Massive genomic variation and strong selection in
Arabidopsis thaliana lines from Sweden


Supplementary figures

1 Map of sampling locations .................................................. 2
2 Chromosomal distribution of large indels .................................. 3
3 Distribution of flow cytometry estimates .................................... 4
4 FISH results for three lines with large estimated 45S rDNA copy number . .................................................. 5
5 GWAS results for genome size using different samples and phenotypes .................................................. 6
6 Derived allele frequency distributions for indels ............................................ 7
7 Decay of LD for different sub-populations ...................................... 8
8 UPGMA clustering of the Swedish lines based on all SNPs .................................................. 9
9 Isolation by distance .......................................................... 10
10 Sweep statistics for chromosome 1 .................................................. 11
11 Sweep statistics for chromosome 2 .................................................. 11
12 Sweep statistics for chromosome 3 .................................................. 12
13 Sweep statistics for chromosome 4 .................................................. 12
14 Sweep statistics for chromosome 5 .................................................. 13
15 Summaries of the pattern of polymorphism in north and south .................................................. 13
16 Validation of transposition ..................................................... 14
17 LD surrounding putative major rearrangements .................................. 15
18 False positive rates in reference re-sequencing as a function of Q-value cut-off .................................................. 16
19 Distribution of putative false positives in reference re-sequencing .................................................. 16
20 Error rates from comparison with SNP data as function of Q-value .................................................. 16
21 Characterization of novel sequence .................................................. 17
22 Correlation between flow cytometry and repeat copy number estimates .................................................. 18
23 Correlation between flow cytometry and 45S rDNA .................................................. 18
24 Effect of LD transformation ..................................................... 19
25 The robustness of long-range LD .................................................. 19

Supplementary tables

1 Error rates for SNPs and short indels .................................................. 20
2 Validation of long-range LD ..................................................... 20
3 Enrichment of long-range LD SNPs among SNPs associated with climate .................................................. 20
4 Summary of simple regressions of flow cytometry estimates. .................................................. 21
5 Genetic and geographic distance between world-wide populations .................................................. 21

Supplementary note

1 Genome sequencing ..................................................... 22
2 Polymorphism detection ..................................................... 22
3 De novo assembly .......................................................... 25
4 Variation in nuclear DNA content .................................................. 27
5 Population genetic analyses ..................................................... 28
6 Data release ............................................................... 31
Supplementary Figure 1  Map of sampling locations. For a list of lines used, see project download site.
Supplementary Figure 2 Chromosomal distribution of positive (blue) and negative (green) indels, compared with SNP polymorphism levels (black). From top to bottom, chromosomes 2–5. Results for chromosome 1 are in the main text (Fig. 1f).
Supplementary Figure 3 The distribution of flow cytometry estimates for the global (36), southern (90), and northern (38) lines. The mean DNA content for the three samples are 166, 167, and 170, respectively. The northern measurements are significantly larger than those in the other two samples (Mann-Whitney p-value < \(10^{-4}\)), whereas are southern sample is not significantly different from the world-wide sample.
Supplementary Figure 4  FISH results for three lines with large estimated 45S rDNA copy number: 6244 (TRÄ 01); 6043 (Löv-1); and; 5856 (Dör-10). The known 45S rDNA clusters on chromosome 2 and 4 are highlighted (in yellow), as is chromosome 1 (green) and the region on chromosome 1 that contains the main association peak (red). There is no evidence for any novel 45S rDNA clusters.
Supplementary Figure 5  Comparison of GWAS results for genome size using different samples and phenotypes. (a) GWAS using flow cytometry estimates for 128 lines (same as Fig. 2b). (b) GWAS using flow cytometry estimates for 38 northern lines. (c) GWAS using flow cytometry estimates for 90 southern lines. (d) GWAS using 45S coverage estimates for the 128 lines for which flow cytometry estimates are available. (e) GWAS using 45S coverage estimates for the full sample of 180 lines. The association peak discussed in the text is marked by the horizontal orange bar. It is clearly present in a, b, and d, but not in c and e.
Supplementary Figure 6 Derived allele frequency distributions for different parts of the genome. The distributions were estimated by dividing the frequency into 100 bins, correcting the estimates using the hypergeometric distribution\(^1\), and averaging across bins to create bins for the plot. Very conservative criteria were used to polarize the polymorphisms in order to avoid inflation at the right end of the plots. (a) Intergenic regions. (b) Genic regions. (c) Introns. (d) Exons. (e) CDS. (f) Whole genome. There is no evidence for high-frequency deletions in any category.
Supplementary Figure 7  Decay of LD for different sub-populations (cf. Fig. 1 in reference [2] and Fig. 3 in reference [3]). Sub-sampling was used to ensure equal sample sizes in all sub-populations.
Supplementary Figure 8  UPGMA clustering of the Swedish lines based on all SNPs. Northern lines are labeled blue and southern red.
Supplementary Figure 9  Isolation by distance. Sequence divergence (proportion of sites that differ) between all pairs of Swedish lines as a function of the distance between the sample locations. Lines are fitted using least squares regression. The geographic distance between pairs was calculated from the geodesic distance using the Vincenty inverse formula for ellipsoids (function gdist of the R-package lmap).
Supplementary Figure 10  Sweep statistics for chromosome 1. Blue curves are for northern Sweden, red for southern Sweden. For CLR, dashed lines indicate statistical cut-offs from simulations. XP-CLR and CLR were calculated for grid points with 1 kb spacing, XP-EHH was calculated for each SNP, and \( F_{ST} \) (between north and south) and nucleotide diversity were calculated in non-overlapping 100 kb windows.

Supplementary Figure 11  Sweep statistics for chromosome 2. See Supplementary Fig. 10 for details.
**Supplementary Figure 12** Sweep statistics for chromosome 3. See Supplementary Fig. 10 for details.

**Supplementary Figure 13** Sweep statistics for chromosome 4. See Supplementary Fig. 10 for details.
Supplementary Figure 14  Sweep statistics for chromosome 5. See Supplementary Fig. 10 for details.

Supplementary Figure 15  Summaries of the pattern of polymorphism in north and south, in 100 kb windows across the genome. (a) Nucleotide diversity, $\pi$. (b) Watterson’s $\theta$. (c) Tajima’s $D$. Only non-coding sites were used for these plots.
Supplementary Figure 16 Validation of transposition. (a) PCR primers used to confirm breakpoints. (b) PCR from lines 6124 (swept) and 6046 (unswept) showing products of expected length consistent with the predicted arrangement. Results were consistent in 15 swept and 5 unswept lines (not shown). (c) Likely mutational events that generated this transposition. Removal of the 278 kb and rejoining at breakpoint 3 was likely facilitated by 5 bp of microhomology, leading to the deletion of one copy. Insertion/capture of the 273 kb between breakpoint 1 and 2 lead to the deletion of 9 bp. (d) Sequence flanking all 6 breakpoints (1–3 from Col-0; 4–6 Sanger sequenced from 6046), and alignments in support of the predicted arrangement (with identity underlined and repeat structures in color).
Supplementary Figure 17  LD surrounding putative major rearrangements identified by de novo assembly. (a) The 1170kb region between chr4:1612606 and chr4:2782618 appears to be inverted in 171 of our 180 Swedish lines, and comparison with A. lyrata shows that it is the A. thaliana reference genome that carries the inversion. (b) The 28 kb region between chr3:16039026 and chr3:16067070 appears to be missing with frequency of 106/180. The region contains several annotated transposable elements but also At3g44400, a possible resistance protein. (c) The 14–15 kb region between chr5:19412000 and chr5:19426000/19427000 appears to be inverted in 140/180 lines.
**Supplementary Figure 18** False positive rates in reference re-sequencing as a function of Q-value cut-off. (a) SNPs. (b) Indels.

**Supplementary Figure 19** The distribution of putative false positives from reference re-sequencing (for which all polymorphisms are assumed to be false) as a function of chromosomal position and alignment quality. (a) SNPs. (b) Indels. Colors indicate mapping quality (Q value: red is high, yellow low). For SNPs (but not indels), false positives are clustered near centromeres and tend to have low Q values.

**Supplementary Figure 20** Error rates from comparison with SNP data as function of Q-value.
Supplementary Figure 21  Characterization of novel sequence. Note that the light/dark green bars are independent of the olive green bars. See Supplementary Note for details.
Supplementary Figure 22  Correlation between flow cytometry and repeat copy number estimates. For 45S rDNA, see Fig. 2a. Note that the TE count refers to the number of novel insertions detected.

Supplementary Figure 23  Correlation of flow cytometry and 45S rDNA. (a) Correlation between flow cytometry and 45S rDNA copy number estimates before removal (a) and after removal (b) of duplicated reads. Correlation between 45S rDNA copy number and 18S rDNA qPCR estimates (c) before and (d) after removal of duplicated reads.
**Supplementary Figure 24** Effect of LD transformation. (a) The plot shows 16 million pairs of SNPs, selected at random from pairs with $r^2 > 0.05$ in the original data. (b) Average decay of LD.

**Supplementary Figure 25** The robustness of long-range LD. The correction for population structure described in section 5.2.1 was used throughout. (a) LD in Northern Swedish population only. Due to smaller sample size, a higher minor allele frequency cutoff of 0.12 was used. (b) LD in Southern Swedish population only (minor allele frequency cutoff of 0.10). (c) LD in unimputed data, illustrating similarity to plot for imputed data (Fig. 4a in main text). (d) LD calculated from imputed data, but removing all high LD pairs from the previous plot. This “subtracted” LD plot, shows that imputation creates few additional high LD pairs.
Supplementary Table 1  Estimated error rates (%) for SNPs and short indels. In the case of indels, separate estimates are given for variants that are longer/shorter than the reference genome whenever enough polymorphisms were observed. For SNPs, all comparisons are with the data from the Q30 regions.

<table>
<thead>
<tr>
<th>Quality control data</th>
<th>False positives</th>
<th>False negatives</th>
<th>Genotyping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNPs</td>
<td>Indels +/-</td>
<td>SNPs</td>
</tr>
<tr>
<td>Reference re-sequencing</td>
<td>0.21</td>
<td>2.5/1.4</td>
<td>NA</td>
</tr>
<tr>
<td>SNP-chip</td>
<td>NA</td>
<td>NA</td>
<td>4.6</td>
</tr>
<tr>
<td>Sanger-sequenced PCR amplicons</td>
<td>1.8</td>
<td>0.8/1.6</td>
<td>3.7</td>
</tr>
<tr>
<td>Sanger-sequenced random clones</td>
<td>1.8</td>
<td>2.4</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Supplementary Table 2  Validation of long-range LD. Six different sets of SNPs exhibiting long-range LD were tested in crosses: the number give r² in the natural as well as the F2 population (where r² = 0 is expected under independent segregation). “ND” means that PCR amplification failed so that genotyping was not possible.

<table>
<thead>
<tr>
<th>centromeric</th>
<th>SNP 1</th>
<th>SNP 2</th>
<th>SNP 3</th>
<th>1 vs 2</th>
<th>1 vs 3</th>
<th>2 vs 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chr</td>
<td>position</td>
<td>chr</td>
<td>position</td>
<td>chr</td>
<td>position</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>724,571</td>
<td>2</td>
<td>8,985,116</td>
<td>2</td>
<td>9,210,944</td>
</tr>
<tr>
<td>N</td>
<td>4</td>
<td>958,430</td>
<td>5</td>
<td>7,723,187</td>
<td>5</td>
<td>7,737,565</td>
</tr>
<tr>
<td>Y</td>
<td>3</td>
<td>13,649,143</td>
<td>1</td>
<td>13,341,722</td>
<td>1</td>
<td>13,359,206</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>14,168,125</td>
<td>2</td>
<td>3,086,551</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Y</td>
<td>4</td>
<td>4,738,528</td>
<td>1</td>
<td>12,831,211</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>12,964,602</td>
<td>5</td>
<td>14,083,994</td>
<td>5</td>
<td>14,088,003</td>
</tr>
</tbody>
</table>

Supplementary Table 3  p-Values for enrichment of long-range LD SNPs among SNPs associated with climate.

<table>
<thead>
<tr>
<th>Climate variable</th>
<th>Distance to peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 kb</td>
</tr>
<tr>
<td>consecutive frost free days</td>
<td>0.3476</td>
</tr>
<tr>
<td>daylength</td>
<td>0.7466</td>
</tr>
<tr>
<td>maximum temperature</td>
<td>0.0995</td>
</tr>
<tr>
<td>minimum temperature</td>
<td>0.3534</td>
</tr>
<tr>
<td>length of growing season</td>
<td>0.3224</td>
</tr>
<tr>
<td>consecutive cold days</td>
<td>0.3467</td>
</tr>
<tr>
<td>relative humidity</td>
<td><strong>0.0034</strong></td>
</tr>
<tr>
<td>photosynthetically active radiation</td>
<td>0.4332</td>
</tr>
<tr>
<td>temperature seasonality</td>
<td>0.5397</td>
</tr>
<tr>
<td>maximum precipitation</td>
<td>0.4922</td>
</tr>
<tr>
<td>precipitation seasonality</td>
<td>0.0924</td>
</tr>
<tr>
<td>minimum precipitation</td>
<td><strong>0.0002</strong></td>
</tr>
<tr>
<td>aridity</td>
<td>0.3373</td>
</tr>
</tbody>
</table>

|                                               | 10 kb            |
| consecutive frost free days                   | 0.2751           |
| daylength                                     | 0.6952           |
| maximum temperature                           | 0.1151           |
| minimum temperature                           | 0.4414           |
| length of growing season                      | 0.3682           |
| consecutive cold days                         | 0.4693           |
| relative humidity                             | **0.0033**       |
| photosynthetically active radiation           | 0.2992           |
| temperature seasonality                       | 0.4812           |
| maximum precipitation                         | 0.4503           |
| precipitation seasonality                     | 0.0968           |
| minimum precipitation                         | **0.0011**       |
| aridity                                       | 0.3855           |
**Supplementary Table 4** Summary of simple regressions of flow cytometry estimates.

<table>
<thead>
<tr>
<th></th>
<th>$R$</th>
<th>$R^2$</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>45S rDNA</td>
<td>0.62</td>
<td>0.39</td>
<td>5.6 \times 10^{-15}</td>
</tr>
<tr>
<td>5S rDNA</td>
<td>0.28</td>
<td>0.078</td>
<td>1.4 \times 10^{-3}</td>
</tr>
<tr>
<td>Centromeres</td>
<td>0.14</td>
<td>0.019</td>
<td>0.12</td>
</tr>
<tr>
<td>TEs</td>
<td>0.15</td>
<td>0.023</td>
<td>0.086</td>
</tr>
</tbody>
</table>

**Supplementary Table 5** Genetic and geographic distance between world-wide populations. The distance between the population (in km) is above the diagonal, and $F_{ST}$ is below.

<table>
<thead>
<tr>
<th>N. Sweden</th>
<th>S. Sweden</th>
<th>Spain</th>
<th>S. Italy</th>
<th>Tübingen</th>
<th>S. Tyrol</th>
<th>E. Europe</th>
<th>Caucasus</th>
<th>Russia</th>
<th>C. Asia</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. Sweden</td>
<td>—</td>
<td>796</td>
<td>3072</td>
<td>2589</td>
<td>1701</td>
<td>1884</td>
<td>2118</td>
<td>3129</td>
<td>2355</td>
</tr>
<tr>
<td>S. Sweden</td>
<td>0.118</td>
<td>—</td>
<td>2361</td>
<td>1828</td>
<td>914</td>
<td>1092</td>
<td>1538</td>
<td>2907</td>
<td>2585</td>
</tr>
<tr>
<td>Spain</td>
<td>0.110</td>
<td>0.038</td>
<td>—</td>
<td>1822</td>
<td>1546</td>
<td>1571</td>
<td>2582</td>
<td>4302</td>
<td>4705</td>
</tr>
<tr>
<td>S. Italy</td>
<td>0.171</td>
<td>0.056</td>
<td>0.048</td>
<td>—</td>
<td>1113</td>
<td>836</td>
<td>985</td>
<td>2554</td>
<td>3291</td>
</tr>
<tr>
<td>Tübingen</td>
<td>0.161</td>
<td>0.036</td>
<td>0.040</td>
<td>0.044</td>
<td>—</td>
<td>280</td>
<td>1337</td>
<td>3035</td>
<td>3187</td>
</tr>
<tr>
<td>S. Tyrol</td>
<td>0.152</td>
<td>0.045</td>
<td>0.048</td>
<td>0.067</td>
<td>0.051</td>
<td>—</td>
<td>1139</td>
<td>2861</td>
<td>3152</td>
</tr>
<tr>
<td>E. Europe</td>
<td>0.130</td>
<td>0.027</td>
<td>0.029</td>
<td>0.030</td>
<td>0.029</td>
<td>0.027</td>
<td>—</td>
<td>1801</td>
<td>2369</td>
</tr>
<tr>
<td>Caucasus</td>
<td>0.158</td>
<td>0.057</td>
<td>0.039</td>
<td>0.043</td>
<td>0.055</td>
<td>0.064</td>
<td>0.029</td>
<td>—</td>
<td>1657</td>
</tr>
<tr>
<td>Russia</td>
<td>0.185</td>
<td>0.104</td>
<td>0.079</td>
<td>0.116</td>
<td>0.120</td>
<td>0.087</td>
<td>0.049</td>
<td>0.101</td>
<td>—</td>
</tr>
<tr>
<td>C. Asia</td>
<td>0.236</td>
<td>0.145</td>
<td>0.112</td>
<td>0.148</td>
<td>0.161</td>
<td>0.133</td>
<td>0.087</td>
<td>0.131</td>
<td>0.068</td>
</tr>
</tbody>
</table>
1 Genome sequencing

Genomic DNA was extracted from the roots of young Arabidopsis seedlings (5-8 plants for each line were pooled) grown on sterile 1/2 MS plates with 1% sugar at room temperature. Roots were ground to a fine powder in liquid nitrogen and mixed with 2X CTAB DNA extraction buffer (100 mM Tris-HCl pH 8, 1.4 M NaCl, 20 mM EDTA pH 8.0, 2% CTAB). After incubation at 65°C for 30 minutes, DNA was extracted with equal volume of chloroform and precipitated with 0.8 volume of isopropanol. Libraries were prepared using slightly modified Illumina Genomic DNA Sample Prep protocol. Briefly, DNA was fragmented by sonication with Bioruptor (Diagenode); the peak of fragment sizes was about 400 bp. End-repair of sheared DNA fragments, A-tailing and adapter ligation were carried out with NEBNext DNA Sample Prep Reagent Set 1 (BioLabs). Adapter-modified DNA was resolved on 2% low melt agarose (Peqlabs) gel (including SybrGold nucleic acid gel stain, Invitrogen), run for 90 minutes at 100 V. Library DNA was size-selected to 450–800 bp via gel extraction with a MinElute Gel Extraction Kit (Qiagen). The paired-end DNA libraries were amplified by PCR for 14–18 cycles with Illumina supplied PCR primers 1.1 and 1.2. Libraries were sequenced on Illumina GAII and HiSeq Analyzers using manufacturer’s standard cluster generation and sequencing protocols, with either 76 or 100 bases read length.

2 Polymorphism detection

2.1 Initial read mapping

2.1.1 Read mapping and SNP discovery

We first mapped all reads to the TAIR10 reference genome using BWA (version 0.5.9)\(^4\), allowing up to 4% mismatches and 1 gap. We tried trimming reads using different parameters before finally choosing the default parameters of BWA. After that, we used the rmdup function of Samtools (version 0.1.6)\(^5\) to remove reads that are duplicated in library preparations or sequencing. Since it turned out that not removing duplicated reads in highly duplicated regions, i.e., ribosome repeats and centromere repeats, improved the quality of coverage estimates (see Supplementary Fig. 23), we retained the reads that are potentially library duplications in those regions.

SNPs and indels were initially called using GATK\(^6\) with the default parameters of UnifiedGenotyper and IndelGenotyperV2, respectively. With these variants called, we run the GATK local realignment (function IndelRealigner) to refine the reads mapping in the presence of the variants. After that, we call SNPs using the pileup function of Samtools and run the varFilter function provided by samtools.pl in the Samtools package to filter low-quality calls. We found that the version of Samtools we were using had a small bug when setting heterozygosity to 0 (as is necessary for inbred lines). We worked around this by taking the SNP file provided by samtools (in pileup format) and counting the coverage of both alleles, then calling heterozygote when the minor allele count was at least 40% of the total. We do not make use of mpileup since we have our own population based pipeline (see below). Commands and detailed parameters used are listed in Supplementary Command Listing 1.

2.1.2 Quality calculations and filtering

The mapping quality (Q-value) calculated by BWA is intended to capture repetitiveness more than mismatches. For each base in the reference genome, we average the quality of all reads in all lines covering this position, resulting in a quality map. We then provide two versions of data: (i) the original version with Q-values attached; (ii) a filtered version where only Q ≥ 30 is retained (see Section 6 for details). The fraction of the genome retained in the filtered version is 87% and the fraction of SNPs and indels retained is 86% and 85%, respectively.

2.2 Reference-based structural variant discovery

We used several different reference-based methods to call indels and other structural variants (SVs), and we made use of population sharing to call variants in low-coverage regions. The different methods cover different sizes. We divide the indel/SV calling procedure into two phases: discovery and genotyping. The former is described in this section, and the latter in Section 2.3.

2.2.1 Short indels through local realignment

GATK calls short indels using local realignment. First-pass mapping will map different reads independently, which means that the same indel can be called in different ways for different reads within the same individual due to the mapper’s local optimization. Re-alignment based on the first-pass information can alleviate this problem.\(^6\) Thus, after running BWA, Samtools, and GATK, we use Samtools pileup and varFilter functions to call indels. Commands and detailed parameters are in Supplementary Command Listing 1.

It should be noted that we did not attempt to realign indels across lines. Instead, we simply filtered out with indels with more than two alleles.

2.2.2 Large SVs through paired-end reads

This method calls the big size structural variants based on abnormal read pairs. We ran BreakDancer\(^7\) version 0.0.1r81 with default settings (Supplementary Command Listing 1).

2.2.3 All sizes of SVs through split reads

This method calls SVs of all lengths by re-mapping the reads that cross the breakpoints of the events. We applied Pindel\(^8\)
BWA, Samtools and GATK (for each line)
% bwa aln -I -o 1 $ref $name.1.fastq > $name.1.sai
% bwa aln -I -o 1 $ref $name.2.fastq > $name.2.sai
% bwa sampe -a $insert -r $tag $ref $name.1.sai $name.2.sai $name.1.fastq $name.2.fastq
  -f $name.sam
% samtools view -bh -t $reflist -o $name.bam $name.sam
% samtools sort $name.bam $name.sort
% samtools index $name.sort.bam
% samtools rmdup $name.sort.bam $name.rmdup.bam
% samtools index $name.rmdup.bam
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T UnifiedGenotyper
  -I $name.rmdup.bam -o $name.gatk.snp.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T IndelGenotyperV2
  -I $name.rmdup.bam -o $name.gatk.indel.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -I $name.rmdup.bam
  -T RealignerTargetCreator -o $name.intervals -B:snps,VCF $name.gatk.snp.vcf
  -B:indels,VCF $name.gatk.indel.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -I $name.rmdup.bam
  -T IndelRealigner -o $name.realigned.bam -targetIntervals $name.intervals
% samtools sort $name.realigned.bam $name.realigned.sort
% samtools index $name.realigned.sort.bam
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T UnifiedGenotyper
  -I $name.realigned.sort.bam -o $name.gatk.realigned.snp.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T IndelGenotyperV2
  -I $name.realigned.sort.bam -o $name.gatk.realigned.indel.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -I $name.realigned.sort.bam
  -T RealignerTargetCreator -o $name.intervals -B:snps,VCF $name.gatk.realigned.snp.vcf
  -B:indels,VCF $name.gatk.realigned.indel.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -I $name.realigned.sort.bam
  -T IndelRealigner -o $name.realigned.sort.bam -targetIntervals $name.intervals
% samtools sort $name.realigned.sort.bam $name.realigned.sort
% samtools index $name.realigned.sort.bam
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T UnifiedGenotyper
  -I $name.realigned.sort.bam -o $name.gatk.realigned.snp.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T IndelGenotyperV2
  -I $name.realigned.sort.bam -o $name.gatk.realigned.indel.vcf
% samtools view -bh -t $reflist -o $name.bam $name.sam
% samtools sort $name.bam $name.sort
% samtools index $name.sort.bam
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T UnifiedGenotyper
  -I $name.realigned.sort.bam -o $name.gatk.realigned.snp.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T IndelGenotyperV2
  -I $name.realigned.sort.bam -o $name.gatk.realigned.indel.vcf
% samtools view -bh -t $reflist -o $name.bam $name.sam
% samtools sort $name.bam $name.sort
% samtools index $name.sort.bam
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T UnifiedGenotyper
  -I $name.realigned.sort.bam -o $name.gatk.realigned.snp.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T IndelGenotyperV2
  -I $name.realigned.sort.bam -o $name.gatk.realigned.indel.vcf

Pindel (run for each chromosome separately)
% pindel -f $ref -i realigned.bam_config_files -c Chr$i -w 0.5 -u 0.1 -e 0.1 -b breakDancer_output
  -o output_Chr$i -Q confirm_Chr$i.txt

BreakDancer (for each individual)
%perl /path/to/breakdancer/bam2cfg.pl $name.sort.bam > $out_folder/bkdancer/$name.cfg
%perl /path/to/breakdancer/BreakDancerMax.pl $out_folder/bkdancer/$name.cfg
  > $out_folder/bkdancer/$name.bkd

TE-Locate
%perl TEhierarchy.pl TAIR/TAIR10_GFF3_transposable_element.gff TAIR/family2superfamily.dat Alias
%perl TE_locate.pl 9 SAM/ TAIR/TAIR10_GFF3_transposable_element_HL.gff ref/at.fa TE 1000 5 1
  > temp.out 2 >&1

LAE-finder
  for all BAM files:
% LAE-finder nothing SVs_02.dat data-folder/$name.realigned.sort.bam
% LAE-finder filterAndSVcall SVs_02.dat data-folder/$name.realigned.sort.bam
  followed by:
% perl groupINVs.pl
% perl groupTLs.pl

Mach
  Run through a custom file reformatting and submission script that is available online.

Supplementary Command Listing 1 Command lines used in sequencing pipeline.

2.2.4 Copy number variants through coverage

We called CNVs by calculating the coverage directly after the BWA reads mapping and Samtools pileup. To avoid bias by factors like GC content9, the coverage was normalized. Coverage was estimated by summing the coverage in 1 kb windows and then normalizing them using the total coverage in the surrounding 3 Mb window. Regions with extremely high coverage, e.g., centromeric or rDNA repeats were not used in the normalization.

Nature Genetics: doi:10.1038/ng.2678
Windows were classified as have more copies than the reference genome if the estimated coverage was at least two-fold higher than background in at least 5 individuals. Regions with zero coverage (i.e., no copies, where the reference genome has one) were also called from coverage, but we also required paired-end reads spanning the putative breakpoint, or sharing in at least 5 individuals. Software is available upon request.

2.2.5 Segregating deletions

Given the apparent shrinkage of the *A. thaliana* genome\(^{10}\), it made sense to focus on finding segregating deletions. If the deleted sequence is present in the reference genome, then the polymorphism should be found using the method in the previous section, but if the deleted sequence is not in the reference genome, then we are looking for additional sequence present in some of the other *A. thaliana* lines, and perhaps also in the outgroup species *A. lyrata*.

With this in mind, we took reads that could not be mapped to the *A. thaliana* reference genome, and mapped them onto the *A. lyrata* genome. Once we found a contiguous region this way, we mapped the start and end coordinates back to the *A. thaliana* genome. If they mapped to a homologous region, we concluded that this is a genuine segregating deletion. Software is available upon request.

2.2.6 TE polymorphism

Additional TEs copies not present in the reference genome were identified by looking for paired-end reads with one end in an existing TE, the other in a unique sequence. The software used is described elsewhere\(^ {11}\). Copy number variation was analyzed at the level of superfamily (using the nomenclature in The Gypsy Database (GyDB) of Mobile Genetic Elements).

2.2.7 Large ancestral events

Very large structural rearrangements can be quite difficult to discover. However, we took advantage of the fact that many structural rearrangements separating the *A. thaliana* and *A. lyrata* reference genomes are segregating in the former species, and focused on finding these. An tool, LAE-finder, was developed for this purpose, in particular for finding inversions and translocations. The program, which is available for download, collects breakpoint information from the divergence data, and checks for presence-absence of these using paired-end reads.

2.3 Population-based SNP/indel genotyping

There is a general trade-off when calling SNPs and indels using short reads. If we call the variants aggressively, there will be many false positives; on the other hand, conservative parameters lead to unnecessary false negatives. The default setting of Samtools is quite conservative, at least for our data where the density of variants is high. In this project, we try to minimize the trade-off by separating the pipeline into two phases: discovery and genotyping. During the discovery phase, we discover variants based on the individuals that have good coverage and read mapping quality, and during genotyping, we call the discovered variants with less conservative thresholds.

2.3.1 Recovering low coverage regions

We composed a list of relatively high-quality SNPs discovered by GATK and Samtools as described above, and then revisited the pileup files and genotype SNPs that were filtered out or not called due to local low coverage. Our criterion for calling a SNP was that it was supported by one read, and that the allele was present in at least 5 other individuals. For indels, we also genotyped alternative (i.e., non-reference) alleles from the consensus sequences that have at least only one read, but added the restriction that the number of reads supporting this event should be greater than the number of reads that did not. The results after this step comprise the original of SNPs/indels used in imputation.

2.3.2 Imputation of SNPs and short indels

In addition to allele sharing, we also leveraged LD across markers. We used MaCH\(^{12}\) version 1.0 to impute SNP and short indel sites without sequence coverage. The result is the imputed version of released data (see Section 6, which is primarily intended for GWAS. Before imputation, all heterozygous calls were converted into missing calls, and only SNPs with exactly 2 alleles were kept. The latter step resulted in the removal of 281,942 SNPs.

As input to MaCH, we encoded each line as a homozygous diploid individual in the Merlin format. We then carried out imputation in windows of 20,000 markers, with an overlap of 2,000 markers between consecutive windows. 30 iterations of MCMC were used for each window, and the resulting probabilistic genotypes were converted into homozygous calls at each position.

2.3.3 Large SVs

Large (> 200 bp) SVs were called using several different methods (section 2.2). Since methods have different resolution, we tolerated a 10% shift of breakpoints when combining events called from different individuals or methods. In the final dataset, an event was accepted if it is: (i) called by multiple pipelines, or; (ii) supported by at least 5% of the individuals (using a single method).

Given the inaccuracies inherent in calling large SVs, we did not try to impute them.

2.4 Error estimates and quality control

2.4.1 Re-sequencing the reference genome

The reference line, Col-0, was re-sequenced using the same methods as for our Swedish sample. Under the assumption that all variants called are errors, and restricting ourselves to regions with Q-value ≥ 30, we estimate false-positive rates of 0.21% and 1.9% for SNPs and indels, respectively. The estimated rates decrease dramatically as the quality of the mapping increases (Supplementary Fig. 18). Putatively false positive SNPs are aggregated near the centromeres (Supplementary Fig. 19), and tend to have low Q-values.

2.4.2 Array-based SNP genotyping

In contrast to reference re-sequencing, SNP-chip data provides estimates of the false negative rate (i.e., the rate at which we fail to discover SNPs), and the genotyping error rate (i.e., the rate at which we make the wrong call for the ones we did detect). The overlap between the previously published SNP data\(^ {13}\) and
our data was 173 lines. As expected, both rates decreased with
the alignment Q-value (Supplementary Fig. 20). Restricting
ourselves to regions with Q-value ≥ 30, we found that we failed
to discover 4.6% of SNPs, and made the wrong call for 3.5% of
discovered SNPs. These estimates are conservative in that we
ignore errors in the SNP data14.

2.4.3 Sanger sequencing of PCR products
A subset of 45 of our lines overlapped an old data set of
close to 1,500 manually curated multiple-alignments of Sanger-
sequenced PCR-amplicons from 95 lines15. Since these regions
were PCR-amplified, they do not represent a random sample for
the genome, but they are useful nonetheless in that the quality
of the data is extremely high. After eliminating complex regions
with overlapping SNPs and indels, we estimated false positive
and false negative rates for SNPs to be 1.8% and 3.7%, respec-
tively, and the overall genotyping error rate (conditional on dis-
covery) to be 0.64%. For indels, corresponding rates were 1.3%,
6.9%, and 2.7%.

2.4.4 Sanger sequencing of random clones
To avoid the biases mentioned on the previous section, we also
sequenced randomly generated shotgun clones from a randomly
chosen accession. Genomic DNA from the roots of Arabidopsis
seedlings (DNeasy Plant Mini Kit, Qiagen) was sonicated (size
range 300–1200 bp), gel-extracted (size range 700–800 bp), ran-
domly cloned into the pJET1.2/blunt cloning vector (CloneJet
PCR Cloning Kit, Fermentas) and transformed into competent
E. coli cells. Plasmid DNA was isolated from overnight cul-
tures. Inserts were amplified with T7 Promoter Sequencing
primer and pJET1.2 Reverse Sequencing, and sequenced in both
directions (Applied Biosystems 3130xL Genetic). The resulting
chromatogram files were pre-processed as follows:

1. Remove vector sequence.
2. Apply sequence quality filter with a threshold of 0.0001
   using Richard Mott’s trimming algorithm (as implemented
   in CLC).
3. Eliminate reads shorter than 150 bp.
4. Align reads from complementary strands using SMALT
   (http://www.sanger.ac.uk/resources/software/smalt/).

All sequences were then aligned to the reference with BWA4.
The mapping properties of the data types are consistent, the
main difference being that Sanger reads that could not be aligned
uniquely, could more often be anchored, due to their greater
length (cf. Fig. 1a). Only sequences with a unique hit were
used to calculate error rates. SNPs and indels were called us-
ing Sambtools5 in the same way as for the main data (except that
there is no threshold for minimal coverage, i.e., a single Sanger
sequence is always sufficient).

The above procedure generated ∼250 kb of overlapping
Sanger and Illumina data. After trimming a further 5 bp from
the ends of each alignment to avoid artificial mismatches due
to alignment problems, error rates were calculated for the Q30
data, assuming that the Sanger result were perfect. This analysis
quickly revealed that two further steps of data filtering were re-
quired to obtain reasonable error rates. First, we eliminated pu-
atively heterozygous polymorphisms. Although some tracts of
genuine heterozygosity were observed in the data, the vast ma-
jority of heterozygous calls were shared between lines, which
is extremely unlikely in a highly selfing species. Thus, most of
these calls are dubious. Second, for indels, we eliminated mono-
or di-nucleotide repeats, as these are very difficult to sequence
(and are at least as likely to be called incorrectly in the Sanger
data). After these filtering steps (released with the data, see Sec-
tion 6), the error rates were comparable to the ones described
above, except for the indel false negative rate (Supplementary
Table 1). Note that we defined this rate as fraction of indels ob-
served using the Sanger data that had not been observed using
the Illumina data in any other line (if it had been observed in
lines other than the right one, it would be a genotyping error).
The indels we fail to call are thus most likely singletons.

3 De novo assembly

3.1 Assembly pipeline

We performed de novo assembly of the lines using two sets of
tools: SOAPDenovo16 v1.05 and clc_novo_assemble in the
4.0.1beta version of CLC Assembly Cell, the command-line
backend to the CLCGenomics Workbench. For both tools, we
used the raw fastq files from each line as input. For SOAPDen-
ovo, we first tested a variety of different parameter options on 3
lines. Based on these results, we decided to carry out assembly
of all lines at three different k-mer size setting: 27, 33 and 41.
For other options, we used map_len=32, pair_num_cutoff=2, and
avg_ins (average insert size for paired end reads) estimated from
BreakDancer7 in the configuration file. We enabled all optional
procedures, including using reads to solve small repeats, remove
low-frequency K-mers, remove low frequency edges and intra-
scaffold gap closure. We used the defaults for all other options
and precompiled parameters. See also Supplementary Com-
mand Listing 2.

In order to choose the best k-mer setting for each line, the re-
sulting scaffolds were mapped back to the reference genome us-
ing BLAST, and the alignment results parsed to eliminate mul-
tiple hits. We used several criteria for this step: first, we only
considered blast hits at above 85% similarity; second, if the
same part of the contig aligned to multiple locations, we picked
the highest scoring one (which in BLAST usually corresponds
to the longest alignment); third, if two contigs mapped to the
same location, we picked the longer alignment. We then evalu-
ated the assemblies based on several criteria, the most important
ones being: the proportion of the reference genome covered; the
minimum length of scaffold to cover 50% of the genome (N50); and
whether (based on manual inspection) the total length of the
scaffolds differed too much from the population average. Fi-
nally, we chose the best among the three sets of scaffolds created
using different k-mer settings as the assembly for that line.

For CLC assembly cell, we used the command line shown in
Supplementary Command Listing 2, with insert sizes again
estimated from BreakDancer. The results were comparable to
those from SOAPDenovo, and we will not discuss them further.
3.2 Detection of structural variants

To complement the read-pair and split-read approaches to structural variant detection, we used a novel method based on our \textit{de novo} assembly. Our method is similar in spirit to soapsv\textsuperscript{17}, based on alignment between scaffold and the reference genome, however, we focus on large events (>200 bp), which are poorly captured by standard read alignment algorithms.

3.2.1 Brief outline of algorithm

Our method detects structural variant breakpoints from irregularities in the alignment of scaffolds to the reference genome. For a region harboring a larger structural variant, we would expect to see the scaffolds covering the breakpoints to contain fragments from different regions on the reference genome, or different strands in the case of an inversion. Identifying such patterns, however, is complicated by the fact that, even in the absence of structural variation, a scaffold will sometimes map to multiple locations due to repetitiveness and small polymorphisms.

Our algorithm utilizes dynamic programming to search all possible alignments to identify scaffolds that contains genuine breakpoints, and further tries to discern the nature of each breakpoint, including where the flanking regions came from. The algorithm is still under development, and will be described in detail in a separate publication, however, the software implementation used here is available on request.

3.2.2 Quality of calls

When applied naïvely, our algorithm identified over 200,000 putative distinct breakpoints. However, when applied to our reference re-sequencing data, we obtained around 1/4 the average number of events for a line, indicating a false positive rate of at least 25%, and a rough analysis of the number of “missing” breakpoints in pair suggests a false negative rate of at least 20%. Thus, the results from this algorithm should be interpreted with care, and we include only small subset of the ones judged most reliable (most notably positive length variants) in the data release. All other types of events are released separately in the form of putative breakpoints (see Section 6). Although error rates are high, we note that our algorithm identifies several interesting examples of large structural variation that are readily validated by local patterns of LD, as illustrated in Supplementary Fig. 17.

3.3 Detection of novel sequence

Scaffolds and singleton contigs were first filtered by read coverage (those showing coverage less than 20% of average were eliminated) and then aligned to the reference using BLAST with the set of parameters described in previous section. All scaffolds, contigs, and sufficiently long (>100bp) parts thereof that aligned poorly (or not at all: the criterion was 80% similarity and a minimum BLAST score of 200) to the reference genome were extracted.

The resulting sequences were aligned to genomic sequences from the Refseq-genomic database in order to identify their origin. Any sequence that was found to map to non-plant genomes was considered to be the result of contamination.

15 lines exhibited an abnormally high amount of putative novel sequences. In some cases, this appeared to be due to contamination; in others it was probably due to a lower sequencing quality. After these lines were removed, the rest showed little sign of contamination. We assessed whether the remaining novel sequence was likely to be real \textit{A. thaliana} sequence in two ways. First, we asked whether a given segment was part of a scaffold or contig that also contained sequence that clearly did match the reference genome (using the BLAST criteria given above). Second, we aligned novel sequences from different lines against each other, and looked for sharing between lines (as would be expected for a segregating indel polymorphism). \textbf{Supplementary Fig. 21} summarizes the results of these analyses across all lines. We found 1.5–2.5 Mb of novel sequence per line, almost of which was either anchored in the reference genome or shared among at least 5 lines. As for the origin of the novel sequence, about 250 kb per line was clearly of plant origin (total combined BLAST score for the best linear alignment >400). In general, the greatest similarity was to \textit{A. lyrata}.

Overall, our attempts to identify large indels, identified many more positive (w.r.t. the reference genome) than negative length variants. If this difference were real, then it would imply that the reference genome comes from a line with unusually small genome. The alternative explanation is that it is simply due to bias: from a statistical point of view, it is easier to detect presence (i.e., novel sequence) than absence (missing reads, which could be due to chance). This explanation is supported by the observation that the spatial distribution of novel sequence along the chromosomes closely mirrors that of missing coverage (Fig. 1f and \textbf{Supplementary Fig. 2}).

We examined some polymorphisms in detail by aligning the scaffolds spanning the region with the homologous regions from the \textit{A. thaliana} and \textit{A. lyrata} reference genomes using mafft\textsuperscript{18} (\textbf{Supplementary Command Listing 2}). An example can be seen in \textbf{Fig. 1e} in the main text.

We tested the hypothesis that NB-LRR and F-box proteins, two gene families with high birth and death rates, contribute disproportionately to the novel sequences. First, we used Hmmer (v3.0)\textsuperscript{19} to create hmm profiles using multiple alignments
of the NBS and F-box domains in Arabidopsis thaliana with assistance of previously published sequence of the domains. Then, we searched in the novel sequences for the motifs (hmm-search). We detected similar motif/length of sequence ratio using the reference genome as using either gene family (no significant enrichment). One caveat is that sequences containing the motifs could potentially have been removed due to similarity to the reference genome in the filtering steps for identifying novel sequence.

4 Variation in nuclear DNA content

4.1 Flow cytometry

Flow cytometry was carried out on 129 lines (128 Swedish plus Col-0) split into two sets with 11 overlapping lines. Each set was further divided into three blocks of replicates that were measured on a different days with 1–2 biological replicates per line within each block. In addition, a set of 36 world-wide lines (plus a single line overlapping the set of 129) divided into three blocks of replicates (no replication within blocks) were measured on different days.

Seeds were stratified directly on soil for 5 days. Plants were grown under long day conditions (16 h light at 21°C, and 8 h dark at 16°C), watered twice a week, and rotated within trays daily. At 2 weeks the first two true leaves of each plant were finely chopped with a razor blade together with an approximately 0.125cm² piece of leaf of the internal standard, Solanum lycopersicum cv. Stupicke (2C = 1.96 pg DNA). In 250 µl of extraction buffer (kit PARTEC CyStain PI Absolute P no. 05-5022), 1 ml of staining solution (with 6 µl of propidium iodine (PI) and 3 µl of RNAse [3.33 mg/ml]) was added, and the resulting suspension was passed through a 30-micron filter (Partec CellTrics no. 04-0042-2316). Samples were stored for 2–4 h at 4°C in the dark prior to DNA content evaluation.

Genome size was measured with a LSRFortessa special order research product equipped with a 561 nm yellow-green laser (110 mW) and a 488 nm blue laser (100 mW), for PI (610/20 nm) and side scattering (SSC; 488/10 nm) detection, respectively. Events representing debris were excluded by selecting only the major cluster when plotting the PI-area versus SSC-area for 10,000 events. Data was analyzed with the flowClust R package. The mean position of the 2C peak for each sample was normalized to the 2C peak of the internal standard and converted into base pairs.

Simple linear regression models were fitted for each set in order to account for the block effect and obtain a single flow cytometry estimate for each line. The mean and standard deviation of these estimates are available online (see Section 6). The results were generally highly reproducible, with the standard deviation being on the order of a percent of the mean. The distribution of estimates is shown in Supplementary Fig. 3.

4.2 Repeat-number estimation through coverage

Sequence coverage for each individual was calculated by summing normalized read coverage in 1 kb windows (as described in Section 2.2.4) across the entire genome. Note that, for this analysis, we did not remove reads that were supposedly due to library duplication, as this seemed to removed actual repeats (leading to poorer agreement with flow cytometry estimates, see Supplementary Fig. 23). The contribution to genome size by 45S rDNA, 5S rDNA and centromeric repeat elements was estimated by summing read coverage across the appropriate regions of the reference sequence. For 45S rDNA, the two ~10 Kb 45S rDNA loci, in the beginning of chromosome 2 and at 14.2 Mb of chromosome 3, respectively, were considered. For 5S rDNA, the locations were determined by BLAST-ing the transcribed region consensus sequences identified for the major and minor loci on chromosomes 4 and 5, as well as loci 1, 2 and 3 on chromosome 3. Similarly, centromeric regions were located via BLAST with two centromeric variants, clones 22_At178 (GenBank: EU359499.1) and AS1 (GenBank: X04320.1). For TEs, the total count of novel TEs was used (see Section 2.2.6). All estimates are available online (see Section 6). The correlation between the flow cytometry and repeat copy number estimates was analyzed using standard regression methods as described in the text. The results are presented in Table 1, Supplementary Table 4, Fig. 2a, and Supplementary Fig. 22.

4.3 Estimating rDNA copy number through qPCR

45S rDNA copy number can also be estimated through quantitative PCR (qPCR) of either the 18S or 25S subunit. We carried out qPCR in technical triplicate and biological replicate for each of five A. thaliana lines (ids: 5856, 6099, 6136, 6244, 8386) in a 25 µl total reaction volume using 2X SensiMix SYBR & Fluorescein Kit (Bioline No. QT615-05).

An iQ5 light cycler (Bio-Rad) was employed with the following thermal profile: 95°C for 600 seconds; 40 cycles at 95°C for 10 seconds, 60°C for 30 seconds and 72°C for 30 seconds; and a final cycle at 72°C for 60 seconds. One standard curve based on serial 10-fold dilutions was made for each sample. No primer dimmers were detected in the melting curve.

45S ribosomal DNA (rDNA) abundance was estimated by comparing 18S rDNA to two single copy genes (At3g18780, At178; see relevant file in Section 6 for a list of primers) according to:

\[
\text{rDNA abundance} = 2^{Ct(\text{single copy gene})} - Ct(18S \text{ rDNA})
\]

where Ct(x) stands for the threshold cycle for x. Estimates for all lines were then normalized to the line with the lowest estimated copy number of 18S rDNA. The qPCR results showed excellent agreement with the coverage-based estimates (Supplementary Fig. 23c–d).

4.4 Genome-wide association mapping

Genome-wide association mapping was carried out using the imputed SNP data (Section 6) using an approximation of the kinship model. Minor alleles below 5% MAF were filtered out prior to the analysis, leaving around 1.8M SNPs. Both flow cytometry and 45S copy numbers estimates were used as phenotype, and we tried different subsamples of the lines in order to evaluate the robustness of the association (Supplementary Fig. 5). Analysis of the northern and southern lines separately demonstrated that the association is due to variation among the northern lines: the association is not present in the southern sample (Supplementary Fig. 5a–c). Association mapping using 45S copy number directly as a phenotype revealed that, while
the peak is still present, it is much less distinct (Supplementary Fig. 5d). Furthermore, when we increase the sample size to the full sample for which we have sequence data, the peak vanishes (Supplementary Fig. 5e), which is troubling. However, the 45S-rDNA coverage data vary greatly between lines and replicates, and may be strongly affected by both alignment and sequencing artifacts. Thus, counter-intuitively, the flow cytometry data may actually be a better estimate of the true number of 45S-rDNA repeat copies.

4.5 FISH

An obvious explanation for putatively trans GWAS peaks related to 45S copy number is that they are linked to novel 45S rDNA clusters (i.e., they are, in fact, cis). Thus, although the 100 kb region on chromosome 1 that contains the most significant associations does not show any evidence for large structural variants, we decided to look genome wide. First, TE-Locate was used with the standard settings to map novel 45S DNA repeat insertions, and none were found. Second, we used Fluorescence In Situ Hybridization (FISH) to look for clusters directly.

Actively growing, young roots were pretreated with ice-cold water for 12 hrs, fixed in ethanol:acetic acid (3:1) at 4°C for 24 hrs. Selected root tips were rinsed in distilled water and citrate buffer (10 mM sodium citrate, pH 4.8), and digested by 0.3% cellulase, cytohelicase and pectolyase (all Sigma-Aldrich) for 24 hrs. Selected root tips were rinsed in distilled water and citrate buffer at 37°C for 90 min. Individual root tips were dissected in ca. 10 µl of acetic acid on a microscopic slide. The cell material was covered with a cover slip, evenly spread by tapping, and the slide gently heated over a flame. Then the slide was frozen in liquid nitrogen, cover slip flicked off, fixed in ethanol:acetic acid (3:1) and air-dried. A. thaliana BAC clone T15P10 (AF167571) containing 45S rRNA genes was used to identify 45S rDNA loci. BAC clones F5A8, F1N21, F12A21, T23K23, and F12B7 were used to paint chromosome 1, and localize the candidate region. All DNA probes were labeled either with biotin-dUTP, digoxigenin-dUTP, or Cy3-dUTP by nick translation, pooled, ethanol precipitated and pipetted on ready-to-use slides. The slides were heated to 80°C for 2 min and incubated at 37°C overnight. Hybridized DNA probes were visualized either as the direct fluorescence of Cy3-dUTP (yellow) or through fluorescently labeled antibodies against biotin-dUTP (red) and digoxigenin-dUTP (green). DNA labeling and fluorescence signal detection was carried out using a previously published step-by-step protocol. Chromosomes were counterstained with 4,6-diamidino-2-phenylindole (2 µg/ml) in Vectashield (Vector Laboratories). Fluorescence signals were analyzed and photographed using a Zeiss Axioimager epifluorescence microscope and a CoolCube camera (MetaSystems), and pseudocolored/inverted using Adobe Photoshop CS2 software (Adobe Systems).

Results for three accessions with large estimated 45S rDNA copy number are shown in Supplementary Fig. 4. The known clusters on chromosomes 2 and 4 are clearly visible, and there is no evidence for any other clusters.

5 Population genetic analyses

5.1 Selection on indels

All polymorphisms were annotated with respect to function, using the reference genome. As expected under the assumption that most mutations disrupt function, structural variants are relatively more common outside genes, and relatively rare in exons.

In order to test whether selection is driving deletions to fixation, as has been suggested, we used the global alignment between A. thaliana and A. lyrata to determine the ancestral state of SNPs and indels. The derived allele frequency distribution can be a powerful tool when looking for signal of selection, essentially because it makes it possible to identify the high-frequency derived allele that are expected to be rare unless selectively favored. Very conservative criteria were used in the polarization, the reason being that the conclusions may be severely biased if alleles are misclassified with respect to ancestral status. Because derived alleles are almost always rare, misclassification of them as ancestral will cause a large inflation of supposedly high-frequency derived alleles. With this in mind, a polarization was accepted only if the identity of surrounding region passed stringent criteria. For SNPs, we required that the identity of the surrounding 30 bp window should be at least 90%. Short indels (≤ 200 bp) were classified as deletions if A. lyrata had what appeared to be the longer allele, and the additional sequence was identical in all carriers (including A. lyrata). Indels were classified as insertions if A. lyrata had the shorter allele and the additional, putatively inserted sequence was identical in all carriers. We did not try to polarize other structural variants.

Given that the overall divergence between the two reference genomes is greater than 10%, the above criteria are very stringent. The estimated allele frequency distribution for SNPs, deletions, and insertions is shown in Supplementary Figure 6. Contrary to previous observations, there is no evidence for an excess of high-frequency deletions. It is not clear what to conclude from this. While our analysis is based on many more events, it is worrisome that we polarize only 18% of all observed events. Experimenting with less conservative criteria demonstrated that the allele frequency distribution is extremely sensitive to the procedure used, however, in no cases did we get results consistent with strong selection favoring deletions (not shown).

For the SNP-based selective sweep analysis in Section 5.4, we employed no filtering when estimating ancestral state. Both polarizations, as well as the function annotation, are part of the released data (Section 6).

5.2 Long-range LD

5.2.1 LD in structured populations

Let denote the number of individuals and the number of SNPs. Given a genotype matrix in which each of the SNP vectors, , is coded numerically, i.e., 0, 1, or 2 for each of the diploid individuals, we can obtain normalized SNPs . With mean 0 and variance 1. Under the assumption that these SNPs are independently sampled from a distribution with covariance matrix , we can ob-
tain an unbiased estimate of this covariance matrix as

\[ \text{Cov}_{est}(X_i) = \hat{V} = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})(X_i - \bar{X})^\top. \] (1)

If we assume that the average allele value for each accession is equal (i.e., alleles are labeled so that all accessions have roughly equal numbers of 0 and 1 alleles), then this covariance matrix estimate is exactly Fisher's correlation kinship matrix. Like Mangin et al., we then obtain pseudo-SNPs, \( T_i = \hat{V}^{-1/2} X_i \), with values that are expected to be independent across individuals. Now we can proceed to obtain \( r^2 \) values that have been corrected for genetic correlations between individuals, or, in other words, population structure. The correlation becomes strikingly simple when written out as

\[ \text{Cor}_{est}(T_i, T_j) = \frac{\text{Cov}_{est}(T_i, T_j)}{\sqrt{\text{Var}_{est}(T_i)\text{Var}_{est}(T_j)}}, \] (2)

where

\[ \text{Var}_{est}(T_i) = \frac{1}{m}(T_i - \frac{1}{m} \sum_k T_{ik})^\top (T_i - \frac{1}{m} \sum_k T_{ik}), \] (3)

and

\[ \text{Cov}_{est}(T_i) = \frac{1}{m}(T_i - \frac{1}{m} \sum_k T_{ik})^\top (T_j - \frac{1}{m} \sum_k T_{jk}). \] (4)

If the genotype data does not contain any missing values then we can speed up the calculation for the adjusted correlation by calculating the pseudo-SNPs and normalizing them beforehand. Consider the normalized pseudo-SNP

\[ W_i = \frac{T_i - \frac{1}{m} \sum_k T_{ik}}{\sqrt{\text{Var}_{est}(T_i)}}. \] (5)

We can now obtain the correlation estimate by simple vector multiplication

\[ \text{Cor}_{est}(T_i, T_j) = W_i^\top W_j, \] (6)

and hence \( r^2(T_i, T_j) = \text{Cor}_{est}(T_i, T_j)^2 \). For data with no missing values (or with missing data imputed), the time complexity for estimating \( r^2 \) for all SNP pairs is \( O(mn^3 + m^2n) \). An obvious extension to this would be to use some of the approximations proposed by Lippert et al., to obtain an approximate \( r^2 \) in sub-cubic \( n \) time.

The transformed LD estimates are generally lower than the original ones, since the inflation caused by population structure is removed (Supplementary Fig. 24), however, large numbers of long-range LD pairs remain (Fig. 4a).

5.2.2 Potential causes of long-range LD

Major population subdivision Since the divergence between north and south in our sample is substantial (see Section 5.3), we evaluated its effect on LD separately, by applying the relatedness correction separately to the northern and southern subsamples. The LD pattern in the north contain many more high \( r^2 \) pairs (Supplementary Fig. 25a), a result of smaller sample size as well as higher relatedness, whereas the LD pattern in the South looks very similar to that of the full sample (Supplementary Fig. 25b). Thus we conclude that the vast excess of long-range LD is not simply due to the north-south division.

Imputation We used imputed data for the LD calculation, primarily to speed up computation, but also to make sample size even across all pairs. To ensure that this did not cause the pattern observed, we re-calculated \( r^2 \) for all high-LD pairs in the unimputed data. The results show that imputation is clearly not a source of long-range LD (Supplementary Fig. 25c-d).

Other artifacts Long-range LD might also various kinds of mapping and genotyping artifacts, i.e., the SNP loci do not segregate normally. To try to eliminate these problems, we applied several stringent filtering steps. To begin with, all analysis was based on the high-quality Q30 SNPs. To be even more stringent, we aligned 75 bp surrounding each SNP to the reference genome using BLAST, using a less stringent criterion (90% similarity) than was used in the original read mapping. Any SNP that could be aligned to more than one region on the reference genome was filtered out. This should remove simple mapping artifacts.

After transforming and filtering, we were still left with over 70,000 SNP pairs exhibiting strong long-range LD, especially between centromeric regions (Fig. 4a). To test whether they segregated normally, 4 centromeric and 2 non-centromeric set of SNPs were genotyped in informative F2 crosses (Supplementary Table 2). Two informative crosses (6035 × 9433 and 6136 × 6064) were carried out, with 10 F2 seedlings genotyped using PCR and dideoxy-sequencing. Of the four centromeric pairs, only one yielded reliable PCR fragments for both SNPs (Supplementary Table 2). In summary, 2 out of 2 between-chromosome, non-centromeric comparisons segregated normally, and 1 out of 1 between-chromosome, centromeric comparison did not.

Based on these results, we conservatively decided to remove all centromeric SNPs, leaving only 2509 pairs. To further test whether some of these might also be closely linked, we aligned short sequences (both 75 and 150 bp was tried) surrounding each SNP to the scaffolds generated by de novo assembly. When sequences flanking both SNP in a pair mapped near each other one the same scaffold, we considered them linked in that line. Out of the 2509 pairs tested, we found 17 that could be due to this kind of structural variation (Fig. 4b).

Selection For the remaining long-range LD pairs, we first tested whether the corresponding pairs of loci were overrepresented in published protein interaction data. No significant overrepresentation was found. Next, we looked for overrepresentation of individual loci among those identified as having signals of local adaptation. We asked whether SNPs involved in long-range LD were close to the peak SNPs (those with \( p < 0.001 \)) for 13 climatic traits, and calculated the \( p \)-value using a permutation scheme that maintain the LD structure in the data. We detected significant (Bonferroni-corrected \( p < 0.05 \)) enrichment in two of the traits, relative humidity and minimum precipitation (Supplementary Table 3). The same method was used to test for overlap with SNPs implicated in selective sweeps (Section 5.4). In northern Sweden, there is a significant overrepresentation of long-range LD SNPs and SNPs within 1 kb of a SNP with a Sweepfinder CLR above 50 (\( p = 0.0336 \)).
5.3 Population structure

5.3.1 Analysis

Global population structure in A. thaliana has been described several times\textsuperscript{13,15,34}. We compared our Swedish sample (which we divided into a northern and a southern population based on latitude 60° N, see Supplementary Fig. 1) with other samples for which whole-genome sequencing data are available, namely the 8 smaller populations (10 individuals in each) sequenced by Cao et al.\textsuperscript{3}. Three different statistics were calculated: PCA, t-SNE, and $F_{ST}$. The results confirmed the distinctiveness of the northern Swedish sample (Supplementary Table 5)\textsuperscript{35}. Clustering based on the 250k SNP data\textsuperscript{13} identified two likely contaminants among the northern lines: 6180 and 1435. These two lines clearly cluster with southern (or even other European) lines, and were excluded from further analysis. These conclusions are supported by standard hierarchical clustering analysis as well (Supplementary Fig. 8).

To further characterize the pattern of variation in Sweden, we plotted the sequence divergence between all pairs of lines in our samples as a function of the distance separating the original sampling locations (Supplementary Fig. 9), and tested for isolation by distance using the Mantel test (function mantel of the R-package vegan). We found a significant correlation within both the northern and the southern sample, although the correlation within northern Sweden was much stronger (Spearman’s $r=0.6109$, $p<0.001$ and $r=0.4525$, $p<0.001$, respectively). The 95% bootstrap confidence interval for the difference in Spearman’s $r$ between north and south was (0.115, 0.173). The north also has lower levels of polymorphism and higher Tajima’s $D$\textsuperscript{36} (Supplementary Fig. 15), as well as more extensive LD (Supplementary Fig. 7). Taken together, these results are consistent with the observation that the north seem to have a much more patchy population structure, with quite small local populations.

5.3.2 Methods

$F_{ST}$ We used only bi-allelic SNPs for which we had complete information for all 260 (180+80) lines. Each SNP was coded as 0 for the major allele, and 1 for the minor allele. The standard method was used\textsuperscript{37}. In addition to genome-wide averages, we also calculated $F_{ST}$ between northern and southern Sweden in non-overlapping windows of 100 kb.

$\theta$, $\pi$, and $D$. We calculated three standard statistics describing aspects of nucleotide diversity, separately for north and south, and in 100 kb windows across the genome: Watterson’s $\theta$; nucleotide diversity, $\pi$; and Tajima’s $D$. All three statistics are intended for complete sequence data rather than SNPs, and we therefore used BamTable\textsuperscript{38}, which essentially tries to integrate over the uncertainty in polymorphism detection, generating probabilistic calls. We ran BamTable on sorted BAM files to call SNPs, using the standard options. The most common base was called at each site and each line. Bases that were supported by less than 5 reads were excluded. If two different bases were supported with more than 4 reads for a certain site, then this site was excluded as potentially due to incorrect alignment or genuine heterozygosity site.

In order to accommodate missing data, we calculated the summary statistics separately for all sample sizes, then carried out weighted averaging similar to what has previously been suggested for Watterson’s $\theta$\textsuperscript{39}. For the other two statistics, we use

$$\pi = \frac{\sum_{i=2}^{n} L_i \pi_i}{\sum_{i=2}^{n} L_i}$$

and

$$D = \frac{\sum_{i=2}^{n} L_i a_i D_i}{\sum_{i=2}^{n} L_i a_i},$$

where $\pi_i$ and $D_i$, $i = \{2, ..., 180\}$, are the values of the sample-size specific statistics, $L_i$ is the sequence length of sites with sample size $i$, and $a_i$ is the $i$th harmonic number. These estimates turned out to be superior in terms of bias and root mean squared error when compared to alternative formulas in neutral simulations with random missing data added in a way that reflects the observed data (results not shown).

5.4 Selective sweeps and local adaptation

5.4.1 Genome scans

CLR Sweeprider\textsuperscript{40} is intended for polarized SNPs, but can handle missing data. SNPs were polarized as described in Section 5.1, but without the conservative filtering. The 46% of SNPs that could not be polarized were nonetheless used in the analysis. The output of Sweeprider is a CLR (composite likelihood ratio) statistic for a grid of positions with distance of 1,000 bp between successive positions. To arrive at a significance threshold, we use the coalescent simulator msms\textsuperscript{41}, which also allows selection.

We used standard neutral simulations to determine the critical CLR values (Supplementary Command Listing 3). The population mutation rate was set to 0.005/bp, which corresponds to the average observed diversity. An average recombination rate of 4.6 cM/Mb was assumed, and a selving rate of 97%\textsuperscript{34}. There is no evidence that recombination rates in the sweep regions deviate strongly from this value, and Sweeprider has been shown to be relatively robust to deviations from the true recombination rate\textsuperscript{40}. Recombination and diversity were scaled to correspond to a sequence length of 1Mb. The sample size was set to the mean northern and southern sample size (averaged over all SNPs): 43 and 111, respectively. The total number of simulations was 12,000, corresponding to the simulation of about 100 whole genomes. The cut-off was then calculated for a family-wise error rate of 5% per genome, so that a false positive signal only occurs in one out of 20 whole genome analyses on average.

Simulations with selection at a single locus showed that a single selective sweep leads to multiple significant peaks in the CLR in about 80% of cases. In those cases, significant regions span an average of 170 kb, with a 99% quantile of 430 kb. The 95% confidence interval of the true position of the selected site is 160 kb centered around the largest peak. Thus, it is not unlikely that multiple significant peaks created by a single sweep are relatively far apart from each other, although the largest peak is almost always nearer to the true selected locus than smaller peaks. Therefore, we treat multiple peaks within a region of 430 kb as single events in our analyses, and designate the peak with the largest CLR as the center of the sweep.
msms (neutral model)
% msms -ms 111 12000 -t 5000 -r 994
msms (with selection)
% msms -ms 50 1000 -t 5000 -r 800 -SAA 1000 -SaA 500 -N 100000 -SF 0 -Sp .5

XP-CLR (repeat for chromosomes 2–5)
% XPCLR -xpclr SouthChr1.geno NorthChr1.geno Chr1.map Chr1.xpclr -wl 0.005 200 1000 1 -p1 0.95
% XPCLR -xpclr SouthChr1.geno NorthChr1.geno Chr1.map Chr1.xpclr -wl 0.005 200 1000 1 -p1 0.95

XP-EHH (repeat for chromosomes 2–5)
% xpehh -m Chr1.map -h SouthChr1.hap NorthChr1.hap > Chr1.xpehh

Supplementary Command Listing 3 Command lines used in sweep finding.

In total there were 22 significant sweep locations in the northern and 1 in the southern Swedish sample. The single significant southern Swedish signal overlaps with the strongest signal in the north. For each location, 160 kb large regions centered on the largest CLR peak were selected, and subjected to further analysis (see below).

The sweep haplotype corresponding to the shared sweep signal on chromosome 1 turned out to be associated with an intrachromosomal transposition of 278 kb to a new position 486 kb away. During the selective sweep, this configuration likely prevented any recombination event between the ancestral and rearranged haplotypes in 764 kb region. To ensure that the sweep signal was not simply due to repression of recombination, we ran Sweeplander on a data set where the entire 764 kb region was replaced by a single base pair. The sweep signal from this analysis was still the strongest in both northern Sweden and southern Sweden, and Sweeplander CLR value for northern Sweden was still 165 fold larger than the average CLR value in the rest of the genome.

XP-CLR and XP-EHH XP-CLR was calculated between the northern and the southern populations, looking for sweeps in the north with the south as reference, as well as vice versa. XP-EHH (downloaded from http://hgdp.uchicago.edu/Software/) just returns a single value for the comparison of the two populations. Since it cannot handle missing data, all SNPs with missing individuals were removed.

5.4.2 Environmental correlations and GO terms
A file available online (see Section 6) summarizes the 22 significant sweep regions. For each region, GO terms were collected, and a test for GO category enrichment was carried out using func(1). Terms that reached reached a significance threshold of \( p < 0.01 \) and had at least three genes in the category are reported. Only biological process (GO:0008150) sub-categories were used.

The data from Hancock et al. were used to look for enrichment of significant SNP-environment correlations in the swept regions. For each environmental variable (aridity, consecutive cold days, consecutive frost-free days, day-length, length of growing season, maximum precipitation, maximum temperature, minimum precipitation, minimum temperature, PAR, precipitation seasonality, relative humidity and temperature seasonality), the 1% tail of largest correlation coefficients was selected and tested for enrichment in each sweep region. The \( p \)-value was calculated by deriving an empirical null distribution of the proportion of tail signals (number of SNPs that are in the tail of the correlation coefficient distribution, divided by all the SNPs in that interval) for a randomly placed interval. The \( p \)-value is the probability of having an as high or higher proportion of tail signals in a randomly placed interval compared to the actual sweep region.

5.4.3 Dating the sweep
The sweep was dated utilizing the average amount of polymorphism separating two sweep haplotypes. To do this we utilized sequence data from a segment within the transposition (20.35–20.45Mb). To account for the fact that we only use SNPs without missing data for the age estimation, we reduce the sequence length by the ratio of the number of SNPs without missing data (1158) and the total number (2221), leading to an effective sequence length of 100000 × 1158/2221 = 52139 bp. The SNPs for which there were no missing individuals, and which were are also monomorphic in the six lines with the ancestral (unswept) configuration were then used to calculate the average number of differences for a randomly selected pair of northern and southern lines, respectively. These numbers were 12.7 and 31.6.

Assuming an approximately star-like tree, we estimate the age of the sweep by calculating the average divergence time of two sweep haplotypes. We do this by dividing the average number of differences with a factor of \( 2 \times 52139 \times 7 \times 10^{-9} \), where \( 7 \times 10^{-9} \) is the estimated mutation rate per bp and generation. This resulted in an average divergence time of 17,390 years for northern Sweden, and 43,282 years for southern Sweden.

6 Data release
The raw data has been deposit to NCBI trace SRA with accession number SRP012869. Processed data are available through the project download site.

6.1 Genotype files
In total we identified around 4.5M SNPs, 576k short indels, 23k transposable elements, 7.7k CNVs, as well as 3.8k other structural variants (larger than 200 bp). The following files are available:

- SNPs (Original and imputed)
- SNP mapping information
- SNP annotations
- Small indels (Original and imputed)
• Small indel annotations
• Large structural variants (several files from multiple pipelines)

6.2 Other files

The following files are also available from the website:

• list of lines used in this study (an interactive version is available).
• alignment and assembly statistics for all sequenced lines.
• flow cytometry and repeat copy number estimates for the 180 lines.
• flow cytometry estimates for 36 worldwide lines.
• multiple alignments for the candidate genes from Fig. 2.
• summaries of putative sweep regions.
• all PCR primers used in this study.
• predicted genotype for 1306 lines with respect to the transposition on chromosome 1, and the large inversion on chromosome 4 (Supplementary Fig. 17).
• all genes in the swept transposition on chromosome 1.

References


