agronomy

MDPI

Article

Multi-Species Transcriptome Assemblies of Cultivated and Wild Lentils (*Lens* sp.) Provide a First Glimpse at the Lentil Pangenome

Juan J. Gutierrez-Gonzalez *, Pedro García, Carlos Polanco, Ana Isabel González, Francisca Vaquero, Francisco Javier Vences, Marcelino Pérez de la Vega and Luis E. Sáenz de Miera

Departamento de Biología Molecular, Universidad de León, Campus de Vegazana s/n, 24071 León, Spain; pgarg@unileon.es (P.G.); carlos.polanco@unileon.es (C.P.); aigonc@unileon.es (A.I.G.); fvaqr@unileon.es (F.V.); fjvenb@unileon.es (F.J.V.); mperv@unileon.es (M.P.d.l.V.); luis.saenzdemiera@unileon.es (L.E.S.d.M.)

* Correspondence: jgutg@unileon.es; Tel.: +34-987-293-195

Abstract: Lentils (*Lens* sp.) are one of the main sources of protein for humans in many regions, in part because their rusticity allows them to withstand semi-dry climates and tolerate a wide spectrum of pests. Both are also highly sought-after attributes to face climate change. Wild accessions, rather than cultivated varieties, are typically the holders of most influential alleles for rusticity traits. However, most genomic and transcriptomic research conducted in lentils has been carried out on commercial accessions (*L. culinaris*), while wild relatives have been largely neglected. Herein, we assembled, annotated, and evaluated the transcriptomes of eight lentil accessions, including the cultivated *Lens culinaris* and the wild relatives: *L. orientalis*, *L. tomentosus*, *L. ervoides*, *L. lamottei*, *L. nigricans*, and two *L. odemensis*. The assemblies allowed, for the first time, a comparison among different lentil taxa at the coding sequence level, providing further insights into the evolutionary relationships between cultivated and wild germplasm and suggesting a grouping of the seven accessions into at least three conceivable gene pools. Moreover, orthologous clustering allowed a first estimation of the lentil pan-transcriptome. It is composed of 15,910 core genes, encoded in all accessions, and 24,226 accessory genes. The different pan-transcriptome clusters were also screened for Pfam-domain enrichment. The present study has a high novelty, as it is the first pan-transcriptome analysis using six wild species in addition to cultivated species. Because of the amount of transcript sequences provided, our findings will greatly boost lentil research and assist breeding efforts.

Keywords: lentils; transcriptome; lentil wild relatives; pangenome; pan-transcriptome; *Lens culinaris*

Citation: Gutierrez-Gonzalez, J.J.; García, P.; Polanco, C.; González, A.I.; Vaquero, F.; Vences, F.J.; Pérez de la Vega, M.; Sáenz de Miera, L.E. Multi-Species Transcriptome Assemblies of Cultivated and Wild Lentils (*Lens* sp.) Provide a First Glimpse at the Lentil Pangenome. *Agronomy* **2022**, *12*, 1619. https://doi.org/10.3390/agronomy12071619

Academic Editors: Thomas Hartwig and Diego Rubiales

Received: 12 April 2022
Accepted: 29 June 2022
Published: 5 July 2022

1. Introduction

Lentils are one of the main sources of carbohydrates, protein, and iron for humans in many areas worldwide [1]. They also contribute to maintaining a healthy lifestyle, as they are an important dietary source of minerals, vitamins, antioxidants, and soluble fiber. Recent data also suggest that lentil consumption is growing in importance [2]. In the last couple of decades, the total world production has doubled, growing from 2,949,018 tons in 1999 to 5,734,201 tons in 2019 [3]. Apart from reasons related to adopting a healthy eating pattern, other factors, such as the need to feed a steadily growing world population or the need to adapt to a changing climate, may be at play. In fact, lentils are traditionally valued for their rusticity and tolerance to drought, one of the most recognized effects of climate change in many regions of the world. It is important to note that the most frequent way of growing lentils is as a rainfed crop.

Because conflicting findings are sometimes the outcome of classifications based on both genetic data and hybridization among the different taxa, the taxonomy of *Lens* continues to be debated [2,4]. Indeed, depending on the criteria followed, the genus *Lens* has been split between a minimum of four and a maximum of seven species. These are: *L. culinaris*, *L. orientalis*, *L. tomentosus*, *L. odemensis*, *L. lamottei*, *L. ervoides*, and *L. nigricans*. However, another

widely accepted taxonomy recognizes the *orientalis*, *tomentosus,* and *odemensis* as subspecies of *L. culinaris* [5,6]. Leaving taxonomy disputes apart, there is a somewhat broader agreement when the grouping is driven by their ability to make crosses. In this regard, the genus *Lens* is often divided into four related "gene pools", although three gene pools have been also proposed [1,6]. The primary gene pool would be comprised by *L. culinaris*, *L. orientalis*, and *L. tomentosus*; the secondary by *L. odemensis* and *L. lamottei*; the tertiary by *L. ervoides*; and the quaternary by *L. nigricans*. Members of each pool can typically produce seeds following interspecific crossing. Commercially cultivated lentils are almost exclusively from the primary gene pool. Non-*culinaris* lentils are nonetheless major sources of genetic material in breeding programs.

Lentils went through a marked genetic bottleneck during domestication, which, together with the monopoly that certain elite varieties have in growing regions and their high percentage of self-pollination, led to an extensive LD and a narrow genetic base of the breeding material [7,8]. In the search for unexploited variability for breeding programs, wild crop relatives are often selected as sources of novel alleles to be introgressed into elite cultivars. For instance, wild relatives of cultivated lentils are native to areas that suffer from frequent droughts, and thus, they could bear beneficial alleles for drought resistance. In general, wild relatives of crops are considered a good source of biotic and abiotic stress resistance genes. Thus, deepening the knowledge of wild lentils at the genomic and transcriptome levels emerges as a necessity for the success of breeding programs. Genomic studies are more straightforward when a reference genome is available. Regrettably, lentils have a large and repetitive genome of about 4 Gb, which has greatly hampered its assembly. The first published draft of the genome [9] has boosted genomic studies, despite being far from complete. Several improvements over the last years have led to the *Lens culinaris* cv 'CDC Redberry' v2.0, which is currently the newest polished draft version of the genome [10]. Nevertheless, improved sequencing platforms have constituted an ultimate tool to advance genomic studies, allowing high-throughput genome-wide studies in species with large genomes.

In one of the few comprehensive studies of the genus *Lens*, Gorim and collaborators [11] performed a phenotypic evaluation of both cultivated and wild lentil species to gauge their ability to be used as an unexploited genetic resource to improve drought tolerance in commercial lentil varieties. The authors tested root and shoot traits of diverse genotypes and found that root distribution into different soil horizons varied among wild lentil genotypes. They also found that wild lentils employed diverse strategies such as delayed flowering, reduced transpiration, short plant height, and deep root systems to either escape, evade, or tolerate drought conditions. In a changing climate scenario, drought has the potential to be one of the most devastating abiotic stresses that affect crops. Yet, the use of genome-wide 'omics' approaches in lentils to highlight drought-related genes is largely unexplored.

In fact, despite some fruitful steps forward, lentil 'omics' are lagging behind those of other legumes. A couple of studies have attempted to assess gene diversity in lentils [6,12]. They used high-confidence SNP markers found in collections of wild and cultivated accessions from diverse geographical regions to understand the genetic relationships among different species/subspecies. Nevertheless, while SNPs in the number of thousands can be informative, they do not explore the diversity enclosed in longer transcript sequences. Certainly, the length and duplication of the lentil genome, together with the lack of a reference-quality genome assembly, are obstacles to genomic research. Until a reference-quality genome sequence becomes available, de novo transcriptome assemblies are strategic in gene discovery, EST sequencing, marker finding, and transcript profiling [13,14]. Lentil transcriptomics has allowed, for instance, the unraveling of plant–pathogen interactions [15,16], understanding the role of heat-responsive genes in regulatory mechanisms involving different combinations of heat stress [17,18], or finding candidate genes for improving drought tolerance [8,19].

Not surprisingly, most of the studies at the transcriptome level have been conducted on *L. culinaris*, and very few on the other species of *Lens*. In one of the few exceptions [20], researchers assembled the transcriptomes of *L. culinaris* and *L. nigricans* to study differential expressed genes and pathways under Al^{3+} stress. They completed the study with trait-associated SNPs and SSR markers that could be used in breeding programs to improve resistance to Al^{3+} in lentils and

other crops. Recently, Garcia-Garcia et al. [21] were able to highlight the pathways that are most affected following *Ascochyta lentis* pathogen infection by using massive analysis of cDNA ends (MACE). They explored *L. orientalis* as the source of resistance.

Even though one-accession transcriptomics are informative, they do not explore the collective gene repertoire of a certain species, known as the pangenome. It consists of a core genome, containing sequences shared between all individuals of those species, and an 'accessory' genome, shared only by some of them. While core genes are conserved, accessory loci comprise a significant portion of a plant's genetic diversity [22]. It has been suggested that using accessions from all available species of a certain genus, including wild relatives, may produce a more complete and more comprehensive pangenome [23]. In the era of high-throughput sequencing technologies, pangenome-based analyses have emerged as important tools to unravel the complex interrelationship between biotic/abiotic stressors and plants [24], supplying invaluable information about gene families involved in important agronomic traits [25,26]. In the absence of well-annotated genome assemblies, pan-transcriptomes can also be used to estimate the pangenomes at an affordable cost [22,27–29].

Herein, we used lentil plantlets to carry out, for the first time, the transcriptome assemblies of diverse genotypes distributed throughout the known taxonomic variability within the genus *Lens*. The main objectives were to (i) assemble and annotate the transcriptomes of eight different lentil accessions, (ii) make a comparative analysis among them, and (iii) estimate, for the first time, the lentil pan-transcriptome. The amount of data generated will pave the path for using wild relatives in lentil breeding programs.

2. Results

The transcriptomes of eight lentil genotypes from different taxa within the genus *Lens* were sequenced at high depth and de novo assembled (Supplementary Dataset S1). The accessions included 'Alpo', a Spanish cultivar of *L. culinaris*, the common lentil, and seven wild relatives (see Materials and Methods for a detailed description). The total number of reads sequenced varied between 50.5 and 134.2 million, for a coverage between 42× and 84×. A summary of the assemblies and quality metrics is presented in Supplementary Table S1. For the sake of conciseness, only the species name, but not the genus *Lens*, is used hereafter.

2.1. Transcriptome Assemblies and Qualities

The assembler groups transcripts into putative gene clusters based on shared sequence content. Such clusters are composed of one or more 'transcript isoforms', which loosely correspond to the several splicing events within a gene. The number of total assembled transcripts per accession varied between 84,196 (*orientalis*) and 115,224 (*culinaris*). The transcriptome of *culinaris* also contained the maximum number of assembled genes (58,375); however, the *nigricans* transcriptome had the minimum number (46,742).

To evaluate the quality and completeness of the assemblies, several metrics were taken (Supplementary Table S1). First, we assessed how many of those transcripts corresponded to full-length or nearly full-length (>80%) unique proteins by comparing the gene set of each assembly with the Uniprot-Viridiplantae plants database [30] and the *L. culinaris* CDC Redberry draft genome [10]. Discontinuous blast hit alignments to the same protein were grouped, and the aggregated hit length was calculated. Only combined hits larger than 80% of the protein length were considered. The Uniprot proteins represented in the transcriptomes ranged between 22,243 (*culinaris*) and 19,313 (*tomentosus*). Similarly, between 16,052 (*culinaris*) and 14,394 (*tomentosus*) CDC Redberry annotated coding sequences (CDS) were assembled. Transcriptome completeness was also estimated with BUSCO [31], which recovers single-copy orthologs that are present across higher taxonomic groups. The percentage of complete genes, single-copy or duplicated, present in the transcriptomes ranged between 87.1% (*tomentosus*) and 91.4% (*culinaris*) (Supplementary Table S1 and Figure S1).

To further benchmark the assemblies, several other metrics were computed. First, the percentage of the reads that mapped back to each assembly was above 99.1% for all the assemblies, of which more than 96.2% mapped as proper pairs at least once. Second, the

N50 value for the transcripts ranged between 1906 bp (*ervoides*) and 2206 bp (*culinaris*). Third, the average transcript length was between 1214 and 1396 bp, which correspond to *ervoides* and *culinaris*, respectively. Taken together, the results suggest a high quality and completeness of all transcriptomes. Results also indicate that the transcriptomes of *culinaris* and *odemensis* are the most comprehensive, as they not only contain the largest numbers of genes and transcripts, but they also have the highest percentages of recovered complete BUSCOs.

2.2. Comparison among Transcriptomes and with the Reference

To assist transcriptome comparisons, the *L. culinaris* cv 'CDC Redberry' draft genome assembly [10] was used to anchor the de novo assembled transcriptomes with a common reference (Supplementary Data). For this, we first aimed to establish what would be the best transcript sequence representative to use in the comparisons. Here, we employ the term 'transcript isoform' or 'transcript' to refer to each individual sequence in a transcriptome assembly. Similar transcript isoforms are grouped together into genes by the assembler, based on sequence variations detected as the assembly process progresses. To assess which transcript isoform best represents each gene's identity, three types of transcript representatives were tested: (i) the longest transcript isoform from each gene; (ii) the isoform that was most expressed (computed as TPM); and (iii) the SuperTranscript, a collapsed version of all transcript isoforms within a gene [32]. The *culinaris* transcriptome assembly was selected for this assessment of the best representative, as it had the largest number of transcripts, genes, and complete BUSCO genes. First, those candidate representatives were aligned to the CDC Redberry genome, and the density functions of each alignment identity percentage were plotted (Figure 1). Similarly, density functions were also plotted for the aligned lengths of query transcripts (*qlen*). Overall, the density functions point at the most expressed transcript isoform as the best transcript representative, followed by the longest and the SuperTranscript.

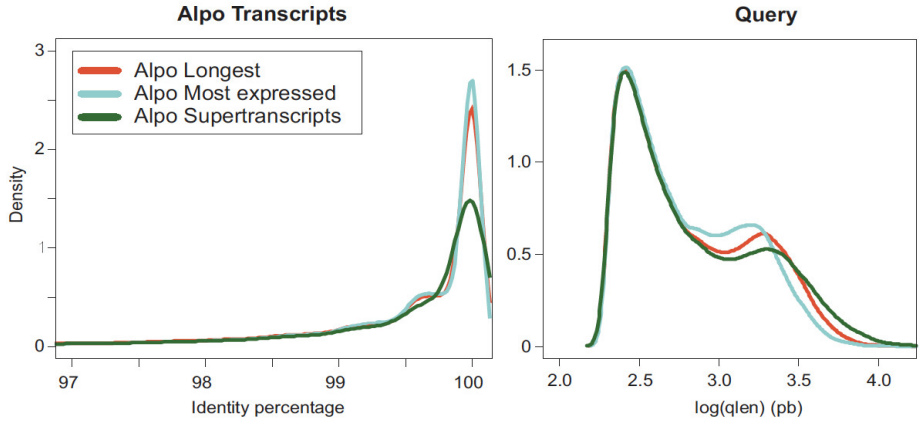

Figure 1. Alignments to the Lens culinaris reference. Alignments of three types of transcript representatives of the Alpo (*Lens culinaris*) transcriptome to the CDC Redberry reference genome. These representatives are: the longest transcript isoform of each gene; the isoform that is most expressed; and the SuperTranscripts. The density functions of the identity percentages of the alignments (**left**) as well as the aligned query lengths (*qlen*) (**right**).

Second, we determined to what extent each candidate transcript representative aligned to their homologous sequence in the CDC Redberry reference. For this, query lengths (*qlen*) of each candidate representative in the *culinaris* transcriptome were plotted against the alignment length (*alen*) of the corresponding homologous sequence in the draft reference. This was performed for the longest, the most expressed, and the SuperTranscript representatives of each gene (Figure 2). In this type of plot, the diagonal represents the perfect alignment, that is, when the entire sequence of the query aligns to the reference with no gaps. Conversely, the larger the distance of a datapoint to the diagonal, the more divergent the query and the reference are. This is because

only a fraction of the query aligns. A metric was computed based on the averaged summation of all distances from each datapoint to the diagonal. Results revealed distances of 17.5, 27.3, and 65.4 for the most expressed, the longest, and the SuperTranscript, respectively. Taking this and the previous breakdown together, the transcript isoform with the highest expression appears to be the best representative of each gene.

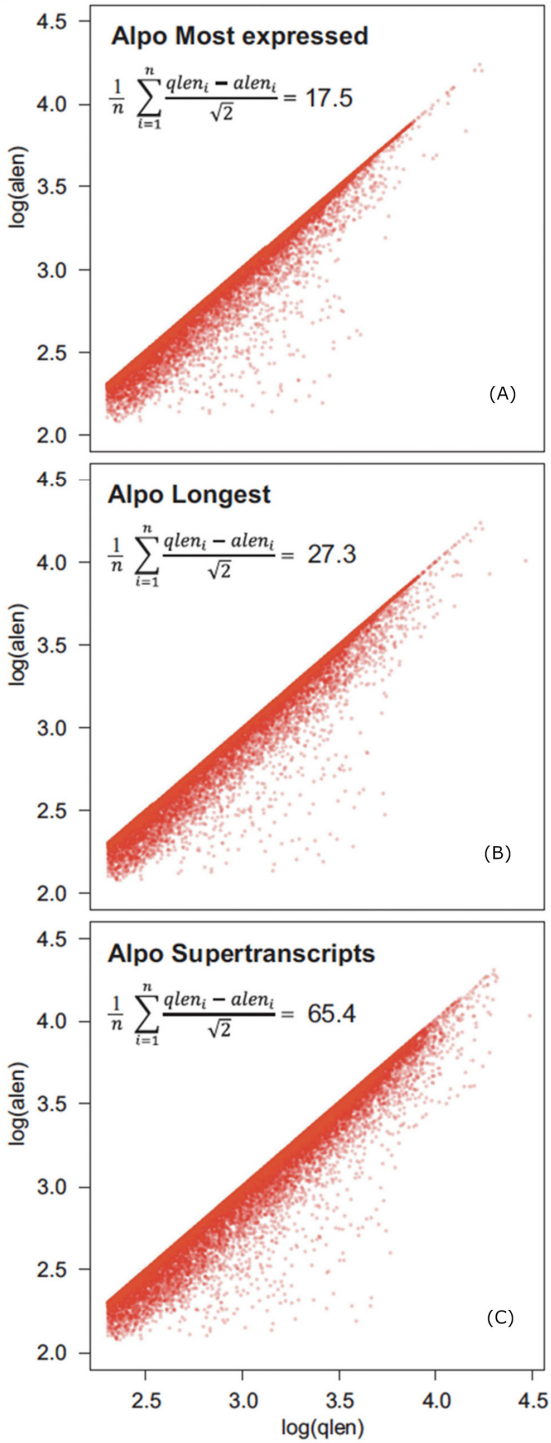

Figure 2. Alpo-CDC Redberry length alignments. The query lengths (*qlen*) of the most expressed transcript (**A**), the longest transcript (**B**), and the SuperTranscript (**C**) were plotted against the length of the aligned sequence in CDC Redberry (*alen*). An averaged summation function was also computed with the distances of each point to a perfect alignment (diagonal).

Subsequently, once the best representative was established, the transcriptome assemblies of all eight lentil accessions were aligned to the CDC Redberry draft genome assembly using the transcript with the highest expression. According to the density functions of their respective alignment identities (Figure 3A), *culinaris* has the highest similarity to the CDC Redberry, with *orientalis* and *tomentosus* next, followed by *ervoides*, *odemensis1* and *odemensis2*, *lamottei,* and *nigricans*. It is worth noting that the graphs of *nigricans–lamottei*, and *odemensis1–odemensis2* overlap with each other to a great extent. The graphs of *orientalis* and *tomentosus* are also almost identical. When the plotted density function is that of the query alignment length, the separation of accessions is less clear (Figure 3B). Here, two peaks can be appreciated. The absolute maximum, around 250–280 bp, is likely a product of the redundant and fragmented transcript sequences usually obtained from the RNA-seq assemblies. A local maximum, centered around 1700–1800 bp, probably reflects the average full transcript length.

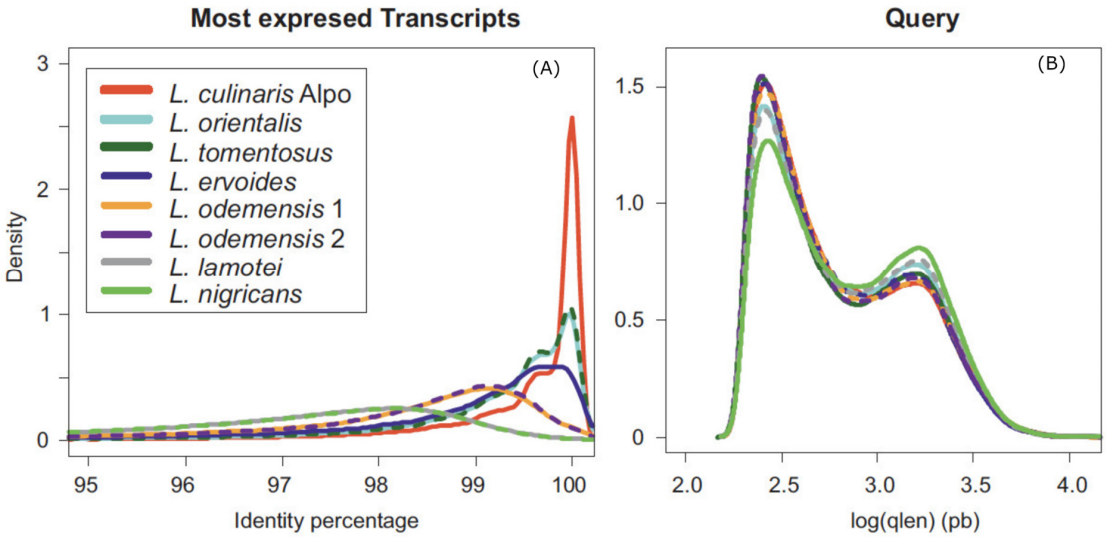

Figure 3. Density functions. (**A**) Density functions for the eight accessions (y-axis) are plotted against the percentage of identity between the most expressed transcripts of each accession and the Lens culinaris reference CDC Redberry. (**B**) Density functions of the length of the alignments between each query and CDC Redberry.

A similar analysis to that displayed in Figure 2 for the *culinaris* accession was conducted for the other seven transcriptomes (although only one of the two *odemensis* accessions is shown). Thus, the query lengths (*qlen*) for the most expressed representative of each transcriptome were plotted against the length of the aligned homologous sequence in CDC Redberry (*alen*) (Figure 4). The distance metrics provided the following average values: 19.0 (*tomentosus*), 20.8 (*orientalis*), 24.6 (*ervoides*), 29.3 and 33.2 (*odemensis2* and *odemensis1*, respectively), 41.8 (*lamottei*), and 45.0 (*nigricans*). Taken together with the alignment density functions (Figure 3), these results further suggest that the transcriptome of *culinaris* has the most similar sequences to CDC Redberry, followed by *orientalis* and *tomentosus*. In contrast, the transcripts of *lamottei* and *nigricans* were the least similar. Lastly, *ervoides* and the two *odemensis* display an intermediate degree of similarity.

To gather more information on the variability and dispersion of the transcriptome data, box plots of the relative distributions of the query alignment lengths over the reference were drawn without outliers (Figure 5). Remarkably, the median of the distribution (Q2) was lower than 0.05 for all accessions, which reveals that the most typical alignment covers more than 95% of the length of the query. The highest percentage corresponds to *culinaris* and the lowest to *nigricans*. Overall, the plots display a marked positive skewness of the data, with data points higher than the median more spread out than the ones lower. Indeed, the fourth quartile (Q4) collects most variability, followed by the third quartile (Q3). The long upper whiskers in the plots indicate high alignment dissimilarity amongst the values in Q4, which corresponds to alignments with the highest differences between *qlen* and *alen*. Conversely, they are very similar for the lowest

quartile group, or alignments with the smallest *qlen–alen* values. In fact, the second (Q2) and first (Q1) quartiles are barely noticeable in the graph. Therefore, the interquartile range (IQR), the distance between the upper (Q3) and lower (Q1) quartiles, is driven primarily by the Q3 quartile. Taken together, the results from the box plots suggest that the transcript representatives of *culinaris*, *orientalis,* and *tomentosus* are the most similar to their homologues in the CDC Redberry reference draft, while *lamottei* and *nigricans* are the least similar.

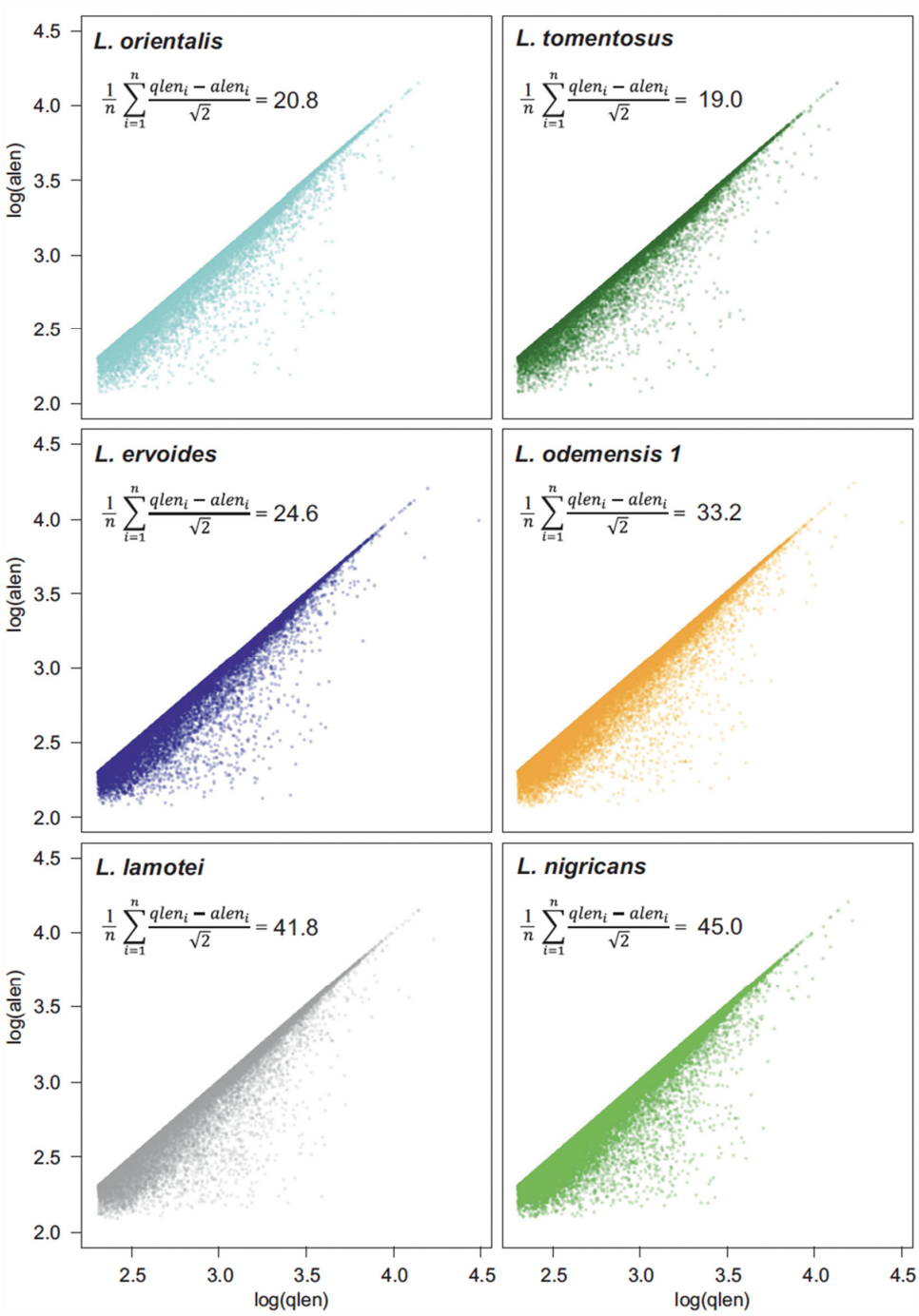

Figure 4. Accession–CDC Redberry length alignments. The query lengths (*qlen*) of each most expressed representative for the seven transcriptomes were plotted against the length of the aligned sequence in CDC Redberry (*alen*). Their average distances from a perfect alignment were calculated (formulas on the top left corner of each plot). The alignment plot and distance metrics (29.3) for *L. odemensis2* are not shown.

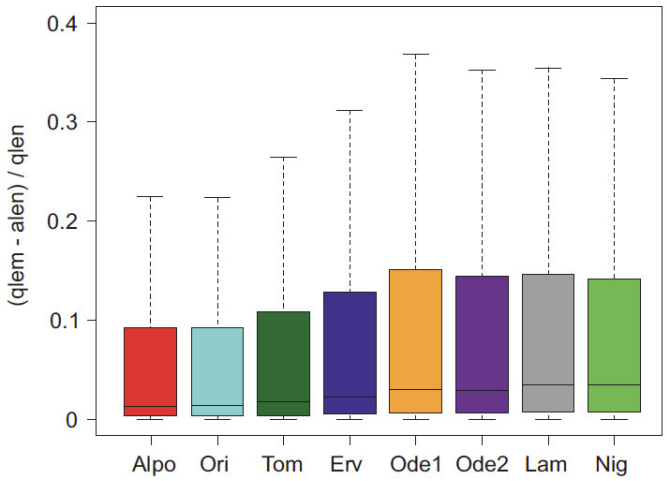

Figure 5. Boxplots displaying the variation in the relative distribution of the query alignment length over the CDC Redberry reference. Boxplots were drawn without outliers. Alpo (*culinaris*), Ori (*orientalis*), Tom (*tomentosus*), Erv (*ervoides*), Lam (*lamottei*), Nig (*nigricans*), Ode1 and Ode2 (*odemensis*).

2.3. Transcriptome Annotation and Multi-Species Pan-Transcriptome

Transcriptomes were annotated against the Universal Protein Resource (UniProt) database, a comprehensive resource for high-quality protein sequence and functional information data. Between a minimum of 60.7 (*culinaris*) and a maximum of 65.5 (*nigricans*) percent of the transcripts could be annotated using the Viridiplantae subset of proteins (Supplementary Dataset S2). The transcriptomes were also annotated against the CDC Redberry gene set, in this case using only the best representative of each gene (Supplementary Dataset S3).

We then aimed to estimate the *Lens* multi-species pangenome for the first time. Pangenome studies aim to dissect the entire set of genes from all accessions within a clade. A pangenome is usually broken down into a 'core pangenome' that contains core genes shared by all individuals, a 'shell pangenome' with genes present in several accessions, and a 'cloud pangenome' or 'accessory genome' that contains 'dispensable' genes only found in one or two accessions. Sometimes a 'soft core' is also computed to include genes present in a majority of accessions. Putative coding regions within transcript sequences (Supplementary Dataset S4) and their corresponding in-frame amino acids were searched and used to estimate the extent of the pan-transcriptome and its components. For this, transcript coding sequences of all different taxa were clustered with OthoMCL [33], which uses a Markov Cluster algorithm to group possible orthologs and paralogs. The resulting clusters were post-processed to produce average nucleotide identity matrices as well as to estimate pan-, cloud, shell, soft-core, and core transcriptomes [22]. Clustering of the CDS sequences of the seven *Lens* taxa supported a pan-transcriptome composed of 41,414 clusters with a core transcriptome of 15,910 genes encoded in all accessions and a total of 24,226 accessory genes (Figure 6A). The pan-transcriptome included clusters with inparalogues (species-specific duplications), which correspond to sequences with the best hits in their own genomes. Accessory sequences were allocated to the cloud (20,787) and shell (3439) core occupancy classes. In this context, occupancy refers to the number of taxa present in each cluster. There is also a softcore pan-transcriptome with 1278 genes that are present in all but one accession. While core genes are valuable to understand key metabolic pathways and to infer phylogenies, accessory genes may include key adaptive genes, such as those involved in biotic and abiotic stresses. To facilitate the survey of the accessory component, the protein domain frequencies within this group were calculated and used to estimate the accessory-set enrichment for protein domains (Supplementary Dataset S5).

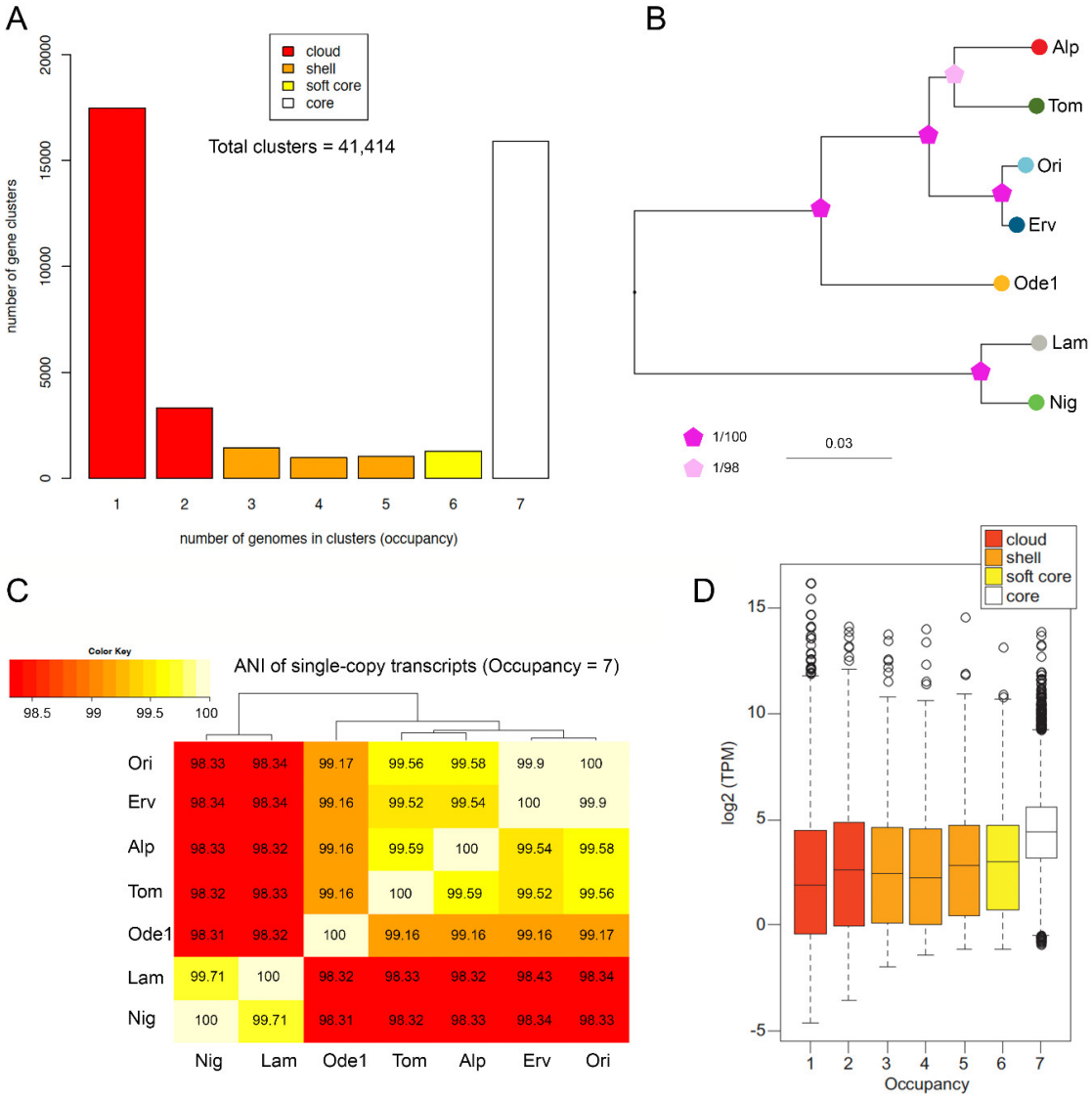

Figure 6. (**A**) Pan-transcriptome distribution of the different orthologous clusters related to the occupancy or the number of taxa present in each cluster. For instance, core loci are found in all seven accessions and hence have occupancy = 7. (**B**) Best maximum-likelihood pan-transcriptome phylogeny for the genus *Lens* using concatenated nucleotide sequences. The branch tips are colored according to the color code maintained for the accessions. The scale bar represents the number of expected substitutions per site under the best-fitting GTR + ASC + F model found. The values on the left of the legend correspond to approximate Bayes branch support values, and those to the right correspond to the UFBoot values. (**C**) Heatmap of average nucleotide identity (ANI) summarizing intra-specific genetic diversity. The dendrograms were computed by complete linkage clustering and Euclidean distances. (**D**) Mean expression of CDS in transcripts per million (TPM) as a function of their occupancy. Alp (*culinaris*), Ori (*orientalis*), Tom (*tomentosus*), Erv (*ervoides*), Lam (*lamottei*), Nig (*nigricans*), Ode1 and Ode2 (*odemensis*).

We then used the filtered non-redundant sequence clusters to simulate pan-transcriptome and core-transcriptome growth curves and thus were able to estimate the degree of completeness of the pan-transcriptome (Supplementary Figure S2). The fitted functions estimated a core set between 15,823 and 16,152 transcripts encoding proteins, and the mean pan-transcriptome size after merging all seven accessions was 29,968 ± 296 non-redundant CDSs. This estimate is smaller than the total size of 41,414 due to the removal of redundant sequences, mostly allocated in the cloud transcriptome. Since the fitting curve does not

reach a plateau, the simulations also suggest that seven accessions are not enough to sample 100% of the *Lens* species pan-transcriptome.

To further analyze pan-transcriptome sequence clusters, we aimed to estimate the best maximum-likelihood phylogeny for the genus *Lens*. For this, we focused on the core clusters since every gene in this group has a homologous version in each one of the seven accessions. The corresponding tree is plotted in Figure 6B. In addition, the clusters were further processed to compute their average percentage nucleotide identity. Average nucleotide identity (ANI) matrices summarizing the genetic distances among pairs of accessions and dendrograms are shown in Figure 6C. Both the phylogenetic and ANI trees generated similar groupings. The percentage of conserved sequence clusters among pairs of taxa was also calculated (Supplementary Table S2). The values summarize how many clusters of one accession also contain sequences from another, and ranged between 10.42% for *odemensis–culinaris*, and 12.02% for *orientalis–tomentosus*.

Lastly, it is important to evaluate up to what extent gene expression is correlated with the number of accessions present in a sequence cluster (occupancy). For this, the expression of each orthologous cluster within each occupancy class was estimated as the average transcript per million (TPM) of all coding sequences in each cluster (Figure 6D). A box plot of each cluster's average expression reveals that cloud genes are less expressed, while core genes show a higher mean expression than any other occupancy level. In addition, the range within expression data points appears to be narrower as the occupancy increases, with the narrowest range within core genes.

3. Discussion

We assembled the transcriptomes of eight lentil genotypes, including the commercial cultivar 'Alpo' (*L. culinaris*), and seven wild accessions of *L. orientalis*, *L. tomentosus*, *L. ervoides*, *L. lamottei*, *L. nigricans*, and two *L. odemensis*, with a depth of sequencing ranging from 42× and 84×. To our knowledge, this study provides the first comprehensive comparison among the main lentil taxa at the transcriptome level. The transcript assemblies developed here constitute a major advance, as they greatly expand the number of lentil ESTs. They will be a useful tool for both the study of genes and the improvement of the crop in general. In fact, a main scheme to introduce new variability into cultivars is through hybridization with wild accessions, which in turn is leveraged by an increased understanding of their transcriptome composition.

3.1. Transcriptome Quality Assessment

Commonly used metrics on transcriptome accuracy and completeness showed that all assemblies were of reasonably high quality. Sequenced reads were meaningfully represented by the assemblies, as they had the vast majority (>99.1%) of all reads mapping back to the corresponding assembly. Besides, more than 96.2% of the mapped fragment pairs were aligned as proper pairs, that is, with the correct distance and orientation. Read coverage is one of the most important parameters in transcriptome profiling, as it is a main factor determining their contiguity, completeness, correction, and the ability to assemble close paralogues and other highly similar sequences [34,35]. This is especially relevant in species with large and repetitive genomes. In fact, the two transcriptomes with the highest coverage, *culinaris* (84×) and *odemensis* (75×), showed improved metrics compared with the other six, with coverages only between 42× and 59×. For instance, *culinaris* had the highest N50 value, the highest percentage of complete BUSCO genes, and the highest number of Uniprot and CDC Redberry proteins represented by nearly full-length transcripts (>80% alignment coverage). On the contrary, transcriptomes with shallower coverage showed inferior numbers. Overall, the transcriptomes of *culinaris* and *odemensis* appeared to be the most comprehensive, as they not only contained the largest numbers of genes and transcripts, but they also had the highest percentage of recovered complete BUSCOs.

3.2. Divergency between Transcriptomes and the Lentil Reference Genome

Assemblers typically group several transcript versions together into each putative gene. Selecting which one of those versions is the best representative of the gene is a necessary step to reduce redundancy within transcriptomes and to facilitate comparisons among them. Three putative representatives of each gene cluster were preselected for an in silico evaluation: the longest transcript isoform, the most expressed, and the SuperTranscript. All these representatives were subsequently aligned to the CDC Redberry reference draft. The alignments in Figure 1 clearly point at the most expressed transcript isoform as the best representative of each putative gene, followed by the longest and the SuperTranscripts. The same conclusions were drawn with a metric that measures the distance of each query transcript from a perfect alignment (Figure 2). The observation that the most expressed is highlighted as the best sequence variant that describes a group of related alternative sequences is not surprising, and it may be due to several factors. First, the fact that an in silico transcript is highly expressed provides further evidence that it constitutes a true sequence fragment and not an assembly artifact. Second, either the longest transcript or the SuperTranscript are more prone to miss-assemblies, in particular to form chimeric fusion transcripts. SuperTranscripts are atypical transcripts because, although they provide a genome-like reference and a gene-like view of the transcriptional complexity of a gene [36], this all-exon arranged gene-like structure may not reflect a true CDS.

To compare all transcriptomes among each other, we first anchored them to a common reference, the *L. culinaris* cv CDC Redberry assembly. Anchoring markers, ESTs, or any kind of genomic sequence to a common reference has been successfully used in other crops with modest amounts of genomic data to circumvent the lack of full genome assemblies in the species/accessions to compare [37,38]. Using this strategy (Figure 3), it was determined that the *culinaris* transcripts were the most similar to the CDC Redberry, used here as the anchoring reference. Conversely, *lamottei* and *nigricans* were the least similar. *Orientalis*, *tomentosus*, *ervoides*, and *odemensis1–odemensis2* had intermediate similarities in decreasing order. The high homology of *culinaris* to the reference was much anticipated, as the accession chosen to be the reference genome for lentils is also an *L. culinaris*. In addition, the accessions of *nigricans* and *lamottei* were found to be phylogenetically the most distant to *culinaris*. A very similar arrangement was obtained when the relative distance of each of the query transcripts to a perfect alignment was considered (Figures 2A and 4). Here, the closest to a perfect alignment was that of the *culinaris* transcriptome, followed by *tomentosus*, *orientalis*, *ervoides*, *odemensis2*, *odemensis1*, *lamottei*, and *nigricans*. In this distance metric, the perfect alignment would give a summation of zero and would correspond to the alignment of CDC Redberry. Both the phylogenetic and ANI trees of Figure 6 generated slightly different groupings, with *culinaris* closer to *tomentosus* and *orientalis* closer to *ervoides*. This could be a consequence of the different sequences used. While CDSs were the input sequences for the phylogenetic and ANI trees, full transcripts were used for the rest.

Overall, the alignment metrics, together with the percentage of identity and relative alignment box plots (Figure 5), would support a classification of the seven *Lens* accessions into three closely related groups, which is compatible with three gene pools. The first would comprise *culinaris*, *orientalis*, *tomentosus*, and *ervoides*; the second would comprise *odemensis*; and the third would comprise *lamottei* and *nigricans*. A divergent position of *nigricans* from the remaining *Lens* species has been frequently described [2]. For instance, in two comprehensive studies involving the same seven accessions, *nigricans* was also identified as the most distantly related to *culinaris* [6,12]. Using thousands of high-confidence SNP markers, Dissanayake and collaborators [12] calculated allele frequencies to assess genetic differences at the species level. Among all cultivated and wild lentil genotypes, they found that the accessions of *nigricans* exhibited the greatest allelic differentiation across their genome compared to all other species/subspecies. A similar conclusion was drawn by Wong et al. [6] using GBS-derived SNP markers. They found *nigricans* to be the most distant relative, although they placed *lamottei* separate from *nigricans* and together with *odemensis*.

Our results diverge from the studies that placed *lamottei*/*odemensis* in the secondary gene pool, *ervoides* in the tertiary, and *nigricans* in the fourth [6,12]. Another focus of discrepancy is the primary gene pool. While other classifications exclude *ervoides* [6,12], our analyses better explain a scenario with four species. It is important to note that *tomentosus* was not within the species included in the study of Dissanayake et al. [12]. The fact that hybrids have been obtained from crosses of *culinaris* with *ervoides* [39] would suggest a grouping with an extended primary pool with the inclusion of *ervoides* and its repositioning closer to *culinaris*. However, the *culinaris–ervoides* hybrids needed a rescue of the embryo by in vitro culture, which according to the definition of gene pools, would place *ervoides* in a separate pool. Further studies are needed to clarify whether *ervoides* should constitute an independent gene pool.

In this regard, although *odemensis* has been considered a gene pool apart from the others, not only in our analysis but also by other classifications [6,12], this affirmation may be controversial, as crosses between *culinaris* and *odemensis* have been made [40]. Again, more research is needed to confirm or refute considering *odemensis* a separate gene pool or part of the primary gene pool.

Furthermore, the structure analysis of Dissanayake et al. [12] placed *orientalis* and *culinaris* separately, which is not consistent with the results of Wong et al. [6] and also in disagreement with our transcriptome comparison. The divergence between the studies of Wong et al. [6] and Dissanayake et al. [12] may be due to the difference in the number of accessions and/or the number of SNP markers used. It is important to remark that, while other comparatives have used SNP genomic markers, our homology analysis involved much longer transcript sequences, which allow for better precision estimates. Overall, the separation of *Lens* species remains controversial, as hybridization barriers are sometimes thin.

3.3. First Glimpse at the Lentil Pangenome

Relying on a single species' reference genome can have an adverse effect on our understanding of the genomic basis of diverse traits. Hence, the latest genomic research tends to focus on more than one species at a time. Thus, the pangenome is the union of all coding and non-coding sequences found in all individuals from a particular species. Sequences within a pangenome are typically classified as core, if present in all accessions, or accessory, if absent in some of them. The term of shell pangenome is often used to include sequences present in all but one or two accessions. Cloud genes have very low occupancy, as they are annotated in just one or two accessions. The accurate assembly of large and repetitive plant genomes is challenging and expensive. Although a comprehensive knowledge of a pangenome can only be achieved through whole genome assemblies, transcriptome sequencing can provide reasonable estimates in cases where there is no reference-quality genome available [28,29]. Still, even the most complete pan-transcriptomes, targeting several plant tissues and developmental stages, have been found to underestimate pangenome sizes by at least 10% [22].

In one of the pioneer pan-transcriptome studies, Hirsch and collaborators [29] sequenced mRNA from seedlings of 503 maize (*Zea mays*) inbred lines to characterize the maize pan-transcriptome. They identified 8681 representative transcript assemblies, with 16.4% of them expressed in all lines and 82.7% expressed in subsets of the lines. In search of the first account of the *Lens* pan-transcriptome, we sampled the transcriptomes of seven *Lens* accessions: *culinaris*, *orientalis*, *tomentosus*, *ervoides*, *lamottei*, *nigricans*, and *odemensis* and uncovered a pan-transcriptome supported by 41,414 gene clusters, including inparalogues. Of those, 15,910 were present in all seven accessions, likely a group of genes with essential roles. There was also a softcore pan-transcriptome with 1278 genes that were present in all but one accession. These could be genes that were not expressed in a particular accession at the time of sampling but also reflect a true absence in one species. The rest (24,226) were categorized as accessory genes. Care must be exerted with the cloud clusters, as they have been proven to be the most unreliable in pan-transcriptome benchmarks [41]. In fact, cloud genes, annotated in only one accession, might include significant numbers of artifacts or

pseudogenes. In contrast, shell genes are more likely to be biologically relevant. Because certain genes might not be expressed in some accessions, soft-core genes are especially relevant in transcriptome-estimated pangenomes.

Dissecting the *Lens* pangenome will have profound implications in understanding evolution, local adaptation, and population structure, with clear-cut applications in areas such as plant breeding and crop genetic studies. For instance, it has been found that core genes are under significantly stronger purifying selective pressure than accessory genes. Conversely, sequences under positive selection seem to be more frequent among accessory genes [22]. It is also well-known that accessory genes may play important roles in evolution and biotic and abiotic stresses [24,29,42]. For example, many agronomically important genes in plant species are most often found in the dispensable (accessory) genome [29]. Through the Pfam-domain enrichment test we found that the NB-ARC domain was overrepresented in the shell and cloud pan-transcriptome. The NB-ARC domain is a functional ATPase domain, and its nucleotide-binding state is proposed to regulate the activity of the disease-resistance proteins (R proteins) that are involved in pathogen recognition and the subsequent activation of plant immune responses. Domains from R genes, such as NB-ARC, have also appeared among accessory genes in cultivated barley and wheat [22,24,27] and are consistent with their duplicated nature, with the presence of frequent paralogues, and also with the expected accumulation of R genes in some wild and cultivated accessions.

Conversely, the core pan-transcriptome was enriched in pentatricopeptide repeat (PPR) family domains, which mediate several aspects of gene expression, primarily in organelles but also in the nucleus, facilitating mechanisms such as splicing, editing, stability, and the translation of RNAs. Their presence could be related to the sampling time point (seedling), a stage with a high rate of cellular division and differentiation.

We also studied the pan-transcriptome expression levels as a function of the occupancy. Interestingly, the average expression of genes in the pan-transcriptome appears to be influenced by the occupancy. In fact, core genes showed, on average, a higher expression than genes in the cloud. Other researchers have also reported the same behavior [22], which could be explained by at least two factors. First, cloud genes often accumulate non-biologically relevant sequences, such as artifacts or pseudogenes, with no expression. Second, on the other hand, the core fraction often represents genes with biologically relevant functions and thus are found to be frequently expressed in most accessions.

Overall, we assembled and annotated the transcriptomes of eight accessions from seven species of lentils, including cultivated and wild relatives. This allowed the first assessment of the *Lens* pan-transcriptome and the first comparison among all recognized *Lens* taxa in which long sequences were used. That, together with the vast amount of sequences generated, greatly increased the accuracy of the estimations over prior studies that used SNPs or other less informative markers. The findings presented here support a classification of the genus *Lens* with seven species and three gene pools. Nonetheless, the taxonomy of the genus is complex and has been subjected to various adjustments. Further wide-ranging hybridization studies are needed, mainly between wild relatives, to confirm or refute the postulates presented here.

4. Materials and Methods

4.1. Plant Material

Plants from eight different lentil genotypes were used for the study, including the Spanish cultivar 'Alpo' (*Lens culinaris*, Alp) and seven wild accessions. The wild relatives of *L. culinaris* were *L. orientalis* (Ori), *L. tomentosus* (Tom), *L. ervoides* (Erv), *L. lamottei* (Lam), *L. nigricans* (Nig), and two *L. odemensis* (Ode1 and Ode2) accessions. The bare name *odemensis* refers to Odemensis1', unless otherwise stated. Plants were grown for 16 days in capped glass tubes containing three plantlets under sterile conditions. Each tube was treated as a replicate, that is, the aerial parts of the three plants of each tube were pooled and frozen for RNA extraction. Except for Alpo and Odemensis1, three replicates of each accession

were extracted for sequencing. For Alpo and Odemensis1, seven and six replicates were sequenced, respectively, as they are the parental lines in a population of RILs developed for mapping [40]. For the same reasons, they were also sequenced at a higher depth.

4.2. RNA Extraction, Library Construction, and Sequencing

Total RNA was isolated from the pooled aerial samples using the method of Chang et al. [43], as detailed in [40]. Briefly, RNA qualities were validated using a High Sensitivity Chip on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and a Qubit 3.0 fluorometer (Thermo Fisher Scientific, Wilmington, DE, USA). Sequencing libraries were prepared following Illumina's recommendations. The quality of the libraries was analyzed by the High Sensitivity RNA assay in a 4200 TapeStation (Agilent Technologies), and the quantity of the libraries was determined by real-time PCR using a LightCycler 480 (Roche Applied Science, Penzberg, Germany). After cDNA library preparation, the samples were barcoded, multiplexed in lanes, and sequenced in an Illumina HiSeq 2500 machine (Illumina Inc., San Diego, CA, USA) using paired-end reads with 101 cycles. The output paired-end reads ranged between 50 and 134 M. However, for the reasons explained above, Alpo and odemensis1 were sequenced at a deeper coverage (see Supplementary Table S1). Raw fastq reads were quality-filtered with Trimmomatic v.0.38 [44] and quality-checked with Fastqc [45] before and after filtering.

4.3. Transcript Assemblies and Quality Parameter Estimation

Genotype-specific filtered reads were de novo assembled with Trinity v.2.10 [32] with the default parameters, except to also output SuperTranscripts. They are constructed by collapsing common and unique regions among the transcript isoforms into a single linear sequence, which ideally included all possible sequence fragments in which a gene could be processed [36]. For transcriptome assembly quality assessment, we first examined the RNA-Seq read representation of each assembly. Ideally, at least ~80% of the input RNA-Seq reads should be represented by the transcriptome assemblies. The remaining unassembled reads likely correspond to weakly expressed transcripts with insufficient coverage to enable assembly or are low-quality or aberrant reads. Reads were mapped back to the assemblies using bowtie2 v.2.3.5.1, and the statistics were visualized with the help of Samtools view v.1.10 [46].

Second, the representations of full-length or nearly full-length reconstructed protein-coding genes were examined by searching the assembled transcripts against both the Uniprot-Viridiplantae database of curated protein sequences and the CDC Redberry set of annotated coding sequences using the blastx program of BLAST+ tools v.2.11.0 [47]. It is sometimes the case that a transcript aligns to a single protein sequence with several discontinuous alignments, that is, a BLAST hit containing multiple high-scoring segment pairs (HSPs). Those multiple HSPs per transcript and database hit were first grouped and the alignment coverage was computed based on the grouped HSPs following Trinity-recommended downstream analyses and scripts. Another classical method for quality evaluation is through BUSCO [31], which determines whether assembly contigs are orthologous with a particular BUSCO dataset, providing quantitative measures of the completeness of transcriptome assemblies. BUSCO v5.0.0 with default parameters and the lineage dataset fabales_odb10 (Creation date: 5 August 2020 number of species: 10, number of BUSCOs: 5366) was used to assess the completeness of the transcriptomes.

4.4. Transcript Abundance Estimation

To compute transcript abundance, the sequenced paired-end reads were first mapped to the corresponding de novo assembled transcriptome using the bowtie aligner v.1.2.3 and the default configuration. Next, the generated bam files were used to estimate gene and transcript isoform expression levels using RSEM v.1.1.3 [48]. Three abundance estimators

were calculated and used for downstream analysis. First, the expected counts, or the sum of the posterior probability of each read coming from each transcript over all reads. Second, two relative measures of transcript abundance: TPM (transcripts per million) and FPKM (fragments per kilobase of transcript per million mapped reads). The NCBI Magic-BLAST [49] algorithm was used to align the transcript representatives against the *Lens culinaris* CDC Redberry reference draft genome assembly [10] with the default configuration, except that it was instructed not to report unaligned reads. Magic-BLAST is a tool especially designed for mapping large next-generation RNA or DNA sequencing runs against a whole genome or transcriptome. Statistical analyses were conducted in R [50]. R with built-in packages was also used to plot the results.

4.5. Transcriptome Annotation and Multi-Species Pan-Transcriptome

The blastx program of the BLAST+ tools [47] was used to search and functionally annotate the transcriptomes against the Uniprot-Viridiplantae database [30]. The -evalue parameter was set to 1×10^{-20} and -max_target_seqs was set to 1 to report only the top alignment with otherwise default parameters. TransDecoder v.3.0.1 software (https://github.com/TransDecoder, accessed on 21 June 2021) was used to identify candidate coding regions within the transcript sequences of Trinity assemblies.

The detection of orthologs and core, soft-core, and pan-genomes was computed with the GET_HOMOLOGUES-EST clustering suite of scripts [22], a branch of GET_HOMOLOGUES [51] for clustering homologous transcript sequences of related plant species. GET_HOMOLOGUES-EST was adapted to adequately handle redundant and fragmented transcript sequences from the large sizes of plant genomic data sets. Transdecoder-predicted CDS nucleotide sequences were used in homologous grouping by the calculation of overlapping sets of CDSs. The identification of orthologous groups across the transcriptomes of the different taxa was accomplished with the OthoMCL algorithm [33], which uses a Markov Cluster algorithm to group putative orthologs and paralogs (-M parameter) with no cluster size restrictions (-t 0). In addition, -c and -z options were included to request pan-, soft-core-, and core-genome analyses of the input sequences, and -A was included to compute the average% sequence identity values among pairs of genomes. Otherwise, default parameters were used. The soft-core genome tends to produce a composition analysis more robust to assembly or annotation errors than the core genome. The accompanying scripts *compare_clusters.pl* and *plot_matrix_heatmap.sh* were used to compile the corresponding pangenome matrix and draw the dendrograms, respectively. CDS clusters were also annotated for Pfam domains, and their enrichment in accessory or core clusters was computed with the *pfam_enrich.pl* script by using Fisher's exact test. Simulations of core and pan-genome sizes were conducted by sorting and adding the sequences in random order. New added sequences required an identity with less than 70% similarity to the sequences already in the pool. Data points were fitted using a Tettelin function [52]. GET_PHYLOMARKERS [53] was used to compute the pan-transcriptome maximum-likelihood (ML) tree using IQT v1.6.3 [54] with 100 independent IQT runs. It first called ModelFinder [55] using the JC2 and GTR2 base models for binary data. The best-fitting base model + ascertain bias correction + among-site rate variation parameters were selected using the Akaike information criterion (AIC). IQT was then called to perform an ML tree search under the selected model with branch support estimation. These were estimated using approximate Bayesian posterior probabilities (aBypp) as well as the ultrafast-bootstrap2 (UFBoot2) test [56].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agronomy12071619/s1, Table S1: Main assembly parameters and quality metrics; Figure S1: BUSCO summary results of the eight assembled transcriptomes. Figure S2: Pan-transcriptome simulations modelling the size versus the number of transcriptomes; Table S2: Percentage of conserved sequence clusters shared by pairs of species.

Author Contributions: Conceptualization, J.J.G.-G., P.G., C.P., A.I.G., F.V., F.J.V., M.P.d.l.V. and L.E.S.d.M.; Data curation, J.J.G.-G., P.G., C.P., A.I.G., F.V., F.J.V., M.P.d.l.V. and L.E.S.d.M.; Formal analysis, J.J.G.-G. and L.E.S.d.M.; Funding acquisition, M.P.d.l.V.; Writing – original draft, J.J.G.-G.; Writing—review & editing, P.G., F.V., M.P.d.l.V. and L.E.S.d.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Spanish Ministerio de Economía y Competitividad (grant AGL2013-44714-R), and by the Junta de Castilla y León, Spain (grant LE005G18).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Zenodo at DOI 10.5281/zenodo.6672086 (https://zenodo.org/record/6672086#.YrC6yRNBzjA (accessed on 10 April 2022)).

Acknowledgments: We would like to thank Supercomputación Castilla y León (SCAYLE) for its support and access to the high-performance computing clusters. The project was co-financed with FEDER funds.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Perez de la Vega, M.; Fratini, R.; Muehlbauer, F. Lentil. In *Genetics, Genomics and Breeding of Cool Season Grain Legumes*; Perez de la Vega, M., Torres, A.M., Cubero, J.I., Kole, C., Eds.; CRC Press: Boca Raton, FL, USA, 2011; pp. 98–150.
2. Pérez de la Vega, M.; García García, P.; Gutierrez-Gonzalez, J.J.; Sáenz de Miera, L.E. Tackling Lentil Biotic Stresses in the Genomic Era. In *Genomic Designing for Biotic Stress Resistant Pulse Crops*; Kole, C., Ed.; Springer: Cham, Switzerland, 2022. [CrossRef]
3. FAOSTAT 2019. Available online: https://www.fao.org/faostat (accessed on 15 March 2021).
4. Liber, M.; Duarte, I.; Maia, A.T.; Oliveira, H.R. The History of Lentil (*Lens culinaris* subsp. *Culinaris*) Domestication and Spread as Revealed by Genotyping-by-Sequencing of Wild and Landrace Accessions. *Front. Plant Sci.* **2021**, *12*, 628439. [CrossRef]
5. Koul, P.M.; Sharma, V.; Rana, M.; Chahota, R.K.; Kumar, S.; Sharma, T.R. Analysis of genetic structure and interrelationships in lentil species using morphological and SSR markers. *3 Biotech* **2017**, *7*, 83. [CrossRef]
6. Wong, M.M.L.; Gujaria-Verma, N.; Ramsay, L.; Yuan, H.Y.; Caron, C.; Diapari, M.; Vandenberg, A.; Bett, K.E. Classification and Characterization of Species within the Genus Lens Using Genotyping-by-Sequencing (GBS). *PloS ONE* **2015**, *10*, e0122025. [CrossRef]
7. Kumar, H.; Singh, A.; Dikshit, H.K.; Mishra, G.P.; Aski, M.; Meena, M.C.; Kumar, S. Genetic dissection of grain iron and zinc concentrations in lentil (*Lens culinaris* Medik.). *J. Genet.* **2019**, *98*, 66. [CrossRef]
8. Singh, D.; Singh, C.K.; Taunk, J.; Tomar, R.S.S.; Chaturvedi, A.K.; Gaikwad, K.; Pal, M. Transcriptome analysis of lentil (*Lens culinaris* Medikus) in response to seedling drought stress. *BMC Genom.* **2017**, *18*, 206. [CrossRef]
9. Ramsay, L.; Koh, C.; Konkin, D.; Cook, D.; Penmetsa, V.; Dongying, G.; Coyne, C.; Humann, J.; Kaur, S.; Dolezel, J.; et al. *Lens culinaris* CDC Redberry Genome Assembly v2.0. 2019. Available online: https://knowpulse.usask.ca/genome-assembly/Lcu.2RBY (accessed on 7 April 2021).
10. Ramsay, L.; Koh, C.S.; Kagale, S.; Gao, D.; Kaur, S.; Haile, T.; Gela, T.S.; Chen, L.-A.; Cao, Z.; Konkin, D.J.; et al. Genomic rearrangements have consequences for introgression breeding as revealed by genome assemblies of wild and cultivated lentil species. *bioRxiv* **2021**. [CrossRef]
11. Gorim, L.Y.; Vandenberg, A. Evaluation of Wild Lentil Species as Genetic Resources to Improve Drought Tolerance in Cultivated Lentil. *Front. Plant Sci.* **2017**, *8*, 1129. [CrossRef]
12. Dissanayake, R.; Braich, S.; Cogan, N.O.I.; Smith, K.; Kaur, S. Characterization of Genetic and Allelic Diversity Amongst Cultivated and Wild Lentil Accessions for Germplasm Enhancement. *Front. Genet.* **2020**, *11*, 546. [CrossRef]
13. Gutierrez-Gonzalez, J.J.; Garvin, D.F. Subgenome-specific assembly of vitamin E biosynthesis genes and expression patterns during seed development provide insight into the evolution of oat genome. *Plant Biotechnol. J.* **2016**, *14*, 2147–2157. [CrossRef]
14. Hu, H.; Gutierrez-Gonzalez, J.J.; Liu, X.; Yeats, T.H.; Garvin, D.F.; Hoekenga, O.A.; Sorrells, M.E.; Gore, M.A.; Jannink, J.-L. Heritable temporal gene expression patterns correlate with metabolomic seed content in developing hexaploid oat seed. *Plant Biotechnol. J.* **2020**, *18*, 1211–1222. [CrossRef]
15. Khorramdelazad, M.; Bar, I.; Whatmore, P.; Smetham, G.; Bhaskarla, V.; Yang, Y.; Bai, S.H.; Mantri, N.; Zhou, Y.; Ford, R. Transcriptome profiling of lentil (*Lens culinaris*) through the first 24 hours of Ascochyta lentis infection reveals key defence response genes. *BMC Genom.* **2018**, *19*, 108. [CrossRef]
16. Mishra, G.P.; Aski, M.S.; Bosamia, T.; Chaurasia, S.; Mishra, D.C.; Bhati, J.; Kumar, A.; Javeria, S.; Tripathi, K.; Kohli, M.; et al. Insights into the Host-Pathogen Interaction Pathways through RNA-Seq Analysis of *Lens culinaris* Medik. in Response to *Rhizoctonia bataticola* Infection. *Genes* **2022**, *13*, 90. [CrossRef]

17. Singh, D.; Singh, C.K.; Taunk, J.; Jadon, V.; Pal, M.; Gaikwad, K. Genome wide transcriptome analysis reveals vital role of heat responsive genes in regulatory mechanisms of lentil (*Lens culinaris* Medikus). *Sci. Rep.* **2019**, *9*, 12976. [CrossRef]
18. Sohrabi, S.S.; Ismaili, A.; Nazarian-Firouzabadi, F.; Fallahi, H.; Hosseini, S.Z. Identification of key genes and molecular mechanisms associated with temperature stress in lentil. *Gene* **2022**, *807*, 145952. [CrossRef]
19. Morgil, H.; Tardu, M.; Cevahir, G.; Kavakli, I.H. Comparative RNA-seq analysis of the drought-sensitive lentil (*Lens culinaris*) root and leaf under short- and long-term water deficits. *Funct. Integr. Genom.* **2019**, *19*, 715–727. [CrossRef]
20. Singh, C.K.; Singh, D.; Taunk, J.; Chaudhary, P.; Tomar, R.S.S.; Chandra, S.; Singh, D.; Pal, M.; Konjengbam, N.S.; Singh, M.P.; et al. Comparative Inter- and IntraSpecies Transcriptomics Revealed Key Differential Pathways Associated with Aluminium Stress Tolerance in Lentil. *Front. Plant Sci.* **2021**, *12*, 693630. [CrossRef]
21. García-García, P.; Vaquero, F.; Vences, F.J.; De Miera, L.E.S.; Polanco, C.; González, A.I.; Horres, R.; Krezdorn, N.; Rotter, B.; Winter, P.; et al. Transcriptome profiling of lentil in response to *Ascochyta lentis* infection. *Span. J. Agric. Res.* **2019**, *17*, e0703. [CrossRef]
22. Contreras-Moreira, B.; Cantalapiedra, C.P.; García-Pereira, M.J.; Gordon, S.P.; Vogel, J.P.; Igartua, E.; Casas, A.M.; Vinuesa, P. Analysis of Plant Pan-Genomes and Transcriptomes with GET_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species. *Front. Plant Sci.* **2017**, *8*, 184. [CrossRef]
23. Khan, A.W.; Garg, V.; Roorkiwal, M.; Golicz, A.A.; Edwards, D.; Varshney, R.K. Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends Plant Sci.* **2020**, *25*, 148–158. [CrossRef]
24. Walkowiak, S.; Gao, L.; Monat, C.; Haberer, G.; Kassa, M.T.; Brinton, J.; Ramirez-Gonzalez, R.H.; Kolodziej, M.C.; Delorean, E.; Thambugala, D.; et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **2020**, *588*, 277–283. [CrossRef]
25. Gao, L.; Gonda, I.; Sun, H.; Ma, Q.; Bao, K.; Tieman, D.M.; Burzynski-Chang, E.A.; Fish, T.L.; Stromberg, K.A.; Sacks, G.L.; et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **2019**, *51*, 1044–1051. [CrossRef]
26. Li, Y.-H.; Zhou, G.; Ma, J.; Jiang, W.; Jin, L.-G.; Zhang, Z.; Guo, Y.; Zhang, J.; Sui, Y.; Zheng, L.; et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **2014**, *32*, 1045–1052. [CrossRef]
27. Golicz, A.; Bayer, P.E.; Barker, G.C.; Edger, P.P.; Kim, H.; Martinez, P.A.; Chan, C.K.K.; Severn-Ellis, A.; McCombie, W.R.; Parkin, I.A.P.; et al. The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* **2016**, *7*, 13390. [CrossRef]
28. Gusev, A.; Ko, A.; Shi, H.; Bhatia, G.; Chung, W.; Penninx, B.W.J.H.; Jansen, R.; de Geus, E.J.C.; I Boomsma, D.; Wright, F.A.; et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **2016**, *48*, 245–252. [CrossRef]
29. Hirsch, C.N.; Foerster, J.M.; Johnson, J.M.; Sekhon, R.S.; Muttoni, G.; Vaillancourt, B.; Peñagaricano, F.; Lindquist, E.; Pedraza, M.A.; Barry, K.; et al. Insights into the Maize Pan-Genome and Pan-Transcriptome. *Plant Cell* **2014**, *26*, 121–135. [CrossRef]
30. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef]
31. Manni, M.; Berkeley, M.R.; Seppey, M.; A Simão, F.; Zdobnov, E.M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **2021**, *38*, 4647–4654. [CrossRef]
32. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.D.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef]
33. Li, L.; Stoeckert, C.J., Jr.; Roos, D.S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **2003**, *13*, 2178–2189. [CrossRef]
34. Gutierrez-Gonzalez, J.J.; Tu, Z.J.; Garvin, D.F. Analysis and annotation of the hexaploid oat seed transcriptome. *BMC Genom.* **2013**, *14*, 471. [CrossRef]
35. Kukurba, K.R.; Montgomery, S.B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, *2015*, 951–969. [CrossRef]
36. Davidson, N.M.; Hawkins, A.D.K.; Oshlack, A. SuperTranscripts: A data driven reference for analysis and visualisation of transcriptomes. *Genome Biol.* **2017**, *18*, 148. [CrossRef]
37. Gutierrez-Gonzalez, J.J.; Garvin, D.F. Reference Genome-Directed Resolution of Homologous and Homeologous Relationships within and between Different Oat Linkage Maps. *Plant Genome* **2011**, *4*, 178–190. [CrossRef]
38. Li, G.; Serba, D.D.; Saha, M.C.; Bouton, J.H.; Lanzatella, C.L.; Tobias, C.M. Genetic Linkage Mapping and Transmission Ratio Distortion in a Three-Generation Four-Founder Population of *Panicum virgatum* (L.). *G3 Genes | Genomes | Genet.* **2014**, *4*, 913–923. [CrossRef]
39. Bhadauria, V.; Ramsay, L.; Bett, K.E.; Banniza, S. QTL mapping reveals genetic determinants of fungal disease resistance in the wild lentil species *Lens ervoides*. *Sci. Rep.* **2017**, *7*, 3231. [CrossRef]
40. Polanco, C.; de Miera, L.E.S.; González, A.I.; García, P.G.; Fratini, R.; Vaquero, F.; Vences, F.J.; De La Vega, M.P. Construction of a high-density interspecific (*Lens culinaris* × *L. odemensis*) genetic map based on functional markers for mapping morphological and agronomical traits, and QTLs affecting resistance to Ascochyta in lentil. *PLoS ONE* **2019**, *14*, e0214409. [CrossRef]
41. Vinuesa, P.; Contreras-Moreira, B. Robust Identification of Orthologues and Paralogues for Microbial Pan-Genomics Using GET_HOMOLOGUES: A Case Study of pIncA/C Plasmids. In *Bacterial Pangenomics*; Mengoni, A., Galardini, M., Fondi, M., Eds.; Humana Press: New York, NY, USA, 2015; Volume 1231, pp. 203–232. [CrossRef]
42. Marroni, F.; Pinosio, S.; Morgante, M. Structural variation and genome complexity: Is dispensable really dispensable? *Curr. Opin. Plant Biol.* **2014**, *18*, 31–36. [CrossRef]

43. Chang, S.; Puryear, J.; Cairney, J. A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* **1993**, *11*, 113–116. [CrossRef]
44. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
45. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; Babraham Institute: Cambridge, UK, 2010. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed on 15 March 2021).
46. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
47. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef] [PubMed]
48. Li, B.; Dewey, C.N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **2011**, *12*, 323. [CrossRef] [PubMed]
49. Boratyn, G.M.; Thierry-Mieg, J.; Thierry-Mieg, D.; Busby, B.; Madden, T.L. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinform.* **2019**, *20*, 405. [CrossRef] [PubMed]
50. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: https://www.R-project.org/ (accessed on 15 April 2021).
51. Contreras-Moreira, B.; Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* **2013**, *79*, 7696–7701. [CrossRef]
52. Tettelin, H.; Masignani, V.; Cieslewicz, M.J.; Donati, C.; Medini, D.; Ward, N.L.; Angiuoli, S.V.; Crabtree, J.; Jones, A.L.; Durkin, A.S.; et al. Genome Analysis of Multiple Pathogenic Isolates of *Streptococcus agalactiae*: Implications for the Microbial 'pan-genome'. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13950–13955. [CrossRef]
53. Vinuesa, P.; Ochoa-Sánchez, L.E.; Contreras-Moreira, B. GET_PHYLOMARKERS, a Software Package to Select Optimal Orthologous Clusters for Phylogenomics and Inferring Pan-Genome Phylogenies, Used for a Critical Geno-Taxonomic Revision of the Genus Stenotrophomonas. *Front. Microbiol.* **2018**, *9*, 771. [CrossRef]
54. Nguyen, L.-T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef]
55. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; Von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **2017**, *14*, 587–589. [CrossRef]
56. Hoang, D.T.; Chernomor, O.; Von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **2017**, *35*, 518–522. [CrossRef]