Genome assembly and annotation of *Arabidopsis halleri*, a model for heavy metal hyperaccumulation and evolutionary ecology

ROMAN V. BRISKINE,* TIMOTHY PAAPE,* RIE SHIMIZU-INATSUGI,* TOMOAKI NISHIYAMA,† SATORU AKAMA,‡ JUN SESE§ and KENTARO K. SHIMIZU*§

*Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, Zurich CH-8057, Switzerland, †Advanced Science Research Center, Kanazawa University, 13-1 Takara-machi, Kanazawa 920-0934, Japan, ‡Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi Koto-ku, Tokyo 135-0064, Japan, §Kihara Institute for Biological Research, Yokohama City University, 642-12 Maioka, Totsuka-ward, Yokohama 244-0813, Japan

Abstract

The self-incompatible species *Arabidopsis halleri* is a close relative of the self-compatible model plant *Arabidopsis thaliana*. The broad European and Asian distribution and heavy metal hyperaccumulation ability make *A. halleri* a useful model for ecological genomics studies. We used long-insert mate-pair libraries to improve the genome assembly of the *A. halleri* ssp. *geminifera* Tada mine genotype (W302) collected from a site with high contamination by heavy metals in Japan. After five rounds of forced selfing, heterozygosity was reduced to 0.04%, which facilitated subsequent genome assembly. Our assembly now covers 196 Mb or 78% of the estimated genome size and achieved scaffold N50 length of 712 kb. To validate assembly and annotation, we used synteny of *A. halleri* Tada mine with a previously published high-quality reference assembly of a closely related species, *Arabidopsis lyrata*. Further validation of the assembly quality comes from synteny and phylogenetic analysis of the HEAVY METAL ATPASE4 (HMA4) and METAL TOLERANCE PROTEIN1 (MTP1) regions using published sequences from European *A. halleri* for comparison. Three tandemly duplicated copies of HMA4, key gene involved in cadmium and zinc hyperaccumulation, were assembled on a single scaffold. The assembly will enhance the genomewide studies of *A. halleri* as well as the allopolyploid *Arabidopsis kamchatica* derived from *A. lyrata* and *A. halleri*.

**Keywords**: *Arabidopsis halleri*, de novo assembly, functional annotation, heavy metal hyperaccumulator, Tada mine

Received 4 April 2016; revision received 4 August 2016; accepted 16 September 2016

Introduction

Ecological genomics studies in plant species can now be extended to close relatives of model and to nonmodel species by improvements in *de novo* assembly and gene annotation methods (Slotte et al. 2013; Liu et al. 2014; Lobréaux et al. 2014). *Arabidopsis halleri* (2n = 16) is a close relative of the model plant *A. thaliana* (2n = 10) (Novikova et al. 2016) and is itself becoming a model for ecological and evolutionary genetics studies. Available genetic tools include transgenic techniques using tissue culture (Hanikenne et al. 2008), complementation test using viable F1 hybrid plants with *A. thaliana* mutants (Shimizu 2002), and QTL maps using *Arabidopsis lyrata* (2n = 16) (Willems et al. 2007). Self-incompatibility in *A. halleri* maintains high genetic diversity in the species, on average an order of magnitude over *A. thaliana* (Castric et al. 2008; Roux et al. 2011). A high-quality genome assembly of *A. lyrata* was recently published (Hu et al. 2011) and is the second well-assembled *Arabidopsis* genome along with *A. thaliana*. The *A. lyrata* assembly was generated from a self-compatible accession, but no self-compatible population of *A. halleri* has been reported.

*Arabidopsis halleri* has attracted the study of speciation and ecological adaptation (Ramos-Onsins et al. 2004). The split of *A. halleri* and *A. lyrata* is estimated to coincide with the tandem duplication of *HEAVY METAL ATPASE4* (HMA4), suggesting it was an ecological speciation event where *A. halleri* evolved heavy metal tolerance (Roux et al. 2011). Although gene flow between the two species is limited, occasional introgression events including 5-haplogroups were detected (Ramos-Onsins et al. 2004; Castric et al. 2008). The self-compatible
alotetraploid species Arabidopsis kamchatrica (Tsuchimatsu et al. 2012) is derived from the hybridization of A. hallieri and A. lyrata and has a broad climatic niche compared with parental species (Hoffmann 2005; Shimizu et al. 2005; Shimizu-Inatsugi et al. 2009; Schmickl et al. 2010; Novikova et al. 2016). Other molecular ecological and evolutionary studies using A. hallieri include defence against herbivores by heavy metals (Kazemidinan et al. 2014) and by trichomes (Shimizu 2002; Kawagoe et al. 2011), local adaptation (Fischer et al. 2013; Kubota et al. 2015), self-incompatibility and mating systems (Shimizu & Purugganan 2005; Bechsgaard et al. 2006; Gouvet et al. 2012; Durand et al. 2014), and gene expression in natural environments (in natura) (Aikawa et al. 2010; Shimizu et al. 2011; Kudoh 2016).

A major area of research in A. hallieri has focused on the study of heavy metal tolerance and hyperaccumulation (Bert et al. 2000; Chiang et al. 2006; Hanikenne et al. 2008; Krämer 2010). Heavy metal hyperaccumulation is a constitutive phenotype in A. hallieri, and all tested genotypes are able to accumulate high levels of cadmium and zinc in leaves (Kubota & Takenaka 2003; Chiang et al. 2006; Pauwels et al. 2006; Talke et al. 2006). QTL studies using crosses between nonhyperaccumulator A. lyrata and A. hallieri showed that the two major loci that cosegregate with cadmium and zinc tolerance are the HEAVY METAL ATPASE4 (HMA4) and the cation diffusion facilitator (CDF) protein METAL TOLERANCE PROTEIN1 (MTP1 also called ATCDF1 or ZAT1) (Courbot et al. 2007; Willems et al. 2007). The sequencing of a BAC assembly in the A. hallieri Langelsheim genotype, which includes three paralogous HMA4 copies (Hanikenne et al. 2008), has led to the estimation of a hard selective sweep at HMA4 in A. hallieri based on surrounding genetic diversity (Hanikenne et al. 2013). BAC sequences of the A. hallieri genotype from the Auby mine site showed that MTP1 is duplicated with up to five copies in this population (Shahzad et al. 2010). Overall, conserved sequence diversity of duplicated copies of HMA4 and MTP1 along with exceptionally high additive expression likely explains constitutive hyperaccumulation in A. hallieri.

Among the largest genomic regions of A. hallieri that have been published are BAC assemblies of the HMA4 region (290 kb) from the Langelsheim accession (Hanikenne et al. 2008, 2013), the MTP1-A region (110 kb) from the Auby mine accession (Shahzad et al. 2010) and the S-locus of multiple haplotypes (ranging from 25 kb to 121 kb) (Gouvet et al. 2012). The first two regions harbour genes involved in heavy metal hyperaccumulation and tolerance. Previously, we constructed a medium-quality de novo assembly of A. hallieri ssp. gemmifera (W302) collected at the Tada mine site in Japan to establish a next-generation read sorting pipeline (HomeoRoq) for distinguishing homolog origins of RNA-seq reads in synthesized A. kamchatrica (Akama et al. 2014). That assembly also contained several long scaffolds over 100 kb. More recently, another accession of A. hallieri ssp. gemmifera (IBO380) from Mt. Ibuki, Japan, was sequenced for genomewide selection scans for altitudinal adaptation (Kubota et al. 2015). While the two reference assemblies reported in Akama et al. (2014) and Kubota et al. (2015) covered ~88 and 98% of the A. hallieri genome, respectively, they were comprised mostly of short contigs (N50 length ~ 18 kb and 5 kb, respectively). Another study, focusing on local adaptation of five populations in the Swiss Alps, analysed pooled resequencing data using the genome of A. thaliana as a reference due to the lack of a high-quality genome of A. hallieri, which limited their analyses to conserved coding regions (Fischer et al. 2013). For such studies to identify candidate single nucleotide polymorphisms (SNPs) under selection, long scaffolds would be valuable for detecting long-range linkage disequilibrium caused by recent selection. Finally, as sequence data accumulates for A. hallieri along with phenotypic data, long scaffolds are essential for identifying the genetic architecture of quantitative traits.

We present a long-scaffold genome assembly of A. hallieri ssp. gemmifera (Tada mine) that was constructed using long mate-pair libraries (insert sizes ranging from 2.5 kb to 22 kb) along with the existing short-insert paired-end libraries from Akama et al. (2014). The Tada mine site in Japan produced silver and copper for about 1000 years (Kobata 1968) until its commercial shutdown in 1973 (Tada Silver Mine Historic Site Preservation Association 2007). It served as a resource for the Toyotomi dynasty during the 16th century (Azuchi–Momoyama period) and then experienced the peak of production around 1660s (Edo period). This site is contaminated with both cadmium and zinc (Paape et al. 2016) (see Supporting Information for details), and A. hallieri plants are abundant in this area along with another hyperaccumulator, Athyrium yokoscense, which was traditionally used as an indicator species to find ore deposits in Japan (Miyake 1897). While A. hallieri is obligately outcrossing, we drastically reduced heterozygosity in the Tada mine accession by forced self-fertilization (five times through bud pollination), making assembly more straightforward with higher homozygosity. We compared the synteny of the Tada mine assembly to the fully assembled A. thaliana (The Arabidopsis Genome Initiative 2000) and A. lyrata (Hu et al. 2011) genomes, and to duplicated HMA4 and MTP1 regions from previously published BAC sequences of A. hallieri (Shahzad et al. 2010; Hanikenne et al. 2013). The successful assembly of complex gene duplications in our Tada mine reference assembly may facilitate evolutionary studies of duplicated genes that led to a multilocus adaptive phenotype in A. hallieri.
Materials and methods

Study species and samples

Arabidopsis halleri (L.) O’Kane and Al-Shehbaz (basionym Arabis halleri) is a diploid species (2n = 16) distributed in Europe and East Asia (O’Kane & Al-Shehbaz 1997; Al-Shehbaz & O’Kane 2002). Populations of A. halleri in both European and East Asian regions are distributed across highly variable lowland and alpine areas (Fischer et al. 2013; Kubota et al. 2015) that experience extremes in temperature, solar radiation and precipitation. In addition, A. halleri is also found across highly variable soil types and is able to tolerate extreme heavy metal contamination in Europe and Asia (Hanikenne et al. 2013). The split of A. halleri from A. thaliana has long been considered to be about 5 Mya based on the estimation of evolutionary rate by Koch et al. (2000), while more recent estimation by Ossowsky et al. (2010) suggested 13.0 or 17.9 Mya (reviewed by Shimizu & Tsuchimatsu 2015). Subsequently, the speciation of A. halleri and A. lyrata occurred ~337 kya or 2.5 Mya depending on assumptions (Castric et al. 2008; Roux et al. 2011). Allopolyploid origins of A. kamchatica by the hybridization of A. halleri and A. lyrata occurred much more recently (~20 kya) (Tsuchimatsu et al. 2012).

A number of subspecies of A. halleri have been proposed based on morphology. Al-Shehbaz & O’Kane recognized three subspecies (ssp. halleri and ssp. ovirensis in Europe and ssp. gemmifera in East Asia). Kolnik & Marhold (2006) added ssp. tatraica and ssp. dacica from East Europe. We consider East Asian plants as A. halleri ssp. gemmifera (Matsumura) O’Kane & Al-Shehbaz (also called Arabis gemmifera or Cardamine gemmifera) (O’Kane & Al-Shehbaz 1997; Hoffmann 2005). We sequenced an accession called “Tada mine” (W302 in our stock number) of A. kamchatica (effort Cov) indicates the coverage from the reads kept by the assembler.

Genome size estimation by flow cytometry

We measured the nuclear DNA content of Tada mine accession by flow cytometry. Flower petals were processed together with internal reference standard tomato leaf (Lycopersicon esculentum cv. ‘Stupicke’; assumed

| Type           | Insert Size | Reads     | Cov | % Used | Eff Cov |
|----------------|-------------|-----------|-----|--------|---------|
| Paired End     | 200 bp      | 140 506 146 | 56.8 | 70.4   | 40.0    |
| Paired End     | 500 bp      | 128 033 686 | 51.7 | 70.8   | 36.6    |
| Paired End     | 800 bp      | 39 614 066 | 16.0 | 71.7   | 11.5    |
| Total          |             | 308 153 898 | 124.5 | 88.1   |         |
| Paired End     |             | 171 385 818 | 69.2 | 52.9   | 36.6    |
| Mate Pair      | 3–5 kb      | 184 082 138 | 74.4 | 51.4   | 38.2    |
| Mate Pair      | 7–11 kb     | 172 373 480 | 69.6 | 16.5   | 11.5    |
| Mate Pair      | 11–15 kb    | 84 648 858  | 34.2 | 16.3   | 5.6     |
| Mate Pair      | 15–22 kb    | 85 893 936  | 34.7 | 46.7   | 16.2    |
| Mate Pair      | 22–38 kb    | 14 873 342  | 6.0  | 6.4    | 0.4     |
| Total          |             | 713 257 572 | 288.2 |        | 108.5   |
| Mate Pair      |             | 1 021 411 470 | 412.7 |        | 196.6   |
ALLPATHS-LG was executed with default parameters and expected insert sizes. Then, the insert size parameters were changed to the values calculated in the first step, and ALLPATHS-LG was run again. The assembly job completed in 66 h using 20 cores on a Linux cluster with the peak memory utilization of 126 GB.

Improving assembly using synteny

Because *A. halleri* and *A. lyrata* diverged recently (Schmickl et al. 2010; Roux et al. 2011) and each has eight chromosomes (Al-Shehbaz & O’Kane 2002), we used the previously published *A. lyrata* reference genome (Hu et al. 2011) to perform genome-wide synteny analysis (see Fig. S1, Supporting Information for assembly pipeline). The complete genome, coding sequences and gene annotation of *A. lyrata* strain MN47 v1.07 were downloaded from the PHYTOZOME v9.0 website (http://phytozome.jgi.doe.gov). Coding sequences of *A. lyrata* were aligned to the *A. halleri* assembly using BLAT v3.5 (Kent 2002) with default parameters except maximum intron size. Because the longest intron in the *A. lyrata* assembly was 44,703 bp, we set the maximum intron size to 50 kb. Hits were filtered, sorted and merged into syntenic regions using custom Perl scripts (see the Data Accessibility section). We only considered the hits covering at least 85% of the query sequence and accepted the hit from a syntenic gene even when it did not have the highest score for the locus. If an *A. halleri* scaffold contained two neighbouring loci that were syntenic in *A. halleri* to two *A. lyrata* regions located on different chromosomes or more than 100 kb apart, the scaffold was split into two parts by removing the sequence of unknown nucleotides. Scaffolds were only split if the sequence of unknown nucleotides at the cut site spanned at least 50 bp, assuming that longer N-stretches would indicate that the support for contig splicing came only from longer-insert libraries with lower quality and without gap-filling alignments. After this correction, the scaffolds were sorted by length in descending order and named sequentially beginning with scaffold_1. We also used published BAC sequences (see phylogeny sections below) for the *HMA4* region from the Langelsheim accession and the *MTP1* region from the Auby accession to determine synteny for loci containing known duplications in other *A. halleri* accessions. Because these BAC sequences are among the longest published scaffolds in *A. halleri*, the synteny analysis also serves as validation of our assembly for complex regions.

Heterozygosity estimation

To obtain an estimate of genomewide heterozygosity, we aligned all reads from 200-bp and 500-bp insert libraries against the assembly using BWA v0.7.2 (Li & Durbin 2009) and called variants using HaplootypeCaller from GATK package v3.4-0 (McKenna et al. 2010) following established best practices (DePristo et al. 2011; Van der Auwera et al. 2013). Low-quality variants and variants in known repetitive elements were discarded (see Supporting Information for more details). The number of the remaining variants was divided by the total count of nonmissing bases with nonzero coverage in the assembly to estimate the heterozygosity level.

Annotation

To annotate the genome of *A. halleri*, we integrated RNA-seq data from leaves and roots (Paape et al. 2016) with the AUGUSTUS gene prediction program (Stanke et al. 2006; AUGUSTUS Development Team 2014) (see Fig. S2, Supporting information for annotation pipeline). Unstranded paired-end 100-bp reads from the *A. halleri* W302_L4 (leaf) and W302_R1 (root) libraries were individually aligned to the *A. halleri* W302 reference genome using STAR v2.4.0i (Dobin et al. 2013) with nondefault parameters. (For the complete list of utilities and the detailed description of parameters used at each step, see the readme file in the online code repository at https://gitlab.com/rbrisk/ahalassembly.) Compared to TorHAT v2.0.13 (Trapnell et al. 2009) ran with modified parameters, STAR yielded more unique alignments ultimately resulting in a higher number of hints for AUGUSTUS (Table S1, Supporting information). Intron hints were extracted from the alignment and merged with repetitive element (nonexonpart) hints derived from the REPEATMASKER v4.0.5 (Smit et al. 1996) output. The merged hints were used for the preliminary AUGUSTUS v3.0.3 run. Introns were extracted from the output and used to generate exon–exon junction database. The original reads were aligned against exon–exon junction sequences using bowtie2 v2.2.4 (Langmead & Salzberg 2012) rather than STAR because splice-aware alignment was not necessary in this case. Spliced reads were removed from the STAR output, and the remaining reads were merged with reads that aligned to exon–exon junctions. The merged reads were filtered to include only the read pairs with high-quality alignments. The filtered alignments were used to generate intron hints for the final AUGUSTUS run. Human readable functional descriptions were added using the AHRD tool and following its documentation (Tomato Genome Consortium 2012). Reciprocal best BLAST hits were calculated by aligning all coding sequences (the longest transcript per gene) corresponding to one annotation version against all coding sequences corresponding to another annotation version (see Table S3, Supporting information for the list of annotations) both ways using NCBI BLAST+ v2.2.29 and comparing the scores for hits longer than 200 bp.
Phylogeny of HMA4 duplications

Previously published BAC sequences (GenBank accessions: EU382072.1 and EU382073.1) covering the tandemly duplicated HMA4 gene in the Langelsheim accession of A. halleri ssp. halleri (Hanikenne et al. 2008, 2013) were used to assess synteny within the scaffold containing the HMA4 region in our A. halleri Tada mine assembly. Because the two Langelsheim BAC sequences overlap, they were spliced together using the minimus2 application from the AMOS A. halleri ssp. halleri sequence assembly. Long-insert libraries were used in conjunction with reads from three paired-end libraries constructed for the previous draft assembly v1.0 of A. halleri Tada mine (Akama et al. 2014) in an attempt to improve scaffold length. While version 1.0 covered 88% of the estimated genome size (250 Mb by flow cytometry), the highly fragmented (low N50) assembly limited analyses to gene-coding sequences. We constructed six mate-pair libraries with insert sizes ranging from 3 kb to 38 kb and sequenced them to obtain more than 700 million additional reads (Table 1). A large proportion of the reads were either duplicates or had low complexity and were discarded by the assembler (see ‘% Used’ column in Table 1). The effective coverage (i.e. the coverage of the kept reads) for the mate-pair libraries was over 100×, and the total effective coverage encompassing three paired-end and the six mate-pair libraries was almost 200× (Table 1).

The Tada mine v2.2 assembly has a smaller total size compared to v1.0 (196 Mb and 221 Mb, respectively; Table 2). The smaller size of v2.2 is due to the contig size filtering automatically performed by the assembler. The shortest scaffold length in v1.0 was 100 bp, while it was 932 bp in v2.2. When we removed all scaffolds shorter than 932 bp from v1.0, its total length decreased to 180 Mb. Likewise, the percentage of missing nucleotides is lower in v1.0 because longer scaffolds often represent concatenation of contigs with long stretches of Ns inserted to preserve expected distance between the contigs. Meanwhile, very short sequences in the previous assembly represented individual contigs without missing data.

Using flow cytometry, we estimated the genome size of Tada mine to be 250 Mb, which is slightly smaller than 255 Mb (Johnston et al. 2005) and 279 Mb (Wolf et al. 2005) and 279 Mb (Wolf et al. 2005). NG50 values are based on the expected nuclear genome size of 250 Mb as determined by flow cytometry (Earl et al. 2011). NG50 values are based on the expected nuclear genome size of 250 Mb as determined by flow cytometry.
bly, the HMA4 duplicated identity of the assembly. That this high homozygosity contributed to the high quality genome is very low (0.0402%). We suggest using the assembly and the genomewide data of the heterozygosity into about 1/32. In agreement with heterozygosity into about half, we should have reduced self-fertilization and each selfing is expected to reduce sequenced a genotype that experienced five rounds of A. halleri instead. For example, BAC sequencing showed that assemble from short reads necessitating BAC sequencing the high accuracy of the assembly.

Generally, high heterozygosity has a negative effect on assembly quality (Schatz et al. 2012). Because we sequenced a genotype that experienced five rounds of self-fertilization and each selfing is expected to reduce heterozygosity into about half, we should have reduced the heterozygosity into about 1/32. In agreement with this expectation, the heterozygosity that was estimated using the assembly and the genomewide data of A. halleri Tada mine genome is very low (0.0402%). We suggest that this high homozygosity contributed to the high quality of the assembly.

Annotation

Using the hints derived from RepeatMasker (Smit et al. 1996) output and RNA-seq data, AUGUSTUS (Stanke et al. 2006) identified 34 553 putative transcripts corresponding to 32 553 loci (Table S2, Supporting information). The number of genes is comparable to A. lyrata, for which 32 670 genes were predicted but alternative transcripts were not reported by Hu et al. (2011). The gene number is, however, higher than 28 775 genes in A. thaliana TAIR10 annotation (The Arabidopsis Genome Initiative 2000). This can be explained in part by larger genome size of A. halleri compared to A. thaliana (250 Mb in eight chromosomes vs. 125 Mb in five chromosomes, respectively). In addition, our annotation may contain more pseudogenes than the more meticulously curated A. thaliana annotation. On the other hand, the total of 35 286 transcripts were reported for A. thaliana, suggesting that AUGUSTUS did not report some of the alternative transcripts in A. halleri. Overall, 25 328 coding sequences could be aligned against A. thaliana TAIR10 gene models and 21 433 of them were reciprocal best BLAST hits (Table S3, Supporting information). The remaining alignable genes could be duplicates of genes that only have a single copy in A. thaliana or homologous genes that are less diverged in A. halleri than A. thaliana. It is also possible that some alleles appear as separate sequences due to misassembly.

With hints based on RNA-seq data from root and leaf tissues of A. halleri Tada mine, AUGUSTUS identified more gene models than without the RNA-seq hints (32 553 and 29 628, respectively; Table S2, Supporting information). The two annotations have 22 733 genes with identical coding sequences but only 9098 identical gene models. The rest of the genes with matching coding sequence have different UTRs or include extra transcripts. In addition, using RNA-seq hints resulted in 120 instances of gene model fusion (when a gene model in one version included exons from multiple gene models in another version) and 1358 instances of gene model splits (when a gene model was split into multiple gene models). The RNA-seq-based annotation also yields more reciprocal best BLAST hits (21 433 vs. 21 018, respectively) with A. thaliana TAIR10 annotation.

Synteny-based adjustments

Arabidopsis halleri is a close relative of A. thaliana and A. lyrata, both of which have high-quality published genomes (The Arabidopsis Genome Initiative 2000; Hu et al. 2011). Unlike A. thaliana which has only five chromosomes, both A. halleri and A. lyrata have eight chromosomes (reviewed in Hunter & Bomblies 2010) and extensive synteny is expected between these two species. Successful QTL mapping (Willems et al. 2007) also suggests large-scale synteny. To assess the quality of our assembly, we extracted the coding sequences from the A. lyrata reference genome (Hu et al. 2011) and aligned them against our assembly.
Overall, we found 1303 syntenic regions. The number is smaller than the total number of scaffolds in our *A. halleri* assembly because some scaffolds did not have any alignment hits. On the other hand, we discovered 725 cases where two neighbouring regions located on the same scaffold were syntenic to loci located either on two different chromosomes or more than 100 kb apart. While some of those instances may represent genuine structural rearrangements, the close relationship between the species suggests that many of them are caused by misassembly. Therefore, we took a conservative approach and split a scaffold whenever two such regions were joined by 50 or more Ns. Based on these criteria, we made cuts at 454 sites on 162 scaffolds. To differentiate the updated assembly from the assembler's original output, we denote them as v2.2 (cut) and v2.0 (uncut).

**Heavy metals in soil**

The metal concentrations in soil samples of the Tada mine site were sufficiently high to be considered heavy metal contaminated (metalliferous) according to Bert *et al.* (2002) and exceeded several legislative thresholds (Appendix S1, Supporting Information). Accordingly, W302 showed the high level of zinc hyperaccumulation (Paape *et al.* 2016). Both of these data sets underline the functional importance of the heavy metal hyperaccumulation genes such as *HMA4* and *MTP1* for individuals growing at the site.

**HMA4 (cadmium and zinc ATPase transporter) region**

We compared the Langelsheim *HMA4* BAC assembly (Hanikenne *et al.* 2013) and our assembled *HMA4* scaffold for Tada mine (Fig. 1 and Table S4, Supporting Information); we will call them haplotypes hereafter. With slight differences in physical distances among three duplicated *HMA4* gene copies, surrounding gene synteny is highly similar with some minor differences in flanking genes (see Supporting Information for details). Our assembly of the *HMA4* flanking region is also consistent with that of *A. thaliana* chromosome 2 (TAIR10) and *A. lyrata* LG3 chromosome 3 (Shahzad *et al.* 2010) with minor rearrangements (Supporting Information).

The phylogeny of the *HMA4* homologs (Fig. 2) did not show pairing of syntenic copies (for example, *HMA4-1* of Tada mine and Langelsheim accessions did not cluster), but rather the three copies clustered within *A. halleri* accessions. The clustering of the three tandem duplicates of Tada mine accession was highly supported (1.00 posterior probability in the MrBayes tree in Fig. 2, 100 bootstrap support in maximum-likelihood analysis). These copies show greater per cent identity within Tada mine *A. halleri* than between the two accessions (Fig. 2). Because the presence of three *HMA4* copies is widespread in *A. halleri* samples (Hanikenne *et al.* 2013; Paape *et al.* 2016), the data are consistent with gene conversion events homogenizing the three copies in the Tada mine accession (see Supporting Information for details). To conduct a formal test for gene conversion, additional data from several *A. halleri* accessions would be necessary (Mansai & Innan 2010).

**MTP1 (metal tolerance protein 1) region**

We also checked for synteny in scaffolds containing *MTP1* orthologs in *A. halleri* Tada mine and an accession...
of *A. halleri* ssp. *halleri* from Auby mine, which was shown to possess five paralogous copies of *MTP1* located on three separate chromosomes (Shahzad et al. 2010). These five duplicated copies of *MTP1* genes were captured on four BAC sequences in the Auby mine accession. Using a BLAST homology search, we detected three rather than five *MTP1* paralogs on three separate scaffolds in our *A. halleri* Tada mine assembly. This is unsurprising considering that the study of Shahzad et al. (2010) did not detect five copies in all genotypes of *A. halleri* ssp. *halleri* where some genotypes showed three or four copies rather than five. Dräger et al. (2004) also identified only three *MTP1* copies in the Langelsheim accession.

The longest of the four Auby BAC sequences contains a tandem duplication of *MTP1* (*AhMTP1-A1* and –A2) where both copies are nearly identical at the nucleotide level (Shahzad et al. 2010). We can infer that the *MTP1-A* copy in *A. halleri* is the ancestral copy of the other duplicated copies as this region is syntenic with both *A. thaliana* and *A. lyrata* (Fig. 3; Table S5, Supporting information), each containing only a single *MTP1* copy. In addition to synteny with corresponding *A. thaliana* and *A. lyrata* chromosomal regions containing *MTP1*, the order of genes surrounding *MTP1-A* on the Auby BAC assembly and Tada mine scaffold_22 (g09643) is also highly similar with the same adjacent gene models as in *A. thaliana* and *A. lyrata*. Unlike the Auby mine *MTP1-A* region, the Tada mine scaffold_22 does not contain a tandem duplication of *MTP1-A*, which indicates an independent gene duplication in the Auby lineage (Fig. 4). Despite marginal support for some nodes in the *MTP1* phylogeny, it is important to note that other *MTP1-B* and *C* copies in Tada mine form lineages with the corresponding Auby mine *MTP1* copies. This is supported by synteny of the surrounding genes in our scaffolds containing each *MTP1* copy with the *A. lyrata* genome (Table S5, supporting information; also see reciprocal best BLAST hits deposited to Dryad) and by QTL maps generated by crosses of *A. halleri* and *A. lyrata* (Willems et al. 2007; Shahzad et al. 2010). The *A. lyrata* assembly analysis, along with the markers used in Shahzad et al. (2010), indicates that the three Tada mine genes g10163 (on scaffold_22), g28207 (scaffold_154) and g18715 (scaffold 61) are syntenic with *MTP1-A*, *MTP1-B* and *MTP1-C*, respectively, in Auby mine *A. halleri*.
Conclusion

The recent interest in developing other model Brassicaceae systems has been facilitated by advances in genome assembly and DNA polymorphism data sets. This family possesses many species with unique global distributions, mating systems, life histories and adaptations. Within the genus Arabidopsis, A. thaliana, A. lyrata, A. halleri and A. kamchatka provide opportunities to study ecological genomics under different temperature regimes, along latitudinal and longitudinal clines, variable altitudes and variable soil types. The long-scaffold assembly of the A. halleri Tada mine accession can now be used to identify long-range patterns of polymorphism.

Fig. 3 Synteny among MTP1-A regions of A. halleri Tada mine, A. halleri Auby (Shahzad et al. 2010), A. thaliana (TAIR10) and A. lyrata MN47 (Hu et al. 2011). Gene connections are based on all-vs-all BLAST hits among all coding sequences in the region. MTP1 is highlighted in orange and is orthologous to AT2G46800.1 in A. thaliana. A. halleri Auby contains two copies of the gene.

Fig. 4 Phylogenetic relationships between A. halleri Tada mine assembly and A. halleri Auby MTP1 orthologs and paralogs. Colours correspond to the MTP1 copies on linkage groups and relative to syntenic A. thaliana regions in Shahzad et al. (2010). Outgroup sequences are A. lyrata ssp. petraea (AJ704807.1), A. lyrata ssp. lyrata (XM_002880219.1), A. lyrata mRNA (AY483147.1) and A. thaliana (AT2G46800.1).
and diversity, and further genotype–phenotype association studies where knowledge of the genetic architecture of complex phenotypes such as flowering time and heavy metal tolerance is needed.

Acknowledgements

We thank the Functional Genomic Center Zurich for technical support; Hiroshi Kudoh, Kiyotaka Okada, Nobuyoshi Nakajima and Heidi Lischer for discussion and logistic support; Eiji Izawa and Mika Aoki for the information on the history of Tada mine. The study was supported by Swiss National Science Foundation to KKS, the University Research Priority Program of Evolution in Action of the University of Zurich to KKS and RSI, NEXT KAKENHI Grant Number 16H06469, 26113709 and Young Investigator Award of Human Frontier Science Program to KKS and JS, European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no GA-2010-267243 – PLANT FELLOWS to RVB and TP, Marie-Heim Völkmann grant by Swiss National Science Foundation to RSI.

References

Aikawa S, Kobayashi MJ, Satake A, Shimizu KK, Kudoh H (2010) Robust control of the seasonal expression of the Arabidopsis FLC gene in a fluctuating environment. Proceedings of the National Academy of Sciences of the United States of America, 107, 11632–11637.

Akama S, Shimizu-Inatsugi R, Shimizu KK, Sese J (2014) Genome-wide quantification of homoeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. Nucleic Acids Research, 42, e66.

Al-Shehbaz IA, O’Kane SL (2002) Taxonomy and phylogeny of Arabidopsis (Brassicaceae). The Arabidopsis Book / American Society of Plant Biologists, 1, e0001.

AUGUSTUS Development Team (2014) Incorporating RNAseq data into AUGUSTUS with TopHat. URL http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.Tophat.

Bechsgaard JS, Castric V, Charlesworth D, Vekemans X, Schierup MH (2008) Repeated adaptive introgression at a gene under multiallelic balancing selection. Proceedings of the National Academy of Sciences of the United States of America, 105, 1052–1056.

Beprimo MA, Banks E, Poplin R et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43, 491–498.

Dobin A, Davis CA, Schlesinger F et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15–21.

Dolezel J, Sgorbati S, Lucretti S (1992) Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. Physiologia Plantarum, 85, 625–631.

Dräger DB, Desbrosses-Fonrouge A-G, Krach C et al. (2004) Two genes encoding Arabidopsis halleri MTP1 metal transport proteins co-segregate with zinc tolerance and account for high MTP1 transcript levels. The Plant Journal, 39, 425–439.

Durand E, Meheust R, Soucaze M et al. (2014) Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. Science, 346, 1200–1205.

Earl D, Bradnam K, St. John J et al. (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Research, 21, 2224–2241.

Fischer MC, Relistab C, Tedder A et al. (2013) Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis from the Alps. Molecular Ecology, 22, 5994–5997.

Goubet PM, Bergès H, Bélec A et al. (2012) Contrastet patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in Arabidopsis. PLoS Genetics, 8, e1002495.

Guindon S, Dufayard J-F, Lefort V et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology, 59, 307–321.

Hanikenne M, Talke IN, Haydon MJ et al. (2008) Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. Nature, 453, 391–395.

Hanikenne M, Kroymann J, Trampczynska A et al. (2013) Hard selective sweep and ectopic gene conversion in a gene cluster allowing environmental adaptation. PLoS Genetics, 9, e1003707.

Hoffmann MH (2005) Evolution of the realized climatic niche in the genus: Arabidopsis (Brassicaceae). Evolution, 59, 1425–1436.

Hu TT, Pattyn F, Bakker EG et al. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nature Genetics, 43, 476–481.

Hunter B, Bombly K (2010) Progress and promise in using Arabidopsis to study adaptation, divergence, and speciation. The Arabidopsis Book, 8, e0138.

Johnston JS, Pepper AE, Hall AE et al. (2005) Evolution of genome size in Brassicaceae. Annals of Botany, 95, 229–235.

Kawagoe T, Shimizu KK, Kakutani T, Kudoh H (2011) Coexistence of tri-rome variation in a natural plant population: a combined study using ecological and candidate gene approaches. PLoS ONE, 6, e22184.

Kazemi-Dinan A, Thomaschky S, Stein RJ, Kramer U, Müller C (2014) Zinc and cadmium hyperaccumulation act as deterrents towards specialist herbivores and impede the performance of a generalist herbivore. New Phytologist, 202, 628–639.

Kent WJ (2002) BLAT—the BLAST-Like alignment tool. Genome Research, 12, 656–664.

Kobata A (1968) Nihon Kazanshi-no Kenkyu (Study of History of Japanese Mines). Iwanami-shoten, Tokyo (in Japanese).

Koch MA, Hauold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). Molecular Biology and Evolution, 17, 1483–1498.

Kolnik M, Marhold K (2006) Distribution, chromosome numbers and nomenclature concept of Arabidopsis halleri (Brassicaceae) in the Carpathians. Biologia, 61, 41–50.

Kramer U (2010) Metal hyperaccumulation in plants. Annual Review of Plant Biology, 61, 517–534.

© 2016 The Authors. Molecular Ecology Resources Published by John Wiley & Sons Ltd.
Kubota H, Takenaka C (2003) Field note: Arabis genniferi is a hyperaccumulator of Cd and Zn. International Journal of Phytoremediation, 5, 197–201.

Kubota S, Iwasaki T, Hanada K et al. (2015) A genome scan for genes underlying microgeographic scale local adaptation in a wild Arabidopsis species. PLoS Genetics, 11, e1005361.

Kudoh H (2016) Molecular phylogeny in plants: in natura systems biology for the comprehensive understanding of seasonal responses under natural environments. New Phytologist, 210, 399–412.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods, 9, 357–359.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25, 1717–1720.

Liu S, Liu Y, Yang X et al. (2014) The Brassica oleracea genome reveals the asymmetrical evolution of polyplid genomes. Nature Communications, 5, 3030.

Lobréaux S, Manel S, Melodelima C (2014) Development of an Arabis alpina genomic contig sequence data set and application to single nucleotide polymorphisms discovery. Molecular Ecology Resources, 14, 411–418.

Mansai S, Innan H (2010) The power of the methods for detecting interlocus gene conversion. Genetics, 184, 517–527.

McKenna A, Hanna M, Banks E et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20, 1297–1303.

Miyake K (1897) Hebinonegoza to koushitsutono kankei (Relationship between Asthenium yokosono Fr. Et Sav. and, the) The Botanical Magazine, Tokyo, 11, 404–406.

Novikova PY, Hohmann N, Nizhynska V et al. (2016) Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nature Genetics, 48, 1077–1082.

O’Kane SL, Al-Shebbaz IA (1997) A synopsis of Arabidopsis (Brassicaceae). Novon, 7, 323–327.

Ossowski S, Schneeberger K, Lucas-Lledó JJ et al. (2010) The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science, 327, 92–94.

Paape T, Hatakeyama M, Shimizu-Inatsugi R et al. (2016) Conserved but attenuated parental gene expression in allopolyploids: constitutive zinc hyperaccumulation in the allotetraploid Arabidopsis kamchatka. Molecular Biology and Evolution, Advanced online publication. doi:10.1093/molbev/msw141.

Pavvels M, Frérot H, Bonnin I, Saumitou-Laprade P (2006) A broad-scale analysis of population differentiation for Zn tolerance in an emerging model species: Arabidopsis halleri. Molecular Ecology, 15, 4024–4048.

Sato Y, Kawagoe T, Sawada Y, Hirai M, Kudoh H (2014) Frequency polymorphisms discovery. Molecular Biology and Evolution, 31, 3930.

Shahzad Z, Costi F, Frérot H et al. (2010) The five AbMT1P1 zinc transporters undergo different evolutionary fates towards adaptive evolution to zinc tolerance in Arabidopsis halleri. PLoS Genetics, 6, e1000911.

Shimizu KK (2002) Ecology meets molecular genetics in Arabidopsis. Population Ecology, 44, 0221–0233.

Shimizu KK, Furugganan MD (2005) Evolutionary and ecological genomics of Arabidopsis. Plant Physiology, 138, 578–584.

Shimizu KK, Tsuchimatsu T (2015) Evolution of selving; recurrent patterns in molecular adaptation. Annual Review of Ecology Evolution and Systematics, 46, 593–622.

Shimizu KK, Fujii S, Marhold K, Watanabe K, Kudoh H (2005) Arabidopsis kamchatka (Fisch. ex DC.) K. Shimizu & Kudoh and A. kamchatka subsp. kawasakiuna (Makino) K. Shimizu & Kudoh, New Combinations. Acta Phytotaxonomica et Geobotanica, 56, 163–172.

Shimizu KK, Kudoh H, Kobayashi MJ (2011) Plant sexual reproduction during climate change: Gene function in nature studied by ecological and evolutionary systems biology. Annals of Botany, 108, 777–787.

Shimizu-Inatsugi R, Lihoiv J, Iwana H et al. (2009) The allopolyploid Arabidopsis kamchatka originated from multiple individuals of Arabidopsis lyrata and Arabidopsis halleri. Molecular Ecology, 18, 4024–4048.

Slotte T, Hazzouri KM, Ageron JA et al. (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nature Genetics, 45, 831–835.

Smii A, Houbly B, Green P (1996) RepeatMasker Open-3.0. URL http://www.repeatmasker.org.

Sommor D, Delcher A, Salzberg S, Pop M (2007) Minimus: a fast, light-weight genome assembler. BMC Bioinformatics, 8, 64.

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

Tada Silver Mine Historic Site Preservation Association (2007) Tada ginzan: Shinkyoku no kotsu: Ogunon demetsu no sato (Tada silver mine: Historic site and lore: Village of golden legend). Tada ginzan shiseki hozon kensho kai, Inagawa-cho (in Japanese). Japanese National Bibliography Number 21410864. URL http://www.tadaginzankenshoukai.com/.

Talke IN, Hanikenne M, Krämer U (2006) Zinc-dependent global transcriptional control, transcriptional deregulation, and higher gene copy number for genes in metallocoastis of the hyperaccumulator Arabidopsis halleri. Plant Physiology, 142, 148–167.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 408, 796–815.

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

Trappnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111.

Tsuchimatsu T, Suwabe K, Shimizu-Inatsugi R et al. (2010) Evolution of self-compatibility in Arabidopsis by a mutation in the male specificity gene. Nature, 464, 1342–1346.

Tsuchimatsu T, Kaiser P, Yew C-L, Bachelier JB, Shimizu KK (2012) Recent loss of self-incompatibility by degradation of the male component in allotetraploid Arabidopsis kamchatka. PLoS Genetics, 8, e1002838.

Van der Auwera GA, Carneiro MO, Hartl C et al. (2013) From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics, 43, 11.10.1–11.10.33.

Willems G, Dräger DB, Courbot M et al. (2007) The genetic basis of zinc tolerance in the mallophytite Arabidopsis halleri ssp. halleri (Brassicaceae): an analysis of quantitative trait loci. Genetics, 176, 659–674.

Wolf DE, Steets JA, Houlston GJ, Takebayashi N (2014) Genome size variation and evolution in allotetraploid Arabidopsis kamchatka and its parents, Arabidopsis lyrata and Arabidopsis halleri. AoB Plants, 6, pft025.
R.V.B., T.P., R.S.I., J.S. and K.K.S. designed research; R.S.I. and T.N. performed experiments; R.V.B., T.P., S.A. analysed the data with inputs from all others; R.V.B., T.P., R.S.I. and K.K.S. wrote the manuscript with inputs from all others.

Data accessibility
The raw Illumina sequences of the mate-pair libraries were deposited to the DNA Data Bank of Japan (DDBJ) under the project number PRJDB4382 and were given the accession nos DRX045069–DRX045074. Paired-end libraries were previously deposited to DDBJ by Akama et al. (2014) under accession nos DRX012199–DRX012201.

Assembled sequence was deposited to the European Nucleotide Archive (ENA) and was given accession no. ERZ270560. Functional annotation, lists of reciprocal best BLAST hits between *A. halleri* and individually *A. thaliana* and *A. lyrata*, and the list of transposable elements were deposited to DryAd (doi:10.5061/dryad.rn4hh).

Workflow documentation and all custom scripts used in the project have been deposited to GitLab (https://gitlab.com/rbrisk/ahalassembly).

Supporting Information
Additional Supporting Information may be found in the online version of this article:

Fig. S1 Assembly pipeline.

Fig. S2 Annotation pipeline.

Appendix S1 Chemical analysis of soil samples from Tada mine locality in Japan, the origin site of the sequenced individual.

Table S1 Alignment results for RNA-seq data.

Table S2 Gene annotation statistics.

Table S3 Reciprocal best BLAST hits.

Table S4 Synteny in *HMA4* region.

Table S5 Synteny in *MTP1* region.