the plant journal

The Plant Journal (2022) 109, 278–294 doi: 10.1111/tpj.15558

RESOURCE

The gene space of European mistletoe (*Viscum album*)

Lucie Schröder[1], Natalija Hohnjec[2], Michael Senkler[1], Jennifer Senkler[1], Helge Küster[2] and Hans-Peter Braun[1]*

[1]*Plant Proteomics, Institute of Plant Genetics, Leibniz Universität Hannover, Herrenhäuser Str. 2, 30419, Hannover, Germany, and*

[2]*Plant Genomics, Institute of Plant Genetics, Leibniz Universität Hannover, Herrenhäuser Str. 2, 30419, Hannover, Germany*

Received 28 July 2021; revised 28 September 2021; accepted 1 October 2021; published online 29 October 2021.
*For correspondence (e-mail braun@genetik.uni-hannover.de).

SUMMARY

European mistletoe (*Viscum album*) is a hemiparasitic flowering plant that is known for its very special life cycle and extraordinary biochemical properties. Particularly, *V. album* has an unusual mode of cellular respiration that takes place in the absence of mitochondrial complex I. However, insights into the molecular biology of *V. album* so far are very limited. Since the genome of *V. album* is extremely large (estimated 600 times larger than the genome of the model plant *Arabidopsis thaliana*) it has not been sequenced up to now. We here report sequencing of the *V. album* gene space (defined as the space including and surrounding genic regions, encompassing coding as well as 5′ and 3′ non-coding regions). mRNA fractions were isolated from different *V. album* organs harvested in summer or winter and were analyzed via single-molecule real-time sequencing. We determined sequences of 39 092 distinct open reading frames encoding 32 064 *V. album* proteins (designated *V. album* protein space). Our data give new insights into the metabolism and molecular biology of *V. album*, including the biosynthesis of lectins and viscotoxins. The benefits of the *V. album* gene space information are demonstrated by re-evaluating mass spectrometry-based data of the *V. album* mitochondrial proteome, which previously had been evaluated using the *A. thaliana* genome sequence. Our re-examination allowed the additional identification of nearly 200 mitochondrial proteins, including four proteins related to complex I, which all have a secondary function not related to respiratory electron transport. The *V. album* gene space sequences are available at the NCBI.

Keywords: SMRT sequencing, viscotoxins, lectins, mitochondria, oxidative phosphorylation, complex I, *Viscum album*, *Arabidopsis thaliana*.

INTRODUCTION

European mistletoe (*Viscum album*) is an obligate hemiparasitic flowering plant that grows on branches of various trees. It is supplied with water, minerals and organic compounds from the host. At the same time, *V. album* carries out photosynthesis and produces energy-rich compounds. *Viscum album* is widely distributed in central and northern Europe. It nicely is visible from November to March because it belongs to the few angiosperms that do not discard their leaves in the European winter. In fact, *V. album* is photosynthetically active at temperatures below the freezing point. *Viscum album* can cause problems in tree vitality, especially in combination with water stress. However, under favorable growth conditions, host trees are only moderately affected and can well coexist with the hemiparasite. European mistletoe has important ecological functions. Its flowers and berries ripe in winter and are a nutritional source for several insects and birds.

Compared to other flowering plants, the life cycle of *V. album* is characterized by numerous remarkable features (reviewed, e.g., in Glatzel and Geils, 2009): (i) *V. album* does not germinate in soil but on branches of trees, which requires particularly 'sticky' fruits (berries) that stably attach to tree bark; (ii) seeds consist of an embryo but lack a seed coat; (iii) embryos can germinate directly from the berry (without a dormancy phase); (iv) the direction of initial shoot growth is not determined by positive but rather negative phototropism, which guides the shoot onto the surface of the branch of the host tree; (v) the shoot

afterwards penetrates the branch and gets connected to the xylem of the vascular system, where it forms a haustorium for uptake of water, minerals and organic compounds; (vi) the dichotomous mistletoe plant, which afterwards develops, forms one pair of shoot segments per year per shoot apical meristem and two comparatively simply organized leaves, which resemble primary leaves; (vii) shoots grow into all directions, giving rise to the typical ball-like shape of the adult plant (overall, the growth rate of *V. album* is low); (viii) in contrast to the leaves of the host tree, mistletoe leaves do not close stomata during water shortage (which may dramatically increase water stress of host plants); (ix) older leaves of the previous growth periods are discarded in September without preceding chlorophyll recycling; (x) leaves of the current growth period are kept during winter and perform photosynthesis; and (xi) fruit ripening and seed dispersal take place in winter.

Viscum album also has a particular biochemical composition. It is known for its rich content in phenolic acids, phenylpropanoids, flavonoids, triterpenes and phytosterols (Jäger *et al.*, 2021; Urech and Baumgartner, 2015). It contains low-molecular-mass proteins designated viscotoxins as well as characteristic lectins (viscolectins), both of which contribute to its biotic defense system. The glue-like substances present in mistletoe berries mainly consist of hemicellulose compounds (Azuma *et al.*, 2000). It is clear but hardly addressed by scientific investigations that the development of mistletoe is based on a very unusual distribution of phytohormones. Extracts of *V. album* have cytotoxic and immune-stimulating effects and are used in medicine (Nazaruk and Orlikowski, 2016).

On a molecular scale, *V. album* has been less characterized to date. Its mitochondrial and chloroplast genomes have been sequenced (Petersen *et al.*, 2015a,b; Skippington *et al.*, 2015, 2017) and were surprisingly found to lack some genes previously considered to be essential for multicellular eukaryotes, like genes encoding subunits of complex I of the mitochondrial respiratory chain. In contrast, the sequence of the nuclear genome has not been analyzed. The *V. album* genome consists of $2n = 20$ chromosomes, is exceptionally large and is estimated to have a mass of 160 pg (80 pg for the haploid genome; average taken from the 'Plant DNA C-values Database', Pellicer and Leitch, 2019 [https://cvalues.science.kew.org/search]; original data from Nagl *et al.*, 1983 [53.5 pg], Ulrich *et al.*, 1988 [79.3 pg], Marie and Brown, 1993 [76 and 77.5 pg] and Zonneveld, 2010 [102.9 pg]). Indeed, the *V. album* genome is one of the largest genomes of any flowering plant known to date (Novák *et al.*, 2020; Zonneveld, 2010). Its size has been estimated to be in the range of 88×10^9 base pairs (approximately 90 Gbp; Novák *et al.*, 2020), which is 600 times the size of the genome of the model plant *Arabidopsis thaliana* (approximately 0.15 Gbp). Correspondingly, chromosomes of *V. album* are very large. Structural rearrangements in the chromosomes occur frequently and may cause large chromosome assemblies during meiosis (Barlow, 1981). The GC content is in the range of 39%, which is about average for flowering plants (Novák *et al.*, 2020). An initial transcriptome analysis of *V. album* haustorium tissue has been performed and yielded sequences of 3044 open reading frames (Ko *et al.*, 2014).

The gene content of seed plants (angiosperms and gymnosperms) is considered to be similar and amounts to approximately 0.03 Gbp (Novák *et al.*, 2020). This implies that the gene content of *V. album* only covers 0.03% of its genome (20% in *A. thaliana*) and that the size of the intergene space is enormous. In general, genome size of eukaryotes correlates with the amount of repetitive DNA (Elliott and Gregory, 2015). Interestingly, this does not hold true for especially large genomes of seed plants (>10 Gbp), which unexpectedly were found not to have a further increased amount of repetitive DNA. In *V. album*, the genome proportion of repeats (copy number > 20) is 55% (Novák *et al.*, 2020). This leaves much space for non-repetitive and low-copy DNA (excluding protein-coding genes).

Due to genome size and the amount of repetitive DNA, determination of the *V. album* genome sequence remains challenging. We therefore decided to firstly characterize the *V. album* gene space. The mRNA fraction was extracted from various organs of *V. album*, reverse-transcribed into cDNAs and subsequently used for systematic sequence determination by single-molecule real-time (SMRT) sequencing. We developed a database including >39 000 *V. album* gene sequences, which contain complete open reading frames encoding *V. album* proteins. Several new insights into the molecular biology of *V. album* are provided by an initial analysis of the deduced protein sequences and by re-evaluation of previously published *V. album* proteome data. The database is publicly available.

RESULTS AND DISCUSSION

SMRT sequencing of the *V. album* gene space

SMRT sequencing was used to analyze the full-length transcriptome of a pooled *V. album* RNA sample (representing stems, leaves and male and female flower buds; harvested in summer and winter). Quality control of our RNA was performed by gel electrophoresis, Qubit fluorometry and Nanodrop spectrophotometry (Table S1). The pooled RNA sample was reverse-transcribed into cDNA and subsequently converted into double-stranded cDNA. Two SMRTbell libraries (termed libraries A and B) were constructed for sequencing without size selection. SMRT sequencing was performed using both libraries. The analysis workflow is given in Figure 1 and a summary of the primary results is given in Table 1.

Overall, SMRT sequencing of libraries A and B revealed 321 472 and 343 119 circular consensus sequences (CCSs;

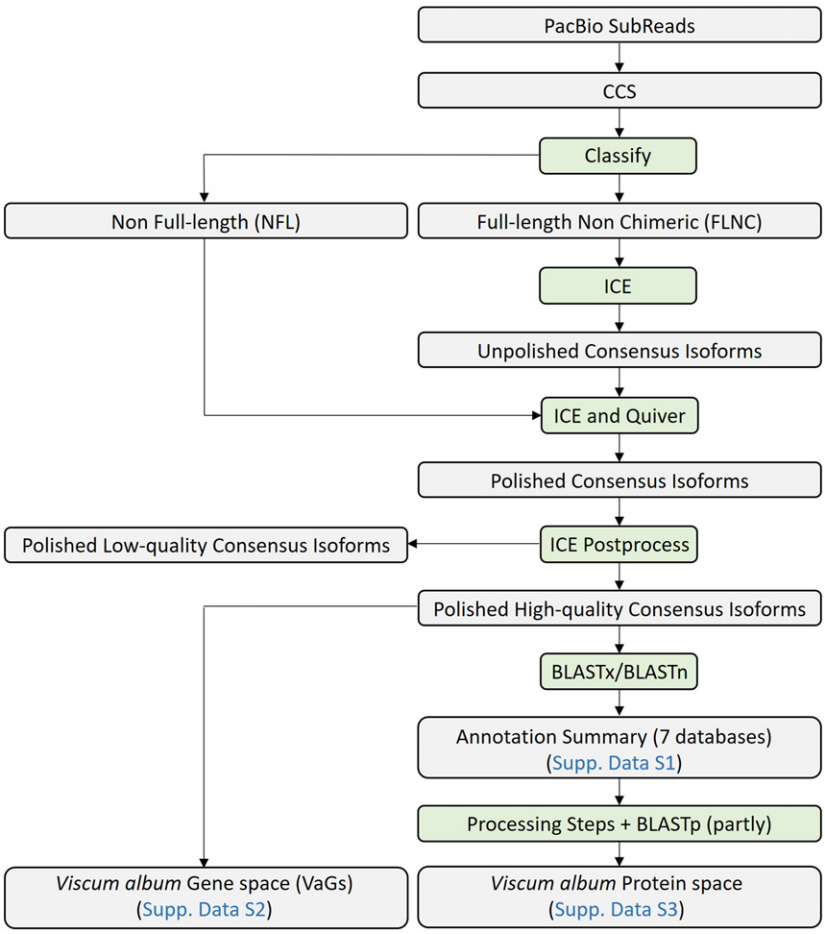

Figure 1. *Viscum album* gene space annotation summary. The different processing steps are shown in green. CCS, circular consensus sequence; ICE, Iterative Clustering for Error correction. For further information, see the Experimental Procedures section. Data are presented in Data S1–S3 (the Supplemental Data are highlighted in blue in the figure).

Table 1 Summary of reads from PacBio SMRT sequencing (for analysis workflow see Figure 1)

	Library A	Library B
Number of subreads	11 894 129	10 838 444
Circular consensus sequence (CCS) number	321 472	343 119
Average CCS length	1746	1846
Full-length reads	286 599	306 147
Non-full-length reads	27 518	30 762
Short reads	7355	6210
Full-length non-chimeric reads	253 284	268 386
Average length of full-length non-chimeric reads	1522	1595
Collected final consensus after polishing	161 841	
High-quality (hq) sequences	39 092	
Low-quality (lq) sequences	122 749	

Table 1). In total, 89% of the sequences were classified as full-length transcripts (including 5′ and 3′ adapters as well as poly(A) tails). In a next processing step, full-length non-chimeric sequences were defined for both libraries. The Iterative Clustering for Error correction (ICE) algorithm was used to define unpolished consensus isoforms, which afterwards were polished using the Quiver algorithm. Based on sequence accuracy, resulting polished consensus sequences were divided into high- (hq) and low-quality (lq) sequences. As a result, 39 092 hq sequences were defined. Length profiles of the sequences at the different processing steps are given in Figure 2.

The coding domain sequences (CDSs) of the 39 092 hq sequences were predicted by BLAST and ESTscan analyses using the current releases of the Swiss-Prot (https://www.uniprot.org/) and NR (https://www.ncbi.nlm.nih.gov/) databases. Functional annotation of all sequences was carried out using seven databases (see the Experimental Procedures section; Data S1). Accession numbers were assigned to all sequences, which range from VaGs00001 to VaGs39092 (VaGs, *V. album* gene space). For ease of use, a table was prepared which includes all hq nucleotide sequences, their accession numbers, sequence length

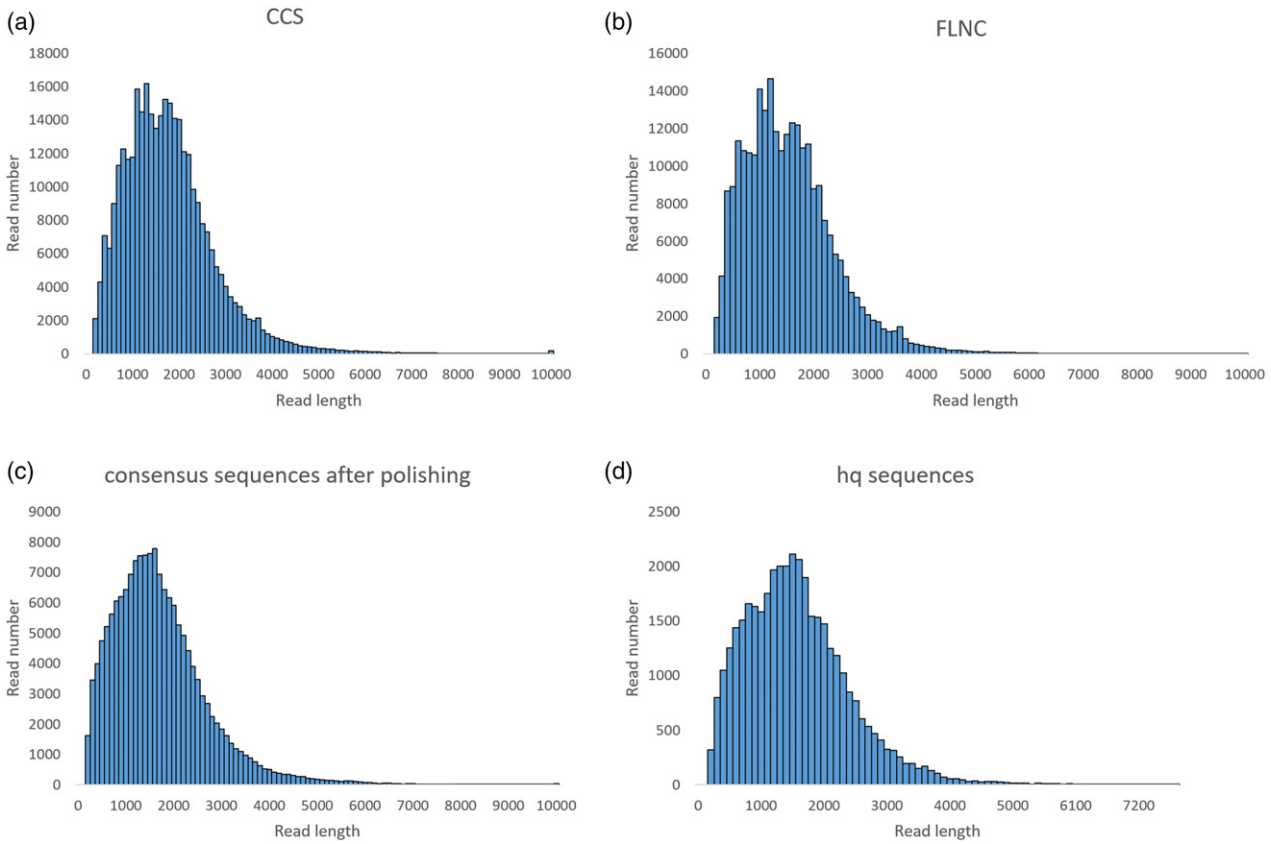

Figure 2. Length profiles of *V. album* sequences defined along the different data processing steps. (a) Length profile of circular consensus (CCS) sequences. (b) Length profile of full-length non-chimeric (FLNC) sequences. (c) Length profile of the consensus sequences after polishing. (d) Length profile of the 39 092 high-quality (hq) sequences. See Table 1 and Figure 1 for information on the processing steps.

information and encoded amino acid sequences as well as information on functional predictions (*V. album* gene space, Data S2). For all accessions, the table also includes the most similar protein of the model plant *A. thaliana*. Finally, a more focused table is presented that lists all *V. album* proteins encoded by the 39 092 hq sequences (*V. album* protein space, Data S3). The overall number of distinct proteins is 32 064, because some of the hq DNA sequences slightly differ but encode proteins with identical amino acid sequences. This might indicate the presence of isogenes and/or allelic variation.

The completeness of the presented gene space with respect to the entire *V. album* transcriptome lies at approximately 78%, as revealed by an evaluation using the Benchmarking Universal Single-Copy Orthologs (BUSCO) software (Seppey *et al.*, 2019; Figure 3).

Transcriptome properties in *V. album*

The sequences of the 39 092 full-length transcripts offer new insights into the transcriptome structure and composition of *V. album*. Overall, the GC content of the coding regions within *V. album* transcripts lies at 50.0%, which is well above the 39.4% determined before by cytometric analyses of the entire *V. album* genome (Marie and Brown, 1993). An increased GC content in coding regions in comparison to intergenic regions is common in plants. For instance, the GC content of coding regions in *A. thaliana* is 44%, but it is only 34% in non-coding regions (Arabidopsis Genome Initiative, 2000). Similarly, the GC content of coding regions of several other dicotyledonous plants is in the range of 43–45% (Singh *et al.*, 2016). Within the clade of dicotyledonous plants, the GC content of *V. album* is strikingly high. A high GC content has positive effects on genome stability but comes at the price of increased energy demand for transcription and genome replication, which both require opening of the double helix.

The codon usage in *V. album* does not differ fundamentally from that in *A. thaliana*, with a few exceptions. For example, the CCC codon (which encodes proline) has a frequency of 12.9/1000 codons in *V. album* but only 5.2/1000 codons in *A. thaliana*; similarly, the GGG codon (which encodes glycine) has a frequency of 20.1/1000 codons in *V. album* but only 10.2/1000 codons in *A. thaliana* (Table S2). Overall, several codons of increased abundance in *V.*

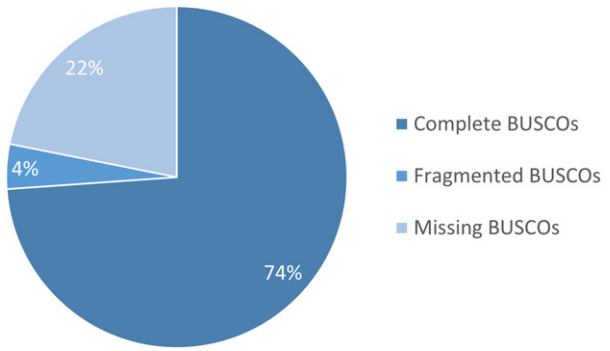

Figure 3. Completeness of the *V. album* gene space as revealed by 'Benchmarking Universal Single-Copy Orthologs' (BUSCO) analysis (Seppey *et al.*, 2019).

album are rich in G and/or C, which contributes to the increased GC content of transcripts in *V. album*.

Proteome properties in *V. album*

The proteins encoded by the *V. album* gene space have an average molecular mass of 40.4 kDa (Figure 4). The average molecular mass of proteins encoded by the *A. thaliana* genome has been reported to be 45.9 kDa based on evaluation of The Arabidopsis Information Resource (TAIR) 7 genome release (Baerenfaller *et al.*, 2008). Recalculation of the average molecular mass of proteins in *A. thaliana* using the TAIR10 genome release revealed an average molecular mass of 45.8 kDa. Hence, the average molecular mass of *V. album* proteins is slightly lower. However, this result has to be treated with caution because we cannot rule out the possibility that a low percentage of transcripts in the *V. album* gene space code for incomplete proteins, which would affect our calculation. At the same time, the overall similar average molecular mass of the proteins in *V. album* and *A. thaliana* can be taken as evidence that a high percentage of our *V. album* transcripts can be considered to be complete.

The average isoelectric point (IEP) of *V. album* proteins encoded by our gene space is 7.43. However, a plot of IEPs of all *V. album* proteins encoded by our gene space shows a bimodal IEP distribution with two peaks at pH 5.8 and 9.2 and a prominent minimum at pH 7.5 (Figure 4). This IEP distribution has been reported before for several other species, including *A. thaliana* (Kiraga *et al.*, 2007; Schwartz *et al.*, 2001; van Wijk *et al.*, 2021), and is interpreted to reflect that solubility of proteins in aqueous solutions is low close to their isoelectric points. The hydrophobicity of the *V. album* proteins peaks at a GRAVY value of −0.2, which again is similar to the value calculated for *A. thaliana* (−0.3 based on analyses using the TAIR10 genome release; Figure 4).

To estimate the average amino acid identity between proteins from *V. album* and *A. thaliana*, sequence comparisons were carried out for selected proteins (Table 2). As

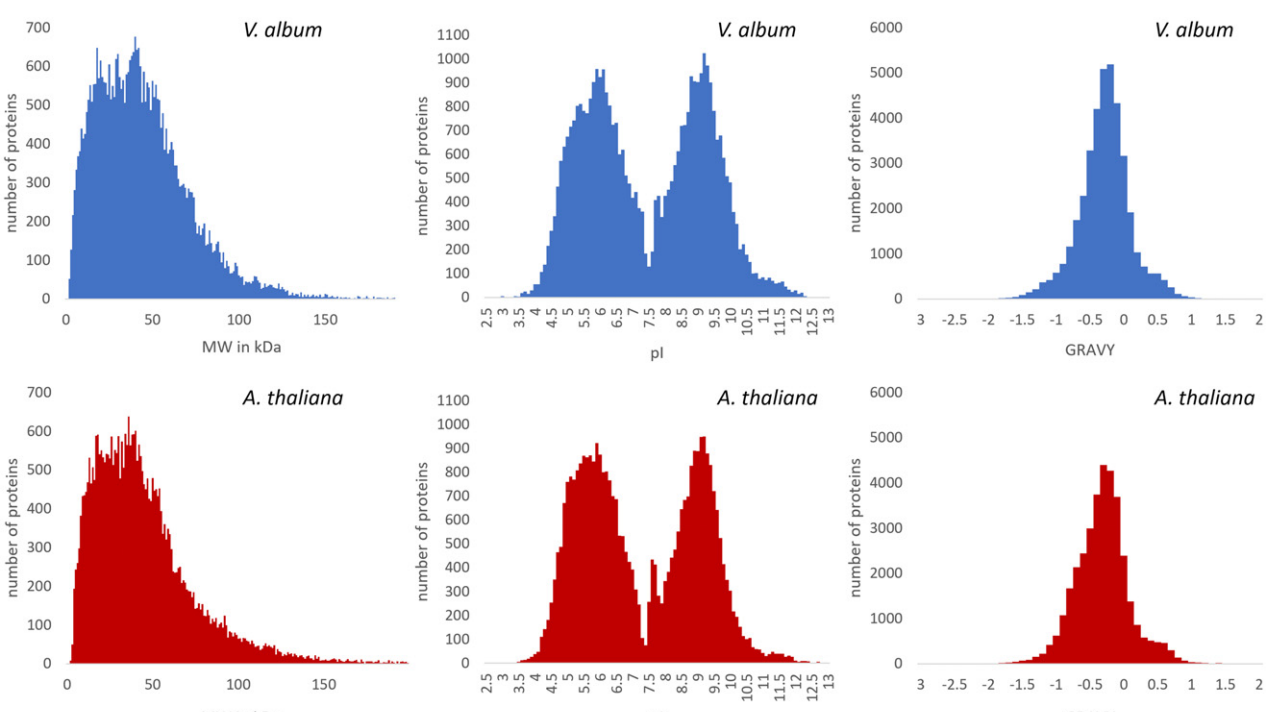

Figure 4. Physicochemical properties of proteins from *V. album* and *A. thaliana*.

expected, some proteins like histones are highly conserved (99%), whereas others are more divergent (e.g., the small subunit of ribulose bisphosphate carboxylase/oxygenase, 65%). Proteins involved in cellular respiration like phosphofructokinase 5 (glycolysis), citrate synthase (tricarboxylic acid cycle) and cytochrome *c* (mitochondrial respiratory chain) exhibit approximately 80% sequence identity between *V. album* and *A. thaliana*. On average, sequence identity between the two plant species lies in the range of 75%.

Transcripts encoding lectins and viscotoxins

Viscum album contains characteristic lectins as well as amphiphilic micro-proteins called viscotoxins (Nazaruk and Orlikowski, 2016). Both classes of proteins are subjects of considerable attention because they contribute to cytotoxic and immune-stimulating effects of the mistletoe extracts used in medicine. Three types of *V. album* lectins were biochemically and structurally characterized, termed mistletoe lectins I, II and III (MLI, MLII and MLIII) (Krauspenhaar *et al.*, 2002; Niwa *et al.*, 2003). All three types of lectins are synthesized as precursor proteins and post-translationally cleaved into an alpha and a beta chain. The two chains are linked via a disulfide bridge. The beta chain has lectin activity and specifically binds to sugar residues of membrane proteins, thereby inducing its endocytic uptake by target cells. The alpha subunit has RNA glycosidase activity. It has been shown to cleave off the adenine of nucleotide A4325 of the 28S ribosomal RNA, thereby inactivating the ribosome and inducing apoptosis (Endo and Tsurugi, 1987).

The *V. album* gene space includes full-length sequences encoding the precursors of MLI (VaGs17673), MLII (VaGs17667) and MLIII (VaGs17674). Sequence conservation between the three proteins is in the range of 76–81% (Figure 5). They highly resemble sequences determined previously for MLI, MLII and MLIII (Kourmanova *et al.*, 2004; Sudarkina *et al.*, 2007) but are not identical (Table 3). Possibly, the previously determined *V. album* gene sequences are from a different *V. album* subspecies or regional variants, which is likely since these studies were not performed with a standardized model line. Alternatively, the *V. album* genome may contain isogenes and/or alleles encoding additional lectin isoforms. Indeed, our gene space includes five further transcripts, which all are highly similar to MLI but slightly shorter. More targeted investigations on the genomic level will be needed to fully characterize the *V. album* lectin gene family.

Viscotoxins have a molecular mass of about 5 kDa. Like the *V. album* lectins, they are synthesized as larger precursors, which are proteolytically processed. Their 3D structures are stabilized by the formation of disulfide bridges. Viscotoxins are functionally less defined but are considered to bind to bio-membranes. Five different transcripts encoding viscotoxin precursors are included in our gene space: VaGs38671, VaGs35165, VaGs38197, VaGs36524 and VaGs36525 (Figure 6). They resemble viscotoxin sequences determined previously (e.g., viscotoxin A3, Swiss-Prot entry P01538) but again are not identical to previously published sequences (Figure S1). Further genes encoding viscotoxins might be transcribed in *V. album* berries, which

Table 2 Sequence identity of selected proteins from *Viscum album* and *Arabidopsis thaliana*

Protein[a]	Accession[b]	Sequence length[c]	Sequence range compared[d]	Identical amino acids[e]	Identity in %[f]
RuBisCO small chain	At1g67090 (At)	180	180	117	65.0
	VaGs21968 (Va)	191			
cytochrome *c*	At5g40810 (At)	307	273	217	79.5
	VaGs25594 (Va)	381			
phosphofructokinase 5	At2g22480 (At)	537	497	395	79.5
	VaGs15964 (Va)	547			
citrate synthase	At2g44350 (At)	474	470	376	80.0
	VaGs13003 (Va)	474			
histone H4	At1g07660 (At)	103	92	91	98.9
	VaGs37578 (Va)	92			
PsaD	At1g03130 (At)	204	202	144	71.3
	VaGs36078 (Va)	224			
Average (Ø)		310	286	223	79.0

[a]Protein names.
[b]Accession numbers according to our *Viscum album* gene space database or The Arabidopsis Information Resource (TAIR, www.arabidopsis.org); At, *A. thaliana*; Va, *V. album*.
[c]Sequence length of the entire protein (number of amino acids).
[d]Length of the sequences compared.
[e]Number of identical amino acids.
[f]Sequence identity.

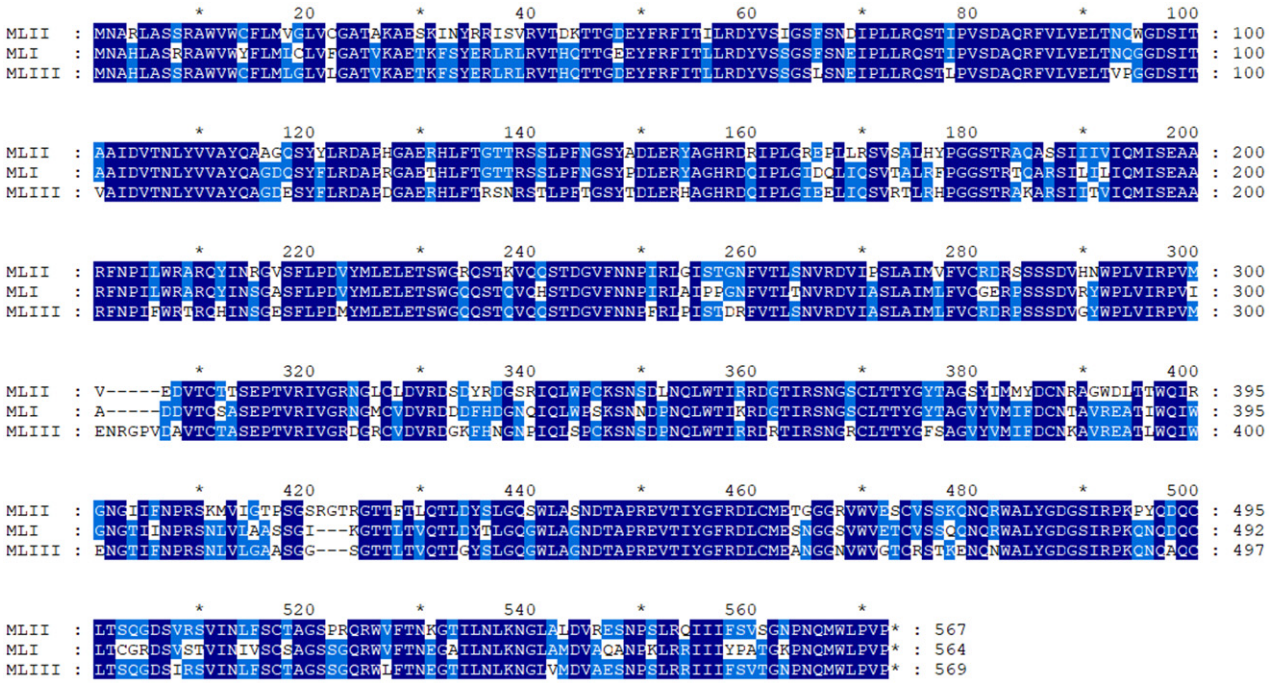

Figure 5. Alignment of the *V. album* lectins MLI (VaGs17673), MLII (VaGs17667) and MLIII (VaGs17674). Dark blue amino acid positions are conserved in all three sequences; medium blue amino acid positions are conserved in two of the three sequences.

Table 3 *Viscum album* lectin sequences from the *V. album* gene space and from UniProt

Protein[a]	Accession[b]	Sequence length[c]	Sequence range compared[d]	Identical amino acids[e]	Identity in %[f]
MLI	VaGs17673	564	564	556	98.6%
	P81446	564			
MLII	VaGs17667	567	567	560	98.8%
	Q6H266	567			
MLIII	VaGs17674	569	569	502	88.2%
	P82683	569			

[a]Protein names. MLI, mistletoe lectin I; MLII, mistletoe lectin II; MLIII, mistletoe lectin III.
[b]Accession numbers according to our *V. album* gene space database or UniProt.
[c]Sequence length of the entire protein (number of amino acids).
[d]Length of the sequences compared.
[e]Number of identical amino acids.
[f]Sequence identity in percent.

were not included in our starting material for SMRT sequencing.

Transcripts encoding *V. album* proteins localized in the mitochondria

Viscum album has unusual mitochondria. Its mitochondrial genome lacks some genes encoding proteins considered to be essential for mitochondrial function, most notably subunits of the NADH dehydrogenase complex (complex I) of the respiratory chain (Petersen *et al.*, 2015a; Skippington *et al.*, 2015, 2017). While it initially could not be excluded that the corresponding genes had been overlooked (due to far-going sequence alterations during evolution or translocation of sequences to the nuclear genome), it later became clear by proteome analyses of isolated mitochondria that complex I indeed is absent in *V. album* (Maclean *et al.*, 2018; Senkler *et al.*, 2018). This was a surprising finding, because it is the only known example of a multicellular eukaryote that can carry out cellular respiration in the absence of complex I (Busch, 2018; da Fonseca-Pereira *et al.*, 2018). In *V. album*, complexes III and IV of the respiratory chain form stable supercomplexes; furthermore,

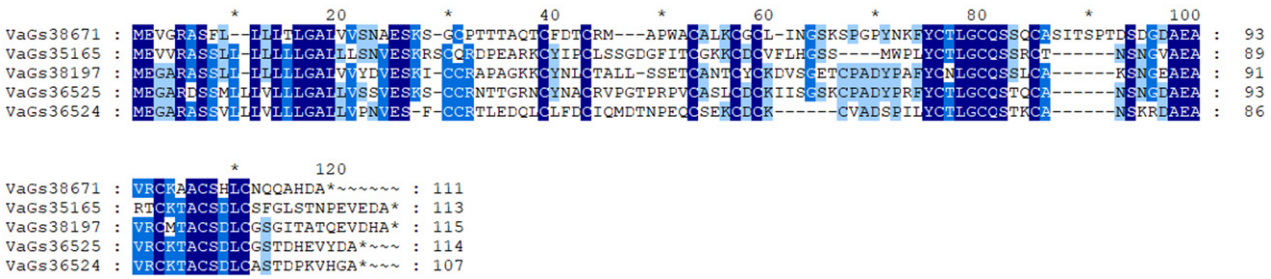

Figure 6. Alignment of the viscotoxin precursors encoded by VaGs38671, VaGs35165, VaGs38197, VaGs36525 and VaGs36524. Dark blue amino acid positions are conserved in all sequences; medium blue amino acid positions are conserved in four of the five sequences; light blue amino acid positions are conserved in three of the five amino acid positions.

numerous alternative oxidoreductases occur (Maclean *et al.*, 2018; Senkler *et al.*, 2018).

Proteome analyses of *V. album* mitochondria were so far greatly hindered due to the very limited genome sequence information for *V. album* or any other species of the *Viscum* genus or the Santalaceae family (which includes the *Viscum* genus and several related genera). Specifically, mass values of tryptic peptides from *V. album* proteins obtained by mass spectrometry could not be matched with peptide sequences encoded by the corresponding genes. In an attempt to evaluate the quality of our *V. album* gene space database, we therefore re-evaluated a mitochondrial proteome dataset from *V. album*. The following experiment has been carried out by Senkler *et al.* (2018):

Mitochondria were isolated from *V. album* leaves, mitochondrial membranes were solubilized with the detergent digitonin and the resulting protein fraction was separated by 2D Blue-native/SDS-PAGE. The result of the electrophoretic separation was visualized by staining the gel using Coomassie blue. The most prominent 182 protein spots were excised and trypsinized, and the masses of the tryptic peptides were subsequently determined by mass spectrometry. Due to the lack of *V. album* genome information, the data had to be evaluated using the genome sequence of the model plant *A. thaliana* (TAIR10 genome release) and the few *V. album* sequences available at NCBI in 2018. This evaluation was very restricted, because only few tryptic peptides are completely conserved between *V. album* and *A. thaliana* (considering about 75% sequence identity, on average 2.5 amino acids are exchanged per peptide of 10 amino acid length, which is about the average length of tryptic peptides). Overall, 3129 peptides could be defined, which were assigned to 427 different mitochondrial proteins (Senkler *et al.*, 2018). The obtained data are accessible as a web-based GelMap project (https://gelmap.de/1327), which offers protein identification information by simply clicking on protein spots of interest (Figure 7a).

Data evaluation of this experiment was now repeated using the sequences of the *V. album* gene space database (Table 4). The new data analysis allowed identification of 11 736 unique peptides (versus 3129 peptides based on TAIR evaluation; +257%). The number of identified proteins increased from 427 to 612 (+43%). Coverage of identified proteins by peptides increased from 7.3 to 19.2 peptides/protein (+163%). A new GelMap has been created (Figure 7b), which is accessible at https://gelmap.de/2274. Comparison of the two GelMaps nicely allows visualization of the much increased identification rate especially of small proteins, which, upon trypsinization, only account for few peptides.

For a more detailed comparison between TAIR10 and *V. album* gene space-based data evaluation, we specifically focused on the mitochondrial oxidative phosphorylation (OXPHOS) system (Figure 8). Blue-native/SDS-PAGE is especially suitable for separating subunits of the protein complexes involved in OXPHOS. Overall, based on the new evaluation, 163 out of 182 analyzed protein spots included at least one OXPHOS protein (75 out of 170 based on the TAIR10 evaluation). The five proteins involved in the ubiquinone biosynthesis pathway were exclusively identified by *V. album* gene space evaluation.

In another attempt to evaluate the completeness of our *V. album* gene space, we directly searched our database for genes encoding subunits of the OXPHOS complexes II, III, IV and V. In *A. thaliana*, these complexes have been characterized in depth and their subunit compositions are well defined (Braun, 2020). Amino acid sequences of all OXPHOS proteins from *A. thaliana* were used to probe the *V. album* gene space database. The gene space includes a close to complete set of nuclear genes encoding OXPHOS proteins (except those encoding subunits of complex I) (Table 5). Interestingly, we also found some OXPHOS sequences transcribed from mitochondrial genes. It originally was anticipated that transcripts of mitochondrial (and chloroplast) genes lack poly(A) tails and therefore are not present in cDNA libraries produced from mRNA (which usually is amplified using poly(T) primers at the 3′ end). However, it later became clear that mitochondrial transcripts in plants can be polyadenylated and that polyadenylation targets organellar transcripts for degradation (Lang *et al.*, 2009; Schuster and Stern, 2009).

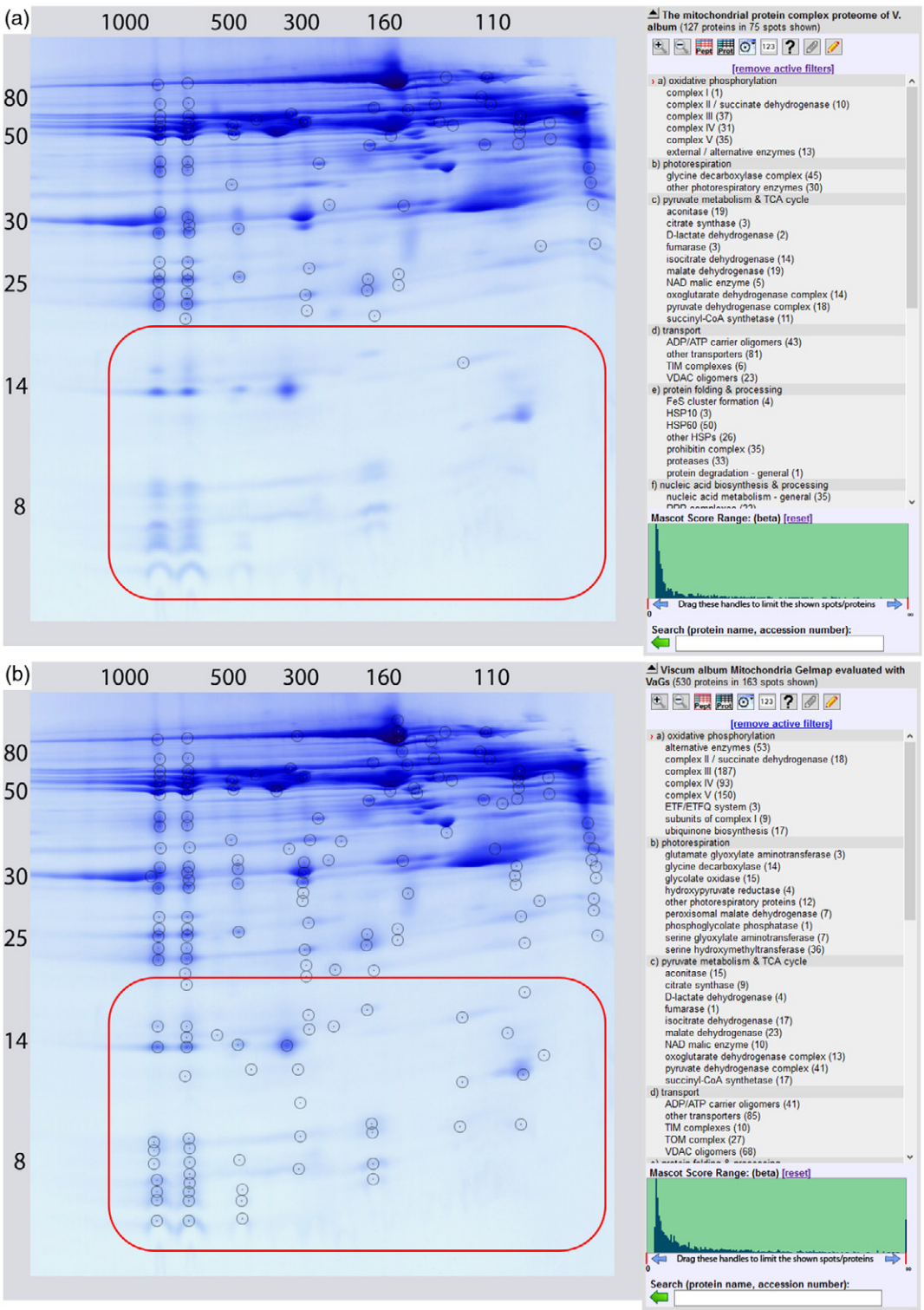

Figure 7. Identified OXPHOS proteins of *V. album* on a 2D Blue-native/SDS-PAGE gel (Senkler *et al.*, 2018). Mass spectrometry data were evaluated using the *A. thaliana* TAIR10 database (a) or the *V. album* gene space database (b). Interactive data presentations are available at https://gelmap.de/1327 (TAIR evaluation) and https://gelmap.de/2274 (*V. album* gene space evaluation). For both parts of the figure, displayed OXPHOS proteins were selected by using the filter menu given to the right. Black circles indicate identified OXPHOS proteins. The red frames on the two 2D gels highlight gel regions containing small proteins (<20 kDa).

Table 4 Numbers of proteins and peptides identified for a mitochondrial fraction of *V. album* as revealed by TAIR10 evaluation (Senkler *et al.*, 2018) and evaluation using the *V. album* gene space (VaGs; this study)

	TAIR10 (2018)[a]	VaGs (2021)[b]
Successfully analyzed protein spots (out of 182)	170	182
Unique peptides	3129	11 736
Proteins identified	1245	2318
Average per spot	*7.3*	*12.7*
Unique proteins	427	612
Average number of peptides per protein	*7.3*	*19.2*
Protein spots with OXPHOS proteins	75	163

[a]Results published in 2018 using the TAIR10 database (https://gelmap.de/1327).
[b]Re-evaluated data using the *V. album* gene space database (this study; https://gelmap.de/2274).

Transcripts encoding subunits of mitochondrial complex I

Arabidopsis thaliana complex I consist of 48 subunits, 39 of which are encoded by the nuclear and nine by the mitochondrial genome (Klusch *et al.*, 2021). The mitochondrial genome of *V. album* lacks the nine genes encoding complex I subunits (Petersen *et al.*, 2015a; Skippington *et al.*, 2015, 2017) and the enzyme complex indeed is absent in the mitochondria as revealed by proteome investigations (Maclean *et al.*, 2018; Senkler *et al.*, 2018). It hence is supposed that all nuclear complex I genes also are absent in *V. album*. This hypothesis was now tested by systematically probing the *V. album* gene space database using complex I sequences from *A. thaliana*.

In contrast to transcripts encoding subunits of complexes II, III, IV and V, transcripts encoding subunits of complex I indeed are absent in our *V. album* gene space. Only a few exceptions were found: a transcript that encodes a complex I-integrated gamma-type carbonic anhydrase (VaGs39093; Table 5) and two transcripts which encode two distinct mitochondrial acyl carrier proteins (VaGs37160/VaGs37159 and VaGs36982; Table 5), which are part of complex I in *A. thaliana* (Klusch *et al.*, 2021).

Peptides of the complex I-integrated gamma-type carbonic anhydrase were also identified in the mitochondrial proteome of *V. album* upon data evaluation using the *V. album* gene space (see GelMap at https://gelmap.de/2274). In plants, a carbonic anhydrase domain is attached to the membrane arm of complex I on its matrix-exposed side (Sunderhaus *et al.*, 2006). It is composed of three gamma-type carbonic anhydrase subunits. The domain binds a metal ion and is considered to be enzymatically active (Klusch *et al.*, 2021). Plant complex I cannot assemble if these proteins are absent (Fromm *et al.*, 2016). A direct role of the gamma-type carbonic anhydrase proteins for complex I function during OXPHOS so far is elusive, but it has been suggested that they might integrate a secondary function into plant complex I, which may be related to photorespiration (Soto *et al.*, 2015). The presence of a transcript of the complex I-integrated gamma-type carbonic anhydrase

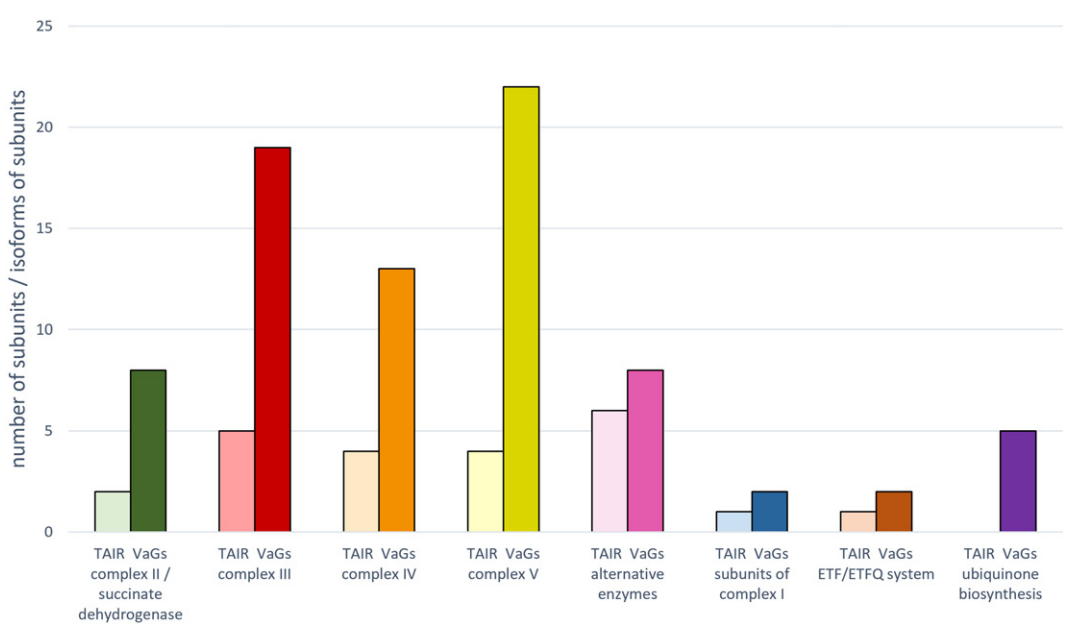

Figure 8. Number of identified OXPHOS components within the mitochondrial proteome dataset of Senkler *et al.*, 2018 upon data evaluation based on TAIR10 (light colors) and the *V. album* gene space (dark colors) database.

Table 5 Proteins involved in the mitochondrial OXPHOS system in *A. thaliana* and *V. album*. Amino acid sequences of all known OXPHOS proteins from *A. thaliana* were used to probe the *V. album* gene space database. Selection of proteins of *A. thaliana* is based on Braun, 2020, but was extended by recently identified additional OXPHOS proteins (Maldonado *et al.*, 2021 for complex IV, Zancani *et al.*, 2020 for complex V and Klusch *et al.*, 2021 for complex I). The data for *V. album* are based on the *V. album* protein space (this publication). Mitochondrially (mt) encoded proteins are partially not included in the *V. album* gene space (but supplemented from NCBI in cases they have been annotated previously)

Annotation	Functional category	No.	Accession(s) in *A. thaliana*	Accession(s) in *V. album*
Complex II				
SDH1	complex II	1	At5g66760, At2g18450	VaGs15504, VaGs18431, VaGs15593, VaGs15594, VaGs15446, VaGs15505
SDH2	complex II	2	At3g27380, At5g40650, At5g65165	VaGs29615, VaGs30174, VaGs34595
SDH3	complex II	3	At5g09600, At4g32210	(mt encoded)
SDH4	complex II	4	At2g46505	(mt encoded)
SDH5	complex II	5	At1g47420	VaGs33410, VaGs33411
SDH6	complex II	6	At1g08480	VaGs36736, VaGs34875, VaGs36436
SDH7	complex II	7	At3g47833, At5g62575	VaGs38038
SDH8	complex II	8	At2g46390	---
Complex III				
cytochrome *b*	complex III	9	AtMg00220	(mt encoded); YP_009220377.1
cytochrome *c*1	complex III	10	At5g40810, At3g27240	VaGs25595, VaGs25661, VaGs24535, VaGs24536, VaGs25594
FeS	complex III	11	At5g13430, At5g13440	VaGs24540, VaGs24539, VaGs24541, VaGs33119
MPPalpha	complex III	12	At1g51980, At3g16480	VaGs13131, VaGs29795, VaGs13256, VaGs00721
MPPbeta	complex III	13	At3g02090	VaGs04283, VaGs01538, VaGs04991
QCR10	complex III	14	At2g40765	VaGs17923
QCR6	complex III	15	At1g15120, At2g01090	VaGs23198, VaGs35349, VaGs23201
QCR7	complex III	16	At4g32470, At5g25450	VaGs39075, VaGs39072, VaGs38385
QCR8	complex III	17	At3g10860, At5g05370	VaGs35998
QCR9	complex III	18	At3g52730	VaGs36144, VaGs03304, VaGs36125, VaGs02923, VaGs39020, VaGs36037, VaGs36038
Complex IV				
COX1	complex IV	19	AtMg01360	VaGs08501, YP_009220376.1
COX2	complex IV	20	AtMg00160	VaGs31628, YP_009220375.1
COX3	complex IV	21	AtMg00730	(mt encoded); YP_009220379.1
COX4 (=COX-X2)	complex IV	22	At4g00860, At1g01170	VaGs37231
COX5b	complex IV	23	At3g15640, At1g80230	VaGs29247, VaGs29249, VaGs29246, VaGs29248
COX5c	complex IV	24	At2g47380, At3g62400, At5g61310	VaGs03020
COX6a	complex IV	25	At4g37830	VaGs37400
COX6b	complex IV	26	At1g22450	VaGs38369, VaGs33922, VaGs33924, VaGs35619, VaGs33923
COX7a (=COX-X4)	complex IV	27	At4g21105	VaGs34759, VaGs34760, VaGs36004
COX7b (=COX-X3)	complex IV	28	At1g72020	VaGs34763, VaGs34761
Complex V (ATP Synthase)				
alpha subunit	complex V	29	AtMg01190	VaGs21852, YP_009220384.1
beta subunit	complex V	30	At5g08670, At5g08680, At5g08690	VaGs14513, VaGs19745, VaGs15576, VaGs14512, VaGs15180, VaGs19744, VaGs14514
gamma subunit	complex V	31	At2g33040	VaGs09754, VaGs09752, VaGs36320
delta subunit	complex V	32	At5g47030	VaGs25731, VaGs25732
epsilon subunit	complex V	33	At1g51650	VaGs38057, VaGs38055, VaGs38056, VaGs38054
subunit a (=ATP6)	complex V	34	AtMg00410, AtMg01170	VaGs28647, YP_009220378.1

(continued)

Table 5. (continued)

Annotation	Functional category	No.	Accession(s) in *A. thaliana*	Accession(s) in *V. album*
subunit b	complex V	35	AtMg00640	(mt encoded)
subunit c (=ATP9)	complex V	36	AtMg01080	(mt encoded)
subunit d	complex V	37	At3g52300	VaGs20371, VaGs20376, VaGs20237
subunit e (=ATP21)	complex V	38	At5g15320	VaGs37459
subunit f (=ATP17)	complex V	39	At4g30010	VaGs34973, VaGs34978,
subunit g (=ATP20)	complex V	40	At2g19680, At4g26210, At4g29480	VaGs33118, VaGs38387, VaGs33861, VaGs38389, VaGs38808, VaGs33863, VaGs38388
FAD (24 kDa)	complex V	41	At2g21870	VaGs32332
Inhibitory factor	complex V	42	At2g27730, At5g04750	VaGs23857, VaGs36801
OSCP	complex V	43	At5g13450	VaGs25651, VaGs25474, VaGs25652, VaGs25473
subunit 8	complex V	44	AtMg00480	(mt encoded)
6 kDa subunit	complex V	45	At5g59613, At3g46430	VaGs37919, VaGs37922, VaGs37923, VaGs37590
Complex I				
13 kDa subunit	complex I	46	At3g03070	---
15 kDa subunit	complex I	47	At3g62790, At2g47690	---
18 kDa subunit	complex I	48	At5g67590	---
24 kDa subunit	complex I	49	At4g02580	---
39 kDa subunit	complex I	50	At2g20360	---
51 kDa subunit	complex I	51	At5g08530	---
75 kDa subunit	complex I	52	At5g37510	---
AGGG	complex I	53	At1g76200	---
ASHI	complex I	54	At5g47570	---
B12	complex I	55	At1g14450, At2g02510	---
B13	complex I	56	At5g52840	---
B14	complex I	57	At3g12260	---
B14.5a	complex I	58	At5g08060	---
B14.5b	complex I	59	At4g20150	---
B14.7	complex I	60	At2g42210	---
B15	complex I	61	At2g31490	---
B16.6	complex I	62	At1g04630, At2g33220	---
B17.2	complex I	63	At3g03100	---
B18	complex I	64	At2g02050	---
B22	complex I	65	At4g34700	---
B8	complex I	66	At5g47890	---
B9	complex I	67	At2g46540	---
CA1	complex I	68	At1g19580	VaGs39093
CA2	complex I	69	At1g47260	---
CA3	complex I	70	At5g66510	---
CAL1/CAL2	complex I	71	At5g63510, At3g48680	---
ESSS	complex I	72	At2g42310, At3g57785	---
C1-FDX	complex I	73	At3g07480	---
GLDH	complex I	74	At3g47930	VaGs17436, VaGs17437
KFYI	complex I	75	At4g00585	---
MNLL	complex I	76	At4g16450	---
MWFE	complex I	77	At3g08610	---
ND1	complex I	78	AtMg00516	---
ND2	complex I	79	AtMg00285	---
ND3	complex I	80	AtMg00990	---
ND4	complex I	81	AtMg00580	---
ND4L	complex I	82	AtMg00650	---
ND5	complex I	83	AtMg00665	---
ND6	complex I	84	AtMg00270	---
ND7	complex I	85	AtMg00510	---
ND9	complex I	86	AtMg00070	---

(continued)

Table 5. (continued)

Annotation	Functional category	No.	Accession(s) in *A. thaliana*	Accession(s) in *V. album*
P1	complex I	87	At1g67350	---
P2	complex I	88	At2g27730	VaGs23857
PDSW	complex I	89	At3g18410, At1g49140	---
PGIV	complex I	90	At3g06310, At5g18800	---
PSST	complex I	91	At5g11770	---
SDAP-1	complex I	92	At2g44620	VaGs37160, VaGs37159
SDAP-2	complex I	93	At1g65290	VaGs36982
SGDH1	complex I	94	At1g67785	---
TYKY	complex I	95	At1g79010, At1g16700	---
Alternative respiratory enzymes				
AOX1A, AOX1B, AOX1C, AOX1D, AOX2	AOX	96-100	At3g22370, At3g22360, At3g27620, At1g32350, At5g64210	VaGs06791, VaGs06681, VaGs06620, VaGs06230, VaGs06621
NDA1	altNDH	101	At1g07180	VaGs26116, VaGs27510
NDA2	altNDH	102	At2g29990	---
NDB1	altNDH	103	At4g28220	VaGs10450, VaGs10451, VaGs08998
NDB2	altNDH	104	At4g05020	VaGs18309, VaGs18311, VaGs19303, VaGs19304, VaGs19328, VaGs19364, VaGs19366, VaGs19369
NDB3	altNDH	105	At4g21490	---
NDB4	altNDH	106	At2g20800	---
NDC1	altNDH	107	At5g08740	VaGs17536, VaGs17959, VaGs17961, VaGs17962
Cytochrome *c*				
Cytc	Cytc	108	At4g10040, At1g22840	VaGs34883, VaGs34884
Other enzymes contributing electrons to the respiratory chain				
Electron transfer flavoprotein α	ETF	109	At1g50940	VaGs28996, VaGs28997
Electron transfer flavoprotein β	ETF	110	At5g43430	VaGs31920
ETF:ubiquinone oxidoreductase	ETFQO	111	At2g43400	VaGs15522
D-Lactate dehydrogenase	D-LDH	112	At5g06580	VaGs18761
Proline dehydrogenase 1	ProDH	113	At3g30775, At5g38710	VaGs18295
Glyceraldehyde 3-phosphate dehydrogenase	GPDH	114	At3g10370	VaGs10872
Dihydroorotate dehydrogenase	DHODH	115	At5g23300	VaGs16765, VaGs16766

in *V. album* now strongly supports a secondary role of this protein in the mitochondria of plants.

Two distinct acyl carrier subunits are part of complex I in fungi and animals (Dobrynin *et al.*, 2010; Runswick *et al.*, 1991). They carry a fatty acid and are essential for assembly and stability of complex I. In *A. thaliana*, three acyl carrier proteins are present in the mitochondrial matrix and assumed to be involved in mitochondrial fatty acid biosynthesis (Meyer *et al.*, 2007). Two of them, termed SDAP-1 and SDAP-2 (or mtACP1 and mtACP2), were recently found to be subunits of plant complex I (Klusch *et al.*, 2021). The presence of homologous transcripts in *V. album* again indicates an essential secondary function of these complex I subunits, probably in mitochondrial fatty acid biosynthesis.

Finally, our *V. album* gene space includes a transcript encoding L-galactono-1,4-lactone dehydrogenase (GLDH). This protein is localized in the mitochondrial intermembrane space and catalyzes the terminal step of the mitochondrial ascorbate biosynthesis pathway (Bartoli *et al.*, 2000). At the same time, this protein has been found to catalyze complex I assembly in plants (Pineau *et al.*, 2008; Schertl *et al.*, 2012; Schimmeyer *et al.*, 2016). However, GLDH is not considered to be a complex I subunit because it does not form part of the holo complex (Soufari *et al.*, 2020). In *V. album*, GLDH is considered to be in charge in ascorbic acid biosynthesis.

We conclude that transcripts encoding complex I subunits are absent in *V. album*, except for transcripts of a few bifunctional complex I components: a gamma-type carbonic anhydrase, two acyl carrier subunits and GLDH.

Concluding remarks

We present a *V. album* gene space comprising 39 092 transcripts. This considerably extends our knowledge on the genome of *V. album*. Currently (July 12th, 2021), 270 *V. album* proteins are annotated at NCBI, in comparison to

35 386 proteins of the model plant *A. thaliana* (TAIR10 database; release July 11th, 2019). The *V. album* protein space now comprises 32 064 proteins. Coverage of the *V. album* gene space with respect to the total coding capacity is estimated to be in the range of 78% as revealed by BUSCO analysis. The more abundant enzymes related to primary metabolism should be covered almost completely, which is supported by the evaluation of transcripts encoding components of the OXPHOS system. The *V. album* gene space is accessible at NCBI (BioProject ID: PRJNA765163). Its further evaluation will offer new insights into the molecular biology of a very unusual flowering plant.

EXPERIMENTAL PROCEDURES

Plant material and sample preparation

Mistletoes (European mistletoe, *V. album*) grown on an apple tree (*Malus* sp.) on our university campus (Leibniz Universität Hannover, Herrenhäuser Str. 2, Hannover, Germany; Figure S2) were harvested in July 2019 (summer sample) and January 2020 (winter sample). Leaves, stems and flower buds of male and female plants were used. Directly after harvesting, the plant material was shock-frozen using liquid nitrogen and stored at −80°C until use.

RNA sample preparation

Frozen plant material (50 μg per sample) was pulverized using a swinging mill pre-chilled with liquid nitrogen. Isolation of total RNA was carried out using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany), including DNase I treatment, as described by Hohnjec *et al.* (2015). In the final purification step, RNA fractions were eluted using 50 μl RNase-free water. RNA samples were kept at −80°C until use.

RNA quality control

All RNA samples went through quality control procedures to determine the concentration, purity and integrity of the RNA. In addition to agarose gel electrophoresis, RNA quality control was based on Qubit fluorometer (Thermo Fischer, Waltham, MA, USA), Nanodrop spectrophotometer (Thermo Fischer), and Bioanalyzer measurements (Agilent, Santa Clara, CA, USA), all according to the manufacturers' instructions. The final quality values determined prior to cDNA synthesis are summarized in Table S1.

cDNA synthesis

After quality control, RNA samples (summer/winter) were pooled at a ratio of 1:1. Five micrograms of the pooled RNA sample was used for cDNA synthesis using the Clontech SMARTer PCR cDNA Synthesis Kit (Takara Bio Inc., Kusatsu, Japan). At the final step, single-stranded cDNA was PCR-amplified to generate double-stranded cDNA, according to a protocol used by Novogene (Cambridge, UK).

Library preparation and SMRT sequencing

Two SMRTbell libraries were constructed using the PacBio SMRTbell® Template Prep Kit 1.0 (Pacific Biosciences, Menlo Park, CA, USA). SMRT sequencing was performed with the PacBio Sequel System using the Sequel® Binding Kit 3.0 Insert Kit (Pacific Biosciences).

Processing of SMRT reads

After SMRT sequencing, data analysis steps were carried out as outlined in Figure 1. Raw data processing was performed with SMRTlink (version 6.0.0.47841). Subread BAM files were used to generate CCSs by setting minFullPasses = 2 and minPredictedAccuracy = 0.9. For this, the subreads from a single zero-mode waveguide (ZMW) were aligned to each other and afterwards self-corrected. Next, 5′ and 3′ adapters and the poly(A) tail of the CCSs were identified and on that basis they were classified into full-length (containing all three elements) and non-full-length reads. Out of the full-length reads the full-length non-chimeric reads were extracted. During the classification step the poly(A) tails, primers and artificial concatemers (caused by PCR amplification due to the low SMRT adapter concentration) were removed. Using the iterative clustering for error correction (ICE) algorithm, the consensus isoforms from the full-length non-chimeric sequences were identified. Subsequently, the consensus isoforms were polished with the non-full-length reads for a higher accuracy using the Quiver algorithm. The polished consensus isoforms were then divided during the ICE post-process into hq (accuracy $>$ 99%; full-length coverage $\geq$ 2) and low quality (lq) consensus isoforms. For further processing steps (e.g., transcriptome database annotation and CDS prediction) the high quality (hq) consensus isoforms were used.

Prediction of coding sequences – Data S1

BLAST and ESTscan were used for automated prediction of the CDSs of the 39 092 hq sequences. BLAST was used to search for matching consensus sequences of the NR (NCBI non-redundant protein sequences) (https://www.ncbi.nlm.nih.gov/) and Swiss-Prot databases (https://www.uniprot.org/). Matching nucleotide sequences were translated via the standard codon table into amino acid sequences. If BLAST analyses did not allow finding a matching consensus sequence, sequences were re-analyzed with ESTscan (3.0.3) to predict coding regions. For 243 hq sequences, neither the standard, automated BLAST searches nor EST Scan analyses predicted a CDS. In these cases, homology searches were carried out by BLAST searches of six-frame translated reading frames against the current release of the NCBI NR database using CLC Main Workbench (Qiagen Digital Insights, Aarhus, Denmark). However, the identified protein sequences were often very short because the corresponding open reading frames include multiple stop codons. We therefore decided to exclude these sequences from further evaluations, as they rather seem to be pseudogenes. The remaining 38 849 hq sequences of our *V. album* gene space were used to predict the *V. album* protein space.

Functional annotation of transcripts

Functional annotation of all sequences was carried out using seven different databases: NR (NCBI non-redundant protein sequences), NT (NCBI nucleotide sequences), KO (KEGG ORTHOLOG), Swiss-Prot, PFAM (Protein family), GO (Gene Ontology) and KOG (euKaryotic Orthologous Groups). The results of the functional annotation are compiled in Data S1. For defining the *V. album* gene space (see below), functional annotation is mainly based on Swiss-Prot. NR annotation was used to further complement our annotation. Sequences insufficiently defined were re-analyzed by comparison to Viridiplantae sequences at NCBI or further manual sequence evaluation.

The *V. album* gene space – Data S2

The *V. album* gene space includes all hq sequences, GenIDs and gene length information. Novel accession numbers were assigned to all sequences starting with '*V. album* gene space', which range from VaGs00001 to VaGs39092. Furthermore, the amino acid sequences encoded by all genes are given, as well as properties of the corresponding proteins (molecular mass, isoelectric point and hydrophobicity). Finally, functional annotation information is added, as well as information on the most similar protein of the model plant *A. thaliana*. In some cases, several slightly differing genes encode identical proteins. This information is given in column k of Data S2. The resulting number of physically distinct *V. album* protein sequences is 32 064.

The *V. album* protein space – Data S3

The *V. album* protein space includes all 32 064 distinct *V. album* proteins deduced from the *V. album* gene space and information on their functional annotation.

Re-evaluation of proteomic mass spectrometry data using the *V. album* gene and protein space

Mass spectrometry data evaluation and annotation was carried out with ProteinScape (Bruker Daltonics) using an in-house Mascot server (Matrix Science; http://www.matrixscience.com/) for searches of our *V. album* gene space database (for details see Senkler *et al.*, 2018). For selected proteins, searches were additionally carried out using the *V. album* database including polished lq consensus isoforms.

ACKNOWLEDGMENTS

We thank Dr. Sabine Reichwein for project advice, Marius Schrader for taking care of the apple tree used for *V. album* cultivation and Matthias Döring for critical reading of the manuscript. PacBio SMRT sequencing was carried out by Novogene, Cambridge, UK. This research has been supported by the Deutsche Forschungsgemeinschaft (grant BR 1829/16-1 to HPB). Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

HPB initiated the project with advice from HK. *Viscum album* was harvested by LS. RNA isolation and quality evaluation were carried out by NH. Data processing and database development were accomplished by LS with support from MS. Data annotation was performed by LS, JS and MS. HK and HPB supervised the project. LS and HPB wrote the manuscript.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally shareable data necessary to reproduce the reported results. The data are available at https://www.ncbi.nlm.nih.gov/bioproject (BioProject ID: PRJNA765163) and at the GelMap portal (https://gelmap.de/2274).

DATA AVAILABILITY STATEMENT

Viscum album nucleotide sequences (raw data) are available at NCBI (https://www.ncbi.nlm.nih.gov/bioproject; BioProject ID: PRJNA765163). The coding sequences are available at the same BioProject at DDBJ/ENA/GenBank, accession GJLG00000000. Primary protein identification data related to the 2D Blue-native/SDS-PAGE gel presented in Figure 7(b), which were obtained by mass spectrometry, are accessible at the GelMap portal at https://gelmap.de/2274.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Alignment of VaGs36525 with viscotoxin A3 from Swiss-Prot (P01538).

Figure S2. *Viscum album* growing on an apple tree at the campus of Leibniz University Hannover.

Table S1. Quality control of the pooled RNA sample, which was used for library preparation and SMRT sequencing.

Table S2. Codon usage in *V. album* and *A. thaliana* (frequency per 1000 codons).

Table S3. Proteins involved in the mitochondrial OXPHOS system in *A. thaliana* and *V. album*.

Data S1. Prediction of *V. album* coding sequences.

Data S2. The *V. album* gene space: nucleotide sequences of 39 093 *V. album* transcripts and their functional annotation.

Data S3. The *V. album* protein space: amino acid sequences of 32 064 *V. album* proteins and their functional annotation.

REFERENCES

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.

Azuma, Ji., Kim, NH., Heux, L., Vuong, R. & Chanzy, H. (2000) The cellulose system in viscin from mistletoe berries. *Cellulose*, **7**, 3–19. https://doi.org/10.1023/A:1009223730317

Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S. *et al.* (2008) Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science*, **320**(5878), 938–941. https://doi.org/10.1126/science.1157956

Barlow, B.A. (1981) *Viscum album* in Japan: Chromosomal translocations, maintenance of heterozygosity and the evolution of dioecy. *The Botanical Magazine Tokyo*, **94**, 21–34. https://doi.org/10.1007/BF02490200

Bartoli, C.G., Pastori, G.M. & Foyer, C.H. (2000) Ascorbate biosynthesis in mitochondria is linked to the electron transport chain between complexes III and IV. *Plant Physiology*, **123**(1), 335–344. https://doi.org/10.1104/pp.123.1.335

Braun, H.P. (2020) The Oxidative Phosphorylation system of the mitochondria in plants. *Mitochondrion*, **53**, 66–75. https://doi.org/10.1016/j.mito.2020.04.007

Busch, K.B. (2018) Respiration: life without complex I. *Current Biology*, **28**(10), R616–R618. https://doi.org/10.1016/j.cub.2018.04.030

da Fonseca-Pereira, P., Silva, W.B., Araújo, W.L. & Nunes-Nesi, A. (2018) How does European mistletoe survive without complex I? *Trends in Plant Science*, **23**(10), 847–850. https://doi.org/10.1016/j.tplants.2018.07.008

Dobrynin, K., Abdrakhmanova, A., Richers, S., Hunte, C., Kerscher, S. & Brandt, U. (2010) Characterization of two different acyl carrier proteins in complex I from *Yarrowia lipolytica*. *Biochimica Et Biophysica Acta*, **1797** (2), 152–159. https://doi.org/10.1016/j.bbabio.2009.09.007

Elliott, T.A. & Gregory, T.R. (2015) What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **370** (1678), 20140331. https://doi.org/10.1098/rstb.2014.0331

Endo, Y. & Tsurugi, K. (1987) RNA N-glycosidase activity of ricin A-chain. *Journal of Biological Chemistry*, **262**, 8128–8130. PMID: 3036799.

Fromm, S., Braun, H.P. & Peterhansel, C. (2016) Mitochondrial gamma carbonic anhydrases are required for complex I assembly and plant reproductive development. *New Phytologist*, **211**(1), 194–207. https://doi.org/10.1111/nph.13886

Glatzel, G. & Geils, B.W. (2009) Mistletoe ecophysiology: host–parasite interactions. *Botany-Botanique*, **87**, 10–15. https://doi.org/10.1139/B08-096

Hohnjec, N., Czaja-Hasse, L.F., Hogekamp, C. & Küster, H. (2015) Preannouncement of symbiotic guests: transcriptional reprogramming by mycorrhizal lipochitooligosaccharides shows a strict co-dependency on the GRAS transcription factors NSP1 and RAM1. *BMC Genomics*, **16**, 994. https://doi.org/10.1186/s12864-015-2224-7

Jäger, T., Holandino, C., Melo, M.N.O., Peñaloza, E.M.C., Oliveira, A.P., Garrett, R. *et al.* (2021) Metabolomics by UHPLC-Q-TOF reveals host tree-dependent phytochemical variation in *Viscum album* L. *Plants*, **10**(8), 1726. https://doi.org/10.3390/plants10081726

Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N. *et al.* (2007) The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics*, **8**, 163. https://doi.org/10.1186/1471-2164-8-163

Klusch, N., Senkler, J., Yildiz, Ö., Kühlbrandt, W. & Braun, H.P. (2021) A ferredoxin bridge connects the two arms of plant mitochondrial complex I. *The Plant Cell*, **33**, 2072–2091. https://doi.org/10.1093/plcell/koab092

Ko, S.M., Kwon, Y.K., Kim, J.H., Song, I.J., Lee, H.Y., Choi, D.W. *et al.* (2014) Transcriptome analysis of mistletoe (Viscum album) haustorium development. *Horticulture, Environment, and Biotechnology*, **55**, 352–361. https://doi.org/10.1007/s13580-014-0033-6

Kourmanova, A.G., Soudarkina, O.J., Olsnes, S. & Kozlov, J.V. (2004) Cloning and characterization of the genes encoding toxic lectins in mistletoe (*Viscum album L*). *European Journal of Biochemistry*, **271**(12), 2350–2360. https://doi.org/10.1111/j.1432-1033.2004.04153.x

Krauspenhaar, R., Rypniewski, W., Kalkura, N., Moore, K., DeLucas, L., Stoeva, S. *et al.* (2002) Crystallisation under microgravity of mistletoe lectin I from *Viscum album* with adenine monophosphate and the crystal structure at 1.9 A resolution. *Acta Crystallographica. Section D, Biological Crystallography*, **58**(Pt 10 Pt 1), 1704–1707. https://doi.org/10.1107/s0907444902014270

Lang, H., Sement, F.M., Canaday, J. & Gagliardi, D. (2009) Polyadenylation-assisted RNA degradation processes in plants. *Trends in Plant Science*, **14**, 497–504. https://doi.org/10.1016/j.tplants.2009.06.007

Maclean, A.E., Hertle, A.P., Ligas, J., Bock, R., Balk, J. & Meyer, E.H. (2018) Absence of complex I is associated with diminished respiratory chain function in European mistletoe. *Current Biology*, **28**(10), 1614–1619.e3. https://doi.org/10.1016/j.cub.2018.03.036

Maldonado, M., Guo, F. & Letts, J.A. (2021) Atomic structures of respiratory complex III2, complex IV, and supercomplex III2-IV from vascular plants. *eLife*, **10**, e62047. https://doi.org/10.7554/eLife.62047

Marie, D. & Brown, S.C. (1993) A cytometric exercise in plant DNA histograms, with 2C values for 70 species. *Biological Cell*, **78**, 41–51. https://doi.org/10.1016/0248-4900(93)90113-s

Meyer, E.H., Heazlewood, J.L. & Millar, A.H. (2007) Mitochondrial acyl carrier proteins in Arabidopsis thaliana are predominantly soluble matrix proteins and none can be confirmed as subunits of respiratory Complex I. *Plant Molecular Biology*, **64**(3), 319–327. https://doi.org/10.1007/s11103-007-9156-9

Nagl, W., Jeanjour, M., Kling, H., Kuhner, S., Michels, I., Muller, T. *et al.* (1983) Genome and chromatin organization in higher plants. *Biologisches Zentralblatt*, **102**, 129–148.

Nazaruk, J. & Orlikowski, P. (2016) (2016) Phytochemical profile and therapeutic potential of Viscum album L. *Natural Product Research*, **30**(4), 373–385. https://doi.org/10.1080/14786419.2015.1022776

Niwa, H., Tonevitsky, A.G., Agapov, I.I., Saward, S., Pfüller, U. & Palmer, R.A. (2003) Crystal structure at 3 A of mistletoe lectin I, a dimeric type-II ribosome-inactivating protein, complexed with galactose. *European Journal of Biochemistry*, **270**(13), 2739–2749. https://doi.org/10.1046/j.1432-1033.2003.03646.x

Novák, P., Guignard, M.S., Neumann, P., Kelly, L.J., Mlinarec, J., Koblížková, A. *et al.* (2020) Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nature Plants*, **6**(11), 1325–1329. https://doi.org/10.1038/s41477-020-00785-x

Pellicer, J. & Leitch, I.J. (2019) The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist*, **226**, 301–305. https://doi.org/10.1111/nph.16261

Petersen, G., Cuenca, A., Møller, I.M. & Seberg, O. (2015a) Massive gene loss in mistletoe (Viscum, Viscaceae) mitochondria. *Scientific Reports*, **5**, 17588. https://doi.org/10.1038/srep17588

Petersen, G., Cuenca, A. & Seberg, O. (2015b) Plastome evolution in hemiparasitic mistletoes. *Genome Biology and Evolution*, **7**(9), 2520–2532. https://doi.org/10.1093/gbe/evv165

Pineau, B., Layoune, O., Danon, A. & De Paepe, R. (2008) L-galactono-1,4-lactone dehydrogenase is required for the accumulation of plant respiratory complex I. *Journal of Biological Chemistry*, **283**(47), 32500–32505. https://doi.org/10.1074/jbc.m805320200

Runswick, M.J., Fearnley, I.M., Skehel, J.M. & Walker, J.E. (1991) Presence of an acyl carrier protein in NADH:ubiquinone oxidoreductase from bovine heart mitochondria. *FEBS Letters*, **286**(1–2), 121–124. https://doi.org/10.1016/0014-5793(91)80955-3

Schertl, P., Sunderhaus, S., Klodmann, J., Grozeff, G.E., Bartoli, C.G. & Braun, H.P. (2012) L-galactono-1,4-lactone dehydrogenase (GLDH) forms part of three subcomplexes of mitochondrial complex I in Arabidopsis thaliana. *Journal of Biological Chemistry*, **287**(18), 14412–14419. https://doi.org/10.1074/jbc.m111.305144

Schimmeyer, J., Bock, R. & Meyer, E.H. (2016) L-Galactono-1,4-lactone dehydrogenase is an assembly factor of the membrane arm of mitochondrial complex I in Arabidopsis. *Plant Molecular Biology*, **90**(1–2), 117–126. https://doi.org/10.1007/s11103-015-0400-4

Schuster, G. & Stern, D. (2009) RNA polyadenylation and decay in mitochondria and chloroplasts. *Progress in Molecular Biology and Translational Science*, **85**, 393–422. https://doi.org/10.1016/s0079-6603(08)00810-6

Schwartz, R., Ting, C.S. & King, J. (2001) Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Research*, **11**(5), 703–709. https://doi.org/10.1101/gr.gr-1587r

Senkler, J., Rugen, N., Eubel, H., Hegermann, J. & Braun, H.P. (2018) Absence of complex I implicates rearrangement of the respiratory chain in European mistletoe. *Current Biology*, **28**(10), 1606–1613.e4. https://doi.org/10.1016/j.cub.2018.03.050

Seppey, M., Manni, M. & Zdobnov, E.M. (2019) BUSCO: Assessing Genome Assembly and Annotation Completeness. In: Kollmar, M. (Ed.) *Gene prediction. Methods in molecular biology*, vol 1962. Humana, New York, NY. 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Singh, R., Ming, R. & Yu, Q. (2016) Comparative analysis of GC content variations in plant genomes. *Tropical Plant Biology*, **9**, 136–149. https://doi.org/10.1007/s12042-016-9165-4

Skippington, E., Barkman, T.J., Rice, D.W. & Palmer, J.D. (2015) Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. *Proceedings of the National Academy of Sciences of the United States of America*, **112** (27), E3515–E3524. https://doi.org/10.1073/pnas.1504491112

Skippington, E., Barkman, T.J., Rice, D.W. & Palmer, J.D. (2017) Comparative mitogenomics indicates respiratory competence in parasitic Viscum despite loss of complex I and extreme sequence divergence, and reveals horizontal gene transfer and remarkable variation in genome size. *BMC Plant Biology*, **17**(1), 49. https://doi.org/10.1186/s12870-017-0992-8

Soto, D., Córdoba, J.P., Villarreal, F., Bartoli, C., Schmitz, J., Maurino, V.G. *et al.* (2015) Functional characterization of mutants affected in the carbonic anhydrase domain of the respiratory complex I in Arabidopsis thaliana. *The Plant Journal*, **83**(5), 831–844. https://doi.org/10.1111/tpj.12930

Soufari, H., Parrot, C., Kuhn, L., Waltz, F. & Hashem, Y. (2020) Specific features and assembly of the plant mitochondrial complex I revealed by cryo-EM. *Nature Communications*, **11**(1), 5195. https://doi.org/10.1038/s41467-020-18814-w

Sudarkina, O.J., Kurmanova, A.G. & Kozlov, J.V. (2007) Production and characterization of the B chains of mistletoe toxic lectins. *Molecular Biology*, **41**, 601–608. https://doi.org/10.1134/S0026893307040127

Sunderhaus, S., Dudkina, N.V., Jänsch, L., Klodmann, J., Heinemeyer, J., Perales, M. *et al.* (2006) Carbonic anhydrase subunits form a matrix-exposed domain attached to the membrane arm of mitochondrial complex I in plants. *Journal of Biological Chemistry*, **281**(10), 6482–6488. https://doi.org/10.1074/jbc.m511542200

Ulrich, I., Fritz, B. & Ulrich, W. (1988) Application of DNA fluorochromes for flow cytometric DNA analysis of plant protoplasts. *Plant Science,* **55**, 151–158. https://doi.org/10.1016/0168-9452(88)90171-9

Urech, K. & Baumgartner, S. (2015) Chemical Constituents of *Viscum album* L.: Implications for the Pharmaceutical Preparation of Mistletoe. In: Zänker, K.S. & Kaveri, S.V. (Eds.), *Mistletoe: from mythology to evidence-based medicine*. Transl. Res. Biomed. Basel: Karger, vol 4, pp 11–23. https://doi.org/10.1159/000375422

van Wijk, K.J., Leppert, T., Sun, Q., Boguraev, S.S., Sun, Z., Mendoza, L. & Deutsch, E.W. (2021) The Arabidopsis thaliana PeptideAtlas; harnessing world-wide proteomics data for a comprehensive community proteomics resource. *The Plant Cell*, **33**(11), 3421–3453. https://doi.org/10.1093/plcell/koab211

Zancani, M., Braidot, E., Filippi, A. & Lippe, G. (2020) Structural and functional properties of plant mitochondrial F-ATP synthase. *Mitochondrion*, **53**, 178–193. https://doi.org/10.1016/j.mito.2020.06.001

Zonneveld, B.J.M. (2010) New record holders for maximum genome size in eudicots and monocots. *Journal of Botany*, **2010**, 1–4. https://doi.org/10.1155/2010/527357