Molecular Genetics and Genomics (2021) 296:677–688
https://doi.org/10.1007/s00438-021-01768-z

ORIGINAL ARTICLE

Tracing foreign sequences in plant transcriptomes and genomes using OCT4, a POU domain protein

Adeleh Saffar[1] · Maryam M. Matin[1,2]

Received: 21 September 2020 / Accepted: 10 February 2021 / Published online: 18 March 2021

Abstract

Contaminations in sequencing data, especially in reference genomes, lead to inevitable errors in downstream analyses. Similarly, presence of contaminants in transcriptomes, misrepresents the molecular basis of various interactions. In this study, we report the presence of a large number of plant transcriptomes contaminated with RNAs encoding POU domain proteins; a family of proteins that has not been reported in plants and fungi. Besides, our findings illustrated that there are four POU domain protein-coding sequences in the reference genome of *Rhodamnia argentea*. It turned out that the existing foreign fragments are related to arthropods that are considered as plant pests. We also identified two contaminated draft genomes, *Humulus lupulus* and *Cannabis sativa* that contained complete rDNA sequences originating from *Tetranychus* species. As a result, careful screening of sequencing data before releasing them in public databases or checking existing genomes for possible contaminations is recommended.

Keywords Plant transcriptomes · POU domain protein · Contamination · Insect

Introduction

Genomic and transcriptomic data in public databases are the basis of many studies. With the invention of faster and cheaper sequencing methods, the number of draft genomes and transcriptomes in databases has dramatically increased. Draft genomes include several contigs or scaffolds that are presented in an unarranged form, not in the form of chromosomes and contain internal gaps (Breitwieser et al. 2019). Transcriptomes provide useful information about the gene expression profile in different conditions and show the basis for molecular functions of organisms. Sometimes, genome or transcriptome sequencing data contain fragments of foreign DNA from other organisms known as contaminations (Lafond-Lapalme et al. 2017). Contamination in sequencing projects is a common problem that results from sampling methods or sequence preparation (Orosz 2015). These contaminants must be identified and removed before genome or transcriptome assembly. Otherwise it can lead to deposition of erroneous data into public databases, which would cause problems in future data mining. However, the detection of these contaminants is not easy especially when there is no reference genome or transcriptome, that is, in the de novo assembly (Lafond-Lapalme et al. 2017).

In this study, contamination of plant transcriptomes with POU domain encoding RNAs was investigated. POU domain genes encode a family of transcription factors. Members of this family are characterized with the presence of a highly conserved bipartite DNA binding domain, a specific subdomain (POUs) in amine terminus, and a homeo-subdomain (POUh) in carboxyl terminus (Zhao 2013). These factors have important roles in embryogenesis, cell differentiation, immune function and neurogenesis. POU domain genes are present in a wide range of organisms, including *Caenorhabditis elegans*, *Drosophila*, *Xenopus*, zebrafish and human and have not been reported in plants and fungi (Ilia 2004).

To identify POU domain homologues in organisms that have not been completely sequenced or annotated, we searched nucleotide sequences available at the National Center for Biotechnology Information (NCBI) website, including expressed sequence

Communicated by Stefan Hohmann.

✉ Maryam M. Matin
matin@um.ac.ir

[1] Department of Biology, Faculty of Science, Ferdowsi University of Mashhad, Mashhad, Iran

[2] Novel Diagnostics and Therapeutics Research Group, Institute of Biotechnology, Ferdowsi University of Mashhad, Mashhad, Iran

Table 1 Sequences containing POU domain in plants. Plants are divided into two parts, the second part which comes after the separating line in the table includes those with low or zero coverage

Plant Species	Accession no	Insect Species	Accession no. 2	Cover, Identity %
Allium cepa	GBGJ01098742.1	*Frankliniella occidentalis*	XM_026418002.1	100%- 89.04%
Ambrosia trifida	GEOH01236939.1	*Tetranychus urticae*	XM_015925687.1	99%- 99.12%
Carya illinoinensis	GGRT01049572.1	*Bombyx mandarina*	XM_028171131.1	100%- 99.25%
Chrysanthemum x morifolium	IABW01189106.1	*Tetranychus urticae*	JR693665.1	100%- 97.13%
Diospyros lotus	GBSJ01218704.1	*Inostemma sp*	GBVB01035794.1	91%- 84.34%
Echinochloa colona	GFJI01013889.1	*Tetranychus urticae*	XM_015932808.2	98%- 93.16%
Echinochloa colona	GFJI01205559.1	*Tetranychus urticae*	JR693665.1	99%- 94.6%
Echinochloa colona	GFJI01427617.1	*Tetranychus urticae*	GT998146.1	92%- 83.58%
Eremochloa ophiuroides	GGKP01115460.1	*Tetranychus urticae*	XM_015930106.2	87%- 89.07%
Eremochloa ophiuroides	GGKP01148349.1	*Tetranychus urticae*	XM_015940011.2	99%- 91.71%
Eremochloa ophiuroides	GGKP01206424.1	*Tetranychus urticae*	XM_015930105.2	99%- 92.47%
Ginkgo biloba	GHLL01366405.1	*Thrips tabaci*	IADS01015141.1	100%- 91.22%
Glehnia littoralis	GGSB01091639.1	*Tetranychus urticae*	XM_015930105.2	100%- 94.75%
Glehnia littoralis	GGSB01094253.1	*Tetranychus urticae*	XM_015925687.1	100%- 96.34%
Glehnia littoralis	GGSB01094376.1	*Tetranychus urticae*	XM_015930106.2	98%- 92.16%
Glehnia littoralis	GGSB01104469.1	*Tetranychus urticae*	XM_015932808.2	100%- 96.96%
Humulus lupulus	GAAW01134144.1	*Tetranychus urticae*	XM_015925687.1	100%- 99.26%
Humulus lupulus	GAAW01035890.1	*Tetranychus urticae*	XM_015930106.2	100%- 99.77%
Humulus lupulus	GAAW01110159.1	*Myzus persicae*	XM_022304451.1	100%- 97.86%
Humulus lupulus	GAAW01119665.1	*Tetranychus urticae*	XM_015932808.2	100%- 100%
Humulus lupulus	GAAW01124668.1	*Myzus persicae*	XM_022304451.1	100%- 98.44%
Humulus lupulus	GAAW01131874.1	*Myzus persicae*	XM_022307304.1	100%- 97.34%
Humulus lupulus	LA341136.1	*Tetranychus urticae*	XM_015925687.1	100%-99.48%
Humulus lupulus	LA710056.1	*Tetranychus urticae*	XM_015925687.1	100%- 99.15%
Humulus lupulus	LA714473.1	*Tetranychus urticae*	XM_015932808.2	100%- 98.8%
Ipomoea purpurea	GALY01035227.1	*Ecdyonurus insignis*	GCCL01023636.1	91%- 81.38%
Ipomoea purpurea	GABG01053872.1	*Ecdyonurus insignis*	GCCL01023636.1	90%- 81.94%
Jasminum sambac	GHOY01040882.1	*Contarinia nasturtii*	XM_031763638.1	100%- 97.64%
Medicago sativa	GGKA01014557.1	*Trichoplusia ni*	XM_026874060.1	99%- 85.89%
Picea glauca	GCHX01016064.1	*Achipteria coleoptrata*	GEXX01038414.1	100%- 90.48%
Pinus sylvestris	GHLA01525201.1	*Frankliniella occidentalis*	XM_026438387.1	98%- 86.7%
Pinus sylvestris	GHKW01492915.1	*Frankliniella occidentalis*	XM_026438387.1	98%- 86.7%
Poa pratensis	GEBH01191041.1	*Schizaphis graminum*	GGMR01010813.1	95%- 98.65%
Poa pratensis	GEBH01003771.1	*Rhopalosiphum maidis*	XM_026954777.1	97%-98/22%
Salvia pomifera	GDKL01026355.1	*Tetranychus urticae*	XM_015930106.2	100%- 99.45%
Selaginella bryopteris	GEMU01081348.1	*Trichogramma pretiosum*	XM_014381976.2	100%- 97.78%
Silene conica	GDJK01041227.1	*Thrips tabaci*	IADS01015141.1	98%- 98.97%
Vaccinium virgatum	GGAB01003286.1	*Drosophila suzukii*	XM_017079329.1	100%- 98.10%
Sesamum indicum	JK069820.1	*Zeugodacus cucurbitae*	XM_011191139.2	98%- 80%
Citrus clementina	FC917318.1	*Ostrinia furnacalis*	XM_028317874.1	100%- 83%
Actinidia deliciosa	GEYI01009310.1	*Altererythrobacter marensis*	CP011805.1	2%-100%
Actinidia deliciosa	GEYI01013624.1	*Diaphanosoma celebensis*	GGQP01027771.1	15%- 81.4%
Actinidia deliciosa	GEYI01058905.1	*Dermacentor variabilis*	GGTZ01012657.1	56%-74.3%
Arabidopsis halleri	GFUL01013257.1	*Peromyscus leucopus*	XM_028865804.1	38%- 75.58%
Camellia sinensis	GFQB01021210.1	*Trichogramma pretiosum*	XM_023459974.1	2%- 96.88%
Camellia sinensis	GFQC01063922.1	*Trichogramma pretiosum*	XM_023459974.1	2%- 96.88%
Ceanothus thyrsiflorus	GGXO01024450.1	*Mania lunus*	GCOU01002324.1	58%- 92.31%
Chromolaena odorata	GACH01063862.1			0

Table 1 (continued)

Citrus clementina	DY279403.1	*Ostrinia furnacalis*	XM_028317874.1	77%- 82.97%
Diospyros lotus	GBSJ01266976.1			0
Echinochloa colona	GFJI01418603.1	*Tetranychus urticae*	XM_015925687.1	70%- 91.46%
Echinochloa colona	GFJI01427616.1	*Tetranychus urticae*	XM_015925687.1	78%- 83.77%
Eremochloa ophiuroides	GGKP01141248.1	*Tetranychus urticae*	XM_015925687.1	59%- 88.76%
Eremochloa ophiuroides	GGKP01034720.1	*Drosophila yakuba*	XM_002093923.2	28%- 77.66%
Eremochloa ophiuroides	GGKP01034721.1	*Ictalurus punctatus*	XM_017474674.1	46%- 82.55%
Eremochloa ophiuroides	GGKP01141443.1	*Drosophila miranda*	GALP01003242.1	18%- 87.05%
Eremochloa ophiuroides	GGKP01141445.1	*Apis dorsata*	XM_006612180.2	6%- 76.4%
Eremochloa ophiuroides	GGKP01202375.1	*Nilaparvata lugens*	IACV01112726.1	47%- 80.68%
Eremochloa ophiuroides	GGKP01204565.1	*Aceria tosichella*	GGYP01004680.1	25%- 90.32%
Eremochloa ophiuroides	GGKP01304111.1			0
Eremochloa ophiuroides	GGKP01304156.1	*Aethina tumida*	XM_020022743.1	42%- 79.31%
Eremochloa ophiuroides	GGKP01317440.1	*Aethina tumida*	XM_020022743.1	26%- 85%
Fagus crenata	GHGX01009058.1			0
Ipomoea trifida	GFXN01120848.1	*Penaeus vannamei*	XM_027368252.1	35%- 83.59%
Iris domestica	GGPC01036853.1	*Tetranychus urticae*	XM_015930431.2	78%- 75.86%
Lathyrus odoratus	GO319296.1	*Steatoda triangulosa*	AF273264.1	12%- 82%
Olea europaea	GBKW01086959.1	*Polistes dominula*	XR_001476078.1	3%- 100%
Picea glauca	GCHX01394126.1	*Steganacarus magnus*	GEYQ01024701.1	57%- 84.93%
Pinus massoniana	GFHB01189396.1	*Odontomachus brunneus*	XM_032829785.1	21%- 81.16%
Pinus massoniana	GFHB01178357.1	*Aceria tosichella*	GGYP01004680.1	56%- 82.59%
Rhodamnia argentea	XP_030539202.1	*Aceria tosichella*	GGYP01003014.1	30%- 78.85%
Rhodamnia argentea	XP_030540859.1	*Steganacarus magnus*	GEYQ01024696.1	50%- 72.63%
Rhodamnia argentea	XP_030540867.1	*Pardosa pseudoannulata*	GGRD01225477.1	29%- 79%
Rhodamnia argentea	XP_030541605.1	*Protophormia terraenovae*	IABT01080171.1	45%-68%
Salix integra	GEYA01023573.1			0
Salix integra	GEYA01025525.1	*Dendroctonus ponderosae*	GGKQ01101541.1	54%- 79.62%
Salix integra	GEYA01052535.1	*Drosophila navojoa*	XM_030382922.1	40%- 88%
Salix integra	GEYA01056204.1	*Nedyus quadrimaculatus*	GDON01021781.1	28%- 89.29%
Salix viminalis	HACX01005526.1	*Centruroides sculpturatus*	XM_023377148.1	4%- 87.5%
Salix viminalis	HACX01005526.1	*Centruroides sculpturatus*	XM_023377148.1	4%-87.5%
Salix viminalis	HACX01018601.1	*Bicyclus anynana*	XM_024090474.1	2%- 100%
Solanum melongena	GAYR01069607.1			0
Ulmus minor	GFUU01015552.1	*Drosophila ficusphila*	XM_017186619.1	2%- 100%

tags (ESTs), transcriptome shotgun assemblies (TSAs) and whole-genome shotgun contigs (WGSs). Surprisingly, despite the lack of POU domain genes in plants, we found such fragments in some EST and TSA databases. We also used Basic Local Alignment Search Tool (BLAST) in plant protein databases and found four sequences containing the POU domain in *Rhodamnia argentea*. As we investigated further, we discovered that these plant sequences are very similar to arthropod proteins. Moreover, we detected fragments that are highly similar to *Tetranychus* ribosomal RNA in WGS sequences of *Humulus lupulus* and *Cannabis sativa*. Since the *Tetranychus* species are considered as plant pests, we concluded that plant samples used in the sequencing have been likely contaminated with the arthropods genome.

Materials and methods

The accession numbers mentioned in this study are related to NCBI GenBank database. The *Homo sapiens* POU domain protein was used as a query in BLAST. Protein databases, and nucleotide sequences including ESTs, TSAs and WGSs, were searched to find similar sequences in plants using BLASTP 2.11.0 + and TBLASTN 2.11.0 + analyses (Altschul 1997), respectively. Plant sequences were used as queries and BLASTN 2.11.0 + analysis (Zhang et al. 2000) was carried out on arthropod EST and TSA databases.

The clustalW was used to align the sequences and the alignments were applied for phylogenetic tree construction. Maximum-likelihood (ML) analysis, with bootstrap values, using GTR_G model was performed by MEGA-X10.1

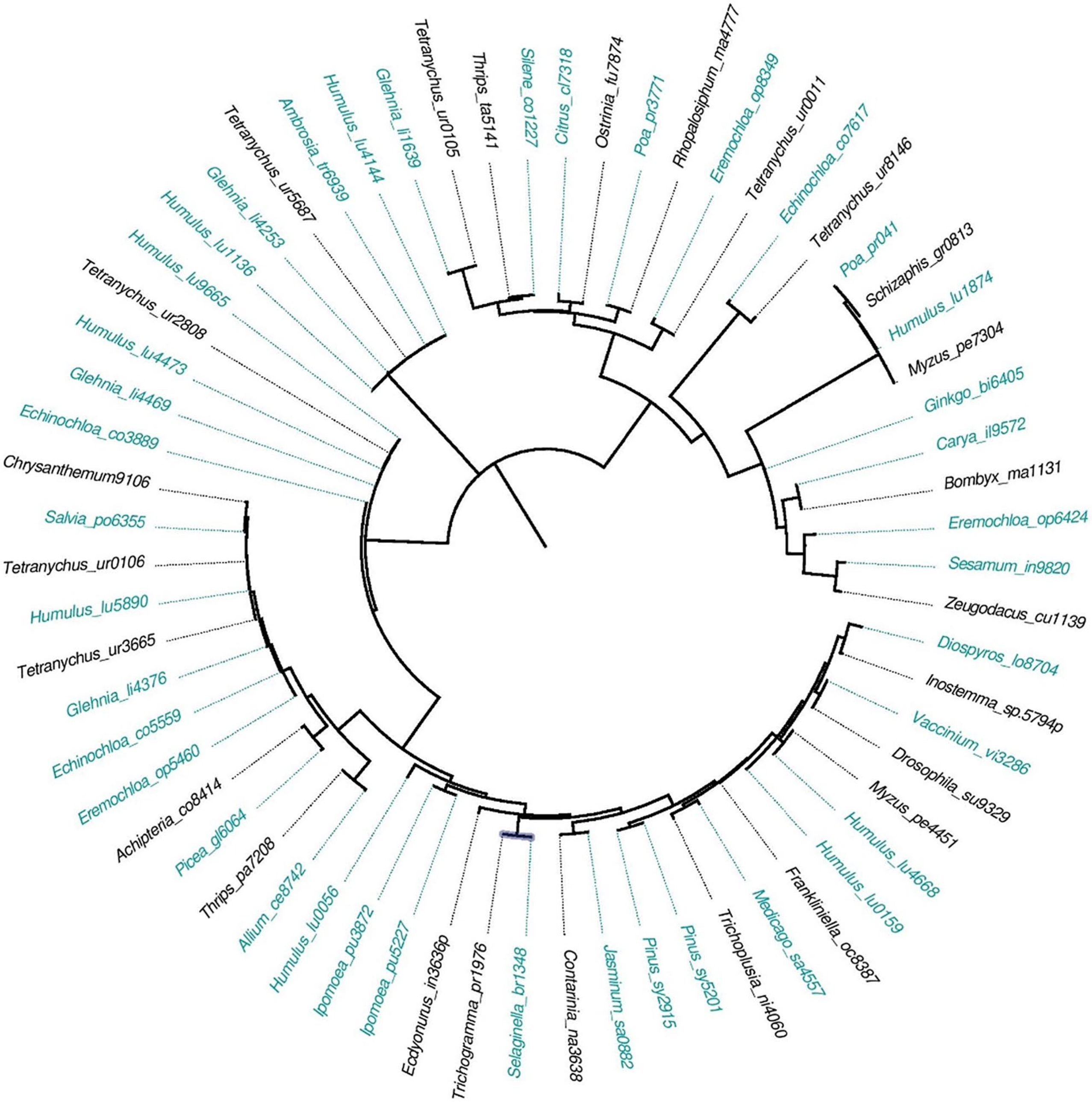

Fig. 1 Phylogenetic tree of sequences containing POU domain in plants and most similar matches in Arthropods. The names are in the form of the genus name of the plant (green) or Arthropods (black), along with the first two letters of the species name and the last four digits of the accession number

(Kumar et al. 2018). The phylogenetic tree was drawn with the software FIGTREE v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Results

The study began with searching for human OCT4 protein homologues in plant nucleotide sequences in the NCBI database. The results indicated that there are similar sequences with this protein in different plants. Since there is no report of the presence of this group of proteins, with the conserved POU domain, in plants, these observations rang an alarm

Table 2 *Rhodamnia argentea* genome containing POU domain proteins

Assembly	Chromosome	Location	mRNA	Protein	Name
Rarg10x-PRISCAF (GCF_900635035.1)	Unplaced Scaffold	NW_022060798.1 (2759099..2761297, complement)	XM_030683342.1	XP_030539202.1	ecdysone-induced protein 74EF-like
	Unplaced Scaffold	NW_022060798.1 (218385..219656)	XM_030684999.1	XP_030540859.1	POU domain protein 2-like
	Unplaced Scaffold	NW_022060798.1 (253088..255268)	XM_030685007.1	XP_030540867.1	POU domain, class 2, transcription factor 1-like
	Unplaced Scaffold	NW_022060887.1 (621048..622604, complement)	XM_030685745.1	XP_030541605.1	POU domain, class 4, transcription factor 1-like

bell which indicates the presence of contamination in plant sequences. Therefore, these sequences were evaluated more carefully.

In the first step, we adopted OCT4 protein (accession number: NP_002692) as a query and after using BLAST in protein, EST, transcriptome and WGS databases, we found similar sequences in plants. Since the mentioned protein has two domains, one of which belongs to the homeobox family (POUh), which is common in living organisms, the hits were further examined, to also include POUs domain. Then in the second step, plant sequences resulted from the previous step were used as queries to search in nucleotide collection database to identify the origin of sequences. The results indicated that these sequences had an arthropod or insect origin.

In insects, several proteins containing POU domain were initially recognized as essential transcription factors for expression of embryonic pattern formation genes. For example, *drifter* as a POU transcription factor functions in the neuroendocrine system development, cell fate determination of imaginal discs, neuronal lineage and wiring, differentiation and migration of tracheal cells and neurons, regulation

Table 3 Identity of *Tetranychus* 18S rRNA genes to *Humulus lupulus* var. lupulus contig

Genus	Species	Max score	Query cover (%)	E value	Identity (%)	base pair	GenBank accession no.
Tetranychus	*T. bambusae Wang & Ma*	3149	100%	0.0	97.26%	1859	AB926294
	T. evansi Baker & Pritchard	3360	100%	0.0	99.30%	1858	AB926295
	T. ezoensis Ehara	3421	100%	0.0	99.89%	1858	AB926296
	T. huhhotensis Ehara,Gotoh & Hong	3410	100%	0.0	99.78%	1858	AB926297
	T. kanzawai Kishida	3415	100%	0.0	99.84%	1858	AB926298
	T. lombardinii Baker & Pritchard	3338	100%	0.0	99.09%	1859	AB926299
	T. ludeni Zacher	3371	100%	0.0	99.41%	1858	AB926300
	T. macfarlanei Baker & Pritchard	3277	100%	0.0	98.49%	1859	AB926301
	T. merganser Boudreaux	3343	100%	0.0	99.14%	1858	AB926302
	T. misumaiensis Ehara & Gotoh	3398	100%	0.0	99.68%	1858	AB926303
	T. neocaledonicus Andre	3376	100%	0.0	99.46%	1859	AB926304
	T. okinawanus Ehara	3288	100%	0.0	98.60%	1859	AB926305
	T. parakanzawai Ehara	3421	100%	0.0	99.89%	1858	AB926306
	T. phaselus Ehara	3382	100%	0.0	99.52%	1858	AB926307
	T. piercei McGregor	3387	100%	0.0	99.57%	1858	AB926308
	T. pueraricola Ehara & Gotoh	3415	100%	0.0	99.84%	1858	AB926309
	T. truncatus Ehara	3415	100%	0.0	99.84%	1858	AB926310
	T. turkestani Ugarov & Nikolski	3415	100%	0.0	99.84%	1858	AB926311
	T. urticae Koch (green form)	3415	100%	0.0	99.84%	1858	AB926312
	T. urticae Koch (red form)	3415	100%	0.0	99.84%	1858	AB926313
	T. zeae Ehara, Gotoh & Hong	3410	100%	0.0	99.78%	1858	AB926314

Table 4 The most similar hits of *Tetranychus* 28S rRNA genes to *Humulus lupulus* var. lupulus

Genus	Species	Max score	Query cover (%)	E value	Identity (%)	base pair	GenBank accession no.
Tetranychus	*T. bambusae Wang & Ma*	878	100%	0.0	89.90%	685	AB926385
	T. evansi Baker & Pritchard	1232	100%	0.0	98.84%	696	AB926386
	T. ezoensis Ehara	1258	100%	0.0	99.57%	694	AB926387
	T. huhhotensis Ehara Gotoh & Hong	1242	100%	0.0	99.13%	694	AB926388
	T. kanzawai Kishida	1258	100%	0.0	99.57%	694	AB926389
	T. lombardinii Baker & Pritchard	1243	100%	0.0	99.13%	696	AB926390
	T. ludeni Zacher	1221	100%	0.0	98.55%	696	AB926391
	T. macfarlanei Baker & Pritchard	1238	100%	0.0	98.99%	696	AB926392
	T. merganser Boudreaux	1243	100%	0.0	99.13%	696	AB926393
	T. misumaiensis Ehara & Gotoh	1258	100%	0.0	99.57%	694	AB926394
	T. neocaledonicus Andre	1253	100%	0.0	99.42%	694	AB926395
	T. okinawanus Ehara	1216	100%	0.0	98.41%	696	AB926396
	T. parakanzawai Ehara	1264	100%	0.0	99.71%	694	AB926397
	T. phaselus Ehara	1249	100%	0.0	99.28%	696	AB926398
	T. piercei McGregor	1249	100%	0.0	99.28%	696	AB926399
	T. pueraricola Ehara & Gotoh	1258	100%	0.0	99.57%	694	AB926400
	T. truncatus Ehara	1258	100%	0.0	99.57%	694	AB926401
	T. turkestani Ugarov & Nikolski	1258	100%	0.0	99.57%	694	AB926402
	T. urticae Koch (green form)	1258	100%	0.0	99.57%	694	AB926403
	T. urticae Koch (red form)	1258	100%	0.0	99.57%	694	AB926404
	T. zeae Ehara, Gotoh & Hong	1258	100%	0.0	99.57%	694	AB926405

of sericin1, fibroin and dopa decarboxylase (DDC) in *Drosophila* (Deng et al. 2012).

Given that the initial BLAST search was performed with human OCT4 protein which is relatively more distant from the POU domain proteins available in insects, therefore, produced hits may show low coverage and identity values.

Since in the second BLAST, most of the plant sequences matched with the CF1A protein of the arthropods, we used that as our first query in BLAST against protein, EST, TSA and WGS databases, limiting our search to plants, to find more contaminated plants. Therefore, CF1A protein (accession number: XP_015781173) of *Tetranychus urticae* was used, which is a type of spider mite whose genome has been sequenced (Zhu et al. 2016). This increased the number of produced hits and the alignment scores. In this step, a total of 83 plant hits were obtained that were used as queries to find similar sequences in proteins, ESTs, and TSAs of arthropods. For each plant sequence we chose the most similar sequence found in arthropods. However, among plant hits containing POU domain, some cases had no similar sequences in arthropods or showed low coverage values. Furthermore, some plant sequences had significant similarities and coverage values to those in arthropods. These cases indicate that the contamination is probably related to other close arthropod species that their genome and transcriptome information is not available in the databases (Zhu et al. 2016). Table 1 shows the plants containing POU domain along with the most similar sequences in arthropods. A phylogenetic tree was constructed using sequences listed in Table 1 (Fig. 1). It is worth mentioning that only those plant sequences were used that had the identity and coverage values higher than 80% with arthropod sequences.

The CF1A protein was searched in the *Arabidopsis thaliana*, *Oryza sativa* and *Marchantia polymorpha* genomes using BLAST to ensure that sequences containing POU domain do not exist in the plants. We found no sequences in the model plants; which strengthens the possibility of contamination of those plants with foreign fragments.

Our results indicate that the presence of contaminants is not limited to plant transcriptomes. BLAST search using CF1A protein against plant nucleotide sequences in the GenBank database revealed that the reference genome (GCA_900635035.1) of *Rhodamnia argentea* has four POU domain sequences. These sequences contain protein annotations and encode four proteins. Search using BLASTP against the non-redundant protein database showed that the best match for each plant protein belongs to arthropod species with low coverage value. However, due to the annotation of one of these plant proteins called ecdysone-induced protein 74EF-like, we first speculated

Table 5 Identity of *Tetranychus* 18S rRNA genes to *Cannabis sativa* contig

Genus	Species	Max score	Query cover (%)	E value	Identity (%)	base pair	GenBank accession no.
	T. bambusae Wang & Ma	3155	100%	0.0	97.31%	1859	AB926294
	T. evansi Baker & Pritchard	3365	100%	0.0	99.35%	1858	AB926295
	T. ezoensis Ehara	3426	100%	0.0	99.95%	1858	AB926296
	T. huhhotensis Ehara Gotoh & Hong	3415	100%	0.0	99.84%	1858	AB926297
	T. kanzawai Kishida	3432	100%	0.0	100%	1858	AB926298
	T. lombardinii Baker & Pritchard	3343	100%	0.0	99.14%	1859	AB926299
	T. ludeni Zacher	3382	100%	0.0	99.52%	1858	AB926300
	T. macfarlanei Baker & Pritchard	3282	100%	0.0	98.55%	1859	AB926301
	T. merganser Boudreaux	3360	100%	0.0	99.30%	1858	AB926302
	T. misumaiensis Ehara & Gotoh	3404	100%	0.0	99.73%	1858	AB926303
Tetranychus	*T. neocaledonicus Andre*	3387	100%	0.0	99.57%	1859	AB926304
	T. okinawanus Ehara	3299	100%	0.0	98.71%	1859	AB926305
	T. parakanzawai Ehara	3426	100%	0.0	99.95%	1858	AB926306
	T. phaselus Ehara	3387	100%	0.0	99.57%	1858	AB926307
	T. piercei McGregor	3393	100%	0.0	99.62%	1858	AB926308
	T. puerariicola Ehara & Gotoh	3432	100%	0.0	100%	1858	AB926309
	T. truncatus Ehara	3432	100%	0.0	100%	1858	AB926310
	T. turkestani Ugarov & Nikolski	3432	100%	0.0	100%	1858	AB926311
	T. urticae Koch (green form)	3432	100%	0.0	100%	1858	AB926312
	T. urticae Koch (red form)	3432	100%	0.0	100%	1858	AB926313
	T. zeae Ehara, Gotoh & Hong	3415	100%	0.0	99.84%	1858	AB926314

that this protein might be related to the phytoecdysteroid producing plants that use it in invertebrate deterrence (Dinan et al. 2020). Phytoecdysteroids are analogues of ecdysone that is regulating moulting and metamorphosis in insects (Dinan et al. 2001). We did not find any similar proteins in the plants that produces this compound. On the other hand, none of the isoforms of the ecdysone-induced protein 74EF (A-E) in *Drosophila melanogaster* has any POU domain in their structure. Table 2 lists the location of these genes on the genome. Whole-genome sequencing and de novo assembly of the Australian rainforest tree, *Rhodamnia argentea*, have been published in 2019 (accession no.: PRJEB30444), from which the reference sequence genome assembly was derived. The above draft genome has 15,781 scaffolds and 23,310 contigs and the contaminated contigs include CAAAGQ010000092.1 and CAAAGQ010000003.1. The presence of such contaminants in the reference genome leads to erroneous results in subsequent studies that use these sequences as a reference for comparison purposes.

Our analysis also showed that the *Humulus lupulus* has a large number of contaminated sequences in its transcriptome. We used the CLC main workbench (QIAGEN) to examine the presence of contaminations in the genome of the mentioned plant. We performed a multi-sequence BLAST search using all of the transcripts of *Tetranychus urticae* as queries, limited to the *Humulus lupulus* WGS database. We found that several sequences related to large and small subunit ribosomal RNAs of the *Tetranychus urticae* are similar to a contig in the whole genome shotgun sequences of this plant. Although ribosomal RNAs are relatively conserved, since the coverage and identity values of *Tetranychus urticae* rRNAs to *Humulus* hit are much more than those to other *plants* such as *Arabidopsis thaliana*, it is sufficient to distinguish between contamination and endogenous genes. Given that ribosomal RNA sequences are the basis of phylogenetic studies, many of these rRNAs have been identified in different species and stored in databases (Zhao et al. 2012), so it is easy to detect the exact *Tetranychus* species causing the contamination in the *Humulus lupulus* genome. Thus, 18S and 28S rRNA genes were used as queries in the BLASTN algorithm against WGS sequences of *Humulus lupulus* (Matsuda et al. 2014). Both genome assemblies of the *Humulus lupulus* varieties (GCA_000831365.1 and GCA_000830395.1) contain contaminant sequences found on the BBPC01017300.1 and BBPB01017300.1 contigs. Table 3 and 4 show the identity of the 18S and 28S rRNA genes of *Tetranychus* species with the best match in the WGS sequence of *Humulus lupulus* var. lupulus. Assembly of the genomic sequences of *H. lupulus* varieties was

Table 6 Highest identity of *Tetranychus* 28S rRNA genes to *Cannabis sativa* contig

Genus	Species	Max score	Query cover (%)	E value	Identity (%)	base pair	GenBank accession no.
Tetranychus	*T. bambusae Wang & Ma*	891	100%	0.0	90.10%	685	AB926385
	T. evansi Baker & Pritchard	1256	100%	0.0	99.28%	696	AB926386
	T. ezoensis Ehara	1271	100%	0.0	99.71%	694	AB926387
	T. huhhotensis Ehara,Gotoh & Hong	1266	100%	0.0	99.57%	694	AB926388
	T. kanzawai Kishida	1271	100%	0.0	99.71%	694	AB926389
	T. lombardinii Baker & Pritchard	1267	100%	0.0	99.57%	696	AB926390
	T. ludeni Zacher	1245	100%	0.0	98.99%	696	AB926391
	T. macfarlanei Baker & Pritchard	1262	100%	0.0	99.43%	696	AB926392
	T. merganser Boudreaux	1267	100%	0.0	99.57%	696	AB926393
	T. misumaiensis Ehara & Gotoh	1282	100%	0.0	100%	694	AB926394
	T. neocaledonicus Andre	1277	100%	0.0	99.86%	694	AB926395
	T. okinawanus Ehara	1243	100%	0.0	98.99%	696	AB926396
	T. parakanzawai Ehara	1277	100%	0.0	99.86%	694	AB926397
	T. phaselus Ehara	1273	100%	0.0	99.71%	696	AB926398
	T. piercei McGregor	1273	100%	0.0	99.71%	696	AB926399
	T. pueraricola Ehara & Gotoh	1282	100%	0.0	100%	694	AB926400
	T. truncatus Ehara	1282	100%	0.0	100%	694	AB926401
	T. turkestani Ugarov & Nikolski	1282	100%	0.0	100%	694	AB926402
	T. urticae Koch (green form)	1282	100%	0.0	100%	694	AB926403
	T. urticae Koch (red form)	1282	100%	0.0	100%	694	AB926404
	T. zeae Ehara, Gotoh & Hong	1282	100%	0.0	100%	694	AB926405

Table 7 Identity of *Tetranychus* ITS genes to *Humulus lupulus* contig

Genus	Species	Max score	Query cover (%)	E value	Identity (%)	base pair	GenBank accession no.
Tetranychus	*T. evansi Baker and Prichard*	1140	100%	0.0	91.90%	824	AB735996
	T. ezoensis Ehara	1511	100%	0.0	100.00%	818	AB735998
	T. kanzawai Kishida	1500	100%	0.0	99.76%	818	AB736000
	T. ludeni Zacher	1164	100%	0.0	92.49%	814	AB736008
	T. misumaiensis Ehara & Gotoh	1382	100%	0.0	97.08%	822	AB736011
	T. neocaledonicus Andre	1116	99%	0.0	91.65%	809	AB736012
	T. okinawanus Ehara	660	89%	0.0	82.75%	912	AB736015
	T. parakanzawai Ehara	1506	100%	0.0	99.88%	818	AB736017
	T. phaselus Ehara	1214	100%	0.0	93.37%	826	AB736023
	T. piercei McGregor	1290	100%	0.0	94.95%	831	AB736025
	T. pueraricola Ehara & Gotoh	1450	100%	0.0	98.66%	819	AB736028
	T. truncatus Ehara	1417	100%	0.0	97.92%	819	AB736031
	T. urticae Koch (green form)	1458	100%	0.0	98.78%	820	AB736033
	T. urticae Koch (red form)	1463	100%	0.0	98.90%	821	AB736036
	Oligonychus coffeae (Nietner)	278	30%	7e-72	87.92%	778	AB757829
	Oligonychus gotohi Ehara	276	27%	2e-71	89.86%	783	AB757828

Table 8 Identity of *Tetranychus* ITS genes to *Cannabis sativa* contig

Genus	Species	Max score	Query cover (%)	E value	Identity (%)	base pair	GenBank accession no.
	T. evansi Baker and Prichard	1136	100%	0.0	91.80%	824	AB735996
	T. ezoensis Ehara	1469	100%	0.0	99.03%	818	AB735998
	T. kanzawai Kishida	1458	100%	0.0	98.78%	818	AB736000
	T. ludeni Zacher	1155	100%	0.0	92.28%	814	AB736008
	T. misumaiensis Ehara & Gotoh	1378	100%	0.0	96.97%	822	AB736011
	T. neocaledonicus Andre	1127	99%	0.0	91.87%	809	AB736012
Tetranychus	*T. okinawanus Ehara*	673	89%	0.0	82.99%	912	AB736015
	T. parakanzawai Ehara	1463	100%	0.0	98.90%	818	AB736017
	T. phaselus Ehara	1205	100%	0.0	93.15%	826	AB736023
	T. piercei McGregor	1299	100%	0.0	95.09%	831	AB736025
	T. pueraricola Ehara & Gotoh	1465	100%	0.0	98.90%	819	AB736028
	T. truncatus Ehara	1426	100%	0.0	98.05%	819	AB736031
	T. urticae Koch (green form)	1509	100%	0.0	99.88%	820	AB736033
	T. urticae Koch (red form)	1511	100%	0.0	99.88%	821	AB736036
	Oligonychus coffeae (Nietner)	278	28%	7e-72	89.04%	778	AB757829
	Oligonychus gotohi Ehara	276	27%	2e-71	89.86%	783	AB757828

performed by Natsume, et al. in 2014 and because *H. lupulus* var. lupulus (Cultivar: Shinshu Wase) was used as the reference genome in genome sequence construction of var. cordifolius (Natsume et al. 2015), this genome also shows similar contaminations (Tables S1 and S2).

As investigated further, we discovered a similar contamination in the WGS sequence of the *Cannabis sativa* plant. The genome (accession no.: PRJNA575581) has recently been obtained by Mckernan, et al. using PacBio single-molecule sequencing in 2020. Tables 5 and 6 show the *Tetranychus* 18S and 28S rRNA genes with the highest identity to contigs in the WGS sequences of *Cannabis sativa*. These contaminants have been located on the JAATIR010000607.1, JAATIR010000576.1, JAATIR010000555.1, JAATIR010000517.1 and JAATIR010000515.1 contigs.

Due to the similarity of the 18S and 28S ribosomal genes of *Humulus* and *Cannabis* with *Tetranychus* species, we assumed that the internal transcribed spacer (ITS) of the nuclear ribosomal DNA (rDNA) might also be similar between them. Therefore, we performed a BLAST search using ITS genes of *Tetranychus* species as queries in the WGS database of mentioned plants and found contigs with significant identity to *Tetranychus* genes. This examination revealed that another genome sequencing and assembly project of the *Cannabis sativa* (GCA_001509995.1) is contaminated with these *Tetranychus* ITSs. The contamination is in cultivar Chemdog91 on contig26666 (GenBank: LKUB01105710.1). This finding is also mentioned in a recently published article (Steinegger and Salzberg 2020). ITS genes are used as barcodes to discriminate *Tetranychus* species (Matsuda et al. 2013), and we also used these genes for phylogenetic tree construction. Tables 7 and 8 list similarities between ITS genes of *Tetranychus* species and *Humulus lupulus* and *Cannabis sativa* contigs.

Construction of the phylogenetic tree was carried out using ITS sequences of *Tetranychus* species (Tables 7 and 8) and corresponding sequences of *Humulus lupulus* (GenBank: BBPC01017300.1) and *Cannabis sativa* (GenBank: JAATIR010000607.1) and also *Oligonychus coffeae* (Nietner) and *Oligonychus gotohi Ehara* sequences as outgroups (Fig. 2). According to the results of the phylogenetic tree, the *Cannabis sativa* contig is adjacent to the *Tetranychus urticae* and the *Humulus lupulus* contig is located closest to the *T. kanzawai* and *T. ezoensis*. Moreover, the BLAST search of complete rDNA sequences (18S rRNA, ITS1, 5.8S rRNA, ITS2, 28S rRNA) showed the highest identity of *Cannabis sativa* contigs to *Tetranychus urticae* (99.88%) and *Humulus lupulus* contigs to *T. parakanzawai* (100%).

Discussion

In the present study, we found highly similar sequences between plants and arthropods. Relevant sequences contained POU domain that has not been reported in plants and fungi. We found these sequences in the transcriptome, reference sequence and NR protein databases of plants. We also detected sequences similar to *Tetranychus* rDNA

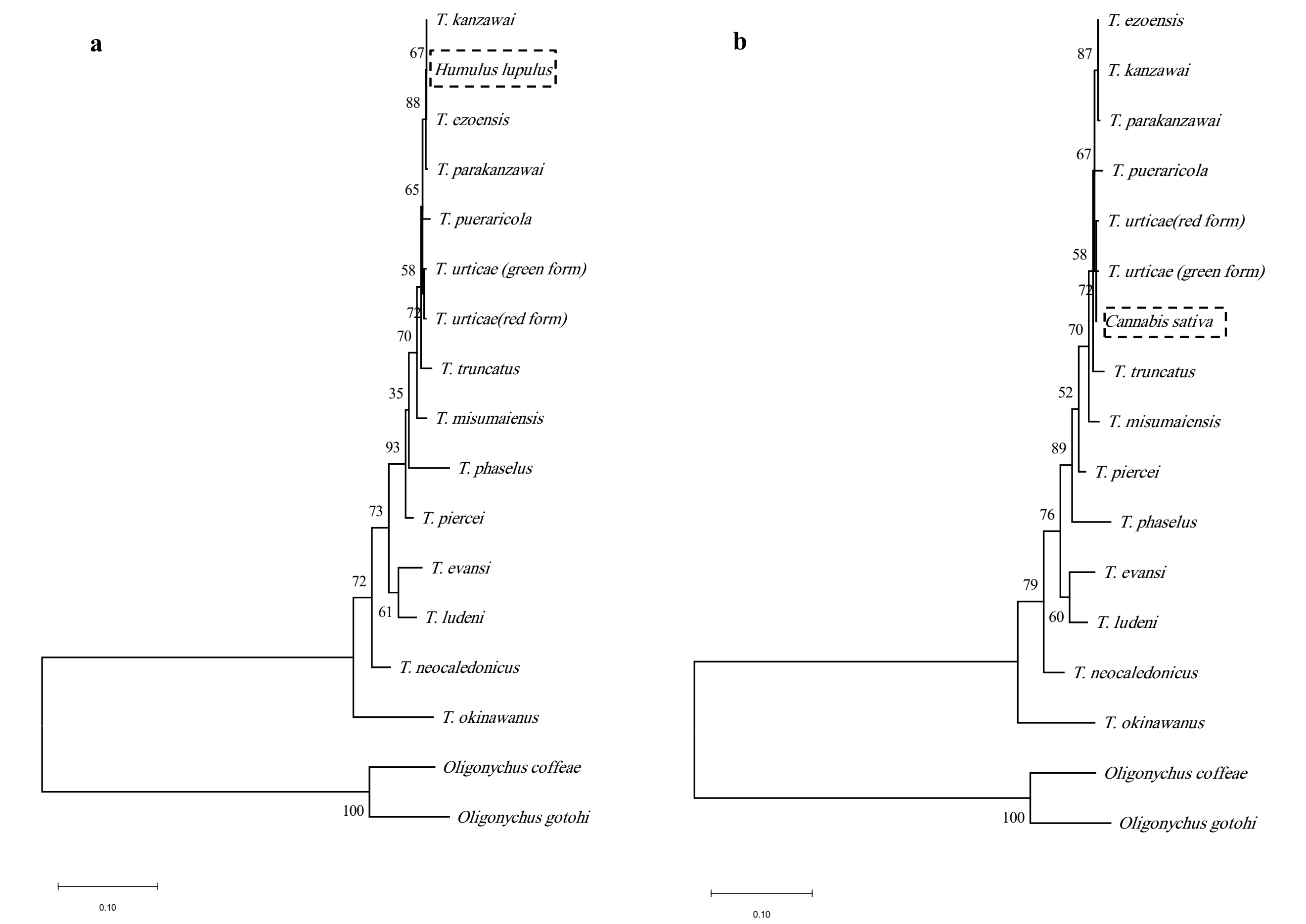

Fig. 2 Maximum likelihood (ML) tree of the *Tetranychus* species based on the internal transcribed spacer region using GTR_G model. Bootstrap values based on 1000 Replications are shown at nodes. The tree was rooted with *Oligonychus coffeae* and *Oligonychus gotohi*. The GenBank accession numbers are available in Tables 7 and 8. **a** The corresponding sequence of *Humulus lupulus* used included 13,291–12,474 nucleotides of BBPC01017300.1 contig. **b** The corresponding sequence of *Cannabis sativa* used included 5827–5007 nucleotides of JAATIR010000607.1 contig

genes in the whole genome shotgun sequences of two plants (*Humulus lupulus* and *Cannabis sativa*). These similarities can result from different hypotheses, such as conservation, horizontal gene transfer (HGT) and sequencing data contamination. Due to the observed similarities between plant rDNA genes (*Humulus* and *Cannabis*) and arthropod (*Tetranychus*), it was possible to study their homology using a phylogenetic tree. Therefore, we selected the ITS gene homologues in *Tetranychus* species and two species of another genus (*Oligonychus*) in this sub-family and performed multiple alignments with the plant ITS genes and then built maximum likelihood trees. If the homology between plant and arthropod sequences was because of conservation, we would expect the plant sequence to be located outside of arthropods in the tree. The plant sequence is located deep within the *Tetranychus* cluster, while even the two species of genus *Oligonychus* of this sub-family form a separate group. This observation convinces us to accept that such identities are beyond the conservation hypothesis and to look for other reasons, including horizontal gene transfer and contamination.

These arthropod species are typically pests of above plants (Matsuda et al. 2018). *Tetranychus urticae*, for example, feeds on 1,100 plant species (Grbić et al. 2011). Some articles have not completely ruled out the possibility of horizontal transfer of genes from insects to plants through viruses and fungi. For example, Zhu et al. (2016) reported that arthropods *OBP* and *CSP* genes might have been introduced to plant transcriptomes by viruses and other vectors, while we found approximately 7072 sequences in *Humulus* transcriptome (common plant in both studies) similar to *Tetranychus* TSA sequences with an identity higher than 93%

(Table S3). Given that some of these sequences are animal-specific and HGT is not common among distantly related organisms (Cornet et al. 2018), the question arises whether all of them have been transferred by HGT.

However, the contamination hypothesis which considers these sequences to be due to the cell-containing secretions left from the saliva of nectar-sucking insects (Zhu et al. 2016) seems more probable. The presence of cross-species contamination in sequencing data deposited in public databases is not uncommon. Contaminants may be introduced during various stages, including sampling or sequencing. Such contaminants are easy to find and remove if they belong to predictable sources such as bacterial cloning vectors and human sequences (Cornet et al. 2018). But detection of contaminants from organisms without fully sequenced genomes is challenging. In this study, the presence of arthropod genes in the plant transcriptomes and genomes, in addition to revealing contaminations, can provide useful information about the relationship between insects and other arthropods and their host plants (Zhu et al. 2016).

Overall, the results of the present study indicate the need for careful screening of sequencing data in terms of possible contamination before their release in the common databases. Of course, sometimes this may not be possible due to the transposition of sequencing projects; however, updating the databases may solve some of these problems.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00438-021-01768-z.

Acknowledgements This work was supported by Ferdowsi University of Mashhad.

Availability of data All sequences used in this study were downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov/genome/).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest with the contents of this article.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Altschul S (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402. https://doi.org/10.1093/nar/25.17.3389

Breitwieser FP, Pertea M, Zimin AV, Salzberg SL (2019) Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res 29(6):954–960. https://doi.org/10.1101/gr.245373.118

Cornet L, Meunier L, Van Vlierberghe M, Léonard RR, Durieu B, Lara Y, Misztak A, Sirjacobs D, Javaux EJ, Philippe H, Wilmotte A, Baurain D (2018) Consensus assessment of the contamination level of publicly available cyanobacterial genomes. PLoS ONE 13(7):1–26. https://doi.org/10.1371/journal.pone.0200323

Deng H, Zhang J, Li Y, Zheng S, Liu L, Huang L, Wei Hua Xu, Palli SR, Feng Q (2012) Homeodomain POU and Abd-A proteins regulate the transcription of pupal genes during metamorphosis of the Silkworm, Bombyx Mori. Proc Natl Acad Sci USA 109(31):12598–12603. https://doi.org/10.1073/pnas.1203149109

Dinan L, Savchenko T, Whiting P (2001) On the distribution of phytoecdysteroids in plants. Cell Mol Life Sci 58(8):1121–1132. https://doi.org/10.1007/PL00000926

Dinan L, Balducci C, Guibout L, Lafont R (2020) Small-scale analysis of phytoecdysteroids in seeds by HPLC-DAD-MS for the identification and quantification of specific analogues, dereplication and chemotaxonomy. Phytochem Anal 31(5):643–661. https://doi.org/10.1002/pca.2930

Grbić M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, Osborne EJ, Dermauw W, Ngoc PCT, Ortego F, Hernández-Crespo P, Diaz I, Martinez M, Navajas M, Sucena É, Magalhães S, Nagy L, Pace RM, Djuranović S, Smagghe G, Iga M, Christiaens O, Veenstra JA, Ewer J, Villalobos RM, Hutter JL, Hudson SD, Velez M, Yi SV, Zeng J, Silva A-D, Roch F, Cazaux M, Navarro M, Zhurov V, Acevedo G, Bjelica A, Fawcett JA, Bonnet E, Martens C, Baele G, Wissler L, Sanchez-Rodriguez A, Tirry L, Blais C, Demeestere K, Henz SR, Ryan Gregory T, Mathieu J, Verdon L, Farinelli L, Schmutz J, Lindquist E, Feyereisen R, Van De Peer Y (2011) The genome of tetranychus urticae reveals herbivorous pest adaptations. Nature 479(7374):487–492. https://doi.org/10.1038/nature10640

Ilia, Maria. 2004. "Oct-6 TRANSCRIPTION FACTOR." 59:471–89.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35(6):1547–1549. https://doi.org/10.1093/molbev/msy096

Lafond-Lapalme J, Duceppe MO, Wang S, Moffett P, Mimee B (2017) A new method for decontamination of de novo transcriptomes using a hierarchical clustering algorithm. Bioinformatics 33(9):1293–1300. https://doi.org/10.1093/bioinformatics/btw793

Matsuda T, Fukumoto C, Hinomoto N, Gotoh T (2013) DNA-based identification of spider mites: molecular evidence for cryptic species of the Genus Tetranychus (Acari: Tetranychidae). J Econ Entomol 106(1):463–472. https://doi.org/10.1603/EC12328

Matsuda T, Kozaki T, Ishii K, Gotoh T (2018) Phylogeny of the spider mite sub-family Tetranychinae (Acari: Tetranychidae) inferred from RNA-Seq data. PLoS ONE 13(9):1–14. https://doi.org/10.1371/journal.pone.0203136

Matsuda T, Morishita M, Hinomoto N, Gotoh T (2014) Phylogenetic analysis of the spider mite sub-family tetranychinae (Acari: Tetranychidae) based on the mitochondrial COI gene and the 18S and the 59 end of the 28S RRNA genes indicates that several genera are polyphyletic. PLoS ONE. https://doi.org/10.1371/journal.pone.0108672

Natsume S, Takagi H, Shiraishi A, Murata J, Toyonaga H, Patzak J, Takagi M, Yaegashi H, Uemura A, Mitsuoka C, Yoshida K, Krofta K, Satake H, Terauchi R, Ono E (2015) the draft genome of hop (*Humulus lupulus*), an essence for brewing. Plant Cell Physiol 56(3):428–441. https://doi.org/10.1093/pcp/pcu169

Orosz F (2015) Two recently sequenced vertebrate genomes are contaminated with apicomplexan species of the sarcocystidae family. Int J Parasitol 45(13):871–878. https://doi.org/10.1016/j.ijpara.2015.07.002

Steinegger M, Salzberg SL (2020) Terminating contamination: large-scale search identifies more than 2,000,000 contaminated Entries in GenBank. Genome Biol 21(1):115. https://doi.org/10.1186/s13059-020-02023-1

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A Greedy algorithm for aligning DNA sequences. J Comput Biol 7(1–2):203–214. https://doi.org/10.1089/10665270050081478

Zhao FQ (2013) Octamer-binding transcription factors: genomics and functions. Front Biosci 18(3):1051–1071. https://doi.org/10.2741/4162

Zhao YE, Li Ping Wu, Li Hu, Yang Xu, Wang ZH, Liu WY (2012) Sequencing for complete RDNA sequences (18S, ITS1, 5.8S, ITS2, and 28S RDNA) of demodex and phylogenetic analysis of acari based on 18S and 28S RDNA. Parasitol Res 111(5):2109–2114. https://doi.org/10.1007/s00436-012-3058-8

Zhu J, Wang G, Pelosi P (2016) Plant Transcriptomes reveal hidden guests. Biochem Biophys Res Commun 474(3):497–502. https://doi.org/10.1016/j.bbrc.2016.04.134

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.