nature genetics
OPEN

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

Annabelle Haudry[1,2,17], Adrian E Platts[3,4,17], Emilio Vello[3,4], Douglas R Hoen[5], Mickael Leclercq[3,4], Robert J Williamson[1], Ewa Forczek[5], Zoé Joly-Lopez[5], Joshua G Steffen[6], Khaled M Hazzouri[1], Ken Dewar[7], John R Stinchcombe[1], Daniel J Schoen[5], Xiaowu Wang[8], Jeremy Schmutz[9,10], Christopher D Town[11], Patrick P Edger[12], J Chris Pires[12], Karen S Schumaker[13], David E Jarvis[13], Terezie Mandáková[14], Martin A Lysak[14], Erik van den Bergh[15], M Eric Schranz[15], Paul M Harrison[5], Alan M Moses[1], Thomas E Bureau[5], Stephen I Wright[1,16] & Mathieu Blanchette[3,4]

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 regions[1]. 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[2,3].

Comparative genomic studies have led to the identification of thousands of CNSs in, among others, vertebrates[4–6], fruit flies[7] and yeast[8]. 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 development[9–13]. 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 rice[14–16], *Brachypodium distachyon* versus rice[17], *A. thaliana* versus *Brassica oleracea*[18,19] and sets of diverse angiosperms[20]). This approach limits either the specificity provided by large divergence times or the sensitivity provided by the comparison of more closely related species[10,11,21]. Comparisons of paralogous noncoding regions flanking duplicated genes have also provided key insights into functional noncoding elements[16,22], 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 phylogeny[23] 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

[1]Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada. [2]Université Lyon 1, Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche (UMR) 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France. [3]School of Computer Science, McGill University, Montreal, Quebec, Canada. [4]McGill Centre for Bioinformatics, McGill University, Montreal, Quebec, Canada. [5]Department of Biology, McGill University, Montreal, Quebec, Canada. [6]Nature Sciences Department, Colby-Sawyer College, New London, New Hampshire, USA. [7]Department of Human Genetics, McGill University, Montreal, Quebec, Canada. [8]Institute of Vegetables and Flowers (IVF), Chinese Academy of Agricultural Sciences (CAAS), Beijing, China. [9]US Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [10]HudsonAlpha Institute of Biotechnology, Huntsville, Alabama, USA. [11]J. Craig Venter Institute, Rockville, Maryland, USA. [12]Division of Biological Sciences, University of Missouri, Columbia, Missouri, USA. [13]The School of Plant Sciences, University of Arizona, Tucson, Arizona, USA. [14]Plant Cytogenomics, Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czech Republic. [15]Biosystematics Group, Plant Sciences, Wageningen University, Wageningen, The Netherlands. [16]Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada. [17]These authors contributed equally to this work. Correspondence should be addressed to M.B. (blanchem@cs.mcgill.ca), S.I.W. (stephen.wright@utoronto.ca), T.E.B. (thomas.bureau@mcgill.ca) or A.M.M. (alan.moses@utoronto.ca).

Received 13 October 2012; accepted 4 June 2013; published online 30 June 2013; doi:10.1038/ng.2684

Table 1 Assembly statistics and gene content

	Genome assembly			Protein-coding gene content			Transposable elements
	Assembly size (Mb)/ expected genome size	Scaffold size N50 (kb)/max (Mb)	Called base (%)	Number of predicted genes	*A. thaliana* gene orthologs (%)	Complete CEGs[g] (%)	Genome coverage[h] (%)
A. thaliana	120/135[a]–157[b]	23,459/30.4	100	28,710	100	98.4	15
A. lyrata	207/230[b]–245[c]	24,464/33.1	89	27,379	92.0	98.0	32
C. rubella	135/210–216[c]	15,060/19.6	96	26,521	88.0	97.6	18
L. alabamica	174/316[e]	70/0.5	88	30,343	67.7	98.8	27
E. salsugineum	243/314[c]	13,441/21.8	98	26,521	82.7	98.4	50
S. parvula	114/140[d]	16,150/19.8	100	28,901	80.2	98.0	13
B. rapa	273/529	46/0.45	100	41,174	78.2	96.4	31
S. irio	259/262[b]	135/1.7	83	28,917	82.9	97.6	38
A. arabicum	203/240[f]	118/1.5	83	23,167	72.4	98.4	37

Unless mentioned otherwise, assembly and gene content statistics come from the genome paper cited in the text.
[a]Statistic from Lamesch *et al.*[75]. [b]Statistic from Johnston *et al.*[76]. [c]Statistic from Lysak *et al.*[77]. [d]Statistic from Dassanayake *et al.*[30]. [e]Statistic from B. Husband (personal communication). [f]Estimated from flow cytometry data. [g]Percentage of 248 ultraconserved core eukaryotic genes (CEGMA[34]) found in a complete form. [h]The incomplete assembly of the genomes of *L. alabamica*, *B. rapa*, *S. irio* and *A. arabicum* means that the reported TE coverage should be seen as lower bound.

plant *A. thaliana*[24]; (ii) *Arabidopsis lyrata*, a congener of *A. thaliana* with a more ancestral karyotype and genome size[25]; (iii) *Capsella rubella*, which falls in the sister group to the genus *Arabidopsis*[26]; (iv) *Brassica rapa* (Chinese cabbage[27]), one of the several closely related *Brassica* crop species in the tribe Brassiceae that share a recent genome triplication event (Br-α)[28]; (v) *Eutrema salsugineum* (previously *Thellungiella halophila*) of the tribe Eutremeae, an extremophile adapted to saline habitats[29]; and (vi) *Schrenkiella parvula* (previously *Thellungiella parvula*)[30], another extremophile of uncertain tribal placement.

To complement the set of previously published Brassicaceae genomes, we have sequenced the genomes of three additional species selected, on the basis of previously published phylogenetic analyses[31,32], 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 Camelineae), a model plant species with recently lost self-incompatibility in some populations; (ii) *Sisymbrium irio* (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[32]. 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[34] (**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.

Protein-coding genes and transposable elements (TEs) were predicted across all nine genomes (**Table 1**, Online Methods and **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[28]). 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 lone 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*[27,35], 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 family[36,37].

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*[38] as an outgroup was consistent with previously published phylogenies[31,39]. 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)[40] used in the identification of conserved noncoding regions. The divergence between *L. alabamica* paralogs (~0.3 substitutions per site) was similar

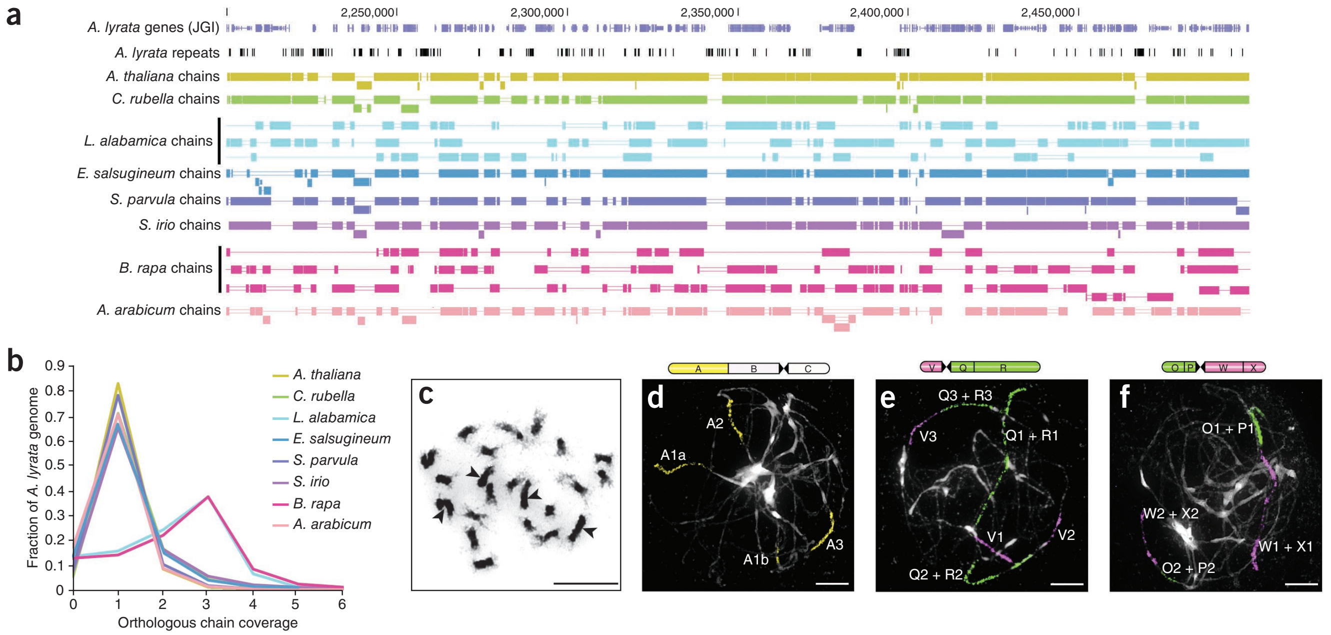


Figure 1 Genome alignments and whole-genome triplications. (**a**) Typical example of a 160-kb *A. lyrata* region aligning against the other eight Brassicaceae genomes. Alignment blocks (solid rectangles) are linked to form collinear chains. In alignments with most species, most positions in *A. lyrata* are contained within a single alignment chain, except in alignments with *B. rapa* and *L. alabamica*, where three alignment chains each were seen, suggesting whole-genome triplications in these two species. (**b**) Genome-wide level of chain coverage by each of the eight species of the *A. lyrata* genome. (**c**–**f**) Comparative chromosome painting (CCP) analysis in *L. alabamica* (2*n* = 22 chromosomes). (**c**) DAPI-stained mitotic chromosomes. Chromosomes of different size and heterochromatin content suggest an allopolyploid origin of the species. Arrowheads mark the four largest chromosomes. (**d**–**f**) CCP analysis of Cardamineae-specific ancestral chromosomes on pachytene (meiotic) chromosome spreads. (**d**) Three genomic copies of the genomic block A (the A1 copy is split between two different chromosomes). (**e**) Three genomic copies of chromosome AK6/8. (**f**) Two genomic copies of chromosome AK8/6; the third copy is absent. In **e**,**f**, copy 1 being substantially longer than the other homeolog(s) suggests an allohexaploid origin or differential fractionation of the three subgenomes. All scale bars, 10 µm.

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-α 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. 3a** and **Supplementary Fig. 1**). 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.

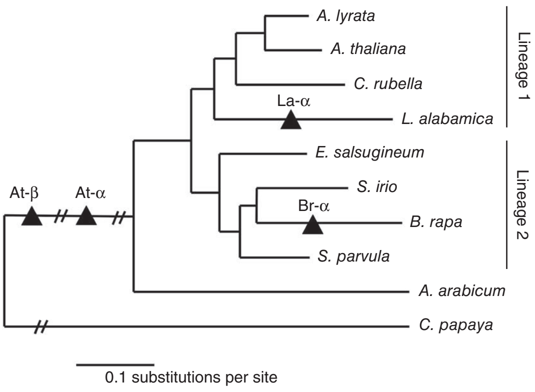


Figure 2 A phylogenetic tree obtained using a set of 1,048,889 fourfold-degenerate sites in PhyML (general time-reversible (GTR) substitution model)[73]. The positions (triangles) of the At-α duplication event and of the two whole-genome triplication events (Br-α and La-α) are approximate.

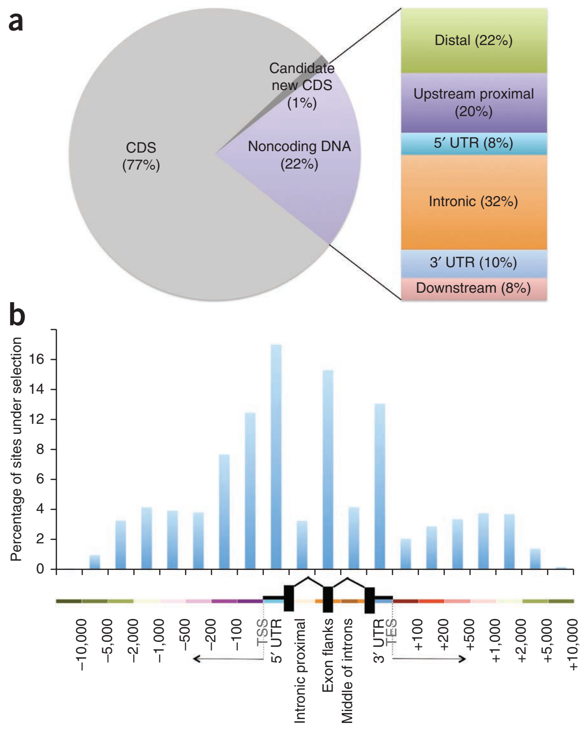


Figure 3 Estimation of the fraction of sites under selection in the *A. thaliana* genome. (**a**) Breakdown of sites under selection between coding (CDS) and noncoding regions (left) and among different types of noncoding regions (right). (**b**) Percentage of sites under selection for different sets of regions of the *A. thaliana* genome. Noncoding region categories include UTRs: 5′ (blue) and 3′ (light blue); intronic regions: 500 bp of intron 1 (tan), 30-bp intronic regions flanking exons (orange) and middle of introns other than intron 1 (brown); and intergenic regions: bases are assigned to the closest annotated TSS or TES.

Outside protein-coding transcripts, the proportion of sites under selection decreased with the distance from the TSS. Nonetheless, as observed in *Drosophila*[49], 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[50], 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 noncoding 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[5]. 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[51] and a set of 13 outbred individuals (26 haplotypes) of *C. grandiflora* (S.I.W., unpublished data), a close relative of *C. rubella*. 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 π (ref. 52) and Watterson's estimator θ_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[54,55].

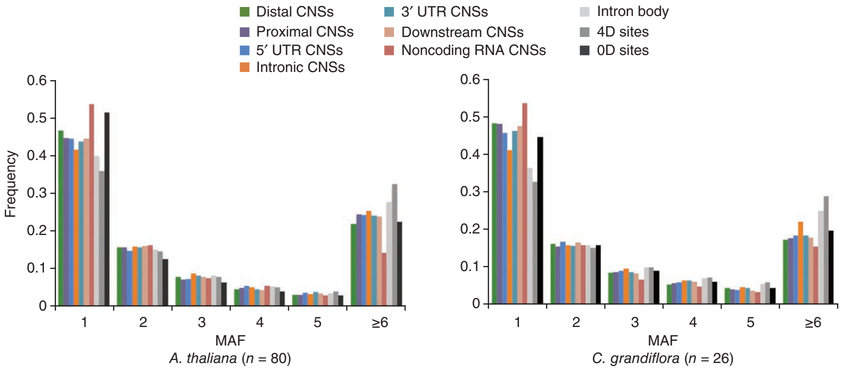
Distribution of CNSs in Brassicaceae and other plants

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

Figure 4 Evidence of selection on CNSs in populations. MAF distributions are shown at polymorphic sites within a population of 80 *A. thaliana* individuals (left) and a population of 13 *C. grandiflora* individuals with 26 haplotypes (right). All types of CNSs show 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. 0D, zero-fold degenerate.



after independent genome amplifications in closely related species[56]. Alignment chain data indicated that 38% of *A. lyrata* genes had a unique ortholog in *B. rapa*, whereas 25% and 7% had two and three copies, respectively. Despite similar estimates for the timing of Br-α and La-α, *B. rapa* seems to have lost gene copies faster, having kept two (three) copies of only 18% (3%) of *A. lyrata* genes. Notably, genes kept in three copies in *B. rapa* were five times more likely to also be kept in three copies in *L. alabamica* than would be expected if losses had occurred randomly. As previously observed, those genes were enriched for dosage-sensitive pathways[57] (for example, the response to environmental cadmium ions, false discovery rate (FDR) = 5 × 10^{-9}) or stoichiometry-dependent protein complexes[58] (for example, the ribosome, FDR = 2.6 × 10^{-6}).

Similarly, CNSs that were preserved in three copies in *B. rapa* were, depending on their type, two to four times more likely to be preserved in three copies in *L. alabamica* (**Supplementary Fig. 8**). Intronic CNSs showed the highest degree of convergent retention, and 5′ UTR CNSs showed the weakest retention. Notably, 136 of the 428 CNSs kept in 3 copies each in *B. rapa* and *L. alabamica* were also kept in 2 copies after the At-α duplication, 89% more than would be expected by chance.

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* chromatin immunoprecipitation chip (ChIP-chip) and sequencing (ChIP-seq) data[59] were combined, and the extent of their overlap with CNSs was assessed (**Supplementary Table 4**). Seventeen percent of bases in distal CNSs were found to overlap a region bound by at least one transcription factor, representing a 13-fold enrichment. Reciprocally, CNSs covered more than 35% of bases in distal bound regions. Similar overlaps were observed for proximal and downstream CNSs, although at slightly lower levels. However, the situation was quite different for intronic and 3′ UTR CNSs, where only 3–4% of bases overlapped a bound region. Additional evidence points to a role in post-transcriptional regulation for many of these intronic and 3′ UTR CNSs (**Supplementary Fig. 9** and **Supplementary Note**).

DNase I footprinting provides a more global perspective on protein binding to the genome. Overall, 44% of CNSs overlapped a recently published set of DNase I–hypersensitive sites (HSSs)[60], four times more than expected by chance. HSS-overlapping CNSs were mostly found in intergenic regions, with 51% (65%) of distal (proximal) CNSs overlapping an HSS but with only 9% of intronic CNSs and 23% of 3′ UTR CNSs showing overlap. Taken together, these results again suggest that CNSs in intronic regions are generally less likely than intergenic CNSs to mediate protein-DNA interactions. SNP density and MAF spectra clearly point to the portions of HSSs that overlap CNSs as the most strongly constrained subregions, whereas HSSs that had no overlap with CNSs showed reduced evidence of selective pressure (**Supplementary Fig. 10**).

We identified a set of 971 5-kb regions that were enriched for intergenic CNSs (coverage above 15%). These CNS-rich regions tended to be located next to genes involved in responses to hormonal stimuli, the regulation of transcription and organ development (**Supplementary Table 5**), matching previous reports in plants[14,22] and vertebrates[4]. Intergenic regions surrounding such genes, as well as their introns, were sometimes covered by up to 25–50% by CNSs. In these cases, it is likely that intronic CNSs have a role in transcriptional rather than post-transcriptional regulation.

Two CNS-rich regions surround the *MIR166A* and *MIR166B* loci, which harbor some of the rare CNSs for which homologs can be found outside Brassicaceae (**Supplementary Fig. 11**). These miRNA genes, which are conserved across eudicot and monocot plants, target the stability of members of the homeobox family of transcription factors[61] and are crucial to the

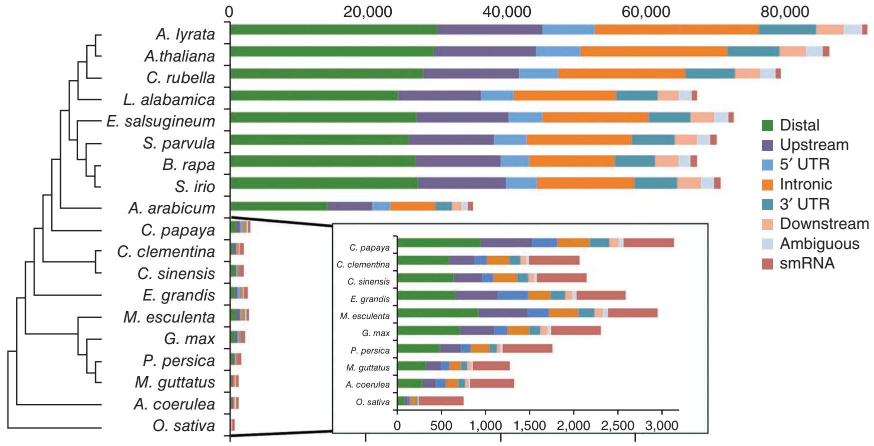
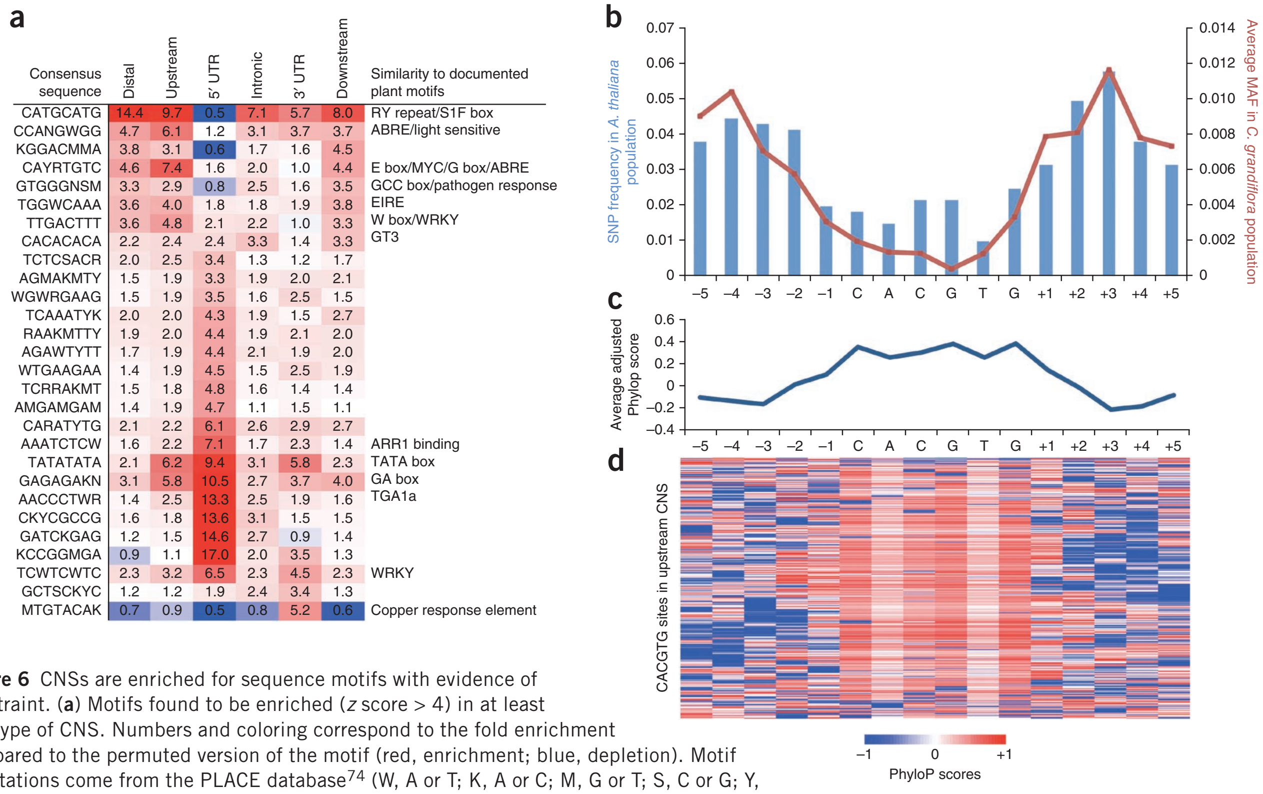


Figure 5 The majority of *A. lyrata* CNSs are shared with most other Brassicaceae, but few are conserved outside that clade, with the exception of those corresponding to smRNAs. *A. lyrata* CNSs were first symmetrically extended to at least 120 bp, except when this reached into coding exons. These sequences 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*.



Consensus sequence	Distal	Upstream	5′ UTR	Intronic	3′ UTR	Downstream	Similarity to documented plant motifs
CATGCATG	14.4	9.7	0.5	7.1	5.7	8.0	RY repeat/S1F box
CCANGWGG	4.7	6.1	1.2	3.1	3.7	3.7	ABRE/light sensitive
KGGACMMA	3.8	3.1	0.6	1.7	1.6	4.5	
CAYRTGTC	4.6	7.4	1.6	2.0	1.0	4.4	E box/MYC/G box/ABRE
GTGGGNSM	3.3	2.9	0.8	2.5	1.6	3.5	GCC box/pathogen response
TGGWCAAA	3.6	4.0	1.8	1.8	1.9	3.8	EIRE
TTGACTTT	3.6	4.8	2.1	2.2	1.0	3.3	W box/WRKY
CACACACA	2.2	2.4	2.4	3.3	1.4	3.3	GT3
TCTCSACR	2.0	2.5	3.4	1.3	1.2	1.7	
AGMAKMTY	1.5	1.9	3.3	1.9	2.0	2.1	
WGWRGAAG	1.5	1.9	3.5	1.6	2.5	1.5	
TCAAATYK	2.0	2.0	4.3	1.9	1.5	2.7	
RAAKMTTY	1.9	2.0	4.4	1.9	2.1	2.0	
AGAWTYTT	1.7	1.9	4.4	2.1	1.9	2.0	
WTGAAGAA	1.4	1.9	4.5	1.5	2.5	1.9	
TCRRAKMT	1.5	1.8	4.8	1.6	1.4	1.4	
AMGAMGAM	1.4	1.9	4.7	1.1	1.5	1.1	
CARATYTG	2.1	2.2	6.1	2.6	2.9	2.7	
AAATCTCW	1.6	2.2	7.1	1.7	2.3	1.4	ARR1 binding
TATATATA	2.1	6.2	9.4	3.1	5.8	2.3	TATA box
GAGAGAKN	3.1	5.8	10.5	2.7	3.7	4.0	GA box
AACCCTWR	1.4	2.5	13.3	2.5	1.9	1.6	TGA1a
CKYCGCCG	1.6	1.8	13.6	3.1	1.5	1.5	
GATCKGAG	1.2	1.5	14.6	2.7	0.9	1.4	
KCCGGMGA	0.9	1.1	17.0	2.0	3.5	1.3	
TCWTCWTC	2.3	3.2	6.5	2.3	4.5	2.3	WRKY
GCTSCKYC	1.2	1.2	1.9	2.4	3.4	1.3	
MTGTACAK	0.7	0.9	0.5	0.8	5.2	0.6	Copper response element

Figure 6 CNSs are enriched for sequence motifs with evidence of constraint. (**a**) Motifs found to be enriched (*z* score > 4) in at least one type of CNS. Numbers and coloring correspond to the fold enrichment compared to the permuted version of the motif (red, enrichment; blue, depletion). Motif annotations come from the PLACE database[74] (W, A or T; K, A or C; M, G or T; S, C or G; Y, C or T; N, any base). (**b**–**d**) Characteristics of bases in and around proximal upstream instances of the CACGTG motifs located in CNSs. (**b**) SNP density in the *A. thaliana* population (blue) and MAF distribution in the *C. grandiflora* population (red). (**c**) Position-specific average PhyloP scores over the set of motif instances. (**d**) Position-specific PhyloP scores for every instance of the set of motifs, ranging from –1 (non-conserved, blue) to +1 (highly conserved, red).

establishment of biological axes in stem, root and leaf[62,63]. Both *MIR166A* and *MIR166B* were surrounded by a cluster of intergenic CNSs, including six with clear homologs beyond the Brassicaceae as well as conservation between the two *A. thaliana* loci. Other miRNA genes that were associated with clusters of CNSs are listed in **Supplementary Table 6**.

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 motifs[7] (**Fig. 6a** and **Supplementary Table 7**). 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 (CACGTG) and the calcium signaling–sensitive E-box (CATGTG) 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-Slim[64], 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 species[20,65] or on intragenomic comparisons[22], 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 rate[66] (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 genomes 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[4]. 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/; RepeatModeler, http://www.repeatmasker.org/RepeatModeler.html; pybloomfaster, https://github.com/brentp/pybloomfaster; 1001 Genomes Project, http://1001genomes.org/data/MPI/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.

Note: Supplementary information is available in the online version of the paper.

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. salsugineum* 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. Joly 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/Génome 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 P501/12/G090). 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., S.I.W., 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. rapa* genome sequence data were provided by X.W., and *E. salsugineum* genome data were provided by J.S., D.E.J. 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.L. identified structural RNAs. Multiple-genome alignments and identification and analysis of CNSs were performed by A.H., A.E.P., 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.

1. Duret, L. & Bucher, P. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**, 399–406 (1997).
2. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391–1394 (2003).
3. Hong, R.L., Hamaguchi, L., Busch, M.A. & Weigel, D. Regulatory elements of the floral homeotic gene *AGAMOUS* identified by phylogenetic footprinting and shadowing. *Plant Cell* **15**, 1296–1309 (2003).
4. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
5. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
6. Margulies, E.H. *et al.* Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17**, 760–774 (2007).
7. Stark, A. *et al.* Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).
8. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
9. Adrian, J. *et al.* cis-Regulatory elements and chromatin state coordinately control temporal and spatial expression of *FLOWERING LOCUS T* in *Arabidopsis*. *Plant Cell* **22**, 1425–1440 (2010).

10. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53**, 661–673 (2008).
11. Freeling, M. & Subramaniam, S. Conserved noncoding sequences (CNSs) in higher plants. *Curr. Opin. Plant Biol.* **12**, 126–132 (2009).
12. Zou, C. *et al.* *Cis*-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **108**, 14992–14997 (2011).
13. Hsieh, T.F. *et al.* Regulation of imprinted gene expression in *Arabidopsis* endosperm. *Proc. Natl. Acad. Sci. USA* **108**, 1755–1762 (2011).
14. Inada, D.C. *et al.* Conserved noncoding sequences in the grasses. *Genome Res.* **13**, 2030–2041 (2003).
15. Guo, H. & Moose, S.P. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* **15**, 1143–1158 (2003).
16. Kaplinsky, N.J., Braun, D.M., Penterman, J., Goff, S.A. & Freeling, M. Utility and distribution of conserved noncoding sequences in the grasses. *Proc. Natl. Acad. Sci. USA* **99**, 6147–6151 (2002).
17. Bossolini, E., Wicker, T., Knobel, P.A. & Keller, B. Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: implications for wheat genomics and grass genome annotation. *Plant J.* **49**, 704–717 (2007).
18. Colinas, J., Birnbaum, K. & Benfey, P.N. Using cauliflower to find conserved non-coding regions in *Arabidopsis*. *Plant Physiol.* **129**, 451–454 (2002).
19. Haberer, G. *et al.* Large-scale *cis*-element detection by analysis of correlated expression and sequence conservation between *Arabidopsis* and *Brassica oleracea*. *Plant Physiol.* **142**, 1589–1602 (2006).
20. Hupalo, D. & Kern, A.D. Conservation and functional element discovery in 20 angiosperm plant genomes. *Mol. Biol. Evol.* published online; doi:10.1093/molbev/mst082 (27 May 2013).
21. Reineke, A.R., Bornberg-Bauer, E. & Gu, J. Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. *Nucleic Acids Res.* **39**, 6029–6043 (2011).
22. Thomas, B.C., Rapaka, L., Lyons, E., Pedersen, B. & Freeling, M. *Arabidopsis* intragenomic conserved noncoding sequence. *Proc. Natl. Acad. Sci. USA* **104**, 3348–3353 (2007).
23. Schranz, M.E., Lysak, M.A. & Mitchell-Olds, T. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci.* **11**, 535–542 (2006).
24. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
25. Hu, T.T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
26. Slotte, T. *et al.* The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **45**, 831–835 (2013).
27. Wang, X. *et al.* The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039 (2011).
28. Cheng, F. *et al.* Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. *Plant Cell* published online; doi:10.1105/tpc.113.110486 (7 May 2013).
29. Yang, R. *et al.* The reference genome of the halophytic plant *Eutrema salsugineum*. *Front. Plant Sci.* **4**, 46 (2013).
30. Dassanayake, M. *et al.* The genome of the extremophile crucifer *Thellungiella parvula*. *Nat. Genet.* **43**, 913–918 (2011).
31. Schranz, M.E., Song, B.H., Windsor, A.J. & Mitchell-Olds, T. Comparative genomics in the Brassicaceae: a family-wide perspective. *Curr. Opin. Plant Biol.* **10**, 168–175 (2007).
32. Couvreur, T.L. *et al.* Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Mol. Biol. Evol.* **27**, 55–71 (2010).
33. Bowers, J.E., Chapman, B.A., Rong, J. & Paterson, A.H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433–438 (2003).
34. Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289–297 (2009).
35. Lysak, M.A., Koch, M.A., Pecinka, A. & Schubert, I. Chromosome triplication found across the tribe Brassiceae. *Genome Res.* **15**, 516–525 (2005).
36. Schnable, J.C., Wang, X., Pires, J.C. & Freeling, M. Escape from preferential retention following repeated whole genome duplications in plants. *Front. Plant Sci.* **3**, 94 (2012).
37. Edger, P.P. & Pires, J.C. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* **17**, 699–717 (2009).
38. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
39. Bailey, C.D. *et al.* Toward a global phylogeny of the Brassicaceae. *Mol. Biol. Evol.* **23**, 2142–2160 (2006).
40. Thomas, J.W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
41. Yang, Y.W., Lai, K.N., Tai, P.Y. & Li, W.H. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J. Mol. Evol.* **48**, 597–604 (1999).
42. Town, C.D. *et al.* Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* **18**, 1348–1359 (2006).
43. Yang, L. & Gaut, B.S. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Mol. Biol. Evol.* **28**, 2359–2369 (2011).
44. Ponting, C.P. & Hardison, R.C. What fraction of the human genome is functional? *Genome Res.* **21**, 1769–1776 (2011).
45. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
46. Margulies, E.H. & Birney, E. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat. Rev. Genet.* **9**, 303–313 (2008).
47. Sorek, R. & Ast, G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* **13**, 1631–1637 (2003).
48. Halligan, D.L., Eyre-Walker, A., Andolfatto, P. & Keightley, P.D. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res.* **14**, 273–279 (2004).
49. Halligan, D.L. & Keightley, P.D. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* **16**, 875–884 (2006).
50. Blanchette, M. *et al.* Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16**, 656–668 (2006).
51. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011).
52. Nei, M. & Li, W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273 (1979).
53. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
54. Katzman, S. *et al.* Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).
55. Casillas, S., Barbadilla, A. & Bergman, C.M. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Mol. Biol. Evol.* **24**, 2222–2234 (2007).
56. Feldman, M. & Levy, A.A. Allopolyploidy—a shaping force in the evolution of wheat genomes. *Cytogenet. Genome Res.* **109**, 250–258 (2005).
57. Thomas, B.C., Pedersen, B. & Freeling, M. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* **16**, 934–946 (2006).
58. Luo, F., Liu, J. & Li, J. Discovering conditional co-regulated protein complexes by integrating diverse data sources. *BMC Syst. Biol.* **4** (suppl. 2), S4 (2010).
59. Muiño, J.M., Hoogstraat, M., van Ham, R.C. & van Dijk, A.D. PRI-CAT: a web-tool for the analysis, storage and visualization of plant ChIP-seq experiments. *Nucleic Acids Res.* **39**, W524–W527 (2011).
60. Zhang, W., Zhang, T., Wu, Y. & Jiang, J. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*. *Plant Cell* **24**, 2719–2731 (2012).
61. Kim, J. *et al.* microRNA-directed cleavage of *ATHB15* mRNA regulates vascular development in *Arabidopsis* inflorescence stems. *Plant J.* **42**, 84–94 (2005).
62. Nogueira, F.T. *et al.* Regulation of small RNA accumulation in the maize shoot apex. *PLoS Genet.* **5**, e1000320 (2009).
63. Nogueira, F.T., Madi, S., Chitwood, D.H., Juarez, M.T. & Timmermans, M.C. Two small regulatory RNAs establish opposing fates of a developmental axis. *Genes Dev.* **21**, 750–755 (2007).
64. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
65. Lyons, E. *et al.* Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* **148**, 1772–1781 (2008).
66. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
67. Schultz, S.T., Lynch, M. & Willis, J.H. Spontaneous deleterious mutation in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **96**, 11393–11398 (1999).
68. Ahituv, N., Prabhakar, S., Poulin, F., Rubin, E.M. & Couronne, O. Mapping *cis*-regulatory domains in the human genome using multi-species conservation of synteny. *Hum. Mol. Genet.* **14**, 3057–3063 (2005).
69. Lockton, S. & Gaut, B.S. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet.* **21**, 60–65 (2005).
70. Margulies, E.H., Blanchette, M., Haussler, D. & Green, E.D. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).
71. Hong, X., Scofield, D.G. & Lynch, M. Intron size, abundance, and distribution within untranslated regions of genes. *Mol. Biol. Evol.* **23**, 2392–2404 (2006).
72. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
73. Guindon, S., Lethiec, F., Duroux, P. & Gascuel, O. PHYML Online—a web server for fast maximum likelihood–based phylogenetic inference. *Nucleic Acids Res.* **33**, W557–W559 (2005).
74. Higo, K., Ugawa, Y., Iwamoto, M. & Korenaga, T. Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* **27**, 297–300 (1999).
75. Lamesch, P. *et al.* The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
76. Johnston, J.S. *et al.* Evolution of genome size in Brassicaceae. *Ann. Bot.* (*Lond.*) **95**, 229–235 (2005).
77. Lysak, M.A., Koch, M.A., Beaulieu, J.M., Meister, A. & Leitch, I.J. The dynamic ups and downs of genome size evolution in Brassicaceae. *Mol. Biol. Evol.* **26**, 85–98 (2009).

ONLINE METHODS

Sequencing and assembly. The genomes of *L. alabamica*, *A. arabicum* and *S. irio* 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 assembler[78] using a *K*-mer size of 31 to 41 (optimized per genome). Mate-pair reads were filtered for duplicates using a Bloom filter (pybloomfaster; 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 were then scaffolded using SOAPdenovo[79] with a *K*-mer size of 61 and with gap filling enabled.

The genomes of *B. rapa* (ssp. Chiifu-401-42; pre-publication data from Wang *et al.*[27]), *S. parvula* (PRJNA63667), *A. lyrata* (ssp. *lyrata*; PRJNA41137), *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. irio* 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 sequences[80] 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 blocks[23] on *L. alabamica* chromosomes: block A (31 BACs: T25K16–T29M8; 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: F7J8–T6G21; 7.4 Mb), block V (22 BACs: MBD2–K23F3; 2.4 Mb), block W (56 BACs: K21P3–MMN10; 4.3 Mb) and block X (26 BACs: MUP24–K9I9; 2.5 Mb).

Genome annotation. All nine genomes were annotated for genic regions using Maker[82] in conjunction with FGENESH and FGENESH+ (ref. 83), Augustus[84], SNAP[85] and BLAT[86] for transcript mapping. Repetitive regions and TEs were annotated with RepeatMasker using repeat models determined on a per-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 mRNA[87] (NCBI Sequence Read Archive (SRA) SRR019209, SRR019183 and SRR064165) and small RNA sequence data from both SOLiD and Illumina platforms (SRR040402, SRR072809, SRR034856 and SRR051926). Reads were aligned both with Novoalign (Novocraft) with high-alignment stringency and SpliceMap[88] 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 lastz[89], and chaining[90] 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 Multiz[91] 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.*[92], adapted to use site-specific conservation scores. PhyloP[45] 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 $\mathrm{R_S}$ 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 $f_{\mathrm{NS}}(x)$ 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 $f_{\mathrm{RS}}(x)$ be the observed distribution of PhyloP scores in $\mathrm{R_S}$. We express $f_{\mathrm{RS}}(x)$ as a mixture of $f_{\mathrm{NS}}(x)$ and $f_{\mathrm{FS}}(x)$, the unknown distribution of scores for sites under selection in $\mathrm{R_S}$. Specifically, we estimated α_{RS} so that $f_{\mathrm{RS}}(x) = \alpha_{\mathrm{RS}}\, f_{\mathrm{FS}}(x) + (1 - \alpha_{\mathrm{RS}})\, f_{\mathrm{NS}}(x)$. Let $F_{\mathrm{NS}}(x)$ and $F_{\mathrm{RS}}(x)$ be the cumulative distributions of $f_{\mathrm{NS}}(x)$ and $f_{\mathrm{NS}}(x)$, respectively. Let x^* be the value for which $F_{\mathrm{NS}}(x)/F_{\mathrm{RS}}(x)$ is maximized (excluding values of x for which either of the two cumulative distributions has a value less than 0.1), and let $r_{\max} = F_{\mathrm{NS}}(x^*)/F_{\mathrm{RS}}(x^*)$. We obtain $\alpha_{\mathrm{RS}} = \Sigma_{x = x^* \ldots 1{,}000}\, f_{\mathrm{RS}}(x) - (f_{\mathrm{NS}}(x)/r_{\max})$. Finally, the fraction under selection for region *R* is determined as $\Sigma_{\mathrm{S}}\, \alpha_{\mathrm{RS}}\, |\mathrm{R_S}|/|\mathrm{R}|$. Note that, because not all fourfold-degenerate sites are truly unconstrained, 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[5] 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 RNAcode[93], the absence of splice sites[88] 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 extensions 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[51] 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[94]. 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[95].

78. Boisvert, S., Laviolette, 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. Chantha, S.C., Herman, A.C., Platts, 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. Korf, 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).