The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2015, ••, ••–••. With 4 figures.

Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section *Lycopersicon*: Solanaceae)

STEVEN DODSWORTH[1,2]*, MARK W. CHASE[2,3], TIINA SÄRKINEN[4], SANDRA KNAPP[5] and ANDREW R. LEITCH[1]

[1]*School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK*
[2]*Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3DS, UK*
[3]*School of Plant Biology, The University of Western Australia, Crawley, WA, 6009, Australia*
[4]*Royal Botanic Garden, Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK*
[5]*Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK*

Received 17 February 2015; revised 7 May 2015; accepted for publication 21 May 2015

High-throughput sequencing data have transformed molecular phylogenetics and a plethora of phylogenomic approaches are now readily available. Shotgun sequencing at low genome coverage is a common approach for isolating high-copy DNA, such as the plastid or mitochondrial genomes, and ribosomal DNA. These sequence data, however, are also rich in repetitive elements that are often discarded. Such data include a variety of repeats present throughout the nuclear genome in high copy number. It has recently been shown that the abundance of repetitive elements has phylogenetic signal and can be used as a continuous character to infer tree topologies. In the present study, we evaluate repetitive DNA data in tomatoes (*Solanum* section *Lycopersicon*) to explore how they perform at the inter- and intraspecific levels, utilizing the available data from the 100 Tomato Genome Sequencing Consortium. The results add to previous examples from angiosperms where genomic repeats have been used to resolve phylogenetic relationships at varying taxonomic levels. Future prospects now include the use of genomic repeats for population-level analyses and phylogeography, as well as potentially for DNA barcoding. © 2015 The Authors. Biological Journal of the Linnean Society published by John Wiley & Sons Ltd on behalf of Linnean Society of London, *Biological Journal of the Linnean Society*, 2015, **00**, 000–000.

ADDITIONAL KEYWORDS: genome skimming – high-throughput sequencing – molecular systematics – next-generation sequencing – phylogenetics – phylogenetic signal – repetitive elements.

INTRODUCTION

One of the simplest approaches to using high-throughput sequencing for phylogenetics is to randomly sequence a small proportion of total genomic DNA. The sequences of reads present in these datasets are biased towards sequences with the greatest numbers of copies in the genome (Straub *et al.*, 2012); this includes not only high-copy organellar DNA, such as the plastid and mitochondrial genomes, but also ribosomal DNA and the many kinds of repeats, particularly retrotransposon sequences (Dodsworth *et al.*, 2015). Molecular systematics relies on the alignment of homologous DNA sequences, whether coding or noncoding, and subsequent phylogenetic trees are inferred based on patterns of differences in these alignments. Repetitive elements are not suitable for such analyses in exactly the same way. For example, although retrotransposons have homologous protein domains involved in element mobility, the sequence divergence of these domains between taxa is not sufficient to resolve phylogenetic relationships. What does vary, and in many cases drastically, is the abundance of particular retrotransposons and other repeat types. This abundance of homologous repeats can then be used as a quantitative character for phylogenetic reconstruction.

*Corresponding author. E-mail: steven.dodsworth@qmul.ac.uk

Recent tools have been developed that allow us to analyze, quickly and efficiently, the repetitive portion of the genome from low-coverage genome sequencing data, and then to use these data for phylogenetic inference (Fig. 1) (Novák, Neumann & Macas, 2010; Novák *et al.*, 2013; Dodsworth *et al.*, 2015). This methodology has been shown to be effective for inferring phylogenetic relationships in well-studied groups of angiosperms in several different families (Apocynaceae, Fabaceae, Liliaceae, Orobanchaceae, and Solanaceae). Typically, the method does not work well above the level of genus because there are often no repeats in common (and therefore no shared characters on which to infer phylogenetic relationships). Understanding how repetitive elements could be used in phylogeographical and population genetic studies, as well as in resolving difficult phylogenetic problems at the species-level, is now a focus for future research.

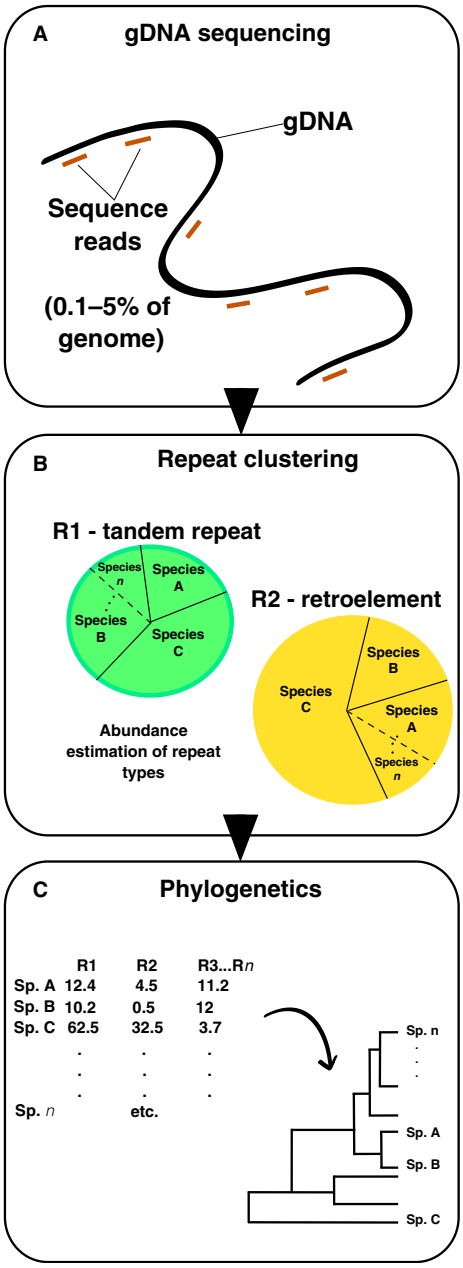

Figure 1. Schematic illustrating the workflow for building trees from repetitive DNA abundances. A, low-coverage genomic DNA sequencing using next-generation sequencing methods (NGS; e.g. Illumina). B, clustering of NGS reads using RepeatExplorer pipeline, resulting in abundance estimates of different repeat families. C, phylogenetic analysis in TNT using cluster abundances as continuous phylogenetic characters.

In the present study, we test the usefulness and power of nuclear repeat regions at inter- and intraspecific levels. We test this using wild and cultivated tomato species, including multiple cultivars as a case study to explore intraspecific variation in genomic repeats and the subsequent performance of these datasets in phylogenetic inference. The wild tomatoes present an excellent case study as a result of the availability of genomic and genetic data, and extensive previous analyses of phylogenetic relationships using plastid markers, low-copy nuclear markers, nuclear ribosomal internal transcribed spacers, and amplified fragment length polymorphisms (Peralta, Spooner & Knapp, 2008; Grandillo *et al.*, 2011). Four informal groups are recognized within the section: (1) '*Lycopersicon* group' with *Solanum lycopersicum*, *Solanum cheesmaniae*, *Solanum galapagense*, and *Solanum pimpinellifolium* (the 'red/orange fruit' clade); (2) '*Arcanum* group' with *Solanum arcanum*, *Solanum chmielewskii*, and *Solanum neorickii* (the 'green fruit' clade); (3) '*Eriopersicon* group' with *Solanum huaylasense*, *Solanum chilense*, *Solanum corneliomulleri*, *Solanum peruvianum*, and *Solanum habrochaites*; and (4) '*Neolycopersicon* group' containing only *Solanum pennellii*, which was considered to be sister to the rest of the section by (Peralta *et al.*, 2008) based on its lack of the sterile anther appendage that occurs as a morphological synapomorphy in *S. habrochaites* and the rest of the core tomatoes. More recent studies using conserved orthologous sequence markers (COSII; Rodriguez *et al.*, 2009) and genome-wide single nucleotide polymorphisms (SNPs) (Aflitos *et al.*, 2014; Lin *et al.*, 2014) have largely supported previous hypotheses with respect to major clades within the tomatoes, although individual species relationships are less clear cut for some taxa. The extent to which multiple evolutionary histories can be recovered through the analysis of different genomic fractions has been explored in tomatoes, and concordance analysis revealed significant discordance possibly as a result of biological processes such as hybridization or incomplete lineage

sorting (Rodriguez *et al.*, 2009). There are no reported polyploids in this clade, and there is not much variation in genome size, although macro- and microgenome rearrangements are reported (Tang *et al.*, 2008; Szinay *et al.*, 2010, 2012; Verlaan *et al.*, 2011).

MATERIAL AND METHODS

Taxa sampled

We sampled material from 20 accessions, including all currently recognized species of the core tomato clade (section *Lycopersicon*) and *Solanum tuberosum* L. (potato) as the outgroup. Representatives of *Solanum* sect. *Lycopersicon* (Table 1) (Peralta, Knapp & Spooner, 2005; Peralta *et al.*, 2008) included in the analyses were: *S. lycopersicum* L., *S. arcanum* Peralta, *S. corneliomulleri* J.F. Macbr., *S. cheesmaniae* (L. Riley) Fosberg, *S. chilense* (Dunal) Reiche, *S. chmielewskii* (C.M.Rick, Kesicki, Fobes & M.Holle) D.M.Spooner, G.J.Anderson & R.K.Jansen, *S. galapagense* S.C.Darwin & Peralta, *S. habrochaites* S.Knapp & D.M.Spooner, *S. huaylasense* Peralta, *S. neorickii* D.M.Spooner, G.J.Anderson & R.K.Jansen, *S. pennellii* Correll, *S. peruvianum* L., and *S. pimpinellifolium* L. Seven accessions representing different cultivars of *S. lycopersicum* were also included.

High-throughput sequence data acquisition

Illumina sequence data from the 100 Tomato Genome Sequencing Consortium (Aflitos *et al.*, 2014) were downloaded from the NCBI Short Read Archive (SRA), with the accession numbers: ERR418040 – *S. lycopersicum* 'Alisa Craig' LA2838A; ERR418039 – *S. lycopersicum* 'Moneymaker' LA2706; ERR418048 – *S. lycopersicum* 'Sonata' LYC1969; ERR418055 – *S. lycopersicum* 'Large Pink' EA01049; ERR418056 – *S. lycopersicum* LYC3153; ERR418058 – *S. lycopersicum* PI129097; ERR418078 – *S. lycopersicum* LYC2962; ERR418093 – *S. arcanum* LA2172; ERR418061 – *S. corneliomulleri* LA0118; ERR418087 – *S. cheesmaniae* LA0483; ERR418098 – *S. chilense* CGN15530; ERR418085 – *S. chmielewskii* LA2663; ERR418121 – *S. galapagense* LA1044; ERR410244 – *S. habrochaites* LYC4; ERR418096 – *S. huaylasense* LA1365; ERR418091 – *S. neorickii* LA0735; ERR410253 – *S. pennellii* LA716; ERR418084 – *S. peruvianum* LA1278; and ERR418082 – *S. pimpinellifolium* LA1584. 454 sequence data for the outgroup *Solanum tuberosum* (ERR023045) were also downloaded from the SRA because appropriate

Table 1. Taxa sampled including accession details, short read archive accession number for genomic data, and genome size (http://data.kew.org/cvalues)

Species	Accession	Cultivar	Short Read Archive accession number	Genome size (1C – Mbp)
Solanum arcanum	LA2172	NA	ERR418093	1125*
Solanum cheesmaniae	LA0483	NA	ERR418087	905
Solanum chilense	CGN15530	NA	ERR418098	1125*
Solanum chmielewskii	LA2663	NA	ERR418085	NA
Solanum corneliomulleri	LA0118	NA	ERR418061	NA
Solanum galapagense	LA1044	NA	ERR418121	905–1002*
Solanum habrochaites	LYC4	NA	ERR410244	905
Solanum huaylasense	LA1365	NA	ERR418096	1125*
Solanum lycopersicum	LYC2962	NA	ERR418078	1002
Solanum lycopersicum	PI129097	NA	ERR418058	–
Solanum lycopersicum	LYC3153	NA	ERR418056	–
Solanum lycopersicum	LYC1969	Sonata	ERR418048	–
Solanum lycopersicum	LA2706	Moneymaker	ERR418039	–
Solanum lycopersicum	LA2838A	Alisa Craig	ERR418040	–
Solanum lycopersicum	EA01049	Large Pink	ERR418055	–
Solanum neorickii	LA0735	NA	ERR418091	NA
Solanum pennellii	LA716	NA	ERR410253	1198
Solanum peruvianum	LA1278	NA	ERR418084	1125
S. pimpinellifolium	LA1584	NA	ERR418082	831
S. tuberosum	DH Kuba 48/6	NA	ERR023045	856

*Values assumed based on previous intraspecific status within other taxa. NA, not available.

Illumina data were unavailable. There are different sequencing biases based on 454 or Illumina technologies (and library preparation protocols) and, ideally, they should not be mixed; however, the outgroup has been clearly defined based on extensive literature and therefore any difference in this one taxon should not have any impact upon the ingroup taxa results.

Dataset preparation and subsampling of read data

SRA files were unpacked into FASTQ using the FASTQ-DUMP executable from the SRA Toolkit. FASTQ files were then filtered with a minimum quality of 10 and converted to FASTA files. For the 454 data, reads were trimmed to 100 bp and filtered. All samples were assumed to have a genome size of approximately 1 Gb based on data available on the Plant C-values Database that shows little variation in genome size between species within section *Lycopersicon* (831–1198 Mbp) (Table 1) (http://data.kew.org/cvalues). Each accession was then sampled for 0.2% of the genome by randomly subsampling each Illumina/454 dataset. This resulted in 20 000 reads of 100 bp per sample from all *Solanum* accessions. The reads in each sample were labelled with a unique nine-character prefix, making a total combined dataset of 400 000 reads. In addition, a further dataset was compiled to test the above assumption that genome size is comparable between species of section *Lycopersicon*. The 20 taxa were randomly shuffled and half were down-sampled to 14 000 reads, representing 0.7 of the original sample. This proportion was chosen because it reflects the genome size variation currently found within the section (approximately 831/1198).

Clustering analysis using RepeatExplorer (RE)

Clustering of Illumina/454 reads was performed using the RE pipeline, implemented in a GALAXY server environment (http://www.repeatexplorer.org) as described in Dodsworth *et al.* (2015). RE clustering was used to identify genomic repeat clusters within each dataset, with default settings (minimum overlap = 55 and cluster size threshold = 0.01%). Briefly, using a BLAST threshold of 90% similarity over 55% of the read length, RE identifies similarities between all sequence reads and then identifies clusters based on a principle of maximum modularity. To identify and discard any potential plastid repeat clusters, we used the *S. lycopersicum* plastid genome (HG975525.1) as a custom repeat database. Plastid repeats are not considered informative in a phylogenetic context because their high abundance is likely linked to the dynamics of photosynthesis in different tissue types and species, and therefore is not indicative of evolutionary history. Hence, plastid regions need to be identified prior to using genomic repeat data in phylogenetic analyses. In our case, none of the clusters were identified as belonging to the plastid genome and hence no regions were removed. Finally, we used RE to identify the 1000 most abundant repeats for phylogenetic analyses, as measured by read numbers per cluster.

Phylogenetic analysis using cluster abundances

The top 1000 most abundant clusters were used to create a matrix for phylogenetic inference. Cluster abundances were used as input characters. To make the cluster abundance values smaller based on requirements of input data for TNT, we divided all abundances by a factor of 18.5 (= largest cluster abundance/65) so that all data would fall within the range 0–65 (as required by the TNT software).

Tree topologies were inferred using maximum parsimony as implemented in the TNT software with continuous character states enabled (Goloboff & Mattoni, 2006; Goloboff, Farris & Nixon, 2008) following settings in Dodsworth *et al.* (2015). Continuous characters are not recoded in any way and are used as 'normal' additive characters, except that count changes can now be non-integer differences (i.e. numerical). Tree searches were performed using implicit enumeration (branch- and bound) with 10 000 symmetric bootstrap (BS) replicates.

To explore reticulation in the dataset, a network approach was employed. SPLITSTREE4 (Huson & Bryant, 2006) was used to create a filtered supernetwork from the 10 000 bootstrap trees from the maximum parsimony analysis, with filtering set at 10% of all input trees (i.e. 1000 trees).

RESULTS

Phylogenetic relationships in *Solanum* sect. *Lycopersicon*

The single most parsimonious tree from our analysis of genomic repeats recovers *S. habrochaites* and *S. pennellii* as the first branching taxa within section *Lycopersicon* (Fig. 2). The '*Eriopersicon* group' (*sensu* Peralta *et al.*, 2008; *S. corneliomulleri*, *S. peruvianum*, *S. huaylasense*, and *S. chilense*) is recovered with high branch support (100 BS). *Solanum neorickii* ('*Arcanum* group') is recovered as sister to all remaining species (99 BS) (Fig. 2). Members of the '*Arcanum* group' (*S. chmielewskii* and *S. arcanum*) are found to be nested within the clade consisting of all the members of the '*Lycopersicon*

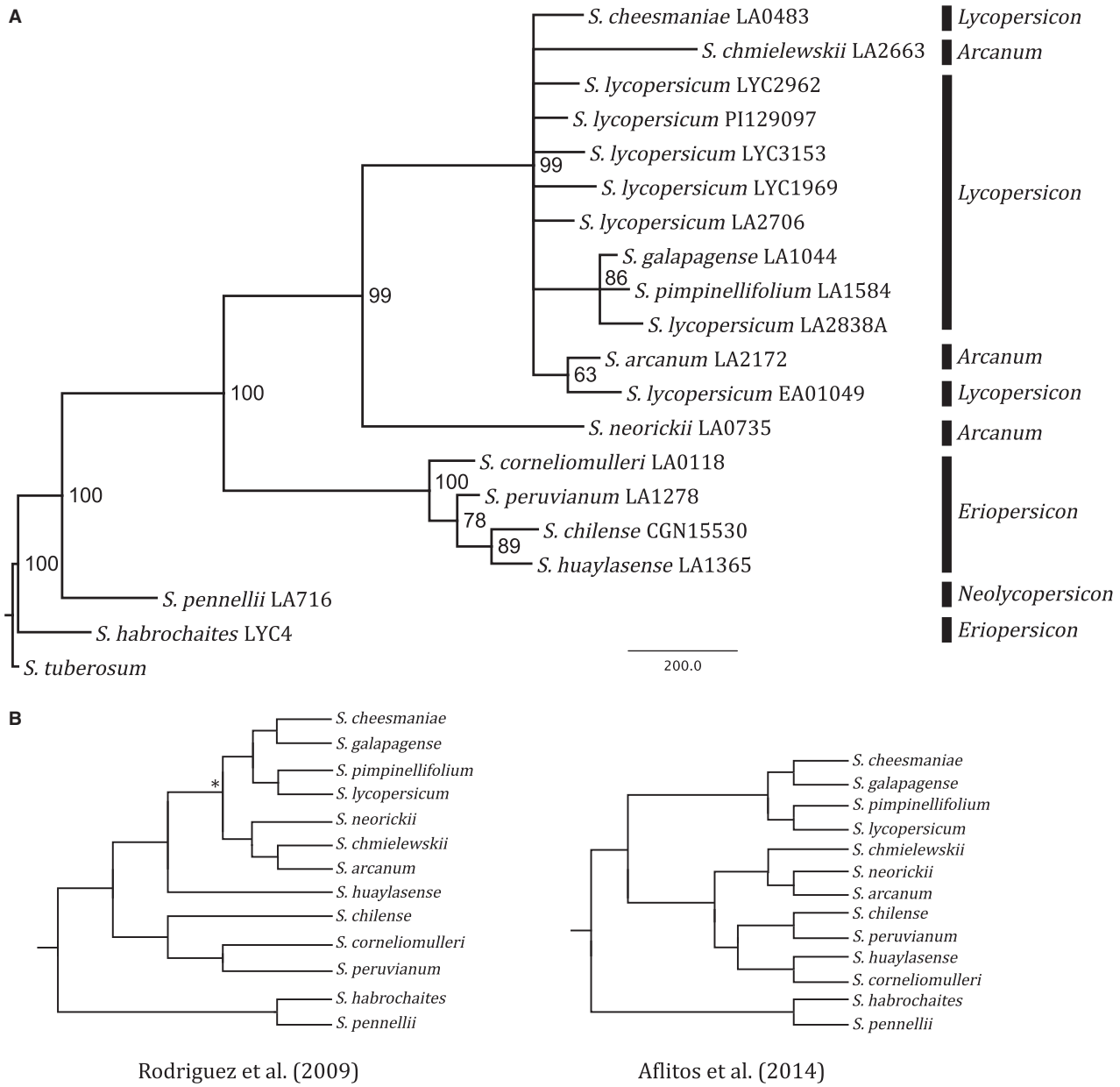

Figure 2. Phylogenetic relationships in *Solanum* section *Lycopersicon*. A, the single most parsimonious tree topology is shown based on abundance values of the 1000 most abundant repeats identified in Illumina/454 next-generation sequencing runs. A total of 0.2% of the genome for each accession was used. Bootstrap values are shown for each node (10 000 symmetric bootstrap replicates). Branch lengths are proportional to numerical step changes in repeat abundances (scale bar). Accession numbers are given for each sample. Current taxonomic grouping is indicated according to informal groups *sensu* Peralta *et al.* (2008). B, summarized phylogenetic hypotheses from Rodriguez *et al.* (2009) and Aflitos *et al.* (2014); low support is indicated by asterisks.

group' (Fig. 2). Our results do not recover a red–orange fruited clade but do find three of the species bearing red or orange coloured fruits in a strongly supported clade (*S. lycopersicum* LA2838A, *S. pimpinellifolium* and *S. galapagense*; 86 BS) within a polytomy including all other red- and orange-fruited accessions.

The additional analysis testing the effect of genome size variation on tree inference is presented in the Supporting information (Fig. S1). Overall, the phylogenetic results are consistent with those based on equal sampling of 20 000 reads, although there are some differences in the large clade containing all *S. lycopersicum* accessions. However, this clade is

still largely unresolved. Network analyses show evidence of reticulation in this clade, as indicated by the presence of splits present in the filtered supernetwork (Fig. 3).

Each accession had a unique combination of repeat percentages, as reflected in the difference in terminal branch lengths. Some accessions also had unique repeat types not found in any other accession (Fig. 4); the largest numbers of unique repeats were found in *S. habrochaites* and *S. pennellii*, with 239 and 301 clusters, respectively, out of the 1000 most abundant clusters. One accession of cultivated *Solanum lycopersicum* (EA01049) had one unique repeat type (Fig. 4).

DISCUSSION

Relationships in *Solanum* section *Lycopersicon*

The taxonomy and estimates of phylogenetic relationships within the core tomato clade (*Solanum* section *Lycopersicon s.s.*) have begun to be stabilized in recent years (Peralta *et al.*, 2005, 2008) using a variety of different markers from both the plastid and nuclear genomes. Rodriguez *et al.* (2009) used a suite of COSII nuclear markers to identify five strongly supported clades within the broader tomato group (incl. sections *Juglandifolia* and *Lycopersicoides*). Their results supported monophyly of section *Lycopersicon* as treated in the present study, and did not resolve either the position of their strongly supported *S. arcanum*+*S. chmielewskii*+*S. neorickii* or the relationships of these species with each other. Our data show a similar lack of resolution regarding these three taxa and, additionally, place them within a large polytomy including all the red- and orange-fruited taxa.

The latest rounds of genome sequencing are likely to add to the robust placement of some species (Aflitos *et al.*, 2014), and the current informal grouping within the section as defined by Peralta *et al.* (2008)

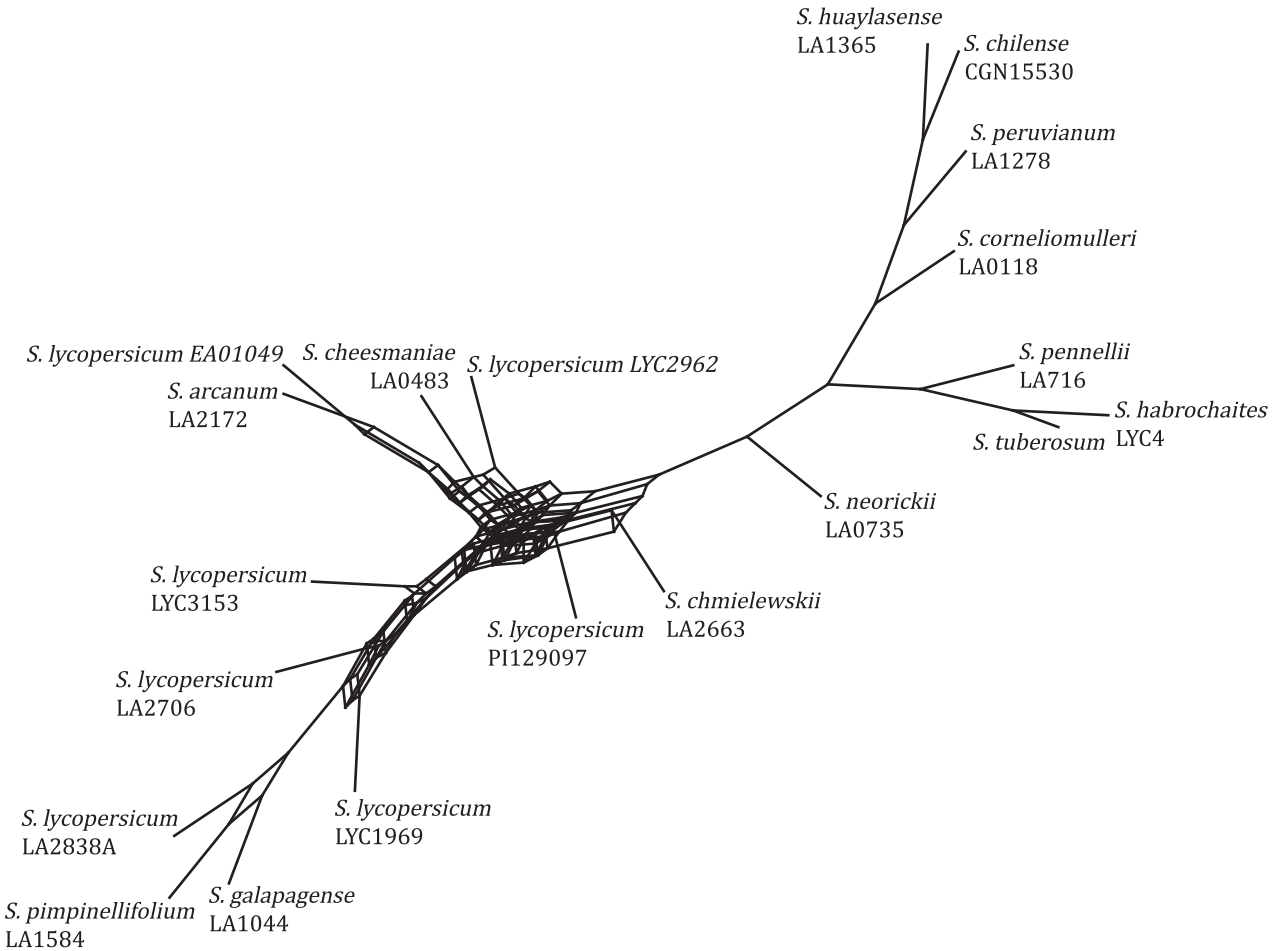

Figure 3. Relationships in *Solanum* section *Lycopersicon* shown as a filtered supernetwork. Splits present in 10% of all bootstrap trees are displayed. Conflict in the network, particularly within the '*Lycopersicon*' group clade, suggests the occurrence of reticulation in the dataset and incongruence between genomic repeat clusters.

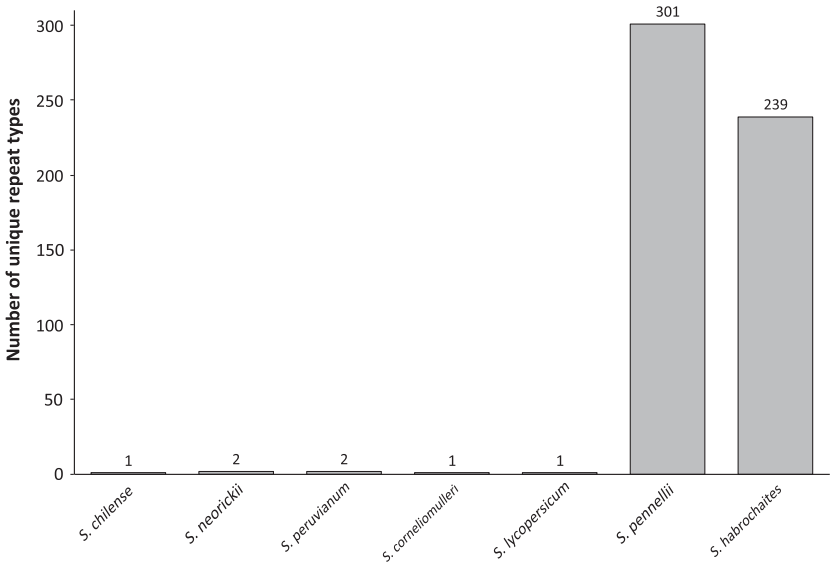

Figure 4. Number of unique repeat types (clusters) for the seven accessions that included them. Note *Solanum lycopersicum*, *Solanum corneliomulleri*, and *Solanum chilense* each have a single unique repeat type.

does appear to reflect what is otherwise known about these taxa/accessions. All recent studies have broadly recovered the four informal groups as defined by Peralta *et al.* (2008): (1) '*Lycopersicon* group' with *S. lycopersicum*, *S. cheesmaniae*, *S. galapagense*, and *S. pimpinellifolium*; (2) '*Arcanum* group' with *S. arcanum*, *S. chmielewskii*, and *S. neorickii*; (3) '*Eriopersicon* group' with *S. huaylasense*, *S. chilense*, *S. corneliomulleri*, *S. peruvianum*, and *S. habrochaites*; and (4) '*Neolycopersicon* group' consisting of *S. pennellii*.

Rodriguez *et al.* (2009) also recovered these groups, although *S. huaylasense* was sister to the '*Arcanum* group', rather than being a member of the '*Eriopersicon* group', and *S. pennellii* and *S. habrochaites* were sister taxa. The groups of Peralta *et al.* (2008) were found to be clades based on genome-wide SNP data (Aflitos *et al.*, 2014), except that, similar to Rodriguez *et al.* (2009), they found *S. pennellii* and *S. habrochaites* to be sister taxa, thus restricting the concept of the '*Eriopersicon* group' to *S. huaylasense*, *S. chilense*, *S. corneliomulleri*, and *S. peruvianum*. This could represent a loss of the anther appendage in *S. pennellii* or a parallel gain of the appendage in *S. habrochaites* and the rest of the core tomatoes. It is clear that further studies on the development of these characters are necessary to examine this result.

In our analyses, most of these major groups were also identified. There were three notable differences: (1) *Solanum habrochaites* was recovered as sister to the rest of the section not as sister to *S. pennellii*; (2) two species of the '*Arcanum* group', *S. chmielewskii*, and *S. arcanum*, were nested within the '*Lycopersicon* group'; and (3) *Solanum neorickii* was recovered as sister to the '*Lycopersicon* group' (including *S. arcanum* and *S. chmielewskii*).

Our recovery of *S. habrochaites* as sister to the rest of the core tomatoes differs from the results based on genome-wide SNPs (Aflitos *et al.*, 2014; Lin *et al.*, 2014), although it is perhaps not unexpected given the relatively unstable position of *S. habrochaites* and *S. pennellii* in previous analyses (Peralta *et al.*, 2008). It highlights the need for further developmental analysis of the sterile anther appendage long considered to be the synapomorphy of the core tomatoes (Peralta *et al.*, 2008).

The nesting of *S. chmielewskii* and *S. arcanum* within the '*Lycopersicon* group' and the sister relationship of *S. neorickii* to this larger group are more unexpected results that require further investigation. The analyses of Aflitos *et al.* (2014) provided strong support for the '*Arcanum* group', including *S. arcanum*, *S. chmielewskii*, and *S. neorickii* and for its sister relationship with the '*Eriopersicon* group' (minus *S. habrochaites*), as reported by Rodriguez *et al.* (2009). The unusual placement of *S. arcanum* and *S. chmielewskii* in our phylogenetic analysis may be the result of the repetitive portion of the genome evolving under non-neutral processes, such as targeted repeat amplification/deletion or potentially horizontal gene transfer. Further characterization of repeat dynamics and additional taxon sampling could help to clarify this. The polytomy involving these taxa and the cultivated tomatoes could also be the result of extensive use of wild species in tomato

breeding in the past, where gene regions from wild species have been introgressed in different cultivars of *S. lycopersicum* (Grandillo *et al.*, 2011). The filtered supernetwork based on 10% of all bootstrap trees (Fig. 3) shows clear evidence of potential reticulation and non-treelike evolution in this clade. Thus, the placement of *S. arcanum* and *S. chmielewskii* could reflect the use of these wild species in previous tomato breeding.

The reference tomato genome ('Heinz 1706'; Tomato Genome Consortium, 2012) contains multiple introgressions from *S. pimpinellifolium*. Lin *et al.* (2014) found exotic fragments containing resistance genes in inbreeding lines, processing tomatoes, and fresh market hybrids that remained intact after several generations of backcrossing. This prospect of introgression, as with hybridization, would affect the analyses of genomic repeats, with some repeats specific to one parental lineage and some to the other parental lineage (or introgressed species). Network approaches do indeed provide some evidence for the involvement of introgression and/or hybridization in such scenarios, as indicated in the present study. However, this is complex and variable depending on the timeframe within which these processes occurred (e.g. polyploids of *Nicotiana*; Dodsworth *et al.*, 2015).

Genome skimming for molecular systematics

The 'genome skimming' approach (*sensu* Straub *et al.*, 2012) involves low-coverage sequencing of genomic DNA using high-throughput technologies such as Illumina. The resulting data represents random sequences distributed throughout the genome but, because the coverage is low, the data will only represent the fraction of the genome that is in relatively high copy number. Notably, this includes ribosomal DNA from the nuclear genome (present in typically hundreds or thousands of copies) and organellar DNA (the plastid and mitochondrial genomes). A current surge in using genome skimming approaches focuses on the plastid and mitochondrial sequences that can be assembled from low-coverage Illumina sequence data (Kane *et al.*, 2012; Steele *et al.*, 2012; Haran, Timmermans & Vogler, 2013; Njuguna *et al.*, 2013; Gillett *et al.*, 2014) and this approach is proving successful in both animals and plants. It has advantages over other methods of high-throughput sequencing for phylogenetics because it requires no prior enrichment or complicated laboratory procedures; the downside is that it is currently limited by the cost of library preparation kits (unless custom protocols are developed). Reduced representation sequencing such as RADseq (Wagner *et al.*, 2013) and hybridization/pull-down methods (Guschanski *et al.*, 2013) both require extensive optimization and/or molecular laboratory work prior to the actual sequencing. A further advantage to genome skimming approaches is that they produce several datasets in one run: plastid, mitochondrial, and nuclear, which provide separate forms of evidence from 3 genomes that complement one another. In terms of nuclear markers, repetitive elements can be easily quantified using the RE pipeline and used in phylogeny reconstruction as shown in the present study. This provides additional evidence that may complement organellar and nuclear ribosomal cistron analyses.

Genome skimming, and in particular utilizing genomic repeats, may be useful for tapping into the genomic resources held in museum collections. Such DNA is often highly degraded, either simply because of age or a combination of age and the method by which specimens were initially dried. Collections have also been subject to various chemical treatments which impact upon DNA quality. These factors have previously hindered polymerase chain reaction success and still limit the availability of some high-throughput sequencing methodologies (such as amplicon sequencing or pull-down approaches). However, because genomic repeats are the most abundant sequences in genomic DNA samples, present in many copies, these will likely be adequately represented even in the most degraded of museum samples.

Conclusions: future prospect for genomic repeats

In the *Solanum* example reported, *S. lycopersicum* samples formed a strongly supported group that included *S. cheesmaniae*, *S. galapagense*, and *S. pimpinellifolium* ('red/orange' fruited clade), as found in all previous studies (Peralta *et al.*, 2008; Aflitos *et al.*, 2014; Lin *et al.*, 2014), which indicates the utility of these data as phylogenetic markers at the intraspecific level. Despite this result, there were two unexpected placements within the *Lycopersicon* clade that require further investigation. Nonetheless, this is an important first result presenting the use of these data for low-level phylogenetic studies, such as phylogeography and investigations of widespread species and species complexes.

Genomic repeats could also serve as markers for DNA barcoding, although a crucial first step will be to determine whether there is a 'barcoding gap' (Meyer & Paulay, 2005; Meier, Zhang & Ali, 2008) in further datasets that include many samples of each species. Future developments including model-based inference in a custom Bayesian framework will add rigour to the analysis of these quantitative characters;

this method can then be fully extended to some of the applications proposed in the present study at the intraspecific and interspecific levels.

ACKNOWLEDGEMENTS

This work was supported by a NERC studentship to SD. We thank two anonymous reviewers for their helpful comments that improved the manuscript. We would also like to thank Sven Buerki, Bill Baker, and other organizers for their support at the 'Collections-based research in the genomic era' meeting held at the Linnean Society, April 2014.

REFERENCES

Aflitos S, Schijlen E, de Jong H, de Ridder D, Smit S, Finkers R, Wang J, Zhang G, Li N, Mao L, Bakker F, Dirks R, Breit T, Gravendeel B, Huits H, Struss D, Swanson-Wagner R, van Leeuwen H, van Ham RCHJ, Fito L, Guignier L, Sevilla M, Ellul P, Ganko E, Kapur A, Reclus E, de Geus B, de van Geest H, te Lintel Hekkert B, van Haarst J, Smits L, Koops A, Sanchez-Perez G, van Heusden AW, Visser R, Quan Z, Min J, Liao L, Wang X, Wang G, Yue Z, Yang X, Xu N, Schranz E, Smets E, Vos R, Rauwerda J, Ursem R, Schuit C, Kerns M, van den Berg J, Vriezen W, Janssen A, Datema E, Jahrman T, Moquet F, Bonnet J, Peters S. 2014. Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *The Plant Journal* **80:** 136–148.

Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, Piednoël M, Weiss-Schneeweiss H, Leitch AR. 2015. Genomic repeat abundances contain phylogenetic signal. *Systematic Biology* **64:** 112–126.

Gillett CPDT, Crampton-Platt A, Timmermans MJTN, Jordal BH, Emerson BC, Vogler AP. 2014. Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution* **31:** 2223–2237.

Goloboff PA, Mattoni CI. 2006. Continuous characters analyzed as such. *Cladistics* **22:** 589–601.

Goloboff PA, Farris JS, Nixon KC. 2008. TNT, a free program for phylogenetic analysis. *Cladistics* **24:** 774–786.

Grandillo S, Chetelat R, Knapp S, Spooner D, Peralta E, Cammareri M, Perez O, Termolino P, Tripodi P, Chuisano ML, Ercolano MR, Frusciante L, Monti L, Pignone D. 2011. *Solanum* section *Lycopersicon*. In: Kole C, ed. *Wild crop relatives: genomic and breeding resources*. Berlin & Heidelberg: Springer Verlag, 129–215.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT, Lenglet G, Mayer F, Savolainen V. 2013. Next-generation museomics disentangles one of the largest primate radiations. *Systematic Biology* **62:** 539–554.

Haran J, Timmermans MJTN, Vogler AP. 2013. Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. *Molecular Phylogenetics and Evolution* **67:** 156–166.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23:** 254–267.

Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, Cronk Q. 2012. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* **99:** 320–329.

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, Zhang Z, Lun Y, Li S, Wang X, Huang Z, Li J, Zhang C, Wang T, Zhang Y, Wang A, Zhang Y, Lin K, Li C, Xiong G, Xue Y, Mazzucato A, Causse M, Fei Z, Giovannoni JJ, Chetelat RT, Zamir D, Städler T, Li J, Ye Z, Du Y, Huang S. 2014. Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics* **46:** 1220–1226.

Meier R, Zhang G, Ali F. 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the 'barcoding gap' and leads to misidentification. *Systematic Biology* **57:** 809–813.

Meyer CP, Paulay G. 2005. DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology* **3:** e422.

Njuguna W, Liston A, Cronn R, Ashman TL, Bassil N. 2013. Insights into phylogeny, sex function and age of *Fragaria* based on whole chloroplast genome sequencing. *Molecular Phylogenetics and Evolution* **66:** 17–29.

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11:** 378.

Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29:** 792–793.

Peralta IE, Knapp S, Spooner DM. 2005. New species of wild tomatoes (*Solanum* Section *Lycopersicon* : Solanaceae) from Northern Peru. *Systematic Botany* **30:** 424–434.

Peralta IE, Spooner DM, Knapp S. 2008. Taxonomy of wild tomatoes and their relatives (*Solanum* sect. *Lycopersoides*, sect. *juglandifolia*, sect. *Lycpersicon*; Solanaceae). *Systematic Botany Monographs* **84:** 1–186.

Rodriguez F, Wu F, Ané C, Tanksley S, Spooner DM. 2009. Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? *BMC Evolutionary Biology* **9:** 191.

Steele PR, Hertweck KL, Mayfield D, McKain MR, Leebens-Mack J, Pires JC. 2012. Quality and quantity of data recovered from massively parallel sequencing: examples in Asparagales and Poaceae. *American Journal of Botany* **99:** 330–348.

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* **99:** 349–364.

Szinay D, Bai Y, Visser R, De Jong H. 2010. FISH applications for genomics and plant breeding strategies in tomato and other solanaceous crops. *Cytogenetic and Genome Research* **129:** 199–210.

Szinay D, Wijnker E, van den Berg R, Visser RGF, de Jong H, Bai Y. 2012. Chromosome evolution in *Solanum* traced by cross-species BAC-FISH. *New Phytologist* **195:** 688–698.

Tang X, Szinay D, Lang C, Ramanna MS, Van Der Vossen EAG, Datema E, Lankhorst RK, De Boer J, Peters SA, Bachem C, Stiekema W, Visser RGF, De Jong H, Bai Y. 2008. Cross-species bacterial artificial chromosome-fluorescence in situ hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. *Genetics* **180:** 1319–1328.

Verlaan MG, Szinay D, Hutton SF, De Jong H, Kormelink R, Visser RGF, Scott JW, Bai Y. 2011. Chromosomal rearrangements between tomato and *Solanum chilense* hamper mapping and breeding of the TYLCV resistance gene Ty-1. *Plant Journal* **68:** 1093–1103.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* **22:** 787–798.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Figure S1. Phylogenetic tree based on randomized down-sampling half the dataset (10 taxa with 14 000 reads versus remaining 10 taxa with 20 000 reads).

