Van Hoeck *et al. Biotechnol Biofuels (2015) 8:188*
DOI 10.1186/s13068-015-0381-1

Biotechnology for Biofuels

RESEARCH

Open Access

The first draft genome of the aquatic model plant *Lemna minor* opens the route for future stress physiology research and biotechnological applications

Arne Van Hoeck[1,2*], Nele Horemans[1,3], Pieter Monsieurs[4], Hieu Xuan Cao[5], Hildegarde Vandenhove[1] and Ronny Blust[2]

Abstract

Background: Freshwater duckweed, comprising the smallest, fastest growing and simplest macrophytes has various applications in agriculture, phytoremediation and energy production. *Lemna minor*, the so-called common duckweed, is a model system of these aquatic plants for ecotoxicological bioassays, genetic transformation tools and industrial applications. Given the ecotoxic relevance and high potential for biomass production, whole-genome information of this cosmopolitan duckweed is needed.

Results: The 472 Mbp assembly of the *L. minor* genome ($2n = 40$; estimated 481 Mbp; 98.1 %) contains 22,382 protein-coding genes and 61.5 % repetitive sequences. The repeat content explains 94.5 % of the genome size difference in comparison with the greater duckweed, *Spirodela polyrhiza* ($2n = 40$; 158 Mbp; 19,623 protein-coding genes; and 15.79 % repetitive sequences). Comparison of proteins from other monocot plants, protein ortholog identification, OrthoMCL, suggests 1356 duckweed-specific groups (3367 proteins, 15.0 % total *L. minor* proteins) and 795 *Lemna*-specific groups (2897 proteins, 12.9 % total *L. minor* proteins). Interestingly, proteins involved in biosynthetic processes in response to various stimuli and hydrolase activities are enriched in the *Lemna* proteome in comparison with the *Spirodela* proteome.

Conclusions: The genome sequence and annotation of *L. minor* protein-coding genes provide new insights in biological understanding and biomass production applications of *Lemna* species.

Keywords: *Lemna minor*, Whole-genome sequencing, Duckweed, Biomass production, Ecotoxicology, Toxicogenomics

Background

Duckweed species comprise a group of aquatic monocotyledons macrophytes consisting of floating plant bodies or "fronds." The family *Lemnaceae* consists of five genera, *Landoltia*, *Lemna*, *Spirodela*, *Wolffia* and *Wolffiella* among which 37 species have been identified so far [1–3]. Frond as well as root structures of duckweed have been morphologically simplified likely by natural selection to only those necessary to survive as floating aquatic plants. Duckweed species are of ecological significance as they are primary producers being a source of food for waterfowl, fish and small invertebrates and provide habitat for a number of small organisms. They are adapted to a wide variety of climatic regions where they, under favorable conditions, can grow extremely rapidly predominantly via asexual reproduction [4]. Such a vegetative growth results in genetically uniform clones, thereby eliminating potential effects due to genetic variability through meiosis. The natural characteristics of this plant family are mainly the basis why duckweed species are attractive

*Correspondence: avhoeck@sckcen.be
[1] Biosphere Impact Studies, SCK•CEN, Boeretang 200, 2400 Mol, Belgium
Full list of author information is available at the end of the article

for economic applications: duckweeds have been used as feed resource for fish, poultry, cattle's and other animals [5, 6] and are utilized for wastewater treatment [7, 8]. Nowadays, duckweeds are genetically modified to improve industrial applications that seem to have great potential for bioenergy production [9, 10] and pharmaceutical applications [8, 11]. Recent work on duckweed species was highlighted in a special issue in Plant Biology to commemorate Dr. Elias Landolt and his contribution to modern duckweed research [12–14].

Of all the duckweed species, Lemna species are probably the best known because of their extensive use in lab-based tests. Lemna species are smaller than *Spirodela* and *Landoltia* facilitating experimentation under limited spatial conditions but large enough to observe morphologic alterations without use of a microscope. Therefore, *L. minor* (Fig. 1a) was put forward as a model system to study fundamental plant research and has been shown to contribute to the understanding of the photoperiodic control of flowering [15, 16] and the discovery of

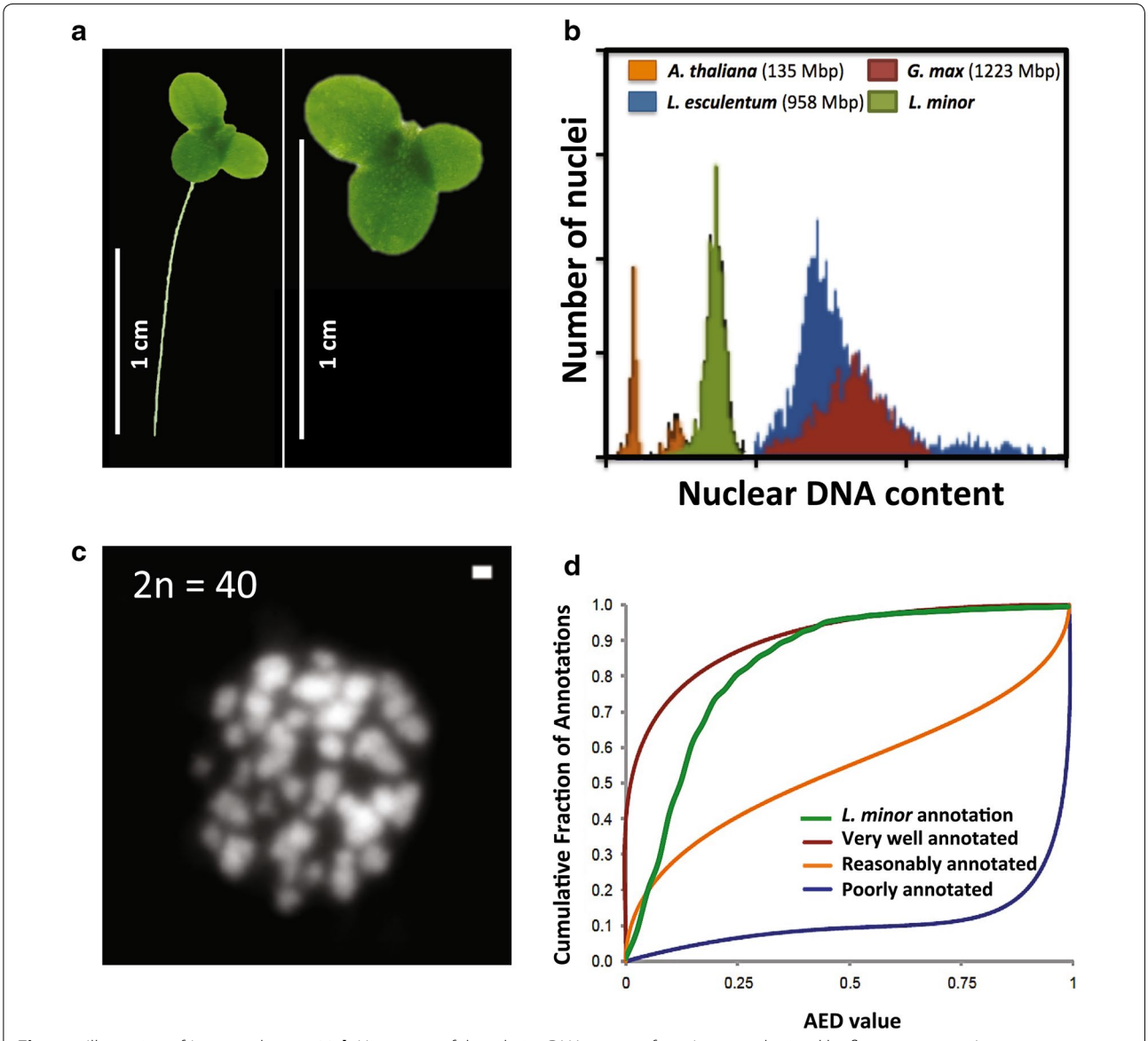

Fig. 1 **a** Illustration of *L. minor* clone 5500. **b** Histogram of the relative DNA content from *L. minor* obtained by flow cytometry. A genome content of 481 Mbp was determined for *L. minor* genome using the DNA content of different plant species (Arabidopsis, tomato and soybean) as internal standard. **c** Picture of *L. minor* chromosomes of *L. minor* ($2n = 40$). The number of chromosomes was determined by DAPI staining. *Scale bar* 1 µm. **d** Evaluation of the *L. minor* genome annotation with AED annotation scoring value's. The *L. minor* genome annotation is compared to reference curves from Maker-P annotation tool

auxin biosynthesis and sulfur assimilation pathways [17, 18]. Besides, their variety in growth habitats and their sensitivity to toxicants allowed the use of Lemna species in ecotoxicological research as representative of higher aquatic plants and standardized guidelines on how to perform a growth inhibition test were developed [19–26]. Testing procedures can follow the ASTM [24], ISO:20079 [26], OECD [19], Environment Canada [25], or EPA [23] test method guidelines. It is clear that duckweeds, and more in particular Lemna species, are gaining importance as physiological and ecotoxicological model organisms [27]. Survival, growth and reproduction at individual organisms are commonly used end points in ecotoxicity tests to support current risk assessments. However, as these end points are regarded as stringent, analyses at molecular level could provide a broader understanding in the alteration of biological pathways, including responses induced from environmental stressors [28]. Therefore, studies focusing on physiological mechanisms and genetic improvements among duckweed species await support at molecular level [3, 29].

The chloroplast genome of *L. minor* has been sequenced for phylogenetic analyses among other monocot plants [30]. Later on, Wang and Messing started with DNA barcoding duckweed strains [31] and sequenced chloroplast DNA from three other duckweed species [32], which allowed a taxonomic comparison between different duckweed ecotypes since their minute size, simple anatomy and the lack of flowering make it very harsh to analyze systematic relationships purely on morphological characteristics. To obtain nuclear DNA information for *Lemnaceae*, a whole-genome high-throughput sequencing project started in 2009 that aimed at sequencing the genome of the giant duckweed, *Spirodela polyrhiza*. Spirodela was chosen for its basal taxonomic position in the *Lemnaceae* and because of its small genome size of 158 Mbp, which is similar to that of Arabidopsis. The comprehensive genomic study of *S. polyrhiza*, published in 2014, provided insights into how this plant is adapted to rapid growth and an aquatic lifestyle [33]. Furthermore, physical mapping of the *S. polyrhiza* genome revealed that its 20 chromosomes are likely derived from seven ancestral chromosome blocks after two successive rounds of whole-genome duplication events [34].

In an effort to strengthen duckweed research, we employed the Illumina high-throughput sequencing technology to construct a draft genome of *L. minor* (Fig. 1a). Except for a small number of nuclear genes from small-scale individual studies [35], DNA information for *L. minor* is today limited to its chloroplast DNA [30]. An extensive analysis of genome sizes in different *L. minor* strains revealed that genome content and composition is highly variable within these species [36]. The estimated size of the haploid genome for a set of tested *L. minor* plants varied between 323 and 760 Mbp, thus two to three times larger than the genome size of *S. polyrhiza.* The primary objective was to characterize the protein-coding genes for *L. minor* clone 5500, a Lemna strain widely used in ecotoxicological research [37–42]. This genomic platform is expected to support the molecular basis of fundamental research in, e.g., ecotoxicogenomics and to facilitate the genetic improvement of this economically important plant, especially in biomass production.

Results and discussion

De novo assembly of *L. minor* genome with greater 100× of Illumina coverage

Genome of *L. minor* clone 5500 was estimated as 481 Mbp by flow cytometry (Fig. 1b) and is compacted in 20 chromosome pairs ($2n = 40$, Fig. 1c). In order to obtain the reference sequence of the *L. minor* genome, total genomic DNA was isolated to create two paired-end libraries for Illumina platform. A high-coverage 2 × 100 HiSeq library was supplemented with longer reads from a 2 × 300 MiSeq library. No gaps were included between both ends of the fragments resulting in paired-end reads having a nominal fragment length of 200 and 600 bp, respectively. HiSeq library consisted of 215,721,669 reads (43 Gbp) representing approximately a 90× genome coverage, while the Miseq library contained 26,270,063 (15 Gbp) reads equivalent to a genome coverage of 30×. After removing adaptors and reads containing unknown or low-quality nucleotides, the remaining 207,985,822 and 24,416,556 high-quality reads (coverage of 87× and 29× respectively) were used to assemble the *L. minor* genome (Additional file 1: Table S1). To obtain the best possible draft sequence, three different assembly programs were evaluated for the de novo assembly namely SOAPdenovo2 [43] and CLC bio, both using a *de Bruijn* graph-based algorithm and MaSuRCA [44] that uses an overlap-based assembly algorithm for the so-called super-reads. Such super-reads are uniquely extended short reads from high-coverage paired-end reads to significantly compress the data. Subsequently, the obtained assemblies were further processed with SSPACE [45] to scaffold, and Gapcloser [43] to close the gaps in a final step. With respect to number of contigs/scaffolds, corresponding N50 values and mismatch error frequency, it was found that draft genome generated by MaSuRCA generated a more robust genome sequence compared to the genomes generated by SOAPdenovo2 and CLC bio (Additional file 2: Table S2). MaSuRCA's error-correction and super-reads processes reduced the raw paired-end reads to 2,145,090 super-reads that were applied to compute pairwise overlap between these reads. From these

super-reads, the MaSuRCA pipeline generated 49,027 contigs (N50 contig size 20.9 kbp) and 46,105 scaffolds (N50 scaffold size 23.6 kbp) with a minimum length of 1000 bp (Additional file 2: Table S2). Therefore, scaffolds resulted from MaSuRCA were used for further downstream analysis.

Using the CEGMA pipeline [46], 233 protein-coding genes (94 %) of a set of highly conserved eukaryotic genes (248) were recognized within the MaSuRCA assembled genome of which 215 genes (86 %) were completely (>70 % of their length) covered (Additional file 3: Table S3). To assess the accuracy of the de novo assembly, a de novo generated set of transcripts coming from the same *L. minor* strain was aligned to the scaffolds. Using BLAT software [47], it was found that ~97 % of the cleaned transcripts aligned to at least one scaffold, with ≥95 % coverage and ≥90 % sequence identity (Additional file 4: Table S4). The final assembled sequence spanned 472,128,703 bases embedded in 46,047 scaffolds, with an N50 length of 23,801 bases when scaffolds of 1000 bp or smaller are excluded. This length is similar to the predicted genome size using Kmergenie [48] that estimated the assembly size to 475 Mbp based on k-mer statistics, or to 481 Mbp using flow cytometry (Fig. 1b). Therefore, as a proportion of the nuclear DNA content, the *L. minor* genome sequence was almost fully (98.15 %) covered by the assembled scaffolds. Scaffolds having a sequence length of 2 kbp or more covered about 96 % in size of the de novo genome assembly sequence of which 17 scaffolds had a minimum sequence length of 0.5 Mbp (Additional file 5: Figure S1). Using the available *L. minor* chloroplast DNA data, the full chloroplast genome of *L. minor* clone 5500 was obtained here by aligning NGS reads using BWA with Genbank *L. minor* chloroplast genome as reference (NC_010109.1) [49]. This chloroplast genome was 165.9 Mbp and contained 48 variants related to 117 bp (0.07 %) compared to the Genbank reference sequence which is originally from a different clone/ecotype (Additional file 6: Table S5).

In this study, a whole-genome shotgun approach was used to sequence *L. minor* genome using de novo assembly of exclusively paired-end read libraries which resulted in a moderate N50 value. The lack of mate-pair libraries makes a significant difference in the size of scaffolds and thus also to the N50 value. Libraries of paired-end reads simply cannot span many of the repetitive sequences in a genome, especially in plant genomes, which are known to have a high amount of repetitive sequences [50]. The involvement of a set of mate-pair libraries would produce longer scaffolds making N50 values 10–100 times higher [51]. Our genome assembly contains a scaffold N50 value of more than 20 kbp, which is comparable to the scaffold N50 value of the genome assemblies from *Cannabis sativa* [52] and *Phoenix dactylifera* [53]. Furthermore, the generated N50 values of other sequenced plant genome assemblies at which no mate-pair libraries are included (scaffold N50 value) are also in line to the here obtained scaffold N50 value [51]. This suggests that the produced *L. minor* assembly covers most of the non-repeated sequences. New sequencing libraries together with mapping information such as physical maps, optical maps, or cytogenetic maps [34] may be needed to improve the quality of genome sequence in order to analyze comparative genomics, whole-genome duplications, or genome evolution in duckweed species. However, current assembly enables us to characterize the basic elements (e.g., repeat and gene content) of the *L. minor* genome.

Repetitive sequences comprise 62 % of the *L. minor* genome assembly

Homology-based comparisons revealed that 62 % of the *L. minor* genome assembly consisted of repetitive sequences (Table 1). The repeats were categorized in retrotransposons (31.20 %), DNA transposons (5.08 %), tandem repeats (3.91 %) and other unclassified repeats (21.27 %). Long terminal repeat (LTR) retrotransposons are the predominant class of transposable elements (29.57 %), which is consistent with other plant genomes.

The most abundant transposon families were *gypsy* and *copia*, contributing to 10.59 and 18.79 % of the genome, respectively. For the DNA transposable elements, it was found that DNA_*hAT-Ac* elements were most abundant spanning almost 2.7 % of the nuclear genome. The high proportion of repetitive sequences could explain for the dispersed distribution of heterochromatin signatures of the *L. minor* clone 8623 (377 Mbp, [54]). Given that the plasticity of genome size in different *L. minor* clones (ranging from 323 to 760 Mbp) [36, 55] could result from different repetitive amplification and/or recent whole-genome duplications, it is interesting to study repeat content and karyotype of different *L. minor* geographical clones. In comparison with the *S. polyrhiza* genome [33] which is the most ancient duckweed, repeat amplification in *L. minor* could explain 94.5 % of the genome size difference between two duckweed reference genomes. Surprisingly, the LTR *copia* is more abundant than LTR *gypsy* in the *L. minor* genome. The *gypsy/copia* ratio in *L. minor* is 0.56, whereas the corresponding ratio in *S. polyrhiza* is 3.5 [33]. Although our repeat identification method is assembly dependent, implying the repeat content could be underestimated and high unclassified repeat proportion (34.37 % repeat content, Table 1), repeat content in *L. minor* suggests that the amplification of LTR retrotransposons played an important role in duckweed genome evolution. More detailed repeat characterization in published or ongoing duckweed

Table 1 De novo identification of sequence repeats in the genome of *L. minor*

Class	Number of elements	Elements percentage (%)	Sequence occupied (bp)	Sequence percentage of transposable elements (%)	Sequence percentage of genome (%)
Retrotransposons	223,595	34.09	152,153,999	50.76	31.20
LTR *Copia*	124,171	18.93	91,641,466	30.57	18.79
LTR *Gypsy*	81,828	12.48	51,647,357	17.23	10.59
LTR other	4532	0.69	943,224	0.31	0.19
LINE	10,193	1.55	6,598,659	2.20	1.35
SINE	2871	0.44	1,323,293	0.44	0.27
DNA transposons	54,699	8.34	24,778,060	8.27	5.08
Tandem repeats	152,112	23.19	19,062,495	6.36	3.91
Satellite	7243	1.10	2,821,147	0.94	0.58
Low complexity	18,075	2.76	1,464,636	0.49	0.30
Simple repeats	126,794	19.33	14,776,712	4.93	3.03
Unclassified	225,397	34.37	103,759,727	34.61	21.27
Total	655,803	100	299,754,281	100	61.46

genomes sequencing projects could shed more light on this interesting story.

L. minor 5500 contains a similar number of protein-coding genes as *S. polyrhiza* 7498

Scaffolds of 2 kbp or longer were selected for gene prediction, as gene predictors require a certain amount of sequence upstream and downstream of a gene to work accurately. Therefore, scaffolds smaller than 2 kbp were skipped in order to reduce the false positive errors and fragmented gene models in gene prediction. The CEGMA tool was utilized to assess the completeness on this selection of scaffold sequences. It was found that still 213 full-length genes were completely aligned meaning that the final number of the gene annotation represents at least 85 % of the true number of genes (Additional file 3: Table S3). Gene models from masked *L. minor* genome sequences were predicted and annotated with the ab initio and homology-based gene prediction pipeline MAKER-P [56] (Additional file 7: Table S6). To obtain a comprehensive set of *L. minor* gene models, RNA was isolated and sequenced from *L. minor* plants cultivated under healthy growth conditions and from *L. minor* plants exposed to various stress conditions (including uranium, gamma radiation and Sr-90 treatment). Using the Illumina HiSeq platform, approximately, 592,326,402 clean sequencing reads were obtained after adapter and low-quality reads trimming (Additional file 8: Table S7). 530,159 transcripts were produced with Trinity de novo assembler, including different isoforms per transcript [57]. These transcriptomic data of *L. minor*, together with all available transcripts from duckweed species *Landoltia punctata*, *Lemna gibba* and *S. polyrhiza* and supplemented with nine proteomes from monocotyledon plants, served as evidence for the gene prediction tools SNAP [58] and Augustus [59] inside Maker-P pipeline. In total, 22,382 protein-coding genes were annotated whereof 18,744 genes (84 %) contained an AED (Annotation Edit Distance) score under 0.25 which can be regarded as highly accurate (Fig. 1d). Although the number of genes is lower than the number found in other sequenced monocot plants, it was very similar to that of the closely related *S. polyrhiza*. This supports the hypothesis that the small and structurally simple anatomy of duckweed species allowed to loose a number of genes. On average, gene models consisted of 1934 bp and means of 4.8 exons per gene (Table 2; Additional file 9: Figure S2). The exon length distribution was consistent with other species, although *L. minor* intron length tended to be shorter than that of other species used in the comparison (Table 2). To assess the accuracy of the obtained annotation, the complete set of the *L. minor* proteins from the National Center of Biotechnology Information (NCBI) was blasted to the *L. minor* proteins. It turned out that 60 of the 61 NCBI accessions (downloaded 11-09-2015) could be aligned to at least one of the *L. minor* proteins (BLASTP [60], e-value of 1e−10) (Additional file 10: Table S8).

Since the *L. minor* genome has been sequenced using a WGS approach without the use of mate-pair libraries or the construction of a physical map, it is not excluded that some alleles may have been annotated as individual genes. Heterozygosity is namely more prevalent in asexual individuals compared to sexual species through mutation accumulation in

Table 2 Overview of gene features from *L. minor* and three other monocotyledonous plants

Species	*L. minor*	*S. polyrhiza*	*O. sativa*	*Z. mays*
Genome size (Mbp)	481	158	430	2.067
No. Of genes	22,382	19,623	39,045	63,480
Mean gene length (bp)	2738	3458	2853	4653
Median gene length (bp)	1934	2245	1654	2397
Mean CDS length	1332	1108	1064	1206
Median CDS length	1146	903	849	996
Mean exon length	208	213	259	333
Median exon length	138	121	139	178
Mean exons per mRNA	4.8	5.2	4.9	5.5
Median exons per mRNA	3	4	3	4
Mean intron length	209	560	418	878
Median intron length	103	178	170	144

clonal lineages [61]. A study of Cole and Voskuil revealed that this was also true for a population of *L. minor* [55]. However, when using the MaSuRCA pipeline instead of *de Bruijn* graph-based assembly approach, it overcomes the repeat sequences, errors, low-coverage regions and small structural differences caused by heterozygosity because of its overlap-layout-consensus approach [62]. To assess the accuracy of the de novo annotation, we examined the proportion of de novo created transcripts represented in the annotated transcriptome. A total of 179,736 different RNA transcripts were made by Transdecoder of which 179,734 could be mapped to the annotated transcripts (BLASTN [60], e-value of 1e−30).

Lemna proteome is mostly (66.2 %) shared with the Spirodela proteome

To study gene content of *L. minor* and duckweed in general, we examined the sequence similarities between *L. minor* and *S. polyrhiza* genes and two other highly annotated monocot plants. Therefore, the 22,382 gene products of *L. minor* were clustered into orthologous and paralogous groups with 107,716 gene products from *S. polyrhiza*, *Oryza sativa* and *Zea mays* using OrthoMCL [63]. Although the three sets of gene annotation contain different numbers of gene models reflecting the different annotation history, this comparison provided an indication of the overall completeness of our assembly. In summary, 8202 orthologous groups were conserved in all four species containing 39 % of the submitted genes (Fig. 2a). In addition to 3546 *L. minor* singleton genes (not grouped by OrthoMCL, 15.8 % of total *L. minor* genes), a total of 795 paralogous groups representing 2897 genes (12.9 %) were unique to *L. minor* (Additional file 11: Table S9). These 6443 genes from two groups are further referred to as Lemna-specific genes in this study. The more closely related species would be expected to have a higher numbers of similar gene models. As a result, 14,830 *L. minor* genes (66.2 %) have orthologs in *S. polyrhiza,* whereas other 1109 *L. minor* genes (4.9 %) have orthologs in either *O. sativa*, *Z. mays,* or both but not *S. polyrhiza* (Fig. 2b). Furthermore, it was found that 1821 genes (8.13 %) of *L. minor* shared a unique similarity with at least one gene from *S. polyrhiza*, which are further referred to as duckweed-specific genes.

It has been shown in the *S. polyrhiza* genome that there have been two ancient rounds of whole-genome duplications during evolution (ca. 90 Mya) [33]. In the comparison of gene families between *S. polyrhiza* and four representative plant species (Arabidopsis, tomato, banana and rice), a low gene copy number in *S. polyrhiza* indicated preferred gene losses of duplicated genes [33]. It would be interesting to study the gene number and relation of gene families of other Lemna genomes which are in progress, such as *L. gibba* G3 DWC131 (450 Mbp) and *Lemna minor* clone 8627 (800 Mbp) [64, 65]. It is conceivable that the ancestor genome of Lemna species contained at least one recent whole-genome duplication after the split between *L. minor* and *S. polyrhiza* genera followed by different degree of gene removal processes of duplicated genes resulting to different Lemna species with the genome size ranging from 323 to 760 Mbp [36, 55]. The most extensive gene loss can result to a reduced total gene numbers such as the case of *L. minor* 5500. An alternative hypothesis, on the other hand, could be that *L. minor* 5500 represents the Lemna ancestor genome which contains the similar gene content as the Spirodela genome. Other larger genome Lemna species could have been evolved from larger repeat expansion or very recent and independent whole-genome duplications. This hypothesis could be tested by future work, which studies macro-synteny relationship between *S. polyrhiza* 7498 genome ($2n = 40$, 158 Mbp) and *L. minor* 5500 genome ($2n = 40$, 481 Mbp).

Gene annotation information supports further genome functional analysis and biomass production applications

To identify the putative functions of the *L. minor* gene models, a sequence similarity search was carried out against the Swiss-Prot protein sequences of *Arabidopsis thaliana* and *O. sativa* (BLASTP [60], e-value of 1e−5). Subsequently, the transcripts were annotated with Gene Ontology (GO) and Pfam terms using a local installation of Interproscan 5 [66] and KEGG pathway mapping using the KEGG Automatic Annotation Server (KAAS) [67]. The pfam-A database provides profile hidden Markov models of over 13,672 conserved protein families [68]. The GO project provides an ontology of defined

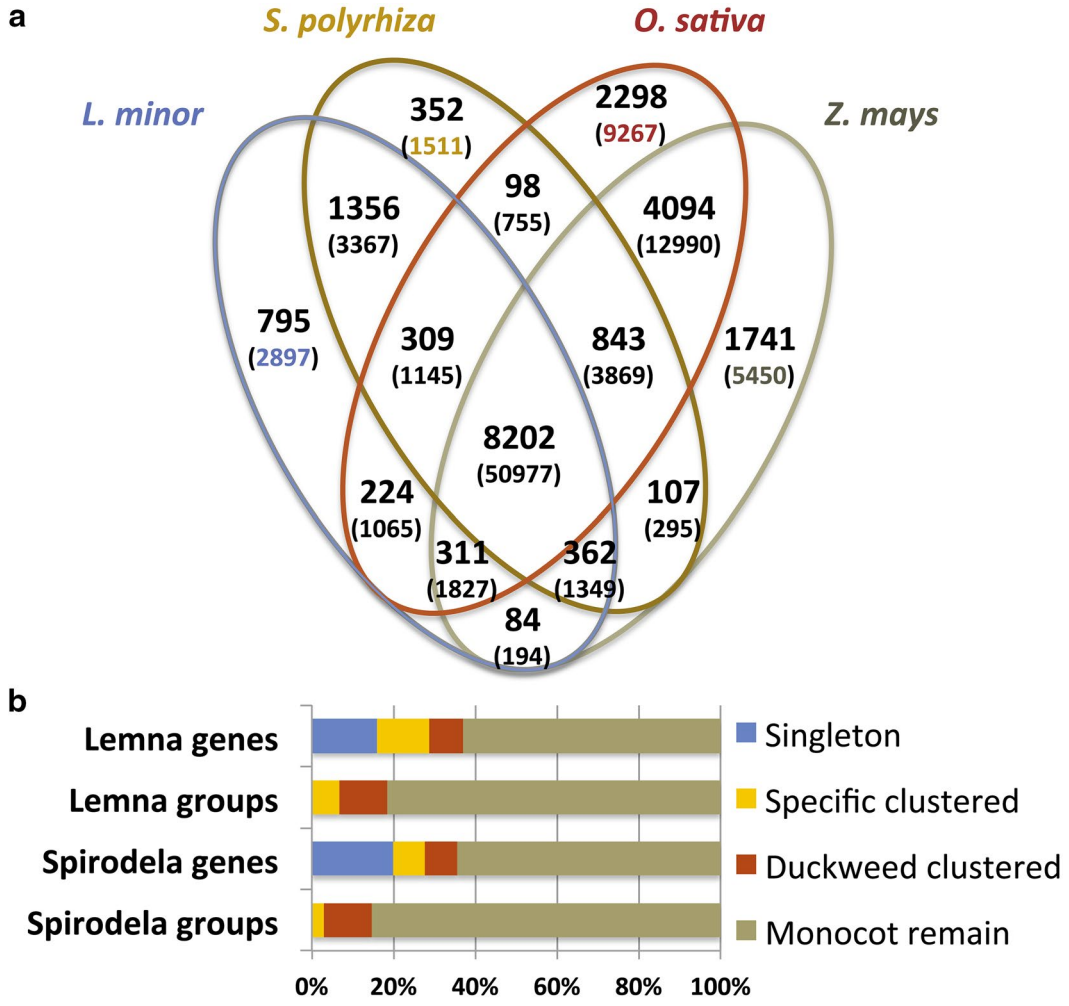

Fig. 2 A venn diagram showing clusters of orthologous and paralogous gene families in *L. minor*, *S. polyrhiza*, *Z. mays* and *O. sativa* as identified by OrthoMCL. Gene family number is listed in each of the components; the number of genes within the families for all of the species within the component is noted within parentheses

terms representing gene product properties, which covers three domains: cellular component, molecular function and biological process. The result of KAAS contains KO (KEGG Orthology) assignments and automatically generated KEGG pathways. In total, 21,263 gene models (95 %) received an annotation link with at least one of the included databases of which 18,597 (83.1 %) were assigned to one or more Pfam domains, 7329 (32.7 %) to KEGG ontology term and 15,512 (69.3 %) of the proteins were successfully annotated with Gene Ontology terms. The GO terms of *L. minor* present overall similarity to the GO annotations of *S. polyrhiza*, *O. sativa* and *Z. mays* (Fig. 3, Additional file 12: Figure S3; Additional file 13: Table S10). The GO enrichment analysis between the two duckweed species reveals that the *L. minor* proteome contains 24 overrepresented and 15 underrepresented GO terms with significant FDR <0.05 (Fig. 3; Additional file 14: Table S11). Enriched proteins in *L. minor* 5500 included (1) enzymes involved in catabolic processes (GO:9056, 422 proteins), hydrolase activity (GO:16787, 2739 proteins); (2) proteins in response to various stimulus (e.g., stress (GO:6950, 529 proteins), abiotic stimulus (GO:9628, 86 proteins), extracellular stimulus (GO:9991, 19 proteins), endogenous stimulus (GO:9719, 55 proteins); and (3) biosynthesis processes (e.g., precursor metabolites and energy (GO:6091, 258 proteins), DNA metabolic process (GO:6259, 350 proteins), carbohydrate metabolic process (GO:5975, 776 proteins). These proteins could contribute to *L. minor* ability for (1) the removal of surplus nutrients from waste water, (2) adaptation to various climate conditions resulting in their world-wide distribution, and (3) providing nutritional value and high biomass productivity.

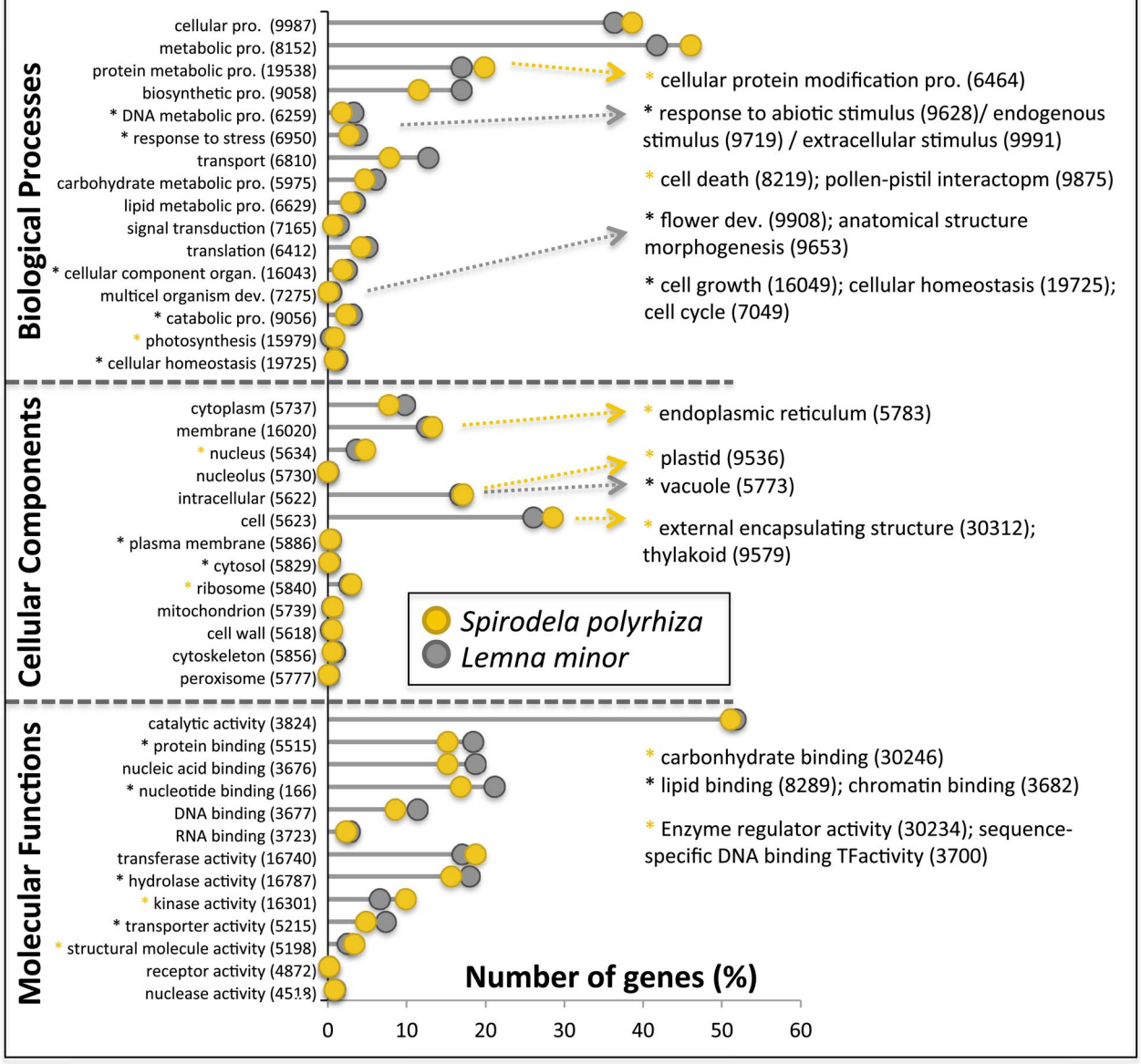

Fig. 3 Comparison of the most relevant plant GO slim terms for three structured ontologies between *L. minor* (*black*) and *S. polyrhiza* (*yellow*). More specific GO terms over over/under represented in *L. minor* are shown on the right side. *Asterisk symbols* indicate that these GO terms are significantly enriched (Fisher exact test, FDR <0.05) in *L. minor* (*black*) or *S. polyrhiza* (*yellow*) (Fisher exact test, FDR <0.05). *pro* process, *organ.* organization, *dev.* development, *TF* transcriptional factor

Interestingly, 2381 *L. minor* specific genes (36.9 %) and 326 *L. minor* tandem duplicated genes (17.4 %) are present in the overrepresented GO terms. Furthermore, *L. minor* contains sequences coding for 12 glutamine synthetases (GS) and 21 glutamate synthases (GOGAT) in comparison with 7 and 11 sequences in *S. polyrhiza*, respectively (Additional files 15, 16: Fig. S4, S5; Additional file 17: Table S12). Both enzymes regulate ammonium assimilation which is an important biochemical pathway for the use of *L. minor* in wastewater remediation, possibly in combination with energy production [69]. Therefore, these amplified genes, which may diverge to produce novel functions via neofunctionalization, could be potential candidates for further functional studies, since efficient transformation protocols for *L. minor* are available [70, 71].

Conclusions

In the present study, the *L. minor* genome has been sequenced using exclusively paired-end sequencing reads. Given an estimated genome size of 481 Mbp, the draft genome represented 98 % of the *L. minor* genome and contained protein-coding genes proportional to *S. polyrhiza*. Functional characterization and comparison supported the accuracy of the obtained predicted protein-coding genes. Therefore, the present *L. minor* genomic resource is highly beneficial for understanding the biological and molecular mechanisms in *L. minor* and will facilitate future genetically improvements and biomass production applications of duckweed species.

Methods

Plant material

Lemna minor cv. Blarney plants (Serial number 1007, ID number 5500) were collected from a pond in Blarney, Co. Cork, Ireland, (University College Cork, Ireland) and are a gift from Prof. Dr. M. Jansen. The plants were aseptically cultured in a growth chamber in 250-mL glass Erlenmeyer flasks containing half-strength Hütner medium [27] under continuous light (Osram 400 W HQI-BT daylight, OSRAM GmbH, Augsburg, Germany, 102 ± 1 µmol m^{-2} s^{-1}) at 24.0 ± 0.5 °C. Plants were subcultured every 10–12 days by transferring three plants to 100 mL of fresh growth medium.

DNA extraction and genome assembly

Genomic DNA was extracted from fresh *L. minor* plants with the DNAeasy Plant kit mini prep (Qiagen, Venlo, Limburg, Netherlands) according to the manufacturer's recommendations. Aliquots of the extraced DNA were used for measuring DNA quantity with NanoDrop ND1000 and for genome sequencing.

We sequenced the *L. minor* genome by exclusively using Illumina platforms which were performed at VIB Nucleomics Core, Leuven, Belgium. Two sequencing runs were performed: A 2X100 paired-end library was used to generate a high-coverage library using the HiSeq 2000 platform, while a 2X300 paired-end library was used for the MiSeq platform (Additional file 1: Table S1). The produced reads were cleaned using the FastX tool [72] and cutadapt [73]. High-quality paired-end sequence reads were used as input for three different assembly programs: SOAPdenvo2 [43], CLC Bio (genomic workbench 7-Qiagen, Aarhus, Denmark), and MaSuRCA pipeline [44]. For SOAPdenovo2 and CLC Bio, overlapping paired-end reads were first merged before assembly using Flash (1.2.8). A k-mer length of 63 was selected for both SOAPdenovo2 and CLC Bio. Using the software package MaSuRCA pipeline, a K-mer length of 81 has been selected by jellyfish, included in the MaSuRCa pipeline, which was close to the k-mer length of 87 suggested by Kmergenie. Since this pipeline generates super-reads, raw sequencing reads were used as input data. The scaffolds/contigs resulted from three different assemblers were further processed with SSPACE [45] to scaffold the contigs, and Gapcloser [43] to close the gaps in a final step. Afterwards, The 46,047 scaffolds that could be mapped with >95 % of their sequence length to the *L. minor* chloroplast genome were excluded from the scaffold pool resulting in 45,990 scaffolds. The chloroplast genome was assembled using BWA with genbank *L. minor* chloroplast sequence as reference (NC_010109.1). GATK tools were used for variant calling.

Accession numbers and availability of material

All the raw sequence data of the *L. minor* genome and transcriptome have been deposited at the NCBI Sequence Read Archive (SRA) repository under the SRA accession number SRP065561. The *L. minor* genome with annotation is available at CoGe database with the genome identifier ID 27408 (https://genomevolution.org/r/ik6h) or upon request (avhoeck@gmail.com; nhoreman@sckcen.be).

Annotation of repeat sequences

Putative transposable elements of the *L. minor* genome assembly were identified using REPEATMASKER version 3.3.0 (http://www.repeat-masker.org) using Dfam library (Dfam 1.3), RepBase library (13 July 2014) and a de novo *L. minor* library. RepeatModeler was used to build a *L. minor* de novo repeat library.

RNA extraction and de novo assembly

RNA was extracted from plants exposed to different concentrations of uranium-238, gamma radiation, and strontium-90. For the gamma radiation treatment, plants were exposed to dose rates of 15, 53, 120, 232 mGy h^{-1} using a cesium-137 gamma source (1.25E + 12 Bq) in a modified OECD medium [42]. For the beta-radiation treatment [41], plants were cultured in a modified growth medium containing ^{90}Sr activity concentrations of 25, 250, 2500, and 25,000 kBq L^{-1} added as $SrCl_2$ (3.7 MBq stock solution, IDB Belgium), while for uranium treatment [37], plants were exposed to a growth medium containing uranium concentrations of 0.5, 4, 6.5, and 10 µM added as $UO_2(NO_3)_2$ H_2O. For each treatment, control plants were grown simultaneously under non-exposure conditions. After 7 days of exposure, plants from each exposure condition (3 replicas) were snap frozen in liquid nitrogen and stored at −80 °C until total RNA extraction. Total RNA was extracted from *L. minor* plants with RNeasy Plant Mini Kit (Qiagen, Venlo, Limburg, Netherlands) according to the manufacturer's recommendations. RNA quality and quantity control was performed using NanoDrop ND1000 and BioAnalyzer (Agilent Technologies).

Transcriptomes of full *L. minor* plants were sequenced on the Illumina HiSeq 2000 platform at the University of Antwerp using the Truseq™ RNA sample prep kit (version 2—single read—50 bp) according to the manufacturer's recommendation. The reads were cleaned using FASTX-Toolkit [72] and cutadapt [73]. The reads of all exposure concentration per stress condition (five different concentrations) and of control plants were pooled and served as input for the de novo transcriptome assembly program Trinity [57]. TransDecoder [50] was used to identify likely coding sequences within Trinity generated transcript sequences after BLASTP (threshold 1e−5) selection with Swiss-Prot database [74]. The obtained Transcdecoder_transcripts.fasta file was used within the gene prediction tool to serve as transcript evidence.

Gene predictions

Gene prediction for *L. minor* was conducted by various methods available in MAKER-P version 2.31.8 [56]. The MAKER-P annotation pipeline consists of different steps to generate high-quality annotations by taking transcriptomic as well as proteomic evidence. Only scaffolds with a minimum sequence size of 2000 bp were considered for gene prediction. Hence, the 32,348 scaffolds were masked with RepeatMasker using the same libraries as described above. MAKER-P was run on *L. minor* using assembled transcriptomic data (Transcdecoder_transcripts.fasta), and proteins from monocotyl plants (*Musa acuminata*, *Musa balbisiana*, *Sorghum bicolor*, *O. sativa* subsp. *japonica*, *Brachypodium distachyon*, *Elaeis guineensis*, *S. polyrhiza* and *P. dactylifera* (downloaded february 17, 2014 from Swiss-Prot database). The gene prediction tools SNAP [58] and Augustus (3.0.1) [59] were trained to generate species-specific gene models. The first run for initial prediction was made by de novo assembled Transcdecoder transcripts data. The output results were used to retrain SNAP inside MAKER-P pipeline using both assembled transcdecoder transcripts and protein evidence. The resulting second gene set was used for the final training of SNAP. Augustus was trained using the output of CEGMA with WebAugustus. An ad hoc filtering procedure was applied to the genes consisting of one exon as single-exon alignments often result from spurious alignments, library contamination, background transcription of the genome, pseudogenes, and repeat elements [56]. Therefore, the proteins of the single-exon gene models that did not supported on transcript level were aligned to the proteins from the monocot plants included in the gene prediction tool. Single-exon genes were retained when its protein length is 90 % or more compared to at least one protein from the monocot plants (BLASTP, e-value of 1e−5).

Functional annotation and GO enrichment analysis

Putative gene function was assigned to the *L. minor* genes based on the best alignment to the protein sequences of Swiss-prot using BLASTP [35] with a threshold of 1e−5. Gene ontology terms and Pfam domains were assigned to the genes by software Interproscan 5 [66]. GO slim identification and representation of the GO terms were performed using BINGO [75]. Two-sided Fisher's exact test with false discovery rate (FDR) threshold of <0.05 was used for enrichment analysis in Blast2GO v3.1.3 [76]. The enriched GO terms were further filtered to reduce to the most specific GO terms.

Identification of orthologous groups and tandem duplicated genes

Proteins of *L. minor*, *S. polyrhiza*, *O. sativa*, and *Z. mays* were selected to perform an all-against-all comparison using BLASTP [35]. The results were fed into the standalone OhrthoMCL [63] program using a default MCL inflation parameter of 1.5. Tandem duplicated genes were detected in 10 kb or longer scaffolds of *L. minor* by using SynMap tool of CoGe website [77]. Cladogram analyses were performed using default parameters using advanced workflow of phylogeny.fr [78].

Flow cytometry and chromosome counting

Flow cytometry was used to experimentally estimate DNA genome size of *L. minor*, by comparing it to the DNA content of *Lycopersicon esculentum*, *Glycine max,* and *Arabidopsis thaliana*. The isolation of nuclei was performed on fresh plant material (fronds) immediately after harvesting using the Cystain PI Absolute P kit (Partec). *L. minor* fronds were chopped with a fresh razor blade in a petri dish containing 250 µl ice-cold extraction buffer. After 1 min incubation, the solution was filtered through a 50-µm nylon filter (Celltrics), and 1 mL staining solution, consisting of 1 mL staining buffer, 120 µL propidium iodide (PI) solution, and 6 µL RNase per sample, was added to the flow-through. The samples were then incubated in the dark for at least 1 h, and their nuclear DNA content was analyzed on the BD Accuri C6 Flow Cytometer (BD Biosciences) with a FL2 585/40 nm filter. Reference DNA standards suitable for plant genome size estimation were received from the lab of J. Dolezel. At least, 500 nuclei counts per plant were analyzed.

Metaphase chromosome preparation of *L. minor* clone 5500 was performed as described in [34] with a small modification. Root tips which were macerated in a drop of 75 % acetic acid were treated further with nine acetic acid:one methanol solution for 3 min before squashing. Chromosome counting was evaluated by DAPI staining (2 µg mL^{-1} in VectaShield) by using a 100× objective of a Zeiss Axioplan 2 epifluorescence microscope equipped with a cooled CCD camera (Diagnostic Instruments, Inc.).

Additional files

Additional file 1: Table S1. Illumina libraries statistics for genome assembly.

Additional file 2: Table S2. Summary statistics of the *L. minor* genome assemblies.

Additional file 3: Table S3. Number of CEGMA hits for different genome assemblies.

Additional file 4:Table S4. Statistics of transcriptome and whole genome assembly of *L. minor*..

Additional file 5: Figure S1. Assembled scaffold length distribution of *L. minor* genome.

Additional file 6: Table S5. Variant calling of *L. minor* (strain 5500) chloroplast genome vs. *L. minor* GenBank reference genome NC_01019.

Additional file 7: Table S6. Summary of the masked and unmasked genome assembly.

Additional file 8: Table S7. Illumina libraries statistics for transcriptome assembly.

Additional file 9: Figure S2. Distribution of the number of exons per gene.

Additional file 10: Table S8. BLAST results of *L. minor* GenBank genes vs *L. minor* 5500.

Additional file 11: Table S9. Overview of OrthoMCL gene clusters for *L. minor* with *S. polyrhiza*, *Z. mays* and *O. sativa*.

Additional file 12: Figure S3. Distribution of top plant GO slim categories for *L. minor*, *S. polyrhiza*, *Z. mays*, and *O. sative* proteomes. GO slim categories were assigned to *L. minor* by InterProscan5. GO slim categories of *S. polyrhiza* were extracted from Phytozome10 and *O. sativa* and *Z. mays* from Plaza 3.0.

Additional file 13: Table S10. Overview of gene ontology classification.

Additional file 14: Table S11. Overview of the over and underrepresented GO terms for *L. minor* genes compared to S. polyrhiza genes and for L. minor and duckweed specific genes in *L. minor* genome. The number of tandem genes for each GO term are also included.

Additional file 15: Figure S4. Cladogram of glutamate synthase isoforms in *L. minor* (Lminor), *S. polyrhiza* (Spipo), *A. thaliana* (AT), O. indica (OSINDICA) and *B. distachyon* (BD). The phylogentic relationships of glutamate synthase from different species have been calculated under default parameters using phylogeny.fr.

Additional file 16: Figure S5. Cladogram of glutamine synthetase isoform in *L. minor* (Lminor), *S. polyrhiza* (Spipo), *A. thaliana* (AT), O. indica (OSINDICA) and *B. distachyon* (BD). The phylogentic relationships of glutamine synthetase from different species have been calculated under default parameters using phylogeny.fr.

Additional file 17: Table S12. Blast hit results of *L. minor* glutamine synthetase and glutamate synthase genes.

Abbreviations

ASTM: American Society for Testing and Materials; OECD: Organisation for Economic Co-operation and Development; EPA: Environmental Protection Agency; LTR: long terminal repeats; AED: annotation edit distance; GO: gene ontology; KEGG: kyoto encyclopedia of genes and genomes.

Authors' contributions

AVH carried out the sequencing, assembly, and annotation of the genome and wrote the manuscript. NH participated in the design of the study and critically reviewed the manuscript. PM performed statistical analyses and wrote the scripts necessary for high performance computing. HXC carried out the chromosomal determination, participated in the GO enrichment studies, and critically reviewed the manuscript. HV and RB supervised the research. All authors read and approved the final manuscript.

Author details

[1] Biosphere Impact Studies, SCK•CEN, Boeretang 200, 2400 Mol, Belgium. [2] Department of Biology, University of Antwerp, Groenenborgerlaan 171, 2020 Antwerp, Belgium. [3] Centre for Environmental Research, University of Hasselt, Universiteitslaan 1, 3590 Diepenbeek, Belgium. [4] Microbiology, SCK•CEN, Boeretang 200, 2400 Mol, Belgium. [5] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), OT Gatersleben, Corrensstrasse 3, 06466 Stadt Seeland, Germany.

Acknowledgements

The authors thank the Research foundation-Flanders (FWO) (G.A040.11N) and the European Commission Contract Fission-2010-3.5.1-269672 Strategy for Allied Radioecology (http://www.star-radioecology.org) for financial support of this work. Belgian nuclear research institute (SCK•CEN) is further thanked for funding the PhD of AVH. HXC is supported by the German Research Foundation (SCH 951/18-1). The people from CALCUA at the University of Antwerp are acknowledged for assisting high performance computing (http://www.uantwerpen.be/calcua). The authors also thank L. Leus, ILVO, for estimating DNA genome size through flow cytometry.

Competing interests

The authors declare that they have no competing interests.

Received: 29 September 2015 Accepted: 10 November 2015
Published online: 25 November 2015

References

1. Bog M, Baumbach H, Schween U, Hellwig F, Landolt E, Appenroth KJ. Genetic structure of the genus Lemna L. (Lemnaceae) as revealed by amplified fragment length polymorphism. Planta. 2010;232(3):609–19. doi:10.1007/s00425-010-1201-2.
2. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT. Phylogeny and systematics of Lemnaceae, the duckweed family. Syst Bot. 2002;27(2):221–40. doi:10.1043/0363-6445-27.2.221.
3. Appenroth KJ, Borisjuk N, Lam E. Telling duckweed apart: genotyping technologies for the Lemnaceae. Chin J Appl Environ Biol. 2013;19(1):1–10.
4. Lemon GD, Posluszny U, Husband BC. Potential and realized rates of vegetative reproduction in *Spirodela polyrhiza*, *Lemna minor*, and *Wolffia borealis*. Aquat Bot. 2001;70(1):79–87. doi:10.1016/S0304-3770(00)00131-5.
5. Leng RA, Stambolie JH, Bell R. Duckweed—a potential high-protein feed resource for domestic animals and fish FAO Livestock Research for Rural Development. 1995;7(1). http://www.fao.org/ag/aga/agap/frg/lrrd/lrrd7/1/3.htm
6. Rusoff LL, Blakeney EW, Culley DD. Duckweeds (Lemnaceae): a potential source of protein and amino acids. J Agric Food Chem. 1980;28:848–50.
7. Bergmann BA, Cheng J, Classen J, Stomp AM. In vitro selection of duckweed geographical isolates for potential use in swine lagoon effluent renovation. Bioresour Technol. 2000;73(1):13–20. doi:10.1016/S0960-8524(99)00137-6.
8. Zhao Y, Fang Y, Jin Y, Huang J, Bao S, Fu T, et al. Pilot-scale comparison of four duckweed strains from different genera for potential application in nutrient recovery from wastewater and valuable biomass production. Plant Biol (Stuttgart, Germany). 2015;17(Suppl 1):82–90. doi:10.1111/plb.12204.
9. Cui W, Cheng JJ. Growing duckweed for biofuel production: a review. Plant Biol (Stuttgart, Germany). 2015;17(Suppl 1):16–23. doi:10.1111/plb.12216.
10. Wang W, Yang C, Tang X, Gu X, Zhu Q, Pan K, et al. Effects of high ammonium level on biomass accumulation of common duckweed *Lemna minor* L. Environ Sci Pollut Res Int. 2014;21(24):14202–10. doi:10.1007/s11356-014-3353-2.

11. Yamamoto Y, Rajbhandari N, Lin X, Bergmann B, Nishimura Y, Stomp A-M. Genetic transformation of duckweed *Lemna gibba* and *Lemna minor*. Vitro Cell Dev Biol Plant. 2001;37(3):349–53. doi:10.1007/s11627-001-0062-6.
12. Landolt E. The family of Lemnaceae—a monographic study. Vol. 1, biosystematic investigations in the family of duckweeds (Lemnaceae). Veröffentlichungen des Geobotanischen Institutes der ETH, Stiftung Rübel, Zürich, 1986.
13. Landolt E, Kandeler R. The family of Lemnaceae—a monographic study. Vol. 4, Biosystematic investigations in the family of duckweeds (Lemnaceae). Veröffentlichungen des Geobotanischen Institutes der ETH, Stiftung Rübel, Zürich, 1987.
14. Appenroth KJ, Crawford DJ, Les DH. After the genome sequencing of duckweed—how to proceed with research on the fastest growing angiosperm? Plant Biol (Stuttgart, Germany). 2015;17(Suppl 1):1–4. doi:10.1111/plb.12248.
15. Maeng J, Khudairi AK. Studies on the flowering mechanism in Lemna. Physiol Plant. 1973;28(2):264–70. doi:10.1111/j.1399-3054.1973.tb01187.x.
16. Moon HK, Stomp AM. Effects of medium components and light on callus induction, growth, and frond regeneration in *Lemna gibba* (Duckweed). Vitro Cell Dev Biol Plant. 1997;33(1):20–5. doi:10.1007/s11627-997-0035-5.
17. Bruce BD, Malkin R. Biosynthesis of the chloroplast cytochrome b6f complex: studies in a photosynthetic mutant of Lemna. Plant Cell. 1991;3(2):203–12. doi:10.1105/tpc.3.2.203.
18. Cedergreen N, Madsen TV. Nitrogen uptake by the floating macrophyte *Lemna minor*. New Phytol. 2002;155(2):285–92. doi:10.1046/j.1469-8137.2002.00463.x.
19. OECD. *Lemna* sp. growth inhibition tests. Guideline 221. Paris: Organisation for Economic Co-operation and Development; 2006.
20. Moody M, Miller J. *Lemna minor* growth inhibition test. In: Blaise C, Férard J-F, editors. Small-scale freshwater toxicity investigations. Dordrecht: Springer; 2005. p. 271–98.
21. Radic S, Stipanicev D, Cvjetko P, Rajcic MM, Sirac S, Pevalek-Kozlina B, et al. Duckweed *Lemna minor* as a tool for testing toxicity and genotoxicity of surface waters. Ecotoxicol Environ Saf. 2011;74(2):182–7. doi:10.1016/j.ecoenv.2010.06.011.
22. Naumann B, Eberius M, Appenroth KJ. Growth rate based dose-response relationships and EC-values of ten heavy metals using the duckweed growth inhibition test (ISO 20079) with *Lemna minor* L. clone St. J Plant Physiol. 2007;164(12):1656–64. doi:10.1016/j.jplph.2006.10.011.
23. Agency EP. Ecological effects test guidelines: OPPTS 850.4400 aquatic plant toxicity test using Lemna spp., Tiers I and II. Washington, DC: USEPA. United States Environmental Protection Agency, Prevention, Pesticides and Toxic Substances (7101) EPA712-C-96-156. 1996.
24. E. A. 1415-91 standard guide for conducting static toxicity tests with *Lemna gibba* G3. West Conshohocken, PA: American Society for Testing and Materials; 1999.
25. Environment-Canada. Biological test method: test for measuring the inhibition of growth using the freshwater macrophyte *Lemna minor*. Method development and application section. Ottawa, ON: Environmental Technology Centre, Environment Canada, Report EPS. 1999.
26. ISO 20079. Water quality—determination of the toxic effect of water constituents and waste water to duckweed (*Lemna minor*)—duckweed growth inhibition test. Geneva: International Standard ISO 20079: 2004; 2004.
27. Brain RA, Solomon KR. A protocol for conducting 7-day daily renewal tests with *Lemna gibba*. Nat Protoc. 2007;2(4):979–87.
28. Versteeg DJ, Naciff JM. In response: ecotoxicogenomics addressing future needs: an industry perspective. Environ Toxicol Chem/SETAC. 2015;34(4):704–6. doi:10.1002/etc.2843.
29. Lam E, Appenroth K, Michael T, Mori K, Fakhoorian T. Duckweed in bloom: the 2nd international conference on duckweed research and applications heralds the return of a plant model for plant biology. Plant Mol Biol. 2014;84(6):737–42. doi:10.1007/s11103-013-0162-9.
30. Mardanov AV, Ravin NV, Kuznetsov BB, Samigullin TH, Antonov AS, Kolganova TV, et al. Complete sequence of the duckweed (*Lemna minor*) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. J Mol Evol. 2008;66(6):555–64. doi:10.1007/s00239-008-9091-7.
31. Wang WQ, Wu YR, Yan YH, Ermakova M, Kerstetter R, Messing J. DNA barcoding of the Lemnaceae, a family of aquatic monocots. BMC Plant Biol. 2010;10:205. doi:10.1186/1471-2229-10-205.
32. Wang WQ, Messing J. High-throughput sequencing of three Lemnoideae (duckweeds) chloroplast genomes from total DNA. PLoS One. 2011;6(9):e24670. doi:10.1371/journal.pone.0024670.
33. Wang W, Haberer G, Gundlach H, Gläßer C, Nussbaumer T, Luo MC et al. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nat Commun. 2014;5. doi:10.1038/ncomms4311.
34. Cao HX, Vu GTH, Wang W, Appenroth KJ, Messing J, Schubert I. The map-based genome sequence of *Spirodela polyrhiza* aligned with its chromosomes, a reference for karyotype evolution. New Phytol. 2015. doi:10.1111/nph.13592.
35. Akhtar TA, Lampi MA, Greenberg BM. Identification of six differentially expressed genes in response to copper exposure in the aquatic plant *Lemna gibba* (duckweed). Environ Toxicol Chem. 2005;24(7):1705–15. doi:10.1897/04-352r.1.
36. Wang WQ, Kerstetter R, Michael TP. Evolution of genome size in duckweeds (Lemnaceae). J Bot. 2011;2011:9.
37. Horemans N, Van Hees M, Van Hoeck A, Saenen E, De Meutter T, Nauts R, et al. Uranium and cadmium provoke different oxidative stress responses in *Lemna minor* L. Plant Biol. 2015. doi:10.1111/plb.12222.
38. Lahive E, O'Halloran J, Jansen MA. A marriage of convenience; a simple food chain comprised of *Lemna minor* (L.) and *Gammarus pulex* (L.) to study the dietary transfer of zinc. Plant Biol (Stuttgart, Germany). 2015;17(Suppl 1):75–81. doi:10.1111/plb.12179.
39. Lahive E, O'Halloran J, Jansen MAK. Frond development gradients are a determinant of the impact of zinc on photosynthesis in three species of Lemnaceae. Aquat Bot. 2012;101:55–63. doi:10.1016/j.aquabot.2012.04.003.
40. Juhel G, Batisse E, Hugues Q, Daly D, van Pelt FN, O'Halloran J, et al. Alumina nanoparticles enhance growth of *Lemna minor*. Aquat Toxicol (Amsterdam, Netherlands). 2011;105(3–4):328–36. doi:10.1016/j.aquatox.2011.06.019.
41. Van Hoeck A, Horemans N, Van Hees M, Nauts R, Knapen D, Vandenhove H, et al. Beta-radiation stress responses on growth and antioxidative defense system in plants: a study with strontium-90 in *Lemna minor*. Int J Mol Sci. 2015;16(7):15309–27. doi:10.3390/ijms160715309.
42. Van Hoeck A, Horemans N, Van Hees M, Nauts R, Knapen D, Vandenhove H, et al. Characterizing dose response relationships: chronic gamma radiation in *Lemna minor* induces oxidative stress and altered polyploidy level. J Environ Radioact. 2015;150:195–202. doi:10.1016/j.jenvrad.2015.08.017.
43. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1(1):18. doi:10.1186/2047-217x-1-18.
44. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669–77. doi:10.1093/bioinformatics/btt476.
45. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578–9. doi:10.1093/bioinformatics/btq683.
46. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061–7. doi:10.1093/bioinformatics/btm071.
47. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64. doi:10.1101/gr.229202.
48. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;30(1):31–7. doi:10.1093/bioinformatics/btt310.
49. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60. doi:10.1093/bioinformatics/btp324.
50. Hamilton JP, Buell CR. Advances in plant genome sequencing. Plant J. 2012;70(1):177–90. doi:10.1111/j.1365-313X.2012.04894.x.
51. Michael TP, Jackson S. The first 50 plant genomes Plant Genome 2013;6(2). doi:10.3835/plantgenome2013.03.0001in.
52. van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, et al. The draft genome and transcriptome of *Cannabis sativa*. Genome Biol. 2011;12(10):R102. doi:10.1186/gb-2011-12-10-r102.
53. Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, et al. Genome sequence of the date palm *Phoenix dactylifera* L. Nat Commun. 2013;4. doi:10.1038/ncomms3274.

54. Cao HX, Vu GTH, Wang W, Messing J, Schubert I. Chromatin organisation in duckweed interphase nuclei in relation to the nuclear DNA content. Plant Biol (Stuttg). 2015;17:120–4. doi:10.1111/plb.12194.
55. Cole CT, Voskuil MI. Population genetic structure in duckweed (*Lemna minor*, Lemnaceae). Can J Bot Rev Can Bot. 1996;74(2):222–30.
56. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 2014;164(2):513–24. doi:10.1104/pp.113.230144.
57. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. 2013;8(8). doi:10.1038/nprot.2013.084.
58. Korf I. Gene finding in novel genomes. BMC Bioinform. 2004;5(1):59.
59. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19(Suppl 2):215–25. doi:10.1093/bioinformatics/btg1080.
60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. doi:10.1016/s0022-2836(05)80360-2.
61. Birky CW Jr. Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. Genetics. 1996;144(1):427–37.
62. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24(8):1384–95. doi:10.1101/gr.170720.113.
63. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89. doi:10.1101/gr.1224503.
64. Lam E, Appenroth KJ, Michael T, Mori K, Fakhoorian T. Duckweed in bloom: the 2nd international conference on duckweed research and applications heralds the return of a plant model for plant biology. Plant Mol Biol. 2014;84(6):737–42. doi:10.1007/s11103-013-0162-9.
65. Applications ISCoDRa. ISCDRA special issue #10 post kyoto 2015 ICDRA. http://www.lemnapedia.org. 2015;3:139–68.
66. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40. doi:10.1093/bioinformatics/btu031.
67. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35(Web Server issue):W182–5. doi:10.1093/nar/gkm321.
68. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res. 2010;38(Database issue):D211–22. doi:10.1093/nar/gkp985.
69. Landolt E, Kandeler R. Biosystematic investigations in the family of duckweeds (Lemnaceae), Vol. 4: the family of Lemnaceae-a monographic study, Vol. 2 (phytochemistry, physiology, application, bibliography). Veroeffentlichungen des Geobotanischen Instituts der ETH, Stiftung Ruebel (Switzerland). 1987.
70. Cantó-Pastor A, Mollá-Morales A, Ernst E, Dahl W, Zhai J, Yan Y, et al. Efficient transformation and artificial miRNA gene silencing in *Lemna minor*. Plant Biol (Stuttg). 2015;17:59–65. doi:10.1111/plb.12215.
71. Yamamoto Y, Rajbhandari N, Lin X, Bergmann B, Nishimura Y, Stomp A-M. Genetic transformation of duckweed *Lemna gibba* and *Lemna minor*. Vitro Cell Dev Biol Plant. 2001;37(3):349–53. doi:10.1007/s11627-001-0062-6.
72. Gordon A. FASTX-toolkit. http://hannonlab.cshl.edu/fastx_toolkit/index.html. Accessed 18 Nov 2010
73. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2. doi:10.14806/ej.17.1.200.
74. Consortium TU. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(D1):D204–12. doi:10.1093/nar/gku989.
75. Maere S, Heymans K, Kuiper M. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005;21(16):3448–9. doi:10.1093/bioinformatics/bti551.
76. Conesa A, Gotz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008;2008:619832. doi:10.1155/2008/619832.
77. Lyons E, Pedersen B, Kane J, Freeling M. The value of nonmodel genomes and an example using SynMap Within CoGe to dissect the hexaploidy that predates the rosids. Trop Plant Biol. 2008;1(3–4):181–90. doi:10.1007/s12042-008-9017-y.
78. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008;36(Web Server issue):W465–9. doi:10.1093/nar/gkn180.