] In other words, it has accumulated more mutations than any known human tumor resulting in large-scale structural variation (SV) which has affected more than 2,000 genes,[13,15,16] including those related to clonal rearrangements in the *MYC* (MYC Proto-Oncogene, BHLH Transcription Factor),[17] homozygous deletion of suppressor gene *CDKN2A* (Cyclin Dependent Kinase Inhibitor 2A), and homozygous loss of DNA repair gene *SETD2* (SET Domain Containing 2).[13] Therefore, tracking germline variants in CTVT has been very challenging over the years; however, they provide useful insights regarding evolutionary origin and effects of germline variants on tumors.

To present day, there has not been any study that has employed CTVT as a model to study cancer in other species, especially humans, through germline variants, and thus to do so their correspondence must be well determined. Recent studies have revealed that there is a genetic convergence of tumor variations across different species.[18] Although current efforts to study CTVT have revealed certain features

[1]State Key Laboratory of Genetic Resources and Evolution & Yunnan Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650201, China
[2]Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming, Yunnan 650204, China
[3]School of Ecology and Environmental Sciences, Yunnan University, Kunming, Yunnan 650500, China
[4]Institute of Neurological Disease, National-Local Joint Engineering Research Center of Translational Medicine, State Key Lab of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China
[5]Kunming Police Dog Base of the Ministry of Public Security, Kunming, Yunnan 650204, China
[6]State Key Laboratory for Conservation and Utilization of Bio-Resource, Yunnan University, Kunming, Yunnan 650500, China
[7]These authors contributed equally
[8]Lead contact
*Correspondence: liuyanhu@mail.kiz.ac.cn (Y.-H.L.), wanggd@mail.kiz.ac.cn (G.-D.W.), zhangyp@mail.kiz.ac.cn (Y.-P.Z.)
https://doi.org/10.1016/j.isci.2023.108431

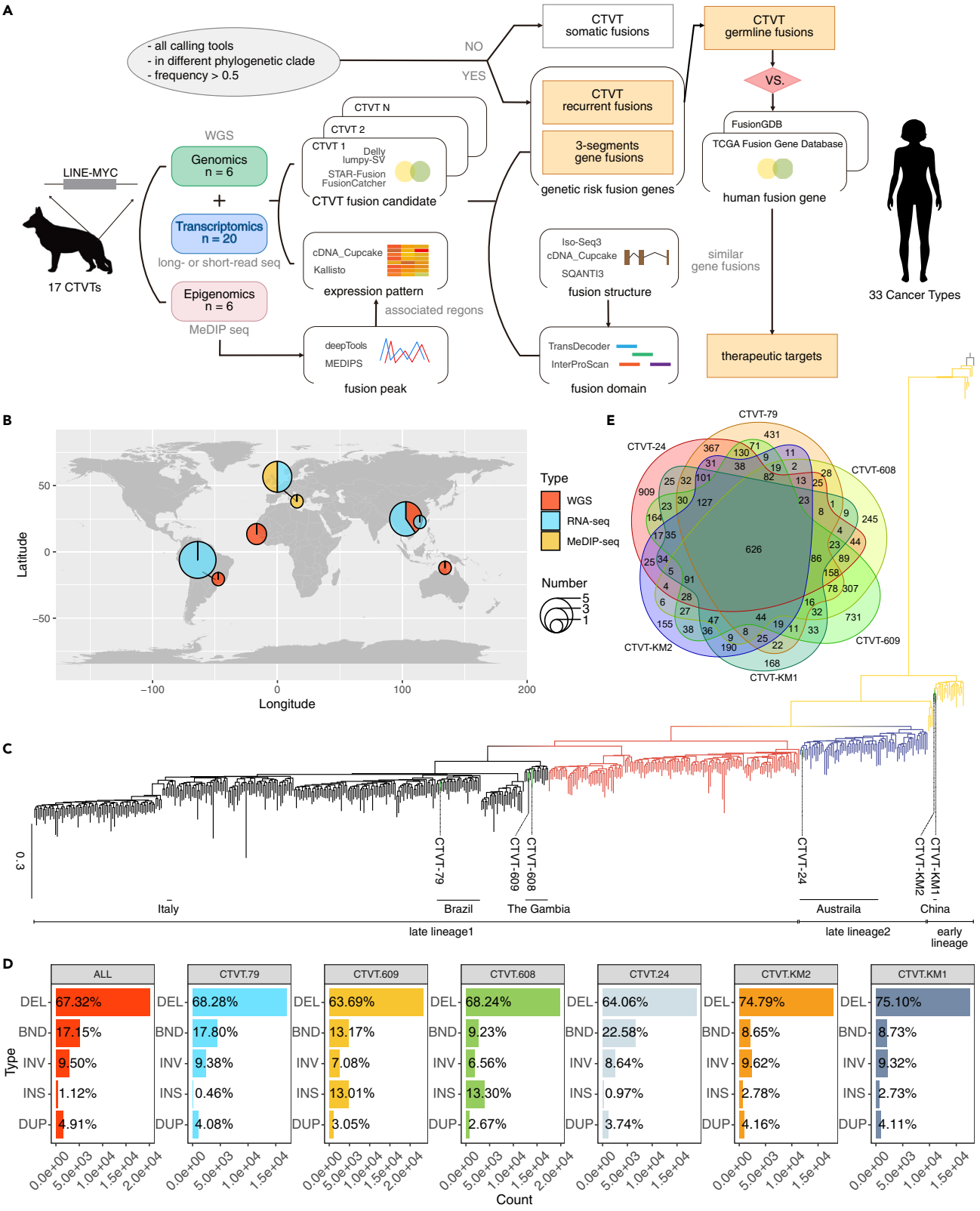

A
- all calling tools
- in different phylogenetic clade
- frequency > 0.5
NO
YES
CTVT somatic fusions
CTVT germline fusions
CTVT recurrent fusions
3-segments gene fusions
genetic risk fusion genes
VS.
FusionGDB
TCGA Fusion Gene Database
human fusion gene
33 Cancer Types
LINE-MYC
17 CTVTs
WGS
Genomics n = 6
+
Transcriptomics n = 20
long- or short-read seq
Epigenomics n = 6
MeDIP seq
CTVT N
CTVT 2
CTVT 1
Delly
lumpy-SV
STAR-Fusion
FusionCatcher
CTVT fusion candidate
cDNA_Cupcake
Kallisto
expression pattern
associated regons
deepTools
MEDIPS
fusion peak
Iso-Seq3
cDNA_Cupcake
SQANTI3
fusion structure
similar gene fusions
TransDecoder
InterProScan
fusion domain
therapeutic targets
B
Latitude
Longitude
Type
WGS
RNA-seq
MeDIP-seq
Number
E
CTVT-79
CTVT-24
CTVT-608
CTVT-609
CTVT-KM2
CTVT-KM1
626
C
0.3
CTVT-79
CTVT-609
CTVT-608
CTVT-24
CTVT-KM2
CTVT-KM1
Italy
Brazil
The Gambia
Austraila
China
late lineage1
late lineage2
early lineage
D
ALL
CTVT.79
CTVT.609
CTVT.608
CTVT.24
CTVT.KM2
CTVT.KM1
Type
DEL
BND
INV
INS
DUP
67.32%
17.15%
9.50%
1.12%
4.91%
68.28%
17.80%
9.38%
0.46%
4.08%
63.69%
13.17%
7.08%
13.01%
3.05%
68.24%
9.23%
6.56%
13.30%
2.67%
64.06%
22.58%
8.64%
0.97%
3.74%
74.79%
8.65%
9.62%
2.78%
4.16%
75.10%
8.73%
9.32%
2.73%
4.11%
Count

Figure 1. Project pipeline overview and WGS-based structural variations (SVs)

(A) Biopsies were sequenced and analyzed according to the schematic diagram after detecting CTVT-specific LINE–*MYC* genomic rearrangement. We begin by calling gene fusions per CTVT sample with multi-omics approach then implement filtering to produce the final CTVT recurrent fusions set with the desired balance of precision and sensitivity. Finally, we performed a cross-species analysis with CTVTs and TCGA tumors to find therapeutic targets.

(B) Geographic distribution of CTVTs. The number of samples from each region is indicated by the pie size, and the proportion of multi-omics from each region is shown in the pie chart (red: Whole-Genome Sequencing; blue: Transcriptome Sequencing; yellow: DNA Methylation Sequencing).

(C) The time-resolved phylogenetic tree of 543 CTVT, which is divided into three super clades: early lineage, late-lineage1, and late-lineage2.

(D) Number of different types of SVs in CTVTs (DUP: duplication, INS: insertion, INV: inversion, TRA: translocation, DEL: deletion), and the proportion of each type of structural variant is shown on the bar.

(E) Venn diagram of WGS-based fusion events. See also Table S1.

in the genome that are very similar to those in the Catalog of Somatic Mutations in Cancer (COSMIC),[16] many RNA-driven alterations such as fusions, splicing, alternative promoters, single-nucleotide variants (SNVs), and RNA-editing may have been overlooked. We thus deployed some of these driver mutations in our study to investigate the similarities of CTVT germline gene fusions (GGFs) to human tumors. These types of alterations which bear striking resemblance to mutations in human cancers are known for their diagnostic and therapeutic significance. Some of the gene fusions (GFs) shared between humans and dogs include *IGK-CCND3* in B cell lymphoma, *MPB-BRAF* in glioma, and *COL3A1-PDGFB* in dermatofibrosarcoma protuberans-like.[19] Intriguingly, the number of GFs discovered to date has increased due to their importance in clinical diagnosis and prognosis.[20,21] In particular, GFs involving kinases are of greater interest in many approved drugs with some of them currently undergoing clinical investigation.[20]

Here, using high-quality short- or long-read RNA sequencing (RNA-seq) on four fresh CTVT samples from China integrated with multi-omics data from other projects, we comprehensively investigated the landscape of GFs in CTVTs. We identified six GGFs that play a role in tumor development, including a GF of 3-segment genes, and provide a description of its mechanism. This demonstrates unprecedented genetic resource. Our study also conducted a comparative analysis where we investigated for the first time the fusion points and proteins of CTVT in relation to other human tumors and constructed a network to explore their relationships. Furthermore, we found similarities in mutated genes, domains, and pathways across species, implying evolutionary conservation of tumor development and the potential for mapping CTVT to other human tumors for the identification of shared susceptibility loci. Moreover, these GGFs were not only significantly upregulated during tumor progression but also demonstrated strong correlations with DNA methylation level.

RESULTS

CTVTs from southern China belong to the founder population

To characterize the GGFs of CTVT founder population, we collected four naturally occurring CTVT samples (designated CTVT-KM1, CTVT-KM2, CTVT-KM3, and CTVT-SZ3) in southern China and performed transcriptome sequencing analysis. About 49.51, 69.38, 70.27, and 69.69 million short reads were obtained on CTVT-KM1, CTVT-KM2, CTVT-KM3, and CTVT-SZ3, respectively (Figure S1A). The sequencing depth of newly collected samples was saturated for junction and expression analysis, as the number of "known junctions" reached a plateau (Figure S1B). A total of thirteen published CTVTs from Italy (CTVT-5, CTVT-6, and CTVT-17),[22] Brazil (CTVT-761, CTVT-765, CTVT-766, CTVT-772, CTVT-774, CTVT-775, and CTVT-79),[13,22] Australia (CTVT-24),[13] and The Gambia (CTVT-608 and CTVT-609)[23] were also added to our analyses (Table S1). We also integrated longitudinal CTVTs multi-omics data (genome, exome, transcriptome, and epigenome) from five continents (Asia, Europe, South America, Oceania, and Africa) with the goal of generating high-confidence GGFs for comparison with 33 different types of human cancer (Figures 1A and 1B).

It is now known that CTVT shares some similarities with highly rearranged chromosomes.[13,15] To identify shared CTVT genomic rearrangements branching trajectories in the CTVT phylogeny, we analyzed publicly available whole-genome sequencing (WGS) data to infer tumor purity, ploidy and SNVs, and copy number variation (CNV) of CTVT,[13,14,23] which is consistent with result form previously published works that CTVT is near diploid.[13] CTVT-KM1, -KM2, -24, -608, and -609 exhibited a ploidy of 2, while -79 had a ploidy of 1.9. Finally, we obtained clock-like 35,468 SNVs from 543 CTVTs in 44 countries using the methods provided by Adrian Baez-Ortega et al.,[16] to construct a comprehensive phylogenetic tree for the CTVT lineage. The topology of the maximum likelihood (ML) tree was found to be highly concordant with the previous exome results.[16] We introduced for the first time CTVTs from southern China and found that they are in the most basal clade (early lineage) in the phylogenetic evolutionary tree (Figure 1C). Our findings suggest that starting ~1,000 years ago, CTVTs in this region may had begun to escape from the founding population to other habitats. The remaining samples represent two sublineages of CTVT with distinct geographic clustering, consistent with previous findings.[16] The Australian lineage represents late sublineage2, while the Gambian and Brazilian lineages represent late sublineage1 (Figure 1C). The CTVT RNA sequences were not included in this ML tree due to their low coverage. However, we only deduced the position of these RNA samples in the phylogenetic tree based on information of their sampling locations. Once the CTVT lineage is established, it is highly likely to remain in situ.[16]

Accurately identifying SVs is critical for genome interpretation.[24] We developed a pipeline to output all types of CTVT SVs and inferred WGS-based GFs for CTVT using the diploid model (see "STAR methods"). Deletion (DEL) was the predominant type of SV in CTVT (Figure 1D). We found that WGS-based GFs for CTVT represent only a small fraction of all SVs, with approximately 2,000–3,000 detectable per sample. Interestingly, these three different lineages still shared up to 626 fusions (Figure 1E), even though they differed in age, sex, breeds, and environment, which suggest that fusions often arise from pre-infection clonal outgrowths.

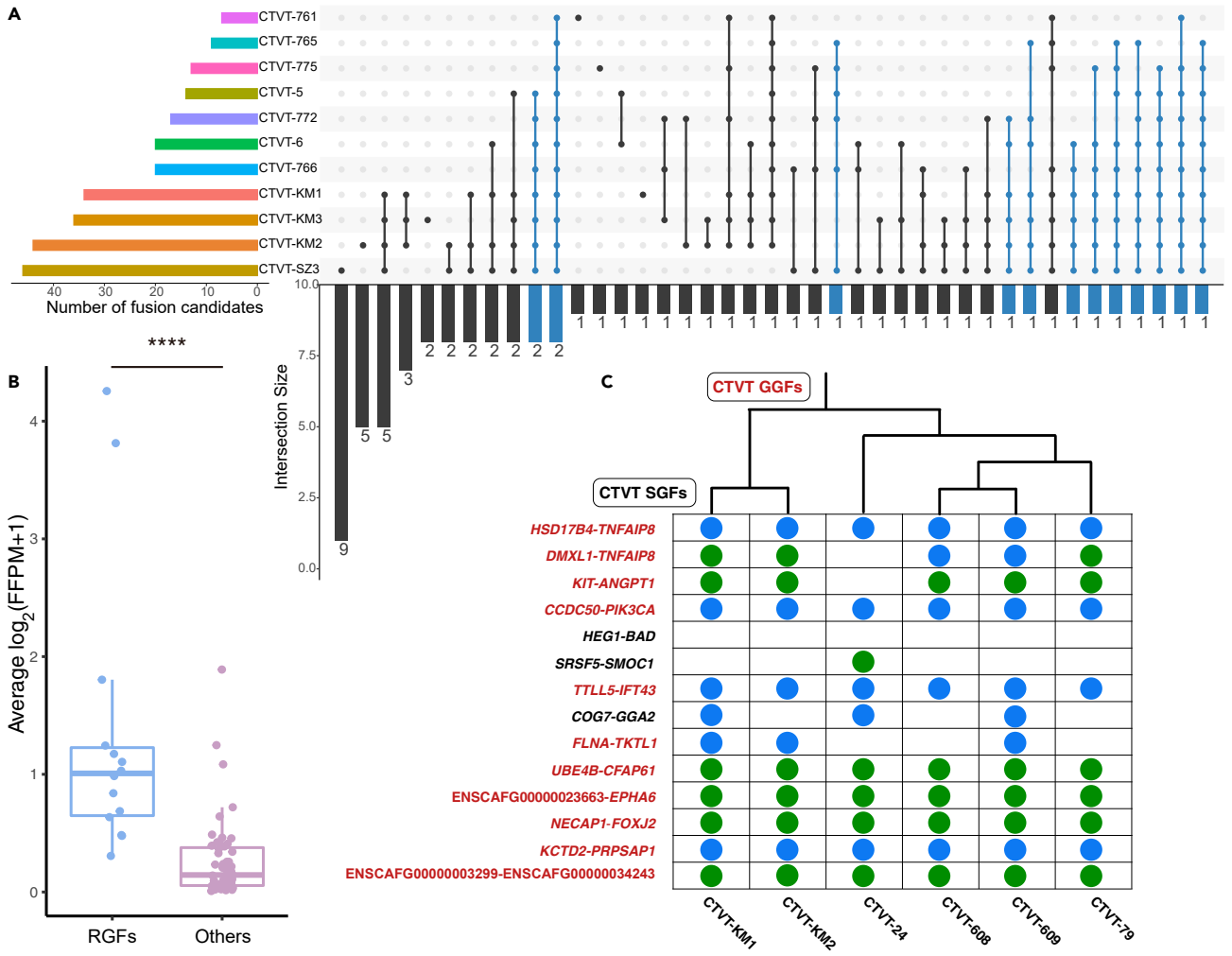


Figure 2. Detection of CTVT gene fusions (GFs)

(A) UpSet plot showing the overlapping fusion events for CTVTs. Recurrent gene fusions (RGFs) are marked in blue; others are non-RGFs, which are marked in black.

(B) Comparison of average FFPM (fusion fragments per million total reads) from the two gene fusion calling tools between RGFs and other gene fusions. **** denotes p value < 0.0001.

(C) Evidence of RGFs at RNA and DNA level. Blue dots indicate fusion point in the coding region, whereas green indicates fusion point in the intergenic region. Otherwise, the fusion is not supported by DNA level. The germline gene fusions (GGFs) are colored in red. See also Figures S1 and S2; Tables S2–S4.

GGFs in CTVT contribute tumor growth

To improve the accuracy of fusions detection, we also combined transcriptomic data for further detection of GGFs. Among the 12 RNA-seq CTVTs from primary biopsies (4 from newly sequenced CTVT samples in this study and 8 from publicly available data), a total of 65 GF candidates were detected (Figure S2A; Table S2), and we defined CTVT recurrent fusion genes (RGFs) as those with frequency greater than 0.5 and shared across different phylogenetic clades (see "STAR methods"). The features of these RGFs include GGFs, somatic gene fusions (SGFs), and background gene fusions (BGFs). Next, we utilized other publicly available transcriptomes from four spontaneous non-transmissible canine tumors (Histiocytic sarcoma, Dermatofibrosarcoma protuberans-like, Anaplastic oligodendroglioma, and Lymphomas)[19,25] from different tissues and embryonic origins to construct BGFs dataset (Table S3). Subsequently, we compared this dataset with the CTVT RGFs from the 65 candidates described earlier. Interestingly, we found 14 RGFs in CTVT, out of which 11 commonly overlapped in WGS-based data (Figures 2A and S2C; Table S4). This suggests that these RGFs possibly contribute more to early developmental stages of CTVT. To avoid any false positives in bioinformatics computations when detecting RGFs, we measured FFPM (fusion fragments per million total reads),[26] a method which is usually used for normalizing fusion-supporting RNA-seq fragments. We observed that the average mean of reads as measured by FFPM was significantly higher in CTVT RGFs when compared with the other GFs (p value = 5.10×10^{-6}, Mann-Whitney U test, two-sided), reflecting the confidence of our computations in the actual detection of RGFs in our samples (Figure 2B).

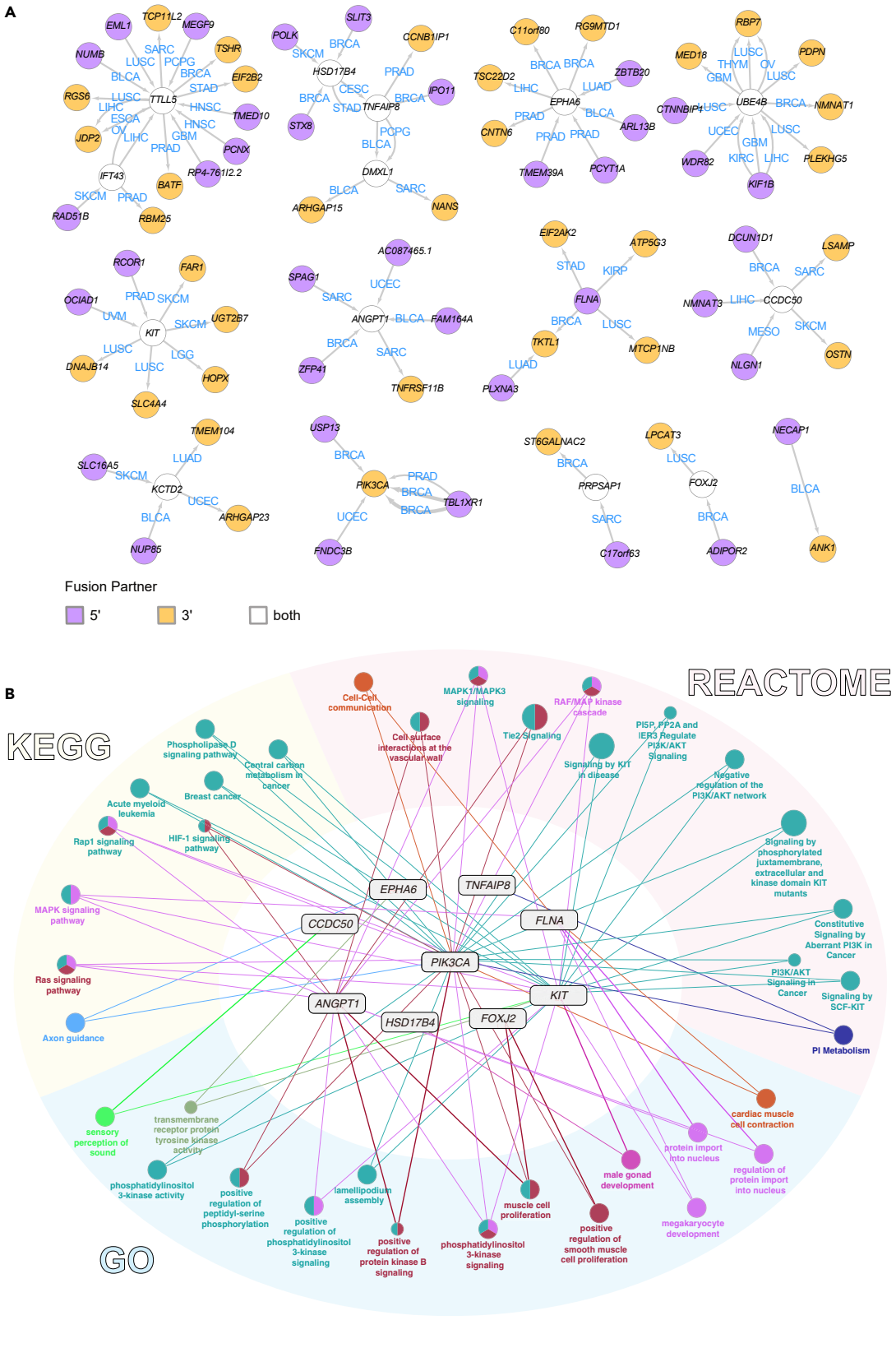

A
EML1
TCP11L2
MEGF9
NUMB
SARC
LUSC
PCPG
TSHR
BLCA
BRCA
EIF2B2
RGS6
LUSC
STAD
TTLL5
LIHC
HNSC
TMED10
ESCA
OV
HNSC
JDP2
LIHC
GBM
PRAD
PCNX
IFT43
RP4-761I2.2
BATF
SKCM
PRAD
RAD51B
RBM25
POLK
SLIT3
SKCM
BRCA
CCNB1IP1
HSD17B4
PRAD
CESC
IPO11
BRCA
STAD
TNFAIP8
BRCA
STX8
PCPG
BLCA
DMXL1
BLCA
SARC
ARHGAP15
NANS
C11orf80
RG9MTD1
BRCA
BRCA
ZBTB20
TSC22D2
LIHC
LUAD
EPHA6
PRAD
BLCA
CNTN6
PRAD
PRAD
ARL13B
TMEM39A
PCYT1A
RBP7
MED18
LUSC
PDPN
THYM
OV
GBM
LUSC
CTNNBIP1
LUSC
UBE4B
BRCA
NMNAT1
UCEC
LUSC
GBM
WDR82
KIRC
LIHC
PLEKHG5
KIF1B
RCOR1
FAR1
OCIAD1
PRAD
SKCM
UVM
KIT
SKCM
UGT2B7
LUSC
LGG
DNAJB14
LUSC
HOPX
SLC4A4
AC087465.1
SPAG1
UCEC
SARC
ANGPT1
BLCA
FAM164A
BRCA
SARC
ZFP41
TNFRSF11B
EIF2AK2
ATP5G3
STAD
KIRP
FLNA
BRCA
LUSC
TKTL1
MTCP1NB
LUAD
PLXNA3
DCUN1D1
LSAMP
BRCA
SARC
NMNAT3
LIHC
CCDC50
MESO
SKCM
NLGN1
OSTN
TMEM104
LUAD
SLC16A5
SKCM
KCTD2
UCEC
BLCA
ARHGAP23
NUP85
USP13
BRCA
PIK3CA
PRAD
BRCA
TBL1XR1
BRCA
UCEC
FNDC3B
ST6GALNAC2
BRCA
PRPSAP1
SARC
C17orf63
LPCAT3
LUSC
FOXJ2
BRCA
ADIPOR2
NECAP1
BLCA
ANK1
Fusion Partner
5'
3'
both
B
KEGG
REACTOME
GO
Cell-Cell communication
MAPK1/MAPK3 signaling
RAF/MAP kinase cascade
Cell surface interactions at the vascular wall
Tie2 Signaling
PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling
Signaling by KIT in disease
Negative regulation of the PI3K/AKT network
Phospholipase D signaling pathway
Central carbon metabolism in cancer
Breast cancer
Acute myeloid leukemia
Rap1 signaling pathway
HIF-1 signaling pathway
Signaling by phosphorylated juxtamembrane, extracellular and kinase domain KIT mutants
EPHA6
TNFAIP8
MAPK signaling pathway
CCDC50
FLNA
Constitutive Signaling by Aberrant PI3K in Cancer
PIK3CA
Ras signaling pathway
PI3K/AKT Signaling in Cancer
KIT
Signaling by SCF-KIT
ANGPT1
HSD17B4
FOXJ2
Axon guidance
PI Metabolism
cardiac muscle cell contraction
sensory perception of sound
transmembrane receptor protein tyrosine kinase activity
protein import into nucleus
regulation of protein import into nucleus
phosphatidylinositol 3-kinase activity
positive regulation of peptidyl-serine phosphorylation
lamellipodium assembly
male gonad development
muscle cell proliferation
positive regulation of phosphatidylinositol 3-kinase signaling
positive regulation of protein kinase B signaling
phosphatidylinositol 3-kinase signaling
positive regulation of smooth muscle cell proliferation
megakaryocyte development

Figure 3. Gene network for recurring gene fusions (RGFs)

(A) Fusion gene partners are represented as nodes. The colors indicate whether a gene is fused to a 5′ (purple) or 3′ (orange) gene or both (white). Two nodes are connected by arrows if one fusion was reported by Tumor Fusion Gene Data Portal. The thickness of the arrow is proportional to the number of cases, and TCGA tumor types are marked on the arrow (blue).

(B) Enrichment analysis identified functional signatures of the 5′ and 3′ fusion partners for RGFs in CTVT. These functionally organized networks are generated based on data from GO, REACTOME, and KEGG databases, and the associated genes are shown in black. The node size indicates the significance of the term. The node colors indicate the proportion of associated genes in each cluster. The links display those genes present in clusters. See also Tables S5 and S6.

We proceeded to perform long-read RNA-seq in CTVT-KM3 to complement the short-read sequencing results. Long-read sequencing has advantages in studying complex SV and GF mechanisms and can improve the accuracy of short-read sequencing results.[27–29] A total of 83,202 high-quality reads were obtained with a filter rate of approximately 0.13% (Figure S1C). 99.30% of all these reads generated mapped successfully on the CanFam3.1 (GCA_000002285.2) reference assembly.[30] Then, we characterized the structure of 340 chimeric transcripts, classifying them into eight distinct categories: 5.42% FSM (Full Splice Match), 28.84% ISM (Incomplete Splice Match), 9.37% NIC (Novel in Catalog), 20.50% NNC (Novel Not in Catalog), 8.05% Antisense (not overlapping with the same strand of the reference gene but overlaps with the annotated gene), 25.04% Intergenic (occurring in intergenic regions), 1.02% Readthrough (where the two genes forming the fusion are on the same gene strand with no known genes in between), and 4.39% Genome (overlapping introns and exons) (Figure S2B). Furthermore, we identified 186 genes that are considered unannotated, of which 160 of them have been previously annotated.[30] Following the application of filtering criteria (see "STAR methods"), we retained 35 non-redundant GFs from a pool of 100 annotated fusion isoforms (Figure S2D).

Additionally, we have also identified complex 3-segment GF which involves the fusion of three different genes or segments from distinct regions in the genome. We identified a type of complex fusion which we named greedy fusion as described in detail in following subsequent sections. Among the cases detected in CTVT, we observed several complex fusions including the fusion between *HSD17B4* (Hydroxysteroid 17-Beta Dehydrogenase 4), *DMXL1* (Dmx Like 1), and *TNFAIP8* (TNF Alpha Induced Protein 8) abbreviated *HSD17B4-DMXL1-TNFAIP8* (ENSCAFG00000000158-ENSCAFG00000000167-ENSCAFG00000000165) and also *RUNX2* (RUNX Family Transcription Factor 2) fusing with two novel genes (ENSCAFG00000031052-*RUNX2*-ENSCAFG00000035853), as well as fusions with three identified genes (ENSCAFG00000014207-ENSCAFG00000025044-ENSCAFG00000032635) that could not be directly discerned through short-read sequencing of full-length chimeric fusion transcripts. Only five CTVT GFs were well supported by both short and long reads (Figure S2E).

Due to the clonally transmissive nature of CTVT, we can define their associated GGFs as follows: (1) those detected in CTVT RGF dataset, (2) those supported by both DNA and RNA data, (3) those being detected within the founder population of CTVT (early lineages), and (4) those occurring in different sublineages. In our findings, we found 11 GGFs that were identified by both RNA and DNA (Figure 2C). Specifically, *HEG1-BAD* (Heart Development Protein With *EGF* Like Domains 1; *BCL2* Associated Agonist Of Cell Death) was only detected in transcriptomes of CTVTs. This is consistent with PCAWG (The Pan-Cancer Analysis of Whole Genomes) results—changes or alterations can be detected at the RNA level which DNA-only approaches cannot detect.[31] Importantly, *KIT* (Proto-Oncogene, Receptor Tyrosine Kinase) mutants are considered gain-of-function mutations where such type of genes predisposes carriers to cancer,[2] and for the case of *KIT*, it fuses with *ANGPT1* (Angiopoietin 1) in ancestral CTVT.

CTVT and human tumors share similar GFs

Alterations shared by both species are more likely to cause cancer than those found in just one species.[32] Therefore, we comprehensively investigated for RGFs (excluded the novel fusion genes in dog) from 33 different tumor types which are publicly accessible through the tumor GF data portal in The Cancer Genome Atlas (TCGA) database.[33] The CTVT GFs detection rate (mean 6 per sample) was very high close to that of sarcoma (SARC) and also higher than the average of 1–3 tumors of mostly human types.[33] Recent studies have also shown that SARC is a type of cancer with enormous chromothripsis.[34] These genes exhibited a susceptibility to tumor development. Notably, the genes that are frequently fused in CTVT also demonstrate a tendency to fuse in other human tumors, as either 5′ or 3′ partners. Approximately 86.36% (19/22) of fusion partner genes in CTVT also had more than one partner in the same or different human tumor types (Table S5). These genes, referred to as "promiscuous genes,"[31] displayed a high degree of connectivity. Network analysis of the promiscuous genes and their fusion partners identified 19 clusters that had an impact on 24 types of tumors in the TCGA dataset (Figure 3A). Notably, within these clusters, the promiscuous genes exhibited fusion events with each other and showed a tendency to act as both 5′ and 3′ partners (Figure 3A; Table S5). One of the most complex clusters involves *HSD17B4*, *DMXL1*, and *TNFAIP8*, which exhibit frequent fusions with each other as well as with different genes in various tumor types. Similarly, the genes *IFT43* (Intraflagellar Transport 43) and *TTLL5* (Tubulin Tyrosine Ligase Like 5) form superclusters and impact up to 13 different tumor types through distinct fusion pairs. These findings highlight the diverse and intricate nature of GF events involving these genes across multiple cancer types.

Hot-fused genes can be used to discover disease pathways

Pathway-specific protein domains not only facilitate pathogenic gene discovery[35] but also represent a novel target class for potential drugs.[36] We then used RGFs to construct functional grouped network based on the Gene Ontology (GO),[37] Kyoto Encyclopedia of Genes and Genomes (KEGG),[38] and REACTOME databases[39] (Figure 3B). A total of 32 significant enrichment terms were detected (p value ≤ 0.05, Bonferroni-correction, two-sided test) including 15 GO terms, 9 KEGG pathways, and 13 REACTOME pathways that are associated with 13 risk genes (Table S6).

Notably, *HSD17B4* and *TNFAIP8* were associated with 5 and 6 cancers with significant risk, respectively (Figure 3A). Their fusions may potentially impact the activity of *PIK3CA* (Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha) through the

phosphatidylinositol (PI) signaling pathway (R-HSA:1483255), thereby promoting tumor growth. *FLNA-TKTL1* (Filamin A; Transketolase Like 1) was related to *MAPK* signaling pathway (KEGG:04010). Besides this, RAS signaling pathway (KEGG:04014) was linked to tumor development through related genes such as *ANGPT1* (Angiopoietin 1), *KIT* (Proto-Oncogene, Receptor Tyrosine Kinase), and *PIK3CA*. Intriguingly, *PIK3CA* plays a key role in the PI3K/AKT signaling pathway (R-HSA:199418, R-HSA:2219528, R-HSA:2219530 and R-HSA:6811558), which is often dysregulated in cancer, and has translational and clinical value.[40] Its fusion partner is *CCDC50* (Coiled-Coil Domain Containing 50), which has been implicated in the progression of a wide range of tumors.[41]

Conserved domains across human tumors reveal carcinogenic risks

The associated fusion partners *HSD17B4*, *DMXL1*, and *TNFAIP8* confer high risk for carcinogenesis.[42–45] *DMXL1* and *TNFAIP8* in human tumors fuse as *TNFAIP8-DMXL1* which is a traditional 2-segment GF in Pheochromocytoma and Paraganglioma (PCPG) and Bladder Urothelial Carcinoma (BLCA).[33,45] And *HSD17B4-TNFAIP8* was also reported in two human tumors Stomach adenocarcinoma (STAD) and cervical squamous cell carcinoma, but also in Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC).[33,45] Noticeably, we found that *TNFAIP8* plays the 5′ or 3′ fusion partner in four human tumors (Figure 4A). To investigate the susceptibility of CTVT GFs in different human tumors, we analyzed the functions of the domains of tumor GFs that were conserved across species. However, we noted that the GF domains of the three tumors CTVT, PCPG, and BLCA were different. Notably, *HSD17B4-DMXL1-TNFAIP8* which encodes only 406 amino acids, conserves one of the features *HSD17B4* in these tumors particularly CTVT and STAD. Our results also indicated that one of the domains NAD(P)-binding superfamily (SSF51735) is a 5′ partner fusion domain, similarly to CADH-Gene3D (G3DSA:3.40.50.720) (Figure 4B). Exon 2 of *TNFAIP8* that encodes the main domains is fully conserved. The unknown function of the domain (PF05527), Tumor Necrosis Factor Induced Protein (PTHR12757), and CATH Superfamily (G3DSA:1.20.1440.160) hardly change after fusion. *TNFAIP8* overexpression in cells reduces cell death and increases tumor growth *in vivo*.[25,44] In addition to this, mobidb-lite (consensus disorder prediction) was a common domain in CTVT, BLCA, and CESC, whereas TMhelix (Region of a membrane-bound protein predicted to be embedded in the membrane) was only common in CTVT and CESC (Figure 4B).

FLNA-TKTL1 has been reported in human breast invasive carcinoma (BRCA).[33,45] We employed the same strategies in CTVT to investigate the conserved domains in *FLNA-TKTL1*. The full-length cDNA sequence to this gene spans 2,735 bp, with an open reading frame (ORF) of 2,679 bp, encoding a polypeptide consisting of 885 amino acids (Figure S3). While both genes had a minimum of 2 fusion partners, *FLNA* was situated at the 5′ position, whereas *TKTL1* was situated at the 3′ position. The majority of the domains (26 out of 27) including two conserved domains, actin-type actin-binding domain signature 1 (PS00019) and domain signature 2 (PS00020), were identified in a corresponding sample of BRCA.

Polygenic fusion mechanisms converge in dogs and humans

Additionally, we investigated the mechanisms of the complex fusion events associated with *HSD17B4*, *DMXL1*, and *TNFAIP8*. In all of these fusion events, the *TNFAIP8* segment was found to fuse with both *DMXL1* and *HSD17B4*. The fusion partners involved in the GFs showed increased copy numbers and frequent inversion rearrangements, which are the main patterns of somatic SV in human cancer genomes.[7] Among them, an extremely high copy number increase occurred in the involved break points, especially in *TNFAIP8* (27–80 copies) (Figures 5A and S4). The copy number gains showed consistent levels across different samples, indicating their genetic properties.

The class of complex GF has been discovered in many human tumor types.[7,31,46] For example, *ETV6-NTRK3* (ETS Variant Transcription Factor 6; Neurotrophic Receptor Tyrosine Kinase 3) is associated with three structural variants in the head and neck thyroid carcinoma sample (i.e., the middle part connects two other fusions and was therefore called a "bridged fusion"[31]). However, instead of producing bridge-like segments, copy number alterations provide more possible combinations in CTVT, including simple events (1) *HSD17B4-DMXL1*, (2) *HSD17B4-TNFAIP8*, (3) *DMXL1-TNFAIP8*, and a complex event (4) *HSD17B4-DMXL1-TNFAIP8* (Table S4). It implies that more than one SV event might lead to the complex fusion. This series of events cannot be satisfactorily explained by deletions, inversions, or translocations. Instead, we infer a mechanism known as copy and paste of local fragments followed by deletion which may be a more concise explanation of these events. So, we propose a fusion mechanism termed "greedy fusion", which we refer to as driving the GF of other fragments by copy number gains.

Firstly, the entire *TNFAIP8* as well as the copy number of the 5′ of *HSD17B4* was duplicated. After that, an additional fragment is inserted in front of *DMXL1*. Finally, a rearranged large fused segment is produced, forming GF alterations *HSD17B4-DMXL1*, *HSD17B4-TNFAIP8*, or *HSD17B4-DMXL1-TNFAIP8* through different deletion events. These rearrangements precluded the generation of readthrough chimeric transcripts. This observation is also supported by long-read sequences. The transcript sequence of *HSD17B4-DMXL1-TNFAIP8* is only 1,132 bp with an ORF of 1,059 bp, which appears to be much shorter than any of the original 2-segment GFs (Figure 5B). It is somewhat similar to a part of copy-and-paste pattern, characterized by the frequent increase in copy number and a mix of inverted and non-inverted breakpoint junctions, in pan-cancer SV studies.[7,46] In addition, we observed that the fusion partner *DMXL1* carries heterozygous breakpoints resulting in the GF *DMXL1-TNFAIP8*, which shows that these complex genomic rearrangements are not originated from a single chromosome. The former rearrangement partner contains exons of *DMXL1* from 25 to 44, and the latter from 1 to 27 (Figure 5B), suggesting that this complex GF was generated by a distinct mechanism. Although we cannot determine the order of the "junctions" between the different "breakpoints", it can be assumed that these fusions originate from the same ancestor.

Downregulation of GGFs after vincristine treatment

The GFs in cancer could be used as targets for treatment design.[20,31] Thus, our study deeply investigated and quantified CTVT RGFs at isoform resolution to reveal pattern of its isoforms. Specifically, the full-length fusion transcript was used as a reference combined together with

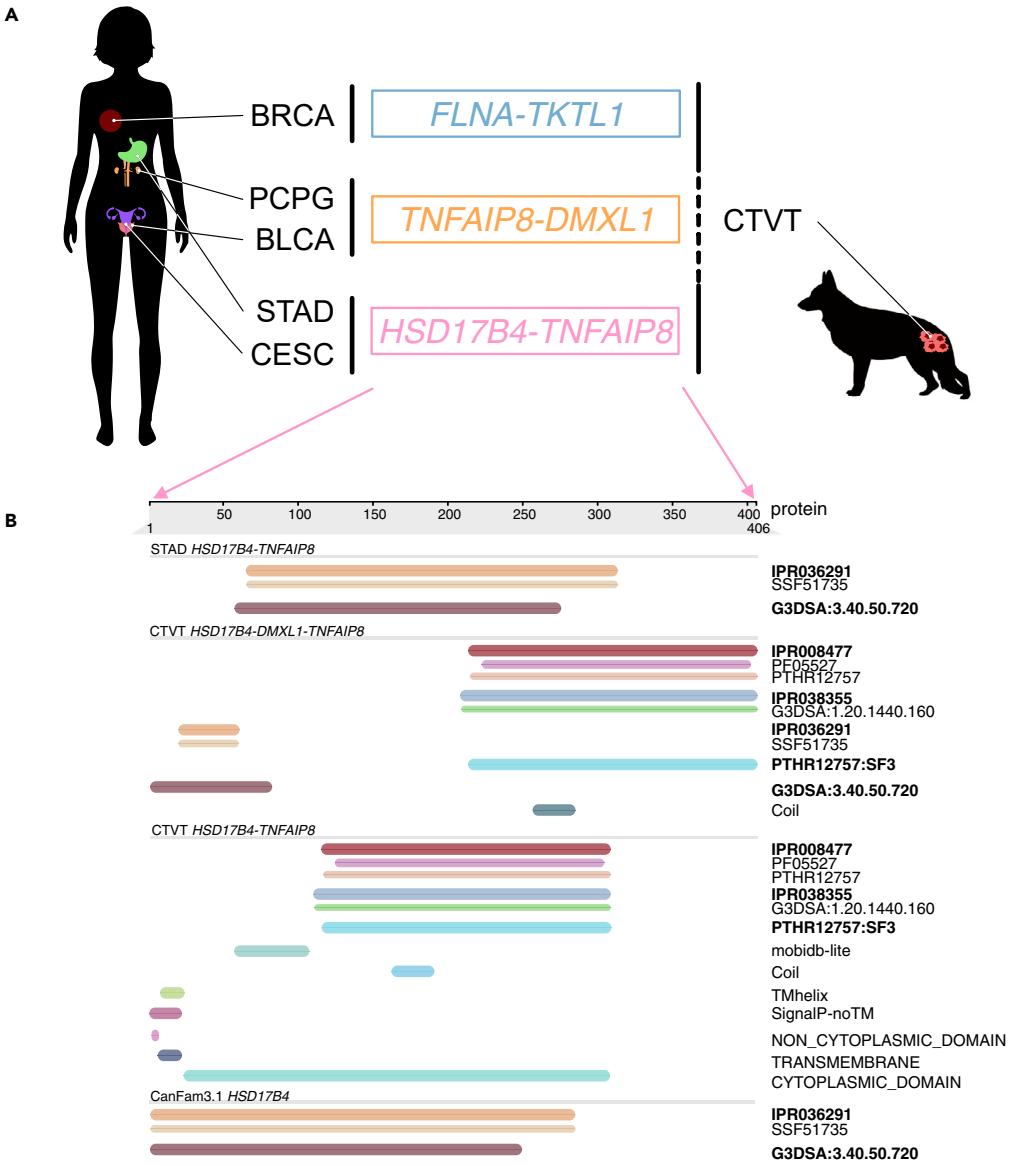


Figure 4. Co-occurrence germline gene fusions (GGFs) in CTVT and human tumors

(A) Three GGFs found in both CTVT and human different tumors. Dashed lines indicate opposite orientations of the fusion in two species. BRCA, Breast invasive carcinoma; PCPG, Pheochromocytoma and Paraganglioma; BLCA, Bladder Urothelial Carcinoma; STAD, Stomach adenocarcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma.

(B) Functional classification and annotation of *HSD17B4-TNFAIP8* fusion proteins. The same domains are displayed using the same color bar. See also Figure S3.

expression data from serial CTVT biopsies (before and after vincristine treatment)[22] and clinical information in order to investigate whether the treatment had any effect on them (Figure S5A). CTVT-5 showed that 91.95% (80/87) of the fusion isoforms were downregulated at 14-day post-therapy, whereas CTVT-6 showed 56.38% (53/94) of the fusion isoforms being downregulated (Figure 6A). Others also showed varying degrees of downregulation after 28-day treatment (Figure 6A). Clinical examination indicated that the post-therapy biopsies of CTVT-5, CTVT-765, CTVT-766, and CTVT-772 were in regressive phase, while others such as CTVT-6, CTVT-761, CTVT-774, and CTVT-775 were in non-regressive phase (but CTVT-775 showed some signatures of entering the regressive phase).[22] The principal-component analysis (PCA) showed a total variance of 66.2% for principal components (PCs) 1 and 2, revealing the relationship of these biopsies (Figure S5B).

In order to assess whether these CTVT GGFs act as favorable drivers of tumorigenesis, we compared their expression levels before and after vincristine treatment. Specifically, greedy fusion (*HSD17B4-DMXL1-TNFAIP8*) was also included in this analysis. CTVT GGFs showed significant changes at different phases (p value < 0.05, Mann-Whitney U test, two-sided). We observed a downregulated expression pattern that correlated with treatment duration (Figures 6B and S5C). But their expression was consistently higher in progressive phase (0-day) than in

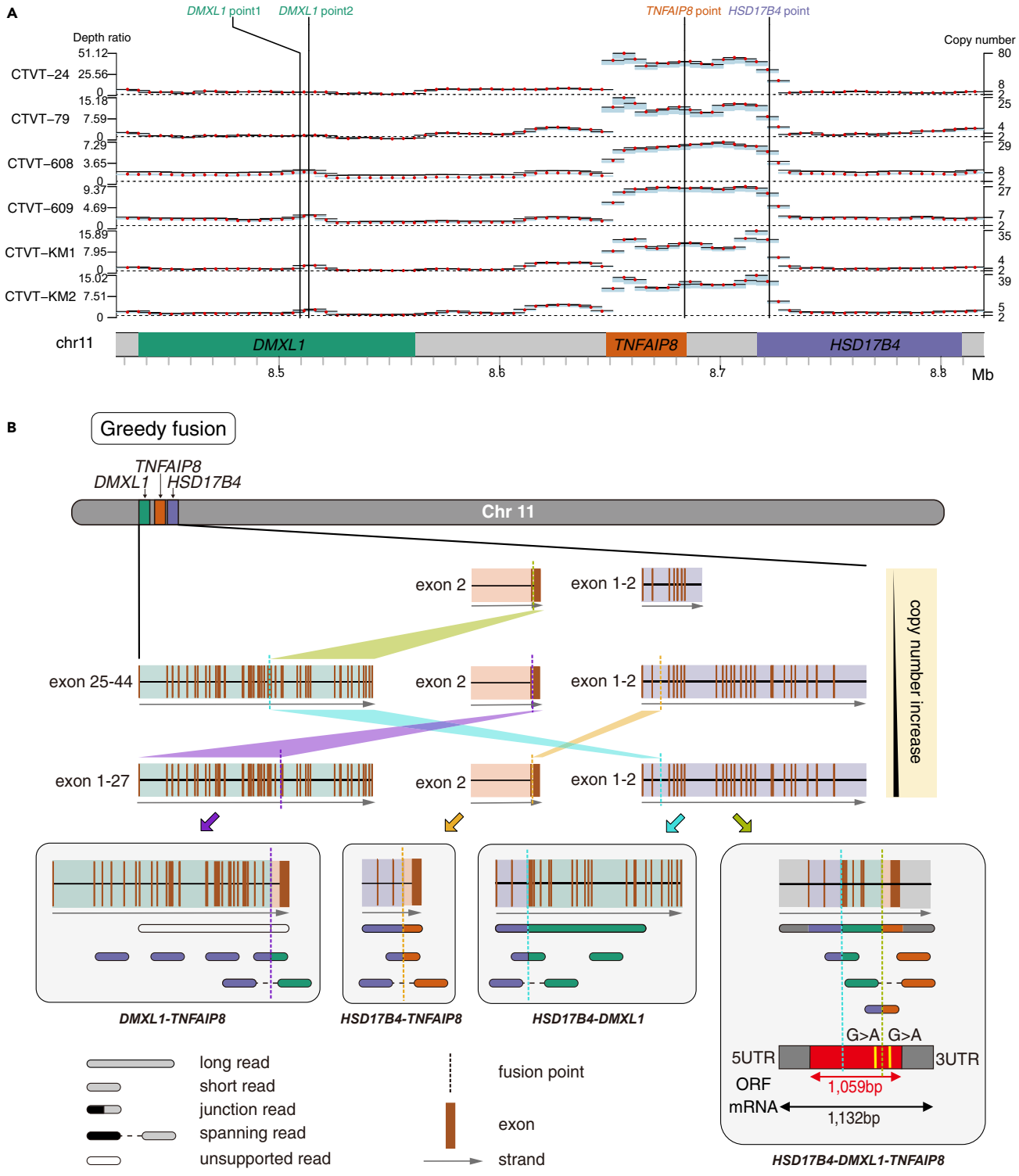


Figure 5. Schematic of greedy fusion

(A) Sequencing depth ratio (tumor versus host) and copy number of the genomic segments involved in the greedy fusion associated genes *DMXL1*, *TNFAIP8*, and *HSD17B4*. From top to bottom, are the results for CTVT-24, CTVT-79, CTVT-608, CTVT-609, CTVT-KM1, and CTVT-KM2. An average depth ratio of genomic position is summarized within 10 kb windows staggered every 5 kb. Red dots indicate the expected number of copies for this segment and the light blue bars indicate the 95% confidence interval. The dotted line indicates that the copy number is 2, followed by the median and maximum. Black vertical lines represent fusion points.

Figure 5. *Continued*

(B) An example of CTVT greedy fusion. The top part shows the location of the related genes on chromosome 11, the middle part shows copy number alteration and the number of exons retained after fusion, and the bottom part shows the final structure of fusions *DMXL1-TNFAIP8*, *HSD17B4-TNFAIP8*, *HSD17B4-DMXL1*, and *HSD17B4-DMXL1-TNFAIP8*. Different breakpoints are indicated by different colors, with polygons corresponding colors were utilized to describe the distinct fusion processes. See also Figure S4.

regressive phase (14-day or 28-day). Compared to the biopsies from the regressive phase, the biopsies from the non-regressive phase exhibited resistance to treatment. The novel GGF (ENSCAFG00000003299-ENSCAFG00000034243) did not show significant downregulation in expression at any phase, which could be attributed to artifacts caused by deficiencies in the annotation files of the canine genome (Figure 6B).

These fusions involving kinase genes, oncogenes, or tumor suppressor genes showed more significant downregulation than others (Figure 6C). A previous study indicated that kinase fusions associated with hematopoietic and malignancies in solid tumors exert an oncogenic role.[47] We then investigated the expression levels of the fusions, especially at the progressive phase involving kinase genes (n = 2), oncogenes (n = 3), and suppressor genes (n = 2). Fusion transcript abundance-associated kinase genes were also higher than oncogene fusions and suppressor GFs, showing a significant difference in CTVT progressive phase (p value < 0.05, Mann-Whitney U test, two-sided). This implies that kinase fusions contribute more to CTVT development.

We further analyzed the association of CTVT GGFs in different phylogenetic clades to understand whether CTVT could lead to drug resistance due to heterogeneity (Figure 1C). As shown in Figure 6D, the expression levels of GGFs from different locations and samples appeared to be consistent due to the strong positive correlation at progression phase (Italy vs. China: R = 0.84; Brazil vs. China: 0.81; Brazil vs. Italy: 0.88, Spearman, Mann-Whitney U test, two-sided). This evidence supports a high degree of similarity or consistency across different CTVT biopsies. But there were slight differences in responses between individuals after treatment (Brazil vs. Italy: R = 0.78).

To compare the differences in the stage of CTVT GFs between treatment and natural conditions, we cultured CTVT primary cells in serial passages and analyzed expression of GFs. The CTVT GGFs showed different expression trends with the increase of culture time. The expression of *HSD17B4-DMXL1-TNFAIP8* gradually decreased, while others did not (Figure S5D).

CTVT GGFs expression pattern correlates with DNA methylation level

Typically, CpG islands (CGIs) are located in the promoter regions of genes, and methylation or demethylation of CpG in these regions is thought to affect the expression of their downstream genes,[48] and even the correlating regions peaked at ~2 kb downstream of transcription start site (TSS) 3′UTRs and near stop codons.[8,49] To investigate whether GGFs expression patterns are associated with epigenetic modifications, we scaled their methylation levels within TSS, transcription end site (TES), 3 kb upstream of TSS, and 3 kb downstream of TES. As expected, we observed a consistent methylation pattern in biopsies before and after CTVT treatment, in line with the expression profile from RNA-seq analysis (Figure S6). The methylation levels of the 6 GGFs remained consistently high from the TSS to TES after treatment, regardless of the days of treatment (Figure 7A). A negative correlation between DNA methylation and mRNA expression was observed in extensive regions downstream of the promoter, sometimes greater than 50 kb.[8] This methylation pattern was particularly consistent for CTVT-17 and CTVT-5. However, CTVT-6 displayed a more intricate methylation pattern. It demonstrated a relatively higher overall methylation level before treatment, and there was an increase in methylation levels near TSS followed by a gradual decrease toward TES after treatment. In contrast, CTVT-5 and CTVT-17 displayed a decrease followed by an increase in methylation levels near the TES. This is consistent with the results of our RNA-seq analysis, where chemotherapy induced changes in the expression levels of GGFs, with more pronounced alterations in methylation in tumors entering the regression phase.

In addition, we compared the methylation level distribution of the differentially methylated region (DMR) between CTVT before and after treatment biopsies. We found 83 DMRs, 51 of which showed significant changes (p value < 0.05) (Figure 7B). These DMRs were not evenly distributed on the chromosomes, with 22.9% (19/83) of them on the chromosome X. Notably, the region where *FLNA* is located (chrX:122070601–122070650) also shows significantly different methylation levels (p value = 0.002), and this gene undergoes fusion with *TKTL1*, with hypomethylation and concomitant upregulation of primary biopsy expression. As previously studied, the genome of CTVT underwent a massive and complex rearrangement,[15] yet only a small fraction of the fusion genes generated were transcribed accordingly, and the expression of these fusions accompanied by tumor development was strongly correlated with DNA methylation.

GGFs can be used as diagnostic molecular biomarkers

GFs have become ideal for diagnostic purposes.[20] By comparing the corresponding hosts (diseased individual), the GGF *UBE4B-CFAP61* (Ubiquitination Factor E4B; Cilia And Flagella Associated Protein 61) is specifically expressed in CTVT (Figures S7A–S7C; Table S7). The molecular weight of this fusion protein was determined to be approximately 20 kDa (Figures S7D and S7E), which aligns with the computed molecular weight based on its amino acid sequence. Previous studies have demonstrated that overexpression of *UBE4B* was referred to as an oncogenic feature, since increasing levels of *UBE4B* would promote *p53* polyubiquitination and degradation, subsequently triggering the transactivation and apoptosis processes through *p53*-dependent inhibition.[50] Although we demonstrated the possibility of translation of the fusion transcripts, the unavailability of CTVT stable cell lines prevented us from gaining a deeper understanding of the regulatory mechanisms of tumor growth and development of all GGFs.

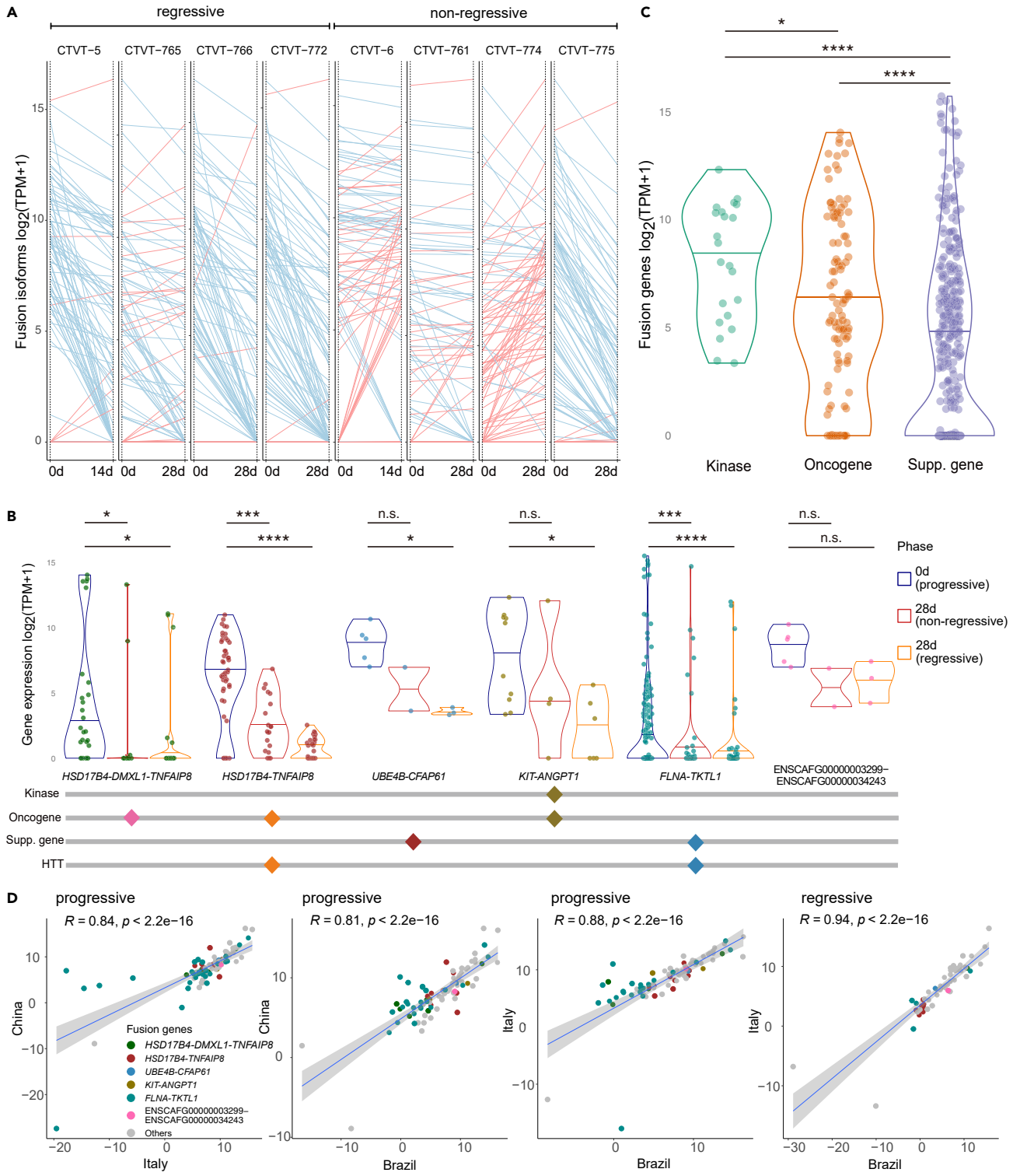


Figure 6. Expression patterns of germline gene fusions (GGFs)

(A) Slope charts show expressions of germline gene fusions (GGFs) related isoforms before (day 0) and after (day 14 or day 28) vincristine treatment. Upregulated fusion isoforms are marked in red, and downregulated in blue.

Figure 6. *Continued*

(B) CTVT GGFs are highly expressed in CTVT progressive phase. The gene fusion partners are kinases, oncogenes, tumor suppressor genes (Supp. genes) or the gene fusion found in other human tumor types (HTT). * denotes p value < 0.05, ** denotes p value < 0.01, *** denotes p value < 0.001, **** denotes p value < 0.0001.

(C) Expression of different GGFs involving kinases, oncogenes, or suppressor genes.

(D) Scatterplot illustrating the correlations of GGFs in CTVT from different CTVT phylogenetic clade. See also Figure S5 and Table S7.

DISCUSSION

For familial and syndromic cancers, more germline variants remain to be discovered. In this study, we focused on GGFs with common susceptibility loci between human tumors and CTVT. We applied different sequencing technologies and combined with multi-omics data to explore fusion patterns, kinase genes, conservative domains, signaling pathways, and expression profiles of CTVTs. Our results support the use of CTVT as a germline variant tumor model and reveal the contribution of non-genetic modifying factors for GGF.

We summarized a set of GGFs in CTVT which could be linked up with early evolutionary events that gradually became prevalent in many parts of the globe. But this set of data may be overestimated, as whole-exome data show only 5 potential drivers of early CTVT that include *CDKN2A*, *MYC*, *PTEN* (Phosphatase And Tensin Homolog), *SETD2*, and *RB1* (RB Transcriptional Corepressor 1).[16] This problem can be improved with increasing sample size. However, there is no doubt some fusions happen directly at the RNA level, such as *trans*-splicing or readthrough events.[31] Our results are in line with this conclusion and found evidence for long non-coding RNA (lncRNA) fused to coding genes whose function has not yet been determined. These insights illustrate the importance and non-negligibility of RNA alterations.

Our cross-species analysis shows that CTVT GGFs share many similarities with human tumors, such as *HSD17B4-TNFAIP8* in STAD and CESC; *TNFAIP8-DMXL1* in BLCA and PCPG; and *FLNA-TKTL1* in BRCA.[19,45] Among the GGFs, the most compelling is *HSD17B4-DMXL1-TNFAIP8*. This fusion mechanism is similar to the mechanism of 3-segments GF in human tumors—copy and paste with multiple simple fusion events.[31] Greedy fusion exemplifies insights for the presence of hot loci on chromosomes that may provide tumors with potential selective growth or survival advantages.

We do not equate CTVT germline variants with human CPGs exactly, but only reveal possible predisposition mechanism of cancer through overlapping germline variants in the absence of familial cancer cases today. It is clear that the CTVT germline variant hits 5 of the 114 CPGs,[2] including 2 fusion partners (*KIT* and *RUNX* Family) and 3 previously published early drivers (*CDKN2A*, *PTEN*, and *RB1*). The shared variants suggest that unstable genes that might have been buried early in mammalian chromosomes could be unraveled through pan-cancer studies that involve multi-type species.[51]

Transmissible tumors have long been on the agenda as models of the cancer stem-cell process.[52] To explore tumor treatment approaches to maximize the use of CTVT as a tumor model, we used an innovative strategy by combining multi-omics and different sequencing technologies to analyze GGFs in response to treatment. These fusions show striking concordance at the level of RNA expression and DNA methylation, emphasizing the high correlation between them. But we still need more evidence to elucidate whether DNA methylation is a cause or a consequence of altered gene expression. In general, vincristine chemotherapy is widely known for its effectiveness in at least 88% of the previously reported cancer cases;[53] however, individual differences cannot be ruled out. For example, CTVT-765 and CTVT-761 showed minimal differences pre- and post-therapy, resulting in fewer downregulated GFs. We believe that chemotherapy is effective since individual differences require more time to reach a state from non-regression to regression. Thus, there is an urgent need to study the expression of specific genes as markers for different tumor growth stages. GFs differ in the extent to which they affect tumor growth status by qualitative and quantitative descriptions. Although the three phases of CTVT growth (progressive, stationary, and regressive) have been identified,[22] there are individual differences in the main influencing factors as well as the arrival times of the different phases. RT-qPCR expression profiles of CTVT cells showed peaks or troughs at a specific time. The expression of GFs did not show a completely linear relationship with tumor development. Different GFs may also play different regulatory roles in different stages of tumor growth. We have successfully established a CTVT primary cell isolation and passaging technique, but more evidence from cellular experiments is needed to demonstrate which downstream targets they affect and which cellular processes they interfere with.

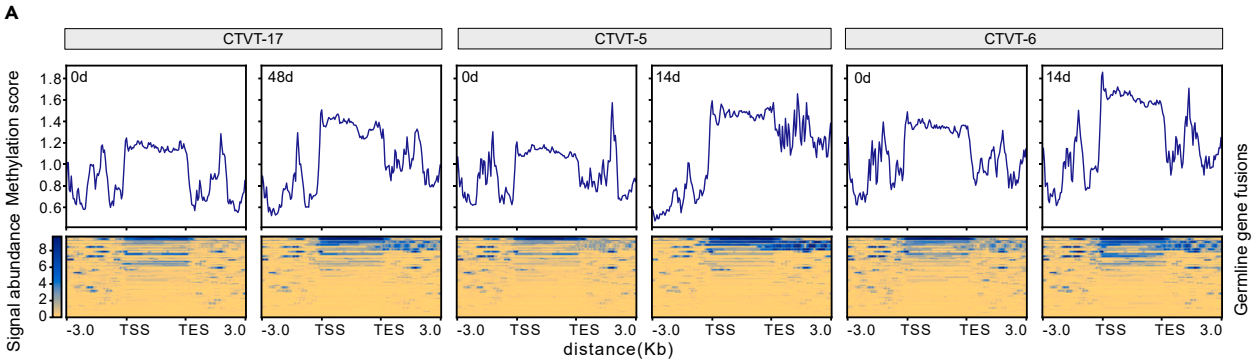
In conclusion, our study sheds light on the landscape of CTVT GGFs and provides a rich resource for future research. In fact, even decades-long cancer surveillance of relatives in a single family can be a very considerable financial outlay. Thus, research on CTVT helps us to understand genetic alterations in cancers as dogs and humans share a common living environment and medical conditions. Accelerating the application of CTVT in translational medicine could provide health benefits and opportunities for cancer prevention.

Limitations of the study

This study was not based on purified tumor cells, although we applied parallel approaches to correct this; and the analysis of CTVT variations still has potential systemic biases. Meanwhile, since CTVT cell lines have not been established, we could not confirm whether the identified GGFs play a key driving role in tumorigenesis. In addition, the identified GFs were based on the current version of the genome assembly and annotation files. With the improvement of genome assembly and annotation, the results may be affected, but we believe this will not undermine the major conclusions of this study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:


A
CTVT-17
CTVT-5
CTVT-6
0d
48d
0d
14d
0d
14d
Signal abundance Methylation score
Germline gene fusions
-3.0 TSS TES 3.0
distance(Kb)

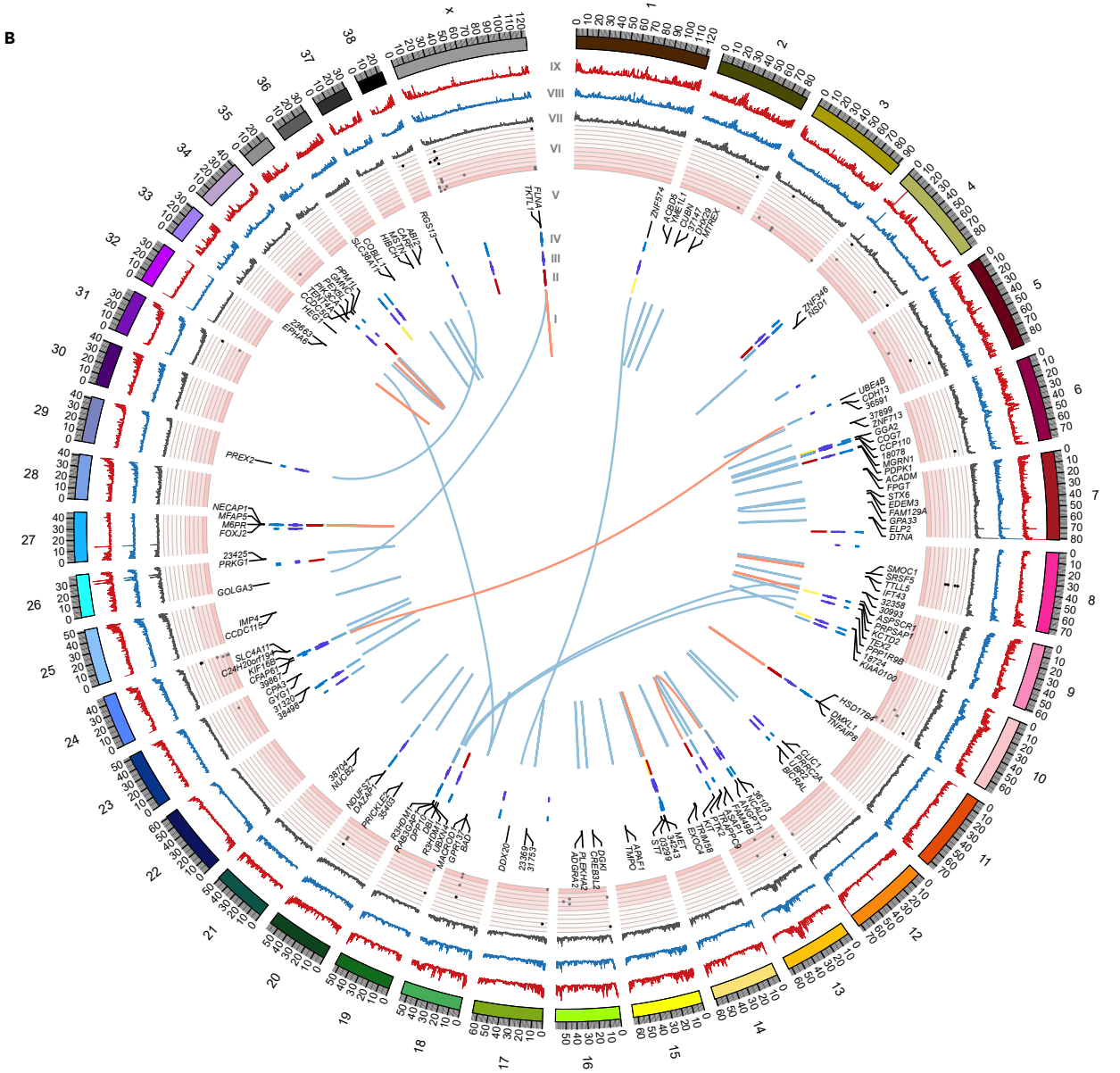

B

Figure 7. CTVT DNA methylation pattern for CTVT germline gene fusions (GGFs)

(A) DNA methylation profiles (top) and heatmaps (bottom) of CTVT-5 (A), CTVT-6 (B), and CTVT-17 (C). Methylation scores were calculated within transcription start site (TSS), transcription end site (TES), 3 kb upstream of TSS, and 3 kb downstream of TES. From left to right are the biopsy before (0 days) and after treatment (14 days or 48 days) for each CTVT.

(B) Circos plot showing high-confidence CTVT GFs landscape. Starting from the inner circle to the outer, (I) Connection of somatic gene fusions (SGFs, blue) and germline gene fusion (GGF, red) partners; (II) Coverage of supported reads by long-read transcriptome sequences (color from warm to cold, the supported counts decrease); (III) Distribution of exons of annotated CTVT GFs; (IV) Distribution of transcripts of annotated CTVT GFs; (V) Gene symbol or Ensembl ID (The last five characters of the Ensembl ID are intercepted) of GF partners; (VI) Differentially methylated regions (DMRs) with adjusted *p* value (dark to light red backgrounds represent decreasing *p* values), DMRs with *p* value < 0.05 are colored black dots, others are gray; (VII-IX) Methylation peaks of different CTVT biopsies. gray: day 0 of vincristine treatment (VII), blue: day 14 (VIII), red: day 48 (IX). The outermost circle represents the chromosomes of the dog. See also Figure S6.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108431.

ACKNOWLEDGMENTS

We thank Yong-bin Chen, Jin-Xiu Li, and Bo-Yang Liu for the invaluable comments and suggestions regarding the experiments, Adrian Baez-Ortega for his assistance in the CTVTs phylogenetic inference, and Ten-Xiao Si for helping us to collect samples. This work was supported by National Key R&D Program of China (2019YFA0707101), STI2030-Major Projects (2021ZD0203900), Spring City Plan: the High-level Talent Promotion and Training Project of Kunming (2022SCP001), Key Research and Development Program of Yunnan province (202203AC100010), and Key Research Program of Frontier Sciences of the Chinese Academy of Sciences (CAS) (ZDBS-LY-SM011). Guo-Dong Wang was kindly supported by Yunnan Fundamental Research Projects (202201AV070011). Yan-Hu Liu was supported by Youth Innovation Promotion Association of Chinese Academy of Sciences. This work was also supported by the Animal Branch of the Germplasm Bank of Wild Species, CAS (the Large Research Infrastructure Funding) and the National Genomics Data Center at the Beijing Institute of Genomics, CAS.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.-P.Z. and G.-D.W.; Investigation, B.-W.Z., Q.-Q.W., S.-R.Z., and C.L.; Formal Analysis and Visualization, B.-W.Z. and Q.-Q.W.; Resources, T.-T.Y. and F.-L.C.; Data Curation, X.W. and T.-T.Y.; Writing - Original Draft, B.-W.Z., Q.-Q.W., and S.-R.Z.; Writing - Review & Editing, D.H.M., Y.-H.L., G.-D.W., and Y.-P.Z.; Supervision, Y.-P.Z., G.-D.W., and Y.-H.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: February 1, 2023
Revised: June 24, 2023
Accepted: November 8, 2023
Published: November 10, 2023

REFERENCES

1. Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. N. Engl. J. Med. *343*, 78–85.
2. Rahman, N. (2014). Realizing the promise of cancer predisposition genes. Nature *505*, 302–308.
3. Zhang, J., Walsh, M.F., Wu, G., Edmonson, M.N., Gruber, T.A., Easton, J., Hedges, D., Ma, X., Zhou, X., Yergeau, D.A., et al. (2015). Germline Mutations in Predisposition Genes in Pediatric Cancer. N. Engl. J. Med. *373*, 2336–2346.
4. Huang, K.-l., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N., et al. (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. Cell *173*, 355–370.e14.
5. Gröbner, S.N., Worst, B.C., Weischenfeldt, J., Buchhalter, I., Kleinheinz, K., Rudneva, V.A., Johann, P.D., Balasubramanian, G.P., Segura-Wang, M., Brabetz, S., et al. (2018). The landscape of genomic alterations across childhood cancers. Nature *555*, 321–327.
6. Fung, Y.K., Murphree, A.L., T'Ang, A., Qian, J., Hinrichs, S.H., and Benedict, W.F. (1987). Structural Evidence for the Authenticity of the Human Retinoblastoma Gene. Science *236*, 1657–1661.
7. Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak, S., Korbel, J.O., Haber, J.E., et al. (2020). Patterns of somatic structural variation in human cancer genomes. Nature *578*, 112–121.
8. Hovestadt, V., Jones, D.T.W., Picelli, S., Wang, W., Kool, M., Northcott, P.A., Sultan, M., Stachurski, K., Ryzhova, M., Warnatz, H.-J., et al. (2014). Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature *510*, 537–541.
9. Wei, Y., Wu, J., Gu, W., Qin, X., Dai, B., Lin, G., Gan, H., Freedland, S.J., Zhu, Y., and Ye, D. (2019). Germline DNA Repair Gene Mutation Landscape in Chinese Prostate Cancer Patients. Eur. Urol. *76*, 280–283.
10. Green, R.C., Berg, J.S., Grody, W.W., Kalia, S.S., Korf, B.R., Martin, C.L., McGuire, A.L., Nussbaum, R.L., O'Daniel, J.M., Ormond, K.E., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet. Med. *15*, 565–574.
11. Murgia, C., Pritchard, J.K., Kim, S.Y., Fassati, A., and Weiss, R.A. (2006). Clonal Origin and Evolution of a Transmissible Cancer. Cell *126*, 477–487.
12. Ostrander, E.A., Davis, B.W., and Ostrander, G.K. (2016). Transmissible Tumors: Breaking the Cancer Paradigm. Trends Genet. *32*, 1–15.
13. Murchison, E.P., Wedge, D.C., Alexandrov, L.B., Fu, B., Martincorena, I., Ning, Z., Tubio, J.M.C., Werner, E.I., Allen, J., De Nardi, A.B., et al. (2014). Transmissible Dog Cancer Genome Reveals the Origin and History of an Ancient Cell Lineage. Science *343*, 437–440.
14. Wang, X., Zhou, B.-W., Yang, M.A., Yin, T.-T., Chen, F.-L., Ommeh, S.C., Esmailizadeh, A., Turner, M.M., Poyarkov, A.D., Savolainen, P., et al. (2019). Canine transmissible venereal tumor genome reveals ancient introgression from coyotes to pre-contact dogs in North America. Cell Res. *29*, 592–595.
15. Decker, B., Davis, B.W., Rimbault, M., Long, A.H., Karlins, E., Jagannathan, V., Reiman, R., Parker, H.G., Drögemüller, C., Corneveaux, J.J., et al. (2015). Comparison against 186 canid whole-genome sequences reveals survival strategies of an ancient clonally transmissible canine tumor. Genome Res. *25*, 1646–1655.
16. Baez-Ortega, A., Gori, K., Strakova, A., Allen, J.L., Allum, K.M., Bansse-Issa, L., Bhutia, T.N., Bisson, J.L., Briceño, C., Castillo Domracheva, A., et al. (2019). Somatic evolution and global expansion of an ancient transmissible cancer lineage. Science *365*, eaau9923.
17. Katzir, N., Rechavi, G., Cohen, J.B., Unger, T., Simoni, F., Segal, S., Cohen, D., and Givol, D. (1985). "Retroposon" insertion into the cellular oncogene c-myc in canine transmissible venereal tumor. Proc. Natl. Acad. Sci. USA *82*, 1054–1058.
18. Wong, K., van der Weyden, L., Schott, C.R., Foote, A., Constantino-Casas, F., Smith, S., Dobson, J.M., Murchison, E.P., Wu, H., Yeh, I., et al. (2019). Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma. Nat. Commun. *10*, 353.
19. Ulvé, R., Rault, M., Bahin, M., Lagoutte, L., Abadie, J., De Brito, C., Coindre, J.-M., Botherel, N., Rousseau, A., Wucher, V., et al. (2017). Discovery of Human-Similar Gene Fusions in Canine Cancers. Cancer Res. *77*, 5721–5727.
20. Mertens, F., Johansson, B., Fioretos, T., and Mitelman, F. (2015). The emerging complexity of gene fusions in cancer. Nat. Rev. Cancer *15*, 371–381.
21. Foltz, S.M., Gao, Q., Yoon, C.J., Sun, H., Yao, L., Li, Y., Jayasinghe, R.G., Cao, S., King, J., Kohnen, D.R., et al. (2020). Evolution and structure of clinically relevant gene fusions in multiple myeloma. Nat. Commun. *11*, 2666.
22. Frampton, D., Schwenzer, H., Marino, G., Butcher, L.M., Pollara, G., Kriston-Vizi, J., Venturini, C., Austin, R., de Castro, K.F., Ketteler, R., et al. (2018). Molecular Signatures of Regression of the Canine Transmissible Venereal Tumor. Cancer Cell *33*, 620–633.e6.
23. Ní Leathlobhair, M., Perri, A.R., Irving-Pease, E.K., Witt, K.E., Linderholm, A., Haile, J., Lebrasseur, O., Ameen, C., Blick, J., Boyko, A.R., et al. (2018). The evolutionary history of dogs in the Americas. Science *361*, 81–85.
24. Ho, S.S., Urban, A.E., and Mills, R.E. (2020). Structural variation in the sequencing era. Nat. Rev. Genet. *21*, 171–189.
25. Evans, J.M., Parker, H.G., Rutteman, G.R., Plassais, J., Grinwis, G.C.M., Harris, A.C., Lana, S.E., and Ostrander, E.A. (2021). Multi-omics approach identifies germline regulatory variants associated with hematopoietic malignancies in retriever dog breeds. PLoS Genet. *17*, e1009543.
26. Haas, B.J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. *20*, 213.
27. Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. Dev. Reprod. Biol. *13*, 278–289.
28. Anvar, S.Y., Allard, G., Tseng, E., Sheynkman, G.M., de Klerk, E., Vermaat, M., Yin, R.H., Johansson, H.E., Ariyurek, Y., den Dunnen, J.T., et al. (2018). Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. Genome Biol. *19*, 46.
29. Namba, S., Ueno, T., Kojima, S., Kobayashi, K., Kawase, K., Tanaka, Y., Inoue, S., Kishigami, F., Kawashima, S., Maeda, N., et al. (2021). Transcript-targeted analysis reveals isoform alterations and double-hop fusions in breast cancer. Commun. Biol. *4*, 1320.
30. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48*, D682–D688.
31. PCAWG Transcriptome Core Group, Calabrese, C., Demircioğlu, D., Demircioğlu, D., Fonseca, N.A., He, Y., Kahles, A., Lehmann, K.V., Liu, F., Shiraishi, Y., et al. (2020). Genomic basis for RNA alterations in cancer. Nature *578*, 129–136.
32. Rowell, J.L., McCarthy, D.O., and Alvarez, C.E. (2011). Dog models of naturally occurring cancer. Trends Mol. Med. *17*, 380–388.
33. Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H., and Verhaak, R.G.W. (2015). The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene *34*, 4845–4854.
34. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Bose, H.K., Ouellette, B.F.F., Li, C.H., et al. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.
35. Shim, J.E., Kim, J.H., Shin, J., Lee, J.E., and Lee, I. (2019). Pathway-specific protein domains are predictive for human diseases. PLoS Comput. Biol. *15*, e1007052.
36. Schapira, M., Tyers, M., Torrent, M., and Arrowsmith, C.H. (2017). WD40 repeat

domain proteins: a novel target class? Nat. Rev. Drug Discov. *16*, 773–786.
37. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.
38. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361.
39. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The Reactome pathway knowledgebase. Nucleic Acids Res. *42*, D472–D477.
40. Thorpe, L.M., Yuzugullu, H., and Zhao, J.J. (2015). PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. Nat. Rev. Cancer *15*, 7–24.
41. Wang, H., Zhang, C.Z., Lu, S.X., Zhang, M.F., Liu, L.L., Luo, R.Z., Yang, X., Wang, C.H., Chen, S.L., He, Y.F., et al. (2019). A Coiled-Coil Domain Containing 50 Splice Variant Is Modulated by Serine/Arginine Rich Splicing Factor 3 and Promotes Hepatocellular Carcinoma in Mice by the Ras Signaling Pathway. Hepatology *69*, 179–195.
42. Huang, H., Liu, R., Huang, Y., Feng, Y., Fu, Y., Chen, L., Chen, Z., Cai, Y., Zhang, Y., and Chen, Y. (2020). Acetylation-mediated degradation of HSD17B4 regulates the progression of prostate cancer. Aging *12*, 14699–14717.
43. van den Boom, J., Wolter, M., Blaschke, B., Knobbe, C.B., and Reifenberger, G. (2006). Identification of novel genes associated with astrocytoma progression using suppression subtractive hybridization and real-time reverse transcription-polymerase chain reaction. Int. J. Cancer *119*, 2330–2338.
44. Niture, S., Gyamfi, M.A., Lin, M., Chimeh, U., Dong, X., Zheng, W., Moore, J., and Kumar, D. (2020). TNFAIP8 regulates autophagy, cell steatosis, and promotes hepatocellular carcinoma cell proliferation. Cell Death Dis. *11*, 178.
45. Kim, P., and Zhou, X. (2019). FusionGDB: fusion gene annotation DataBase. Nucleic Acids Res. *47*, D994–D1004.
46. Carvalho, C.M.B., Ramocki, M.B., Pehlivan, D., Franco, L.M., Gonzaga-Jauregui, C., Fang, P., McCall, A., Pivnick, E.K., Hines-Dowell, S., Seaver, L.H., et al. (2011). Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. Nat. Genet. *43*, 1074–1081.
47. Stransky, N., Cerami, E., Schalm, S., Kim, J.L., and Lengauer, C. (2014). The landscape of kinase fusions in cancer. Nat. Commun. *5*, 4846.
48. Razin, A., and Cedar, H. (1991). DNA methylation and gene expression. Microbiol. Rev. *55*, 451–458.
49. Meyer, K.D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C.E., and Jaffrey, S.R. (2012). Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3′ UTRs and near Stop Codons. Cell *149*, 1635–1646.
50. Antoniou, N., Lagopati, N., Balourdas, D.I., Nikolaou, M., Papalampros, A., Vasileiou, P.V.S., Myrianthopoulos, V., Kotsinas, A., Shiloh, Y., Liontos, M., and Gorgoulis, V.G. (2019). The Role of E3, E4 Ubiquitin Ligase (UBE4B) in Human Pathologies. Cancers *12*, 62.
51. Hutter, C., and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell *173*, 283–285.
52. O'Neill, I.D. (2011). Concise Review: Transmissible Animal Tumors as Models of the Cancer Stem-Cell Process. Stem Cell. *29*, 1909–1914.
53. do Prado Duzanski, A., Flórez, L.M.M., Fêo, H.B., Romagnoli, G.G., Kaneno, R., and Rocha, N.S. (2022). Cell-mediated immunity and expression of MHC class I and class II molecules in dogs naturally infected by canine transmissible venereal tumor: Is there complete spontaneous regression outside the experimental CTVT? Res. Vet. Sci. *145*, 193–204.
54. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
55. Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185.
56. Websites: PacificBiosciences (2019). SMRT Analysis Software Suite. https://www.pacb.com/.
57. De Coster, W., and Rademakers, R. (2023). NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics *39*, btad311.
58. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv *1303*. 3997.
59. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008.
60. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv *1303*.
61. Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., and Eklund, A.C. (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann. Oncol. *26*, 64–70.
62. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.
63. Verhaak, R.G.W., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98–110.
64. Revkov, E., Kulshrestha, T., Sung, K.W.-K., and Skanderup, A.J. (2023). PUREE: accurate pan-cancer tumor purity estimation from gene expression data. Commun. Biol. *6*, 394.
65. Websites, and Ortiz, E.M. (2019). vcf2phylip v2.0: Convert a VCF Matrix into Several Matrix Formats for Phylogenetic Analysis.
66. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.
67. Websites, and Rambaut, A. (2018). FigTree. http://tree.bio.ed.ac.uk/software/Figtree/.
68. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.
69. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.
70. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.
71. Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., and Sedlazeck, F.J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. *8*, 14061.
72. Nicorici, D., Şatalan, M., Edgren, H., Kangaspeska, S., Murumägi, A., Kallioniemi, O., Virtanen, S., and Kilkku, O. (2014). FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. Preprint at bioRxiv *1303*.
73. Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938–2940.
74. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.
75. Websites, and Tseng, E. (2020). cDNA_Cupcake. https://github.com/Magdoll/cDNA_Cupcake.
76. Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. *28*, 396–411.
77. Websites, and Haas, B.J. (2021). TransDecoder (Find Coding Regions within Transcripts). https://github.com/TransDecoder/TransDecoder/.
78. Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., et al. (2021). The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. *49*, D344–D354.
79. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091–1093.
80. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. *13*, 2498–2504.
81. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.
82. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
83. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

84. Lienhard, M., Grimm, C., Morkel, M., Herwig, R., and Chavez, L. (2014). MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. Bioinformatics *30*, 284–286.

85. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

86. Liao, K.-W., Lin, Z.-Y., Pao, H.-N., Kam, S.-Y., Wang, F.-I., and Chu, R.-M. (2003). Identification of Canine Transmissible Venereal Tumor Cells Using in Situ Polymerase Chain Reaction and the Stable Sequence of the Long Interspersed Nuclear Element. J. Vet. Diagn. Invest. *15*, 399–406.

87. Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature *438*, 803–819.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
c-Myc	Santa Cruz Biotechnology, Inc.	Cat# sc-40; RRID: AB_2857941
ACTB Mouse mAb	ABclonal Technology Co., Ltd.	Cat# AC004; RRID: AB_2737399
Anti-mouse IgG, HRP-linked Antibody	Cell Signaling Technology, Inc.	Cat# 7076S; RRID: AB_330924
Biological samples		
CTVT Tissue Biopsies	This paper; Table S1	N/A
Chemicals, peptides, and recombinant proteins		
100×MEM Non-Essential Amino Acids Solution	Gibco, Thermo Fisher Scientific Inc.	Cat# 11140050
2×Phanta® Max Master Mix (Dye Plus)	Vazyme biotech Co., Ltd.	Cat# P312-03
BeyoECL Star	Beyotime Biotechnology Inc.	Cat# P0018AS
DMEM	Gibco, Thermo Fisher Scientific Inc.	Cat# C11995500BT
E.Z.N.A.®Gel Extraction Kit	Omega Bio-tek, Inc.	Cat# D2500
Fetal Bovine Serum	Gibco, Thermo Fisher Scientific Inc.	Cat# 10099141
Halt Protease and Phosphatase Inhibitor Single-Use Cocktail (100×)	Thermo Fisher Scientific Inc.	Cat# 78442
HiScript® III RT SuperMix for qPCR (+gDNA wiper)	Vazyme biotech Co., Ltd.	Cat# R323-01
Lipofectamine® 3000	Invitrogen, Thermo Fisher Scientific Inc.	Cat# L3000-015
McCoy's 5A	Shanghai Yuanpei Biotechnology Co., Ltd.	Cat# L630KJ
PBS buffer	Sangon Biotech (Shanghai) Co., Ltd.	Cat# E607008
Penicillin-streptomycin	Gibco, Thermo Fisher Scientific Inc.	Cat# 15070063
Pierce™ BCA protein assay kit	Thermo Fisher Scientific Inc.	Cat# 23227
polyvinylidene fluoride	Merck Millipore Co., Ltd.	Cat# IPVH00010
QuickBlock™ Blocking Buffer	Beyotime Biotechnology Inc.	Cat# P0252-100mL
RIPA Lysis and Extraction Buffer	Thermo Fisher Scientific Inc.	Cat# 89900
RNAeasy® QIAGEN mini kit	QIAGEN	Cat# 74134
RNAlater™ Stabilization Solution	Invitrogen, Thermo Fisher Scientific Inc.	Cat# AM7021
Sodium Pyruvate 100 mM Solution	Gibco, Thermo Fisher Scientific Inc.	Cat# 11360070
Trypsin	Gibco, Thermo Fisher Scientific Inc.	Cat# 25200-072
type II collagenase	Sigma-Aldrich, Sigma-Aldrich LLC.	Cat# C6885
Critical commercial assays		
Clontech SMARTer® PCR cDNA Synthesis Kit	Takara Bio Inc.	Cat# 634925
Deposited data		
PacBio RNA-seq for CTVT-KM3	This paper	GSA: CRA003315
Illumina RNA-seq for CTVT-KM1, -KM2, -KM3 and -SZ3	This paper	GSA: CRA003315
Illumina RNA-seq for CTVT-5 and CTVT-6	Frampton et al.[22]	ENA: E-MTAB-5488
Illumina RNA-seq for CTVT-761, -765, −766, -772 -774 and −775	Frampton et al.[22]	ENA: E-MTAB-5889
Illumina MeDIP-seq for CTVT-5, -6 and -17	Frampton et al.[22]	ENA: E-MTAB-5495

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Illumina WGS for CTVT-KM1 and -KM2 (tumors and corresponding hosts)	Wang et al.[14]	GSA: CRA000939
Illumina WGS for CTVT-24 and -79 (tumors and corresponding hosts)	Murchison et al.[13]	SRA: PRJEB5068
Illumina WGS for CTVT-608 and -609 (tumors and corresponding hosts)	Leathlobhair et al.[23]	SRA: PRJEB22148
Illumina WES for CTVT tumors and hosts	Baez-Ortega et al.[16]	ENA: ERP109580
Dog reference genome Canfam3.1 (GCA_000002285.2)	Ensembl[30]	http://may2021.archive.ensembl.org/Canis_lupus_familiaris/Info/Index
Raw data and analysis code	This paper	https://doi.org/10.17632/z2xxcw8398.1
Experimental models: Cell lines		
293T cell line	Kunming Cell Bank of the Chinese Academy of Sciences	KCB200744YJ
MDCK cell line	Kunming Cell Bank of the Chinese Academy of Sciences	KCB2006105YJ
Canine transmissible venereal tumor (CTVT) primary cells	This paper	N/A
Oligonucleotides		
Primers used in this study	This paper; Table S7	N/A
Recombinant DNA		
pcDNA3.1™ (+)/UBE4B-CFAP61	Tsingke Biottech Co., Ltd. CC o	N/A
Software and algorithms		
Somatypus (version 1.3)	Baez-Ortega et al.[16]	https://github.com/baezortega/somatypus
STAR-Fusion (version 1.8.1)	Haas et al.[26]	https://github.com/STAR-Fusion/STAR-Fusion/
Trimmomatic (version 0.33)	Bolger et al.[54]	https://github.com/timflutre/trimmomatic/
RSeQC (version 4.0.0)	Wang et al.[55]	http://rseqc.sourceforge.net/
SMRT Tools (version 8.0.0)	Biosciences[56]	http://www.pacb.com/support/software-downloads
NanoPlot (version 1.29.0)	Coster et al.[57]	https://libraries.io/pypi/NanoPlot/
BWA (version 0.7.15-r1140)	Li[58]	https://github.com/lh3/bwa/
samtools (version 1.5)	Danecek et al.[59]	https://github.com/samtools/
GATK (version 4.1.8.0)	Ryan et al.[60]	https://gatk.broadinstitute.org/hc/en-us
Sequenza Utils (version 3.0.0)	Favero et al.[61]	https://sequenzatools.bitbucket.io/#/home
kallisto (version 0.44.0)	Bray et al.[62]	https://pachterlab.github.io/kallisto/
estimate (version 1.0.13)	Yoshihara et al.[63]	https://bioinformatics.mdanderson.org/estimate/
PUREER (version 0.1.0)	Revkov et al.[64]	https://github.com/skandlab/PUREE
vcf2phylip (version 2.0)	Ortiz et al.[65]	https://github.com/edgardomortiz/vcf2phylip
RAxML (version 8.2.9)	Stamatakis[66]	https://github.com/stamatak/standard-RAxML
FigTree (version 1.4.4)	Rambaut[67]	https://github.com/rambaut/figtree
LUMPY (version 0.2.13)	Layer et al.[68]	https://github.com/arq5x/lumpy-sv
Delly (version 1.1.3)	Rausch et al.[69]	https://github.com/dellytools/delly
bedtools (version 2.29.0)	Quinlan et al.[70]	https://bedtools.readthedocs.io/en/latest/index.html#

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SURVIVOR (version 1.07)	Jeffares et al.[71]	https://github.com/fritzsedlazeck/SURVIVOR
FusionCatcher (version 0.99.7b)	Nicorici et al.[72]	https://github.com/ndaniel/fusioncatcher/
UpsetR (version 1.4.0)	Conway et al.[73]	https://cran.r-project.org/web/packages/UpSetR/
minimap2 (version 1.1.0)	Li[73]; Biosciences[56]	https://github.com/PacificBiosciences/pbmm2/
cDNA_Cupcake (version 18.1.0)	Tseng[75]	https://github.com/Magdoll/cDNA_Cupcake
SQANTI3 (version 1.6)	Tardaguila et al.[76]	https://github.com/Magdoll/SQANTI2
TransDecoder (version 5.5.0)	Haas et al.[77]	https://github.com/TransDecoder/TransDecoder
InterProScan (version 5.54–87.0)	Matthias Blum et al.[78]	https://www.ebi.ac.uk/interpro/
ClueGO (version 2.5.7)	Bindea et al.[79]	http://apps.cytoscape.org/apps/cluego
Cytoscape (version 3.9.1)	Paul Shannon et al.[80]	https://cytoscape.org/
fastp (version 0.23.2)	Chen et al.[81]	https://github.com/OpenGene/fastp
Bowtie2 (version 2.3.0)	Langmead et al.[82]	https://github.com/BenLangmead/bowtie2/
DeepTools (version 3.5.1)	Ramírez et al.[83]	https://github.com/deeptools/deepTools
MEDIPS (version 1.50.0)	Matthias et al.[84]	https://bioconductor.org/packages/release/bioc/html/MEDIPS.html
Circos (version 0.69)	Krzywinski et al.[85]	http://circos.ca/software/download/circos/
Other		
PacBio Sequel systems	PacBio, USA	Cat# Sequel II
Illumina HiSeq™ 2000	Illumina, USA	Cat# SY-401-1001
Illumina HiSeq™ X Ten	Illumina, USA	Cat# SY-412-1001

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ya-Ping Zhang (zhangyp@mail.kiz.ac.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The short-read RNA-seq data of samples CTVT-KM1, -KM2, -KM3, -SZ3 and the long-read RNA-seq data of sample CTVT-KM3 have been deposited at the Genome Sequence Archive (GSA) database, and their respective accessions are listed in the key resources table. Publicly accessible data such as WES (for CTVT hosts and tumors), WGS (for the samples CTVT-KM1, -KM2, -24, -79, -608, and -609), and short-read RNA-seq data (for the CTVT samples -5, -6, -761, -765, -766, -772 -774 and -775), as well as MeDIP-seq data (for CTVT-5, -6 and -17), and the Dog reference genome CanFam3.1 (GCA_000002285.2) can be found in their original publications listed in the key resources table.
- The raw data and analysis code used in this study can be accessed at Mendeley Data. The DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Tissue biopsies

All experimental procedures were conducted in compliance with the ethical approvals (certificate numbers: SYDW-20130420-74 and SMKX-20170301-154) granted by the Kunming Institute of Zoology. In brief, CTVT tissue biopsies were collected during anesthesia-assisted surgery at veterinary animal hospitals in Kunming, China. The collected tissues were then preserved in RNAlater Stabilization Solution (Invitrogen, Thermo Fisher Scientific Inc.) containing CTVT-KM1, CTVT-KM2, CTVT-KM3, CTVT-SZ3, CTVT-KM23, CTVT-KM24, CTVT-KM25 and

CTVT-KM26 tissues or PBS buffer (Sangon Biotech Shanghai Co., Ltd.) for CTVT-KM11 tissue as detailed in Table S1. PCR was employed to confirm the CTVT-specific LINE-*MYC* genomic rearrangement for diagnostic purposes.[11,86] Details of the primers and PCR programs can be found in Table S7.

Primary cultures

CTVT primary cells were isolated from the external genitalia of a 1-year-old female dog (CTVT-KM11). The dissociation process followed afterward by cutting the tissue into pieces and then digested for 30 min with two enzymes, 0.25% trypsin (Gibco, Thermo Fisher Scientific Inc.) and 0.2% type II collagenase (Sigma-Aldrich, Sigma-Aldrich LLC.) at 37°C in a 2:1 ratio. The growth medium containing McCoy's 5A (Shanghai Yuanpei Biotechnology) with 10% fetal bovine serum (FBS) (Gibco, Thermo Fisher Scientific Inc.), 1% 100 × MEM Non-Essential Amino Acids Solution (Gibco, Thermo Fisher Scientific Inc.), 1% Sodium Pyruvate 100 mM Solution (Gibco, Thermo Fisher Scientific Inc.) and 1% penicillin-streptomycin (Gibco, Thermo Fisher Scientific Inc.) was added to terminate digestion. CTVT cells were obtained and cultured in growth medium at 37°C in a saturated humidified atmosphere containing 5% CO_2. The CTVT primary cells were passaged every 3–5 days, depending on the confluency of the cells to reach at least 80–90% confluence in the cultured dish.

The control 293T and MDCK cell lines were purchased from Kunming Cell Biobank of the Chinese Academy of Sciences (Kunming, China) and grown in DMEM (Gibco, Thermo Fisher Scientific Inc.) containing 10% FBS (Gibco, Thermo Fisher Scientific Inc.) at 37°C and 5% CO_2

METHOD DETAILS

Nucleic acid extraction and sequencing

Total RNA was extracted from each CTVT tumor biopsy or cell using the RNAeasy QIAGEN mini kit (QIAGEN). The paired-end reads from RNA extracts of CTVT-KM1, CTVT-KM2, CTVT-KM3 and CTVT-SZ3 were 2× 104 bp, 2× 104 bp, 2× 150 bp and 2× 150 bp, respectively. Transcriptomic sequencing was then performed on Illumina HiSeq 2000 or X Ten platforms according to the manufacturer's instructions.

Additionally, the RNA extracts of CTVT-KM3 tumor biopsy were sent to Biolinker Technology (Kunming) Co., Ltd. and Annoroad gene technology (Beijing) Co, Ltd. for long-read transcriptome sequencing using PacBio SMRT sequencing technology. A non-sizable selected Iso-Seq library (< 4 kb) together with a sizable selected Iso-Seq library (> 4 kb) were constructed and pooled using the Clontech SMARTer PCR cDNA Synthesis Kit (Takara Bio Inc.) according to manufacturer's instructions for sequencing on PacBio Sequel II System.

Sequencing reads pre-processing and quality control

For Illumina RNA-seq data, we used *Trimmomatic* (version 0.33)[54] for trimming adapters from the raw sequencing reads with the flag options "*LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36*" and *RSeQC* (version 4.0.0)[55] for visualizing the quality of the filtered reads. For PacBio RNA-seq data, we used the *SMRT Tools* (version 8.0.0)[56] software suite to process the raw data. BAM files obtained from PacBio subreads were polished for circular consensus sequences (CCS) using the command *ccs* with optional flags "*–minPasses 1, –min-rq 0.99 and –skip-polish*". These CCS BAM files were then used by (1) *lima* to remove cDNA primers; (2) *isoseq3 refine* to remove polyA tail and artificial concatemers, (3) *isoseq3 cluster* to *de novo* isoform-level clustering, and (4) *isoseq3 polish* to polish the transcripts. Finally, *NanoPlot* (version 1.29.0)[57] was used to plot the filtered sequences for further quality check and evaluation.

Tumor cellularity, ploidy and CNV analysis

DNA sequencing reads form CTVTs (CTVT-KM1 and -KM2: ~40 ×; −79: ~60 ×; −24, 608 and 609: ~100 ×) were aligned to the CanFam3.1 (GCA_000002285.2) reference genome (downloaded from the Ensembl website[30]) using *BWA mem* (version 0.7.17-r1188)[58] with the flag commands "*-M* (mark shorter split hits as secondary)" and "*-a* (output all alignments for paired-end reads)". Notably, the published sequenced data for CTVT-24 and CTVT-79 integrated in our study were re-aligned against CanFam3.1 reference genome.[13,87] The mapping coverage was assessed with *samtools depth* (version 1.5).[59] Recalibration of the base scores in the bam files was done by *GATK BaseRecalibrator* (Generates recalibration table for Base Quality Score Recalibration) and *ApplyBQSR* (Apply base quality score recalibration) (version 4.1.8.0).[60]

Sequenza Utils (version 3.0.0)[61] was used to estimate the cellularity, ploidy purity and CNV. The GC content of CanFam3.1 reference genome were calculated in the 50 bp window size by *sequenza-utils gc_wiggle*. We applied *sequenza-utils bam2seqz* to generate seqz for each CTVT tumor and its corresponding host BAM file. Subsequently, *sequenza-utils seqz_binning* was used to bin the seqz into 50 bp windows. Finally, cellularity and ploidy were estimated using the R packages *sequenza* (version 3.0.0).[61] The gender parameter was assigned according to the tumor host's gender. CNV profiles were then estimated using the estimated cellularity and ploidy results.

Tumor purity analysis

For WGS data, the minimum value of cellularity obtained from *sequenza* can be used as an estimate of the tumor purity for the CTVT. For RNA-seq data, there is no corresponding host, so we can estimate tumor purity using gene expression matrix. We first used *kallisto index* (version 0.44.0)[62] to build an index for each Ensembl gene (version 97).[30] Then, we utilized *kallisto quant* with the parameter "*-b100*" (Number of bootstrap samples) and "*–fusion*" to calculate the transcripts per million (TPM) for each CTVT sample. Afterward, we converted the canine Ensembl gene ID in the expression matrix to their human homologous gene. We then utilized the R package *estimate* (version 1.0.13)[63] and the Python software *PUREE* (version 0.1.0)[64] to calculate the tumor purity for each CTVT sample. The average value derived from both tools was considered as the estimated tumor purity for that RNA sample.

SNVs calling and genotyping

Genetic variants were called by *Somatypus* (version 1.3)[16] using default settings. RNA samples were excluded from subsequent analyses because of their low coverage. We selected 484 corresponding host dogs from Adrian Baez-Ortega et al.'s results,[16] and included CTVT-KM1H, -KM2H, −24H, −79H, −608H and −609H to construct a reference panel of 488 host dogs. Next, we merged the 539 CTVT tumors published by Adrian Baez-Ortega et al. to create a final dataset of 543 tumors from 44 different countries and ensured that only variants were present in at least one CTVT tumor. Any variants present in the host reference panel were removed, meaning that only CTVT-specific SNVs were retained. These SNVs were then aligned with clocklike exonic somatic variant dataset to infer CTVT phylogenetic tree.

Next, we corrected the genotyping results for each gene variant in each CTVT tumor using the method described by Adrian Baez-Ortega et al.[16] Specifically, we determined the heterozygosity of each variant by calculating the depth of reads (FORMAT/NR, the number of reads covering the variant location in that sample), the number of supported reads (FORMAT/NV, the number of reads containing variants) and the tumor purity at each locus (readdepth × tumorpurity/2).[16]

Conditions	Unphased genotype
NR * purity/2 < 6	GT = 0/1
NR * purity/2 ≥ 6 and NV ≥3	GT = 1/1
others	GT = 0/0

CTVT phylogenetic tree inference

We used *vcf2phylip* (version 2.8)[65] to convert SNV genotypes in VCF format to a matrix for phylogenetic analysis in PHYLIP. Additionally, we introduced the CanFam3.1 reference genome to determine the root of the tree. Phylogenetic tree inference was performed using *RAxML* (version 8.2.9)[66] with the options "*–print-identical-sequences -f d -m GTRGAMMA -p 272730* (a generalised time reversible substitution model with a Gamma model)". Use *FigTree* (version 1.4.4)[67] for visualization.

Structural variants identification and filtering

Structural variants were inferred and genotyped from the WGS data using *LUMPY* (version 0.2.13)[68] and *DELLY* (version 1.1.3).[69] We run *lumpyexpress* jointly on tumor-normal pairs with pre-extracted splitters and discordance factors using default parameters, and then employed *svtyper* to call genotypes using a Bayesian maximum likelihood algorithm on the *LUMPY* output VCF files. We used *delly call* to discover somatic SVs for each tumor-normal pair (using default parameters) and pre-filtered with "*-f somatic*" using *delly filter*. Then, any host variants were removed, and the following filtering strategy was adopted.

LUMPY	DELLY
QUAL ≥ 100	QUAL ≥ 100
Genotype(GT) ! = 0/0	Genotype(GT) ! = 0/0
Read depth (DP) > 10	Per-sample genotype filter (FT) = = 'PASS'
Alternate allele observations (AO) = = 0	Raw high-quality read counts or base counts for the SV (RC) > 0
–	high-quality reference junction reads (RR) > 0
–	high-quality variant junction reads (RV) > 0

After that, we used *bedtools intersect* (version 2.29.0)[70] to obtain the overlapping SVs from *LUMPY* and *DELLY* for each CTVT and annotated the merged VCF file to the canine Ensembl gene (version 97) by *sansa annotate*.[69] WGS-based GFs were analyzed by script *extractFuionsFromSansa.py* (Mendeley Data). Finally, we merged all SVs into a single event using *SURVIVOR merge* (version 1.07)[71] with parameter "*500 2 1 0 0 50* (allows distances < 500 bp, supported by 2 callers, agreed on the same type of SV, different strands, not estimate distance by SV size and only compare SV that are at least 50 bp)". Any SV found in the WGS-derived canine Variation and Systematic Error Catalog (VSEC) which is a dataset containing 31,613 SVs from 186 canids were removed so as to remain with only CTVT-specific SVs.[15]

Gene fusions identification

Discovery of CTVT fusion candidates was done based on raw RNA sequencing data with two independent methods, including *FusionCatcher* (version 0.99.7b)[72] and *STAR-Fusion* (version 1.8.1).[26] For *FusionCatcher*, we used *fusioncatcher-build* to create index files for CanFam3.1 reference genome and called GFs by *FusionCatcher* with parameter "*–aligners=blat, star,bowtie2,bowtie,bwa* (the aligners used)". This software uses ENSEMBL database to find novel/known GFs, COSMIC, TICdb, ChimerDB, Cancer Genome Project (CGP), and ConjoinG to

annotate the obtained GFs. For *STAR-Fusion*, we used a Perl script *prep_genome_lib.pl* to prepare the dog CTAT genome library, which is a resource library containing CanFam3.1 reference genome, annotation files, pfam and dfam database that support fusion-finding. Next, we run *STAR-Fusion* with typical parameter to predict GFs. For each CTVT, we merged the output results from the two methods based on the GF IDs (5′ ENSEMBL ID - 3′ ENSEMBL ID) to generate a unified calling of candidate GFs.

Next, we followed the above-mentioned process to obtain candidate GFs in canine Histiocytic sarcoma.[25] We also incorporated three previously identified canine tumor types (Histiocytic sarcoma, Dermatofibrosarcoma protuberans-like, Anaplastic oligodendroglioma and Lymphomas) from GF databases.[19] This resulted in obtaining comprehensive BGFs dataset comprising of four non-transmissible canine tumor GFs from different tissues and embryonic regions. Any BGFs were excluded to obtain CTVT-specific GFs. To identity CTVT RGFs, we calculated the occurrence frequency of each GF by considering the number of samples in which the fusion gene was detected out of the total number of samples (number of GF samples/total number of samples). We defined RGFs as those GFs with a frequency greater than 0.5. To identity CTVT germline fusion genes, we integrated WGS-based fusion results with the CTVT phylogenetic tree. We defined GGFs if they meet the following criteria: (1) belongs to RGFs, (2) detected in DNA data, (3) present in the CTVT founder population, and (4) detected in multiple CTVT sublineages. These GGFs indicate their ancestral origin and inherited characteristics within the CTVT lineage. Use UpsetR (version 1.4.0)[73] for visualization.

Furthermore, we validated the candidate fusion genes using long-read transcriptome sequencing. Polished CCS BAM files were mapped to CanFam3.1 reference genome by *Minimap2* (version 1.1.0)[56,74] with "*-ax splice -uf –secondary=no -C5* (default parameters)". Fusion transcripts were detected using *fusion_finder.py* with "*-c 0.05* (minimum per-locus coverage in percentage) *-t 0.95* (minimum total coverage) *-d 50* (minimum distance between loci) *–dun-merge-5-shorter* (not collapse shorter 5′ transcripts)" from *cDNA_Cupcake* (version 18.1.0).[75] Fusion transcripts were classified, annotated and curated in GFF format by *sqanti3_qc.py* from *SQANTI3* (version 1.6)[76] with parameter "*–is_fusion*". *fusion_collate_info.py* from *cDNA_Cupcake* was used to collate and filter candidate fusion transcripts. Candidates that share the same gene segments are considered to be the same fusion event.

Gene fusions expression analysis

As mentioned earlier, we used *kallisto index* to build index for the full-length fusion transcript sequences, and then performed expression quantification of each CTVT using *kallisto quant*. The *log* transformation of TPM values was calculated as $log_2(TPM+1)$ and the resulting scores were normalized using *z*-sore. We used the nonparametric Mann-Whitney *U* test to determine the significance of GF expression at different phase. PCA was performed by *prcomp* in *R*. The correlations of fusion transcripts between progressive, regressive and non-regressive CTVTs were calculated using Spearman's rank correlation coefficient. The significance of these statistics was considered at the *p* value threshold level 0.05.

Cross-species comparative analysis

The human GFs were obtained from The Tumor Fusion Database. Using the two partner genes of CTVT RGFs, we searched for all relevant GFs in 33 cancer types in TCGA database and extracted each GF sequences based on the reported fusion points. *TransDecoder* (version 5.5.0)[77] was used to identify candidate coding regions for GFs. The longest protein sequence is considered representative of a particular fusion event. The protein domains of peptides were searched using *InterProScan* (version 5.54–87.0),[78] including all member databases. GFs domains that are shared or similar between human cancers and CTVT were investigated.

Enrichment analysis

Functional enrichment analysis was carried out by *ClueGO* (version 2.5.9)[79] from *cytoscape* (version 3.9.1).[80] We used 0.4 kappa score, 2% GO term/pathway selection and 0.05 *p* value of as significant threshold levels. To test for statistical significance adjusted or corrected *p* value was computed using Bonferroni step down two-sided hypergeometric test.

DNA methylation analysis

The quality of the MeDIP-seq data was performed using *fastp* (version 0.23.2).[81] Reads were also aligned to CanFam3.1 using *bowtie2* (version 2.3.5).[82] *deepTools* (version 3.5.1)[83] was used to analyze the methylation levels in CTVT GGFs. *bamCoverage* within this tool was utilized to convert BAM files to the bigWig format, and reads per kilobase per Million (RPKM) mapped reads was used for normalization. *computeMatrix scale-regions* was used to calculate the methylation scores for each genomic region, with the parameter "*– upstream 3000 –downstream 3000 –binSize 50 – regionBodyLength* 3000".[83] *plotHeatmap* from *deeptools* was used to visualize the analysis results. R-Package *MEDIPS* (version 1.50.0)[84] was used for basic data processing, quality controls, normalization, and identification of differentially methylated regions (DMRs) in CTVT before and after chemotherapy with parameter "*extend=0* (the number of bases by which the region will be extended before the genome vector is calculated), *shift=0* (shift by the specified number of nucleotides with respect to the given strand information), *uniq=1e-3* (cap the number of stacked reads), *window_size=50* (the genomic resolution), *minRowSum=2* (threshold for the sum of counts), *diff.method=edgeR* (method for calculating differential coverage), *paired=T* (paired end reads)". Use *Circos* (version 0.69)[85] for visualization.

PCR and sanger sequencing

cDNA was synthesized from total RNA extracts using HiScript III RT SuperMix for qPCR (+gDNA wiper) (Vazyme Biotech Co., Ltd.). Real Time Quantitative PCR (RT-qPCR) was performed by applying ChamQ Universal SYBR qPCR Master Mix (Vazyme Biotech Co., Ltd.) using QuantStudioTM 5 Real-Time PCR instrument (Applied Biosystems, USA). After the calculation of the cycle threshold (Ct) value for each sample, quantitative expression results were obtained according to the $2^{-\Delta\Delta ct}$ method. The *ACTB* (Actin Beta) was used as an endogenous control gene for normalizing the expression of target genes. Each sample was analyzed in triplicate. Details were shown in the Table S7. Additionally, we performed PCR using 2 × Phanta Max Master Mix (Dye Plus) (Vazyme Biotech Co., Ltd.) to amplify the gene fusion *UBE4B-CFAP61*. Gel recovery used the E.Z.N.A.Gel Extraction Kit (Omega Bio-tek, Inc.).

Plasmid transfection and western blotting

The pcDNA3.1 (+) vector (Invitrogen, Thermo Fisher Scientific, Inc.) was used to construct a recombinant plasmid (Tsingke Biological Technology) that allows the expression of fusion protein *UBE4B-CFAP61* with *Myc*-6 × *His* tag. When CTVT, 293T and MDCK cells reached about 80% confluence, cells were seeded in 6-well culture plates at ~5 × 10^4 cells/well in growth medium. After 24 h, cells were transfected with 2 μg pcDNA3.1(+)/*UBE4B-CFAP61* plasmid, and pcDNA3.1(+) empty plasmid (Mock) served as a control. The process of transfection was conducted using Lipofectamine 3000 (Invitrogen, Thermo Fisher Scientific, Inc.) according to the manufacturer's instructions. The lipofectamine complex was removed 6 h after transfection, and the cells were cultured continuously for 48 h.

RIPA Lysis and Extraction Buffer (Thermo Fisher Scientific Inc.) and Halt Protease and Phosphatase Inhibitor Single-Use Cocktail (100 ×) was used to lyse the cells and extract protein and Pierce BCA protein assay kit (Thermo Fisher Scientific Inc.) was used to determine protein concentration. Protein was separated by SDS-PAGE and transferred to polyvinylidene fluoride (PVDF) membranes (Merck Millipore Ltd.). Membranes were blocked in QuickBlock Blocking Buffer (Beyotime Biotechnology Inc.) for 2 h and incubated overnight at 4 °C with primary antibodies: c-Myc (1:500, Santa Cruz Biotechnology), ACTB Mouse mAb (1:1000, ABclonal Technology). The membranes were then incubated with secondary antibodies for 2 h at room temperature: Anti-mouse IgG, HRP-linked Antibody (1:1000, Cell Signaling Technology). Images were developed with BeyoECL Star (Beyotime Biotechnology Inc.) for 1 min and obtained by chemiluminescence using Gel DocTM XR+ (Bio-Rad, USA).

QUANTIFICATION AND STATISTICAL ANALYSIS

FFPM (fusion fragments per million total reads) was calculated as follows[26]:

$$FFPM = \frac{counts\ of\ junction\ reads + counts\ of\ spanning\ reads}{total\ number\ of\ reads} \times 10^6$$

For expression analyses, *log* transformation of TPM values was calculated as $log_2(TPM+1)$ and *z*-sore was used for normalization. For enrichment analyses, we used 0.4 kappa score and 0.05 *p* value as significant threshold levels. Two-side hypergeometric test was used to correct for p values using Bonferroni step down statistical method. We used the non-parametric Mann-Whitney *U* test to determine the significance of GFs. The correlations of GGFs between CTVTs were calculated using Spearman's rank correlation coefficient. For methylation analysis, we performed differential coverage analysis with edgeR and corrected *p* values with Bonferroni step down method. The minimal number of reads per window was 2. The significance of all statistics was considered at p value < 0.05.