Comprehensive genome annotation of the model ciliate *Tetrahymena thermophila* by in-depth epigenetic and transcriptomic profiling

Fei Ye[1, 2, #], Xiao Chen[3, 4, #], Aili Ju[1, 2, #], Yalan Sheng[5], Lili Duan[1, 2], Khaled A. S. Al-Rasheid[6], Naomi A. Stover[7], Shan Gao[1, 2 *]

[1]MOE Key Laboratory of Evolution & Marine Biodiversity and Institute of Evolution & Marine Biodiversity, Ocean University of China, Qingdao 266003, China

[2]Laboratory for Marine Biology and Biotechnology, Qingdao Marine Science and Technology Center, Qingdao 266237, China

[3]Laboratory of Marine Protozoan Biodiversity & Evolution, Marine College, Shandong University, Weihai 264209, China

[4]Suzhou Research Institute, Shandong University, Suzhou 215123, China

[5]Shum Yiu Foon Shum Bik Chuen Memorial Centre for Cancer and Inflammation Research, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong, SAR, China

[6]Zoology Department, College of Science, King Saud University, Riyadh 11451, Saudi Arabia

[7]Department of Biology, Bradley University, Peoria, Illinois 61625, USA

#These authors contributed equally.

*Corresponding author (Shan Gao, email: shangao@ouc.edu.cn)

Abstract

The ciliate *Tetrahymena thermophila* is a well-established unicellular model eukaryote, contributing significantly to foundational biological discoveries. Despite its acknowledged importance, current *Tetrahymena* biology studies face challenges due to gene annotation inaccuracy, particularly the notable absence of untranslated regions (UTRs). To comprehensively annotate the *Tetrahymena* macronuclear genome, we collected extensive transcriptomic data spanning various cell stages. To ascertain transcript orientation and transcription start/end sites, we incorporated data of epigenetic marks displaying enrichment towards the 5' end of gene bodies, including H3 lysine 4 tri-methylation (H3K4me3), H2A.Z, nucleosomes, and N^6-methyldeoxyadenine (6mA). Additionally, we integrated Nanopore direct sequencing (DRS), strand-specific RNA-seq, and ATAC-seq data. Using a newly-developed bioinformatic pipeline, coupled with manual curation and experimental validation, our work yielded substantial improvements to the current gene models, including the addition of 2,481 new genes, updates to 6,257 existing genes, and the incorporation of 5,917 alternatively spliced isoforms. Furthermore, novel UTR information was annotated for 26,223 high-confidence genes. Intriguingly, 16% of protein-coding genes were identified to have natural antisense transcripts (NATs) characterized by high diversity in alternative splicing, thus offering insights into understanding transcriptional regulation. Our work will enhance the utility of *Tetrahymena* as a robust genetic toolkit for advancing biological research.

Keywords: *Tetrahymena*, genome annotation, untranslated regions (UTRs), epigenetic information, natural antisense transcript (NATs)

INTRODUCTION

Tetrahymena thermophila (hereafter referred to as *Tetrahymena*) is a well-recognized unicellular model eukaryote and serves as a cornerstone for numerous scientific discoveries [1-11]. Like other ciliates, *Tetrahymena* maintains two functionally distinct nuclei, the diploid micronucleus (MIC) containing five pairs of chromosomes and the polyploid macronucleus (MAC) comprising 181 chromosomes [3, 12-16]. Notably, the MIC remains transcriptionally inactive until the occurrence of sexual reproduction (conjugation), while the MAC is active throughout the vegetative stage to meet the cellular demands [17-19]. The MIC and MAC are generated from the same zygotic nucleus during conjugation [18].

The first assembly of the *Tetrahymena* macronuclear genome was reported in 2006, employing the shotgun sequencing technique [20, 21]. Subsequent versions were published in 2008 and 2014, through the application of next-generation sequencing [22, 23]. Most recently, the contiguity of the MAC genome was substantially improved, ultimately leading to a complete assembly, by using the PacBio Single-Molecule Real-Time (SMRT) sequencing technology [24]. Along with advancements in genome assembly, multiple efforts were dedicated to improving gene annotation, using various datasets from complementary DNA (cDNA) library [20], EST [23], microarray [25], and RNA-seq [26], as well as manual curations. All these genome assembly and gene annotation data were deposited in *Tetrahymena* Genome Database (TGD; Ciliate.org) [22], including two major updates TGD2014 [23] and TGD2021 [24].

However, even the most updated gene model (TGD2021) remained incomplete in several aspects. We found many instances where intron-exon boundary junctions were not accurate, annotated protein-coding genes were fusions of two independent transcription units or needed to be fused with others, or the putative genes were not supported by any RNA-seq reads. In addition, the annotation of untranslated regions (UTRs) was strikingly lacking. Few UTRs were included in the original annotation version when microarray was used for genetic target selection [27]. With the help of Illumina RNA-seq data, the number of genes with 5' UTRs and/or 3' UTRs increased to 6,676 [26]; these UTRs data, however, were not integrated into subsequent annotation versions. In TGD2014, only 1,447 genes were annotated with UTRs, representing merely ~5% of all protein-coding genes [23]. In TGD2021, scarcely any genes were annotated with UTR information [24].

In order to optimize the *Tetrahymena* MAC genome annotation, we generated RNA-seq data from different cell stages, accumulating an ultra-deep

sequencing dataset to detect low-expression genes and cell stage-specific genes. Most importantly, we incorporated the distribution information of multiple epigenetic marks, including the histone modification H3K4me3 [5, 28], the histone variant H2A.Z [29], nucleosomes [5, 29], and N^6-adenine DNA methylation (6mA) [28-30]. All these marks displayed the preferential accumulation towards the 5' end of the gene body, and were thus helpful in determining the gene orientation and predicting transcription start sites (TSSs). We also integrated Nanopore direct sequencing (DRS) data, strand-specific transcriptomic data, and ATAC-seq data. Based on computational prediction, manual editing, and experimental verification, we have produced a comprehensive annotation of protein-coding genes in the MAC genome of *Tetrahymena*, offering improved precision in intron-exon boundaries, TSS, transcription end sites (TESs), and UTRs. We also performed a preliminary analysis of natural antisense transcripts (NATs), which will help to better resolve the regulation of transcription in *Tetrahymena*.

RESULTS & DISCUSSION

Optimize gene model with transcriptomic data

To validate the TGD2021 gene models and identify potential novel genes in the *Tetrahymena* MAC genome, we assembled RNA-seq data from different cell stages, including growth, multiple timepoints during starvation and conjugation (Additional File 2: Table S1). We initially employed LoReAn2 [31], an integrated annotation pipeline, to annotate protein-coding genes. The average lengths of predicted coding regions (3,900 bp vs. 2,452 bp) and intergenic regions (5,550 bp vs. 1,456 bp) were both considerably longer than those in TGD2021 (Additional File 3: Table S2). However, the number of protein-coding genes was notably lower (15,355 vs. 26,259) and only 8,351 of these genes contained UTR information. Moreover, these predicted coding regions covered only 37.61% of the entire genome (38.87 Mb out of 103.34 Mb), much lower than the coverage in TGD2021 (64.38 Mb, 62.30%) [24]. The incompleteness of the gene annotation was further manifested by the lower mapping ratio of RNA-seq reads (56.74% vs. 82.07% in TGD2021), suggesting that a large proportion of genes were not annotated by LoReAn2. Collectively, the unsuccessful annotation by LoReAn2 [31] prompted us to develop a more efficient approach for the *de novo* annotation of the *Tetrahymena* MAC genome.

Here, we employed a newly developed pipeline (Fig. 1, Additional File 1: Fig. S1), which enabled us to identify a total of 27,369 gene candidates (draft v1).

Many gene candidates (17,170 out of 27,369, 63%) shared identical intron-exon boundary junctions with TGD2021 and were thus temporarily considered well-annotated genes. For the remaining gene candidates, we further optimized their annotations using full-length transcripts obtained from Nanopore direct RNA sequencing (DRS), strand-specific RNA-seq (ssRNA-seq), and the most highly expressed RNA-seq transcripts among all cell stages. 3,408 new genes were identified, mostly located within intergenic regions as previously defined in TGD2021 (Fig. 2A, Additional File 1: Fig. S2A).

Most importantly, we optimized gene annotations for a substantial number of genes (7,817). These optimizations fell into four major classes. 1) Exon-altered genes: 4,296 genes had altered intron-exon boundaries (Fig. 2B, Additional File 1: Fig. S2B). 2) Fused genes: 2,858 genes were merged accordingly and annotated as 1,314 genes. These mergers were supported by RNA-seq reads and DRS reads spanning two neighboring genes (Fig. 2C, Additional File 1: Fig. S2C). 3) Partitioned genes: 518 genes were separated into 1,036 genes, based on RNA-seq reads that were interrupted in the middle of these genes, with no RNA-seq reads spanning the neighboring genes (Fig. 2D, Additional File 1: Fig. S2D). 4) Orientation-reversed genes: the orientation of 145 single-exon genes was changed according to the strand-specific reads (Fig. 2E).

1,271 genes in TGD2021 were defined as low-confidence genes (Fig. 2F). First, there were 86 genes for which no RNA-seq reads were detected among all our data, suggesting that these genes either have extremely low expression levels or their expression is highly specific to certain conditions. The prediction of these genes was more likely attributable to errors in the *ab initio* annotation, given the absence of supports from previous datasets, including EST, microarray, and Illumina RNA-seq data [20, 23, 25, 26]. Second, for 1,185 genes, their sequenced reads did not align well with the original annotation, especially at intron-exon boundaries. Despite the presence of RNA seq reads, new transcripts failed to be assembled owing to its low sequencing depth and occasionally mixed antisense reads.

At this stage, we identified a total of 28,640 genes (draft v2) (Additional File 1: Fig. S1), encompassing 17,170 well-annotated genes, 3,408 new genes, 4,296 exon-altered genes, 1,314 fused genes, 1,036 partitioned genes, 145 orientation-reversed genes, and 1,271 low-confidence genes.

Refine gene model using epigenetic marks

To further enhance the accuracy of gene models optimized by transcriptomic data, we developed a machine learning algorithm to utilize information from epigenetic marks (Fig. 1B, 3A). These marks, including H3K4me3, H2A.Z, 6mA, and nucleosome positioning, all exhibited the preferential enrichment at the 5' end of actively transcribed genes [5, 28-30, 32], thus providing valuable guidance for predicting TSSs. For model training and evaluation, 10,460 long genes (> 1kb) were selected from a pool of 17,170 well-annotated genes (see more details in Methods and Materials). Using the trained Random Forest (RF) model (Fig. 3B), 27,840 TSS regions were predicted.

ATAC-seq fragments from the nucleosome-free region (NFR) tended to enrich on gene promoters around TSS (Additional File 1: Fig. S3A) [33]. From our ATAC-seq data, we identified 42,469 significant broad peaks at NFR, and the center of each peak was defined as a candidate TSS. Of these candidate TSSs, those located within 200 bp of the TSS regions predicted by our RF model were defined as epigenetic data supported TSSs (eTSSs). Those located within 200 bp of 5' end of genes, lacking support from our RF model, were defined as potential TSSs (pTSSs) (Additional File 1: Fig. S3B).

Among 28,640 genes optimized by transcriptomic data (draft v2), 25,617 possessed either eTSS (21,095) or pTSS (4,522) (Additional File 1: Fig. S1). 3,937 genes had multiple eTSSs and they were subsequently subjected to manual curation (Fig. 5A, B). Interestingly, 885 head-to-head gene pairs (1,670 genes) shared their respective eTSSs, indicating that they utilized a bidirectional promoter for transcription (Additional File 1: Fig. S3C). We also found 1,752 genes with neither eTSS nor pTSS. These genes were either tandem duplicate genes (Additional File 1: Fig. S3D) or duplicated genes with multiple short exons (Additional File 1: Fig. S3E). These duplicated genes were also subsequently subjected to manual curation (Fig. 5C, D). For the 1,271 low-confidence genes poorly supported by RNA-seq reads, neither eTSS nor pTSS were found in close proximity to them (Additional File 1: Fig. S3F), further confirming that they are either silent genes or error genes [20].

Based on eTSSs with high confidence, we reexamined the gene model (draft v2) that has been refined by transcriptomic data (Fig. 1B). 13 genes were identified as new genes based on the presence of eTSSs (Fig. 3C, Additional File 1: Fig. S4A). These genes were lowly expressed, limited to only one developmental stage, and were not originally annotated by our pipeline due to the scarcity of supporting reads in the stage-combined RNA-seq dataset.

Meanwhile, annotations of multiple genes were optimized based on eTSS, complemented by transcriptomic data (Fig. 1B). 1) Orientation-reversed genes: the orientation of 24 single-exon genes was reversed, because their eTSSs were located within the previously annotated 3' UTRs (Fig. 3D). Their orientation could not be determined by ssRNA-seq reads, because they were not expressed during the growth stage when the ssRNA-seq was conducted. 2) TSS-altered genes: the TSSs of 65 genes were altered according to the positions of their eTSSs. The gaps between eTSSs and TSSs predicted by transcriptomic data were attributed to the limited RNA-seq read coverage. Consequently, their TSSs were extended to align with eTSSs, supported by limited yet existing RNA-seq reads (Fig. 3E). 3) Fused genes: 146 genes were merged into 73 genes. These genes were initially misclassified into two separate genes primarily attributed to minor gaps between two clusters of RNA-seq reads. However, only one of the constituent genes contained a well-defined eTSS, while the other lacked any discernible eTSS or pTSS. The surrounding genes each had respective eTSSs, thus eliminating their chances to be merged with other genes (Fig. 3F, Additional File 1: Fig. S4B). 4) Partitioned genes: 67 genes were split into 134 genes. These genes contained two different eTSSs that were divided into three subgroups: (a) co-directional, 19 genes had two eTSSs either simultaneously at the 5' end and the middle of the previously annotated genes or at the 3' end and the middle (Fig. 3G, Additional File 1: Fig. S4C), representing two genes transcribed in the same direction; (b) tail-to-tail, 43 genes had two peaks at both 5' and 3' ends, respectively (Fig. 3G, Additional File 1: Fig. S4D), representing two convergent genes proceeding in opposite directions and towards each other; and (c) head-to-head, 5 genes had two close yet separated peaks in the middle of the gene body (Fig. 3G, Additional File 1: Fig. S4E), representing two divergent genes proceeding in opposite directions and away from each other.

Compared to draft v2, draft v3 (Additional File 1: Fig. S1) contained 13 new genes and 255 optimized genes including 24 TSS-altered genes, 73 fused genes, 134 partitioned genes, and 24 orientation-reversed genes.

UTR annotation and regulatory elements analysis

By employing Nanopore DRS data (Additional File 1: Fig. S5A), we identified TES, defined as the 3' cleavage/polyadenylation site before the poly-A tail [34] [35, 36], in 75% (21,660 out of 28,647) genes. 1,915 genes harbored multiple TESs (Additional File 1: Fig. S5B). For the genes in draft v3 with well-defined TSS and TES, we predicted coding DNA sequences (CDSs) and open reading

frames (ORFs) according to the ciliate genetic code [37] for 28,647 genes. 1,153 genes with no predictable ORF were classified as potential non-coding RNA (Fig. 4A). We then defined the regions on both sides of transcripts, excluding the CDS, as 5' UTRs and 3' UTRs respectively (Fig. 1C, Fig. 4A). In total, 25,898 genes had both 5' UTRs and 3' UTRs, 85 genes had only 5' UTRs, and 240 genes only had 3' UTRs. The 1,271 low-confidence genes did not have annotated UTR information (Fig. 4A). The average lengths of 5' UTRs and 3' UTRs were 192.54 bp and 238.61 bp (Fig. 4B), respectively. Moreover, the inclusion of more precise and reliable UTR information also increased the mapping ratio of RNA-seq reads (82.07% in TGD2021 vs. 91.87% in the updated gene model).

In the proximal promoter sequences surrounding TSS, we identified several core promoter motifs that may play a role as *cis*-elements (Fig. 4C, Additional File 5: Table S4). They contained key motifs involved in transcription activation, such as CCAAT box [38], TATA box [39], and CRE (cAMP response element) (TGACGTCA) [40], and involved in nucleosome positioning (Reb1: CGGGTAA) [41, 42].

In metazoans, a predominant polyadenylation signal (PAS) was observed within the region spanning 0 and 50 bp upstream of the RNA cleavage site [43-45]. In *Tetrahymena*, PAS also consisted of a primary dominant AATAAA motif, along with six variant motifs ATTAAA, AATGAA, AATAGA, CATAAA, GATAAA, and AAAAAG (Fig. 4D) [43]. However, there was a pronounced AT motif in *Tetrahymena* (Fig. 4E), in contrast to the CA motif at the cleavage site in mammals [46]. In metazoans, GT-rich elements (GTGT) were observed both upstream and downstream of the cleavage site [47]. In *Tetrahymena*, however, T-rich sequences were observed within 20 bp downstream and AT-rich beyond 30 bp upstream (Fig. 4E, Additional File 1: Fig. S5C). This suggests that *Tetrahymena* may have distinct mRNA cleavage and polyadenylation mechanisms compared to metazoans.

Additionally, we found that the length of poly-A tails peaked at approximately 18 nt in *Tetrahymena*, similar to *Arabidopsis* (~19 nt), soybean (~19 nt), maize (~18 nt) and rice (~18 nt) [48] (Fig. 4F). When analyzing the longest poly-A sequences of each gene, it was observed that their poly-A length of genes exhibited two prominent peaks at 13-30 nt and 95-100 nt, respectively (Fig. 4G). To investigate whether functional classes of genes are associated with the length of poly-A tails, we sorted all genes by the length of their longest poly-A tails from shortest to longest and divided them into three groups: 1) the first 25% of genes, defined as short-tailed genes, with poly-A lengths ranging from 5-19 nt; 2) the middle 25%-75% of genes, defined as medium-tailed genes, with poly-A lengths

between 19-239 nt; and 3) the remaining 25% of genes, defined as long-tailed genes, with poly-A lengths exceeding 239 nt (Fig. 4H). Gene ontology (GO) analysis revealed that short-tailed genes were highly enriched in the pathway of membrane and ion transport, while long-tailed genes were more prominently enriched in functions related to mitochondria, translation, RNA processing, and ribosome (Fig. 4H, I, Additional File 6: Table S5). This was in strong contrast to *Caenorhabditis elegans* and mammals, wherein short-tailed genes were highly enriched for genes involved in translation, nucleosome, and ribosome [49]. Additionally, we identified a positive correlation between lengths of gene poly-A tails and their expression levels (rho =0.72, *P* <2.2e-16) (Fig. 4J) in *Tetrahymena*, suggesting that long poly-A tails stabilize mRNA [50, 51]. This contrasted with the previous finding in *Caenorhabditis elegans*, where highly expressed mRNAs were observed to have shorter poly-A tails, explained by enhanced translation efficiency and the maintenance of an optimal tail length [49]. No correlation was observed between gene poly-A length and gene length (rho= -0.089, *P* <2.2e-16) (Additional File 1: Fig. S5D). The discrepancy between *Tetrahymena* and other eukaryotes suggested functional diversification of poly-A tails across different species.

In this version of gene models (draft v4) (Additional File 1: Fig. S1), 25,898 genes had both 5' UTRs and 3' UTRs, 85 genes had only 5' UTRs, 240 genes only had 3' UTRs, and 1,271 genes had no UTR.

Manual curation, genome polish, and protein annotation

Subsequently, we performed manual curation in IGV-sRNA [52], conducting two rounds of evaluations across the 180 non-rDNA chromosomes (Fig. 1D, 5A). Firstly, we checked the 3,937 genes with multiple eTSSs. Among the 3,935 genes with two eTSSs, 3,908 were capable of transcribing antisense transcripts, with one eTSS belonging to a protein-coding gene and the other eTSS corresponding to an antisense transcript (Fig. 5A, 7D-F). The remaining 27 genes contained two eTSSs, with one of them serving as an alternative TSS for the protein-coding gene (Fig. 5A, B). Additionally, two genes contained three eTSSs, signifying three alternative TSSs. Secondly, we checked 1,752 duplicated genes with neither eTSS nor pTSS (Fig. 4A). These genes could be categorized into two groups. One group consisted of 1,174 repetitive genes with multiple short exons (mostly <100 bp) distributed across distinct chromosomes (Fig. 5A, 5E). They tended to be misaligned due to the default Smith-Waterman algorithm for Nanopore DRS data analysis [53]. Most of these multi-short-exon

genes belonged to the leucine-rich repeat (LRR) superfamily, which has recently evolved and lacks the transcription activation marks including 6mA [54]. The other group comprised 578 tandem duplicate genes with multiple copies arranged in a linear fashion at a single genomic locus (Fig. 5A, C). Thirdly, there were 15 genes exhibiting super high splicing diversity, with nearly every noncoding exon being subject to alternatively splicing (near-universal AS) (Fig. 5D). This phenomenon was also observed in humans, wherein 69% of human protein-coding exons were classified as alternative, and some functional long-noncoding RNAs (lncRNAs) such as XIST, HOTAIR, GOMAFU, and H19 were observed to be near-universally alternatively spliced at each locus [55]. We annotated these 15 genes with their most dominant isoforms.

While conducting manual curation, we observed sequence errors in certain regions. Therefore, we polished the genome sequence using Illumina sequencing data (Figure 1D, 5A), correcting a total of 3,759 insertions, 135 deletions, 43 transitions, and 48 transversions. The corrections were validated by Sanger sequencing at representative sites (Additional File 1: Fig. S6A, Additional File 7: Table S6). Among these corrected sites, 1,696 were located in genic regions, with 645 in exons and 1,051 in introns (Figure 5A). Errors in gene exons could lead to inaccuracies in the predicted CDS (Figure 5F). Using the polished genome, we re-predicted CDS for 645 genes with errors in their exons, resulting in 438 genes acquiring more accurate and extended CDS.

To update the functional annotation, predicted proteins were blasted against multiple public protein databases (Figure 1D, Additional File 1: Fig. S7A). In total, we annotated 25,846 functional genes, featuring an additional 1,732 functional genes compared to TGD2021 (Additional File 1: Fig. S7A, B). In the case of these newly annotated genes, protein functional annotation revealed their distribution across distinct structural domain families, with a higher prevalence observed in certain families, such as the leucine-rich repeat domain, the cyclic nucleotide-binding domain, and the WD40/YVTN repeat-like-containing domain (Additional File 1: Fig. S7C). Three newly annotated proteins were associated with epigenetic regulation (Additional File 1: Fig. S7D). Two featured a histone H3 K4-specific methyltransferase SET domain homologous to MLL5 (KMT2E) that are critical for gene transcription regulation, cell cycle regulation (G1/S transition), and myoblast differentiation [56-59]. Another exhibited homology to the 16S rRNA m5C methyltransferase NSUN4, characterized by the presence of a RsmB domain [60].

In this version (draft v5) (Additional File 1: Fig. S1), we optimized TSS annotation for 3,937 genes with multiple eTSSs, manually re-annotated 1,752

duplicated genes and 15 universally alternatively spliced genes, and re-predicted CDS for 438 genes.

Annotate transcript isoforms generated by alternative splicing

It has been reported that 1,286 *Tetrahymena* genes generate alternative splicing **(**AS) isoforms [26], but this information was not integrated into previous gene models and TGD2021 contained only 459 AS genes. With the gene model being highly optimized in this study, we identified all six types of alternative splicing, namely exon skipping, alternative last exon, intron retention, mutually exclusive exons, alternative 5' splice site, and alternative 3' splice site, in a total of 3,041 genes, generating 5,917 isoforms (Fig. 1E, 6B). Consistent with the previous report [26], intron retention was the dominant form of AS (Fig. 6B). The numbers of AS genes and isoforms in our annotation were much higher than those in TGD2021 (gene: 5,041 vs. 459, isoform: 5,917 vs. 516) (Fig. 6B). Of these, 876 genes exhibited no less than two AS isoforms. Each AS event was supported by DRS full-length reads spanning the intron-exon junctions (Fig. 6A). Gene loci representing each of the six different AS types were selected for RT-PCR analysis, validating the existence and structure of AS isoforms (Additional File 1: Fig. S8A-F, Additional File 8: Table S7).

To further investigate whether the generation of AS isoforms was stage-specific, we compared AS isoforms in growth, starvation, and different timepoints of conjugation. The results showed that 2,131 out of 5,917 AS isoforms were generated across all periods, while others exhibited a tendency to be highly expressed during specific stages. Specifically, 114 AS isoforms were generated exclusively during growth, 326 during starvation, and 1,146 during conjugation (Fig. 6C). For example, in the case of the starvation- and conjugation-specific gene TTHERM_001026363, its AS isoforms showed a stage-specific pattern, with a gradual increase of the ratio of retained introns and intron-containing reads observed as conjugation progressed (Fig. 6D). In support, GO analysis of the overall functions of AS isoforms revealed a predominant enrichment in processes related to cell cycle and meiosis (Additional File 1: Fig. S8G, Additional File 9: Table S8).

Identify natural antisense transcripts (NATs)

We observed that many gene loci could be transcribed from both sense and antisense strands (Fig. 7D-F). Intriguingly, transcripts originating from the

antisense strand, which were typically shorter in length, were located within or in close proximity to the sense-coding transcripts, characteristic of natural antisense transcripts (NATs) [61]. In total, 4,389 NATs were identified (Fig. 7A). The presence of DRS and RNA-seq reads provided strong support for these NATs, confirming that they were *bona fide* transcripts rather than transcriptional noise. 14% of protein-coding genes (3,908/27,494) showed evidence of antisense transcription. Most NATs lacked a discernable ORF (>100 aa), but 11 NATs were annotated as potential functional protein and 112 displayed high protein-coding potentials (Additional File 1: Fig. S9A).

The resulting set of NATs was categorized into three groups according to their positional relationship with the corresponding sense transcript (protein-coding genes). 1) Intergenic NATs. These NATs did not overlap with other sense transcripts and were further subdivided into: a) 107 promoter NATs originated from shared bidirectional promoters (Fig. 7B), and b) 583 intergenic NATs possessing their own independent promoters (Fig. 7C). 2) Exonic NATs. These NATs were transcribed from loci with sense transcripts and were categorized into: a) 3,460 type 1 exonic NATs located within 1 kb downstream of the TSS of the respective sense transcription unit and shared the epigenetic marks with the latter (Fig. 7D), and b) 417 type 2 exonic NATs located more than 1 kb downstream of the TSS of the sense transcript and possessed the epigenetic marks downstream of their own TSS independently (Fig. 7E). It is worth mentioning that these NATs are not the reverse transcription or replication byproducts of sense transcripts. Instead, their exon-intron boundaries slightly deviate from those of their sense transcripts and they themselves contain canonical GU-AG sites for intron splicing (Additional File 1: Fig. S9B, C). 3) Intronic NATs. There were 31 NATs transcribed within the intronic regions of sense transcripts (Fig. 7F). The distinct genomic locations of NATs and their positional proximity to the sense transcripts may determine their roles in various aspects of gene expression.

Compared to sense transcripts, NATs were generally characterized by shorter lengths (Additional File 1: Fig. S9D) and lower expression levels (Additional File 1: Fig. S9E), while no difference in GC content was observed (Additional File 1: Fig. S9D). Interestingly, the majority of antisense transcripts also carried epigenetic marks (H3K4me3, H2A.Z, 6mA and well-positioned nucleosome) in their transcription units (Figure 7G, Additional File 1: Fig. S9F), similar to the report in *Arabidopsis thaliana* wherein NATs were enriched with H3K4me3 [62]. Most of these epigenetic marks were shared with their

corresponding sense transcripts (Fig. 7D), but they were also possibly involved in regulating NATs expression.

Most importantly, these NATs exhibited temporal-specific expression patterns that were opposite to their corresponding sense coding genes, mirroring the findings in *Arabidopsis thaliana* where sense and antisense transcripts exhibited mutual exclusivity at individual loci [63]. In the instance of the gene TTHERM_00412050 in *Tetrahymena*, the expression of its NATs gradually decreased while the expression of its sense transcripts increased, during the transition from growth to starvation and subsequently to conjugation (Fig. 7H). This phenomenon might induce gene silencing of the corresponding sense genes by degrading the sense mRNA or interfering with its translation, a role that has been reported for NATs in plants [62, 64].

We then assessed the alternative splicing diversity (ASD), defined as the proportion of different types of introns for each NAT loci and its sense gene in DRS data. Intriguingly, the ASD of NATs significantly exceeded that of their sense counterparts (0.96 vs. 0.28, P <0.001) (Fig. 7I) or total sense coding transcripts (0.96 vs. 0.15, P <0.001) (Additional File 1: Fig. S9G). We speculate that, akin to non-coding RNA, NATs appear to be exempt from the evolutionary constraints imposed on protein-coding genes by the preservation of a functional ORF. This exemption might allow their exons to function as modular sections and act as independent units that can be shuffled and rearranged with great flexibility [55].

CONCLUSION

In this study, we established a novel workflow to optimize the genome annotation of *Tetrahymena* that integrated large scale transcriptomic data, including RNA-seq data from different cell stages, strand-specific RNA-seq data, and Nanopore DRS data. Most importantly, epigenetic data including H3K4me3, H2A.Z, 6mA, and nucleosome along with ATAC-seq data were incorporated. This comprehensive dataset enabled the optimization of gene models, the accurate identification of TSS and TES, the augmentation of UTR information, the updated annotation of protein functions, and the addition of alternatively spliced isoforms.

Our updated gene model (TGD2024) (Additional File 1: Fig. S1) comprised a total of 27,494 genes, including 26,223 well-annotated and 1,271 low-confidence genes. Compared to TGD2021, we annotated 2,481 new genes, and optimized 6,428 genes including 3,878 exon-altered genes, 169 orientation-reversed genes,

65 TSS-altered genes, 1,379 fused genes, and 1,136 partitioned genes. We also increased the number of alternatively spliced isoforms to a total of 5,917 and annotated an additional 1,732 functional genes. Furthermore, we identified a large pool of NATs that might generate a diverse and extensive repertoire of potential regulatory RNAs.

Our work will largely facilitate *Tetrahymena* biology studies and the conceptual framework employed here holds substantial promise for facilitating genome annotation in other eukaryotes. Moving forward, we aim to enrich our analysis by incorporating additional epigenic data such as H3K27me1 that is predominantly enriched at the 3' end of gene bodies (unpublished data) and Nanopore DRS data from conjugating cells. We also plan to generate data using QTI-seq and Cap-seq that will help to better determine the transcription start sites and translation start sites [65, 66].

MATERIALS AND METHODS

Cell growth, RNA extraction, and library construction

Tetrahymena wild-type strains (SB210 and CU428) were obtained from the *Tetrahymena* Stock Center (http://tetrahymena.vet.cornell.edu). Cells were grown in SPP medium at 30°C. For conjugation, starved SB210 and CU428 cells were resuspended in 10 mM Tris (pH 7.5) at 2×10^5 cells/ml, mixed in equal volumes, and samples were collected at 4h, 5h, 6h, 8h, 10h after mixing. Total RNA was collected using the RNeasy Plus Mini kit (Qiagen, 74134). The quality and concentration of RNA samples were analyzed by 1% agarose gel electrophoresis and Qubit®3.0 Fluorometer (Thermo Fisher Scientific). Strand-specific RNA sequencing libraries and Illumina sequencing libraries of SB210 were constructed according to manufacturer-recommended protocols and sequenced by Novogene Co. Ltd (Beijing, China). For ATAC-seq libraries, transposase Tn5 from the Nextera DNA Library Preparation Kit (#FC-121-1030, Illumina, USA) was employed to treat 10^5 macronuclei for 1h at 37°C. The DNA from the samples was subsequently recovered using the MinElute Recovery Kit (#28004, Qiagen, Germany). Amplification and library construction of the sample DNA were performed for 13 PCR cycles, with library adapter primers sourced from the Nextera XT Index Kit (#FC-121-1011, Illumina, USA). The DNA from the constructed library was recovered once more using the MinElute Recovery Kit (#28004, Qiagen, Germany).

Gene loci identification based on RNA-seq and strand-specific RNA-seq data

The latest published MAC genome of *Tetrahymena thermophila* downloaded from TGD (http://ciliate.org) was used as the reference for reads mapping.

Adapters and low-quality reads were removed using Trim Galore (http://www.bioinformat-ics.babraham.ac.uk/projects/trim_galore/). Paired-end reads generated by RNA-seq and strand-specific RNA-seq were mapped back to the genome for transcript assembly using Hisat2 [67] with default parameters (--rna-strandness R/RF for strand-specific RNA-seq). Picard Tools (https://broadinstitute.github.io/picard/) were used to remove duplicate reads from PCR. RNA-seq and strand-specific RNA-seq data were divided into three groups (growth, conjugation, starvation) according to the different cell cycle stages of *Tetrahymena*. The RNA-seq reads in each group were assembled and merged into the stage-combined transcripts by Stringtie with default parameters [68]. For each gene locus, only transcripts with the highest expression level (FPKM>1) across three groups were used for gene model optimization.

Gene prediction and UTRs annotation using LoReAn2

The latest published MAC genome of *Tetrahymena* downloaded from TGD (http://ciliate.org) was used as the reference genome for reads mapping [24]. Gene prediction was performed using the LoReAn2 annotation pipeline [31]. In detail, the transcriptomic data were aligned to the genome using the Program to Assemble Spliced Alignments (PASA) [69] and the Genomic Mapping and Alignment Program (GMAP) [70]. For protein alignment, the analysis and annotation tool (AAT) [71] was used to align protein sequence to the genome. Reference genome-guided transcripts were assembled using Trinity [72]. SNAP [73], Augustus [74], and GeneMark-ET [75] were used to generate *de novo* gene annotation individually, which were then combined using EVM [69]. PASA was used to annotate UTRs.

Reverse transcription polymerase chain reaction (RT-PCR)

Total RNA after Dnase treatment (Invitrogen, AM1907) was reverse-transcribed using an oligo-dT primer and M-MLV Reverse Transcriptase (Invitrogen, 28025013) and cDNA was used as a template. RT-PCR was performed using Premix Taq (TaKaRa, RR901A). All PCR primers are listed in Supplementary Data.

ChIP-seq, MNase-seq, SMRT-seq data processing

The preprocessing steps for the epigenetic data (H3K4me3 ChIP-seq, H2A.Z ChIP-seq, MNase-seq) were consistent with those employed for RNA-seq. Only the mono-nucleosome sized fragments (120-260 bp) were analyzed. The unique mapping results (bam files) and the 6mApT sites (SMRT-seq 6mA data) were calculated by custom Perl scripts and plotted by deepTools [76] (bin = 10 bp).

ATAC-seq and data processing

ATAC-seq was performed as previously described [77]. Libraries were sequenced (PE150) on an Illumina HiSeq sequencer. Mapped reads without PCR duplicates were used to retrieve short open chromatin regions (shorter than 100 bp) using deepTools [76], as defined by the Greenleaf research team [78, 79]. Peak calling was performed using MACS2 (v2.1.0) [80] The open chromatin profile distribution around TSS was plotted by deepTools [76]. The human precise TSSs from the RefTSS database [81] were mapped to corresponding genes using custom Perl scripts.

Gene model optimization by a machine learning approach based on epigenetic information

To optimize the gene model predicted by transcriptomic data, a Random Forest (RF) model was developed to further identify TSS based on epigenetic information (H3K4me3, H2A.Z, 6mA, and well-positioned nucleosome). RF classification algorithm was implemented with the randomForest R package [82]. The model training was performed using a dataset containing abundant information of epigenetic marks in regions of 1,000 bp downstream of TSS (positive training set) and in regions of 1,000 bp centered by TES (negative training set). 10,460 well-annotated and longer than 1kb genes were selected, 70% of which were used for model training and the rest were for testing.

The error rate of the RF model was computed based on the out-of-bag (OOB) error, which is the mean prediction error over all Random Forest trees. The importance of each feature was computed as “mean decrease in accuracy” (MDA). Feature importance (MDA) and classification performance (OOB error) measures were further averaged over a collection of five hundred Random Forests to obtain stable results. The genome-wide regions were clustered and divided into different categories regions based on comprehensive consideration

of chromatin states (H3K4me3, H2A.Z, 6mA, and nucleosome) using the DBSCAN cluster algorithm in fpc R package (eps=150, MinPts=3) (https://www.unibo.it/sitoweb/christian.hennig/en/). Subsequently, epigenetic mark signals on regions from clustered categories (scaled to 1kb) was feature engineered to predict TSS-regions using the pre-trained RF model.

CDS prediction, UTR annotation and protein function annotation

Longest ORFs on each strand were predicted from the stage-combined transcripts using ORFfinder [83]. A putative ORF was defined as amino acid sequences exceeding 100 aa in length. The orientation of strand-specific RNA-seq transcripts was used to determine CDSs from longest predicted ORFs. The regions beyond CDSs on transcripts were defined as 5' UTRs and 3' UTRs, respectively. Putative protein coding regions were annotated using EggNOG [84], Interproscan [85], and Pannzer2 [86] by mapping to known proteins, protein domains and signal peptides collected in UniProtKB database [87], Pfam databases [88], and InterPro database [89].

Motif enrichment analysis

For the motif enrichment analysis of open chromatin region upstream of the TSS, only fragments shorter than the mono-nucleosome size (100 bp) were analyzed after mapping to the genome. Peaks were identified using MACS2 (v2.1.0) [80]. Motif enrichment analysis was performed in called peaks using HOMER (v4.8) [90].

For the motif analysis around the TES, sequences spanning 50 bp upstream and downstream of RNA cleavage sites/TES were extracted using bedtools [91]. Subsequently, the extracted fasta sequences were renamed utilizing SeqKit [92]. Motif analysis was conducted on these sequences using MEME-Suite's simple enrichment analysis (SEA) [93].

Nanopore direct RNA sequencing data generation and analysis

Oxford PromethION 2D amplicon libraries for full-length transcriptome sequencing were generated according to the Nanopore community protocol using library preparation kit SQK-LSK109 and were sequenced on R9 flowcells to generate fast5 files. Fastq files were derived from fast5 reads by basecalling using guppy v3.2.10 (default parameters,

https://github.com/metagenomics/denbi-nanopore-training/blob/master/docs/basecalling/basecalling.rst). Reads were filtered using Nanofilt v2.5.0 [94]. Nanopore direct RNA sequencing (DRS) reads were aligned to the genome using minimap2 v2.16 [95]. The alternatively spliced isoforms were identified by customized Perl scripts followed by manual curation. The RNA-seq data (growth, starvation 24h, and conjugation at 4h, 5h, 6h, 8h, and 10h), each with two or three biological replicates, were used to calculate the expression level during different cell stages. The heatmap plot and gene ontology (GO) analysis were plotted by Tbtools [96]. For the poly-A tail analysis, nanopolish-polya version 0.10.2 (https://github.com/jts/nanopolish) was used to estimate polyadenylated tail lengths from Nanopore DRS raw reads.

Identification and classification of natural antisense transcripts

Natural antisense transcripts (NATs) were identified by fulfilling the following criteria: 1) transcribed from the antisense strand of protein-coding genes as evidenced by DRS data, and 2) localized upstream or within protein-coding genes, encompassing intronic or exonic regions. Classification of each transcript as either coding or noncoding was determined using a stepwise filtering pipeline. First, all candidates were scored with LGC [97] to determine their coding potential. All transcripts that were named “non-coding” were retained as potential noncoding candidates. Second, all candidate transcripts were subjected to blastp [98] and HMMER (versus Pfam-A and Pfam-B) [99]. For blastp and HMMER, transcripts were translated in all three sense frames. Transcripts with an E-value less than 1e-4 in any of the three search algorithms were considered as functional-coding; transcripts that were predicted to contain ORF exceeding 200 bp in length, yet lacked identifiable homologous proteins or functional domain, were defined as potential-coding; and the remaining were classified as non-coding. The alternative splicing diversity (ASD) was quantified as the ratio between the number of distinct splice sites and the total reads number captured by Nanopore DRS data for a particular gene. The comparison of ASD for sense protein-coding genes and NATs could be approached in two ways: 1) comparing all NATs (4,398) to protein-coding genes with NATs (4,398); and 2) comparing all NATs (4,398) to all protein-coding genes (27,494).

Supplementary information

Additional file 1: Figure S1. The flowchart for genome annotation. **Figure S2.** RT-PCR validation of gene models optimized by the transcriptomic data. **Figure**

637 **S3.** The eTSSs/pTSSs prediction and gene optimization using epigenetic data.
638 **Figure S4.** RT-PCR validation of gene models optimized by eTSSs. **Figure S5.**
639 The regulatory elements of untranslated regions analysis in *Tetrahymena*. **Figure**
640 **S6.** Four types of error sites polished by Illumina and Sanger sequencing data.
641 **Figure S7.** Protein function annotation on draft v4. **Figure S8.** The experimental
642 validation and functional analysis for alternative splicing (AS) isoforms. **Figure**
643 **S9.** Identification and characterization of natural antisense transcripts (NATs).

644 **Additional file 2: Table S1**. The Information of the sequencing data.

645 **Additional file 3: Table S2**. The comparison between TGD2014, TGD2021, and
646 the LoReAn2 annotated gene model.

647 **Additional file 4: Table S3**. PCR primers for gene optimization.

648 **Additional file 5: Table S4**. Homology protein of the promoter binding protein in
649 *Tetrahymena*.

650 **Additional file 6: Table S5**. Results of GO enrichment analysis for gene with
651 poly-A.

652 **Additional file 7: Table S6**. PCR primers for genome polish.

653 **Additional file 8: Table S7**. PCR primers for alternative splicing.

654 **Additional file 9: Table S8**. Results of GO enrichment analysis for alternative
655 splicing.

656

657 **Data availability**

658 Scripts to generate data and to perform the above analysis are available in github:
659 https://github.com/yefei521/UTR-annotation. Public RNA-seq datasets were
660 available on the Gene Expression Omnibus (GEO) under accession number
661 GSE27971: https://www.ncbi.nlm.nih.gov/gds/?term=GSE27971 [26]. Our SMRT-
662 seq 6mA data was available at the NCBI database (BioProject accession number:
663 PRJNA932808) [30], MNase-seq nucleosome data was available at
664 https://www.ncbi.nlm.nih.gov/sra/SRX5146438[accn] [28]. ChIP-seq data
665 (H3K4me3 and H2A.Z), RNA-seq, strand-specific RNA-seq, Nanopore direct
666 RNA sequencing, and ATAC-seq data from the current work were deposited at
667 the NCBI database (BioProject accession number: PRJNA1048844).
668

669 **Ethics approval and consent to participate**

670 Not applicable.

671

Consent for publication

Not applicable.

Competing interests

All authors declare no potential conflict of interest.

Funding

This work was supported by the Science & Technology Innovation Project of Laoshan Laboratory (LSKJ202203203), the National Natural Science Foundation of China, (32125006 and 32070437 to SG, 32030015 and 32270512 to XC), Program of Qilu Young Scholars of Shandong University, Youth Innovation Team of Shandong Provincial Higher Education Institutions, the Natural Science Foundation of Jiangsu Province (BK20220268), and a supporting project (Project number RSP2024R10) at King Saud University, Saudi Arabia. TGD is supported by US National Institutes of Health Grant P40OD010964. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contribution

Fei Ye conceived and led the project, performed strand-specific RNA-seq and Nanopore direct RNA sequencing, conducted computational and experimental analysis, and wrote and revised the manuscript. Xiao Chen provided analytical framework for TSS prediction and ATAC-seq processing, conducted ATAC-seq, revised the manuscript, and provided funding resources. Yalan Sheng assisted in resolving analytical issues and revised the manuscript. Lili Duan and Aili Ju performed RNA-seq of conjugation samples. Naomi Stover assisted with gene nomenclature and formatted data for presentation on TGD. Khaled A. S. Al-Rasheid revised the manuscript. Shan Gao conceived and supervised the project, wrote and revised the manuscript, and provided funding resources. All authors read and approved the final manuscript.

Acknowledgements

We thank members of the Gao laboratory for discussion. Our special thanks are given to Dr. Weibo Song (Ocean University of China, OUC) for his kind

706 suggestions during preparation of the manuscript and Dr. Jun Wang (Institute of
707 Microbiology, Chinese Academy of Sciences) for his advice on Nanopore direct
708 RNA sequencing and analysis. We appreciate the computing resources provided
709 by the Institute of Evolution & Marine Biodiversity in OUC, the Center for High
710 Performance Computing and System Simulation at Laoshan Laboratory, and
711 Marine Big Data Center of Institute for Advanced Ocean Study of OUC.

712

References

1. Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *Tetrahymena***.** Cell*.* 2002, 110:689-699.
2. Mochizuki K, Gorovsky MA. Conjugation-specific small RNAs in *Tetrahymena* have predicted properties of scan (scn) RNAs involved in genome rearrangement**.** Genes & Development*.* 2004, 18:2068-2073.
3. Collins K, Gorovsky MA. *Tetrahymena thermophila***.** Current Biology*.* 2005, 15:R317-R318.
4. Gao S, Xiong J, Zhang C, Berquist BR, Yang R, Zhao M, Molascon AJ, Kwiatkowsk SY, Yuan D, Qin Z. Impaired replication elongation in *Tetrahymena* mutants deficient in histone H3 Lys 27 monomethylation**.** Genes & Development*.* 2013, 27:1662-1679.
5. Xiong J, Gao S, Dui W, Yang W, Chen X, Taverna SD, Pearlman RE, Ashlock W, Miao W, Liu Y. Dissecting relative contributions of cis- and trans-determinants to nucleosome distribution by comparing *Tetrahymena* macronuclear and micronuclear chromatin**.** Nucleic Acids Research*.* 2016, 44:10091-10105.
6. Feng L, Wang G, Eileen P H, Xiong J, Yan G, Chen K, Chen X, Dui W, Plemens A, Khadr L. A germline-limited piggyBac transposase gene is required for precise excision in *Tetrahymena* genome rearrangement**.** Nucleic Acids Research*.* 2017, 45:9481-9502.
7. Orias E, Singh DP, Meyer E. Genetics and epigenetics of mating type determination in *Paramecium* and *Tetrahymena***.** Annual Review of Microbiology*.* 2017, 71:133-156.
8. Zhao X, Wang Y, Wang Y, Liu Y, Gao S. Histone methyltransferase TXR1 is required for both H3 and H3.3 lysine 27 methylation in the well-known ciliated protist *Tetrahymena thermophila***.** Science China Life Sciences*.* 2017, 60:264-270.
9. Xu J, Li X, Song W, Wang W, Gao S. Cyclin Cyc2p is required for micronuclear bouquet formation in *Tetrahymena thermophila***.** Science China Life Sciences*.* 2019, 62:668-680.
10. Gao Y, Solberg T, Wang C, Gao F. Small RNA-mediated genome rearrangement pathways in ciliates**.** Trends in Genetics*.* 2023, 39:94-97.
11. Wei F, Pan B, Diao J, Wang Y, Sheng Y, Gao S. The micronuclear histone H3 clipping in the unicellular eukaryote *Tetrahymena thermophila***.** Marine Life Science & Technology*.* 2022, 4:584-594.
12. Yan Y, Maurer-Alcalá XX, Knight R, Pond SLK, Katz LA. Single-cell transcriptomics reveal a correlation between genome architecture and gene family evolution in ciliates**.** mBio*.* 2019, 10:10-1128.
13. Cheng T, Wang Y, Huang W, Chen X, Zhao X, Gao S, Song W. Our recent progress in epigenetic research using the model ciliate, *Tetrahymena thermophila***.** Marine Life Science & Technology*.* 2019, 1:4-14.
14. Lyu L, Asghar U, Fu J, Gao Y, Zhang X, Al-Farraj SA, Chen Z, Gao F. Comparative analysis of single-cell genome sequencing techniques toward the characterization of germline and somatic genomes in ciliated protists**.** European Journal of Protistology*.* 2023, 88:125969.
15. Jin D, Li C, Chen X, Byerly A, Stover NA, Zhang T, Shao C, Wang Y. Comparative genome analysis of three euplotid protists provides insights into the evolution of nanochromosomes in unicellular eukaryotic organisms**.** Marine Life Science & Technology*.* 2023, 5:300-315.
16. Fu J, Chi Y, Lu X, Gao F, Al-Farraj SA, Petroni G, Jiang J. Doublets of the unicellular organism *Euplotes vannus* (Alveolata, Ciliophora, Euplotida): the morphogenetic patterns of the ciliary and nuclear apparatuses associated with cell division**.** Marine Life Science & Technology*.* 2022, 4:527-535.
17. Gorovsky MA. Macro- and micronuclei of *Tetrahymena pyriformis*: a model system for studying the structure and function of eukaryotic nuclei**.** Journal of Eukaryotic Microbiology*.* 2010, 20:19-25.
18. Ray C. Meiosis and nuclear behavior in *Tetrahymena pyriformis***.** The Journal of Protozoology*.* 2010, 3:88-96.

19. Karrer KM. *Tetrahymena* genetics: two nuclei are better than one**.** Methods in Cell Biology*.* 1999, 62:127-186.

20. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote**.** PLOS BIOLOGY*.* 2006, 4:1544-9173.

21. Hamilton EP, Dear PH, Rowland T, Saks K, Eisen JA, Orias E. Use of HAPPY mapping for the higher order assembly of the *Tetrahymena* genome**.** Genomics*.* 2006, 88:443-451.

22. Coyne RS, Stover NA, Miao W, Chapter 4 - Whole Genome Studies of *Tetrahymena***.** In Methods in Cell Biology. Volume 109. Edited by Collins K: Academic Press; 2012: 53-81

23. Coyne RS, Thiagarajan M, Jones KM, Wortman JR, Tallon LJ, Haas BJ, Cassidy-Hanley DM, Wiley EA, Smith JJ, Collins K, et al. Refined annotation and assembly of the *Tetrahymena thermophila* genome sequence through EST analysis, comparative genomic hybridization, and targeted gap closure**.** BMC Genomics*.* 2008, 9:562.

24. Sheng Y, Duan L, Cheng T, Qiao Y, Gao S. The completed macronuclear genome of a model ciliate *Tetrahymena thermophila* and its application in genome scrambling and copy number analyses**.** Science China Life Sciences*.* 2020, 63:1534-1542.

25. Miao W, Xiong J, Bowen J, Wang W, Liu Y, Braguinets O, Grigull J, Pearlman RE, Orias E, Gorovsky MA. Microarray analyses of gene expression during the *Tetrahymena thermophila* life cycle**.** PLoS One*.* 2009, 4:e4429.

26. Xiong J, Lu X, Zhou Z, Chang Y, Miao W. Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using deep RNA sequencing**.** PLoS One*.* 2012, 7:e30630.

27. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling**.** American Journal of Obstetrics and Gynecology*.* 2006, 195:373-388.

28. Wang Y, Sheng Y, Liu Y, Zhang W, Cheng T, Duan L, Pan B, Qiao Y, Liu Y, Gao S. A distinct class of eukaryotic MT-A70 methyltransferases maintain symmetric DNA N^6-adenine methylation at the ApT dinucleotides as an epigenetic mark associated with transcription**.** Nucleic Acids Research*.* 2019, 47:11771-11789.

29. Wang Y, Chen X, Sheng Y, Liu Y, Gao S. N^6-adenine DNA methylation is associated with the linker DNA of H2A.Z-containing well-positioned nucleosomes in Pol II-transcribed genes in *Tetrahymena***.** Nucleic Acids Research*.* 2017, 45:11594–11606.

30. Sheng Y, Wang Y, Yang W, Wang X, Lu J, Pan B, Nan B, Liu Y, Li C, Song J, et al. Semi-conservative transmission of DNA N^6-adenine methylation in a unicellular eukaryote**.** bioRxiv*.* 2023:2023.2002.2015.468708.

31. Cook DE, Valleinclan JE, Pajoro A, Rovenich H, Thomma B, Faino L. Long-read annotation: automated eukaryotic genome annotation based on long-read cDNA sequencing**.** Plant Physiology*.* 2019, 179:38-54.

32. Zhao L, Gao F, Gao S, Liang Y, Long H, Lv Z, Su Y, Ye N, Zhang L, Zhao C, et al. Biodiversity-based development and evolution: the emerging research systems in model and non-model organisms**.** Science China Life Sciences*.* 2021, 64:1236-1280.

33. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis**.** Genome Biology*.* 2020, 21:22.

34. Zheng D, Tian B. RNA-binding proteins in regulation of alternative cleavage and polyadenylation**.** Advances in Experimental Medicine and Biology*.* 2014, 825:97-127.

35. Workman RE, Tang AD, Tang PS, Jain M, Tyson JR. Nanopore native RNA sequencing of a human poly(A) transcriptome**.** Nature Methods*.* 2019, 16:1297-1305.

36. Mulroney L, Wulf MG, Schildkraut I, Tzertzinis G, Buswell J, Jain M, Olsen H, Diekhans M, Corrêa IR, Akeson M. Identification of high confidence human poly(A) RNA isoform scaffolds using nanopore sequencing**.** RNA*.* 2020, 28:162-176.

37. Ruehle MD, Orias E, Pearson CG. *Tetrahymena* as a unicellular model eukaryote: genetic and genomic tools**.** Genetics*.* 2016, 203:649-665.

38. Mantovani R. The molecular biology of the CCAAT-binding factor NF-Y**.** Gene*.* 1999, 239:15-27.

39. Lifton RP, Goldberg ML, Karp RW, Hogness DS. The organization of the histone genes in *Drosophila melanogaster*: functional and evolutionary implications**.** Cold Spring Harbor Symp Quant Biol*.* 1978, 42:1047-1051.

40. Impey S, Smith DM, Obrietan K, Donahue R, Wade C, Storm DR. Stimulation of cAMP response element (CRE)-mediated transcription during contextual learning**.** Nature Neuroscience*.* 1998, 1:595-601.

41. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, Segal E. The DNA-encoded nucleosome organization of a eukaryotic genome**.** Nature*.* 2009, 458:362-366.

42. Ju Q, Morrow BE, Warner JR. REB1, a yeast DNA-binding protein with many targets, is essential for growth and bears some resemblance to the oncogene myb**.** Molecular and Cellular Biology*.* 1990, 10:5226-5234.

43. Tian B, Manley JL. Alternative cleavage and polyadenylation: the long and short of it**.** Trends in Biochemical Sciences*.* 2013, 38:312-320.

44. Ozsolak F, Kapranov P, Foissac S, Kim S, Fishilevich E, Monaghan A, John B, Milos P. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation**.** Cell*.* 2010, 143:1018-1029.

45. Shen Y, Liu Y, Liu L, Liang C, Li QQ. Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in *Chlamydomonas reinhardtii***.** Genetics*.* 2008, 179:167-176.

46. Rao P, Yuan W, Krug RM. Crucial role of CA cleavage sites in the cap-snatching mechanism for initiating viral mRNA synthesis**.** The EMBO Journal*.* 2003, 22:1188-1198.

47. Gruber AJ, Zavolan M. Alternative cleavage and polyadenylation in health and disease**.** Nature Reviews Genetics*.* 2019, 20:599-614.

48. Jia J, Lu W, Liu B, Fang H, Yu Y, Mo W, Zhang H, Jin X, Shu Y, Long Y, et al. An atlas of plant full-length RNA reveals tissue-specific and monocots–dicots conserved regulation of poly(A) tail length**.** Nature Plants*.* 2022, 8:1118-1126.

49. Lima SA, Chipman LB, Nicholson AL, Chen Y-H, Yee BA, Yeo GW, Coller J, Pasquinelli AE. Short poly(A) tails are a conserved feature of highly expressed genes**.** Nature Structural & Molecular Biology*.* 2017, 24:1057-1063.

50. Meyer S, Temme C, Wahle E. Messenger RNA turnover in eukaryotes: pathways and enzymes**.** Critical Reviews in Biochemistry and Molecular Biology*.* 2004, 39:197-216.

51. Parker R, Song H. The enzymes and control of eukaryotic mRNA turnover**.** Nature Structural & Molecular Biology*.* 2004, 11:121-127.

52. Chen C, Li J, Feng J, Liu B, Feng L, Yu X, Li G, Zhai J, Meyers BC, Xia R. sRNAanno—a database repository of uniformly annotated small RNAs in plants**.** Horticulture Research*.* 2021, 8:45.

53. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases**.** Nucleic Acids Research*.* 2019, 47:10994-11006.

54. Xiong J, Yang W, Chen K, Jiang C, Ma Y, Chai X, Yan G, Wang G, Yuan D, Liu Y, et al. Hidden genomic evolution in a morphospecies—The landscape of rapidly evolving genes in *Tetrahymena***.** PLOS BIOLOGY*.* 2019, 17:e3000294.

55. Deveson IW, Brunck ME, Blackburn J, Tseng E, Hon T, Clark TA, Clark MB, Crawford J, Dinger ME, Nielsen LK, et al. Universal alternative splicing of noncoding exons**.** Cell Systems*.* 2018, 6:245-255.e245.

56. Zhou P, Wang Z, Yuan X, Zhou C, Liu L, Wan X, Zhang F, Ding X, Wang C, Xiong S, et al. Mixed lineage leukemia 5 (MLL5) protein regulates cell cycle progression and E2F1-

responsive gene expression via association with host cell factor-1 (HCF-1)*. Journal of Biological Chemistry. 2013, 288:17532-17543.

57. Sebastian S, Sreenivas P, Sambasivan R, Cheedipudi S, Kandalla P, Pavlath GK, Dhawan J. MLL5, a trithorax homolog, indirectly regulates H3K4 methylation, represses cyclin A2 expression, and promotes myogenic differentiation. Proceedings of the National Academy of Sciences. 2009, 106:4719-4724.
58. Deng L-W, Chiu I, Strominger JL. MLL5 protein forms intranuclear foci, and overexpression inhibits cell cycle progression. Proceedings of the National Academy of Sciences. 2004, 101:757-762.
59. Cheng F, Liu J, Zhou SH, Wang XN, Chew JF, Deng L-W. RNA interference against mixed lineage leukemia 5 resulted in cell cycle arrest. The International Journal of Biochemistry & Cell Biology. 2008, 40:2472-2481.
60. Zhang H, Zhai X, Liu Y, Xia Z, Xia T, Du G, Zhou H, Franziska Strohmer D, Bazhin AV, Li Z, et al. NOP2-mediated m5C modification of c-Myc in an EIF3A-dependent manner to reprogram glucose metabolism and promote hepatocellular carcinoma progression. Research. 6:0184.
61. Pelechano V, Steinmetz LM. Gene regulation by antisense transcription. Nature Reviews Genetics. 2013, 14:880-893.
62. Kornienko AE, Nizhynska V, Morales AM, Pisupati R, Nordborg M. Population-level annotation of lncRNAs in *Arabidopsis* reveals extensive expression variation associated with transposable element-like silencing. The Plant Cell. 2023, koad233.
63. Rosa S, Duncan S, Dean C. Mutually exclusive sense–antisense transcription at FLC facilitates environmentally induced gene repression. Nature Communications. 2016, 7:13031.
64. Deforges J, Reis RS, Jacquet P, Sheppard S, Gadekar VP, Hart-Smith G, Tanzer A, Hofacker IL, Iseli C, Xenarios I, Poirier Y. Control of cognate sense mRNA translation by cis-natural antisense RNAs. Plant Physiology. 2019, 180:305-322.
65. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. Diversity and dynamics of the *Drosophila* transcriptome. Nature. 2014, 512:393-399.
66. Gao X, Wan J, Liu B, Ma M, Shen B, Qian S-B. Quantitative profiling of initiating ribosomes *in vivo*. Nature Methods. 2015, 12:147-153.
67. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 2019, 37:907-915.
68. Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Computational Biology. 2022, 18:e1009730.
69. Haas BJ, Salzberg SL, Zhu W, Pertea M. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. Genome Biology. 2008, 9:R7.
70. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005, 21:1859-1875.
71. Huang X, Adams MD, Zhou H, Kerlavage AR. A tool for analyzing and annotating genomic sequences. Genomics. 1997, 46:37-45.
72. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Amit I. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology. 2013, 29:644-652.
73. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004, 59:1-9.
74. Stanke M, Tzvetkova A, Morgenstern B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biology. 2006, 7:1-8.
75. Alexandre L, Burns PD, Mark B. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Research. 2014, 42:e119-e119.

913 76. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research. 2016, 44:W160-W165.

916 77. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Current Protocols in Molecular Biology. 2015, 109:21.

918 78. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods. 2013, 10:1213-1218.

921 79. Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Genome Research. 2015, 25:1757-1770.

924 80. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008, 9:R137.

927 81. Abugessaisa I, Noguchi S, Hasegawa A, Kondo A, Kawaji H, Carninci P, Kasukawa T. refTSS: a reference data set for human and mouse transcription start sites. Journal of Molecular Biology. 2019, 431:2407-2422.

930 82. Breiman L. Random Forests. Machine Learning. 2001, 45:5-32.

931 83. Xu H, Wang P, Fu Y, Zheng Y, Tang Q, Si L, You J, Zhang Z, Zhu Y, Zhou L, et al. Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. Cell Research. 2010, 20:445-457.

934 84. Jaime HC, Kristoffer F, Pedro CL, Damian S, Juhl JL, Christian VM, Peer B. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Molecular Biology & Evolution. 2016, 34:2115-2122.

937 85.Philip J, David B, Hsin-Yu C, Matthew F, Weizhong L, Craig MA, Hamish MW, John M, Alex M, Gift N. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014, 30:1236-1240.

940 86. Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids Research. 2018, 46:W84-W88.

942 87. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research. 2003, 31:365-370.

945 88. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J. The Pfam protein families database. Nucleic Acids Research. 2004, 32:D138-D141.

948 89. Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Research. 2015, 43:D213-D221.

951 90. Duttke SH, Chang MW, Heinz S, Benner C. Identification and dynamic quantification of regulatory elements using total RNA. Genome Research. 2019, 29:1836-1846.

953 91. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010, 26:841-842.

955 92. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 2016, 11:e0163962.

957 93. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Research. 2015, 43:W39-W49.

959 94. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018, 34:2666-2669.

961 95. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018, 34:3094-3100.

963 96. Chen C, Chen H, Zhang Y, Thomas HR, Xia R. TBtools: An integrative toolkit developed for
964 interactive analyses of big biological data**.** Molecular Plant*.* 2020, 13:1194-1202.
965 97. Wang G, Yin H, Li B, Yu C, Wang F, Xu X, Cao J, Bao Y, Wang L, Abbasi AA. Characterization
966 and identification of long non-coding RNAs based on feature relationship**.** Bioinformatics*.*
967 2019, 35:2949-2956.
968 98. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+:
969 architecture and applications**.** BMC Bioinformatics*.* 2009, 10:421.
970 99. Eddy SR. Accelerated profile HMM searches**.** PLOS Computational Biology*.* 2011,
971 7:e1002195.

972

973

Figure legends

Figure 1. Schematic overview of gene model optimization by integrating transcriptomic and epigenetic data.

A. Transcripts at different stages of growth, starvation, and conjugation were assembled into draft v1. By comparing newly assembled transcripts with those from TGD2021, well-annotated genes were retained, and error genes were optimized with the assistance of Nanopore DRS and strand-specific RNA-seq (ssRNA-seq) data, resulting in draft v2.

B. Epigenetic data were integrated to predict transcription start sites (TSSs) using a random forest (RF) model, and TSSs were further categorized into eTSSs and pTSSs with the addition of ATAC-seq data. Further optimization of the gene model was achieved using eTSSs, resulting in draft v3.

C. Transcripts in draft v3 were subjected to open reading frame (ORF) prediction, and UTR information was provided based on information of CDS, TSSs and transcription end sites (TESs), resulting in draft v4. Features of regulatory elements including promoters, poly-A sequences, and poly-A signals were analyzed.

D. The draft gene model v4 underwent two rounds of manual curation, followed by additional genome polish and protein function annotation, resulting in the generation of an improved gene model, draft v5.

E. Annotation of alternatively spliced (AS) isoforms was performed by integrating RNA-seq and Nanopore DRS data, resulting in TGD2024 (updated). Natural antisense transcripts (NATs) were annotated based on the updated gene model.

TGD: *Tetrahymena* genome database; NFR: nucleosome free region; PAS: poly-A signal.

Figure 2. IGV snapshots showing five categories of gene models optimized by transcriptomic data, including new gene (A), exon-altered gene (B), fused gene (C), partitioned gene (D), and orientation-reversed gene (E). Low-confidence genes (F) were not supported by RNA-seq data, thus retaining their annotations in draft v2 as in TGD2021. Tracks from top to bottom were RNA-seq (growth, starvation 24h, and conjugation at 4h, 5h, 6h, 8h, and 10h), Nanopore DRS coverage and reads alignment, and the gene models of draft v2 and TGD2021. Reads and gene models in pink represented transcription on the sense strand, and those in purple on the antisense strand.

Figure 3. Gene model optimization using epigenetic information.

A. Distribution profiles of H3K4me3, H2A.Z, 6mA, and nucleosome on the gene body. Genes were scaled to unit length and was extended to each side by 1kb length. Note that all four marks were accumulated downstream of TSS, towards the 5' end of the gene body.

B. The ROC-AUC curve (ROC: Receiver Operating Characteristics, AUC: Area Under the Curve) measuring the performance of our random forest (RF) model. The ROC was a probability curve and AUC represented the degree or measure of separability. The higher the AUC, the better the model was at predicting "TSS-region" classes as "TSS-region" classes or "not-TSS-region" classes. The AUC for both the training data and the testing data was close to 1, indicating excellent performance of our RF model in predicting TSS-region.

C-G. IGV snapshots of seven types of gene models optimized by epigenetic data with the complementation of transcriptomic data, including new gene (C), orientation-reversed gene (D), TSS-altered gene (E), fused gene (F), and partitioned gene (G). Partitioned gene was further subcategorized as co-directional (a), tail-to-tail (b), and head-to-head (c). The tracks from top to bottom were epigenetic information including nucleosome free region (NFR) deduced from ATAC-seq, H3K4me3, H2A.Z, 6mA, and nucleosome, and RNA-seq transcripts of different cell stages. The most highly expressed transcripts in conjugation were selected. Reads and gene models in pink represented transcription on the sense strand, and those in purple on the antisense strand.

Figure 4. UTR annotation and regulatory elements analysis.

A. Schematics for UTR annotation. ORF prediction was conducted on top of draft v3, resulting in a total of 27,494 protein-coding genes. 1,153 genes lacking ORF were defined as potential non-coding RNA. A putative ORF was defined as amino acids sequence longer than 100 aa. UTR information was further supplemented based on predicted TSSs and TESs. 1,271 low-confidence genes defined in Figure 2F lacked UTR annotations. Draft v3 after ORF prediction and UTR annotation generated draft v4.

B. UTR comparisons between draft v4 and TGD2014, the latter of which contained UTR information for 1,477 genes. Student's *t*-test was performed. **** $P<0.0001$.

C. Enriched core promoter motifs in promoter proximal sequences around TSS were identified by Homer [90]. P-values represented the statistical significance of

motif enrichment, indicating the likelihood that the observed frequency of each motif in the specified genomic region was greater than what would be expected by chance.

D. Venn diagram showing the composition of sequence motifs around poly-A signals (PAS). AATAAA was identified as the most predominant motif.

E. Summary of nucleotide frequencies and main regulatory elements around cleavage sites. Cleavage sites were significantly associated with the AT motif. The dashed black line represented positions of cleavage sites.

F. Length distribution of poly-A tails identified in Nanopore DRS (minimum reads >5). The median length was 18nt, illustrated by a dashed black line.

G. Distribution of the maximal poly-A tail length for each gene (number of gene with poly-A = 21,660). All genes were sorted by the length of their longest poly-A tails from shortest to longest and divided into three groups: 1) the first 25% of genes, defined as short-tailed genes, with poly-A tail length ranging from 5-19 nt; 2) the middle 25%-75% of genes, defined as medium-tailed genes, with poly-A tail length between 19-239 nt; 3) the remaining 25% of genes, defined as long-tailed genes, with poly-A tail length exceeding 239 nt.

H. Gene ontology (GO) analysis revealed that short-tailed and long-tailed genes were enriched in distinct functional groups. The colored bars represented the percentage of genes in each tail-length category.

I. Distribution of poly-A tail length in different functional groups. Student's *t*-test was performed. ** P <0.01, *** P <0.001, and ns P >0.05.

J. The Spearman's correlation between poly-A tail length and gene expression level (rho=0.72, P <2.2e-16). The longest poly-A tail was selected as the representative for each gene. Gene expression levels were quantified using the number of Nanopore DRS reads, with the removal of interference from antisense RNA. Both axes were plotted on a logarithmic scale.

Figure 5. Manual curation, genome polish, and protein function annotation.

A. Illustration of manual curation and genome polish on draft v4, resulting in draft v5. Two rounds of manual curation were conducted for all 180 non-rDNA chromosomes, focusing on genes with more than one eTSS, as well as those with neither eTSS nor pTSS. Genome polish was conducted by correcting error sites identified through manual curation using Illumina sequencing data.

B. An IGV snapshot showing the manual curation of a multi-eTSS gene, based on epigenetic and transcriptomic data. The tracks from top to bottom were nucleosome free region (NFR), H3K4me3, H2A.Z, 6mA, nucleosome, RNA-seq transcripts from different cell stages, and Nanopore DRS transcripts. The arrows and dashed lines indicated positions of eTSSs.

C. An IGV snapshot showing the manual curation of a tandem duplicate gene, by incorporating RNA-seq transcripts of different cell stages with its corresponding reads alignment and Nanopore DRS reads. The arrows indicated the chimeric alignment of RNA-seq transcripts.

D. An IGV snapshot showing the manual curation of a universally alternatively spliced gene, by incorporating RNA-seq transcripts of different cell stages with its corresponding reads alignment and Nanopore DRS reads. In the magnified box on the right, arrows indicated the universal alternatively spliced site. These universally alternatively spliced genes were annotated with their most dominant isoforms.

E. An IGV snapshot showing that multi-short-exon genes were always error-assembled when using Nanopore DRS data. This manual curation was performed with the aid of RNA-seq data from multiple stages. In the magnified box on the bottom, arrows indicated error-assembled sites.

F. An IGV snapshot showing an insertion site located in the exon resulted in erroneous CDS predictions. This manual curation was supported by both Illumina and transcriptomic data. The arrow and box indicated the insertion site.

Figure 6. Annotation of alternatively spliced (AS) isoforms in *Tetrahymena*.

A. The representative display of gene models and IGV snapshots of Nanopore DRS reads for six different AS types.

B. Comparative summary of gene and isoform numbers in each of the six different AS types in TGD02021 and TGD2024 (updated).

C. A heatmap depicting the expression profiles of AS transcripts across different stages: growth, starvation for 24h, and conjugation at 4h, 5h, 6h, 8h, and 10h.

D. A representative gene exhibiting a stage-specific tendency for intron retention, supported by transcriptomic data (left), as well as the ratio of retained intron and the ratio of the intron-containing reads (right). The ratio of retained intron was defined as the retained intron number divided by the total sequenced intron

number of the gene. The ratio of the intron-containing reads was defined as the reads aligned to the intron divided by the total reads aligned to the gene.

Figure 7. Identification and characterization of five types of natural antisense transcripts (NATs) in *Tetrahymena*.

A. Schematics for NATs annotation. NATs were identified on the updated gene model (TGD2024) using both transcriptomic and epigenetic data. Identified NATs were further categorized based on their relative positions to corresponding sense transcripts.

B-F. IGV snapshots showing five types of NATs. They included promoter NATs (B), originating from shared bidirectional promoters of the sense transcripts; intergenic NATs (C), transcribed from the upstream or downstream of the sense transcripts and possessing their own promoters; type 1 exonic NATs (D), located within 1 kb downstream of the TSSs of the sense transcripts and sharing epigenetic marks with their sense transcripts; type 2 exonic NATs (E), located more than 1 kb downstream of the TSSs of the sense transcripts; and intronic NATs (F), transcribed from the intronic regions of sense transcripts.

G. Distribution profiles of H3K4me3, H2A.Z, 6mA, and well-positioned nucleosomes on the transcript body of NATs. Transcripts were scaled to unit length and was extended to each side by 1kb length.

H. An IGV snapshot showing the anti-correlation of temporal expression patterns between a NAT and its corresponding sense transcript (left). The line chart (right) depicted the proportion of expression level for sense and antisense transcripts at different time points. The error bar represented the standard deviation (SD).

I. The box plot showing that the alternative splicing diversity (ASD) of NATs exceeded that of their sense transcripts (the median of NATs and sense transcripts were 0.96 and 0.28, respectively). Student's *t*-test was performed. *** P <0.001. ASD was defined as the number of different types of splice sites divided by the total reads aligned to the NATs or sense transcripts.

Figure S1. The flowchart for genome annotation.

Figure S2. RT-PCR validation of gene models optimized by the transcriptomic data, including new genes (A), exon-altered genes (B), fused genes (C), and partitioned genes (D). For each representative gene, gene model in draft v2 and

TGD2021 (top), primers flanking the target region and the expected size of RT-PCR products (bottom left), and the gel electrophoresis image of RT-PCR products (bottom right) were displayed. Orange and green stars indicated corresponding RT-PCR products.

Figure S3. The eTSSs/pTSSs prediction and gene optimization using epigenetic data.

A. Peaks corresponding to nucleosome free regions (NFRs) were identified using *Homo sapiens* ATAC-seq data and were subsequently integrated with TSS data from RefTSS for analysis. The results indicated that ATAC-seq fragments from NFRs tended to be enriched around TSS.

B. The prediction of eTSSs and pTSSs using epigenetic data. Genes were classified into different groups according to the presence or absence of eTSSs and pTSSs.

C-F. IGV snapshots of different types of genes optimized by epigenetic data, complemented by transcriptomic data, including gene pairs sharing their promoters (C), genes with multiple short exons (D), and tandem duplicate genes (E). Low-confidence genes (F) were neither supported by epigenetic marks nor Nanopore DRS reads.

Figure S4. RT-PCR validation of gene models optimized by eTSSs, including new genes (A), fused genes (B), and co-directional partitioned genes (C), tail-to-tail partitioned genes (D), and head-to-head partitioned genes (E). For each representative gene, gene models in draft v2 and draft v3 (top), primers flanking the target region and the expected size of RT-PCR products (bottom left), and the gel electrophoresis image of RT-PCR products (bottom right) were displayed. Orange and green stars indicated corresponding RT-PCR products.

Figure S5. The regulatory elements of untranslated regions analysis in *Tetrahymena*.

A. The workflow of Nanopore sequencing and the prediction of poly-A tails.

B. An IGV snapshot of a gene with multiple TESs, identified according to the position of poly-A tails in DRS reads.

C. Distribution frequency of main regulatory motifs around cleavage sites. The dominant motifs of PASs (AATAAA and ATTAAA) were enriched within 20 bp

upstream of cleavage sites. The regulatory elements such as GTGT (GT-rich elements) and TGTA (AT-rich elements) were enriched more than 30 bp upstream of cleavage sites.

D. The Spearman's correlation showed no correlation between poly-A tail length and gene length (rho= -0.089, *P* <2.2e-16). The longest poly-A tail was selected as the representative for each gene. Both axes were plotted on a logarithmic scale.

Figure S6. Four types of error sites polished by Illumina and Sanger sequencing data, including two examples each for insertion, deletion, transition, and transversion.

Figure S7. Protein function annotation on draft v4.

A. Protein sequences of TGD2021 and draft v4 were blasted against multiple public protein databases, featuring an additional 1,732 functional genes compared to TGD2021.

B. The number of functional genes annotated by different databases for TGD2021 and draft v5.

C. Top 25 functional gene groups for genes with new functions.

D. Domain structure of four newly annotated proteins possibly related to epigenetic regulations, homologous to MLL5 and NSUN4, respectively.

Figure S8. The experimental validation and functional analysis for alternative splicing (AS) isoforms.

A-F. RT-PCR validation of the presence of AS transcripts, including exon skipping (A), alternative last exon (B), intron retention (C), mutually exclusive exons (D), alternative 5' splice site (E), and alternative 3' splice site (F). For each representative gene, gene model in TGD2024 (updated) and isoform, primers flanking the target region and the expected size of RT-PCR products (bottom left), and the gel electrophoresis image of RT-PCR products (bottom right) were displayed. Orange and green stars indicated corresponding RT-PCR products.

G. GO analysis of total AS isoforms revealed that genes possessing AS isoforms were primarily enriched in processes related to cell cycle regulation.

Figure S9. Identification and characterization of natural antisense transcripts (NATs).

A. NATs were divided into non-coding, potential-coding and functional-coding, according to their significantly different coding potential scores. ****, *P*<0.0001.

B. An IGV snapshot showing the splice sites of a NAT (red) and its corresponding sense transcript (purple). Note that this NAT possessed its own canonical GU-AG site for intron splicing and its exon-intron boundary slightly deviated from that of its sense transcript.

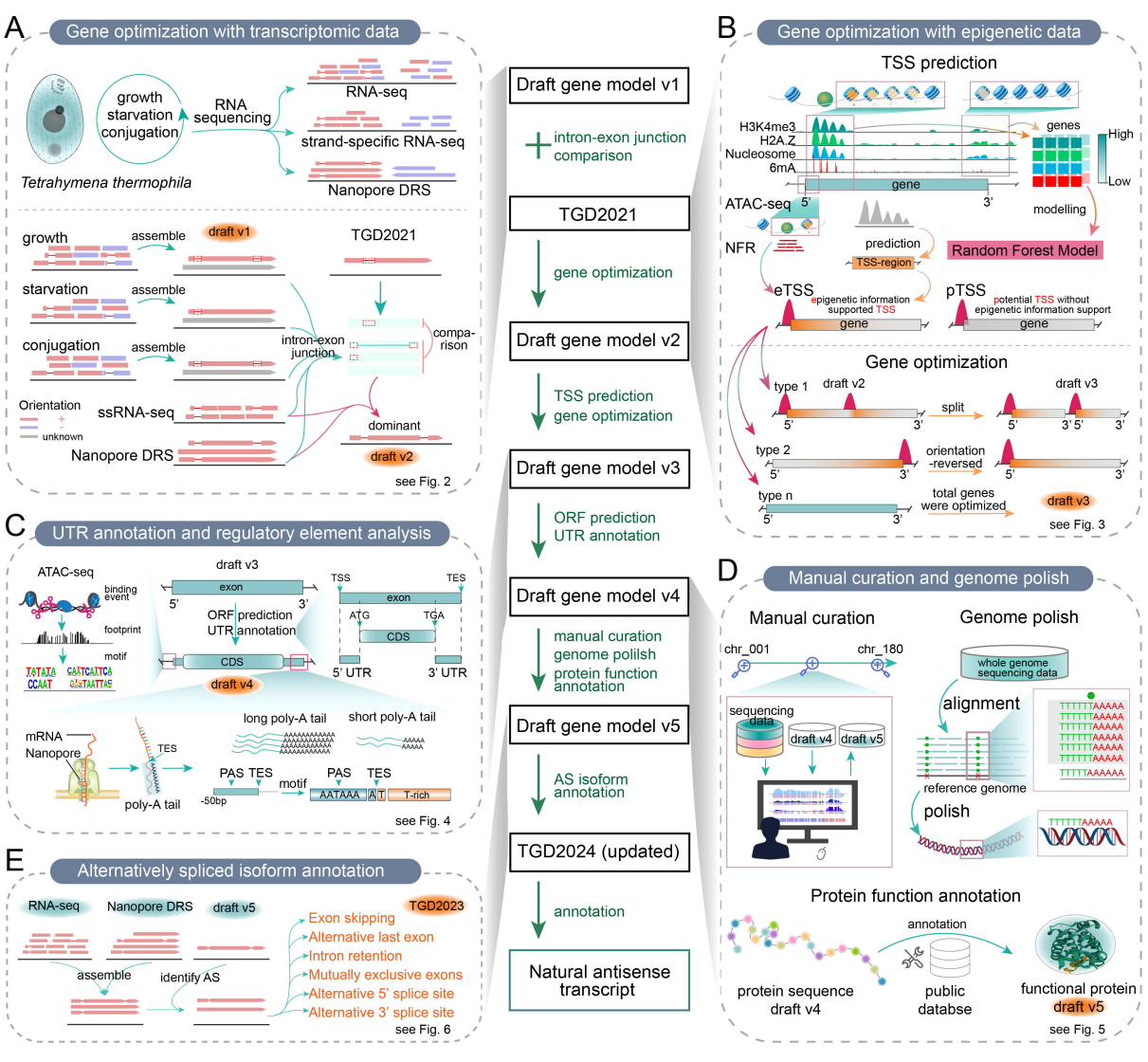
C. Distribution of intron-exon boundaries of NATs in relation to those of sense transcripts. The red dashed line at position 0 marked the location of the sense intron-exon boundaries.

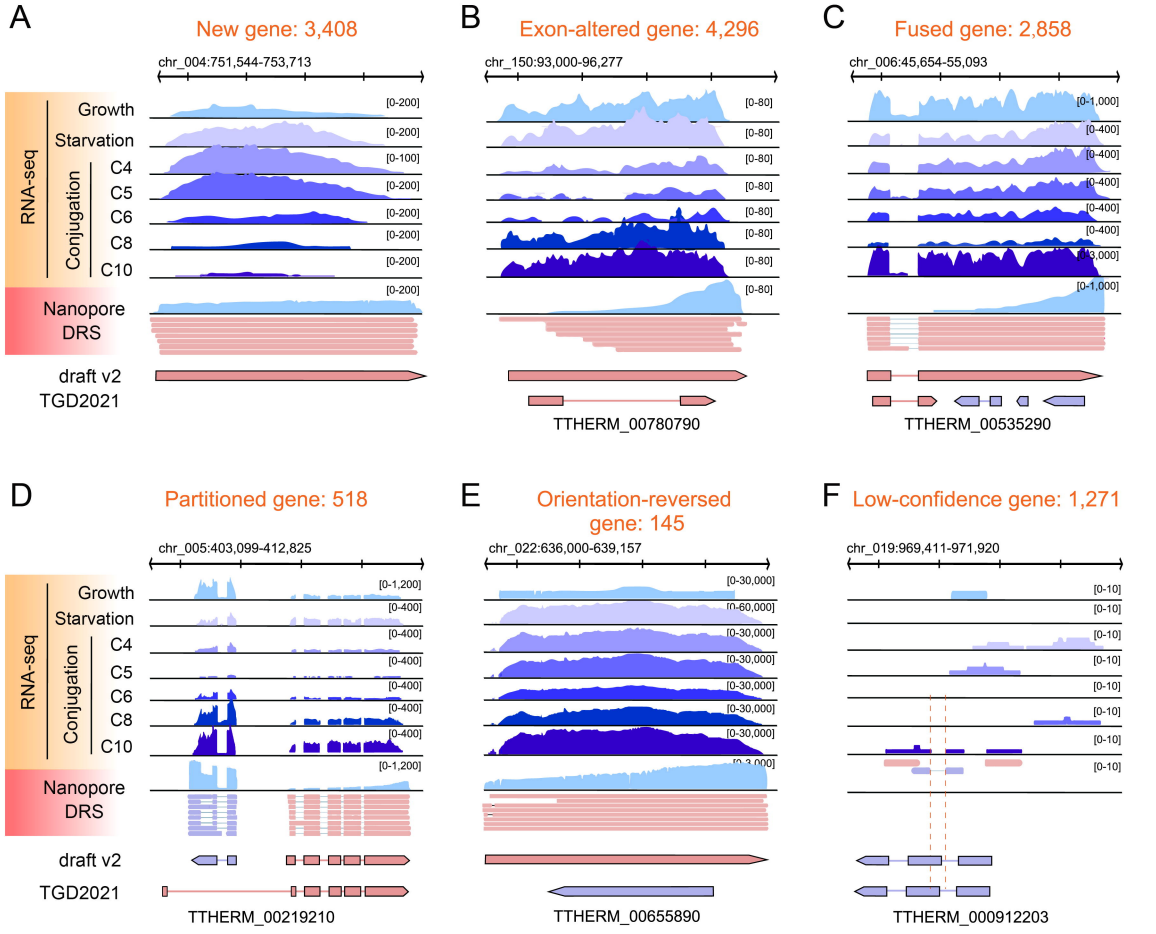
D. The heatmap showing the GC content and length of sense transcripts and antisense transcripts. NATs had shorter length but had no difference in GC content comparing to sense transcripts.

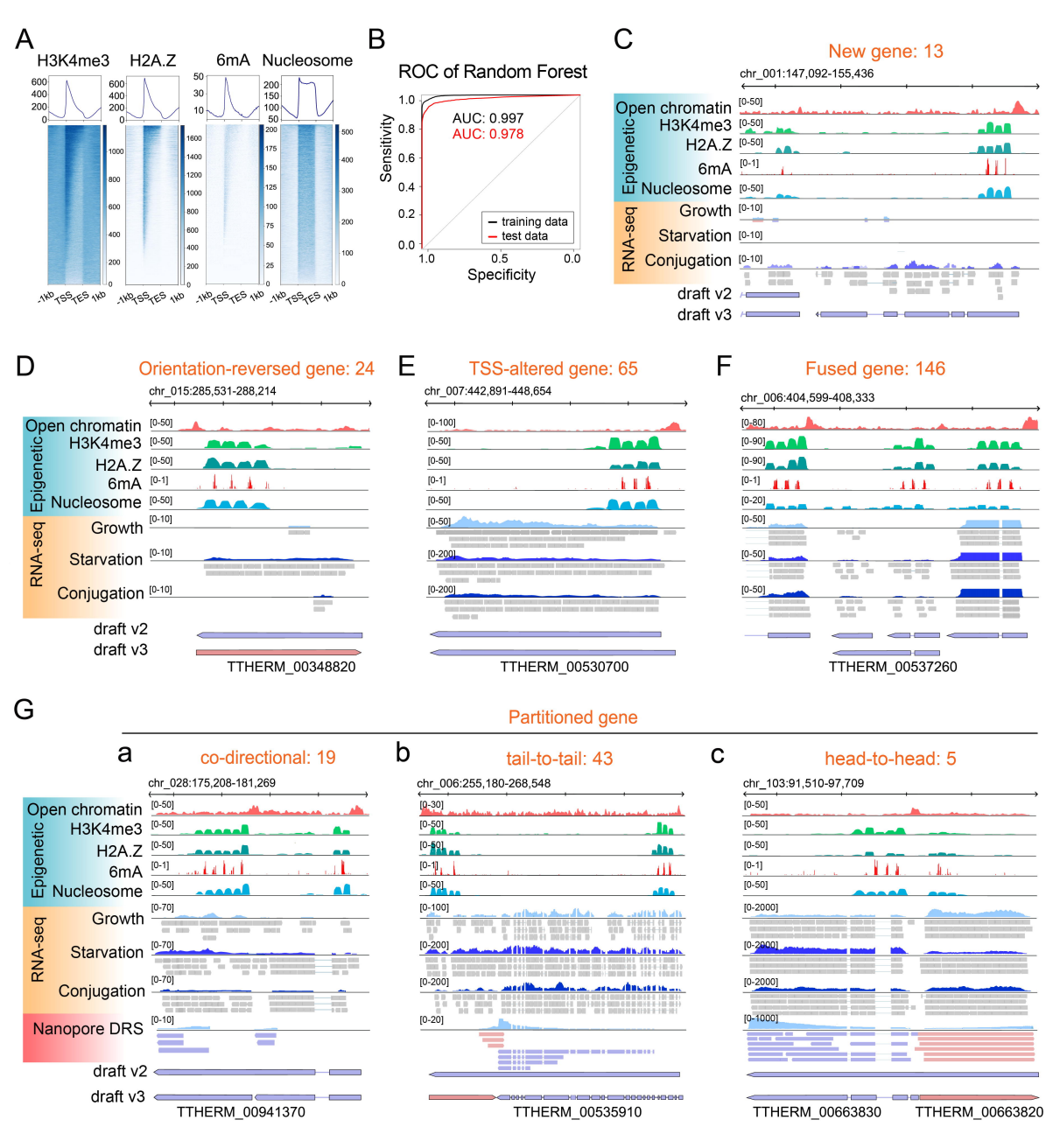
E. The box plot showing that NATs had lower expression level comparing to sense transcripts. The expression level was defined as DRS reads counts of NATs and its corresponding sense transcripts. ***, *P*<0.001.

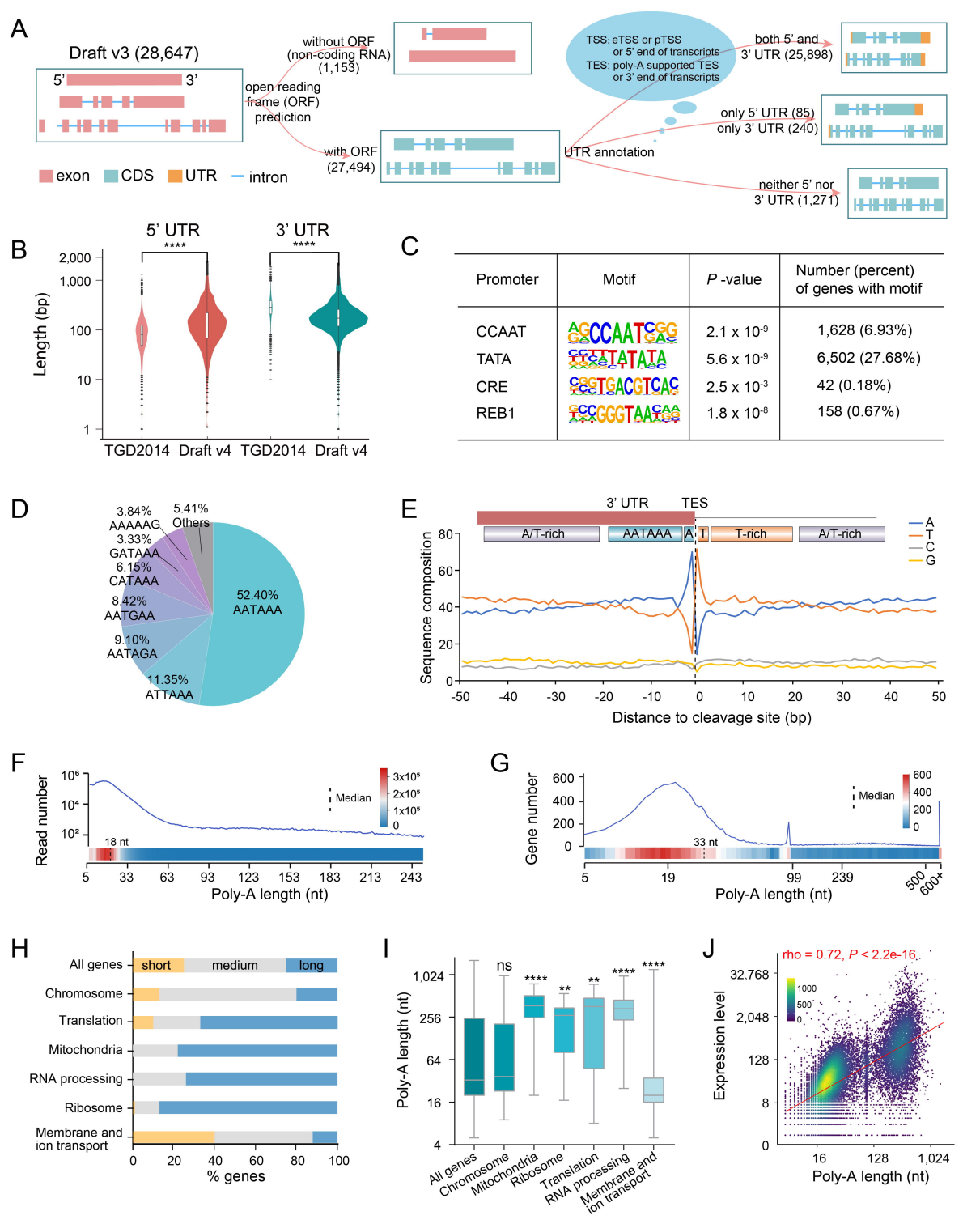
F. Composite distribution profiles of H3K4me3, H2A.Z, 6mA, and nucleosome near the TSS of NATs. The distribution profiles showing that all four marks were accumulated downstream of TSS.

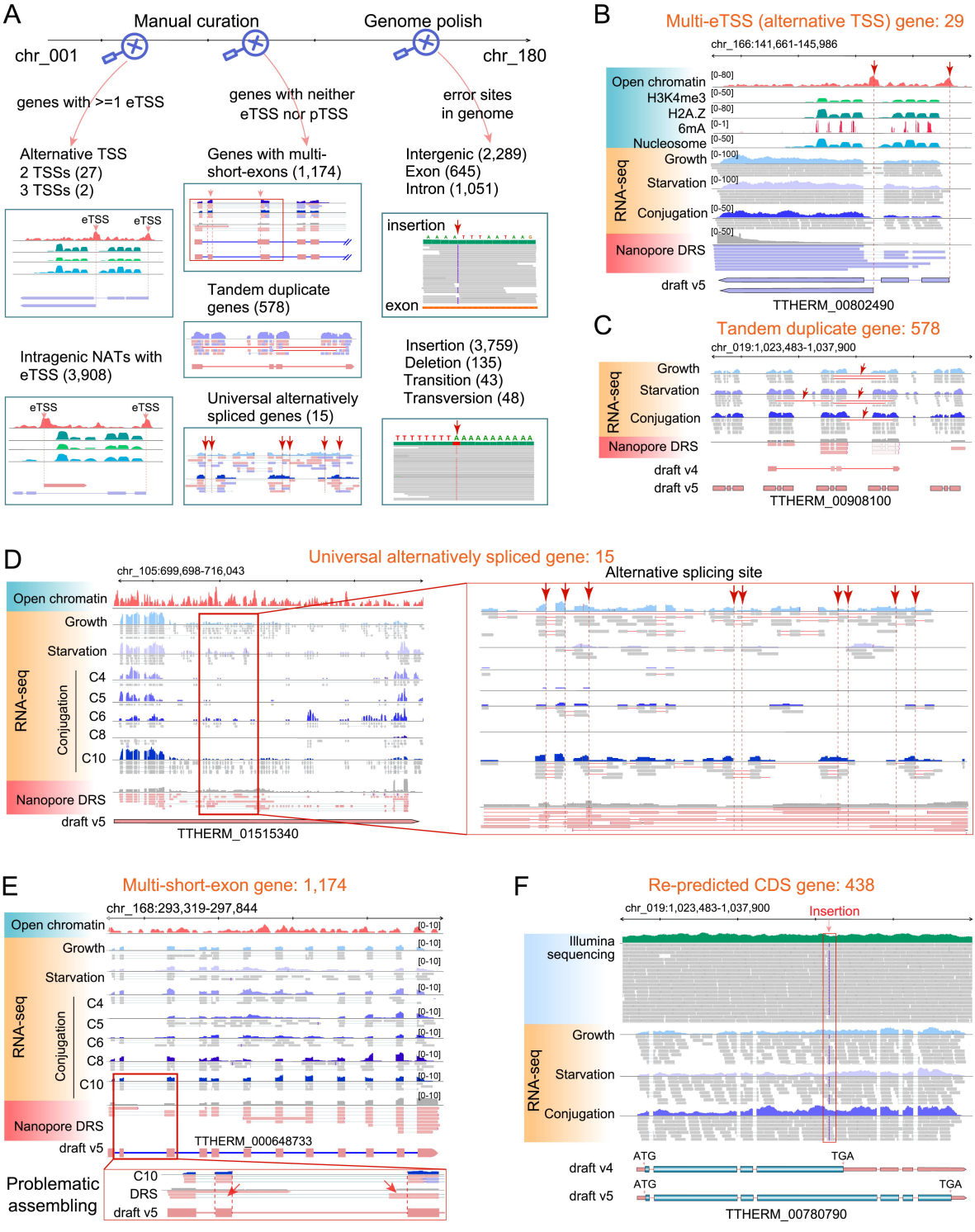
G. The box plot showing that the alternative splicing diversity (ASD) of total NATs exceeded the total sense transcripts (the median of NATs and sense transcripts were 0.96 and 0.15, respectively). Student's *t*-test was performed. ***, *P* <0.001. ASD was defined as the number of different splice sites divided by the total reads aligned to the NATs or sense transcripts.

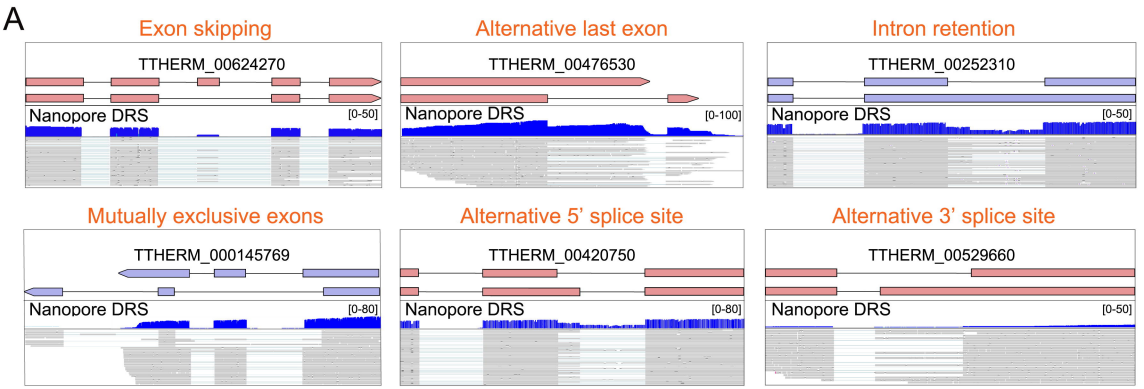

A
Gene optimization with transcriptomic data
growth
starvation
conjugation
RNA sequencing
RNA-seq
strand-specific RNA-seq
Nanopore DRS
Tetrahymena thermophila
growth
assemble
draft v1
starvation
assemble
conjugation
assemble
intron-exon junction
TGD2021
compa-rison
Orientation
unknown
ssRNA-seq
Nanopore DRS
dominant
draft v2
see Fig. 2
Draft gene model v1
intron-exon junction comparison
TGD2021
gene optimization
Draft gene model v2
TSS prediction
gene optimization
Draft gene model v3
ORF prediction
UTR annotation
Draft gene model v4
manual curation
genome polilsh
protein function annotation
Draft gene model v5
AS isoform annotation
TGD2024 (updated)
annotation
Natural antisense transcript
B
Gene optimization with epigenetic data
TSS prediction
H3K4me3
H2A.Z
Nucleosome
6mA
genes
High
Low
gene
ATAC-seq
5'
3'
modelling
NFR
prediction
TSS-region
Random Forest Model
eTSS
epigenetic information supported TSS
gene
pTSS
potential TSS without epigenetic information support
gene
Gene optimization
type 1
draft v2
split
draft v3
type 2
orientation -reversed
type n
total genes were optimized
draft v3
see Fig. 3
C
UTR annotation and regulatory element analysis
ATAC-seq
binding event
footprint
motif
draft v3
exon
ORF prediction
UTR annotation
CDS
draft v4
TSS
TES
ATG
TGA
5' UTR
3' UTR
mRNA
Nanopore
TES
poly-A tail
long poly-A tail
short poly-A tail
PAS
motif
-50bp
AATAAA
T-rich
see Fig. 4
D
Manual curation and genome polish
Manual curation
chr_001
chr_180
sequencing data
draft v4
draft v5
Genome polish
whole genome sequencing data
alignment
reference genome
polish
Protein function annotation
annotation
protein sequence
draft v4
public databse
functional protein
draft v5
see Fig. 5
E
Alternatively spliced isoform annotation
RNA-seq
Nanopore DRS
draft v5
TGD2023
assemble
identify AS
Exon skipping
Alternative last exon
Intron retention
Mutually exclusive exons
Alternative 5' splice site
Alternative 3' splice site
see Fig. 6


A
New gene: 3,408
chr_004:751,544-753,713
RNA-seq
Growth
Starvation
Conjugation
C4
C5
C6
C8
C10
Nanopore DRS
draft v2
TGD2021
B
Exon-altered gene: 4,296
chr_150:93,000-96,277
TTHERM_00780790
C
Fused gene: 2,858
chr_006:45,654-55,093
TTHERM_00535290
D
Partitioned gene: 518
chr_005:403,099-412,825
TTHERM_00219210
E
Orientation-reversed gene: 145
chr_022:636,000-639,157
TTHERM_00655890
F
Low-confidence gene: 1,271
chr_019:969,411-971,920
TTHERM_000912203


A
H3K4me3
H2A.Z
6mA
Nucleosome
-1kb TSS TES 1kb
B
ROC of Random Forest
Sensitivity
Specificity
AUC: 0.997
AUC: 0.978
training data
test data
C
New gene: 13
chr_001:147,092-155,436
Epigenetic
Open chromatin
H3K4me3
H2A.Z
6mA
Nucleosome
RNA-seq
Growth
Starvation
Conjugation
draft v2
draft v3
D
Orientation-reversed gene: 24
chr_015:285,531-288,214
TTHERM_00348820
E
TSS-altered gene: 65
chr_007:442,891-448,654
TTHERM_00530700
F
Fused gene: 146
chr_006:404,599-408,333
TTHERM_00537260
G
Partitioned gene
a
co-directional: 19
chr_028:175,208-181,269
Nanopore DRS
TTHERM_00941370
b
tail-to-tail: 43
chr_006:255,180-268,548
TTHERM_00535910
c
head-to-head: 5
chr_103:91,510-97,709
TTHERM_00663830
TTHERM_00663820

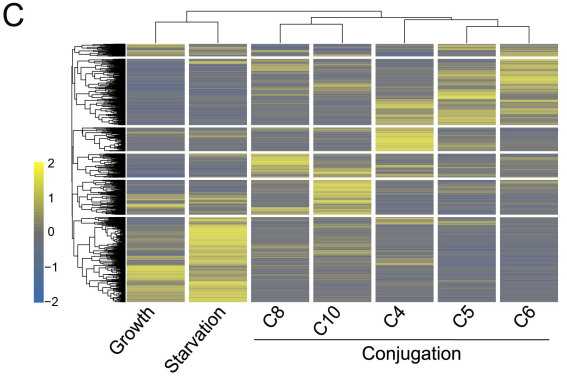

A
Draft v3 (28,647)
5'
3'
exon
CDS
UTR
intron
open reading frame (ORF) prediction
without ORF (non-coding RNA) (1,153)
with ORF (27,494)
TSS: eTSS or pTSS or 5' end of transcripts
TES: poly-A supported TES or 3' end of transcripts
UTR annotation
both 5' and 3' UTR (25,898)
only 5' UTR (85)
only 3' UTR (240)
neither 5' nor 3' UTR (1,271)
B
5' UTR
3' UTR
****
Length (bp)
2,000
1,000
100
10
1
TGD2014
Draft v4
TGD2014
Draft v4
C
Promoter
Motif
P -value
Number (percent) of genes with motif
CCAAT
2.1 x 10-9
1,628 (6.93%)
TATA
5.6 x 10-9
6,502 (27.68%)
CRE
2.5 x 10-3
42 (0.18%)
REB1
1.8 x 10-8
158 (0.67%)
D
3.84% AAAAAG
5.41% Others
3.33% GATAAA
6.15% CATAAA
8.42% AATGAA
9.10% AATAGA
11.35% ATTAAA
52.40% AATAAA
E
3' UTR
TES
A/T-rich
AATAAA
A
T
T-rich
A/T-rich
A
T
C
G
Sequence composition
80
60
40
20
0
-50
-40
-30
-20
-10
0
10
20
30
40
50
Distance to cleavage site (bp)
F
Read number
10^6
10^4
10^2
Median
3x10^5
2x10^5
1x10^5
0
18 nt
5
33
63
93
123
153
183
213
243
Poly-A length (nt)
G
Gene number
600
400
200
0
Median
600
400
200
0
33 nt
5
19
99
239
500
600+
Poly-A length (nt)
H
All genes
Chromosome
Translation
Mitochondria
RNA processing
Ribosome
Membrane and ion transport
short
medium
long
0
20
40
60
80
100
% genes
I
Poly-A length (nt)
1,024
256
64
16
4
ns
****
**
**
****
****
All genes
Chromosome
Mitochondria
Ribosome
Translation
RNA processing
Membrane and ion transport
J
rho = 0.72, P < 2.2e-16
1000
500
0
Expression level
32,768
2,048
128
8
0
16
128
1,024
Poly-A length (nt)

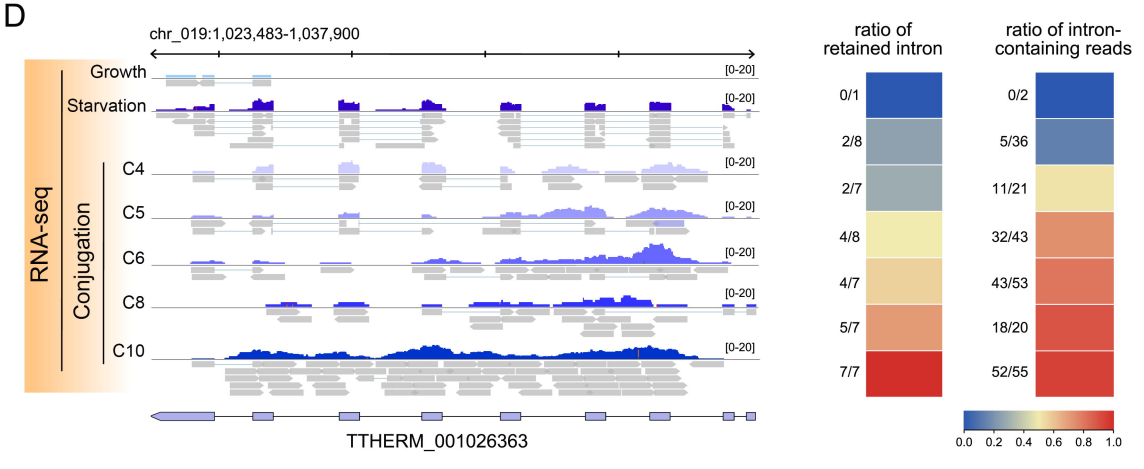

A
Manual curation
Genome polish
chr_001
chr_180
genes with >=1 eTSS
genes with neither eTSS nor pTSS
error sites in genome
Alternative TSS
2 TSSs (27)
3 TSSs (2)
eTSS
eTSS
Genes with multi-short-exons (1,174)
Tandem duplicate genes (578)
Intergenic (2,289)
Exon (645)
Intron (1,051)
insertion
exon
Intragenic NATs with eTSS (3,908)
eTSS
eTSS
Universal alternatively spliced genes (15)
Insertion (3,759)
Deletion (135)
Transition (43)
Transversion (48)
B
Multi-eTSS (alternative TSS) gene: 29
chr_166:141,661-145,986
Open chromatin
H3K4me3
H2A.Z
6mA
Nucleosome
RNA-seq
Growth
Starvation
Conjugation
Nanopore DRS
draft v5
TTHERM_00802490
C
Tandem duplicate gene: 578
chr_019:1,023,483-1,037,900
RNA-seq
Growth
Starvation
Conjugation
Nanopore DRS
draft v4
draft v5
TTHERM_00908100
D
Universal alternatively spliced gene: 15
chr_105:699,698-716,043
Alternative splicing site
Open chromatin
Growth
Starvation
RNA-seq
Conjugation
C4
C5
C6
C8
C10
Nanopore DRS
draft v5
TTHERM_01515340
E
Multi-short-exon gene: 1,174
chr_168:293,319-297,844
Open chromatin
Growth
Starvation
RNA-seq
Conjugation
C4
C5
C6
C8
C10
Nanopore DRS
draft v5
TTHERM_000648733
Problematic assembling
C10
DRS
draft v5
F
Re-predicted CDS gene: 438
chr_019:1,023,483-1,037,900
Insertion
Illumina sequencing
Growth
Starvation
RNA-seq
Conjugation
ATG
TGA
draft v4
ATG
TGA
draft v5
TTHERM_00780790

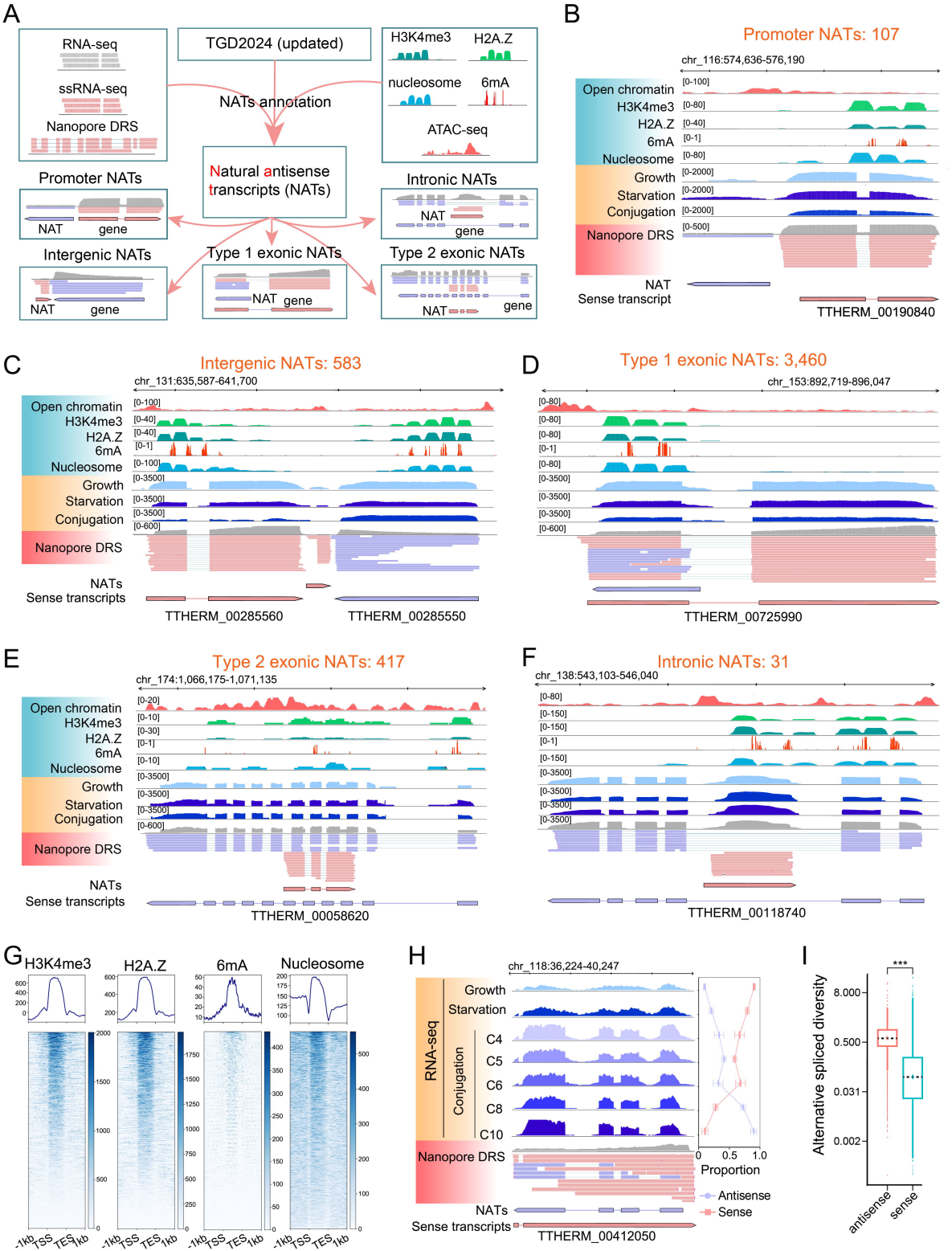


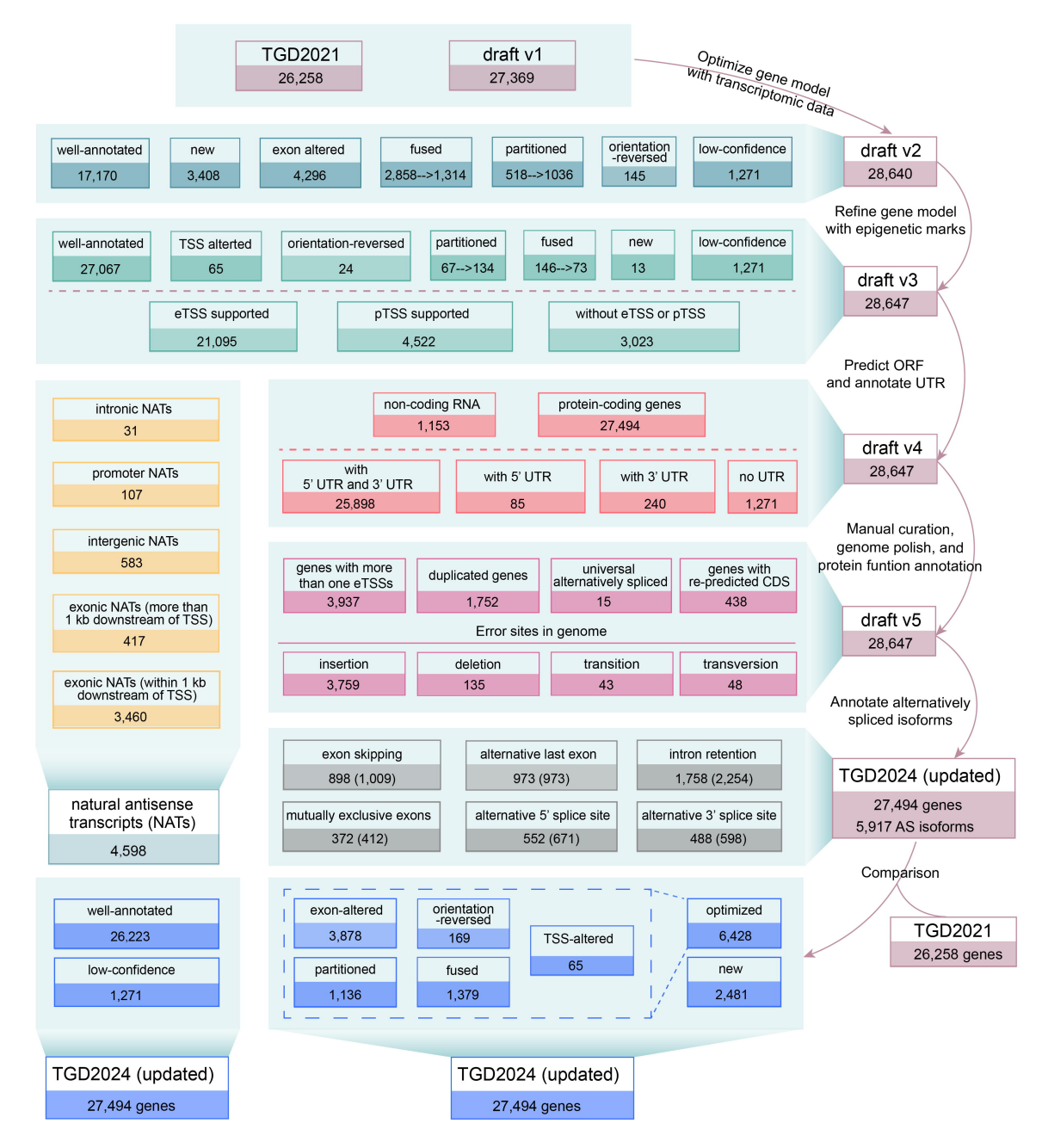
B

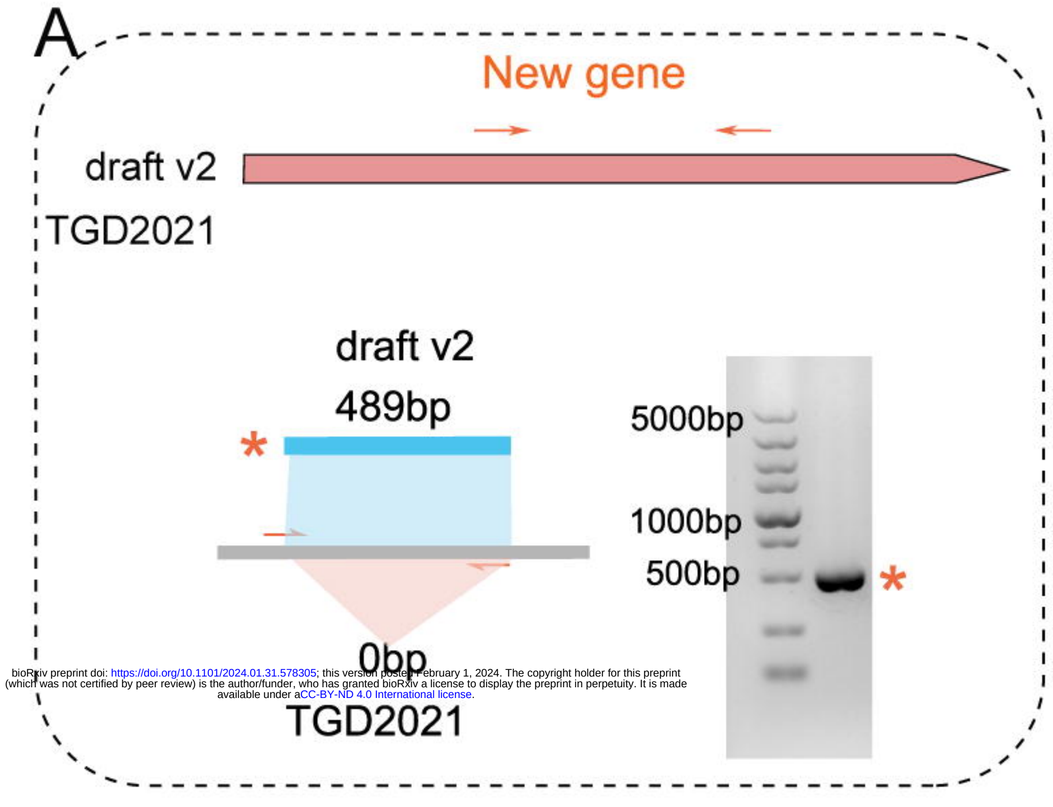
AS type	TGD2021	Updated
Exon skipping	23 (23)	898 (1,009)
Alternative last exon	145 (145)	973 (973)
Intron retention	92 (92)	1,758 (2,254)
Mutually exclusive exons	54 (54)	372 (412)
Alternative 5' splice site	88 (88)	552 (671)
Alternative 3' splice site	51 (51)	488 (598)
		gene (isoform)

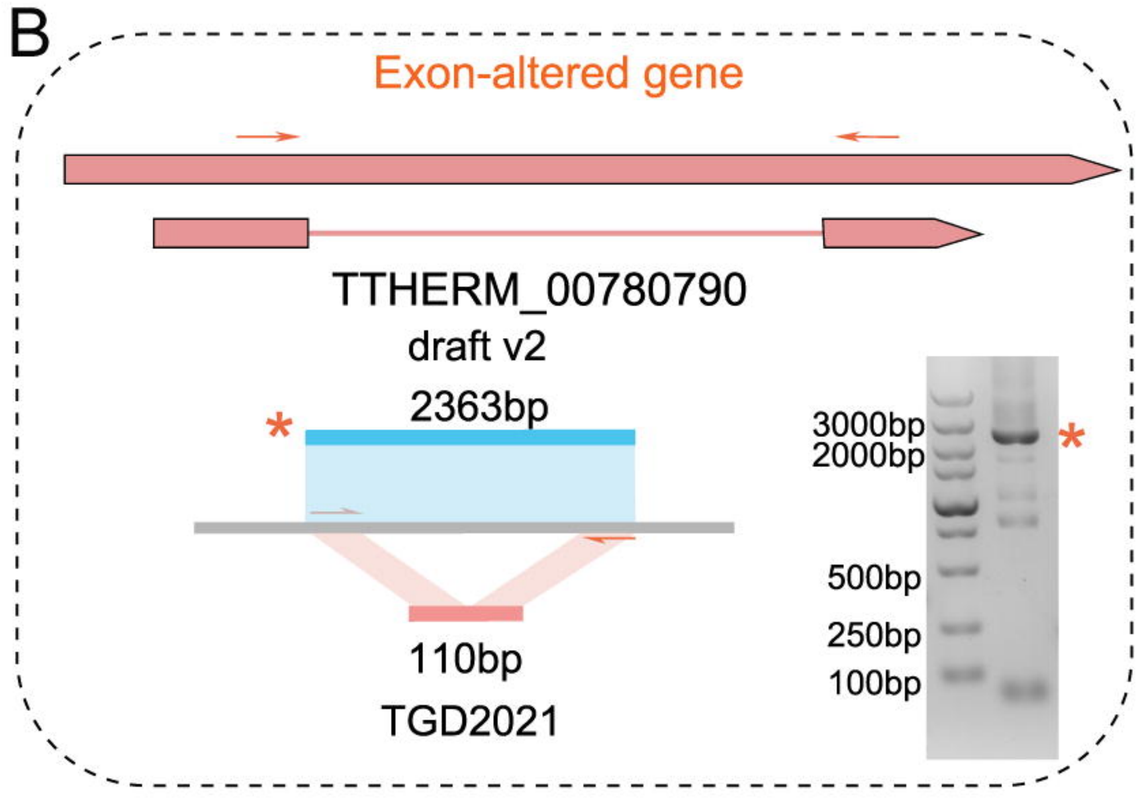


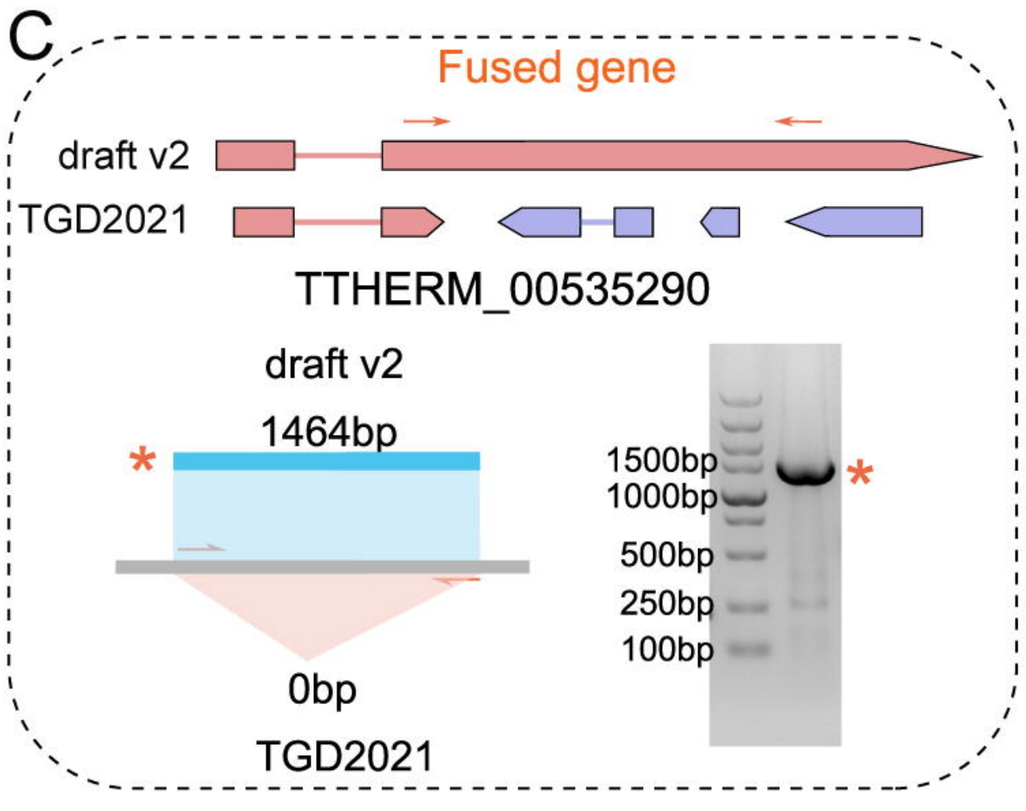


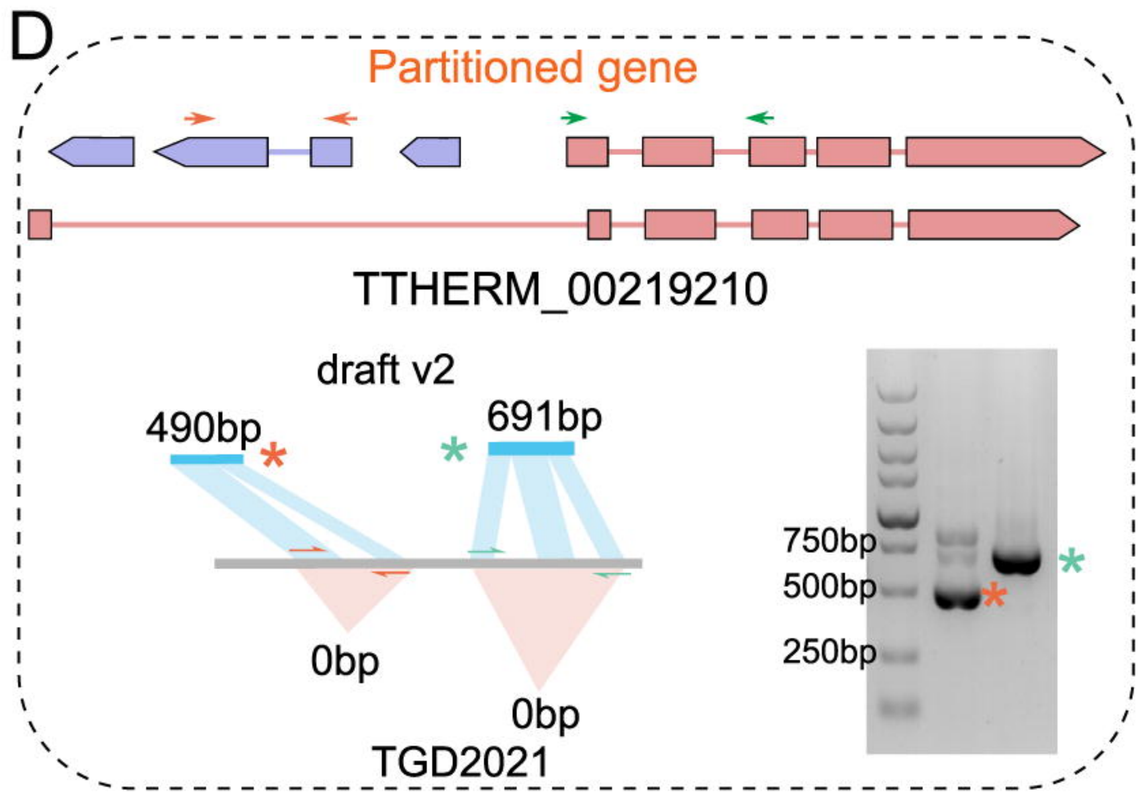

A
RNA-seq
ssRNA-seq
Nanopore DRS
TGD2024 (updated)
NATs annotation
H3K4me3
H2A.Z
nucleosome
6mA
ATAC-seq
Natural antisense transcripts (NATs)
Promoter NATs
NAT
gene
Intronic NATs
Intergenic NATs
Type 1 exonic NATs
Type 2 exonic NATs
B
Promoter NATs: 107
chr_116:574,636-576,190
Open chromatin
Growth
Starvation
Conjugation
Nanopore DRS
NAT
Sense transcript
TTHERM_00190840
C
Intergenic NATs: 583
chr_131:635,587-641,700
NATs
Sense transcripts
TTHERM_00285560
TTHERM_00285550
D
Type 1 exonic NATs: 3,460
chr_153:892,719-896,047
TTHERM_00725990
E
Type 2 exonic NATs: 417
chr_174:1,066,175-1,071,135
TTHERM_00058620
F
Intronic NATs: 31
chr_138:543,103-546,040
TTHERM_00118740
G
H3K4me3
H2A.Z
6mA
Nucleosome
-1kb
TSS
TES
1kb
H
chr_118:36,224-40,247
RNA-seq
Growth
Starvation
Conjugation
C4
C5
C6
C8
C10
Nanopore DRS
NATs
Sense transcripts
TTHERM_00412050
Proportion
Antisense
Sense
I
Alternative spliced diversity
***
antisense
sense

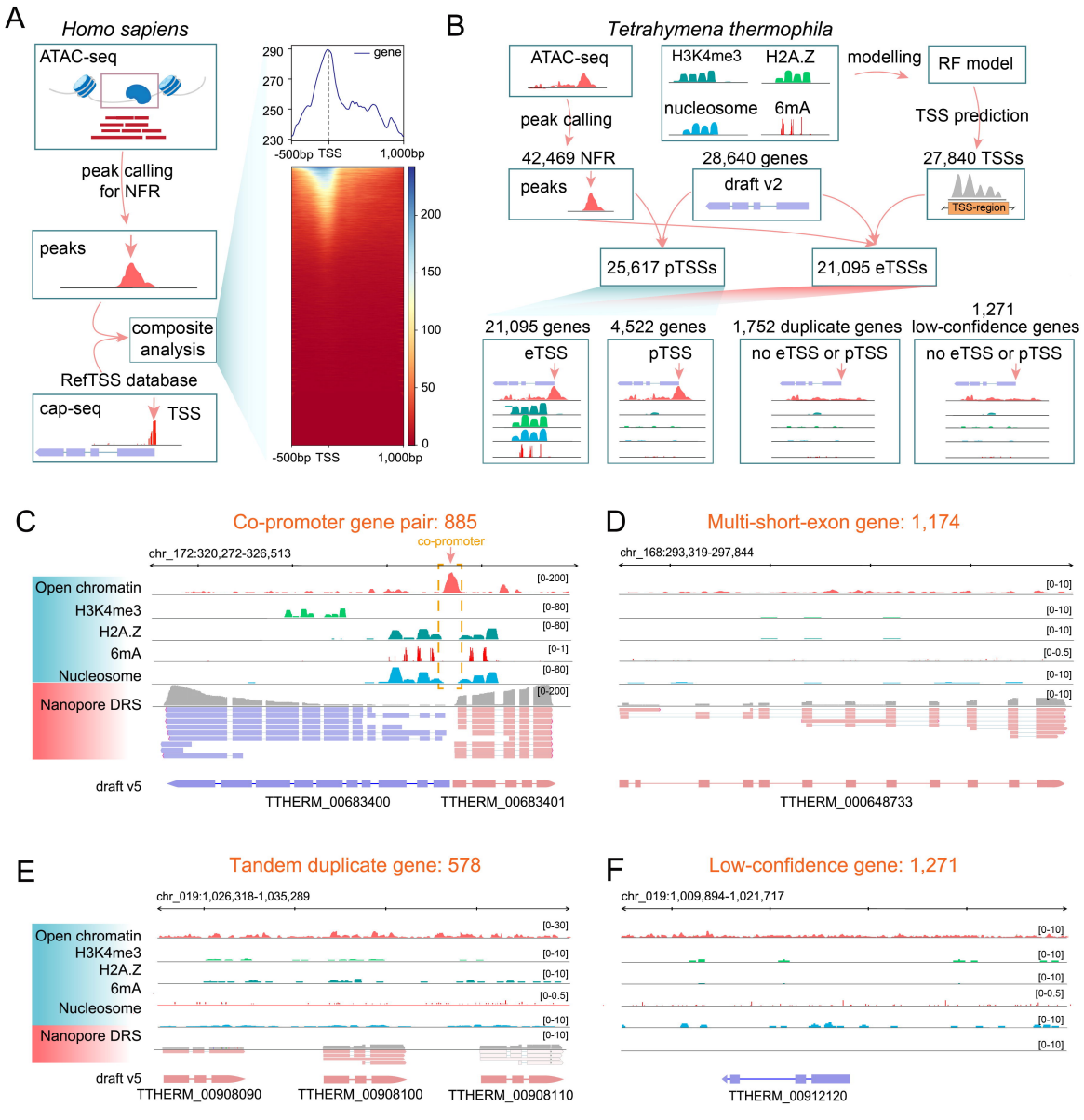

TGD2021
26,258
draft v1
27,369
Optimize gene model with transcriptomic data
well-annotated
17,170
new
3,408
exon altered
4,296
fused
2,858-->1,314
partitioned
518-->1036
orientation -reversed
145
low-confidence
1,271
draft v2
28,640
Refine gene model with epigenetic marks
well-annotated
27,067
TSS alterted
65
orientation-reversed
24
partitioned
67-->134
fused
146-->73
new
13
low-confidence
1,271
eTSS supported
21,095
pTSS supported
4,522
without eTSS or pTSS
3,023
draft v3
28,647
Predict ORF and annotate UTR
non-coding RNA
1,153
protein-coding genes
27,494
with 5' UTR and 3' UTR
25,898
with 5' UTR
85
with 3' UTR
240
no UTR
1,271
draft v4
28,647
Manual curation, genome polish, and protein funtion annotation
genes with more than one eTSSs
3,937
duplicated genes
1,752
universal alternatively spliced
15
genes with re-predicted CDS
438
Error sites in genome
insertion
3,759
deletion
135
transition
43
transversion
48
draft v5
28,647
Annotate alternatively spliced isoforms
intronic NATs
31
promoter NATs
107
intergenic NATs
583
exonic NATs (more than 1 kb downstream of TSS)
417
exonic NATs (within 1 kb downstream of TSS)
3,460
natural antisense transcripts (NATs)
4,598
exon skipping
898 (1,009)
alternative last exon
973 (973)
intron retention
1,758 (2,254)
mutually exclusive exons
372 (412)
alternative 5' splice site
552 (671)
alternative 3' splice site
488 (598)
TGD2024 (updated)
27,494 genes
5,917 AS isoforms
Comparison
TGD2021
26,258 genes
well-annotated
26,223
low-confidence
1,271
exon-altered
3,878
orientation -reversed
169
TSS-altered
65
partitioned
1,136
fused
1,379
optimized
6,428
new
2,481
TGD2024 (updated)
27,494 genes
TGD2024 (updated)
27,494 genes

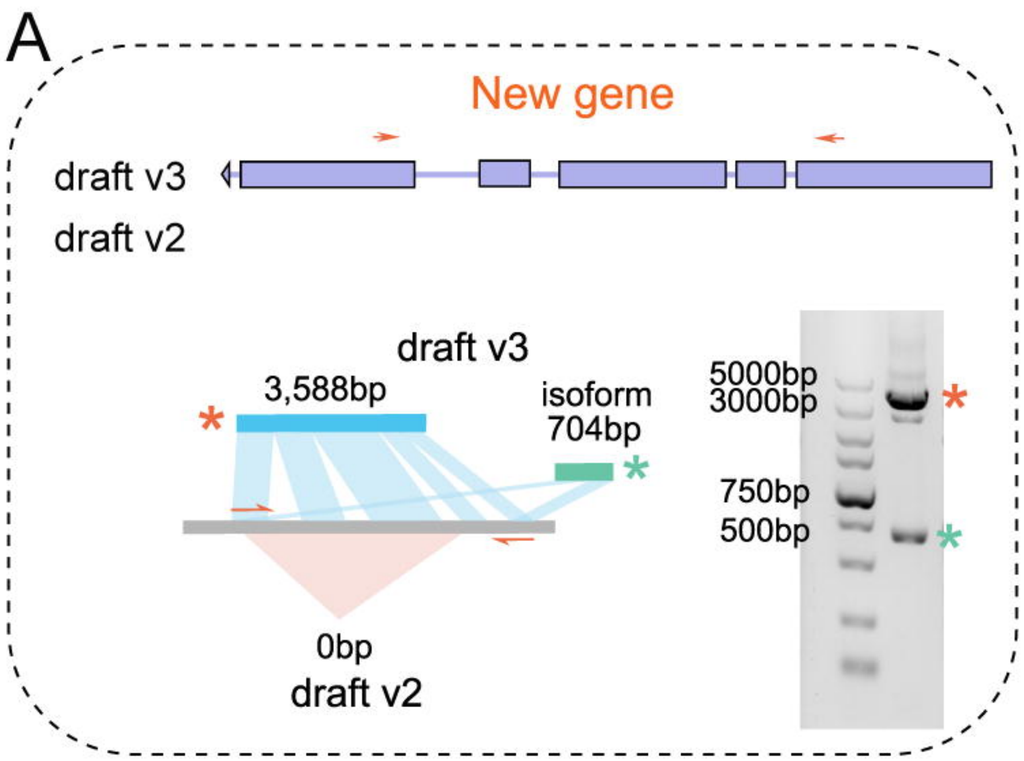


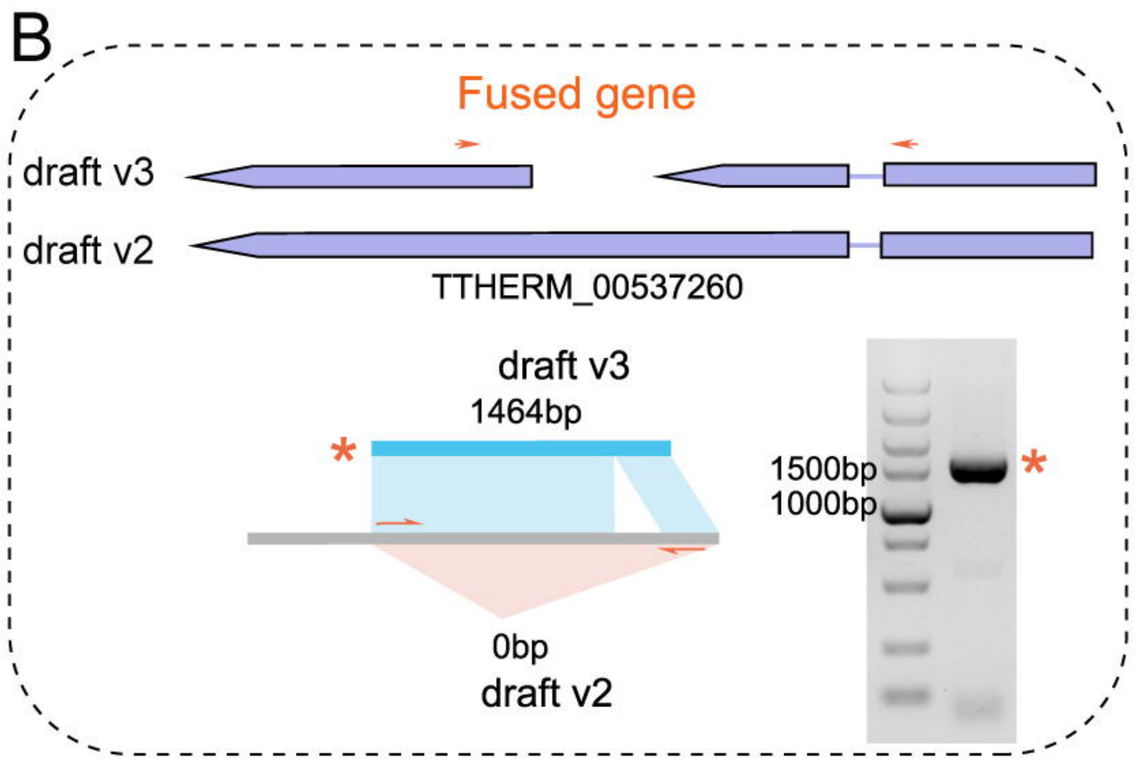


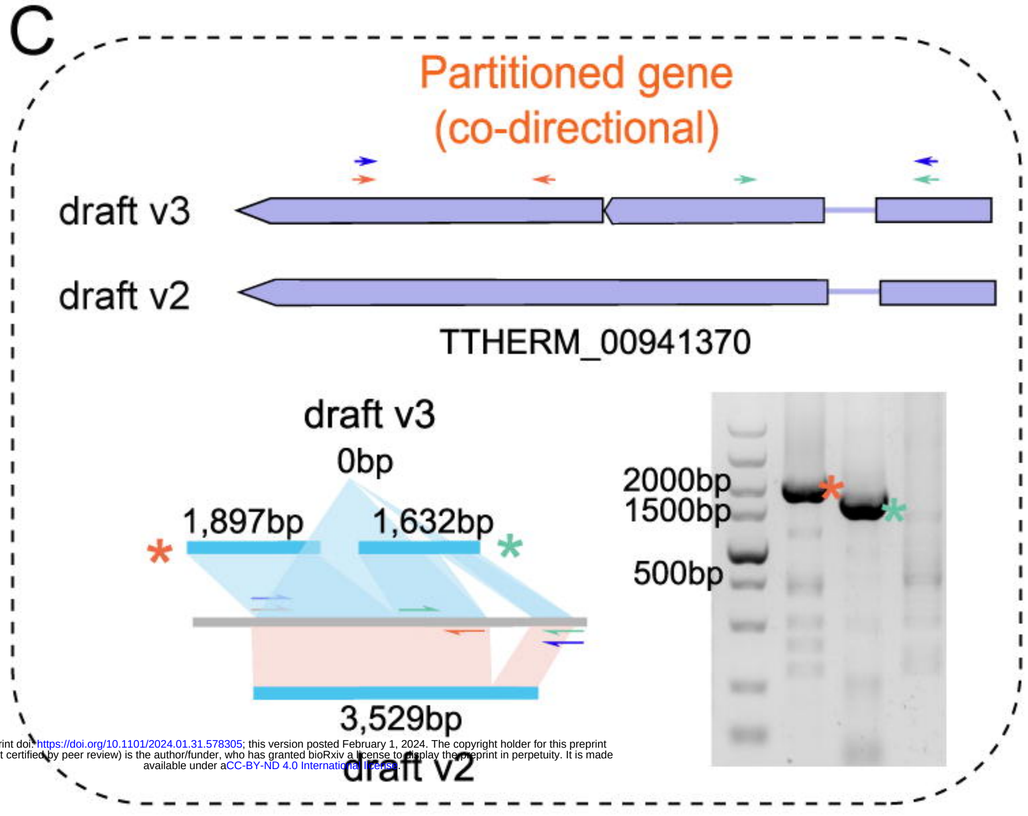


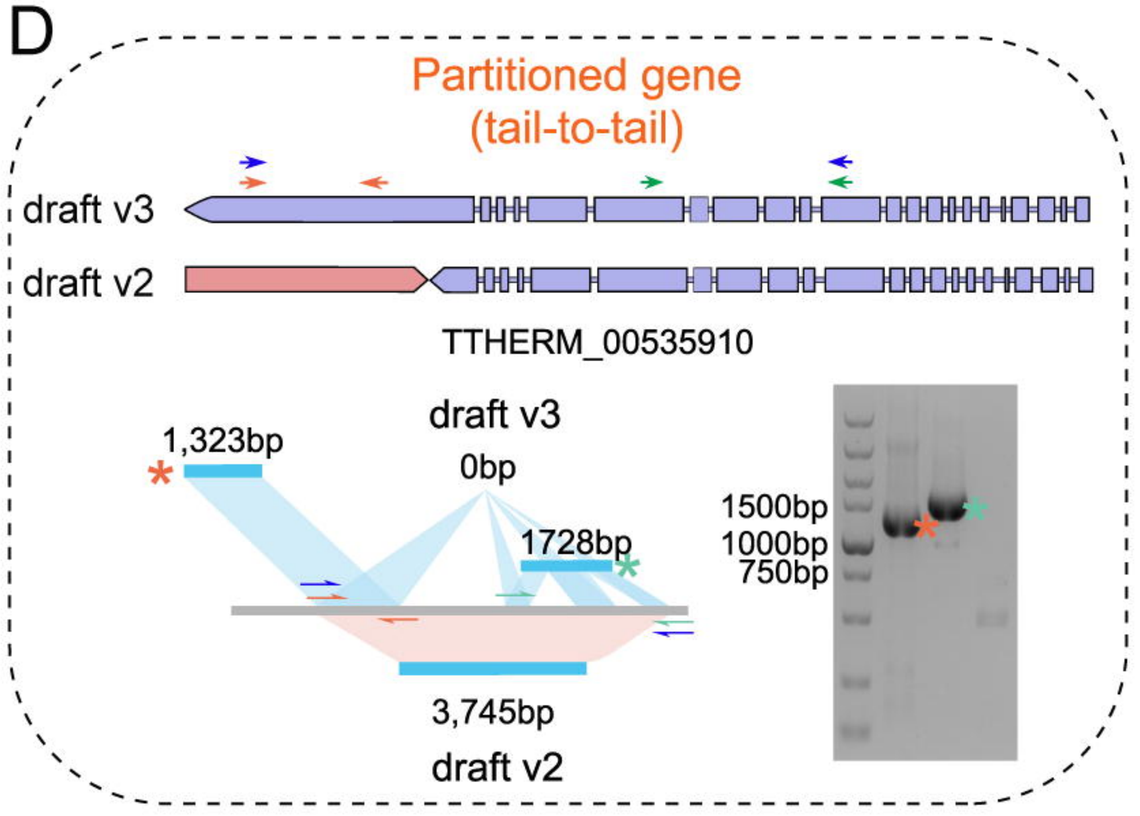


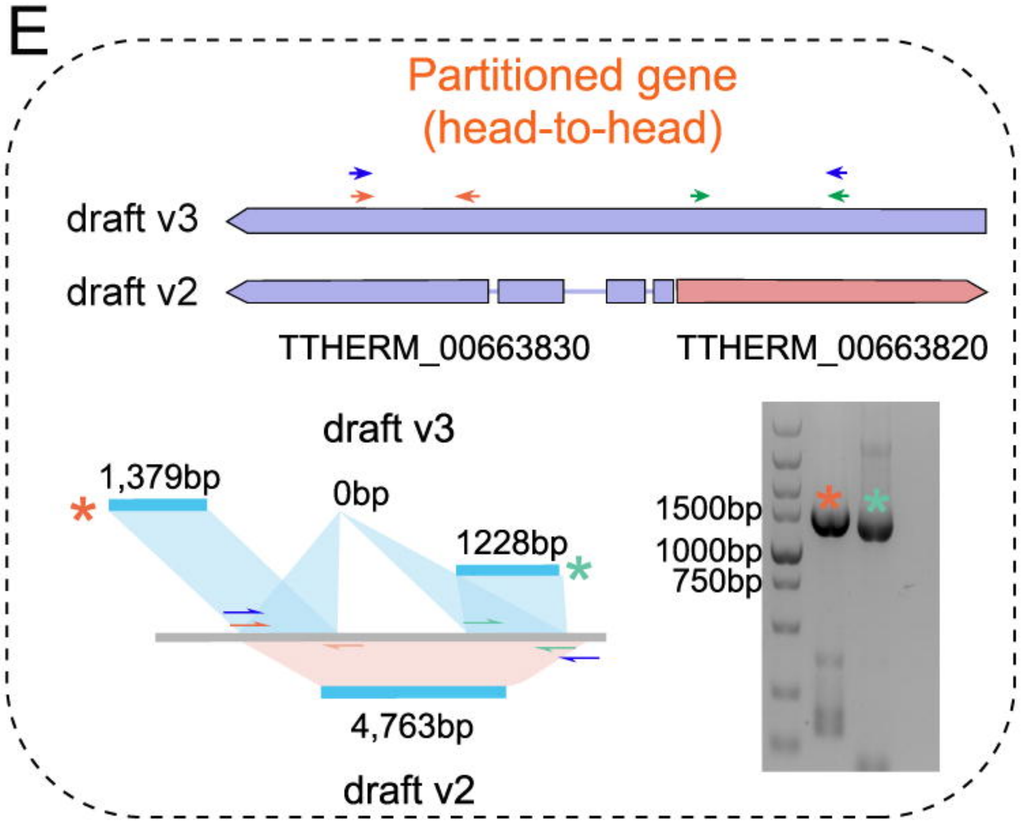

A
Homo sapiens
ATAC-seq
peak calling for NFR
peaks
composite analysis
RefTSS database
cap-seq
TSS
gene
290
270
250
230
-500bp TSS
1,000bp
200
150
100
50
0
B
Tetrahymena thermophila
ATAC-seq
peak calling
42,469 NFR
peaks
H3K4me3
H2A.Z
nucleosome
6mA
modelling
RF model
TSS prediction
28,640 genes
draft v2
27,840 TSSs
TSS-region
25,617 pTSSs
21,095 eTSSs
21,095 genes
eTSS
4,522 genes
pTSS
1,752 duplicate genes
no eTSS or pTSS
1,271 low-confidence genes
no eTSS or pTSS
C
Co-promoter gene pair: 885
chr_172:320,272-326,513
co-promoter
Open chromatin
H3K4me3
H2A.Z
6mA
Nucleosome
Nanopore DRS
draft v5
TTHERM_00683400
TTHERM_00683401
D
Multi-short-exon gene: 1,174
chr_168:293,319-297,844
TTHERM_000648733
E
Tandem duplicate gene: 578
chr_019:1,026,318-1,035,289
TTHERM_00908090
TTHERM_00908100
TTHERM_00908110
F
Low-confidence gene: 1,271
chr_019:1,009,894-1,021,717
TTHERM_00912120

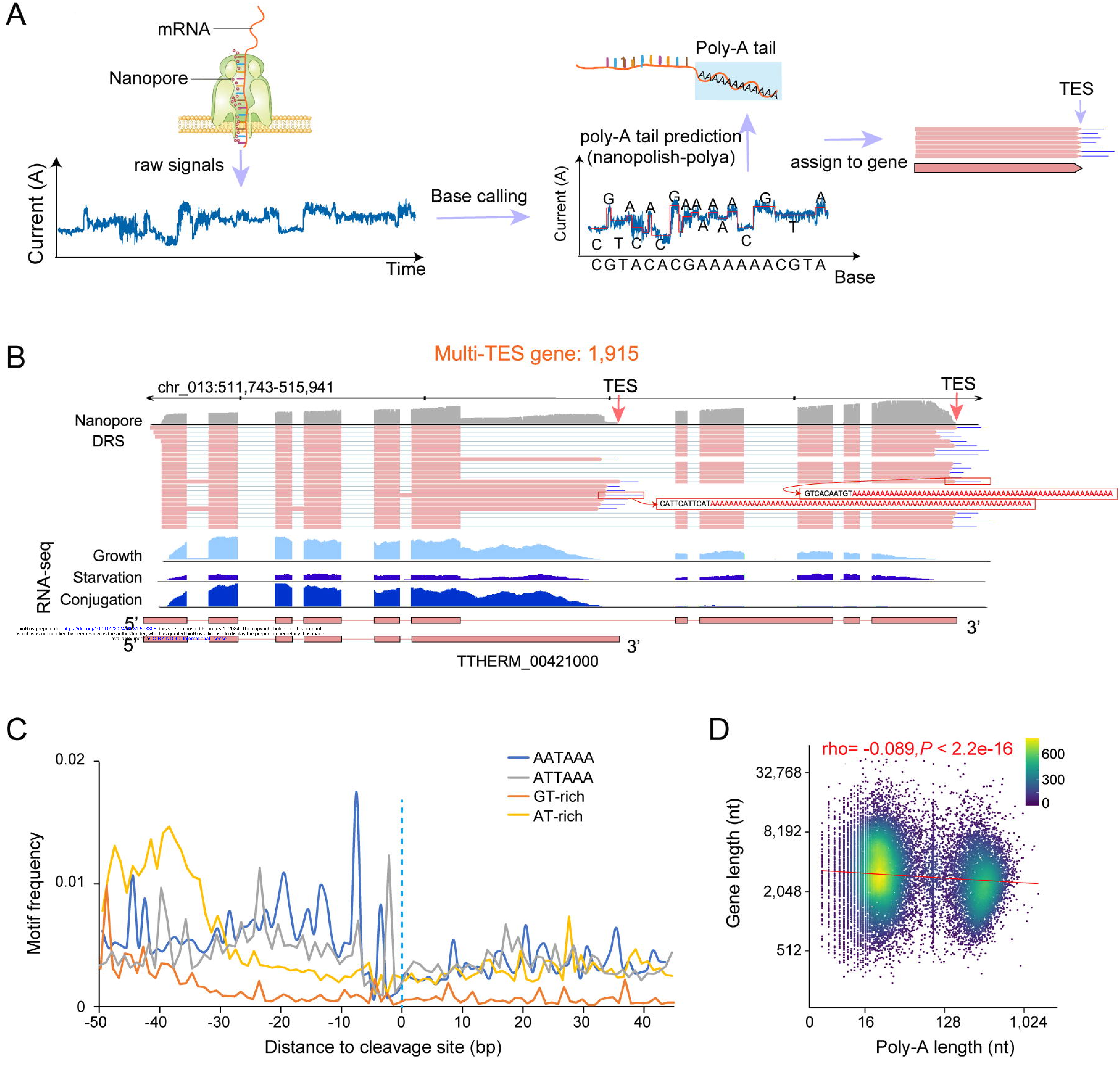




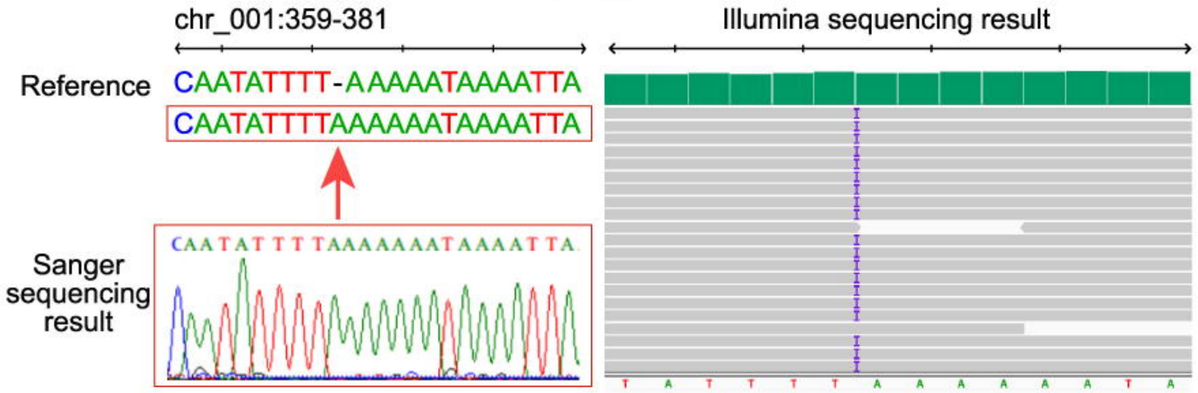




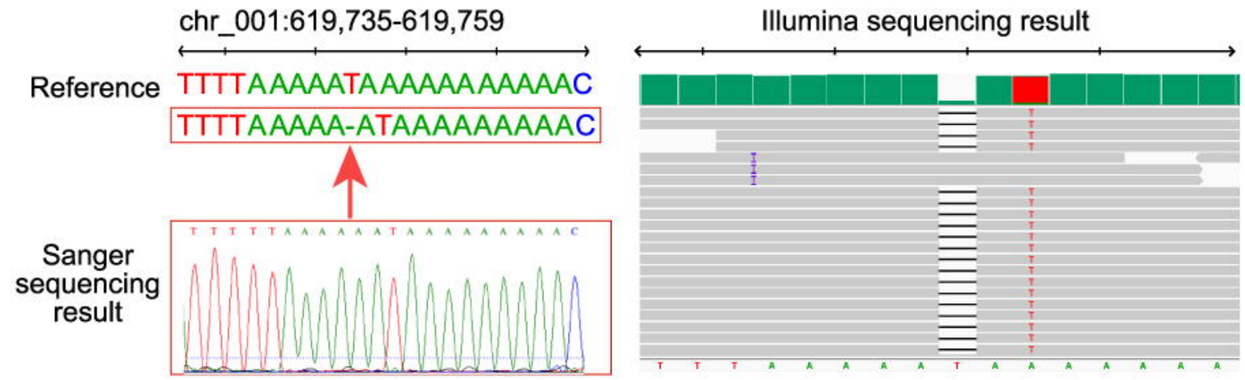




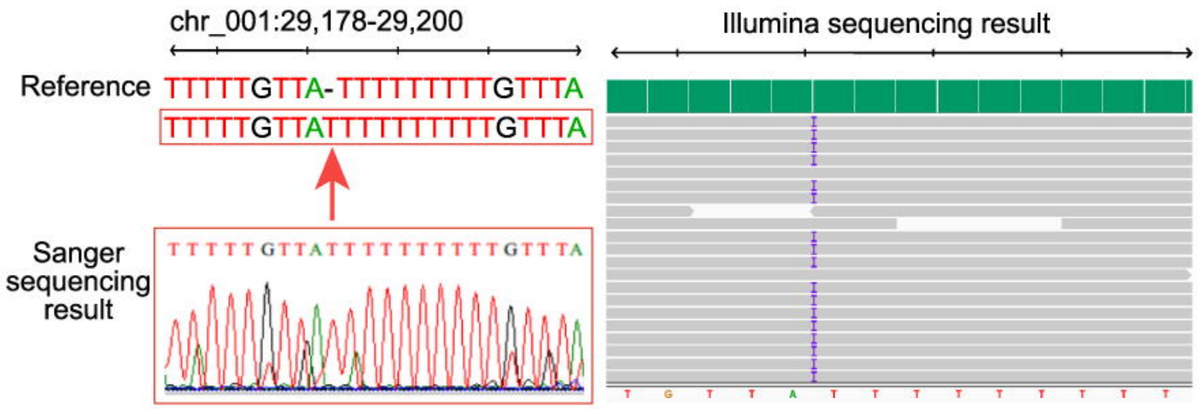
Insertion-1



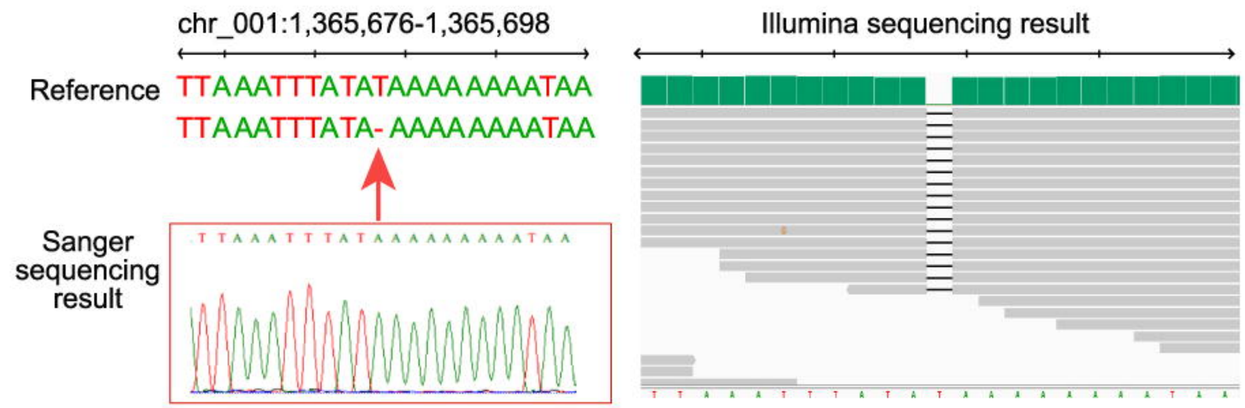
Deletion-1



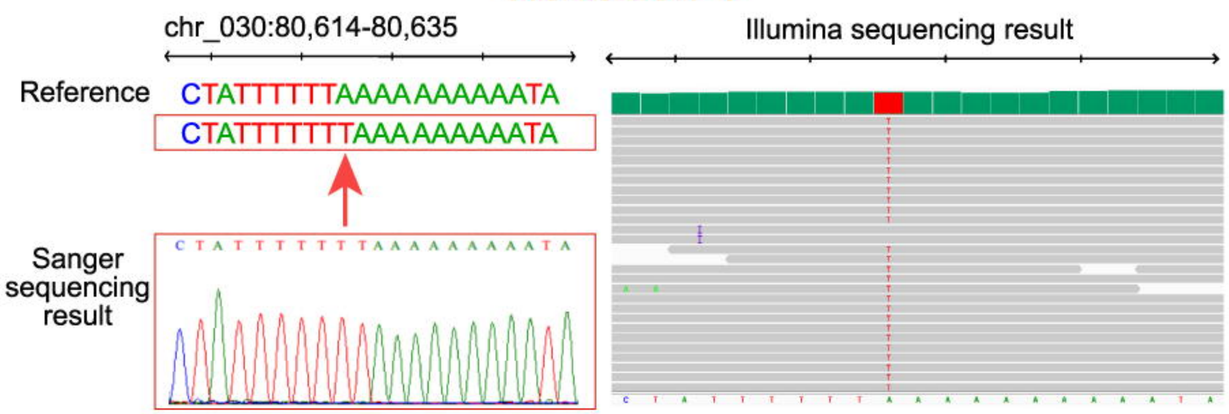
Insertion-2



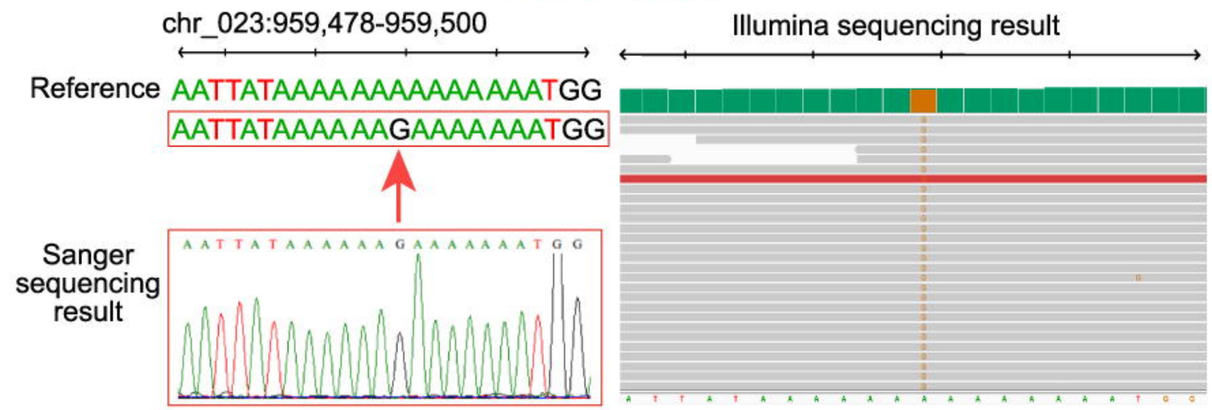
Deletion-2



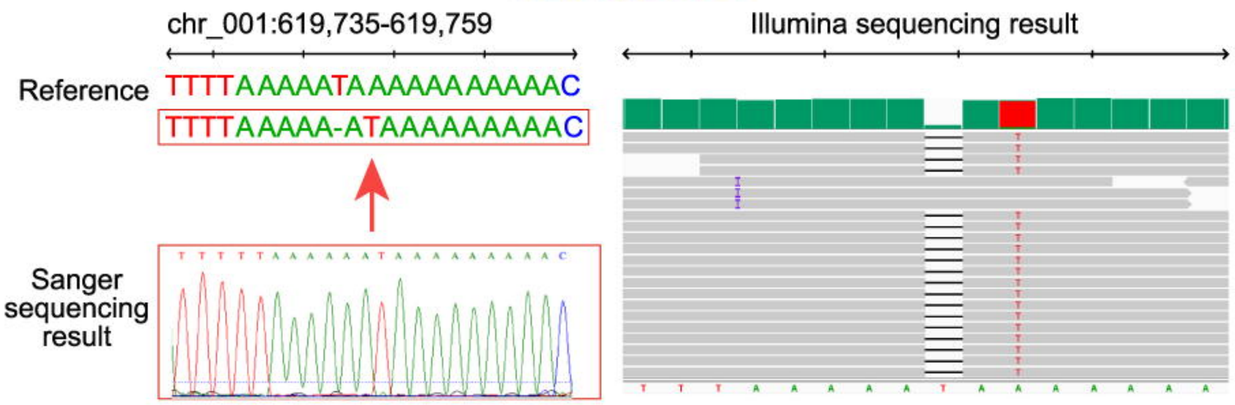
Transition-1



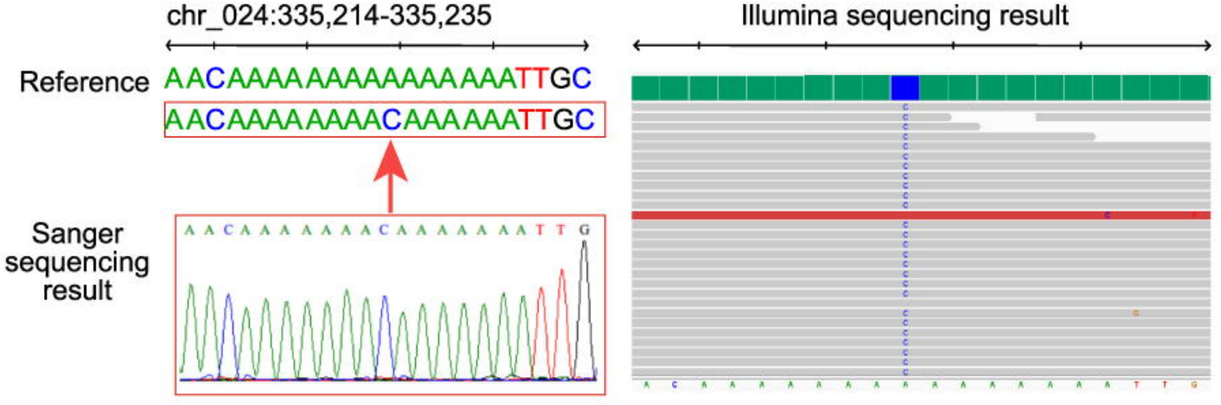
Transversion-1

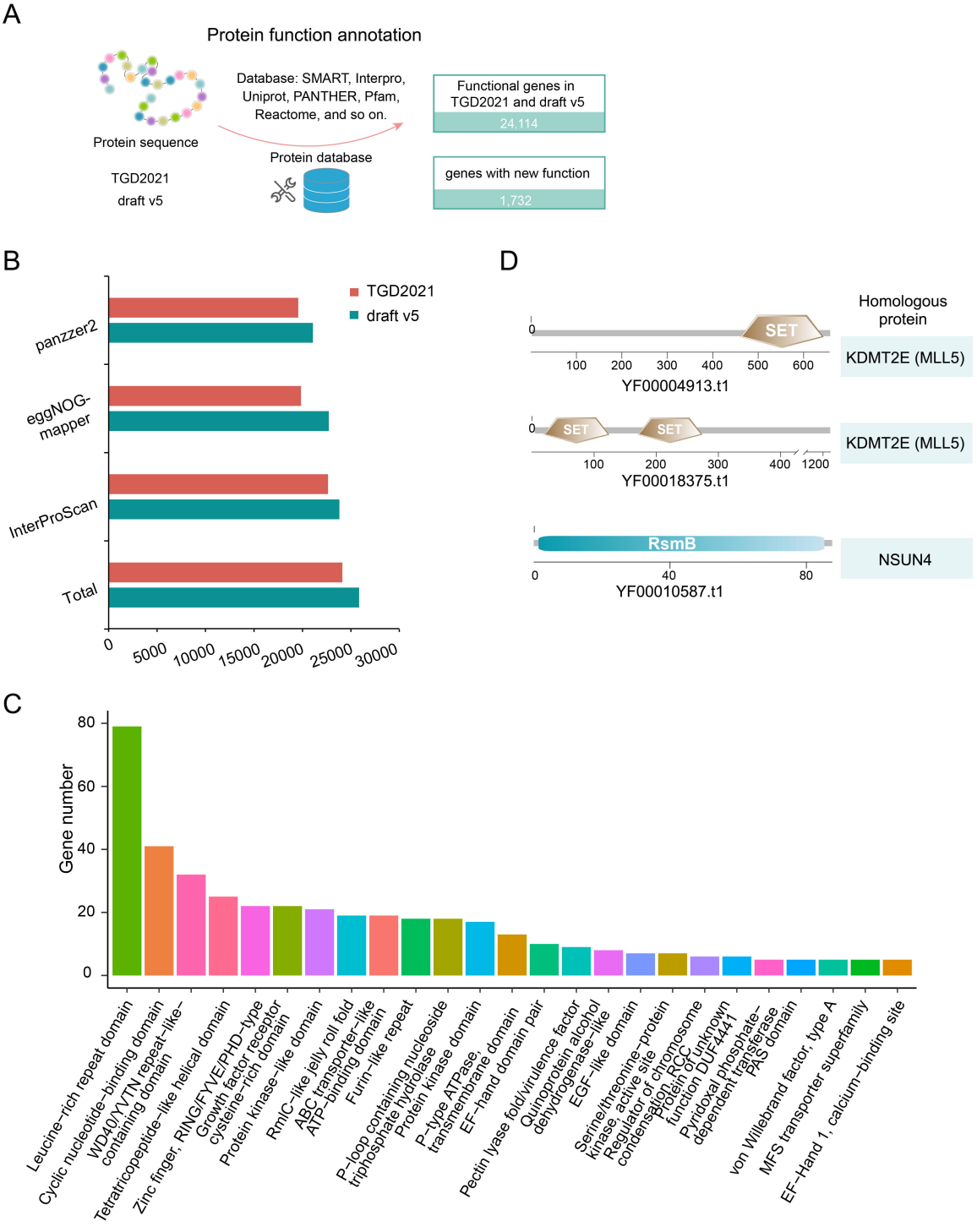


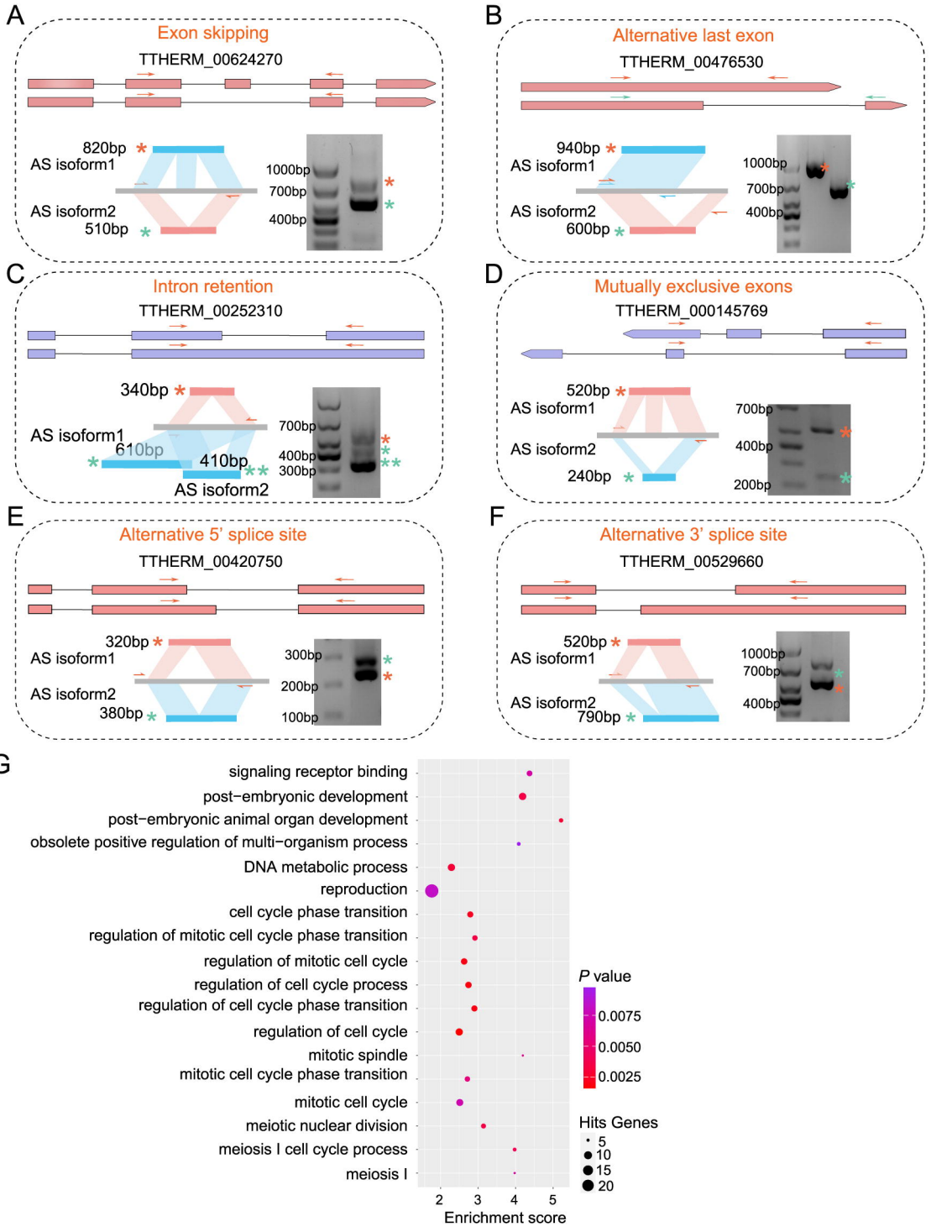
Transition-2

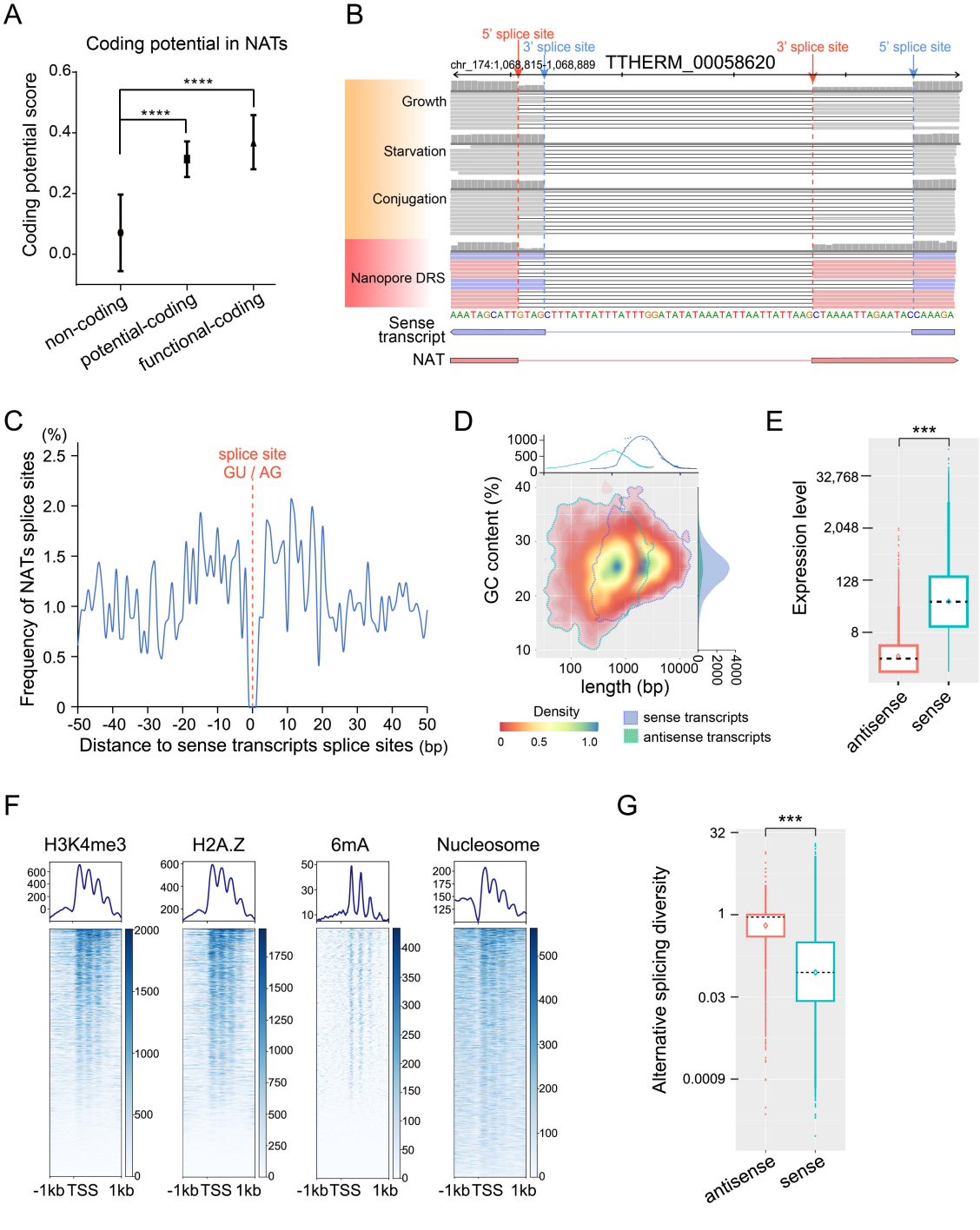


Transversion-2




A
Protein function annotation
Protein sequence
TGD2021
draft v5
Database: SMART, Interpro, Uniprot, PANTHER, Pfam, Reactome, and so on.
Protein database
Functional genes in TGD2021 and draft v5
24,114
genes with new function
1,732
B
TGD2021
draft v5
panzzer2
eggNOG-mapper
InterProScan
Total
0
5000
10000
15000
20000
25000
30000
D
Homologous protein
SET
YF00004913.t1
KDMT2E (MLL5)
SET
SET
YF00018375.t1
KDMT2E (MLL5)
RsmB
YF00010587.t1
NSUN4
C
Gene number
80
60
40
20
0
Leucine-rich repeat domain
Cyclic nucleotide-binding domain
WD40/YVTN repeat-like-containing domain
Tetratricopeptide-like helical domain
Zinc finger, RING/FYVE/PHD-type
Growth factor receptor cysteine-rich domain
Protein kinase-like domain
RmlC-like jelly roll fold
ABC transporter-like
ATP-binding domain
Furin-like repeat
P-loop containing nucleoside triphosphate hydrolase
Protein kinase domain
P-type ATPase, transmembrane domain
EF-hand domain pair
Pectin lyase fold/virulence factor
Quinoprotein alcohol dehydrogenase-like
EGF-like domain
Serine/threonine-protein kinase, active site
Regulator of chromosome condensation, RCC
Protein of unknown function DUF4441
Pyridoxal phosphate-dependent transferase
PAS domain
von Willebrand factor, type A
MFS transporter superfamily
EF-Hand 1, calcium-binding site


A
Exon skipping
TTHERM_00624270
820bp
AS isoform1
AS isoform2
510bp
1000bp
700bp
400bp
B
Alternative last exon
TTHERM_00476530
940bp
AS isoform1
AS isoform2
600bp
1000bp
700bp
400bp
C
Intron retention
TTHERM_00252310
340bp
AS isoform1
610bp
410bp
AS isoform2
700bp
400bp
300bp
D
Mutually exclusive exons
TTHERM_000145769
520bp
AS isoform1
AS isoform2
240bp
700bp
400bp
200bp
E
Alternative 5' splice site
TTHERM_00420750
320bp
AS isoform1
AS isoform2
380bp
300bp
200bp
100bp
F
Alternative 3' splice site
TTHERM_00529660
520bp
AS isoform1
AS isoform2
790bp
1000bp
700bp
400bp
G
signaling receptor binding
post−embryonic development
post−embryonic animal organ development
obsolete positive regulation of multi−organism process
DNA metabolic process
reproduction
cell cycle phase transition
regulation of mitotic cell cycle phase transition
regulation of mitotic cell cycle
regulation of cell cycle process
regulation of cell cycle phase transition
regulation of cell cycle
mitotic spindle
mitotic cell cycle phase transition
mitotic cell cycle
meiotic nuclear division
meiosis I cell cycle process
meiosis I
2
3
4
5
Enrichment score
P value
0.0075
0.0050
0.0025
Hits Genes
5
10
15
20


A
Coding potential in NATs
Coding potential score
0.6
0.4
0.2
0.0
****
****
non-coding
potential-coding
functional-coding
B
5' splice site
3' splice site
3' splice site
5' splice site
chr_174:1,068,815-1,068,889
TTHERM_00058620
Growth
Starvation
Conjugation
Nanopore DRS
AAATAGCATTGTAGCTTTATTATTTATTTGGATATATAAATATTAATTATTAAGCTAAAATTAGAATACCAAAGA
Sense transcript
NAT
C
(%)
Frequency of NATs splice sites
splice site
GU / AG
2.5
2.0
1.5
1.0
0.5
0
-50 -40 -30 -20 -10 0 10 20 30 40 50
Distance to sense transcripts splice sites (bp)
D
1000
500
0
GC content (%)
40
30
20
10
100 1000 10000
length (bp)
0 2000 4000
Density
0 0.5 1.0
sense transcripts
antisense transcripts
E
***
Expression level
32,768
2,048
128
8
antisense
sense
F
H3K4me3
H2A.Z
6mA
Nucleosome
-1kb TSS 1kb
G
***
Alternative splicing diversity
32
1
0.03
0.0009
antisense
sense