A draft genome of *Alliaria petiolata* (garlic mustard) as a model system for invasion genetics

Nikolay Alabi[1], Yihan Wu[1], Oliver Bossdorf[2], Loren H. Rieseberg[3] and Robert I. Colautti[1*]

[1]Biology Department, Queen's University, Kingston, Ontario, Canada

[2]Institute of Ecology and Evolution, University of Tübingen, Tübingen, Germany

[3]Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada

Corresponding author: robert.colautti@queensu.ca

Keywords: *Alliaria petiolata*, EFCC3, invasion genetics, garlic mustard, Illumina, mate pairs

Abstract

The emerging field of invasion genetics examines the genetic causes and consequences of biological invasions, but few study systems are available that integrate deep ecological knowledge with genomic tools. Here we report on the *de novo* assembly and annotation of a genome for the biennial herb *Alliaria petiolata* (M. Bieb.) Cavara & Grande (Brassicaceae), which is widespread in Eurasia and invasive across much of temperate North America. Our goal was to sequence and annotate a genome to complement resources available from hundreds of published ecological studies, a global field survey, and hundreds of genetic lines maintained in Germany and Canada. We sequenced a genotype (EFCC-3-20) collected from the native range near Venice, Italy and sequenced paired-end and mate pair libraries at ~70 × coverage. A *de novo* assembly resulted in a highly continuous draft genome (N_{50} = 121 Mb; L_{50} = 2) with 99.7 % of the 1.1 Gb genome mapping to scaffolds of at least 50 Kb in length. A total of 64,770 predicted genes in the annotated genome include 99 % of plant BUSCO genes and 98 % of transcriptome reads. Consistent with previous reports of (auto)hexaploidy in western Europe, we found that almost one third of BUSCO genes (390/1440) mapped to two or more scaffolds despite < 2 % genome-wide average heterozygosity. The continuity and gene space quality of our draft assembly will enable molecular and functional genomic studies of *A*. *petiolata* to address questions relevant to invasion genetics and conservation strategies.

Introduction

Biological invasions are a threat to global biodiversity with significant impacts to human health and welfare (Mack et al., 2000). They also present opportunities for large-scale 'natural' experiments to study ecological and evolutionary processes in the wild (Mooney and Cleland, 2001; Sax et al., 2005). Despite a large body of ecological research and a growing number of evolutionary studies of invasive species, functional genetic and 'omics approaches have been rare among studies of invasive species until recently (reviewed in (Barrett, 2015; Bock et al., 2015); but see e.g. (Barrett et al., 2016; Boheemen et al., 2017; Bourne et al., 2020)). Apart from studies using neutral markers to assess population structure (reviewed in (Dlugosch and Parker, 2008)), relatively little is known about the genetic causes and consequences of biotic invasions at the molecular level. Perhaps this is partly due to a lack of genomic resources, which has hindered high-resolution genetic studies of most invasive species. Here, we report on a draft genome for the herbaceous biennial plant *Alliaria petiolata*, a plant invader in North America with potential to become a model system for the emerging field of invasion genetics.

Several factors favour *A*. *petiolata* as an emerging model system for invasion genetics (Colautti et al., 2014). First, it is widely distributed with variable ecological impacts throughout its range (Lankau et al., 2009; USDA, 2020). Second, it has a relatively simple two-year life cycle with well-defined life stages (seed, rosette, bolting, senescent). This simple life history facilitates measurements of lifetime fitness, natural selection, and their combined influence on population dynamics in natural populations. Third, *A. petiolata* is a member of the Brassicaceae and therefore benefits from genetic resources available for well-studied species like *Brassica rapa* (canola) and the model plant *Arabidopsis thaliana*, providing opportunities for functional and comparative genomics. Fourth, high selfing rates produce naturally inbred seed families that can be maintained through single-seed descent. Fifth, *A*. *petiolata* has been the focal species in hundreds of field surveys and experimental studies, including influential studies testing the role of natural enemies (Lewis et al., 2006), competition (Prati and Bossdorf, 2004), the 'novel weapons' hypothesis (Callaway et al., 2008), competitive ability (Bossdorf et al., 2004), glucosinolate

metabolism (Haribal et al., 2001), and eco-evolutionary dynamics (Lankau et al., 2009). One mechanism that is especially well-studied is the production of secondary metabolites and their effects on soil microbiota, particularly the suppression of mycorrhizal fungi that form beneficial symbiotic networks among native plant roots (Anthony et al., 2017; Duchesneau et al., 2021). Understanding the genomic basis of such interactions with soil microbes will not only advance basic science but it also has potential applications in plant restoration and agriculture.

Recent efforts to develop *A. petiolata* as a model system include the Global Garlic Mustard Field Survey (GGMFS), which mobilized 164 participants from 16 countries to collect field data and seed samples across Europe and North America, resulting in thousands of seed families from 383 distinct populations (Colautti et al., 2014). A subset of inbred lines collected across North America have been maintained through single-line descent in labs in Germany (Bossdorf Lab, University of Tübingen) and Canada (Colautti Lab, Queen's University). Adding to these resources, we here report on a draft genome of a single *A. petiolata* genotype from Europe, annotated with RNA sequencing of leaf and root tissue.

Methods & Materials

Study Organism and Line Derivation

Alliaria petiolata (M. Bieb.) Cavara & Grande is a biennial herbaceous plant in the Thlaspideae tribe of the Brassicaceae family. It was introduced to North America prior to 1868, when it was discovered in Long Island, New York (Nuzzo, 1993). By 1948 it was reported on the West Coast of North America and has established in at least 37 U.S. States and five Canadian provinces from the Atlantic coast to the Pacific (Cavers et al., 1979; USDA, 2020). As the only species of *Alliaria* with a broad distribution, *A. petiolata* is relatively easy to identify in natural habitats owing to its white flowers and dentated peltate leaves with long petioles. It is considered a noxious weed across most of its introduced range, due in part to impacts on native plants and tree regeneration in deciduous forest ecosystems (Cipollini and Cipollini, 2016; Stinson et al., 2007). Two genome types have been identified, including diploids ($2n$ = 14) in Eastern Europe and Western Asia and hexaploids ($2n$ = 42 and $2n$ = 36) in Central Europe and North America (Esmailbegi et al., 2018; Weiss-Schneeweiss and Schneeweiss, 2003). The haploid genome size was previously estimated to be 1.35 Gb based on flow cytometry (Barow and Meister, 2002).

S_0 generation

All source material for genome and transcriptome sequencing originated from a single individual grown from seed collected as part of the GGMFS (Colautti et al., 2014). The inbred line used in this study is from population EFCC3, collected in 2011 from a small forest fragment (~2,000 m^2) surrounded by agricultural land, about 75 km northwest of Venice, Italy (UTM 45.71 °N × 11.72 °E). The specific seed line was sampled at 3.2 m along a 10 m sampling transect originating at the edge of population EFCC3. This inbred line (code EFCC3-3-20) is currently maintained along with other GGMFS seed collections by two of the coauthors in replicate collections in Tübingen, Germany (Bossdorf) and Kingston, Ontario, Canada (Colautti).

S_1 generation

In July 2012, ten seeds of the EFCC3-3-20 genotype from the original S_0 field collection were removed from cold storage (4 °C), surface washed with a mild detergent and rinsed with distilled H_2O before surface sterilizing in 10 % bleach for 10 minutes. Sterilized seeds were again rinsed with distilled H_2O

101 before placing on filter paper saturated with distilled H_2O and sealed in a petri dish with paraffin wax.
102 We stratified seeds in the dark at 10 °C for ~ 90 days and thereafter inspected weekly until emerging
103 radicles were observed. Germinating seeds were transplanted into 4" plastic pots containing a peat soil
104 mixed with vermiculite that was watered to saturation and placed under shade cloth in the Horticulture
105 Greenhouses at the University of British Columbia. We let seedlings establish in soil, watered as needed,
106 for four weeks before a small amount (~ 5 mm x 2 mm) of young leaf meristem tissue was harvested
107 from a single individual and immediately preserved in liquid nitrogen. Roughly 25 to 50 mm^3 of this
108 tissue was divided into two separate 2 mL screw-cap tubes, each containing two stainless steel ball
109 bearings of 2mm diameter. These samples were used for genomic DNA purification and genome
110 sequencing.

111 A second individual from the same inbred family was transplanted to an 8" plastic pot and fertilized with
112 20/20/20 N/P/K fertilizer to encourage rosette growth before being moved outside for cold
113 vernalization from October 2012 to April 2013 at the University of British Columbia Horticulture
114 Greenhouses. In April 2013, we moved the plant back into the greenhouse and sprayed with 2 %
115 insecticidal soap to remove pests. Once inside the greenhouse, the plant was left to mature and set seed
116 autonomously via self-pollination. Mature siliques were harvested in July 2013 and seeds were stored in
117 paper envelopes at 4 °C.

118 S_2 generation

119 In May 2016, we removed a subset of 10 seeds of the S_1 generation from cold storage and germinated in
120 a 60 mm × 15 mm petri dish containing filter paper covered with a mixture of autoclaved ProMix soil and
121 silica sand (1:9 ratio). We added distilled water until saturation and thereafter petri dishes were sealed
122 with paraffin wax before storing in the dark at 4 °C. Of these, six seedlings germinated and were
123 retained for transcriptome sequencing, divided into one of two treatments. The first true leaf from each
124 of the three plants in the experimental treatment were cut with scissors. We used a Kimwipe tissue
125 saturated with either 0.4 mM jasmonic acid (JA) dissolved in 10 % ethanol (treatment) or 10 % ethanol
126 alone (control), adhered directly to maintain contact the cut site (treatment) or uncut leaf (control). We
127 replaced the saturated Kimwipe every 8 h to maintain the signal. After 48 h of treatment, we harvested
128 seedlings and preserved treated leaves and root tissue in liquid nitrogen, to be used for RNA purification
129 and transcriptome sequencing.

130 DNA Isolation & Library Construction

131 In September 2013, frozen tissue from the S_1 genotype was pulverized and extracted using a
132 Cetyltrimethyl Ammonium Bromide (CTAB) protocol (Clarke, 2009) with the following modifications.
133 After pulverising tissue in a bead mill homogenizer at 60 Hz for 60 s, we added 1 mL chilled wash buffer
134 (200 mM Tris-HCl pH 8.0, 50 mM EDTA, 250 mM NaCl) and incubated on ice for 10 min. The purpose of
135 this wash step is to remove secondary metabolites after disruption of cell walls but prior to cell lysis with
136 CTAB. Following this initial wash step, we spun tubes in a microcentrifuge at 4000 g and 4° C for 10
137 minutes, then discarded the supernatant and added another 1 mL of wash buffer. This was repeated
138 once more for a total of three wash cycles until no coloration was visible in the supernatant. After final
139 discard of the supernatant and addition of warm lysis buffer as per the CTAB protocol, we vortexed
140 tubes briefly to resuspend plant cells. After completion of the CTAB protocol, pellets were dissolved in
141 50 µL of reverse osmosis (RO) H_2O and sent to Centre d'expertise et de services Génome Québec
142 (Génome Québec) for library preparation and sequencing.

143 We used four separate sequencing libraries for genome assembly: (i) One whole-genome shotgun
144 sequencing library using the Illumina TruSeq DNA v1 preparation kit with a target fragment length of 150
145 bp. (ii & iii) Two Illumina Nextera MatePair libraries with target lengths of 5 Kb and 10 Kb. These three
146 libraries were multiplexed and sequenced on a single flowcell of Illumina HiSeq 2000 using 2 × 100 bp
147 paired end (PE) sequencing chemistry. (iv) Target fragment lengths of 450 bp using the Illumina TruSeq
148 DNA v1 and sequenced on Illumina MiSeq with 2 × 250 bp paired-end reads.

149 RNA Isolation & Library Construction

150 For RNA purification, we pulverized frozen leaf and root tissue In March 2017, in the same manner as
151 the DNA extraction protocol outlined in the previous section. After pulverizing the tissue, we extracted
152 whole RNA from each plant separately using Invitrogen's TRIzol reagent, following the manufacturer's
153 protocol (Pub No. MAN0001271 Rev. A.0). We sequenced four of the six extractions with the highest
154 RNA yields at Génome Québec using the KAPA rRNA-depletion Kit (plant) followed by the Illumina
155 TruSeq LT multiplex kit for sequencing on a single lane of Illumina HiSeq 2500 with 2 × 125 bp paired-
156 end reads. Eight separate libraries were sequenced based on tissue and treatment as described
157 above (see S2 Generation): Control Leaf (CL), Control Root (CR), Treated Leaf (TL), and Treated Root (TR),
158 with two replicates each.

159 Data Processing Methods

160 Raw sequencing data was processed and demultiplexed by Génome Québec, and copied to the
161 Frontenac cluster hosted by the Centre for Advanced Computing (CAC) at Queen's University and the
162 Cedar cluster maintained by Simon Fraser University on behalf of Compute Canada. The CAC maintains
163 the Rosalind Franklin Cluster for Analysis of Complex Genomes, which is a 256-core computing cluster
164 with 2 TB of RAM. We used this hardware for the memory-intensive steps of *de novo* genome assembly,
165 with the remaining analyses completed using shared Frontenac and Cedar clusters, and on personal
166 computers. All FASTQ files from both experiments passed quality controls using **fastqc** (version 0.11.5)
167 (Andrews, 2010). We used the raw, demultiplexed FASTQ files for *de novo* assembly, but the
168 transcriptome data were pre-processing using **cutadapt** (Martin, 2011) to trim adapters and remove
169 reads shorter than 25b prior to assembly.

170 Our genome assembly pipeline involved two main steps. First, we used **ALLPATHS-LG** (Gnerre et al.,
171 2011) version R52488 to assemble contigs from both the HiSeq and MiSeq paired-end libraries and then
172 to link contigs into scaffolds using the 5 Kb and 10 Kb MatePair libraries. We filtered reads that aligned
173 to a draft cpDNA genome but otherwise did not pre-process reads as the **ALLPATHS-LG** pipeline includes
174 filtering steps. Mitochondrial DNA sequences were manually removed from the final assembly. The
175 analysis parameters included PLOIDY = 2 and HAPLOIDIFY = TRUE to perform a diploid genome
176 assembly. Although our genome is likely hexaploid, polyploid models are not supported by **ALLPATHS-LG**
177 and low heterozygosity is expected given the high selfing rates in natural populations. Second, we joined
178 scaffolds from **ALLPATHS-LG** into larger mega-scaffolds using **redundans** (Pryszcz and Gabaldón, 2016)
179 version 0.13c, with the following parameters: *identity* = 0.9, *iters* = 5, *joins* = 5, *limit* = 1, *linkratio* = 0.7,
180 *mapq* = 10, *minlength* = 1000 and *overlap* = 0.75. Thus, overlapping scaffolds were collapsed only if they
181 overlapped by at least 75% with an average identity of at least 90%. We repeated this script four times
182 with output scaffolds of the prior run acting as input scaffolds given the long run-time required (~28d).
183 This second round of assembly combines scaffolds with overlapping similarity, resulting in mega-
184 scaffolds that can span across multiple chromosomes, and other regions of high recombination.

185 Sequences from the transcriptome experiment were cleaned and assembled with **trinity** (Grabherr et
186 al., 2011) following protocols outlined on the software documentation and in Haas *et al* (Haas et al.,
187 2013). We used default parameters and the quality of the assembly was analyzed using the custom perl
188 scripts included in the **trinity** package to examine full length transcripts and scaffolds (i.e. Contig Nx
189 lengths in *analyze_blastPlus_topHit_coverage.pl, TrinityStats.pl*). Additionally, we mapped read pairs to
190 the transcriptome assembly to assess read content using **bowtie2** (version 2.3.3.1) with default
191 parameters (Langmead and Salzberg, 2012). We used **TransDecoder** to predict open reading frames in
192 transcripts before using **Trinotate** to annotate and analyze assembled transcripts (Haas et al., 2013).

193 As a first step in annotation, we established a detailed repeat library. Miniature Inverted Transposable
194 Elements (MITES) were identified using **MITE Tracker** (Crescente et al., 2018). Long Terminal Repeat
195 (LTR) elements were identified using the **GenomeTools** package (Gremme et al., 2013). To reduce the
196 number of false positive LTR transposons, only those that contained PPT (poly purine extract) or PBS
197 (primer binding sites) within a 30bp radius were kept and the rest filtered. We further filtered the LTR
198 candidates to eliminate three main sources of false positives: tandem local repeats such as centromeric
199 repeats, local gene clusters derived from recent gene duplications, and two other transposable elements
200 located in adjacent regions. We also identified elements with nested insertions. After processing known
201 MITEs and LTR elements, we identified additional repetitive sequences using **RepeatModeler** against a
202 transposase database and excluded gene fragments using **ProtExcluder**.

203 The annotation pipeline **MAKER** (Campbell et al. 2014) was used to identify gene models and predict
204 functional annotations in the draft genome. Both *est2genome* and *protein2genome* modes were used
205 initially to make *ab initio* gene predictions from EST and protein evidence, respectively. The EST
206 evidence was based on our own transcriptome data whereas the protein evidence was gathered from
207 the reference proteomes of six closely related plants available from the SwissProt database: *Arabidopsis*
208 *thaliana, Glycine max, Brassica oleracea, Medicago truncatula, Brassica napus*, and *Brassica rapa*. From
209 the first round of annotation, high confidence models were predicted by the *maker2zff* command with
210 default minimums (50 % of splice sites confirmed by EST alignment, 50 % of exons match an EST
211 alignment, 50 % of exons overlap any evidence, and maximum AED of 0.5). These predictions were used
212 to train the *ab initio* gene predictor **SNAP** (Korf, 2004). A second round of **MAKER** was run using the
213 hidden Markov model (HMM) from **SNAP** rather than the *est2genome* mode. All other settings were the
214 same as for the first run, with the transcripts now being used only as evidence to support *ab initio* gene
215 predictions. Two more rounds of annotation and gene prediction improvement followed. An **Augustus**
216 gene prediction file was also generated for use as a second *ab initio* prediction (Stanke and
217 Morgenstern, 2005). For the final round of annotation in **MAKER**, the final HMM file from **SNAP**, the *A.*
218 *petiolata* species **Augustus** library and gene prediction file were all used in addition to the following
219 settings: always complete, single exons, and using correct EST fusion.

220 We used Benchmarking Universal Single-Copy Orthologs (**BUSCO**) v4.1 (Simão et al., 2015) to assess the
221 assembly quality of the final draft genome. We used plant lineage delineation from the EmbryophytaDB
222 V10 database, focusing on universal orthologs present in > 90 % of lineages, resulting in a total of 1,440
223 BUSCO orthologs. We used **minimap2** to check for assembly contiguity and synteny with the model
224 plant *Arabidopsis thaliana* using the TAIR 10 assembly and up to 20 % sequence divergence. We used
225 the R package ***pafr*** to generate dot-plots for the four largest scaffolds of *A*. *petiolata* and all
226 chromosomes of *A*. *thaliana*.

227 The genome annotation files were curated through **deFusion** to resolve fused gene annotation problems
228 (Wang et al., 2021). We also used **EvidenceModeler** (Haas et al., 2008) to combine *ab initio* gene
229 predictions and protein and transcript alignments into weighted consensus gene structures. The
230 functional annotations were then created using NCBI **BLAST+** and **InterProScan** (Jones et al., 2014) by
231 adding new names, domains, and putative functions to improve the utility of the genome database.

232 Data Availability

233 Raw data used for genome assembly, transcriptome assembly, and the final draft genome is available
234 from DDBJ/ENA/GenBank under BioProject **PRJNA702530**, including the Whole Genome Shotgun project
235 **AHZTW000000000** from **BioSample SAMN17958863**, with sequence reads deposited in **SRR13765007 -**
236 **SRR13765010**. The transcriptome analysis BioSamples **SAMN17984654 - SAMN17984657** and raw
237 sequence reads are also available (**SRR13765003 - SRR13765006**). The latest genome assembly scripts
238 are available on GitHub (https://github.com/ColauttiLab/Ap_Genome_Assembly).

239 Results and Discussion

240 Most (98.7 %) of the essential single-copy genes from **BUSCO** mapped to our assembly (Fig. 1), with 71.5
241 % occurring only once in the assembled genome. Similarly, 98 % of sequence reads from the
242 transcriptome experiment mapped successfully to the assembled genome. Sequencing of the
243 transcriptome libraries yielded a total of 68.1 Gb from 272.5 million paired reads. Trimming sequence
244 reads for quality reduced usable data by less than 2 %.

245 Whole Genome Shotgun (WGS) sequencing produced 45.8 Gb from 229 million HiSeq reads and an
246 additional 7.9 Gb from 15.8 million MiSeq reads. This represents an estimated 68× average coverage of
247 the genome with an estimated genome size of 1.07 Gb. This is 80 % of the published size of 1.35 Gb,
248 which was a rough estimate based on flow cytometry (Barow and Meister, 2002). Initial assembly with
249 **ALLPATHS-LG** resulted in 16,743 scaffolds longer than 1 Kb. These scaffolds were linked using mate
250 pairs. We further scaffolded by merging heterozygous loci using the **redundans** program. The final
251 assembled genome was 1.08 Gb long across 694 scaffolds larger than 1 Kb; more than 75 % of the
252 genome is contained in the five longest scaffolds (Table 1).

253 Cytological studies of *A. petiolata* suggest variation in chromosome number and ploidy (Cavers et al.,
254 1979; Weiss-Schneeweiss and Schneeweiss, 2003), but are ambiguous to whether the hexaploids are
255 allo- or auto-polyploids. Approximately 11 % of scaffolds (143 of 1,332) remained heterozygous after
256 **redundans** assembly, compared to 36 % (6,045 of 16,743) after **ALLPATHS-LG** but prior to redundancy
257 reduction. The average identity within these scaffolds was 99.61 %, meaning that an average of just 0.4
258 % of sites on the heterozygous scaffolds were still heterozygous after redundans assembly.
259 Heterozygosity was similarly low (98.0 %) following **ALLPATHS-LG** assembly but prior to the **rendundans**
260 step, indicating that homozygosity is a feature of the genome rather than an artifact of redundancy
261 reduction. Finally, 27% of **BUSCO** genes (390 of 1440) were duplicated across the assembly (390 of
262 1440) but just 3.2 % of genes mapped to more than two scaffolds (46 of 1440), indicating that the
263 **redundans** step was successful in collapsing homologous chromosomes.

264 This relatively low level of heterozygosity is consistent with a high-inbred autopolyploid or diploid.
265 Geographic variation in ploidy (Esmailbegi et al., 2018; Weiss-Schneeweiss and Schneeweiss, 2003) with
266 relatively simple genome make *A. petiolata* an appealing species to study the role of polyploidy in local

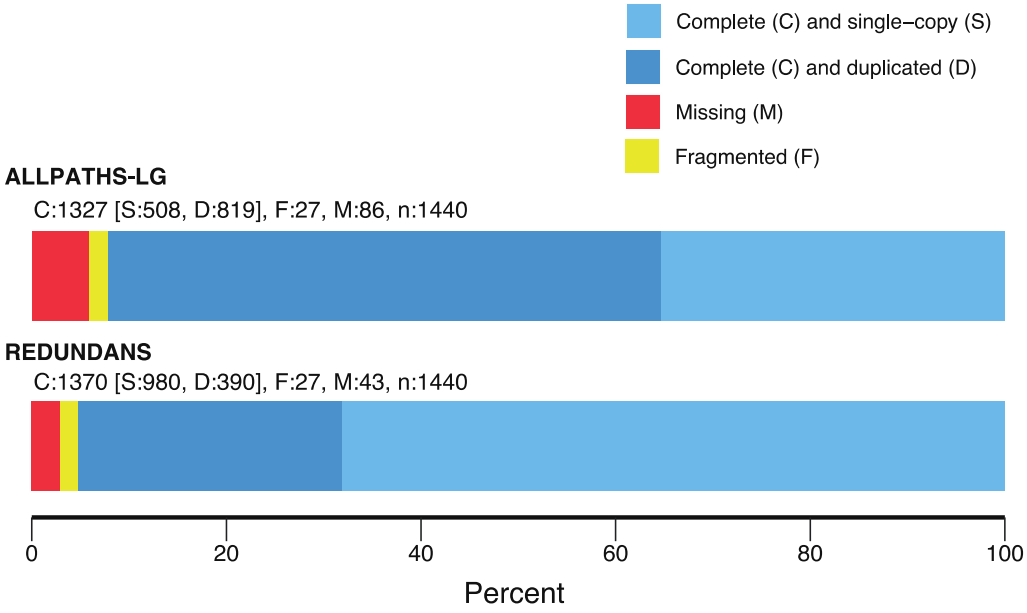
267 adaptation and range expansion, which is an active area of research (e.g. (Pandit et al., 2011; Payseur
268 and Rieseberg, 2016; te Beest et al., 2012)).

269 **Table 1**. Assembly statistics for the *Alliaria petiolata* genome

Statistic	Value
N scaffolds (>= 1000b)	694
N scaffolds (>= 50,000b)	227
Total length	1,075,010,735
Total length (>= 50,000b)	1,071,536,925
Largest scaffold	485,611,451
GC (%)	37.2
N_{50}	121,941,980
N_{75}	40,840,077
L_{50}	2
L_{75}	5
Mean Sequence Length	1,549,006.82

270

271



272

273 **Figure 1.** Percentage of predicted single-copy plant genes from **BUSCO** that are
274 found one (light blue) or more times (dark blue), or are missing (red) or fragmented
275 (yellow) in the annotated genome assembly of *Alliaria petiolata*.

276 Our *de novo* transcriptome assembly included 699,048 putative isoform "transcripts"
277 representing 535 Mb with N50 of 1,233 base pairs. The minimum transcript length was 201 as set
278 by the Trinity default parameter while 1382 (~1.98 %) of transcripts were longer than 5 Kb. These
279 transcripts clustered into 350,672 hypothetical genes with an average of 2.18 isoforms and 26,910
280 (~7.67 %) of hypothetical genes having more than five isoforms. A BLAST search of hypothetical genes to

281 the SwissProt protein database matched 10,352 proteins with at least 90 % coverage of the query
282 sequence, including 7,930 proteins with 100 % coverage (Table 2). The large difference in the
283 number of protein and gene transcript is likely a function of the Trinity algorithm and the
284 ribosome-depletion library preparation method, which includes all RNA transcripts, not just
285 mRNA.

286 Our TE annotation analysis identified 8,220 unique sequences across the *A. petiolata* genome. Of these,
287 112 were classified as LTRs with relatively recent origin (99 % similarity), 240 as relatively old (85 %), and
288 7,137 were classified as miniature inverted transposable elements (MITE). An additional 731 sequences
289 were found to match the DNA transposase family. After masking TE sequences the final gene set from
290 **maker** included 64,770 gene predictions with an average of 6 exons and an average exon length of 251
291 bp (Table 2).

292 A dot-plot comparison of gene synteny with the model plant *Arabidopsis thaliana* (TAIR 10) (Berardini et
293 al., 2015) revealed large blocks of orthologous sequence (Figure 2). However, the arrangement of
294 synteny blocks along chromosomes shows a complete re-arrangement of *A. thaliana* chromosomes
295 when mapped to the *A. petiolata* scaffolds. Despite a high level of asynteny that is characteristic of the
296 Brassicaceae family, the conservation of large synteny blocks holds promise for future research to
297 identify candidate genes and genetic loci of interest for understanding plant invasions. Future research
298 could also investigate whether ploidy variation and gene rearrangements occur among geographically
299 and historically isolated populations of *A. petiolata*, and whether this genomic architecture has played
300 an important role in the spread of the species.

301

302 **Table 2**. Summary statistics of genes annotated for the *Alliaria petiolata* genome assembly.

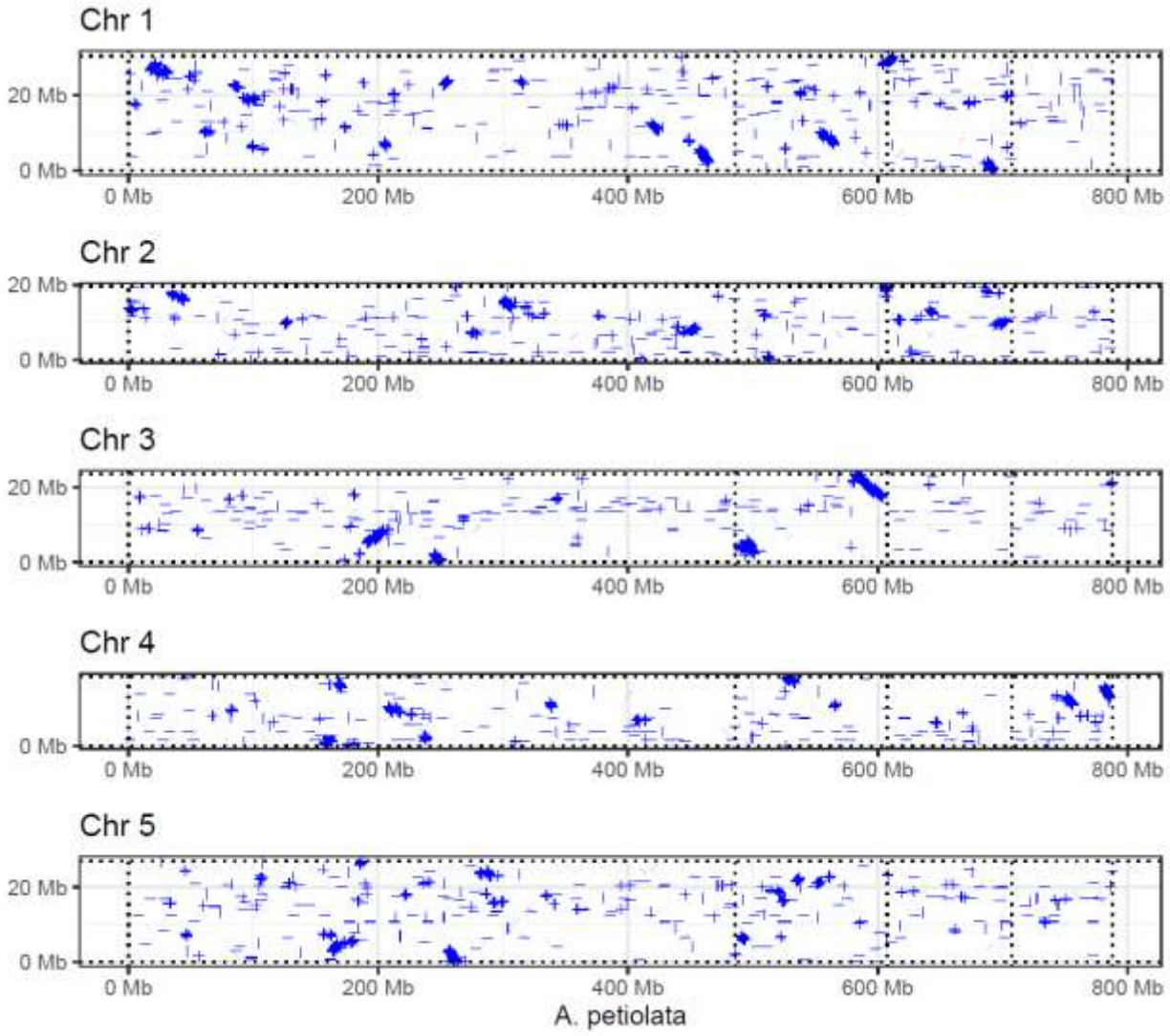
Statistic	Value
Number of genes	64,770
Number of exons	408,155
Number of introns	343,385
Overlapping genes	9,669
Contained genes	1,464
Total gene length	210,804,785
Total exon length	102,316,121
Total intron length	109,175,434
Total CDS length	842,788
% of genome covered by genes	19.6
% of genome covered by CDS	7.8
Mean mRNAs per gene	1
Mean exons per mRNA	6
Mean introns per mRNA	5
Mean gene length	3,255
Mean intron length	318
Mean exon length	251
Mean CDS length	1301

303

304

305

306



307

308 **Figure 2.** Dot-plot showing blocks of synteny between the four largest scaffolds of the *Alliaria petiolata*
309 assembly (x-axis) and five chromosomes of the model plant *Arabidopsis thaliana*. Blue lines show
310 aligned sequences with up to 20 % divergence. Vertical dotted lines denote separation of the major
311 scaffolds of the *A. petiolata* assembly.

312 Conclusions

313 There is a growing interest in the genetic causes and consequences of range expansion and biological
314 invasion. The field of invasion genetics has emerged from ecological and evolutionary studies of invasive
315 species but lacks well-developed model systems. The draft genome and gene annotation reported here
316 represents an important link from the many field and experimental studies of *A. petiolata* to the genetic
317 architecture of adaptation and invasion. High levels of self-fertility and the resultant low levels

heterozygosity observed in the genome will be beneficial for future projects linking ecologically important phenotypes to specific genes. The genomic resources reported here complement available seed resources, experimental findings, and field data to accelerate genomic and molecular studies of *A. petiolata* as a candidate for a model system in invasion genetics.

Acknowledgements

The authors are grateful for bioinformatics support from C Grassa, J Stafford, and H Schmider, hardware support from C MacPhee, and wetlab assistance from M Todesco, W Chen and A Siew. We also thank M Todesco and D Galanti for comments that improved the manuscript. This work was supported by NSERC Discovery grants to RIC and LHR.

Literature Cited

Andrews, S., 2010. FastQC A Quality Control tool for High Throughput Sequence Data [Online] [WWW Document]. URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 11.19.20).

Anthony, M.A., Frey, S.D., Stinson, K.A., 2017. Fungal community homogenization, shift in dominant trophic guild, and appearance of novel taxa with biotic invasion. Ecosphere 8, e01951. https://doi.org/10.1002/ecs2.1951

Barow, M., Meister, A., 2002. Lack of correlation between AT frequency and genome size in higher plants and the effect of nonrandomness of base sequences on dye binding. Cytometry 47, 1–7. https://doi.org/10.1002/cyto.10030

Barrett, S.C.H., 2015. Foundations of invasion genetics: The Baker and Stebbins legacy. Molecular Ecology 24, 1927–1941. https://doi.org/10.1111/mec.13014

Barrett, S.C.H., Colautti, R.I., Dlugosch, K.M., Rieseberg, L.H., 2016. Invasion Genetics: The Baker and Stebbins Legacy. Wiley.

Berardini, T.Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., Huala, E., 2015. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis 53, 474–485. https://doi.org/10.1002/dvg.22877

Bock, D.G., Caseys, C., Cousens, R.D., Hahn, M.A., Heredia, S.M., Huebner, S., Turner, K.G., Whitney, K.D., Rieseberg, L.H., 2015. What we still don't know about invasion genetics. Molecular Ecology 24, 2277–2297. https://doi.org/10.1111/mec.13032

Boheemen, L.A. van, Lombaert, E., Nurkowski, K.A., Gauffre, B., Rieseberg, L.H., Hodgins, K.A., 2017. Multiple introductions, admixture and bridgehead invasion characterize the introduction history of *Ambrosia artemisiifolia* in Europe and Australia. Molecular Ecology 26, 5421–5434. https://doi.org/10.1111/mec.14293

Bossdorf, O., Prati, D., Auge, H., Schmid, B., 2004. Reduced competitive ability in an invasive plant. Ecology Letters 7, 346–353. https://doi.org/10.1111/j.1461-0248.2004.00583.x

Bourne, S.D., Hudson, J., Holman, L.E., Rius, M., 2020. Marine invasion henomics: Revealing ecological and evolutionary consequences of biological invasions, in: Oleksiak, M.F., Rajora, O.P. (Eds.), Population Genomics: Marine Organisms, Population Genomics. Springer International Publishing, Cham, pp. 363–398. https://doi.org/10.1007/13836_2018_21

Callaway, R.M., Cipollini, D., Barto, K., Thelen, G.C., Hallett, S.G., Prati, D., Stinson, K., Klironomos, J., 2008. Novel weapons: Invasive plant suppresses fungal mutualists in America but not in its native Europe. Ecology 89, 1043–1055. https://doi.org/10.1890/07-0370.1

Cavers, P.B., Heagy, M.I., Kokron, R.F., 1979. The biology of Canadian weeds: 35. *Alliaria petiolata* (M. Bieb.) Cavara and Grande. Canadian Journal of Plant Science 59, 217–229.

362 Cipollini, D., Cipollini, K., 2016. A review of garlic mustard (Alliaria petiolata, Brassicaceae) as an
363 allelopathic plant. tbot 143, 339–348. https://doi.org/10.3159/TORREY-D-15-00059
364 Clarke, J.D., 2009. Cetyltrimethyl Ammonium Bromide (CTAB) DNA Miniprep for Plant DNA Isolation.
365 Cold Spring Harb Protoc 2009, pdb.prot5177. https://doi.org/10.1101/pdb.prot5177
366 Colautti, R., Franks, S.J., Hufbauer, R.A., Kotanen, P.M., Torchin, M., Byers, J.E., Pyšek, P., Bossdorf, O.,
367 2014. The Global Garlic Mustard Field Survey (GGMFS): challenges and opportunities of a
368 unique, large-scale collaboration for invasion biology. NeoBiota 21, 29–47.
369 https://doi.org/10.3897/neobiota.21.5242
370 Crescente, J.M., Zavallo, D., Helguera, M., Vanzetti, L.S., 2018. MITE Tracker: an accurate approach to
371 identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics
372 19, 348. https://doi.org/10.1186/s12859-018-2376-y
373 Dlugosch, K.M., Parker, I.M., 2008. Invading populations of an ornamental shrub show rapid life history
374 evolution despite genetic bottlenecks. Ecol. Lett. 11, 701–709. https://doi.org/10.1111/j.1461-
375 0248.2008.01181.x
376 Duchesneau, K., Golemiec, A., Colautti, R.I., Antunes, P.M., 2021. Functional shifts of soil microbial
377 communities associated with Alliaria petiolata invasion. Pedobiologia 84, 150700.
378 https://doi.org/10.1016/j.pedobi.2020.150700
379 Esmailbegi, S., Al-Shehbaz, I.A., Pouch, M., Mandáková, T., Mummenhoff, K., Rahiminejad, M.R.,
380 Mirtadzadini, M., Lysak, M.A., 2018. Phylogeny and systematics of the tribe Thlaspideae
381 (Brassicaceae) and the recognition of two new genera. TAXON 67, 324–340.
382 https://doi.org/10.12705/672.4
383 Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea,
384 T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A.,
385 Nusbaum, C., Lander, E.S., Jaffe, D.B., 2011. High-quality draft assemblies of mammalian
386 genomes from massively parallel sequence data. PNAS 108, 1513–1518.
387 https://doi.org/10.1073/pnas.1017351108
388 Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L.,
389 Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma,
390 F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Trinity:
391 reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol
392 29, 644–652. https://doi.org/10.1038/nbt.1883
393 Gremme, G., Steinbiss, S., Kurtz, S., 2013. GenomeTools: A comprehensive software library for efficient
394 processing of structured genome annotations. IEEE/ACM Transactions on Computational Biology
395 and Bioinformatics 10, 645–656. https://doi.org/10.1109/TCBB.2013.68
396 Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D.,
397 Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N.,
398 Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A.,
399 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for
400 reference generation and analysis. Nature Protocols 8, 1494–1512.
401 https://doi.org/10.1038/nprot.2013.084
402 Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., Wortman, J.R.,
403 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the
404 Program to Assemble Spliced Alignments. Genome Biology 9, R7. https://doi.org/10.1186/gb-
405 2008-9-1-r7
406 Haribal, M., Yang, Z., Attygalle, A.B., Renwick, J.A.A., Meinwald, J., 2001. A Cyanoallyl Glucoside from
407 Alliaria petiolata, as a Feeding Deterrent for Larvae of Pieris napi oleracea. J. Nat. Prod. 64, 440–
408 443. https://doi.org/10.1021/np000534d

409 Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A.,
410 Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R.,
411 Hunter, S., 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics
412 30, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031
413 Korf, I., 2004. Gene finding in novel genomes. BMC Bioinformatics 5, 59. https://doi.org/10.1186/1471-
414 2105-5-59
415 Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357–
416 359. https://doi.org/10.1038/nmeth.1923
417 Lankau, R.A., Nuzzo, V., Spyreas, G., Davis, A.S., 2009. Evolutionary limits ameliorate the negative impact
418 of an invasive plant. Proc. Natl. Acad. Sci. U. S. A. 106, 15362–15367.
419 https://doi.org/10.1073/pnas.0905446106
420 Lewis, K.C., Bazzaz, F.A., Liao, Q., Orians, C.M., 2006. Geographic patterns of herbivory and resource
421 allocation to defense, growth, and reproduction in an invasive biennial, *Alliaria petiolata*.
422 Oecologia 148, 384–395. https://doi.org/10.1007/s00442-006-0380-9
423 Mack, R.N., Simberloff, D., Lonsdale, W.M., Evans, H., Clout, M., Bazzaz, F.A., 2000. Biotic invasions:
424 Causes, epidemiology, global consequences, and control. Ecological Applications 10, 689–710.
425 Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
426 EMBnet.journal 17, 10–12. https://doi.org/10.14806/ej.17.1.200
427 Mooney, H.A., Cleland, E.E., 2001. The evolutionary impact of invasive species. PNAS 98, 5446–5451.
428 https://doi.org/10.1073/pnas.091093398
429 Nuzzo, V., 1993. Distribution and spread of the invasive biennial *Alliaria petiolata* (garlic mustard) in
430 North America., in: McKnight, B. (Ed.), Biological Pollution: The Control and Impact of Invasive
431 Exotic Species. Indiana Academy of Science, Indianapolis, Indiana, pp. 137–145.
432 Pandit, M.K., Pocock, M.J.O., Kunin, W.E., 2011. Ploidy influences rarity and invasiveness in plants.
433 Journal of Ecology 99, 1108–1115. https://doi.org/10.1111/j.1365-2745.2011.01838.x
434 Payseur, B.A., Rieseberg, L.H., 2016. A genomic perspective on hybridization and speciation. Molecular
435 Ecology 25, 2337–2360. https://doi.org/10.1111/mec.13557
436 Prati, D., Bossdorf, O., 2004. Allelopathic inhibition of germination by Alliaria petiolata (Brassicaceae).
437 American Journal of Botany 91, 285–288. https://doi.org/10.3732/ajb.91.2.285
438 Pryszcz, L.P., Gabaldón, T., 2016. Redundans: an assembly pipeline for highly heterozygous genomes.
439 Nucleic Acids Res 44, e113–e113. https://doi.org/10.1093/nar/gkw294
440 Sax, D.F., Stachowicz, J.J., Gaines, S.D., 2005. Species Invasions: Insights into Ecology, Evolution, and
441 Biogeography. Sinauer.
442 Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing
443 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31,
444 3210–3212. https://doi.org/10.1093/bioinformatics/btv351
445 Stanke, M., Morgenstern, B., 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that
446 allows user-defined constraints. Nucleic Acids Res 33, W465–W467.
447 https://doi.org/10.1093/nar/gki458
448 Stinson, K., Kaufman, S., Durbin, L., Lowenstein, F., 2007. Impacts of garlic mustard Invasion on a forest
449 understory community. nena 14, 73–88. https://doi.org/10.1656/1092-
450 6194(2007)14[73:IOGMIO]2.0.CO;2
451 te Beest, M., Le Roux, J.J., Richardson, D.M., Brysting, A.K., Suda, J., Kubešová, M., Pyšek, P., 2012. The
452 more the better? The role of polyploidy in facilitating plant invasions. Ann Bot 109, 19–45.
453 https://doi.org/10.1093/aob/mcr277
454 USDA, N., 2020. The PLANTS Database (http://plants.usda.gov, 10 April 2020). National Plant Data Team,
455 Greensboro, NC.

456 Wang, W., Zhang, X., Dai, D.-Q., 2021. DeFusion: a denoised network regularization framework for multi-
457 omics integration. Brief Bioinform bbab057. https://doi.org/10.1093/bib/bbab057
458 Weiss-Schneeweiss, H., Schneeweiss, G.M., 2003. Karyological investigations of selected angiosperms
459 from Georgia and Azerbaijan. Acta Biologica Cracoviensia Series Botanica 45, 49–56.
460