nature genetics

# Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils

Thomas L Turner[1,2,5], Elizabeth C Bourne[3], Eric J Von Wettberg[4], Tina T Hu[1] & Sergey V Nuzhdin[1]

**A powerful way to map functional genomic variation and reveal the genetic basis of local adaptation is to associate allele frequency across the genome with environmental conditions[1–5]. Serpentine soils, characterized by high heavy-metal content and low calcium-to-magnesium ratios, are a classic context for studying adaptation of plants to local soil conditions[6,7]. To investigate whether *Arabidopsis lyrata* is locally adapted to serpentine soil, and to map the polymorphisms responsible for such adaptation, we pooled DNA from individuals from serpentine and nonserpentine soils and sequenced each 'gene pool' with the Illumina Genome Analyzer. The polymorphisms that are most strongly associated with soil type are enriched at heavy-metal detoxification and calcium and magnesium transport loci, providing numerous candidate mutations for serpentine adaptation. Sequencing of three candidate loci in the European subspecies of *A. lyrata* indicates parallel differentiation of the same polymorphism at one locus, confirming ecological adaptation, and different polymorphisms at two other loci, which may indicate convergent evolution.**





High-throughput sequencing of DNA pooled from multiple individuals is a highly efficient means of locating genomic variation. By pooling DNA from natural populations, the frequencies of each polymorphism within and between populations can be simultaneously estimated and used to study evolutionary processes on a genome-wide scale.

To investigate genomic variation in *Arabidopsis lyrata*, we extracted DNA from 25 plants from each of four populations: two from serpentine soils and two from granitic soils. We pooled the DNA from all plants at each site and sequenced it with two flow cells of the Illumina Genome Analyzer, resulting in a 39-fold coverage of the genome in total (**Supplementary Figs. 1–6**). Excluding the portion of the genome without unique read alignments, we used these sequencing reads to locate all common genetic polymorphisms. After discarding mismatches that were found only once in the full data set, which should eliminate most sequencing errors but also rare variation, we detected 12,398,558 polymorphisms across the genome.

Some polymorphisms have low coverage in at least one population, precluding allele frequency estimates, whereas others have high coverage and are likely to be repetitive. When polymorphisms with less than 3× or more than 30× coverage in any population were excluded, 8,433,201 polymorphisms remained (**Supplementary Fig. 7**). The frequency distribution of this variation within each population is similar, with median heterozygosities between 0.007 and 0.009. On the basis of the allele frequency differences between all pairs of populations, the genetic connectivity between populations was also quite uniform (median pairwise allele frequency difference is between 6.0% and 6.9%; **Supplementary Table 1**). These conditions therefore seem favorable for mapping adaptive variation by simply associating allele frequency with soil type.

Of the 8.4 million polymorphisms detected, 96 have allele frequency differences of greater than 80% between soil types (**Supplementary Table 2**). These soil type–associated polymorphisms are spread throughout the genome and are mostly unlinked to one another: if polymorphisms within 10 kb of one another are considered to be at the same locus, these 96 variants are spread across 82 loci. Using this arbitrarily stringent threshold, we asked whether these polymorphisms showed evidence of local adaptation. Using the initial *A. lyrata* genome annotation supplied by the Department of Energy Joint Genome Institute's community sequencing program (D. Weigel and M.P.I. Tübingen, personal communication), these 96 polymorphisms are in the exons, introns or within 1 kb of the start or stop codons of 81 genes.

To test the hypothesis that these polymorphisms are differentiated as a result of local adaptation to soil type, we compared the predicted functions of these genes to the annotations of all genes in the genome. Several gene ontology terms are over-represented among differentiated loci, including metal ion transmembrane transporter activity (the single most significant term) and calcium ion binding. By using the much more detailed gene annotations from the homologous genes in *Arabidopsis thaliana*, it is clear that many of the polymorphisms most strongly associated with soil type are excellent candidates for serpentine adaptation (**Table 1**).

We also calculated $F_{ST}$ (a metric of DNA differentiation that compares relatedness within and between populations[8]) for all 500-bp windows of the genome: when genes overlapping windows with the highest $F_{ST}$ between soil types were also included (**Table 2** and

[1]Molecular and Computational Biology, University of Southern California, Los Angeles, California, USA. [2]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria. [3]The Macaulay Land Use Research Institute, Craigiebuckler, Aberdeen, UK. [4]Evolution and Ecology, University of California, Davis, California, USA. [5]Present address: Evolution, Ecology, and Marine Biology Department, University of California Santa Barbara, Santa Barbara, California, USA. Correspondence should be addressed to T.L.T. (tturner@lifesci.ucsb.edu).

Received 19 August 2009; accepted 9 December 2009; published online 24 January 2010; doi:10.1038/ng.515

**Table 1 Candidate polymorphisms for serpentine adaptation**

| | Allele frequency | | | | |
|---|---|---|---|---|---|
| Rank | Granitic | Serpentine | Gene | Type | Annotation |
| 3 | 0.00 | 0.98 | *AT5G64560* | 3′ UTR | MRS2-2; magnesium transporter |
| 4 | 0.04 | 1.00 | *AT3G58060* | 6,419 bp 5′ of gene | MTPc3 cation efflux family protein/metal-tolerance protein, putative |
| 10 | 0.07 | 1.00 | *AT4G11880* | Intronic | Transcription factor; expressed in roots; involved in nutritional regulation |
| 12 | 0.00 | 0.92 | *AT4G03560* | Nonsynonymous: Val→Gly | Calcium channel 1; voltage-gated calcium channel in vacuolar membrane |
| 14 | 0.00 | 0.91 | *AT4G25960* | Intronic | PGP2; coupled to transmembrane movement of substances |
| 16 | 0.00 | 0.90 | *AT4G32890* | 321 bp 5′ of gene | Transcription factor, changes expression upon phosphate starvation |
| 18 | 0.00 | 0.90 | *AT2G29870* | 653 bp 5′ of gene | Major intrinsic family protein; water channel activity |
| 24 | 0.03 | 0.92 | *AT4G19960* | Intronic | KUP9; potassium ion transmembrane transporter |
| 27 | 0.00 | 0.89 | *AT4G38920* | Intronic | Vacuolar proton pump; root cap expression; root growth under salt stress |
| 33 | 0.03 | 0.90 | *AT1G35720* | Nonsynonymous: Asp→Glu | Annexin1: osmotic stress response; cadmium response; copper ion binding |
| 37 | 0.06 | 0.93 | *AT4G19680* | 5′ UTR | Iron transporter whose expression is induced by iron and zinc deficiency |
| 38 | 0.00 | 0.87 | *AT5G23980* | Exonic, fourfold degenerate | Ferric-chelate reductase; important in iron-deficient growth; zinc and magnesium accumulation |
| 46 | 0.86 | 0.00 | *AT3G17510* | 5′ UTR | CIPK1; interacts with calcium sensor CBL1; osmotic stress response |
| 49 | 0.06 | 0.92 | *AT5G17850* | Intergenic, between paralogs | CAX7 and CAX8; calcium:sodium antiporters |
| 64 | 0.00 | 0.85 | *AT5G63980* | 445 bp 3′ of gene | Induction of osmotic stress response through inositol triphosphate signaling |
| 68 | 0.84 | 0.00 | *AT5G09690* | Intronic | MRS2-7 magnesium transporter, paralog of MRS2-2 above |
| 70 | 0.11 | 0.94 | *AT3G45060* | 521 bp 5′ of gene | Nitrate transporter |
| 77 | 0.00 | 0.82 | *AT1G31120* | Intronic | KUP10; potassium ion transmembrane transporter; paralog of KUP9 above |
| 80 | 0.10 | 0.92 | *AT5G21274* | 185 bp 5′ of gene | Calmodulin 6; calcium ion binding, calcium-mediated signaling |
| 88 | 0.08 | 0.89 | *AT5G03570* | Nonsynonymous: Ser→Thr | Tonoplast-localized nickel transport protein |





**Supplementary Table 3**), we found 12.5% of the metal ion transmembrane transporters in the genome to be among the most differentiated loci between soils, despite the fact that loci that are strongly associated with soil type constitute only 0.6% of annotated genes (binomial $P = 5.85^{*}10^{-6}$). These genes include several cases of closely related gene family members, in completely different regions of the genome, which each overlap mutations that are highly associated with soil type: *AT4G19960* and *AT1G31120* (also known as *KUP9* and *KUP10*, respectively) encode potassium transporters (the smallest in-group containing these loci contains only three genes[9]); *AT3G58060* and *AT2G47830* (also known as *MTPc3* and *MTPc1*, respectively) encode metal-tolerance proteins (the smallest in-group contains four genes[9]); and *AT5G64560*, *AT5G09690* and *AT3G58970* (also known as *MRS2-2*, *MRS2-7* and *MRS2-6*, respectively) encode CorA-like $Mg^{2+}$ transporters (11 genes in *A. thaliana* are annotated with this domain).

One of these $Mg^{2+}$ transporter genes, *AT5G64560*, contains the third most strongly soil-associated polymorphism in the genome: a single-base-pair indel in the 3′ UTR. To further investigate the possibility of serpentine adaptation at this site, we sequenced this locus in a serpentine and a nonserpentine population of the European subspecies of *A. lyrata*: *A. lyrata petraea*. We found that few of the polymorphisms in the US *A. lyrata lyrata* populations were also polymorphic in the Scottish *A. lyrata petraea* populations. However, four other polymorphisms at this locus are completely differentiated between serpentine and nonserpentine populations in Scotland, including a SNP and 13-bp indel, which are also in the 3′ UTR. For comparison, we genotyped six polymorphic microsatellite loci in these Scottish populations and found $F_{ST}$ values ranging from 0.017 to 0.381 (mean = 0.200); in contrast, $F_{ST}$ at *AT5G64560* is 0.966 (**Table 3**). These values are consistent with convergent local adaptation to serpentine soil in the 3′ UTR of *AT5G64560*.

**Table 2 Candidate loci for serpentine adaptation among the windows of high $F_{ST}$**

| Rank | $F_{ST}$ | Gene | Annotation |
|---|---|---|---|
| 13 | 0.565 | *AT1G07600* | Between MT1a and MT1c; metallothionein proteins |
| 52 | 0.483 | *AT5G37500* | GORK; gated outwardly rectifying $K^+$ channel |
| 65 | 0.470 | *AT3G58970* | MRS2-6; magnesium transporter |
| 88 | 0.450 | *AT2G01770* | VIT1; an iron transporter required for iron sequestration into vacuoles |
| 103 | 0.440 | *AT5G13740* | ZIFL2; zinc homeostasis |
| 204 | 0.403 | *AT5G21326* | CIPK3; osmotic stress response |
| 206 | 0.403 | *AT2G47830* | MTPc1; cation efflux family protein/ metal-tolerance protein |

At some polymorphisms that are strongly associated with soil type, there is notably little differentiation at linked sites (**Fig. 1** and **Supplementary Fig. 8**). For example, the twelfth most strongly soil-associated polymorphism in the genome is a nonsynonymous change in the pore domain of the voltage-gated calcium channel, *Calcium channel 1* (*AT4G03560*). The frequency of the derived glycine is 1.00 and 0.85 on the two serpentine populations and 0.00 on both granitic populations, and the only linked polymorphism with more than 55% allele frequency difference between soil types is a shared polymorphism with *A. thaliana*[10]. One explanation for this lack of linked differentiation is that this valine-to-glycine substitution–causing polymorphism is an old variant that is maintained by spatially varying selection between soils with different calcium content[11]. If this were the case, this mutation might also be associated with serpentine environments in the *A. lyrata petraea* subspecies, despite the evolutionary distance between subspecies[12]. When we sequenced this locus in Scottish plants, we found that the derived glycine was indeed fixed in the serpentine population, whereas the ancestral valine was fixed in the nonserpentine population. Several additional mutations were differentiated between Scottish populations, but these sites were not polymorphic in the US populations. The lack of comparable differentiation at microsatellite loci strongly supports the possibility that this mutation is locally adaptive on serpentine.

A third locus was also sequenced in the Scottish populations: the intergenic region between the heavy-metal detoxification genes

**Table 3 Scottish DNA sequence and microsatellite data**

| | N n | S n | Len/# | $F_{ST}$ | N H | S H | N pi | S pi | Shared/fixed |
|---|---|---|---|---|---|---|---|---|---|
| **DNA sequence data** | | | | | | | | | |
| Ca channel 1 | 14 | 14 | 620 | 0.794 | 0.692 | 0.264 | 0.0014 | 0.0004 | 0/2 |
| MRS2 | 14 | 14 | 518 | 0.966 | 0.000 | 0.264 | 0.0000 | 0.0005 | 0/4 |
| Mt | 12 | 10 | 460 | 0.897 | 0.000 | 0.533 | 0.0000 | 0.0012 | 0/2 |
| **Microsatellite data** | | | | | | | | | |
| Contig65 | 12 | 11 | 2 | 0.158 | 0.000 | 0.309 | – | – | – |
| C10 | 13 | 13 | 3 | 0.017 | 0.481 | 0.519 | – | – | – |
| J03 | 12 | 13 | 3 | 0.381 | 0.432 | 0.590 | – | – | – |
| J17 | 12 | 11 | 6 | 0.249 | 0.670 | 0.682 | – | – | – |
| Ice14 | 13 | 12 | 2 | 0.193 | 0.782 | 0.598 | – | – | – |
| F19 panel | 13 | 13 | 2 | 0.203 | 0.147 | 0.494 | – | – | – |
| L12 panel | 13 | 13 | 1 | — | 0.000 | 0.000 | – | – | – |

N, nonserpentine population; S, serpentine population; n, sample size; Len/#, length of sequenced fragment or number of microsatellite alleles; H, haplotype (gene) diversity; pi, heterozygosity; Shared/fixed, shared polymorphisms or fixed differences.

*Metallothionein 1a* and *Metallothionein 1c*[13], which has the 13th-highest $F_{ST}$ between soil types in the United States (500-bp windows; **Table 2**). As there are many differentiated polymorphisms spread across 4 kb at this locus in the US sample, we hoped that a subset of these polymorphisms would also be differentiated in the *A. lyrata petraea* subspecies, providing a smaller set of candidate adaptive substitutions. Although the region sequenced in Scotland includes two of the three most differentiated US polymorphisms, neither of these sites is polymorphic in the Scottish sample. Similarly to the 3′ UTR of the $Mg^{2+}$ transporter sequenced above, however, other polymorphisms in this region were completely differentiated in Scotland, and these bases were not polymorphic in the US sample. The three candidate loci therefore have much higher levels of differentiation than all six polymorphic microsatellites (overall *t*-test, $P < 0.0001$).

Many plants are endemic to serpentine soils, possibly because mutations that allow plants to thrive on serpentine soil are deleterious on other soil types. If many of these 96 polymorphisms that are most strongly associated with soil type in *A. lyrata* experience such trade-offs, we would expect that the serpentine-associated allele at each differentiated locus would also not match the reference genome. The reference genome was sequenced from a line that was originally collected on sand dunes in Michigan by Charles Langley (C. Langley, personal communication), which is expected to have ion content more similar to granitic populations. Indeed, at 18 out of 20 of the candidates in **Table 1**, and 80% of the 96 most differentiated polymorphisms overall, the nonserpentine allele matches the reference

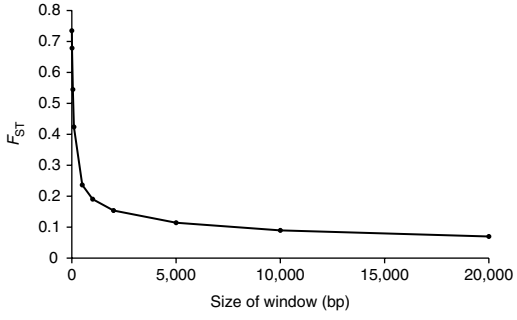


**Figure 1** The decay of average $F_{ST}$ around the 200 most strongly soil-associated polymorphisms. As the window size increases from 1 bp (the focal polymorphism) to 500 bp, the average $F_{ST}$ drops rapidly, indicating little linked differentiation at many sites.

*A. lyrata* genome (binomial $P = 9.31 \times 10^{-10}$). Moreover, many of these mutations occur in potentially functional sites: of the seven polymorphisms that fall within exons in **Table 1**, three are in UTRs, three are nonsynonymous and only one is synonymous. Of the 12 polymorphisms not in exons in **Table 1**, five are within 1 kb of the start codon, one is within 1 kb of the stop codon and six are in introns; only one polymorphism is not immediately adjacent to a gene of interest.

Finally, we scanned the genome for large indel polymorphisms (also referred to as copy number variation, or CNV) differentiated between soil types by comparing the number of aligned reads at each locus in each population. To focus on CNVs unique to serpentine soil, we limited our analysis to regions that seem to be present as a single copy in both granitic populations (75% of 1-kb windows). Of these windows, 177 (0.1%) have greater than twice as much coverage on both serpentine populations than on the granitic populations and are therefore candidate duplications. A larger proportion, 997 windows (0.6%), had less than half as much coverage in the serpentine populations as in the granitic ones, indicating potential deletions. By combining windows within 5 kb of one another, we detected a total of 94 candidate duplications and 373 candidate deletions within serpentine populations. Some 'deleted' regions may also be regions of high divergence, in which reads from serpentine populations are unalignable. Genes within these regions are also found to be a nonrandom subset of genes in the genome. For example, three different candidate deletions include members of the multidrug and toxic compound extrusion (MATE) family ($P = 0.0006$). In addition, a fourth gene in this family, *AT2G04050*, overlaps one of the 100 windows of highest $F_{ST}$ in the genome. Because genes are likely to be duplicated and deleted at different rates owing to variation in mutation rates and functional constraint, and because these forces may be correlated with gene function, the evolutionary significance of nonrandomness among CNV loci is unclear. Notably, however, one of these MATE loci is known as *ABERRANT LATERAL ROOT FORMATION 5* (*AT3G23560*) in *A. thaliana* and is implicated in protection of the roots from inhibitory compounds[14].

Among candidate CNVs, several are large and contain multiple genes. Among candidate serpentine duplications, 27 1-kb windows (20% of candidate duplicated windows) are within a 78-kb region containing approximately 11 genes. Similarly, 27 1-kb windows that are candidate serpentine deletions are within a 63-kb region containing 7 gene annotations. The presence of large CNVs in these regions is verified by genomic tiling array hybridizations, which show substantial differences in probe hybridization intensity in these same regions[15] (**Supplementary Fig. 9**). Similarly large regions of the genome were also found to be associated with climate in populations of *Drosophila melanogaster*[4]—notably, it seems that such mutations might have a role in local adaptation.

Under some circumstances, the long history of gene flow, recombination, selection and mutation in natural populations leads to a strong association between conditionally adaptive mutations and the environmental conditions with which they interact. This association between genetic and environmental variation can also result from neutral demographic processes if, for example, gene flow is higher between similar environments. However, by combining allele frequency data with functional gene annotations, our data point to a role for local

adaptation at many loci. Moreover, we observe the same associations between individual loci and serpentine soil in distantly related populations in Scotland, further implicating soil-mediated selection. A more comprehensive understanding of the relative roles of selection and demography can probably be gained by genotyping many populations, in a broad range of environments, and determining the multivariate relationships between environmental variables and genomic polymorphism. Because allele frequency across the genome can now be efficiently determined using pooled high-throughput sequencing, this approach provides an excellent opportunity to study the genomic basis of adaptation and locate variants with functionally useful properties.

## METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

*Note: Supplementary information is available on the Nature Genetics website.*

### ACKNOWLEDGMENTS

We are grateful to M. Nordborg, R. Pakeman and R. Brooker for their advice and support, and to M. Woodhead, J. Russell and C. Booth for assistance with microsatellite analysis. This work was supported by the US National Science Foundation (NSF) grant DEB-0548991 and US National Institutes of Health (NIH) grant RGM-076643 (S.V.N.), NSF DEB-0723935 (to M. Nordborg), the University of Southern California (S.V.N., T.L.T. and T.T.H.), an NSF predoctoral fellowship (T.L.T.), the Macaulay Development Fund (E.C.B.), the Gregor Mendel Institute (T.L.T.) and an NIH NRSA fellowship (E.J.V.W.).

### AUTHOR CONTRIBUTIONS

T.L.T. and S.V.N. designed experiments; T.L.T., E.C.B. and T.T.H. performed analyses; E.C.B. and E.J.V.W. designed and performed all field collections; T.L.T., E.J.V.W. and S.V.N. wrote the paper.

### COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Oakeshott, J.G. *et al.* Alcohol dehydrogenate and glycerol-3-phosphate dehydrogenate clines in *Drosophila melanogaster* on different continents. *Evolution* **36**, 86–96 (1982).
2. Thompson, E.E. *et al. CYP3A* variation and the evolution of salt-sensitivity variants. *Am. J. Hum. Genet.* **75**, 1059–1069 (2004).
3. Hancock, A.M. *et al.* Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet.* **4**, e32 (2008).
4. Turner, T.L., Levine, M.T., Eckert, M.L. & Begun, D.J. Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. *Genetics* **179**, 455–473 (2008).
5. Young, J.H. *et al.* Differential susceptibility to hypertension is due to selection during the Out-of-Africa expansion. *PLoS Genet.* **1**, e82 (2005).
6. Kruckeberg, A.R. Intraspecific variability in the response of certain native plant species to serpentine soil. *Am. J. Bot.* **38**, 408–419 (1951).
7. Kruckeberg, A.R. *California Serpentines: Flora, Vegetation, Geology, Soils, and Management Problems* (University of California Press, Berkeley, California, USA, 1984).
8. Wright, S. *Evolution and the Genetics of Populations Volume 2: The Theory of Gene Frequencies* (University of Chicago Press, Chicago, 1969).
9. Mäser, P. *et al.* Phylogenetic relationships within cation transporter families of *Arabidopsis*. *Plant Physiol.* **126**, 1646–1667 (2001).
10. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
11. Przeworski, M., Coop, G. & Wall, J.D. The signature of positive selection on standing genetic variation. *Evolution* **59**, 2312–2323 (2005).
12. Ross-Ibarra, J. *et al.* Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS One* **3**, e2411 (2008).
13. Zhou, J. & Goldsbrough, P.B. Functional homologs of fungal metallothionein genes from *Arabidopsis*. *Plant Cell* **6**, 875–884 (1994).
14. Diener, A.C., Gaxiola, R.A. & Fink, G.R. *Arabidopsis ALF5*, a multidrug efflux transporter gene family member, confers resistance to toxins. *Plant Cell* **13**, 1625–1638 (2001).
15. Turner, T.L., von Wettberg, E.J. & Nuzhdin, S.V. Genomic analysis of differentiation between soil types reveals candidate genes for local adaptation in *Arabidopsis lyrata*. *PLoS One* **3**, e3183 (2008).

## ONLINE METHODS

**Sample collection.** We collected North American seeds at two serpentine sites and two granitic sites in close proximity from May to September, 2006 (see ref. 14 for further information). The serpentine sites are part of the Nature Conservancy's state line serpentine barren complex in Chester County, Pennsylvania, USA, and Soldier's Delight Natural Environment Area, Baltimore County, Maryland, USA. Granitic sites were located at the Lock 12 recreation area on the Susquehanna river (York County, Pennsylvania) and Lancaster County Park (Lancaster County, Pennsylvania). We collected Scottish plants in August and September 2008 at Ben Avon, Eastern Cairngorms site of special scientific interest (SSSI) (nonserpentine) and Coyles of Muick, Balmoral Estate (serpentine), which are separated by approximately 23 km. Individual plants were collected in three subareas of each site, with subsequent analysis using individuals representing each subarea.

**DNA isolation and genome sequencing.** We isolated DNA, using the Qiagen Plant Miniprep kit, from approximately 200 tissue samples (for the US samples) or by the CTAB method (for the Scottish samples). For each US population, we selected the 25 cleanest and highest-yield samples for pooling. We assessed the samples using the Nanodrop spectrophotometer, analyzing three or more aliquots per sample. We created a DNA pool for each population by adding 180 ng of DNA from each of the 25 samples, creating a pool of 4.5 μg. We further cleaned these samples by phenol-chloroform purification and ethanol precipitation into equal volumes of water. We constructed a single sequencing library from each DNA pool by Canada's Michael Smith Genome Sciences Centre (GSC) at the BC Cancer Agency, according to standard Illumina protocols. Each DNA library (one from each population) was sequenced with 16 'lanes' of the Illumina/Solexa Genome Analyzer at the GSC. In total, this resulted in 540 million sequencing reads, or 19.44 billion bases of sequence data. We amplified the loci sequenced in the Scottish populations using standard PCR conditions and sequenced them directly. For microsatellite loci, PCR products were amplified with primers from the literature[16,17], diluted and precipitated on an ABI 3730 automated sequencer. We used GENESCAN ROX 500 as an internal size standard and analyzed the samples using GENEMAPPER version 3.7 (Applied Biosystems).

**Read alignment and sequence analysis.** Illumina sequencing reads were aligned to the Araly1 assembly of the *A. lyrata* genome supplied by DOE-JG's community sequencing program (D. Weigel, unpublished data). We mapped the alignments using the Short Oligonucleotide Alignment Program (SOAP) v.1 (ref. 18). For each 36-bp sequencing read, we used SOAP to first search for perfect (zero mismatch) alignments to the 695 nuclear scaffolds of the Araly1 assembly. If no perfect alignments were found, we used SOAP to search for 1-bp mismatches, then 2-bp mismatches and finally alignments with a single insertion or deletion of 1–3 bp. If, after this iterative procedure, no alignments were found, we removed the last 2 bp of the read and repeated the process (beginning with a search for perfect matches to the 34-bp read). In this way, SOAP first searched for alignments using the 36-mer, then the 34-mer, the 32-mer, the 30-mer and finally the 28-mers. This procedure is intended to locate alignments for reads with low-quality bases at the end of the read, where most sequencing errors occur with the Illumina technology (**Supplementary Figs. 1** and **2**). If more than one alignment was found with equal probability (for example, multiple perfect 36-bp matches, multiple 28-bp matches that each have 2 mismatches, and so on), we noted the read as being nonunique and did not use it (**Supplementary Fig. 3**). Using the first release of the *A. lyrata* reference genome, 56% of reads could be aligned to the nuclear genome, with an additional 2.6% of reads aligning to the plastid genomes. Unaligned reads are likely to include those with low quality, bacterial or other contamination and loci (such as heterochromatin) that are not assembled in the Araly1 assembly. Of the alignable reads, 61% align to one locus in the genome with higher probability than any second locus, providing unique alignments that were used to locate sequence polymorphisms.

**Supplementary Figure 4** shows the distribution of sequencing coverage across the genome: data are shown for the eight largest scaffolds in the reference assembly, corresponding to the eight *A. lyrata* nuclear chromosomes, and not for the smaller unplaced scaffolds (this is true throughout, except where stated otherwise). Regardless of whether repetitive alignments are included, the mode genome coverage was 39-fold. When repetitive reads were excluded, 16% of bases have zero coverage, as no unique 36-mers were found in these regions. Coverage varied between the four populations, because yield and quality differed greatly between flow cells, and some flow cells were run entirely with DNA from a single population. The mode coverage values for each population were 9×, 10×, 12× and 5× for granite1, granite2, serpentine1 and serpentine2 soil types, respectively (**Supplementary Fig. 5**).

Excluding the portion of the genome without unique read alignments, we can use these sequencing reads to locate all common genomic polymorphisms in these populations. In the full set of alignments, we found 167 million sequence mismatches between the uniquely aligning sequencing reads and the reference genome. We used two methods to purge sequencing errors from this set of mismatches: quality filtering and frequency trimming. The Illumina sequencing analysis software generates a quality score for each base: **Supplementary Figure 6** shows the quality scores for all single-nucleotide mismatches, where 40 is the best quality and 1 the worst quality. We considered all mismatches with a quality score of less than 10 to be probable errors, and we therefore discarded 79% of all mismatches. Second, we eliminated all mismatches that were found only once in the full data set, as random sequencing errors should therefore be largely eliminated. As the total sequencing coverage is 39-fold, we therefore also discarded rare polymorphisms (those with observed frequencies of less than 2 in 39, or 5%).

After these filtering criteria were applied, we detected 12,398,558 polymorphisms across the genome. Some polymorphic loci showed insufficient coverage in at least one population; others showed very high coverage and are probably repetitive. When polymorphisms with less than 3× or more than 30× coverage in any population were excluded, 8,433,201 polymorphisms remained; the frequency spectra for each population are shown in **Supplementary Figure 7**.

**Copy number variation.** To search for CNVs, we calculated the average read coverage for all nonoverlapping 1,000-bp windows of the genome for each population. Because the average coverage differed between populations, we standardized the windows by dividing by the median coverage of all windows for that population. To focus on regions that seemed to be single copy (relative to the reference) in granitic populations, but duplicated or deleted on serpentine soil, we discarded windows with low or high coverage on granite. We removed windows with coverage of less than 25% of the median or greater than 175% of the median in either granite population and compared the remaining 130,143 windows (75%) to serpentine populations. Of these windows, those with more than twice as much coverage on granite were considered candidate deletions and those with more than twice as much coverage on serpentine were considered candidate duplications. We used nonoverlapping windows of 1 kb, and we considered windows within 5 kb of one another as the same CNV.

**URL.** The *A. lyrata* genome is available at http://genome.jgi-psf.org/Araly1/Araly1.home.html.


16. Woodhead, M. *et al.* Development of EST-derived microsatellite markers for *Arabidopsis lyrata* subspecies *petraea* (L.). *Mol. Ecol. Notes* **7**, 631–634 (2007).
17. Clauss, M.J. & Cobban, H. Cross-species microsatellite markers for elucidating population genetic structure in *Arabidopsis* and *Arabis* (Brassicaceae). *Mol. Ecol.* **11**, 591–601 (2002).
18. Li, R., Li, Y.R., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).





doi:10.1038/ng.515