Extensive RNA editing in tomato nuclear genes

Jia-Yu Xue[1,#], Hai-Yun Fan[1,#], Zhen Zeng[2,#], Yu-Han Zhou[1], Shuai-Ya Hu[1], Sai-Xi Li[2], Ying-Juan Chen[1], Fei Chen[3,4,*], Yves Van de Peer[1,5,6,*], Zhu-Qing Shao[2,*]

[1]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China

[2]State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, 210023, China

[3]College of Tropical Crops, Sanya Nanfan Research Institute, Hainan University, Haikou 570228, China

[4]Hainan Yazhou Bay Seed Laboratory, Sanya 572025, China

[5]Department of Plant Biotechnology and Bioinformatics, Ghent University, VIB-UGent Center for Plant Systems Biology, B-9052 Ghent, Belgium

[6]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa

[#]These authors contributed equally to this work.

*To whom correspondence should be addressed: zhuqingshao@nju.edu.cn (Z.Q.S), yvpee@psb.vib-ugent.be (Y.V.d.P.) or feichen@hainanu.edu.cn (F.C.).

Introduction

RNA editing (RE) refers to a means of posttranscriptional modifications to transcripts, leading to sequence discrepancy between genomic DNA and mature mRNAs. RE were reported to be extensive in squid, pig and fish nuclear genomes (Albertin *et al.*, 2015; Buchumenski *et al.*, 2021; Huang *et al.*, 2021), as well as in plant organellar genomes (Xue *et al.*, 2010; Dong, SS *et al.*, 2019; Lukes *et al.*, 2021). In plant mitochondrial genomes, RE mostly converts cytidines (C) to uridines (U) and occasionally U to C, and usually restores start/stop codons, removes internal stop codons and modifying amino acids (Sloan, 2017). While in animal nuclear genomes, RE mainly changes adenosine (A) to inosine (I), which is recognized as guanosine (G) by the translational machinery, and occurs overwhelmingly in non-coding regions. The RE levels in animals are more dynamic, and associated with different developmental (Wahlstedt *et al.*, 2009; Hwang *et al.*, 2016) and health conditions (Silberberg *et al.*, 2012; Han *et al.*, 2015; Jain *et al.*, 2018), or environmental changes (Garrett & Rosenthal, 2012; Buchumenski *et al.*, 2017). However, RE has been rarely reported in plant nuclear genomes. Is RE also a prevent phenomenon in plant nuclear genomes? What kind of RE occurs most frequently in plant nuclear genomes, A-to-I, C-to-U, or a different pattern? Is RE in plant nuclear genomes functional or evolutionarily meaningful? These questions remain poorly understood.

The microTom tomato has become an emerging model plant system for developmental biology of fruit development (Klee & Giovannoni, 2011; Azzi *et al.*, 2015), arbuscular mycorrhiza symbiosis (Chitarra *et al.*, 2016; Liao *et al.*, 2022) and RNA editing (Yang *et al.*, 2017; Yang *et al.*, 2020; Li *et al.*, 2022) as well, which, however, all focused on organellar RE and the regulators, whereas nuclear RE has not been touched. Therefore in this study, we assembled a high-quality of genome of microTom, together with large amounts of RNA-seq data sequenced in this study and collected from public databases (Li *et al.*, 2020; Zeng *et al.*, 2022), we performed comprehensive analyses of RE events in the microTom tomato nuclear genome across different organs/developmental stages/treatments.

Extensive RE events and biased editing pattern in the microTom tomato nuclear genome

Using a combination of 92.40 Gb Nanopore data and 54.38 Gb Illumina data, we obtained a microTom genome assembly of 799 Mb with 60 contigs (contig N50 =

41.37 Mb), which was further merged into 12 corresponding pseudo-chromosomes using the genome of Heinz as a reference. Protein-coding gene annotation of the microTom assembly captures 98.57 % of the embryophyta BUSCO (odb10) genes, with 97.89 % single copy genes and 0.68% of duplicates. These results indicate that our assembled microTom genome has reached a high standard of quality and completeness (**Table 1**).

With the assistance of three full-length transcriptome and 81 Illumina short read length RNA-seq datasets derived from different organs, under different developmental stages and/or treatments (See Methods and Materials) from this and previous studies (Li *et al.*, 2020; Zeng *et al.*, 2022), a total of 35,213 protein-coding genes were annotated (**Figure 1a**), with 31,891 genes (90.57%) receiving transcriptomic data support from at least one RNA-seq sample. The massive RNA-seq data helped to annotate 4,449 "new" genes that were previously "hiding" in the Heinz tomato genome (gene length coverage and DNA sequence identity > 90%), probably due to inadequate RNA-seq data in the Heinz cultivar. Additionally, a large number of alternatively spliced genes (14,873, 42.24%, **Figure 1b** and **1c**) and untranslated regions (UTRs, 25,180, 71.51%, **Figure 1d**) were recognized. This comprehensive annotation provides fundamental data for the identification of RE sites in microTom nuclear genes.

By comparing the microTom RNA-seq data to its genome, we detected massive RE events (37,984) between the parental genes and their transcripts (**Figure 2a** and **Supplementary Table S1**), and these RE events affected 5,873 genes in the genome. Among the 37,984 RE events, four transitional RE (C-to-U, U-to-C, G-to-A, and A-to-G) showed significantly higher frequencies than other eight types (**Figure 2b**), accounting for 65.90% of all RE events. This biased pattern of transition-outnumbering-transversion seems similar to the RE pattern in zebrafish (Buchumenski *et al.*, 2021).

Regulatory mechanism of RE in plants should be different from that in metazoan

As stated, RE in plant nuclear genomes is rarely reported, and the regulatory mechanism of nuclear RE has not been studied. Even for animals, in which nuclear RE has been discovered in multiple taxa, the RE levels can vary greatly among different taxa. For instance, very limited RE events were detected in the coding

regions of humans, *Drosophila*, mice, sea hares and nautiloids, but were more abundant in zebrafish and cephalopods, despite the fact that animals share a RE machinery medicated by the ADAR (adenosine deaminase that acts on RNA) protein, which was considered to be a metazoan innovation by duplication (Grice & Degnan, 2015). So ADAR specifically regulates A-to-I RNA editing in animals, whereas not in plants. This may explain the exceptionally highest frequency of A-to-I editing in zebrafish. Interestingly, except for the dominance of A-to-I editing, the editing pattern in zebrafish appears to largely agree with that in microTom. In plants, the regulatory complexes for RE in mitochondrial genomes comprise a number of gene families, e.g., pentatricopeptide repeat (PPR), RNA editing factor-interacting proteins (RIPs) / multiple organellar RNA editing factors (MORFs), and organelle RNA recognition motif-containing proteins (ORRM). By analyzing the transcriptomic data of one microTom ORRM mutant (SlORRM4) in a previous study (Yang *et al.*, 2017), we detected changes of nuclear RE levels, with previously 2,604 edited sites becoming no longer edited after knock-out and 1,924 previously unedited sites being edited after (**Supplementary Table S2**), suggesting plant organellar RE regulators may also participate in the RE regulation of nuclear genomes. Despite of possibly different regulatory proteins in plants and animals, these two eukaryotic lineages seem to share a similar pattern of RE - transitions outnumbering transversions. May be through a convergent evolution in mechanism? After all, transitions only involve modifications of (de)amination and (de)hydroxylation, and could be achieved at the original base, but transversions probably have to go through more complicated modifications or even base substitutions. Likewise, the transitions outnumbering transversions manner can be observed for genomic DNA mutations, but not known if they are regulated by the same machinery. The underlying mechanisms for RNA editing and DNA mutation both demand future explorations.

Conserved and conditional RE sites, and possible functional roles of RE

Our statistics also indicates that slightly more RE sites were detected in UTRs than in coding regions (**Figure 2a**). Some RE sites in UTRs may be functional or evolutionarily meaningful, as 50.57% (10,018/19,810) of them were detected in at least two replicates, In the coding regions, about 51.21% (9,307/18,174) of the RE sites were supported by at least two replicates, and 44.89% (8,159/18,174) cause changes of amino acids (**Figure 2a**), suggesting a potential functional role. The

majority of microTom RE sites showed differential organ/developmental stage/treatment-specificity (**Figure 2c**), e.g. 5,208 RE sites only occurred in one organ/developmental stage/treatment (**Supplementary Figure S1** and **Table S1**, detected in at least two replicates, hereafter); 606 RE sites were only detected in fruits, and 179 RE sites were specific to roots, suggesting organ-specificity; and 3,117 RE sites were only detected during arbuscular mycorrhiza symbiosis, suggesting their specific roles in plant-microbe interaction. Other than these conditional RE sites, 79 RE sites were detected present in all tissues and conditions. Five randomly selected RE sites that were detected in all samples were verified by PCR and Sanger sequencing of genomic DNA and cDNAs (**Figure 2e**), validating the RE phenomenon in microTom nuclear genome. The RE sites present in multiple tissues/developmental stages/treatments could be considered as "fixed mutations" at the RNA level, and may be signs of adaptation. For other conditional RE sites, they were supported by replicates and their differential patterns under different conditions seem to indicate their specific functional roles, possibly playing distinct roles in plant growth, organ/tissue differentiation and development, or response to stimuli.

Even experimentally verified, one could still argue whether RE events are biologically meaningful or merely transcriptional errors, because the observed ratio of nonsynonymous and synonymous editing sites does not show a significant difference (**Figure 2a**). However, quite a few examples have shown that synonymous substitution in mRNA sequences can affect the productivity or correctly folding of proteins by either changing the secondary structure or remodeling the codon usage and/or ordering (Cannarrozzi *et al.*, 2010; Buhr *et al.*, 2016). In contrast, a direct effect of nonsynonymous substitution is generating proteins with different amino acid sequences. To further testify whether the transcripts with RE events can be reflected by translated proteins, which may truly lead to functional alterations, we surveyed 12 public proteome data sets generated from microTom tissues, and a total of 1,153 potential RE sites were detected, with at least one mapped peptide fragment in the proteome datasets (**Figure 2d** and **Supplementary Table S3**). Among those, 424 showed amino acid changes in accordance with RE events, including 34 sites with both edited and unedited forms (**Figure 2d**), indicating the RNA level nucleotide alteration can extend sequence diversity of mRNAs and even functional innovation of translated proteins.

Evolutionary significance of RE

A trade-off hypothesis was previously proposed for RE and genome evolution: there should be a balance between the genome evolutionary rate and the degree of RE, and a slow rate could be compensated by RE to generate more sequence diversity, so the organism can evolve to adapt to environmental changes and survive (Liscovitch-Brauer *et al.*, 2017). This theory may explain the extensive RE in cephalopods, as it has a slow genomic evolutionary rate, while there are not adequate cases to verify this hypothesis in plants because of too few studies. From an evolutionary point of view, it makes sense that RE and alternative splicing can both conserve genetic material at the DNA level but create sequence diversity at the RNA level, which may be an evolutionary advantage when encountering survival dilemma and demanding quick responses.

Concluding remarks

To sum up, we conducted comprehensive identification of RE in a plant genome, and for the first time detected extensive RE events in the microTom, is supported by evidence from both experimental validation and proteomic data. The prevalence of RE and the biased pattern (transition-outnumbering-transversion) in microTom rival that in animals, and may imply an ancient and shared mechanism underlying this phenomenon across eukaryotes. We observed conserved RE sites that are always edited under all conditions, that could be considered as “fixed mutations” at the RNA level; also detected conditional RE sites occurring only in specific organ/developmental stage/treatment, possibly suggesting distinct functional roles. As a means of posttranscriptional modification, RE results in sequence diversity at the RNA level and may lead to even functional innovation of proteins, hence may be considered as an additional mechanism to expand gene diversity other than gene duplications and alternative splicing.

Our findings suggest diverse and complicated regulatory mechanisms at the RNA level. It could be anticipated that the identification of extensive REs from nuclear genes of more plants would provide new challenge and opportunity for genetic and molecular studies on many aspects.

Materials and methods

Plant material and data sources

Seedlings of the *Solanum lycopersicum* (cv. 'microTom') were grown in plastic pots filled with a mixture of sterilized sand/gravel (1:1 ratio) in a climate-controlled growth room with 16 h light at 24°C and 8 h dark at 22°C. Roots, stems, leaves, flower, immature green fruits, and mature red fruits were collected from plants grown in nutrient-rich soil for six weeks, and used for RNAome sequencing. Green and red fruits were classified by color chart "USDA Visual Aid TM-L-1" (USDA, 1975).

We also downloaded 75 Illumina RNA-seq datasets and two PacBio RNA-seq datasets involving different tissues/organs and different developmental stages and/or different treatments from our previous study (Zeng *et al.*, 2022), and a study performed by others (Li *et al.*, 2020). Twelve microTom tomato proteomic datasets were retrieved from the PRIDE Archive database (Proteomics identifications database, https://www.ebi.ac.uk/pride/archive/).

Library construction and sequencing

Fresh leaves of four-week old microTom that were grown in half-strength Murashige and Skoog basal medium with sucrose and phytagel were collected, and the genomic DNA was sequenced on the Nanopore platform using a MinION R9 flow cell, and the Illumina NovaSeq6000 platform with an insert size of 450bp at Benagen (Wuhan, China).

Fresh tissues of roots, stems, leaves, flower, immature green fruits and mature red fruits of plants grown in nutrient-rich soil were subjected to total RNA extraction using TRIzol (Invitrogen, Carlsbad, USA) according to the manufacturer's protocol, which were used for subsequent RNA sequencing using different strategies. Samples from each tissue were subjected to construct a ribo-minus RNA library and small RNA library for RNA sequencing, whereas a mixture of the six samples were used to construction a PacBio SMRT library and a circular RNA library. The ribo-minus RNA libraries, small RNA libraries and circular RNA library were sequenced using the Illumina HiSeq platform at Novogene Co., Ltd. (Tianjin, China), and the PacBio SMRT library was sequenced using the PacBio Sequel System at Novogene Co., Ltd.

(Tianjin, China).

Genome assembly and assessment

Firstly, the raw Nanopore reads were corrected and assembled into contigs by NextDenovo-v2.5.0 (https://github.com/Nextomics/NextDenovo) with default parameters. After that, Nanopore reads were further mapped into primary contigs by software Minimap2-v2.17 (Li, 2021) with the parameter '-ax map-ont', followed by Nextpolish-v1.4.1 (Hu *et al.*, 2019) to polish contigs. Meanwhile, the Illumina paired-end reads were processed to remove adaptor and low quality sequences using Trimmomatic-v0.38 (Bolger *et al.*, 2014). Then, the clean Illumina short reads were mapped to the polished contigs using BWA-MEM-v0.7.17 (Li & Durbin, 2009) with default parameters, based on three iterative rounds of polishing with parameters: '--fix all' by Pilon-v1.23 (Walker *et al.*, 2014) (read length N50 = 41.37 Mb). Finally, the polished contigs were aligned to the Heinz genome to form pseudomolecules, by the software RaGOO-v1.1 (Alonge *et al.*, 2019). Embryophyta BUSCO-v5.4.2 (odb10) (Manni *et al.*, 2021) was used to evaluate the integrity of the final assembly.

PacBio transcriptome processing

The SMRTlink-v6.0.0 pipeline was applied to PacBio *lso-seq* raw data. Firstly, the circular consensus sequence reads (CCSs) were extracted from subreads of BAM files with parameters: minLength= 300, minZScore= -999, minPasses= 1, maxDropFraction= 0.8, minPredictedAccuracy= 0.8, minSnr= 4, noPolish= TURE. Secondly, CCS reads were divided into full-length non-chimeric (FLNC) reads and non-full-length (nFL) reads, by identifying whether 5′ and 3′ adapters and the poly(A) tail were present. Then, we obtained consensus isoforms using the ICE algorithm from FLNC reads (Iterative Clustering for Error Correction), which were further polished with nFL reads to get high-quality consensus FLNC reads based on Quiver (parameters: hq_quiver_min_accuracy 0.99, others with default parameters). Finally, the nonredundant FLNC reads were obtained by CD-HIT-v4.8.1 (Fu *et al.*, 2012) with an identity cutoff of 0.95. We also used the software LoRDEC-v0.9 (Salmela &

Rivals, 2014) to correct mismatch and nucleotide indels in FLNC reads.

Repetitive DNA identification and Genome annotation

We identified the repetitive elements in the microTom genome based on a combination of *de novo* and homology-based strategies. A *de novo* library was first built through RepeatModeler-v2.0.1 (Bao & Eddy, 2002) and LTR_retriever-v2.9.0 (Ou & Jiang, 2018). The repeat region in the genome was further masked by scanning *de novo* repeat library. Next, we used RepeatMasker-v4.0.9 (Tarailo-Graovac & Chen, 2009) to classify the repeat sequences using the parameters: -poly -html -gff -lib -no_is -xsmall.

The combination of *ab initio*, homology-based and RNA-Seq-based annotation strategy was applied to predict and annotate the protein coding genes. Firstly, the masked genome was trained by BRAKER2-v2.1.5 (Brůna *et al.*, 2021) and GeneMark-ET-v4.69 (Lomsadze *et al.*, 2014) for *ab initio* gene prediction. Secondly, for finding homologous genes, related species like *Solanum tuberosum* (686_v6.1) and *Solanum lycopersicum* Heinz 1706 (691_ITAG4.0), which downloaded from Phytozome v13, were loaded into GenomeThreader-v1.7.1 (Gremme *et al.*, 2005) with the parameters: -species arabidopsis -gff3out -intermediate. Thirdly, 81 RNA-Seq (**Supplementary Figure S2 and Table S4**) data sets were used to assist the annotation. These were evaluated by FastQC-v0.11.9 and Trimmomatic-v0.38 (Bolger *et al.*, 2014) to trim low-quality and adaptor sequences. Then, these high-quality reads were aligned to the genome by HISAT2-v2.2.1 (Kim *et al.*, 2019). We assembled transcripts with StringTie-v2.1.5 (Pertea *et al.*, 2015) and TransDecoder-v5.5.0 (https://github.com/TransDecoder/TransDecoder). The PacBio *lso-seq* was also applied to forecast complete coding sequences (CDS) by PASA-v2.5.2 pipeline (Haas *et al.*, 2003). Finally, the EVidenceModeler-v1.1.1 pipeline was applied to integrate the above-mentioned prediction strategies (Haas *et al.*, 2008). The GTF annotation file, assembled from the Illumina and PacBio data, was used for identifying AS events using the AStalavista v3.2 tool (Foissac & Sammeth, 2007).

Identification and verification of RNA editing events

Clean RNA-seq reads from 81 transcriptome data sets and genomic DNA sequencing reads generated by Illumina platform were mapped onto the microTom genome assembly separately using HISAT2-v2.2.1 (Kim *et al.*, 2019) and Bowtie2-v2.4.5 (Langmead & Salzberg, 2012) with default parameters. After sorting the bam format files, we employed Picard-v2.26.0 (https://broadinstitute.github.io/picard/) to mark the duplicated bam files. The vcf files were generated by SAMtools-v1.13 (Danecek *et al.*, 2021) and BCFtools-v1.13 (Danecek *et al.*, 2021) with parameter settings: -m -v -V indels. Through a custom perl script to process above vcf files (Dong, S *et al.*, 2019), SNP sites (potential RNA editing sites) were called. We only retained sites that were supported by at least 10 reads and had a mutation proportion (representing a frequency of RNA editing at a certain site) of at least 20%. A custom python script was used to confirm the position of RNA editing events in the microTom genome, and to determine whether a site is in the CDS or UTR region. We also removed genomic SNPs from RNA editing sites to reduce false positives. We constructed a nonredundant dataset from 25 samples with RNA editing sites, from which the sites supported by at least two replicate experiments in one sample were recorded.

To validate the detected RNA editing events, PCR amplification and Sanger sequencing were performed on five randomly selected sites with specific primers (**Supplementary Table S5**). Genomic DNA and total RNA of microTom tomato roots were extracted using CTAB (Beyotime, Shanghai, China) and TRIzol (Invitrogen, Carlsbad, USA), respectively. cDNA synthesis was performed using HiScript III 1st Strand cDNA Synthesis Kit (Vazyme, Nanjing, China). PCR was carried out using designed primers (**Supplementary Table S5**) and I-5™ 2×High-Fidelity Master Mix (MCLAB, South San Francisco, USA). Sanger sequencing was performed to further confirm the edited RNA bases at Sangon Biotech (Nanjing, China). Moreover, peptides of microTom tomato proteomic data from the public database (**Supplementary Table S6**) were mapped to protein sequences of genes with identified RNA editing sites to further validate the RNA editing events (Perez-Riverol *et al.*, 2022).

Data availability

The data supporting the findings of this work are available within the paper and its Supplementary files. All sequencing data generated have been deposited in National Genomics Data Center (https://bigd.big.ac.cn/) with BioProject ID: PRJCA012325 under GSA: CRA008482. The genome assembly, along with the gene models, can be accessed under GWH: WGS029763 and microTomBase (http://eplant.njau.edu.cn/microTomBase).

Author contributions

J.Y.X., Z.Q.S., F.C. and Y.V.d.P. conceived the study. Z.Q.S., Y.J.C. and Z.Z. collected plant samples. H.Y.F. conducted genome assembly and annotation. H.Y.F., Z.Z., S.Y.H., Y.H.Z.and S.X.L conducted analyses of RNA-seq data. J.Y.X. and Z.Q.S. drafted the manuscript. All authors read the manuscript and participated in the revision of the manuscript. All authors approved the final manuscript.

Acknowledgments

This work was supported by grants from the Fundamental Research Funds for the Central Universities (No. KYCXJC2022003), the National Natural Science Foundation of China (32070243) and the Outstanding Young Teacher of “QingLan Project” of Jiangsu Province. Y.V.d.P. acknowledges funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01). The computational resources and services were provided by the Bioinformatics Center of Nanjing Agricultural University.

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supporting information

Supplementary Note S1. Detailed materials and methods used in this study.

Supplementary Fig. S1. The number of shared and unique RE sites in 79 samples (25 tissues) were counted.

Supplementary Fig. S2. Information about transcriptome datasets and sequencing strategies.

Supplementary Table S1. A nonredundant dataset of RNA editing sites from 81 samples data.

Supplementary Table S2. A nonredundant dataset of RNA editing sites from 6 samples data including 3 repeat experiments which knock out SlORRM4.

Supplementary Table S3. Mapped peptide number of potential 1,153 RNA editing sites.

Supplementary Table S4. Summary of transcriptome mapping for microTom tissues.

Supplementary Table S5. Primers used for verification of RE

Supplementary Table S6. The 12 microTom tomato proteome used in this study.

References

Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner S, Ragsdale CW, Rokhsar DS. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524**(7564): 220-+.

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology* **20**(1): 224.

Azzi L, Deluche C, Gevaudant F, Frangne N, Delmas F, Hernould M, Chevalier C. 2015. Fruit growth-related genes in tomato. *J Exp Bot* **66**(4): 1075-1086.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research* **12**(8): 1269-1276.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15): 2114-2120.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics* **3**(1).

Buchumenski I, Bartok O, Ashwal-Fluss R, Pandey V, Porath HT, Levanon EY, Kadener S. 2017. Dynamic hyper-editing underlies temperature adaptation in Drosophila. *Plos Genetics* **13**(7).

Buchumenski I, Holler K, Appelbaum L, Eisenberg E, Junker JP, Levanon EY. 2021. Systematic identification of A-to-I RNA editing in zebrafish development and adult organs. *Nucleic Acids Research* **49**(8): 4325-4337.

Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H, Rodnina MV, Komar AA. 2016. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Molecular Cell* **61**(3): 341-351.

Cannarrozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, Gonnet P, Gonnet G, Barral Y. 2010. A Role for Codon Order in Translation Dynamics. *Cell* **141**(2): 355-367.

Chitarra W, Pagliarani C, Maserti B, Lumini E, Siciliano I, Cascone P, Schubert A, Gambino G, Balestrini R, Guerrieri E. 2016. Insights on the Impact of Arbuscular Mycorrhizal Symbiosis on Tomato Tolerance to Water Stress. *Plant Physiol* **171**(2): 1009-1023.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10**(2).

Dong S, Zhao C, Zhang S, Wu H, Mu W, Wei T, Li N, Wan T, Liu H, Cui J, et al. 2019. The Amount of RNA Editing Sites in Liverwort Organellar Genes Is Correlated with GC Content and Nuclear PPR Protein Diversity. *Genome Biol Evol* **11**(11): 3233-3239.

Dong SS, Zhao CX, Zhang SZ, Wu H, Mu WX, Wei T, Li N, Wan T, Liu H, Cui J, et al. 2019. The Amount of RNA Editing Sites in Liverwort Organellar Genes Is Correlatedwith GC Content and Nuclear PPR Protein Diversity. *Genome Biology and Evolution* **11**(11): 3233-3239.

Foissac S, Sammeth M. 2007. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res* **35**(Web Server issue): W297-299.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *BIOINFORMATICS* **28**(23): 3150-3152.

Garrett S, Rosenthal JJC. 2012. RNA Editing Underlies Temperature Adaptation in K+ Channels from Polar Octopuses. *Science* **335**(6070): 848-851.

Gremme G, Brendel V, Sparks ME, Kurtz S. 2005. Engineering a software tool for gene structure

398 prediction in higher organisms. *Information and Software Technology* **47**(15): 965-978.
399 **Grice LF, Degnan BM. 2015.** The origin of the ADAR gene family and animal RNA editing. *Bmc*
400 *Evolutionary Biology* **15**.
401 **Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB,**
402 **Town CD, et al. 2003.** Improving the Arabidopsis genome annotation using maximal
403 transcript alignment assemblies. *NUCLEIC ACIDS RESEARCH* **31**(19): 5654-5666.
404 **Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008.**
405 Automated eukaryotic gene structure annotation using EVidenceModeler and the program to
406 assemble spliced alignments. *Genome Biology* **9**(1).
407 **Han L, Diao LX, Yu SX, Xu XY, Li J, Zhang R, Yang Y, Werner HMJ, Eterovic AK, Yuan Y, et al. 2015.** The
408 Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer*
409 *Cell* **28**(4).
410 **Hu J, Fan J, Sun Z, Liu S. 2019.** NextPolish: a fast and efficient genome polishing tool for long-read
411 assembly. *Bioinformatics* **36**(7): 2253-2255.
412 **Huang JR, Lin L, Dong ZY, Yang L, Zheng TY, Gu WW, Zhang Y, Yin TL, Sjostedt E, Mulder J, et al. 2021.**
413 A porcine brain-wide RNA editing landscape. *Communications Biology* **4**(1).
414 **Hwang T, Park CK, Leung AKL, Gao Y, Hyde TM, Kleinman JE, Rajpurohit A, Tao R, Shin JH,**
415 **Weinberger DR. 2016.** Dynamic regulation of RNA editing in human brain development and
416 disease. *Nature Neuroscience* **19**(8): 1093-+.
417 **Jain M, Mann TD, Stulic M, Rao SP, Kirsch A, Pullirsch D, Strobl X, Rath C, Reissig L, Moreth K, et al.**
418 **2018.** RNA editing of Filamin A pre-mRNA regulates vascular contraction and diastolic blood
419 pressure. *Embo Journal* **37**(19).
420 **Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019.** Graph-based genome alignment and
421 genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**(8): 907-915.
422 **Klee HJ, Giovannoni JJ. 2011.** Genetics and control of tomato fruit ripening and quality attributes.
423 *Annu Rev Genet* **45**: 41-59.
424 **Langmead B, Salzberg SL. 2012.** Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**(4):
425 357-359.
426 **Li H. 2021.** New strategies to improve minimap2 alignment accuracy. *BIOINFORMATICS* **37**(23):
427 4572-4574.
428 **Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows-Wheeler transform.
429 *BIOINFORMATICS* **25**(14): 1754-1760.
430 **Li J, Wang K, Yang Y, Lu Y, Cui K, Ji Y, Ma L, Cheng K, Ostersetzer-Biran O, Li F, et al. 2022.** SlRIP1b is a
431 global organellar RNA editing factor, required for normal fruit development in tomato plants.
432 *New Phytologist*.
433 **Li Y, Chen Y, Zhou L, You S, Deng H, Chen Y, Alseekh S, Yuan Y, Fu R, Zhang Z, et al. 2020.** MicroTom
434 Metabolic Network: Rewiring Tomato Metabolic Regulatory Network throughout the Growth
435 Cycle. *Mol Plant* **13**(8): 1203-1218.
436 **Liao D, Sun C, Liang H, Wang Y, Bian X, Dong C, Niu X, Yang M, Xu G, Chen A, et al. 2022.**
437 SlSPX1-SlPHR complexes mediate the suppression of arbuscular mycorrhizal symbiosis by
438 phosphate repletion in tomato. *Plant Cell*.
439 **Liscovitch-Brauer N, Alon S, Porath HT, Elstein B, Unger R, Ziv T, Admon A, Levanon EY, Rosenthal JJC,**
440 **Eisenberg E. 2017.** Trade-off between Transcriptome Plasticity and Genome Evolution in
441 Cephalopods. *Cell* **169**(2): 191-202.

442 **Lomsadze A, Burns PD, Borodovsky M. 2014.** Integration of mapped RNA-Seq reads into automatic
443 training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**(15): e119.
444 **Lukes J, Kaur B, Speijer D. 2021.** RNA Editing in Mitochondria and Plastids: Weird and Widespread.
445 *Trends in Genetics* **37**(2): 99-102.
446 **Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021.** BUSCO: Assessing Genomic Data Quality and
447 Beyond. *Curr Protoc* **1**(12): e323.
448 **Ou S, Jiang N. 2018.** LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long
449 Terminal Repeat Retrotransposons. *Plant Physiol* **176**(2): 1410-1422.
450 **Perez-Riverol Y, Bai JW, Bandla C, Garcia-Seisdedos D, Hewapathirana S, Kamatchinathan S, Kundu**
451 **DJ, Prakash A, Frericks-Zipper A, Eisenacher M, et al. 2022.** The PRIDE database resources in
452 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*
453 **50**(D1): D543-D552.
454 **Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015.** StringTie enables
455 improved reconstruction of a transcriptome from RNA-seq reads. *NATURE BIOTECHNOLOGY*
456 **33**(3): 290-+.
457 **Salmela L, Rivals E. 2014.** LoRDEC: accurate and efficient long read error correction. *Bioinformatics*
458 **30**(24): 3506-3514.
459 **Silberberg G, Lundin D, Navon R, Ohman M. 2012.** Deregulation of the A-to-I RNA editing mechanism
460 in psychiatric disorders. *Human Molecular Genetics* **21**(2): 311-321.
461 **Sloan DB. 2017.** Nuclear and mitochondrial RNA editing systems have opposite effects on protein
462 diversity. *Biology Letters* **13**(8).
463 **Tarailo-Graovac M, Chen N. 2009.** Using RepeatMasker to identify repetitive elements in genomic
464 sequences. *Curr Protoc Bioinformatics* **Chapter 4**: Unit 4.10.
465 **USDA. 1975.** Color Classification Requirements in Tomatoes. *Visual Aid TM-L-1*
466 **(Lansing,MI:TheJohnHenryCompany)**.
467 **Wahlstedt H, Daniel C, Enstero M, Ohman M. 2009.** Large-scale mRNA sequencing determines global
468 regulation of RNA editing during brain development. *Genome Research* **19**(6): 978-986.
469 **Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J,**
470 **Young SK, et al. 2014.** Pilon: an integrated tool for comprehensive microbial variant detection
471 and genome assembly improvement. *PloS one* DOI: 10.1371/journal.pone.0112963
472 **Xue JY, Liu Y, Li LB, Wang B, Qiu YL. 2010.** The complete mitochondrial genome sequence of the
473 hornwort Phaeoceros laevis: retention of many ancient pseudogenes and conservative
474 evolution of mitochondrial genomes in hornworts. *Current Genetics* **56**(1): 53-61.
475 **Yang YF, Liu XY, Wang KR, Li JY, Zhu GN, Ren S, Deng ZP, Zhu BZ, Fu DQ, Qu GQ, et al. 2020.** Molecular
476 and functional diversity of organelle RNA editing mediated by RNA recognition
477 motif-containing protein ORRM4 in tomato. *New Phytologist* **228**(2): 570-585.
478 **Yang YF, Zhu GN, Li R, Yan SJ, Fu DQ, Zhu BZ, Tian HQ, Luo YB, Zhu HL. 2017.** The RNA Editing Factor
479 SlORRM4 Is Required for Normal Fruit Ripening in Tomato. *Plant Physiology* **175**(4):
480 1690-1702.
481 **Zeng Z, Liu Y, Feng XY, Li SX, Jiang XM, Chen JQ, Shao ZQ. 2022.** The RNAome landscape of tomato
482 during arbuscular mycorrhizal symbiosis reveals an evolving RNA layer symbiotic regulatory
483 network. *Plant Commun*: 100429.
484

Table 1. Statistics and comparison of microTom and Heinz genomes

	MicroTom				Heinz 1706 (SL4.0)			
	Contig		Pseudomolecule		Contig		Pseudomolecule	
	Size (bp)	Number	Size (bp)	Number	Size (bp)	Number	Size (bp)	Number
N50	41,367,923	8	66,788,879	6	6,007,830	37	65,269,487	6
Maximum length	68,578,620		94,893,795		26,291,688		90,863,682	
Total number	60		12		448		12	
Number of gaps	47				435			
Assembled length	798,935,606				782,475,302			
BUSCO completeness	98.57%				97.40%			
Number of gene models	35,213				34,690			

Figure legends

Figure 1. **Genome annotation of microTom and comparison with Heinz and other species.**

(a) Protein-coding gene numbers for Heinz and microTom varieties. Shared one-to-one orthologs in microTom and Heinz genomes are presented in the cross section of the two tomato cartoon pictures. **(b)** Proportion of genes with multiple transcripts and genes with different transcript numbers. **(c)** Categories and frequency of AS events in microTom. IR, intron retention; AA, alternative adaptor; AD, alternative donor; ES, exon skipping. **(d)** Proportion of genes with annotated UTRs in microTom (MT), Heinz (HZ), *Arabidopsis thaliana* (AT), and *Oryza sativa* (OS).

Figure 2. **RNA editing in the microTom tomato genome.**

(a) Affected regions and detailed classifications of RNA editing events. **(b)** Frequency of 12 types of RNA editing events. Transition (orange): C-to-U, U-to-C, G-to-A, A-to-G; transversion (red): A-to-C, A-to-U, C-to-A, C-to-G, G-to-U, G-to-C, U-to-A, U-to-G. **(c)** Occurring regions (coding/noncoding), types (synonymous/nonsynonymous) and frequency of RNA editing events in different samples. **(d)** 424 out of 1,153 nonsynonymous RNA editing sites causing amino acid sequence changes of translated proteins were detected from 12 proteomes. **(e)** Random-selected cross-sample conserved RNA editing sites examined by experiments.

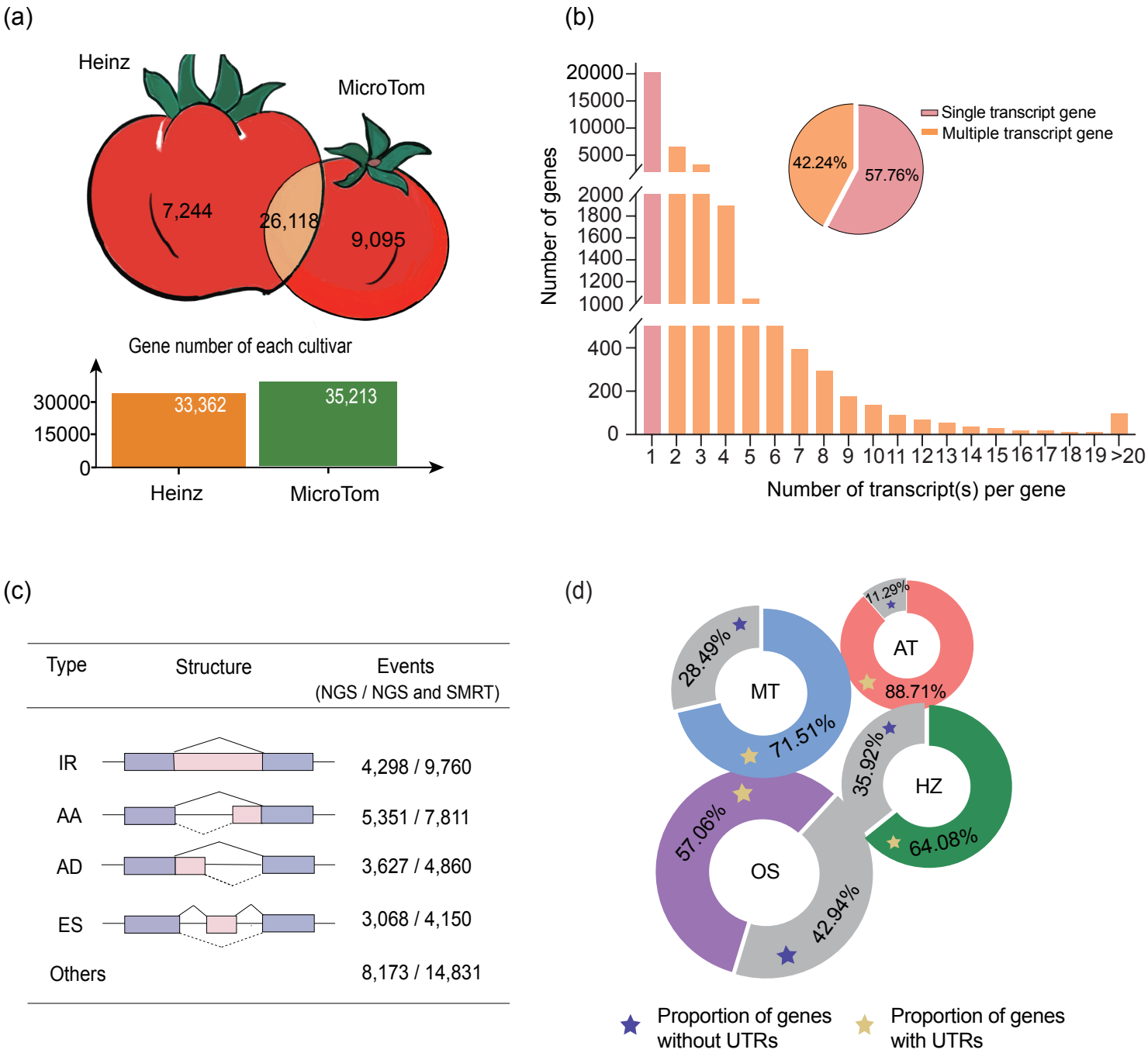

Figure 1. Genome annotation of microTom and comparison with Heinz and other species. (a) Protein-coding gene numbers for Heinz and microTom varieties. Shared one-to-one orthologs in microTom and Heinz genomes are presented in the cross section of the two tomato cartoon pictures. (b) Proportion of genes with multiple transcripts and genes with different transcript numbers. (c) Categories and frequency of AS events in microTom. IR, intron retention; AA, alternative adaptor; AD, alternative donor; ES, exon skipping. (d) Proportion of genes with annotated UTRs in microTom (MT), Heinz (HZ), Arabidopsis thaliana (AT), and Oryza sativa (OS).

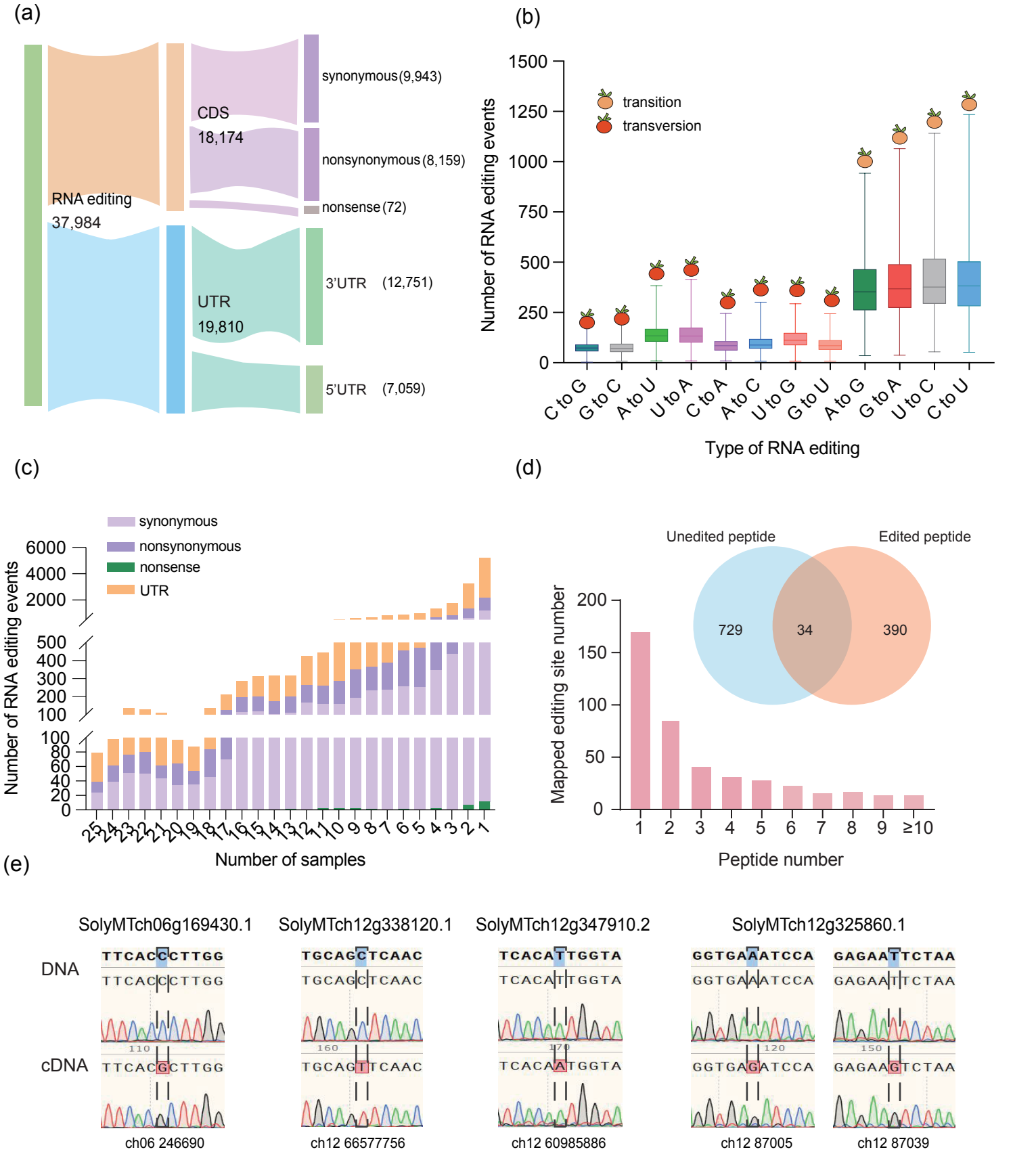

Figure 2. RNA editing in the microTom tomato genome. (a) Affected regions and detailed classifications of RNA editing events. (b) Frequency of 12 types of RNA editing events. Transition (orange): C-to-U, U-to-C, G-to-A, A-to-G; transversion (red): A-to-C, A-to-U, C-to-A, C-to-G, G-to-U, G-to-C, U-to-A, U-to-G. (c) Occurring regions (coding/noncoding), types (synonymous/nonsynonymous) and frequency of RNA editing events in different samples. (d) 424 out of 1,153 nonsynonymous RNA editing sites causing amino acid sequence changes of translated proteins were detected from 12 proteomes. (e) Random-selected cross-sample conserved RNA editing sites examined by experiments.