An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions

Annabelle Haudry1,2,17, Adrian E Platts3,4,17, Emilio Vello3,4,17, Douglas R Hoen5, Mickael Leclercq3,4, Robert J Williamson1, Ewa Forczek5, Zoé Joly-Lopez5, Joshua G Steffen6, Khaled M Hazzouri1, Ken Dewar7, John R Stinchcombe1, Daniel J Schoen5, Xiaowu Wang8, Jeremy Schmutz9,10, Christopher D Town11, Patrick P Edger12, J Chris Pires12, Karen S Schumaker13, David E Jarvis13, Terezie Mandáková14, Martin A Lysak14, Erik van den Bergh15, M Eric Schranz15, Paul M Harrison5, Alan M Moses1, Thomas E Bureau5, Stephen I Wright1,16 & Mathieu Blanchette3,4

Despite the central importance of noncoding DNA to gene regulation and evolution, understanding of the extent of selection on plant noncoding DNA remains limited compared to that of other organisms. Here we report sequencing of genomes from three Brassicaceae species (Leavenworthia alabamica, Sisymbrium irio and Aethionema arabicum) and their joint analysis with six previously sequenced crucifer genomes. Conservation across orthologous bases suggests that at least 17% of the Arabidopsis thaliana genome is under selection, with nearly one-quarter of the sequence under selection lying outside of coding regions. Much of this sequence can be localized to approximately 90,000 conserved noncoding sequences (CNSs) that show evidence of transcriptional and post-transcriptional regulation. Population genomics analyses of two crucifer species, A. thaliana and Capsella grandiflora, confirm that most of the identified CNSs are evolving under medium to strong purifying selection. Overall, these CNSs highlight both similarities and several key differences between the regulatory DNA of plants and other species.

A central challenge in functional and evolutionary genomics has been to determine the parts of a genome that are under selective constraint. Whereas protein-coding regions are relatively straightforward to identify, other functional elements such as transcriptional and post-transcriptional regulatory regions may be short and lacking in clear sequence signatures that would allow them to be detected in a single genome. Comparative genomic analyses across a group of closely related species provide a powerful approach to identify functional noncoding regions1. Over evolutionary time, non-functional sequences are expected to diverge faster than sequences under selective constraint. Patterns of sequence conservation may therefore be used to detect the footprints of functional noncoding elements. It is now widely accepted that the most powerful approach to phylogenetic footprinting is one based on a large number of species that have substantial aggregate divergence yet remain sufficiently closely related that the loss or displacement of functional elements is rare2,3.

Comparative genomic studies have led to the identification of thousands of CNSs in, among others, vertebrates4–6, fruit flies7 and yeast8. These CNSs are thought to be involved in diverse regulatory functions, including transcription initiation and transcript processing (for example, splicing or mRNA localization), as well as being implicated in complex patterning, such as embryonic development9–13. Plant CNSs have previously been identified on a genome-wide scale on the basis of the comparison of few or distant genomes (for example, maize versus rice14–16, Brachypodium distachyon versus rice17, A. thaliana versus Brassica oleracea18,19 and sets of diverse angiosperms20). This approach limits either the specificity provided by large divergence times or the sensitivity provided by the comparison of more closely related species10,11,21. Comparisons of paralogous noncoding regions flanking duplicated genes have also provided key insights into functional noncoding elements16,22, but intraspecies duplicated CNSs may often experience relaxed selective constraints.

The Brassicaceae are an ideal family for the identification of CNSs owing to their relatively small genome sizes, robust phylogeny23 and wealth of genomic data. So far, the genomes of six crucifer species have been partially or completely sequenced, including those of (i) the model species Arabidopsis thaliana24,25, (ii) the Brassicaceae outgroup Capsella grandiflora26,27, (iii) the model grass Brachypodium distachyon28, (iv) Leavenworthia alabamica29,30, (v) Aethionema arabicum31,32, and (vi) Sisymbrium irio33,34. A central challenge in functional and evolutionary genomics has been to determine the parts of a genome that are under selective constraint. Whereas protein-coding regions are relatively straightforward to identify, other functional elements such as transcriptional and post-transcriptional regulatory regions may be short and lacking in clear sequence signatures that would allow them to be detected in a single genome. Comparative genomic analyses across a group of closely related species provide a powerful approach to identify functional noncoding regions. Over evolutionary time, non-functional sequences are expected to diverge faster than sequences under selective constraint. Patterns of sequence conservation may therefore be used to detect the footprints of functional noncoding elements. It is now widely accepted that the most powerful approach to phylogenetic footprinting is one based on a large number of species that have substantial aggregate divergence yet remain sufficiently closely related that the loss or displacement of functional elements is rare.

Comparative genomic studies have led to the identification of thousands of CNSs in, among others, vertebrates4–6, fruit flies7 and yeast8. These CNSs are thought to be involved in diverse regulatory functions, including transcription initiation and transcript processing (for example, splicing or mRNA localization), as well as being implicated in complex patterning, such as embryonic development9–13. Plant CNSs have previously been identified on a genome-wide scale on the basis of the comparison of few or distant genomes (for example, maize versus rice14–16, Brachypodium distachyon versus rice17, A. thaliana versus Brassica oleracea18,19 and sets of diverse angiosperms20). This approach limits either the specificity provided by large divergence times or the sensitivity provided by the comparison of more closely related species10,11,21. Comparisons of paralogous noncoding regions flanking duplicated genes have also provided key insights into functional noncoding elements16,22, but intraspecies duplicated CNSs may often experience relaxed selective constraints.

The Brassicaceae are an ideal family for the identification of CNSs owing to their relatively small genome sizes, robust phylogeny23 and wealth of genomic data. So far, the genomes of six crucifer species have been partially or completely sequenced, including those of (i) the model species Arabidopsis thaliana24,25, (ii) the Brassicaceae outgroup Capsella grandiflora26,27, (iii) the model grass Brachypodium distachyon28, (iv) Leavenworthia alabamica29,30, (v) Aethionema arabicum31,32, and (vi) Sisymbrium irio33,34.
plant *A. thaliana*; (ii) *Arabidopsis lyrata*, a congener of *A. thaliana* with a more ancestral karyotype and genome size; (iii) *Capsella rubella*, which falls in the sister group to the genus *Arabidopsis*; (iv) *Brassica rapa* (Chinese cabbage), one of the several closely related *Brassica* crop species in the tribe Brassicaceae that share a recent genome triplication event (*Br-α*); (v) *Eutrema salsugineum* (previously *Thellungiella halophila*) of the tribe Eutremeae, an extremophile adapted to saline habitats; and (vi) *Schrenkella parvula* (previously *Thellungiella parvula*), another extremophile of uncertain tribal placement.

To complement the set of previously published Brassicaceae genomes, we have sequenced the genomes of three additional species, to provide a broad diversity of lineages within the family. We took advantage of these nine closely related genome sequences to identify and characterize over 90,000 CNSs. The extent of selection acting on them was determined using a combination of comparative and population genomics data. Several lines of computational and experimental evidence point to a large proportion of CNSs having a role in transcriptional or post-transcriptional regulation. A full catalog of CNSs and their associated annotations in *A. thaliana* and *A. lyrata* are available via a genome browser.

## RESULTS

### Genome sequencing, assembly and annotation

To supplement six publicly available crucifer genomes, we sequenced and partially assembled the genomes of three further crucifers (Table 1 and Online Methods): (i) *Leavenworthia alabamica* (lineage 1 in the tribe Camelinae), a model plant species with recently lost self-incompatibility in some populations; (ii) *Sisymbrium trifoliatum* (lineage 2 in the tribe Sisymbrieae), a self-compatible annual closely related to the *Brassica* genus but lacking the derived whole-genome triplication; and (iii) *Aethionema arabicum* (tribe Aethionemeae), a self-compatible, early branching sister group to the remainder of the core Brassicaceae. All species share the ancient whole-genome duplication that occurred at the base of the family (At-α; ref. 33).

Assemblies of these three genomes included orthologs for the majority of *A. thaliana* protein-coding genes (68–83%, on a par with those found in more completely assembled genomes) and 98–99% of the ultraconserved core eukaryotic genes (Table 1), suggesting that the coverage of non-repetitive DNA was high. Furthermore, the scaffold size (N50 of 70–135 kb) was suitable for the identification of orthologous regions through synteny.

### Supplementary Table 1

The number of annotated genes varied between species from 23,167 genes in *A. arabicum* to 41,174 in *B. rapa*. This variation was expected given the rediploidization process following the At-α duplication and several whole-genome amplifications after At-α (for example, the *Brassica* triplication). TEs comprised the majority of the variation in genome size observed across the crucifers, varying in content from 13–15% in the smallest genomes (*A. thaliana* and *S. parvula*) to ~50% in *E. salsugineum*.

### Multiple-genome comparison

Nine-way genome alignments were generated for the crucifer genomes, using either *A. lyrata* or *A. thaliana* as the reference (Online Methods). A region of a non-reference genome was only allowed to align to a single region of the reference genome, although many regions from non-reference genomes were allowed to map to the same reference genome region (Fig. 1a). This ensured that lonesome paralogous genes resulting from the At-α duplication did not contaminate the alignments, while allowing more recent duplication events, such as the whole-genome triplication in *B. rapa*, to be represented. Local pairwise alignment blocks were filtered to only retain those that belonged to long sets of collinear blocks (chains). The vast majority of the *A. lyrata* genome belonged to a single such chain in each of the other species (Fig. 1a and Supplementary Table 2), leaving little doubt about correct orthology. This pattern of alignment also provided strong support for the notion that most of the gene loss after At-α duplication was substantially completed before the divergence of species within the Brassicaceae.

However, two species showed a strong departure from this essentially diploid organization. As expected, most *A. lyrata* regions were spanned by three long alignment chains in the *B. rapa* genome, owing to the established Br-α triplication event (Fig. 1a,b). Notably, the *L. alabamica* genome showed the same clear signs of an independent whole-genome triplication, referred to hereafter as La-α. Comparative chromosome painting of the *L. alabamica* genome (Fig. 1c–f) independently supported the establishment of hexaploidy after At-α, with the patterning of *A. thaliana* BAC probes supporting the retention of some chromosomal regions in two copies and others in three copies.

A phylogenetic tree (Fig. 2) derived from 1,048,889 fourfold-degenerate sites and using *Carica papaya* as an outgroup was consistent with previously published phylogenies. The total branch length of the tree (~1.5 substitutions per site) was similar to that for a set of nine diverse mammals (~1.3 substitutions per site) used in the identification of conserved noncoding regions. The divergence between *L. alabamica* paralogs (~0.3 substitutions per site) was similar...
to that between the paralogs of B. rapa (~0.35 substitutions per site), formed ~24 (18–28) million years ago\(^{41,42}\) by the Br-\(\alpha\) triplication. Although this slightly larger number of substitutions per site may not imply a more recent event, because of the possibility of variation in neutral substitution rates\(^{43}\), it likely indicates a similar era for these independent hexaploidization events.

Selection on noncoding sites in the crucifer genomes

Comparisons of multiple closely related genomes allows the fraction of the genome that is constrained by selection to be estimated\(^{44}\). PhyloP\(^{45}\) was used to measure interspecies conservation of each nucleotide of the A. thaliana and A. lyrata genomes, independent of the flanking nucleotides. Because of the insufficient level of divergence between the nine species considered, these scores could not unambiguously distinguish individual constrained sites from neutral ones. However, the proportion of sites under selection in the whole genome or in any given subset of sites could be estimated by comparing the distribution of PhyloP scores across the genome to that at fourfold-degenerate sites, which are largely unconstrained\(^{46}\) (Online Methods).

At least 17.7% of the assembled A. thaliana genome sequence (21.1 Mb) seemed to be evolving under constraint (Fig. 3a), with close to a quarter of this sequence (4.5 Mb) located outside of protein-coding regions. In the larger TE-rich A. lyrata genome, very slightly more sites seemed to be under selection (22.2 Mb), corresponding to a much smaller fraction of the genome (11.3%). Consequently, the major cause of the difference in genome size between A. lyrata and A. thaliana is probably not the loss of functional sites but rather the loss of effectively unconstrained regions in A. thaliana, likely coupled with higher recent TE activity in A. lyrata\(^{25}\).

In both A. thaliana and A. lyrata, the constrained noncoding sites were divided roughly evenly between transcript-associated sites (introns and UTRs) and intergenic sites (Fig. 3b). The proportion of sites under selection was particularly high in 5’ and 3’ UTRs (17% and 13%, respectively) as well as in intronic regions flanking exons (15% within 30 bp of splice sites) but was much lower in the center of introns (Fig. 3b). Contrary to what is observed in mammals\(^{47}\) and Drosophila melanogaster\(^{48}\), intronic bases located within 500 bp of the transcription start site (TSS) did not seem to be under significantly stronger selective pressure than other intronic bases.
Outside protein-coding transcripts, the proportion of sites under selection decreased with the distance from the TSS. Nonetheless, as observed in Drosophila, more than half of intergenic sites showing signs of selection were >1 kb away from the closest annotated TSS. Notably, and in contrast to findings in mammals, evidence of constraint was lowest immediately downstream of 3′ UTRs, suggesting that regulatory elements are rarely located in those regions.

At least 90,000 conserved noncoding sequences

The PhyloP analyses yielded estimates of the number of individual sites under selection in specific portions of the A. thaliana genome, but they did not pinpoint the location of those sites. Constrained sites can only be reliably identified if they cluster with other such sites, forming CNSs. CNSs are genomic regions that show a reduced mutation rate over contiguous or near-contiguous sets of non-coding bases. A set of 92,421 (90,104) CNSs was identified in the A. lyrata (A. thaliana) genome (Supplementary Fig. 2) using a pipeline based on PhastCons. Evidence of selection was also clearly provided by the relative rarity of insertions and deletions within CNSs (Supplementary Fig. 3). Previously published CNSs obtained from pairwise genome comparisons were typically identified in this screen, but five- to tenfold more conserved regions were also identified, owing to the sensitivity afforded by the use of multiple genomes (Supplementary Note).

Each CNS was annotated in accordance with its position relative to genes in the A. lyrata genome. CNSs without evidence of expression in whole A. thaliana or A. lyrata plants (Online Methods) were classified as proximal upstream or downstream (<500 bp upstream of the TSS or downstream of a TES, respectively), distal (>500 bp away from any gene’s TSS or TES) or ambiguous. Genic CNSs were subdivided on the basis of their location in 5′ UTRs, introns and 3′ UTRs. We also identified 820 CNSs with substantial evidence of short RNA expression in A. lyrata (Online Methods) and labeled these as putative small noncoding RNA CNSs (smRNA CNSs). Conserved regions with evidence of small RNA expression in other plants but not in A. lyrata were set apart as a class of potential smRNA CNSs (Supplementary Table 3).

CNS density in different types of genomic regions closely followed that of the inferred sites under selection (Supplementary Fig. 4). Crucifer CNSs were typically short (median length of 36 bp and slightly shorter in introns and UTRs) and had a GC content similar to that of the noncoding portion of the genome (25–40%; Supplementary Fig. 5). smRNA CNSs, most of which corresponded with known noncoding RNA genes, formed a relatively distinct group, showing higher conservation and GC content and a markedly bimodal size distribution, mostly caused by large numbers of microRNA (miRNA) and tRNA genes.

Evidence of purifying selection on CNSs at the population level

To independently assess evidence for purifying selection acting on CNSs, we analyzed the distribution of sequence diversity in CNSs within the populations of two Brassicaceae species: a recently sequenced set of 80 A. thaliana genomes and a set of 13 outbred individuals (26 haplotypes) of C. grandiflora (S.I.W., unpublished data), a close relative of A. thaliana. Evidence for recent purifying selection acting on CNSs was found in the minor allele frequency (MAF) spectra of both populations. Both species showed an excess of rare variants in the CNS bases (Fig. 4) and reduced levels of population diversity compared to fourfold-degenerate sites, as measured by nucleotide diversity $\pi$ (ref. 52) and Watterson’s estimator $\Theta_W$ (ref. 53) (Supplementary Fig. 6). However, purifying selection on CNSs was not generally as strong as in the highly constrained zero-fold degenerate sites. Similar observations were made for deletion polymorphisms at the population level (Supplementary Fig. 7). Because analyses of MAF spectra only examined segregating variation and ignored the level of polymorphism, these results provide independent validation of the action of purifying selection and limit the possibility of low divergence in CNSs arising from mutation cold spots.

Distribution of CNSs in Brassicaceae and other plants

The fraction of A. lyrata CNSs for which homologs could be detected in other plant genomes was determined on the basis of sequence similarity (Fig. 5). Whereas most Brassicaceae genomes contained homologs for more than 75% of these CNSs, the early branching A. arabicum genome had homologs for only 38%. The two other Brassicaceae with a reduced number of identifiable homologs were those that had undergone whole-genome triplication events, B. rapa and L. alabamica, suggesting increased rates of CNS loss after each triplication. The proportion of A. lyrata CNSs with detectable homologs outside Brassicaceae was relatively low, ranging from 0.8% in the phylogenetically distant Oryza sativa to 3.4% in the more recent neighbor C. papaya. CNSs that seemed to predate Brassicaceae divergence were 75-fold enriched for small noncoding RNAs.

Loss of genes and CNSs after whole-genome triplications

The presence of two whole-genome triplications (Br-α and La-α) offered a further opportunity to study the fate of genes and CNSs...
Many CNSs are transcriptional regulatory elements
To test for a function for intergenic CNSs in transcriptional regulation, the regions bound by 13 transcription factors in A. thaliana showed the weakest retention. Notably, 136 of the 428 CNSs kept in 3 copies each in A. thaliana individuals with 26 haplotypes (right). All types of CNSs showed lower population diversity and higher MAF compared to fourfold-degenerate (4D) and intronic sites, confirming that a substantial fraction of CNSs are under selective pressure. Whereas most types of CNSs have similar diversity levels, intronic CNSs seem to be under the weakest selective pressure, and smRNA CNSs seem to be under the strongest selective pressure. ODs, zero-fold degenerate.

Figure 5 The majority of A. lyrata CNSs are shared with most other Brassicaceae, but few are conserved outside that clade, with the exception of those corresponding to smRNAs. A. lyrata CNSs were first symmetrically extended to at least 120 bp, except when this reached into coding exons. These sequences were then aligned against each genome using BLAST. The number of CNSs with at least one hit with an E value below 0.0001 is shown. Whereas smRNA CNSs constitute only 1.1% of eligible CNSs, they account for more than 18% of CNSs mapped to C. papaya, and this fraction increases to 66% for the most distant species considered, O. sativa.
CNSs are enriched for specific sequence motifs

By highlighting genomic regions of likely regulatory function, CNSs facilitated the identification of regulatory motifs for transcription factors or RNA-binding proteins. Each type of CNS was found to be enriched for particular motifs, on the basis of a z-score calculation that compared the number of occurrences of motifs 6–8 nt in length in CNSs to permuted versions of the same motifs7 (Fig. 6a and Supplementary Table 6). Many of the motifs identified were associated with the binding preferences of ubiquitous transcription factors (G-box, E-box, W-box, EIRE, GT-1 and TATA-binding elements) and were enriched in all types of intergenic CNSs, suggesting that these CNSs may have similar regulatory roles. For example, the abscisic acid–linked G-box (CAGCTG) and the calcium signaling–sensitive E-box (CAGTGT) motifs, when grouped together under the CAYRTGTC motif (with Y representing C or T and R representing A or G), were four- to sevenfold enriched only in intergenic CNSs. The motif enrichment analysis was repeated on subsets of CNSs associated with genes with similar functions, as determined on the basis of GO-Slim4, identifying a large number of known and new putative regulatory motifs (Supplementary Table 8 and Supplementary Note).

Likely reflecting the binding contexts relative to the TSS needed for a functional site, diversity estimated in the population of 80 A. thaliana lines was 7-fold lower at G-box sequences in CNSs compared to G-box sequences located outside CNSs. SNP density within G-box sequences in CNSs showed constraint in A. thaliana and C. grandiflora populations (Fig. 6b) and in interspecies conservation profiles (Fig. 6c.d). Notably, positions immediately flanking most G-box sequences showed a reduced level of conservation compared to overall CNSs, possibly highlighting spatial constraints on the placement of other regulatory elements.

Several motifs of unknown function were strongly enriched in 5′ UTRs but not elsewhere, with some exhibiting strong strand bias, hinting at a role in post-transcriptional regulation. Others were found in both proximal and 5′ UTR CNSs (TATA box and GA track), suggesting a role in transcription initiation.

DISCUSSION

Although the annotation of protein-coding and some small noncoding RNA genes in A. thaliana has become increasingly complete, until now, no high-resolution map of regulatory regions existed. Here we report the first genome-wide high-resolution atlas of noncoding regions under selection in crucifers. Because the detection of CNSs is based on the comparison of a large number of closely related species, the sensitivity of this map is higher than that in previous studies based on pairs or sets of more distant species22,65 or on intragenomic comparisons22, resulting in an eight to tenfold increase in the number of constrained regions identified.

Our analysis shows that at least 5% of noncoding sites in A. thaliana seem to have been evolving under some form of purifying selection in Brassicaceae. Our estimate of the proportion of the A. thaliana genome that is constrained, combined with experimental estimates of substitution rate66 (7 × 10−9 substitution per site per generation), yields a
lower bound on the deleterious substitution rate of at least 0.15 bases per individual per generation, which is comparable to the conservative estimate of 0.1 from a large mutation accumulation experiment\(^67\).

The number of genes analyzed in this study and their divergence relative to each other are comparable to those used in similar comparative genomics studies of mammals\(^4,68\) and fruit flies\(^7\). It is consequently possible to contrast the properties of their CNSs. The regulatory complexity of a genome can be approximated by the number of bases in CNSs, normalized by gene number. In A. thaliana and A. lyrata, this regulatory complexity amounts to 160 bp per gene, slightly more than in yeast (≈110 bp per gene) but substantially less than in animals (worms, ≈600 bp per gene; fruit flies, ≈2,500 bp per gene; mammals, ≈5,000 bp per gene)\(^5\). This finding suggests that noncoding regulatory mechanisms in plants are intermediate in complexity between those of yeast and worms and is consistent with the hypothesis that plants obtain regulatory complexity via gene or entire-genome duplication rather than from noncoding regulation\(^69\). Alternatively, this low CNS-to-gene ratio might be caused by a high rate of turnover or streamlining of regulatory regions, possibly linked to frequent whole-genome duplications, resulting in many such CNSs going undetected by our approach.

The most highly conserved noncoding sequences identified were on average ≈70 bp in length with ≈0.15 substitutions per site, which is much lower than the 100% conservation over 200 bp used to define mammalian ultraconserved elements\(^8\). Nonetheless, the set of 2,012 most highly conserved noncoding sequences (with, at most, 0.5 substitutions per site over at least 50 bp; Supplementary Table 9) stands out from the rest of the CNSs in a manner that is reminiscent of ultraconserved elements, clustering around genes involved in the regulation of embryonic and post-embryonic development. Because the most recent common ancestor of plants and vertebrates was unicellular and, hence, likely lacked developmental patterning, this finding suggests that similar patterns of CNS control may have evolved independently in the two kingdoms—a noteworthy example of convergent evolution of genomic organization.

Crucifer CNSs differ from animal CNSs in the way that they are associated with their putative target genes: distal CNSs are comparatively less frequent in Brassicaceae than in animals, and first introns are generally depleted of CNSs, unlike in mammals where CNSs distribute roughly symmetrically around TSSs, and first introns are enriched for regulatory elements\(^70\). These differences in CNS distribution may reflect some of the differences in intron-exon structures between plants and animals. In vertebrates, the first intron may be relatively extended, whereas, in plants, it tends to be shorter with alternative splicing more frequently driving intron retention\(^71\), thereby potentially limiting the first intron as a site for regulatory CNSs.

Together with findings from other ongoing investigations, the resources introduced here will help to establish the properties of the constrained portion of the noncoding genome of the crucifers, in much the same way as similar projects have in other eukaryotic species. Combined with the application of systematic genome-wide experimental assays\(^72\), this atlas of noncoding selection in the Brassicaceae will further open the door to the detailed characterization of the cis regulome of these species.

### URLs

RepeatMasker, [http://www.repeatmasker.org/](http://www.repeatmasker.org/); RepeatModeler, [http://www.repeatmasker.org/RepeatModeler.html](http://www.repeatmasker.org/RepeatModeler.html); pybloomfaster, [https://github.com/brentp/pybloomfaster](https://github.com/brentp/pybloomfaster); 1001 Genomes Project, [http://1001genomes.org/data/MPJ/MPICao2010/releases/](http://1001genomes.org/data/MPJ/MPICao2010/releases/).

### METHODS

Methods and any associated references are available in the online version of the paper.

### Accession codes

A. arabicum genome, PRJNA202984; S. irio genome, PRJNA202979; L. alabamica genome, PRJNA202983. All sequences, genome annotations, pairwise and multiple alignments, conservation scores and CNSs are available for visualization and download on a local installation of the UCSC Genome Browser at [http://mustang.biol.mcgill.ca:8885](http://mustang.biol.mcgill.ca:8885).

### ACKNOWLEDGMENTS

We would like to thank the US Department of Energy Joint Genome Institute (for the C. rubella genome sequence, produced under a Community Sequencing Program (CSP) proposal submitted by D. Weigel and colleagues, and the E. salugineum genome sequence, produced under a CSP proposal submitted by K. Schumaker, R. Wing and T. Mitchell-Olds) and R. Clark (for A. lyrata mRNA sequencing). We also thank S.-C. Chantha for assistance with genome sequencing in L. alabamica. S. Ioly for suggestions on the genomic DNA isolation protocol and D. Scofield for helpful discussions about intron-exon structure. We thank M. Freeling, D. Weigel, E. Harmsen and I. Lacroix for comments on the manuscript. This project was funded by a Genome Canada/Genome Québec grant to T.E.B., S.I.W., M.B., J.S., A.M.M., D.J.S. and P.M.H. In addition, T.M. and M.A.L. were supported by the European Regional Development Fund (CZ.1.05/1.1.00/02.0068) and by the Czech Science Foundation (excellence cluster PSI01/12/GO90). J.G.S. was supported by National Science Foundation (NSF) award 0929262. M.E.S. and E.V.d.B. were supported by a Vidi grant from the Netherlands Organisation for Scientific Research.

### AUTHOR CONTRIBUTIONS

The study was conceived by M.B., A.M.M., T.E.B., D.J.S., P.M.H. and J.R.S. Computational experiments were designed by A.H., A.E.P., A.M.M., S.I.W., T.E.B. and M.B. E.F., Z.J.-L., J.C.P., M.E.S., D.J.S. and T.E.B. obtained material for genome sequencing for L. alabamica, S. irio and A. arabicum. A.E.P., K.D. and T.E.B. sequenced the DNA, and A.E.P. assembled the genomes, using additional data provided by C.D.T., P.P.E., M.E.S., E.V.d.B. and J.C.P. Additional RNA sequencing data were obtained from J.G.S., B. rupi genome sequence data were provided by S.I.W. and E. salugineum genome data were provided by J.S., D.E.I. and K.S.S. T.M. and M.A.L. performed the multicolor FISH study on L. alabamica. P.M.H. and A.E.P. performed the gene annotation, D.R.H. and T.E.B. annotated TEs, and M.I. identified structural RNAs. Multiple-genome alignments and identification and analysis of CNSs were performed by A.E.P., A.H., E.V. and M.B. Population genetics analyses were performed by A.H., A.E.P., R.J.W., K.M.H., A.M.M., A.E.P. and S.I.W. The manuscript was written primarily by A.H., A.E.P., S.I.W. and M.B., with input from all coauthors.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at [http://www.nature.com/reprints/index.html](http://www.nature.com/reprints/index.html).
ONLINE METHODS

Sequencing and assembly. The genomes of *L. alabamica*, *A. arabicum* and *S. iringo* were assembled from Illumina paired-end reads (2 × 105 nt with a nominal 64-nt gap on the Genome Analyzer IIX platform; 55–110× coverage) and mate-pair reads (2 × 105 nt with 5-kb and 10-kb insert sizes on the Genome Analyzer IIX and HiSeq 2000 platforms). Libraries and reads were generated in accordance with Illumina protocols, with special attention paid to gentle shearing of mate-pair circular DNA. Reads were trimmed for quality (3’ trimming starting at the first position with Q < 32) and assembled with the Ray assembler79 using a K-mer size of 31 to 41 (optimized per genome). Mate-pair reads were filtered for duplicates using a Bloom filter (pyblowlmagefast; see URLs) and for potential false mates on the edges of contigs, resulting in approximately 12× (5×) coverage for 5-kb (10-kb) inserts for each species and see URLs) and for potential false mates on the edges of contigs, resulting in approximately 12× (5×) coverage for 5-kb (10-kb) inserts for each species and with gap filling enabled.

The genomes of *B. rapa* (ssp. Chifu-401-42; pre-publication data from Wang et al.27), *S. iringo* (PRJNA364667), *A. lyrata* (ssp. lyrata; PRJNA411337), *A. thaliana* (Col-0; TAIR9/TAIR10, PRJNA10719), *E. salsugineum* (PRJNA80723) and *C. rubella* (PRJNA13878) were obtained either directly from their assemblers or from data published at the time of the release of the genome.

Genome completeness was assessed relative to the total and expected assembly length, the count of *A. thaliana* orthologs and the percentage of complete, highly conserved eukaryotic genes (Table 1). The *A. arabicum* genome was further validated against a set of physical mapping data (Keygene) that showed near-perfect concordance between assembled scaffolds and BAC contigs. The *S. iringo* genome was compared against BAC and BAC-end data in GenBank. Although sequences were mostly concordant, the BAC data was from a tetraploid species and was consequently not expected to be completely similar. The *L. alabamica* genome was examined in several TE-rich extended loci (for example, the S locus) that had been assembled using long-read fosmid sequences80 and showed near-perfect concordance.

CCP analysis in *L. alabamica*. Preparation of chromosome spreads from young anthers and BAC painting probes, as well as multicolor FISH, followed the protocols described by Mandáková et al.81. In total, 237 chromosome-specific BACs from *A. thaliana* (~23 Mb) were used as painting probes. The following *A. thaliana* BAC contigs were applied to identify 8 ancestral genomic blocks82 on *L. alabamica* chromosomes: block A (31 BACs: T25K16–T29MB; 6.7 Mb), block O (24 BACs: F6N15–T1J1; 2.5 Mb), block P (13 BACs: T3H13–T22B4; 1.3 Mb), block Q (32 BACs: T20O7–T8M17; 2.6 Mb), block R (33 BACs: F7J9–T6G21; 7.4 Mb), block V (22 BACs: MB2D–K23F3; 2.4 Mb), block W (36 BACs: K21P3–MMN10; 4.3 Mb) and block X (26 BACs: MUP24–K919; 2.5 Mb).

Genome annotation. All nine genomes were annotated for genomic regions using Maker82 in conjunction with FGENESH and FGENESH+ (ref. 83), Augustus84, SNAP85 and BLAT86 for transcript mapping. Repetitive regions and TEs were annotated with RepeatMasker using repeat models determined on a species basis obtained from RepeatModeler (Supplementary Table 2).

In addition to annotation of the *A. lyrata* and *A. thaliana* genomes, we combined sequenced mRNA from whole *A. lyrata* plants (Illumina, strand-specific RNA sequencing, 2 × 80 nt; R. Clark, personal communication; PRJNA207497), archived mRNA87 (NCBI Sequence Read Archive (SRA) SRRA019209, SRR019183 and SRR064165) and small RNA sequence data from both SOLID and Illumina platforms (SRR044002, SRR072809, SRR034856 and SRR051926). Reads were aligned both with Novoalign (Novocraft) with high-alignment stringency and SpliceMap88 for exon-spanning reads. Expression tracks were then lifted over between the two reference genomes to aid annotation.

Whole-genome alignments. Each genome was soft masked and aligned to *A. lyrata* and *E. salsugineum* (primary and secondary reference genomes, respectively) using last89 and chaining90 and assembling collinear alignment blocks separated by gaps of <100 kb were then performed. We filtered for orthologous chains by retaining chains in decreasing order of score that did not substantially overlap previously selected chains in the non-reference genome (chains could overlap in the reference genome). This filter effectively separates orthologs from α paralogs, while allowing more recent whole-genome duplications to be properly represented. In the case of the two genomes with whole-genome triplications (*B. rapa* and *L. alabamica*), genomic regions were subdivided into three groups of chains, where each group contained non-overlapping chaining. We obtained a 13-way multiple alignment using the MultiX92 progressive alignment program, following phylogenetic order, using *A. lyrata* as the reference for lineage 1 and *E. salsugineum* as the reference for lineage 2. For the purpose of measuring sequence conservation of a region, the most conserved of the paralogs in *B. rapa* and *L. alabamica* were retained.

Determination of the fraction of sites under selection. Our approach to estimate the fraction of sites under selection is based on that of Watterson et al.93, adapted to use site-specific conservation scores. PhyloP45 was used to measure position-specific conservation levels on the basis of nine-way alignment, using a model of freely evolving sites obtained from fourfold-degenerate sites of the same alignment and on the JGI gene annotation of *A. lyrata*. The fraction of sites under selection in a given set of sites R was obtained as follows. Let S be a subset of the nine species considered and let Rb be a subset of sites from R that has nucleotides in S and gaps in the other species. PhyloP scores were discretized in 1,000 bins. For each S, let f(S) be the distribution of discretized PhyloP scores obtained from fourfold-degenerate sites by replacing nucleotides in species outside S by gaps and calculating the distribution of PhyloP scores. Let f(S)x be the observed distribution of PhyloP scores in Rb. We express f(S)x as a mixture of f(S) and f(S)x, the unknown distribution of scores for sites under selection in Rb. Specifically, we estimated ω* so that f(S)x = ωf(S) + (1 − ω)f(S)x. Let F(S)x and F(S)x be the cumulative distributions of f(S)x and f(S)x, respectively. Let χ be the value for which F(S)x/F(S)x is maximized (excluding values of x for which either of the two cumulative distributions has a value less than 0.1), and let fmax = F(S)x/ F(S)x. We obtain ω* = Σx<χ* ω f(S)x / fmax. Finally, the fraction under selection for region R is determined as Σx ω f(S|x)/ fmax. Note that, because not all fourfold-degenerate sites are truly unconsstrained, our estimate of the fraction under selection in R is a lower bound.

CNS identification. CNSs were identified as regions located beyond annotated coding sequences in the *A. lyrata* reference genome that showed high PhastCons score (>0.82) over an extended length (>7 nt) and did not include a region of more than 12 nt with low PhastCons score (<0.55). To facilitate the comparison of incompletely sequenced genomes, insertions, deletions and missing orthologous sequences were not penalized, and CNSs were not required to be present in all nine species. The parameters were refined relative to an 800,000-base sequence generated from concatenated fourfold-degenerate sites within which a CNS FDR of <1% was required. Independent verifications based on evolutionary signatures of coding sequences using RNAcode93, the absence of splice sites88 and a uniform density of stop codons suggested that very few CNSs correspond with unannotated protein-coding exons.

Candidate CNSs were assigned a location category, and only those in the small noncoding and UTR classes were allowed to overlap expressed regions. CNSs were rejected that formed extended sequences of coding sequences (due to the PhastCons smoothing algorithm) and that overlapped other evidence of a potential coding role. Because UTR annotation, particularly the transition between coding sequences and UTRs, is not error free, we expect a slightly higher false positive rate for the UTR CNSs.

Motif enrichment. The significance of the enrichment of motifs of 5 to 11 nt in length in CNSs was determined by a z score representing a comparison of the frequency of a motif’s occurrence in all CNSs to the distribution of the occurrence of all permutations of the motif sequence in CNSs. This approach was selected to account for substantial base bias at single- and multi-base levels between CNSs and surrounding promoter regions. Enriched motifs were then clustered with those with a minimal edit distance and combined into IUPAC and PWM representations. Motif characterization was determined using the PLACE, TAIR and JASPAR databases. Enrichment of motifs in upstream, intronic, UTR and downstream CNSs relative to ontology groups was determined relative to the GO-Slim ontology annotation of the immediately proximate gene.
Population genomics analyses. Alignments for the genomes of 80 Eurasian *A. thaliana* plants Against the TAIR9/TAIR10 reference were obtained from the 1001 Genomes Project (see URLs). For measures of diversity over specific regions, those locations with base calls in all 80 samples were used, whereas, for more general comparisons with comparative genomics data, calls in at least 40 samples were required. Diversity estimates from the genomes of a population of 13 Greek *C. grandiflora* plants (26 haplotypes) were generated by aligning Illumina paired-end data to the genome of its close relative *C. rubella* using the STAMPY-GATK pipeline. Again, 26 base calls were required for diversity estimates, and regions with extremes of sequence depth or low quality were excluded. Pipelines for population genomics analyses were developed using Perl and Python languages and Bio++ libraries.

78. Boisvert, S., Lavoie, F., & Corbeil, J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol.* 17, 1519–1533 (2010).
79. Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20, 265–272 (2010).
80. Charantha, S.C., Herman, A.C., Platt, A.E., Vekemans, X., & Schoen, D.J. Secondary evolution of a self-incompatibility locus in the brassicaceae genus leavenworthia. *PLoS Biol.* 11, e1001560 (2013).
81. Lysak, M.A. & Mandáková, T. Analysis of plant meiotic chromosomes by chromosome painting. *Methods Mol. Biol.* 990, 13–24 (2013).
82. Cantarel, B.L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–196 (2008).
83. Salamov, A.A. & Solovyev, V.V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* 10, 516–522 (2000).
84. Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* 24, 637–644 (2008).
85. Korfi, I. Gene finding in novel genomes. *BMC Bioinformatics* 5, 59 (2004).
86. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664 (2002).
87. Gan, X. et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477, 419–423 (2011).
88. Au, K.F., Jiang, H., Lin, L., Xing, Y., & Wong, W.H. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.* 38, 4570–4578 (2010).
89. Harris, R.S. Improved Pairwise Alignment of Genomic DNA. PhD thesis, Penn. State Univ. (2007).
90. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 100, 11484–11489 (2003).
91. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715 (2004).
92. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562 (2002).
93. Washietl, S. et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* 17, 578–594 (2011).
94. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Res.* 21, 936–939 (2011).
95. Dutheil, J. et al. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* 7, 188 (2006).