Re-annotated *Nicotiana benthamiana* gene models for enhanced proteomics and reverse genetics

Jiorgos Kourelis[1], Farnusch Kaschani[2], Friederike M. Grosse-Holz[1], Felix Homma[1], Markus Kaiser[2], Renier A. L. van der Hoorn[1]

[1]Plant Chemetics Laboratory, Department of Plant Sciences, University of Oxford, South Parks Road, OX1 3RB Oxford, UK; [2]Chemische Biologie, Zentrum fur Medizinische Biotechnologie, Fakultät für Biologie, Universität Duisburg-Essen, Essen, Germany.

***Nicotiana benthamiana* is an important model organism and representative of the Solanaceae (Nightshade) family. *N. benthamiana* has a complex ancient allopolyploid genome with 19 chromosomes, and an estimated genome size of 3.1Gb. Several draft assemblies of the *N. benthamiana* genome have been generated, however, many of the gene-models in these draft assemblies appear incorrect. Here we present a nearly non-redundant database of 42,855 improved *N. benthamiana* gene-models. With an estimated 97.6% completeness, the new predicted proteome is more complete than the previous proteomes. We show that the database is more sensitive and accurate in proteomics applications, while maintaining a reasonable low gene number. As a proof-of-concept we use this proteome to compare the leaf extracellular (apoplastic) proteome to a total extract of leaves. Several gene families are more abundant in the apoplast. For one of these apoplastic protein families, the subtilases, we present a phylogenetic analysis illustrating the utility of this database. Besides proteome annotation, this database will aid the research community with improved target gene selection for genome editing and off-target prediction for gene silencing.**

Keywords: *Solanaceae // Genome annotation // Nicotiana benthamiana // Proteomics // Subtilases*

Database availability. The database has been uploaded at Oxford Research Archives (ORA) and can be downloaded from this link: https://ora.ox.ac.uk/objects/uuid:f09e1d98-f0f1-4560-aed4-a5147bc7739d. We hope that this database will be accessible via additional databases in the near future.

Introduction

Nicotiana benthamiana has risen to prominence as a model organism for several reasons. First, *N. benthamiana* is highly susceptible to viruses, resulting in highly efficient virus-induced gene-silencing (VIGS) for rapid reverse genetic screens (Senthil-Kumar and Mysore, 2014). This hypersusceptibility to viruses is due to an ancient disruptive mutation in the RNA-dependent RNA polymerase 1 gene (*Rdr1*), present in the lineage of *N. benthamiana* which is used in laboratories around the world (Bally *et al.*, 2015). Reverse genetics using *N. benthamiana* have confirmed many genes important for disease resistance (Wu *et al.*, 2017; Senthil- Kumar *et al.*, 2018). Additionally, *N. benthamiana* is highly amenable to the generation of stable transgenic lines (Clemente, 2006; Sparkes *et al.*, 2006) and to transient expression of transgenes (Goodin *et al.*, 2008). This easy manipulation has facilitated rapid forward genetic screens and has established *N. benthamiana* as the plant bioreactor of choice for the production of biopharmaceuticals (Stoger *et al.*, 2014). Finally, *N. benthamiana* is a member of the Solanaceae (Nightshade) family which includes important crops such as potato (*Solanum tuberosum*), tomato (*Solanum lycopersicum*), eggplant (*Solanum melongena*), and pepper (*Capsicum* ssp.), as well as tobacco (*Nicotiana tabacum*) and petunia (*Petunia* ssp.).

N. benthamiana belongs to the *Suaveolentes* section of the *Nicotiana* genus, and has an ancient allopolyploid origin (>10Mya) accompanied by chromosomal re-arrangements resulting in a complex genome with 19 chromosomes in the haploid genome – reduced from the ancestral allotetraploid 24 chromosomes - and an estimated haploid genome size of ~3.1Gb (Leitch *et al.*, 2008; Goodin *et al.*, 2008; Wang and Bennetzen, 2015). There are four independent draft assemblies of the *N. benthamiana* genome (Bombarely *et al.*, 2012; Naim *et al.*, 2012), as well as a de-novo transcriptome generated from short-read RNAseq (Nakasugi *et al.*, 2014). These datasets have greatly facilitated research in *N. benthamiana*, allowing for efficient prediction of off-targets of VIGS (Fernandez-Pozo *et al.*, 2015) and genome editing using CRISPR/Cas9 (Liu *et al.*, 2017), as well as RNAseq and proteomics studies (Grosse-Holz *et al.*, 2018). These draft assemblies are, however, several years old and gene annotations have not been updated since. Furthermore, in the course of our research we realized that many of the gene models in these draft assemblies are incorrect, and that putative pseudo-genes are often annotated as protein-encoding genes. This is exacerbated because these draft assemblies are highly fragmented and that *N. benthamiana* has a complex origin. Furthermore, the *de-novo* transcriptome assembly has a high proportion of chimeric transcripts. Because of incorrect annotations, extensive processing is required to select target genes for reverse genetic approaches such as gene silencing and editing, or for phylogenetic analysis of gene families. We realized that gene annotation in several other species in the *Nicotiana* genus were much better (Xu *et al.*, 2017; Sierro *et al.*, 2013; Sierro *et al.*, 2014), and decided to re-annotate the available *N. benthamiana* draft genomes using these gene models as a template. The gene models obtained in this way were extracted into a single non-redundant database with improved gene models. Here we show that this database is more accurate and sensitive for proteomics, facilitates phylogenetic analysis of gene families, and may be useful for genome editing and VIGS on- and off-target prediction.

Results and Discussion

Re-annotation of gene-models in the N. benthamiana genome assemblies

For the annotation of gene-models in the different *N. benthamiana* draft genomes, we chose to use Scipio (Keller *et al.*, 2008). Scipio refines the transcription start-site, exon-exon boundaries, and the stop-codon position of protein sequences aligned to the genome using BLAT (Keller *et al.*, 2008). Importantly, given that the input protein sequences are well-annotated, this method is more accurate and sensitive than other gene prediction methods (Keller *et al.*, 2011). Because the efficiency of this process correlates with phylogenetic distance, we took the predicted protein sequences from recently sequenced *Nicotiana* species (**Figure 1**) (Sierro *et al.*, 2013; Xu *et al.*, 2017). We then used CD-HIT at a 95% identity cut-off to reduce the redundancy in this database and additionally to remove partial sequences (**Figure 1**, Step 1). The resulting database – Nicotiana_db95 – contains 85,453 protein sequences from various *Nicotiana* species. We used the protein sequences in this database as an input to annotate gene-models in the four independent draft assemblies of the *N. benthamiana* genome using Scipio (**Figure 1**, Step 2). As the available *N. benthamiana* draft genomes are highly fragmented and each individual draft genome may miss a number of genes, we extracted the gene-models generated by Scipio, filtered for redundancy using CD-HIT-EST and combined the gene-models into a single database (NbA) containing 41,651 gene-models (**Figure 1**, Step 3). We next compared the predicted proteome derived from our NbA database against the published predicted proteomes using a proteomics dataset from a full-leaf extract or apoplastic fluid (samples described further in the manuscript). Proteins for which peptides were identified in the published proteomes but not in our NbA database were extracted and re-annotated in the draft genomes as described above and added to our NbA database resulting in the NbB database containing 42,884 gene-models (**Figure 1**; Step 4, 1233 additional entries). Finally, missing BUSCOs (Benchmarking Universal Single-Copy Orthologues) (Simão *et al.*, 2015; Waterhouse *et al.*, 2018) were re-annotated in our database as above, together with the manual curation of several gene-families in which several duplicated genes were removed (see Material & Methods) to obtain our final NbC database containing 42,853 entries (**Figure 1**; Step 5).

The new proteome database is more complete, more sensitive and accurate, and relatively small

We next compared the predicted proteome database to the published predicted proteomes. We also included our Nicotiana_db95 proteome database in this comparison. The published proteomes included the predicted proteomes from the Niben0.4.4 and Niben1.0.1 draft genomes, a previously described curated database in which gene-models from Niben1.0.1 were corrected using RNAseq reads (Grosse-Holz *et al.*, 2018), and the predicted proteome derived from the *de-novo* transcriptomes (Nakasugi *et al.*, 2014).

We used BUSCO (Simão *et al.*, 2015; Waterhouse *et al.*, 2018) as a quality measure to estimate the completion of our database as compared to the published predicted proteomes (**Figure 2a**). The BUSCO set used contains 1440 highly conserved plant genes which are expected to be predominantly found in a single-copy (Simão *et al.*, 2015). Nicotiana_db95 has one fragmented and nine missing BUSCOs, indicating that at best we should be able to identify 99.3% of the *N. benthamiana* genes using this database (**Figure 2a**). In our NbB database, 1406/1440 (97.6%) BUSCO proteins were identified as complete, of which 762 were single-copy (54.2%), 644 were duplicated (45.8%), nine sequences were fragmented (0.63%), and 25 were missing (1.74%) (**Figure 2a**). The high number of duplication is likely due to either technical duplication generated by small variations between the different draft genomes, or genuinely duplicated genes arising from the allo-tetraploid origin of *N. benthamiana*. By adding missing BUSCOs into the NbC database, we recovered eight of the nine fragmented BUSCOs and ten of the 25 missing BUSCOs. In comparison, the next most complete, previously published proteome is the predicted proteome from the Nbv5.1 primary + alternate transcriptome, which has 12 fragmented and 32 missing BUSCOs, but it also has nearly five times more proteins than our database and 71.8% of BUSCOs are duplicated.

Next, we investigated the number of unique PFAM identifiers found with each entry in each proteome, as an estimation of the number of proteins incorrectly annotated (**Figure 2b**). We expect that miss-annotated sequences and fragmented gene products are less likely to get a PFAM annotation. Indeed, significantly more proteins get at least one PFAM identifier in our three databases as compared to the published proteomes, indicating that proteins in our database are better annotated.

Furthermore, we looked at the length distributions of proteins in the different predicted proteomes (**Figure 2c**). We reasoned that the protein-length distribution should be similar to that of the Nicotiana_db95 database. The proteins in the final proteome are significantly longer than those in the Niben0.4.4, Niben1.0.1, and manually curated proteome (Grosse-Holz et al., 2018) while the proteins in the Nbv5.1 primary + alternate proteome are on average larger than in our final NbC database. We speculate that the Niben0.4.4 and Niben1.0.1 predicted proteomes contain many pseudo-genes which are annotated as protein-encoding as well as partial genes (**Figure 2c**), while the Nbv5.1 primary + alternative proteome has a high proportion of chimeric sequences which due to the short-read sequencing techniques used are biased towards long transcripts (**Figure 2c**). Additionally, the curated proteome has a large proportion of very small proteins and 47.3% of genes do have a PFAM annotation, which we speculate is due to partial sequences or spurious small ORFs being annotated as protein-encoding (**Figure 2b,c**).

Finally, comparing the different proteomes on a proteomics dataset indicates the new database has the highest sensitivity, with the highest percentage of annotated MS/MS spectra in both tested samples, while it has the fewest entries (**Figure 2d**). Additionally, using our new NbC database, we identify the highest number of unique peptides identified in at least 3 out of 4 biological replicates of both proteomes (**Figure 2d**). These metrics combined indicate that the new NbC database is more sensitive and accurate for proteomics than the currently available databases. Importantly, this does not come at the cost of increased redundancy, which would hinder downstream applications.

Since the previous proteomics dataset was also used to re-annotate gene-models (**Figure 1**; Step 4), we independently validated our database on an independent dataset where we re-analysed a previously published apoplastic proteome of agro-infiltrated *N. benthamiana* as compared to non-infiltrated *N. benthamiana* (PRIDE repository PXD006708) (Grosse-Holz et al., 2018). The new NbC database was also more sensitive and accurate than the Curated database on this dataset (18,430 vs 17,960 peptides detected, 22.5%±/-3.1% vs 21.7%±3.0% spectra identified).

Finally, since phylogenetic analysis of gene families in closely related species often relies on gene-annotations, we compared the predicted proteome from our NbB database against the predicted proteomes of Solanaceae species for which genomes have been sequenced (**Figure S1a,b**). Our NbB proteome compares well to the predicted proteomes of other sequences Solanaceae species. Additionally, since the predicted proteomes of some of these species miss a relatively high proportion of genes (up to 28.5% of genes missing or fragmented), care must be taken to not over-interpret results derived from phylogenetic analysis using these sequences.

Improved annotation of the apoplastic proteome of N. benthamiana

Next we used our final NbC database to analyse the extracellular protein repertoire of the *N. benthamiana* apoplast. The plant apoplast is the primary interface in plant-pathogen interactions (Misas-Villamil and van der Hoorn, 2008; Doehlemann and Hemetsberger, 2013) and apoplastic proteins include many enzymes potentially important in plant-pathogen interactions. We found the protein composition of leaf apoplastic fluid (AF) to be distinct from that of a leaf total extract (TE) (**Figure 3a**). We considered proteins apoplastic when only detected in the AF samples or those with a $\log_2$ fold abundance difference ≥1.5 and a p-value cut-off off ≤0.01 (BH-adjusted moderated t-test) in the comparison of AF vs TE (518 proteins). Similarly, we considered proteins intracellular when found only in the TE or those proteins with a $\log_2$ fold abundance difference ≤-1.5 with a p-value cut-off off ≤0.01 in the comparison of AF vs TE (1042 proteins) (**Figure 3b**). The remainder proteins was considered both apoplastic and intracellular (832 proteins). As expected, the apoplastic proteome is significantly enriched for signal peptide containing proteins, while the intracellular proteins and proteins present both in the apoplast and intracellular are significantly enriched for proteins lacking a signal peptide (BH-adjusted hypergeometric test, p<0.001).

Proteins considered predominantly intracellular are enriched for GO-SLIM terms associated with translation, photosynthesis and transport as biological processes (**Figure 3c**), and a similar pattern is seen for the molecular function terms (**Figure 3d**). Proteins present both in TE and in AF are enriched for GO-SLIM terms associated with biosynthetic processes, and homeostasis (**Figure 3c**). These processes usually performed by proteins acting at multiple subcellular localizations. The apoplastic proteome is enriched for proteins acting in catabolic processes and carbohydrate and lipid metabolic processes (**Figure 3c**), which is reflected in the enrichment of peptidases and glycosidases (**Figure 3d, Table S1** for a full list).

To specify which peptidases are enriched in the apoplast, we also annotated the proteome with MEROPS peptidase identifiers (Rawlings *et al.*, 2018). Three of the 15 different families of peptidases detected in the apoplast have significantly more members enriched in the AF as compared to TE, namely the subtilase (S08; 13 members, p<0.001), serine carboxypeptidase-like (S10; 8 members, p<0.01), and aspartic peptidase families (A01; 16 members, p<0.001), while the proteasome is enriched in the intracellular fraction (T01; 27 members, p<0.001) (BH-adjusted hypergeometric test, **Table S2** for a full list).

Pseudogenization in the subtilisin family is consistent with a contracting functional genome

One of the gene families found enriched in the apoplast is the subtilisin family. Several subtilisins are implicated in immunity, notably the tomato P69 clade of subtilisins (Taylor and Qiu, 2017). In order to estimate the completeness of our database, we manually verified and corrected genes belonging to the subtilisin gene family. Our NbC database contains 64 complete subtilisin genes, and one partial gene. By searching the Niben1.0.1 and Niben0.4.4 genome assemblies, we identified an additional 43 putative subtilisin pseudo-genes which had internal stop-codons and are therefore likely non-coding.

Interestingly, phylogenetic analysis shows that close paralogs are often pseudogenised. This pattern of pseudogenization in the subtilisin gene family is consistent with a contracting functional genome upon polyploidization, where for each functional protein-encoding gene there is a corresponding pseudo-gene (**Figure 4**, and **Figure S2**). Remarkably, no SBT3 clade family members were identified in *N. benthamiana* (**Figure S3**). Finally, we looked for the amino acid residue at the pro-domain junction, as the presence of an aspartic acid residue is indicative of phytaspase activity (Reichardt et al., 2018). Three *N. benthamiana* subtilisins may possess phytaspase activity based on the presence of an apartic acid residue at the pro-domain junction as well as a histidine residue in the S1 pocket which is thought to bind to P1 aspartic acid (**Figure S2**, and **Figure S3**, Reichardt et al., 2018).

During this analysis we discovered three subtilisin genes that are missing in our NbB database, and six incomplete sequences lacking 5-107 amino acids. In addition, five putative pseudo-genes were annotated as protein-encoding genes and were removed from the final NbC database, and 18 subtilase genes were found to be duplicated and these duplicates were removed in the final NbC database (**Table S3**). In comparison, the Niben1.0.1 genome annotation predicts 103 different subtilisin gene products. However, we found that these annotated genes correspond to 38 pseudo-genes and 49 protein-encoding genes - none of which are correctly annotated - while 16 subtilisin genes are absent from Niben1.0.1 (**Table S3**). Furthermore, the predicted proteome from the Nbv5.1 primary+alternate transcriptome contains more than 400 subtilisin gene products, largely due to a large number of chimeric sequences. In conclusion, the new database represents a significant improvement over previous genome annotations and facilitates more accurate and meaningful phylogenetic analysis of gene families in *N. benthamiana*.

Improved accuracy for genome editing: the subtilase gene-family

Target selection for genome editing is improved by the use of our new database for several reasons: 1) gene-models in this database are more complete; 2) fewer pseudo-genes are annotated as protein-encoding genes; 3) gene duplication is reduced as compared to the de-novo transcriptome; and 4) the remaining duplication in our database is easily resolved for genes of interest as it mostly involves genes with slight sequence variations between the different draft genomes. These sequence variations may be due to heterozygosity or technical artefacts of the sequencing and assembly. As an example we show the gene-model of one of the subtilisins in the different databases. In our NbC database, this subtilisin is encoded by a single-exon gene-model of 2,268bp encoding for a 756 amino acid protein (**Figure 5a**). This subtilisin is highly fragmented in the Niben1.0.1 genome assembly, with parts of the sequence present on different contigs, while the gene is only partially annotated (**Figure 5b**). The last 90bp of this gene are not annotated in the Nbv0.5 genome (**Figure 5c**). Furthermore, there is a 132bp insertion in the Niben0.4.4 genome assembly resulting in a predicted protein with a 44 amino-acid insertion (**Figure 5d**). Additionally, we identified 13 sequences corresponding to partial or chimeric variants of this subtilisin are present in the Nb5.1 primary + alternate predicted proteome using BLAST with no full match. Finally, this subtilisin differs by three non-synonymous SNPs between the Niben1.0.1 and Nbv0.5 genome assemblies, while two of these non-synonymous SNPs are present in the Niben0.4.4 genome assembly (**Figure 5b,d**). In conclusion, this example displays how combining gene-models derived from different genome assemblies has made our database more complete than annotating any single genome assembly currently available.

Although our NbC database does not contain the genomic context and lacks non-coding genes, this database will vastly improve research on *N. benthamiana*. We trust our NbC database to be useful for the large research community of plant scientists using *N. benthamiana* as a model system, for example to identify novel interactors in Co-IP experiments, but also to facilitate reverse genetic approaches such as genome editing and VIGS.

Material & Methods

Sequence retrieval - The predicted proteomes for *N. attenuata* (GCF_001879085.1) (Xu *et al.*, 2017) (http://nadh.ice.mpg.de/NaDH/), *N. tabacum* TN90 (GCF_000715135.1) (Sierro *et al.*, 2013), *N. sylvestris* (GCF_000393655.1) (Sierro *et al.*, 2013) and *N. tomentosiformis* (GCF_000390325.2) (Sierro *et al.*, 2013), and *Daucus carota* subsp. *sativus* (GCA_001625215.1) (Iorizzo *et al.*, 2016) were downloaded from Genbank. In addition, we retrieved 565 full-length *N. benthamiana* protein sequences from Genbank. The *Arabidopsis thaliana* predicted proteome (Araport11_genes.201606.pep) was obtained from Araport (Cheng *et al.*, 2017). The *Solanum melongena* predicted proteome (SME_r2.5.1_pep) was obtained from the Eggplant Genome DataBase (Hirakawa *et al.*, 2014). The *N. obtusifolia* (NIOBT_r1.0) predicted proteome was obtained from the *Nicotiana attenuata* Data Hub (Xu *et al.*, 2017) (http://nadh.ice.mpg.de/NaDH/). The *Petunia axillaris* N (Petunia_axillaris_v1.6.2_proteins) and *P. inflata* S6 (Petunia_inflata_v1.0.1_proteins) (Bombarely *et al.*, 2016), *Capsicum annuum glabriusculum* (CaChiltepin.pep) and *C. annuum zunla-1* (CaZL1.pep) (Qin *et al.*, 2014), *C. annuum* cv CM334 (Pepper.v.1.55.proteins.annotated) (Kim *et al.*, 2014), *Solanum tuberosum* (PGSC_DM_v3.4_pep) (Consortium, 2011), and *Solanum lycopersicum* (ITAG3.2_proteins) (Consortium, 2012) predicted proteomes were downloaded from Solgenomics. The *N. benthamiana* draft genome builds Niben1.0.1 and Niben0.4.4 - both generated by the Boyce Thompson Institute for Plant Research (BTI) (Bombarely *et al.*, 2012) - were downloaded from Solgenomics, and the Nbv0.5 and Nbv0.3 draft genomes were made available by the Waterhouse lab at the Queensland University of Technology (Naim *et al.*, 2012).

Annotation - In order to extract gene-models from the published *N. benthamiana* draft genomes we combined all the *Nicotiana* protein sequences, except for those from *N. obtusifolia*, in one database, with the addition of 110 genes which we had previously manually curated leading to a database with 226,543 protein sequences. We used CD-HIT (v4.6.8) (Fu *et al.*, 2012) to cluster these sequences at a 95% identity threshold and reduce the redundancy in our database while removing partials (Nicotiana_db95; 85,453 sequences). This database was used to annotate the gene-models in the different *N. benthamiana* genome builds using Scipio version 1.4.1 (Keller *et al.*, 2008) which was run with default settings. After running Scipio we used Augustus (v3.3) (Stanke *et al.*, 2006) to extract complete and partial gene models. Putative pseudo-genes (containing internal stop codons) and genes lacking an ATG start or stop codon were stored separately. Transdecoder (v5.0.2) (Haas *et al.*, 2013) was used to retrieve the single-best ORF on the putative pseudo-genes containing homology to the Nicotiana_db95 database as determined by BlastP searches. If a putative pseudo-gene contained an ORF >90% of the annotated gene length and lacking <30 amino acids it was considered a putative gene. Other putative pseudo-genes were discarded. Next we used CD-HIT-EST to filter the redundancy from this database. First, we used CD-HIT-EST to cluster the CDS derived from the gene-models derived from the different databases at 100% identity. Next, we selected the longest sequence at 99% identity between the different genome builds using CD-HIT-EST-2D in the following order for both the complete and the partial databases: Niben1.0.1 > Nbv0.5 > Niben0.4.4 > Nbv0.3. Since sequences which are smaller are maintained like this we used the reduced databases in the opposite direction to remove partial genes: Nbv0.3 > Niben0.4.4 > Nbv0.5 > Niben1.0.1. Finally we used CD-HIT-EST-2D to remove genes from the partial database with a longer representative in the complete database at 99% identity and vice-versa. This resulted in the NbA database. We compared this database for proteomic analysis on the described proteomics dataset containing apoplastic fluid (AF) samples and full-leaf extract (TE) samples and compared its performance to the other published predicted proteomes. For this analysis we predicted the Nbv5.1 proteome from the transcriptome using Transdecoder and selecting the single-best ORF with homology to the Nicotiana_db95 database, and filtered the database using CD-HIT at 100% identity. Proteins for which peptides were identified in the other databases but absent from the NbA database search were extracted, clustered at 100% using CD-HIT, and re-annotated in the genomes as above. This resulted in the NbB database. Finally, we ran BUSCO (v3.0.2; dependencies: NCBI-BLAST v2.7.1+; HMMER v3.1; Augustus v3.3) (Simão *et al.*, 2015; Waterhouse *et al.*, 2018) on the different *N. benthamiana* predicted proteomes using the plants set (Embryophyta_odb9), extracted the missing BUSCOs and re-annotated these as above. Additionally, we manually inspected the database for the PLCP, subtilisin, VPE, and GH35-domain encoding gene families, and manually removed redundant sequences. This resulted in our final database. This database was annotated using SignalP (v4) (Petersen *et al.*, 2011), ApoplastP (v1.0.1) (Sperschneider *et al.*, 2018), and PFAM (v31) (Finn *et al.*, 2016). Finally we annotated the predicted proteome with GO terms and UniProt identifiers using Sma3s v2 (Casimiro-Soriguer *et al.*, 2017).

Sample preparation for proteomics and definition of biological replicates - Four-week old *N. benthamiana* plants were used. The AF was extracted by vacuum infiltrating *N. benthamiana* leaves with ice-cold MilliQ. Leaves were dried to remove excess liquid, and apoplastic fluid was extracted by centrifugation of the leaves in a 20 ml syringe barrel (without needle or plunger) in a 50 ml falcon tube at 2000x g, 4°C for 25min. Samples were snap-frozen in liquid nitrogen and stored at -80°C prior to use. TE was collected by removing the central vein and snap-freezing the leaves in liquid nitrogen followed by grinding in a pestle and mortar and addition of three volumes of phosphate-buffered saline (PBS) (w/v). One biological replicate was defined as a sample, AF or TE, consisting of one leaf from three independent plants (3 leaves total). Four independent biological replicates were taken for AF and TE.

Protein digestion and sample clean-up - AF and TE sample corresponding to 15µg of protein was taken for each sample (based on Bradford assay). Dithiothreitol (DTT) was added to a concentration of 40mM, and the volume adjusted to 250µl with MS-grade water (Sigma). Proteins were precipitated by the addition of 4 volumes of ice-cold acetone, followed by a 1hr incubation at -20°C and subsequent cetrifugation at 18,000 g, 4°C for 20min. The pellet was dried at room temperature (RT) for 5min and resuspended in 25µL 8M urea, followed by a second chloroform/methanol precipitation. The pellet was dried at RT for 5 min and resuspended in 25µL 8M urea. Protein reduction and alkylation was achieved by sequential incubation with DTT (final 5mM, 30 min, RT) and iodoacetamide (IAM; final 20mM, 30min, RT, dark). Non-reacted IAM was quenched by raising the DTT concentration to 25mM. Protein digestion was started by addition of 1000ng LysC (Wako Chemicals GmbH) and incubation for 3hr at 37°C while gently shaking (800rpm). The samples were then diluted with ammoniumbicarbonate (final concentration 80mM) to a final urea concentration of 1M. 1000ng Sequencing grade Trypsin (Promega) was added and the samples were incubated overnight at 37°C while gently shaking (800rpm). Protein digestion was stopped by addition of formic acid (FA, final 5% v/v). Tryptic digests were desalted on home-made C18 StageTips (Rappsilber *et al.*, 2007) by passing the solution over 2 disc StageTips in 150µL aliquots by centrifugation (600-1200× g). Bound peptides were washed with 0.1% FA and subsequently eluted with 80% Acetonitrile (ACN). Using a vacuum concentrator (Eppendorf) samples were dried, and the peptides were resuspended in 20 µL 0.1% FA solution.

LC-MS/MS - The samples were analysed as in (Grosse-Holz *et al.*, 2018). Briefly, samples were run on an Orbitrap Elite instrument (Thermo) (Michalski *et al.*, 2011) coupled to an EASY-nLC 1000 liquid chromatography (LC) system (Thermo) operated in the one-column mode. Peptides were directly loaded on a fused silica capillary (75µm × 30cm) with an integrated PicoFrit emitter (New Objective) analytical column packed in-house with Reprosil-Pur 120 C18-AQ 1.9 µm resin (Dr. Maisch), taking care to not exceed the set pressure limit of 980 bar (usually around 0.5-0.8µl/min). The analytical column was encased by a column oven (Sonation; 45°C during data acquisition) and attached to a nanospray flex ion source (Thermo). Peptides were separated on the analytical column by running a 140-min gradient of solvent A (0.1% FA in water; ; Ultra-Performance Liquid Chromatography (UPLC) grade) and solvent B (0.1% FA in ACN; UPLC grade) at a flow rate of 300nl/min (gradient: start with 7% B; gradient 7% to 35% B for 120 min; gradient 35% to 100% B for 10 min and 100% B for 10 min) at a flow rate of 300 nl/min.). The mass spectrometer was operated using Xcalibur software (version 2.2 SP1.48) in positive ion mode. Precursor ion scanning was performed in the Orbitrap analyzer (FTMS; Fourier Transform Mass Spectrometry) in the scan range of m/z 300-1800 and at a resolution of 60000 with the internal lock mass option turned on (lock mass was 445.120025 m/z, polysiloxane) (Olsen *et al.*, 2005). Product ion spectra were recorded in a data-dependent manner in the ion trap (ITMS) in a variable scan range and at a rapid scan rate. The ionization potential was set to 1.8kV. Peptides were analysed by a repeating cycle of a full precursor ion scan (1.0 × 106 ions or 50ms) followed by 15 product ion scans (1.0 × 10^4 ions or 50ms). Peptides exceeding a threshold of 500 counts were selected for tandem mass (MS2) spectrum generation. Collision induced dissociation (CID) energy was set to 35% for the generation of MS2 spectra. Dynamic ion exclusion was set to 60 seconds with a maximum list of excluded ions consisting of 500 members and a repeat count of one. Ion injection time prediction, preview mode for the Fourier transform mass spectrometer (FTMS, the orbitrap), monoisotopic precursor selection and charge state screening were enabled. Only charge states higher than 1 were considered for fragmentation.

Peptide and Protein Identification - Peptide spectra were searched in MaxQuant (version 1.5.3.30) using the Andromeda search engine (Cox *et al.*, 2011) with default settings and label-free quantification and match-between-runs activated (Cox and Mann, 2008; Cox *et al.*, 2014) against the databases specified in the text including a known contaminants database. Included modifications were carbamidomethylation (static) and oxidation and N-terminal acetylation (dynamic). Precursor mass tolerance was set to ±20 ppm (first search) and ±4.5 ppm (main search), while the MS/MS match tolerance was set to ±0.5 Da. The peptide spectrum match FDR and the protein FDR were set to 0.01 (based on a target-decoy approach) and the minimum peptide length was set to 7 amino acids. Protein quantification was performed in MaxQuant (Tyanova *et al.*, 2016), based on unique and razor peptides including all modifications.

Proteomics processing in R - Identified protein groups were filtered for reverse and contaminants proteins and those only identified by matching, and only those protein groups identified in 3 out of 4 biological replicates either AF or TE were selected. The LFQ values were $\log_2$ transformed, and missing values were imputed using a minimal distribution as implemented in imputeLCMD (v2.0) (Lazar, 2015). A moderated t-test was used as implemented in Limma (v3.34.3) (Ritchie *et al.*, 2015; Phipson *et al.*, 2016) and adjusted using Benjamini–Hochberg (BH) adjustment to identify protein groups significantly differing between AF and TE. Bonafide apoplastic protein groups were those only detected in AF and those significantly ($p \leq 0.01$) $\log_2$ fold change ≥1.5 in AF samples. Protein groups only detected in TE and those significantly ($p \leq 0.01$) $\log_2$ fold change ≤-1.5 depleted in AF samples were considered intracellular. The remainder was considered both apoplastic and intra-cellular. Majority proteins were annotated with SignalP, PFAM, MEROPS (v12) (Rawlings *et al.*, 2018), GO, and UniProt keywords identifiers. A BH-adjusted Hypergeometric test was used to identify those terms that were either depleted or enriched ($p \leq 0.05$) in the bonafide AF protein groups as compared to bonafide AF depleted proteins or protein groups present both in the AF and TE.

Phylogenetic analysis - Predicted proteomes were annotated with PFAM identifiers, and all sequences containing a Peptidase S8 (PF00082) domain were extracted from the different databases. Additionally, we manually curated the subtilisin gene-family in the Niben1.0.1 draft genome, identifying putative pseudo-genes which were annotated as protein-encoding genes, as well as missing genes and incorrect gene models or genes in which the reference sequence was absent in Niben1.0.1. Tomato subtilisins were retrieved from Solgenomics, and other previously characterized subtilisins (Taylor and Qiu, 2017) were retrieved from NCBI. Clustal Omega (Sievers *et al.*, 2011; Li *et al.*, 2015) was used to align these sequences. The putative pseudo-gene sequences were substituted with the best blast hit in NCBI in order to visualize pseudogenization in the alignment and phylogenetic tree. Determining the best model for maximum likelihood phylogenetic analysis and the phylogenetic analysis was performed in MEGA X (Kumar *et al.*, 2018). The evolutionary history was inferred by using the Maximum Likelihood method based on the Whelan and Goldman model. A discrete Gamma distribution was used to model evolutionary rate differences among sites, and the rate variation model allowed for some sites to be evolutionarily invariable. All positions with less than 80% site coverage were eliminated. Niben101Scf00595_742942-795541 was used to root the phylogenetic trees.

Data Availability - The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaíno et al., 2016) partner repository (https://www.ebi.ac.uk/pride/archive/) with the data set identifier PXD010435. During the review process the data can be accessed via a reviewer account (Username: reviewer17475@ebi.ac.uk; Password: PQSfFZyN). Samples FGH01-04 represent AF and FGH05-08 represent TE.

Acknowledgements. We would like to thank Philippe Varennes-Jutras and Daniela Sueldo for critically reading the manuscript and providing important suggestions for improving the manuscript.

Funding. This work has been supported by 'The Clarendon Fund' (JK), and the ERC Consolidator grant 616449 'GreenProteases' (RvdH, FGH). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests. The authors have declared that no competing interests exist.

Author contributions. Conceptualization: JK, RvdH; Formal analysis: JK; Funding acquisition: RvdH; Wetlab experiments: FHG; Proteomics: FK, MK; Programming: JK, FH; Writing: JK, RvdH.

References

Bally, J., Nakasugi, K., Jia, F., Jung, H., Ho, S.Y.W., Wong, M., Paul, C.M., Naim, F., Wood, C.C., Crowhurst, R.N., Hellens, R.P., Dale, J.L. and Waterhouse, P.M. (2015) The extremophile *Nicotiana benthamiana* has traded viral defence for early vigour. *Nat Plants*, **1**, 15165.

Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C.S., Bliek, M., Boersma, M.R., Borghi, L., Bruggmann, R., Bucher, M., D'Agostino, N., Davies, K., Druege, U., Dudareva, N., Egea-Cortines, M., Delledonne, M., Fernandez-Pozo, N., Franken, P., Grandont, L., Heslop-Harrison, J.S., Hintzsche, J., Johns, M., Koes, R., Lv, X., Lyons, E., Malla, D., Martinoia, E., Mattson, N.S., Morel, P., Mueller, L.A., Muhlemann, J., Nouri, E., Passeri, V., Pezzotti, M., Qi, Q., Reinhardt, D., Rich, M., Richert-Pöggeler, K.R., Robbins, T.P., Schatz, M.C., Schranz, M.E., Schuurink, R.C., Schwarzacher, T., Spelt, K., Tang, H., Urbanus, S.L., Vandenbussche, M., Vijverberg, K., Villarino, G.H., Warner, R.M., Weiss, J., Yue, Z., Zethof, J., Quattrocchio, F., Sims, T.L. and Kuhlemeier, C. (2016) Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. *Nat. Plants*, **2**, 16074.

Bombarely, A., Rosli, H.G., Vrebalov, J., Moffett, P., Mueller, L.A. and Martin, G.B. (2012) A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. *Mol. Plant-Microbe Interact.*, **25**, 1523–1530.

Casimiro-Soriguer, C.S., Muñoz-Mérida, A. and Pérez-Pulido, A.J. (2017) Sma3s: A universal tool for easy functional annotation of proteomes and transcriptomes. *Proteomics*, **17**.

Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S. and Town, C.D. (2017) Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J*, **89**, 789–804.

Clemente, T. (2006) *Nicotiana* (*Nicotiana tobaccum*, *Nicotiana benthamiana*). In K. Wang, ed. *Agrobacterium Protocols*. Methods in Molecular Biology (Clifton, N.J.). Humana Press, pp. 143–154. Available at: http://link.springer.com/protocol/10.1385/1-59745-130-4%3A143 [Accessed January 4, 2015].

Consortium, T.P.G.S. (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.

Consortium, T.T.G. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N. and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics*, **13**, 2513–2526.

Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotech*, **26**, 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V. and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, **10**, 1794–1805.

Doehlemann, G. and Hemetsberger, C. (2013) Apoplastic immunity and its suppression by filamentous plant pathogens. *New Phytol.*, **198**, 1001–1016.

Fernandez-Pozo, N., Rosli, H.G., Martin, G.B. and Mueller, L.A. (2015) The SGN VIGS Tool: user-friendly software to design Virus-Induced Gene Silencing (VIGS) constructs for functional genomics. *Mol. Plant*, **8**, 486–488.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J. and Bateman, A. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*, **44**, D279–D285.

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

Goodin, M.M., Zaitlin, D., Naidu, R.A. and Lommel, S.A. (2008) *Nicotiana benthamiana*: its history and future as a model for plant–pathogen interactions. *MPMI*, **21**, 1015–1026.

Grosse-Holz, F.M., Kelly, S., Blaskowski, S., Kaschani, F., Kaiser, M. and van der Hoorn, R.A.L. (2018) The transcriptome, extracellular proteome and active secretome of agroinfiltrated *Nicotiana benthamiana* uncover a large, diverse protease repertoire. *Plant Biotechnol. J.*, **16**, 1068–1084.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet,

N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N. and Regev, A. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols*, **8**, 1494–1512.

Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., Negoro, S., Ohyama, A., Yamaguchi, H., Sato, S., Isobe, S., Tabata, S. and Fukuoka, H. (2014) Draft genome sequence of eggplant (*Solanum melongena* L.): the representative *Solanum* species indigenous to the old world. *DNA Res*, **21**, 649–660.

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., Bowman, M., Iovene, M., Sanseverino, W., Cavagnaro, P., Yildiz, M., Macko-Podgórni, A., Moranska, E., Grzebelus, E., Grzebelus, D., Ashrafi, H., Zheng, Z., Cheng, S., Spooner, D., Van Deynze, A. and Simon, P. (2016) A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.*, **48**, 657–666.

Keller, O., Kollmar, M., Stanke, M. and Waack, S. (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, **27**, 757–763.

Keller, O., Odronitz, F., Stanke, M., Kollmar, M. and Waack, S. (2008) Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics*, **9**, 278.

Kim, S., Park, M., Yeom, S.-I., Kim, Y.-M., Lee, J.M., Lee, H.-A., Seo, E., Choi, J., Cheong, K., Kim, K.-T., Jung, K., Lee, G.-W., Oh, S.-K., Bae, C., Kim, S.-B., Lee, H.-Y., Kim, S.-Y., Kim, M.-S., Kang, B.-C., Jo, Y.D., Yang, H.-B., Jeong, H.-J., Kang, W.-H., Kwon, J.-K., Shin, C., Lim, J.Y., Park, J.H., Huh, J.H., Kim, J.-S., Kim, B.-D., Cohen, O., Paran, I., Suh, M.C., Lee, S.B., Kim, Y.-K., Shin, Y., Noh, S.-J., Park, J., Seo, Y.S., Kwon, S.-Y., Kim, H.A., Park, J.M., Kim, H.-J., Choi, S.-B., Bosland, P.W., Reeves, G., Jo, S.-H., Lee, B.-W., Cho, H.-T., Choi, H.-S., Lee, M.-S., Yu, Y., Do Choi, Y., Park, B.-S., Deynze, A. van, Ashrafi, H., Hill, T., Kim, W.T., Pai, H.-S., Ahn, H.K., Yeam, I., Giovannoni, J.J., Rose, J.K.C., Sørensen, I., Lee, S.-J., Kim, R.W., Choi, I.-Y., Choi, B.-S., Lim, J.-S., Lee, Y.-H. and Choi, D. (2014) Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.*, **46**, 270–278.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. and Battistuzzi, F.U. (2018) MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol Biol Evol*, **35**, 1547–1549.

Lazar, C. (2015) *imputeLCMD: a collection of methods for left-censored missing data imputation*, Available at: https://cran.r-project.org/web/packages/imputeLCMD/index.html [Accessed February 25, 2018].

Leitch, I.J., Hanson, L., Lim, K.Y., Kovarik, A., Chase, M.W., Clarkson, J.J. and Leitch, A.R. (2008) The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann Bot*, **101**, 805–814.

Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y.M., Buso, N. and Lopez, R. (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res*, **43**, W580–W584.

Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K. and Chen, L.-L. (2017) CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants. *Molecular Plant*, **10**, 530–532.

Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M. and Horning, S. (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol Cell Proteomics*, **10**, M111.011015.

Misas-Villamil, J.C. and van der Hoorn, R.A.L. (2008) Enzyme–inhibitor interactions at the plant–pathogen interface. *Current Opinion in Plant Biology*, **11**, 380–388.

Naim, F., Nakasugi, K., Crowhurst, R.N., Hilario, E., Zwart, A.B., Hellens, R.P., Taylor, J.M., Waterhouse, P.M. and Wood, C.C. (2012) Advanced engineering of lipid metabolism in *Nicotiana benthamiana* using a draft genome and the V2 viral silencing-suppressor protein. *PLOS ONE*, **7**, e52717.

Nakasugi, K., Crowhurst, R., Bally, J. and Waterhouse, P. (2014) Combining transcriptome assemblies from multiple *de novo* assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. *PLOS ONE*, **9**, e91776.

Olsen, J.V., Godoy, L.M.F. de, Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S. and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics*, **4**, 2010–2021.

Petersen, T.N., Brunak, S., Heijne, G. von and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

Phipson, B., Lee, S., Majewski, I.J., Alexander, W.S. and Smyth, G.K. (2016) Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann Appl Stat*, **10**, 946–963.

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., Cheng, J., Zhao, S., Xu, M., Luo, Y., Yang, Y., Wu, Z., Mao, L., Wu, H., Ling-Hu, C., Zhou, H., Lin, H., González-Morales, S., Trejo-Saavedra, D.L., Tian, H., Tang, Xin, Zhao, M., Huang, Z., Zhou, A., Yao, X., Cui, J., Li, Wenqi, Chen, Z., Feng, Y., Niu, Y., Bi, S., Yang, X., Li, Weipeng, Cai, H., Luo, X., Montes-Hernández, S., Leyva-González, M.A., Xiong, Z., He, X., Bai, L., Tan, S., Tang, Xiangqun, Liu, D., Liu, J., Zhang, S., Chen, M., Zhang, Lu, Zhang, Li, Zhang, Yinchao, Liao, W., Zhang, Yan, Wang, M., Lv, X., Wen, B., Liu, H., Luan, H., Zhang, Yonggang, Yang, S., Wang, X., Xu, J., Li, X., Li, S., Wang, J., Palloix, A., Bosland, P.W., Li, Y., Krogh, A., Rivera-Bustamante, R.F., Herrera-Estrella, L., Yin, Y., Yu, J., Hu, K. and Zhang, Z. (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl. Acad. Sci. USA*, **111**, 5135–5140.

Reichardt, S., Repper, D., Tuzhikov, A.I., Galiullina, R.A., Planas-Marquès, M., Chichkova, N.V., Vartapetian, A.B., Stintzi, A. and Schaller, A. (2018) The tomato subtilase family includes several cell death-related proteinases with caspase specificity. *Sci. Rep.* **8**, 10531.

Rappsilber, J., Mann, M. and Ishihama, Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protocols*, **2**, 1896–1906.

Rawlings, N.D., Barrett, A.J., Thomas, P.D., Huang, X., Bateman, A. and Finn, R.D. (2018) The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*, **46**, D624–D632.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, **43**, e47–e47.

Senthil-Kumar, M. and Mysore, K.S. (2014) Tobacco rattle virus–based virus-induced gene silencing in *Nicotiana benthamiana*. *Nat. Protocols*, **9**, 1549–1562.

Senthil- Kumar, M., Wang, M., Chang, J., Ramegowda, V., Pozo, O. del, Liu, Y., Doraiswamy, V., Lee, H.-K., Ryu, C.-M., Wang, K., Xu, P., Eck, J.V., Chakravarthy, S., Dinesh- Kumar, S.P., Martin, G.B. and Mysore, K.S. (2018) Virus-induced gene silencing database for phenomics and functional genomics in *Nicotiana benthamiana*. *Plant Direct*, **2**, e00055.

Sierro, N., Battey, J.N., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., Peitsch, M.C. and Ivanov, N.V. (2013) Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biology*, **14**, R60.

Sierro, N., Battey, J.N.D., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., Goepfert, S., Peitsch, M.C. and Ivanov, N.V. (2014) The tobacco genome sequence and its comparison with those of tomato and potato. *Nat Commun*, **5**, 3833.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D. and Higgins, D.G. (2011) Fast, scalable generation of high- quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, **7**, 539.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Sparkes, I.A., Runions, J., Kearns, A. and Hawes, C. (2006) Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nat. Protocols*, **1**, 2019–2025.

Sperschneider, J., Dodds, P.N., Singh, K.B. and Taylor, J.M. (2018) ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytol*, **217**, 1764–1778.

Stanke, M., Schöffmann, O., Morgenstern, B. and Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Stoger, E., Fischer, R., Moloney, M. and Ma, J.K.-C. (2014) Plant molecular pharming for the treatment of chronic and infectious diseases. *Annual Review of Plant Biology*, **65**, 743–768.

Taylor, A. and Qiu, Y.-L. (2017) Evolutionary history of subtilases in land plants and their involvement in symbiotic interactions. *Mol. Plant-Microbe Interact*, **30**, 489–501.

Tyanova, S., Temu, T. and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protocols*, **11**, 2301–2319.

Vizcaíno, J.A., Csordas, A., Toro, N. del-, Dianes, J.A., Griss, J., Lavidas, I., Mayer,

G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.-W., Wang, R. and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res*, **44**, D447–D456.

Wang, X. and Bennetzen, J.L. (2015) Current status and prospects for the study of *Nicotiana* genomics, genetics, and nicotine biosynthesis genes. *Mol Genet Genomics*, **290**, 11–21.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V. and Zdobnov, E.M. (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*, **35**, 543–548.

Wu, C.-H., Abd-El-Haliem, A., Bozkurt, T.O., Belhaj, K., Terauchi, R., Vossen, J.H. and Kamoun, S. (2017) NLR network mediates immunity to diverse plant pathogens. *Proc. Natl. Acad. Sci. USA*, **114**, 8113–8118.

Xu, S., Brockmöller, T., Navarro-Quezada, A., Kuhl, H., Gase, K., Ling, Z., Zhou, W., Kreitzer, C., Stanke, M., Tang, H., Lyons, E., Pandey, P., Pandey, S.P., Timmermann, B., Gaquerel, E. and Baldwin, I.T. (2017) Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proc. Natl. Acad. Sci. USA*, **114**, 6133–6138.

Figure S1: Comparison of Solanaceae proteomes.

Figure S2: Phylogenetic analysis of the subtilisin gene-family from figure 4 with names.

Figure S3: Phylogenetic analysis of the subtilisin gene-family of tomato and Arabidopsis and including other previously characterized subtilisins.

Table S1: GO-SLIM term enrichment complete at p≤0.05

Table S2: MEROPS family term enrichment complete

Table S3: Gene-model comparison

Supplemental dataset 1: NbC gene-models database fasta nucleotide sequence

Supplemental dataset 2: NbC gene-models database gff3 annotation

Supplemental dataset 3: NbC predicted CDS

Supplemental dataset 4: NbC predicted proteome

Supplemental dataset 5: PFAM, SignalP, ApoplastP, and Sma3 v2 annotation of the NbC predicted proteome.

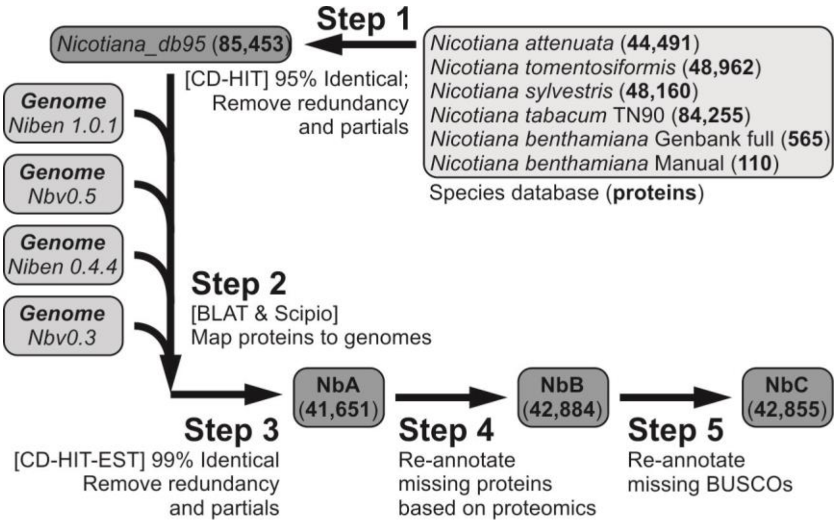
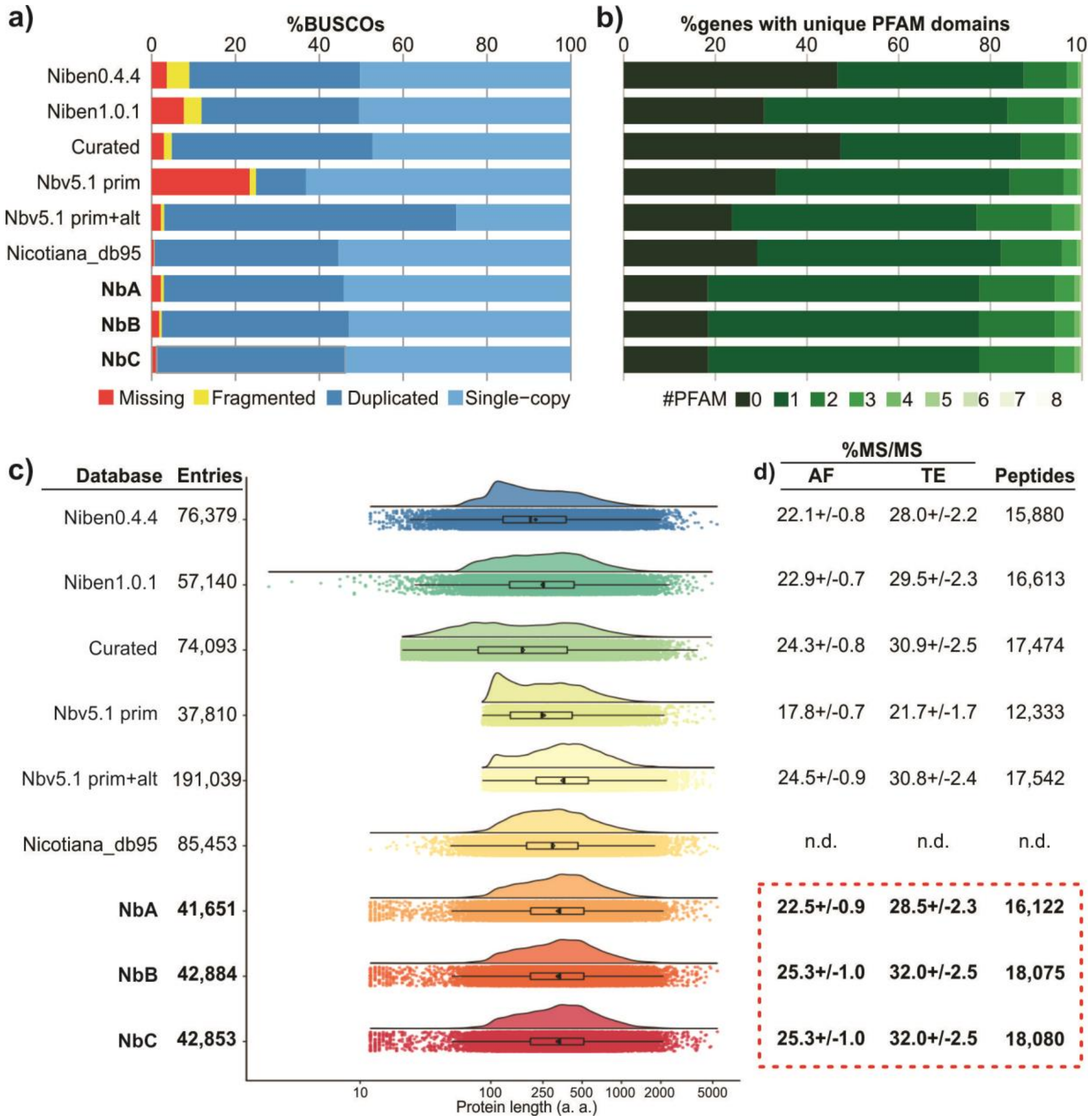


Figure 1. Bioinformatics pipeline for improved *Nicotiana benthamiana* proteome annotation. A database of *Nicotiana* predicted protein sequences was retrieved from Genbank and clustered at 95% identity threshold to reduce redundancy (Step 1), and used to annotate the four available *N. benthamiana* draft genomes (Step 2). The databases derived from the different genome builds were clustered at 99% sequence identity using CD-HIT-EST-2D in the following order: Niben1.0.1 > Nbv0.5 > Niben0.4.4 > Nbv0.3 generating NbA (Step 3). Proteins identified by proteomics in other databases but missing in NbA were added as above to generate NbB (Step 4), and finally missing BUSCOs were added as above to generate the final database NbC (Step 5).



Database	Entries
Niben0.4.4	76,379
Niben1.0.1	57,140
Curated	74,093
Nbv5.1 prim	37,810
Nbv5.1 prim+alt	191,039
Nicotiana_db95	85,453
NbA	**41,651**
NbB	**42,884**
NbC	**42,853**

AF	TE	Peptides
22.1+/-0.8	28.0+/-2.2	15,880
22.9+/-0.7	29.5+/-2.3	16,613
24.3+/-0.8	30.9+/-2.5	17,474
17.8+/-0.7	21.7+/-1.7	12,333
24.5+/-0.9	30.8+/-2.4	17,542
n.d.	n.d.	n.d.
22.5+/-0.9	**28.5+/-2.3**	**16,122**
25.3+/-1.0	**32.0+/-2.5**	**18,075**
25.3+/-1.0	**32.0+/-2.5**	**18,080**

Figure 2. Increased coverage and annotation of *N. benthamiana* proteins.
a) Completeness of the different predicted proteomes was estimated using BUSCO with the embryophyta database. **b)** The percentage of the proteins assigned with a certain number of unique PFAM identifiers (from 0-8) is plotted. **c)** Violin and boxplot graph of $\log_{10}$ protein length distribution of each database. Jittered dots show the raw underlying data. **d)** Percentage of annotated MS/MS spectra in AF or TE samples. Means and standard deviation (n=4) are shown and the number of unique peptides identified in at least of three out of four biological replicates of either AF or TE.

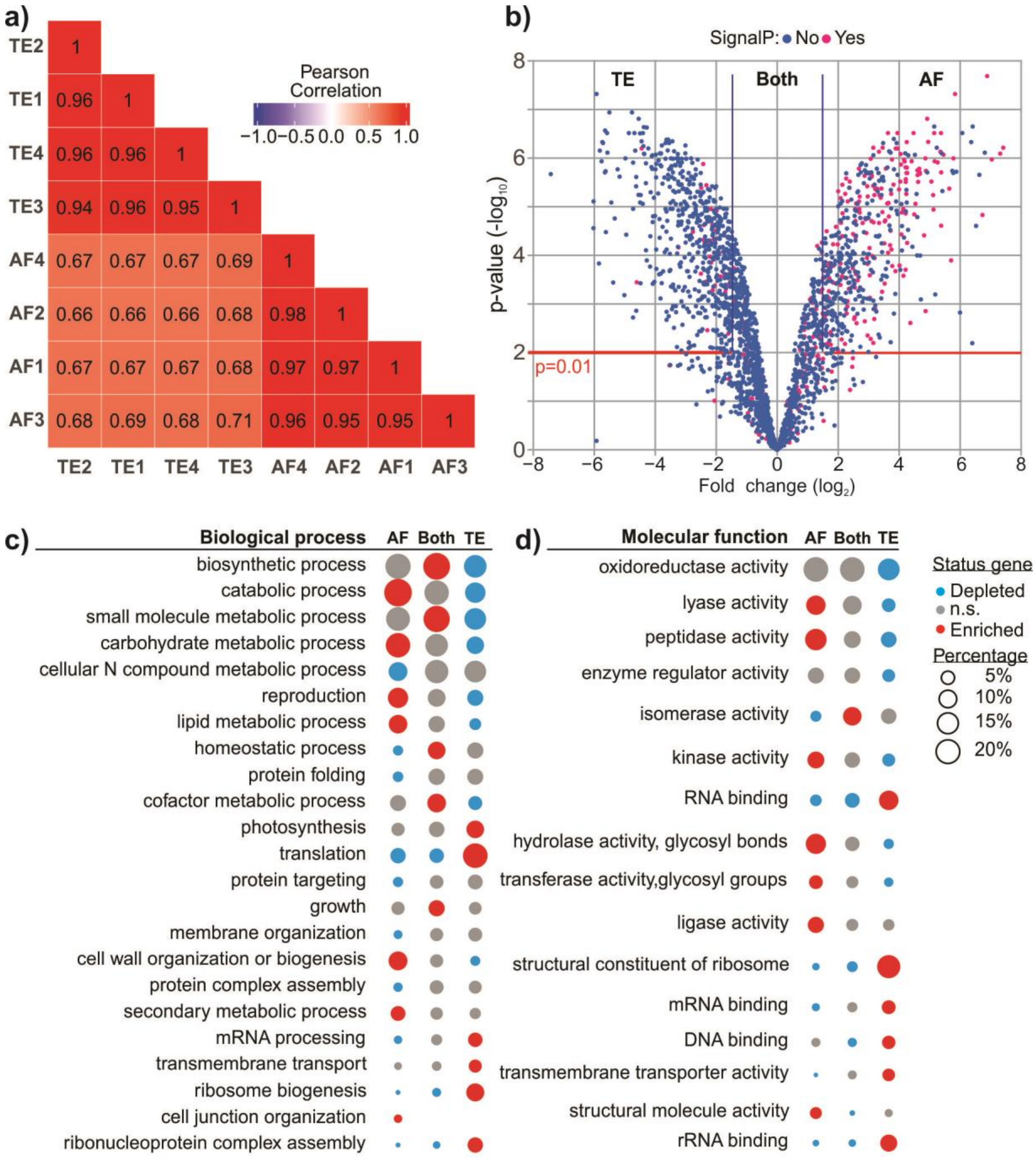


Figure 3. Improved annotation of the *N. benthamiana* apoplastic proteome.
a) Correlation matrix heat map of the LFQ intensity of protein groups in the biological replicates of AF and TE. Biological replicates are clustered on similarity. **b)** A volcano plot is shown plotting $\log_2$ fold difference of AF/TE over $-\log_{10}$ BH-adjusted moderated p-values. Proteins $\log_2$ fold change ≥ 1.5 and $p \leq 0.01$ were considered apoplastic, as well as those only found in AF. Conversely, proteins with a $\log_2$ fold change ≤ -1.5 and $p \leq 0.01$ were considered strictly intracellular, as well as those found only in TE. **c)** and **d)** shows a grid where each row represents an GO-SLIM annotation significantly enriched or depleted (BH-adjusted hypergeometric test, $p<0.05$) in at least one of the fractions (apoplast, intracellular, or both) and each column the compartment. Each bubble indicates the percentage of genes containing that specific GO-SLIM annotation in that compartment. Colours indicate whether the GO-SLIM annotations are enriched or depleted in that fraction ($p<0.05$, n.s. = non-significant). **c)** Percentage of proteins in each fraction annotated with biological process-associated GO-SLIM terms; **d)** Molecular function-associated GO-SLIM terms.

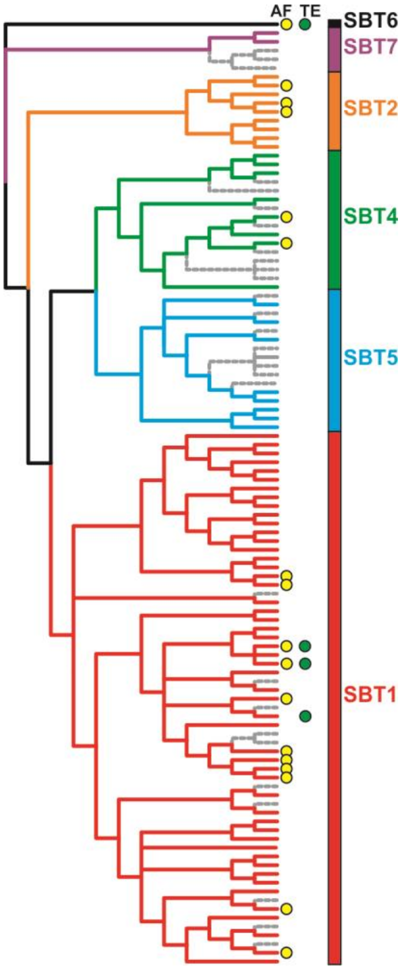


Figure 4. Birth and death of subtilase gene paralogs in *N. benthamiana.*
The evolutionary history of the subtilisin gene family was inferred by using the Maximum Likelihood method based on the Whelan and Goldman model. The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analysed. Putative pseudogenes are indicated in grey. Subtilases identified in apoplastic fluid (AF) and/or total extract (TE) are indicated with yellow and green dots, respectively. Naming of subtilase clades is according to (Taylor and Qiu, 2017). **Figure S2** includes the individual names.

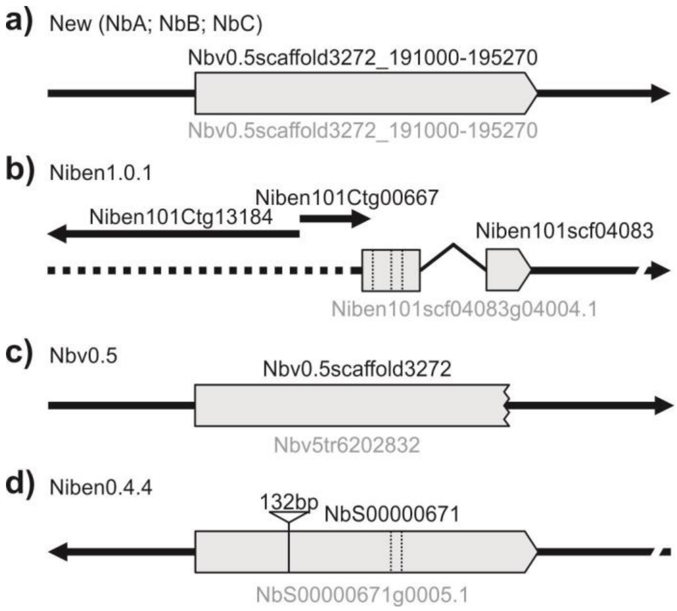


Figure 5. Example of a subtilase gene misannotation in different genome assemblies.
The gene-models corresponding to a subtilase encoded by Nbv0.5scaffold3272_191000-195270 were retrieved from the different databases: **a)** NbC, **b)** Niben1.0.1 genome, **c)** Nbv0.5 genome, and **d)** Niben0.4.4 genome. CDS is annotated in grey boxes. The dotted line indicates a stretch of unknown sequence (Ns). Vertical dotted lines indicate SNPs, and the triangle indicate an insertion. Strand direction is indicated by arrows.