doi:10.1038/nature10414

Supplementary Methods

Multiple Genomes and Transcriptomes of *Arabidopsis thaliana*

Xiangchao Gan[1§], Oliver Stegle[2§], Jonas Behr[3§], Joshua Steffen[4§], Philipp Drewe[3§], Katie L. Hildebrand[5], Rune Lyngsoe[6], Sebastian J. Schultheiss[3], Edward J. Osborne[4], Vipin T. Sreedharan[3], André Kahles[3], Regina Bohnert[3], Gèraldine Jean[3], Paul Derwent[7], Paul Kersey[7], Eric Belfield[8], Nicholas Harberd[8], Eric Kemen[9], Christopher Toomajian[5*], Paula X. Kover[10*], Richard M. Clark[4*], Gunnar Rätsch[3*], Richard Mott[1*]

[1] Wellcome Trust Centre for Human Genetics, University of Oxford, OX3 7BN, UK
[2] MPI for Intelligent Systems & MPI for Developmental Biology, Spemannstr. 38, 72076 Tübingen, Germany
[3] Friedrich Miescher Laboratory, Max Planck Society, Spemannstr. 39, 72076 Tübingen, Germany
[4] Dept Biology, University of Utah, Salt Lake City, 84112-5330 USA
[5] Dept Plant Pathology, Kansas State University, Manhattan KS 66502-5502, USA
[6] Dept of Statistics, University of Oxford, South Parks Rd, Oxford OX1 3TG, UK
[7] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
[8] Dept Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK
[9] The Sainsbury Laboratory, Norwich NR4 7UH, UK
[10] Dept of Biology & Biochemistry, University of Bath, Bath, BA2 7AY, UK

* Corresponding authors.
§ These authors made equal contributions.

Table of Contents

Note: Unless indicated otherwise, all analyses of the reference genome refer to the TAIR9 genome assembly and the TAIR10 annotation of Col-0, which are available at http://www.arabidopsis.org/. For simplicity, we refer only to genome as TAIR10 (no genome sequence changes were made for the TAIR10 annotation release).

1. Portals for Accessing the Data and Analyses

Web Site All assembled genome sequences, BAM files of genomic and transcript reads and annotations are downloadable as files from our web site http://mus.well.ox.ac.uk/19genomes.

ENA The Illumina genomic DNA reads are also available from the EBI ENA under accession number ERP000565.

GEO The Illumina RNA-seq reads are available from GEO under SuperSeries accession number GSE30814.

GBROWSE genome browser We created a public GBROWSE instance http://fml.mpg.de/gbrowse-19g containing all the assembled genome sequence, mapped RNA-seq data and annotations generated by this project.

Ensembl The genetic variation data are also available via EnsemblPlants http://plants.ensembl.org. The Ensembl variation model previously developed for vertebrate genomes[1] has been applied to the *Arabidopsis* resequencing data. Supported features include a query-oriented data warehouse, views of SNPs and indels from the perspective of genes, transcripts, proteins, and individuals, with programmatic access available via a Perl API or direct access to a public database server. This resource is being further developed with the inclusion of additional data from the Arabidopsis 1001 Genomes Project[2].

2. Sequencing of 18 *Arabidopsis* Genomes

2.1 Choice of biological material Our study focuses on 18 accessions (strains) that are, along with the reference accession Col-0, the parents of more than 700 Multiparent Advanced Generation Inter-Cross (MAGIC) lines. The design and generation of the MAGIC lines is described by Kover et al, 2009[3]. Briefly, the accessions are: Bur-0, Can-0, Ct-1, Edi-0, Hi-0, Kn-0, Ler-0, Mt-0, No-0, Oy-0, Po-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0, and Zu-0 (see also Supplementary Table 1). The MAGIC lines are conceptually similar to the Nested Association Mapping (NAM) population in maize[4], and the Collaborative Cross in mouse[5]. The MAGIC population is among the most advanced and extensive cross-based resource available for genetic mapping studies (in *A. thaliana* or otherwise). Informing these lines was the primary consideration in selecting the accessions for sequencing.

The 19 parents of the MAGIC lines were selected before extensive genetic polymorphism data was available to characterize population structure in the species (e.g., Nordborg *et al.,* 2005[6]), and were selected largely to maximize phenotypic variation, and for geographical diversity. Characterization of the 18 accessions by Kover *et al.,* 2009[3] and in the current study (e.g., the analyses described in Supplementary Information section 6), has shown that the lines are representative of both phenotypic and genotypic variation in the global *A. thaliana* population.

2.2 Relation to the A. thaliana 1001Genomes Project Our genomes contribute to the *A. thaliana* 1001Genomes Project[2] (http://www.1001genomes.org/), an effort by multiple groups within the Arabidopsis community, and funded by various sources, to provide dense sequence data for evolutionary and functional genomic studies. Our specific effort is limited to the parental accessions of the MAGIC lines (Supplementary Information section 2.1).

More generally the dense collection of genomic sequences and gene models we describe, as well as the methodological approaches, should be broadly applicable to other studies in *A. thaliana*.

2.3 Genome sequencing Seeds for all accessions were obtained from the Arabidopsis stock centre, and bulked for one generation in the Kover Lab. These stocks are from the same batch of seeds used for construction of the MAGIC lines[3]. DNA was extracted from leaves with DNEasy Kits (Qiagen™), and libraries were prepared using the Illumina Genomic DNA Sample Prep protocol with the cBot PE cluster generation kit (sequencing was performed with 36bp cycle sequencing kits v2, v3 or v4). For most accessions, two paired-end (PE) libraries with different insert sizes and read lengths (~200bp; 32bp PE and ~400bp; 51bp PE) were made from different plants and sequenced with a Genome Analyzer II by the Core Genomics Groups at the Welcome Trust Centre for Human Genetics (WTCHG); exceptions were one library for Mt-0 (GATC Biotech, Germany) and one library of Ws-0 (John Innes Centre, Norwich UK). In total the combined average coverage across accessions of reads aligned to the reference genome (i.e., chromosomes 1-5 plus organelles) was ~46x. Supplementary Table 1 gives the detailed breakdown for each accession, including library-specific values. The sequence reads are available from our web site http://mus.well.ox.ac.uk/19genomes and from the EBI SRA under accession ERP000565.

3. Genome Assembly

3.1 Overview We developed a novel hybrid strategy to assemble the genomes, implemented in a package IMR/DENOM, that combines iterative mapping of reads to the reference genome with *de novo* assembly (software available on request from R.M.). Our reasoning was that whereas the latter procedure has advantages in the assembly of large insertions and deletions in unique regions, it is has difficulties in repetitive genomic regions that might be resolvable when combined with read-mapping information (i.e., a repetitive read can be correctly aligned using information from its uniquely aligned read pair). IMR/DENOM can be used with the read mappers MAQ[7], STAMPY[8] and BWA[9] and the *de novo* assemblers SOAPdenovo[10], ABySS[11] and VELVET[12] and is applicable to the assembly of any genome expected to be predominantly homozygous. In our study, the genome of each accession was assembled independently; because we had high coverage of each genome, and because we expected a large number of private alleles, we reasoned that any gains from a simultaneous assembly across accessions would be marginal.

3.2 Iterative read-mapping and realignment (IMR) Iterative realignment has a potential advantage over a single pass aligner for describing complex loci. Briefly, at each iteration, reads are aligned to the current version of a consensus sequence for a genome, high-confidence SNPs and indels are called, and incorporated into a new consensus. This process is then repeated until additional rounds of iteration produce few (or alternating) changes in the consensus sequence.

For assembling each accession's genome, we used the TAIR10 reference sequence as the consensus for the first iteration, and then aligned reads using STAMPY[8]. In our *Arabidopsis* genomes, we found that convergence occurred after about five iterations when the number of additional variants accepted was less than 2% of the number of the variants detected in the first iteration. At that point, the majority of remaining variants were unresolvable "heterozygotes" or cycled between alleles in successive iterations. These ambiguous positions can result for multiple reasons, including where repetitive read

mappings are not resolvable, where there is copy number variation, or where accessions harbour residual heterozygosity.

We called SNPs by processing the pileup files generated by the SAMTOOLS package[13] using default parameters. At the end of an iteration, a putative SNP was accepted/rejected using the default criteria in VarFilter[13], i.e., it was separated from other variable sites, its coverage of aligned reads was ≥3 and <100 and the root mean square (RMS) of the mapping qualities (PHRED scores) of the aligned reads was >25. SNPs at ill-defined sites where the reference sequence is not A,C,G or T, were only called if the major allele was supported by at least 80% of reads and coverage was within 20% of the mean genome-wide read coverage.

A confounding factor for describing genetic variation with alignment approaches is clusters of apparently heterozygous SNPs and small indels, which are sometimes an artifactual signature of an indel as the alignment of ends of reads over a genuine indel can generate SNP calls in preference to indels (e.g., see Supplementary Fig. 6 in [14]). By allowing for sequence variants to be inserted into a consensus, and removed if they do not agree in further iterations, iterative mapping is potentially useful to resolve such instances. In this study, we generated short indel predictions for consensus modification when indel predictions were reported in the pileup files (up to about 30bp) and were the best call at a genomic position.

In contrast to short insertions or deletions, we detected long deletions (defined as greater than 10bp) by local assembly. First, sites likely to contain an undetected deletion were inferred from read-pair data where the local observed insert size was longer than the expected library size (estimated by predicting the insert size of every pair of mapped reads), and then looking for local departures from the mean using dynamic programming[15]. We estimated the breakpoints (to within ~30bp) based on the difference between the apparent insert size and the global mean value (where multiple libraries with different insert sizes were used, we allowed for this by normalizing the insert size to a common value). Over this region we built left and right temporary consensus sequences by growing inwards from the breakpoints using the read mapping information. If there was a deletion, we expected that the two ends would overlap. To determine the precise breakpoints, we aligned the left and right consensus sequences using the Smith-Waterman algorithm[16].

We paid special attention to clusters of SNPs and indels that may be indicative of imbalanced substitutions (i.e., deletion of the consensus sequence with simultaneous insertion of a novel sequence of the same or, typically, different length). First, the best indel in a cluster was accepted and the nearby variants left for evaluation in the next iteration. Second, if there were no indels, we accepted the SNP in the cluster with the highest RMS of the mapping qualities, if it passed the criteria above for SNPs at ill-defined reference sites. We ignored the other variants nearby, thus allowing those to be considered at the next iteration. We found that SNP and indel predictions that were artefacts due to misalignment of reads often disappeared in subsequent iterations.

When all iterations were finished, the variants were re-evaluated to identify potential errors caused by read alignment errors (i.e., mismapping of repetitive reads). The variants detected in each iteration were mapped to the corresponding coordinate of TAIR10, thereby allowing iteration histories to be constructed. Variants that cycled over several iterations but eventually converged were accepted once they became stable, while unstable cycling variants were discarded.

3.3 De novo assembly based variant-calling (DENOM) We used SOAP denovo[10], which has previously been used in eukaryotic assembly efforts with Illumina data[17], to assemble the short reads for each accession into contigs (although any other assembler could be used). We

aligned the *de novo* contigs to TAIR10, using the BWA short read aligner[9] where contigs were less than 200 bp. The longer contigs were pre-processed using BWA-SW (using default parameters), and contigs containing divergent regions or structural variants were (typically) split into several segments by BWA-SW. The DENOM approach then assessed if the locus was likely to be a divergent region or structural variant, and realigned the contig. We incorporated the contig alignments into BAM files for display and to facilitate processing of variant calls. We called variants against the reference in a similar way to a single iteration of IMR, except that the parameter settings were altered to reflect the fact that there is only one contig covering a locus in most instances. The N50 *de novo* contig length (Supplementary Table 1) varied between accessions, in the range of 1-2kb, which is an effective upper bound on the size of insertions that can be detected. In addition we do not expect to be able to assemble through extended repetitive regions. Interestingly, the N50 contig size for Mt-0 was markedly lower, at 646bp. This accession was the only one that included a mate-pair library (with 1.5kb inserts), so the small N50 value is unexpected.

3.4 Integration of iterative and de novo variant calls Finally, we used the *de novo* contigs to refine the IMR/DENOM variants and assemblies, in particular to resolve insertions that are undetectable by IMR. Both IMR and DENOM produce lists of variants anchored to the original TAIR10 coordinate system, and identical variants predicted by IMR and DENOM (see above) were always accepted. If a SNP was heterozygous in IMR but homozygous in DENOM then the heterozygote call was accepted because heterozygotes are excluded from DENOM by construction.

Next, for more complex regions, we treated a genomic region with many shared alleles between IMR and DENOM as defining a common haplotype. To resolve the small number of discrepant variants, we used the positions of the identical alleles as a scaffold, and then attempted to match variants lying between successive scaffold points in the two data sets. For example, a variant in an IMR scaffold interval can only be compared to DENOM variants in the same interval. Complex combinations of SNPs and indels often appeared different solely as a result of alternative alignments (i.e., the underlying sequence is the same), and we attempted to determine equivalence. If two indels were close to each other, with the same length in DENOM and IMR, then they were regarded as the same and the IMR result was used. If an indel in DENOM overlaps with a set of SNPs in IMR, then the indel was selected since the SNPs likely reflected alignment artefacts. We accepted remaining variants detected solely by IMR, or solely by DENOM (for instance, long insertions).

A simulated set of sequencing reads (of the same read lengths and insert sizes as those used in reality) was generated from TAIR10 and mapped back to the reference using STAMPY (parameter settings were as for mapping actual reads, see above). The alleles called at each site give an estimate of the errors to be expected simply due to read mismapping, as there should be no variants or heterozygotes called. Then, at each SNP identified, we compared the counts of reads supporting each allele call in the real and simulated alignments using Fisher's Exact Test and rejected calls that were not significantly different. We calculated PHRED scores to show whether the SNP is well supported, using only the reads with high-mapping qualities. Since the mapping quality of reads in a repetitive region is usually very low, the score indicates whether the variant call is likely an artefact due to mapping error. We also remapped all the reads to the final version of the genome to derive a further quality score indicating regions in which the reads align cleanly. The average median coverage of reads aligned to the final assembly of each accession was 38x, lower than the initial 46x, due in part to the large number of reads aligned to the chloroplast genomes which are included in the initial estimate but excluded from the final (Supplementary Table 1).

This final step in the assembly process is referred to as "Final" in Supplementary Table 2. This table tracks the decrease in error rate as the assembly progressed through each step, for each of the test data sets described below in Supplementary Information section 4. During the five iterative alignment steps the SNP error rate halved, while the indel error rate fell by 5-9 fold. The integration with *de novo* assemblies improves accuracy further: SNP error rates fall by almost 100-fold (for the Bur-0 divergent data, from 387 to 4), and about 1.5-fold for the Ler-0 transposable element rich sequence. Indel error rates also decrease further, by variable amounts. The table shows that the greatest improvement of our method over a simple approach of aligning reads once and calling variants is for highly divergent sequence, as might be expected. The analysis of the remaining errors in the transposable element rich Ler-0 sequence shows that over half of the errors are where there is a discrepancy between the *de novo* contig and the iterative assembly, and that the latter is usually correct although the former has been used to call the variant. These errors are predominantly in transposable elements. Thus, while including *de novo* contig data improved accuracy in single copy regions (particularly at divergent loci), it can have a detrimental effect in repetitive regions.

3.5 Integration of assemblies across accessions The output of our method – the merged IMR and DENOM genome assemblies – is a table of variants, all of which are anchored to the TAIR10 reference coordinate system, with an additional column containing a quality score that details the rule used to determine the variant. Together with the reference sequence, this output is sufficient to generate pseudochromosome sequences for each accession. Nevertheless, comparing sequences across accessions is not trivial as divergent sequences could be anchored in different ways to the reference sequence (even though the sequences themselves are identical; for instance, a cluster of nucleotide differences can be represented as SNPs, or as an imbalanced substitution). We performed a post-processing step to present identical alleles in a consistent manner across accessions. Based on the resulting alignments, we generated a common pseudo-reference sequence for all the assembled accessions (the shortest common super-sequence from which each genome can be derived only by substitution or deletion, but never by insertion). This defines a common coordinate system. Depending on the analysis that we performed, we used either the original set of variant calls, the pseudochromosome sequences, or this integrated set (see below).

3.6 Identifying polymorphic regions (PRs) Although we attempted to resolve all sequence variants, this is not possible with short read data with small insert sizes. In *Arabidopsis*, microarray and pilot Illumina resequencing projects with a small number of accessions have shown that as much as several Mb in typical accessions is deleted (or highly different) relative to the reference accession, even in non-repetitive regions[14,18,19,20]. These regions that lack hybridization support (arrays) or next-generation read coverage have been termed polymorphic regions, or PRs, denoting their inferred underlying sequence divergence[14,19].

For our IMR approach, the consensus sequence was modified at each iteration step; however, in deleted regions, or regions of extreme sequence divergence, no reads are expected to map in any iteration (the reads do not exist, or differ too greatly from the reference to be aligned). A consequence is that unless a deletion can be predicted explicitly, such regions are not changed during the iteration process, and persist as the reference sequence in the final assembly even though they are (most likely) absent. To identify these regions, which we also call PRs for consistency with earlier studies, we aligned all read data for each accession to the final pseudochromosomes (read alignments were performed with STAMPY as described above). By parsing pileup files generated with SAMtools from the "self" read mappings, we identified PRs as regions of no (or low) coverage in the final

pseudochromosome sequences (Supplementary Figs. 1 and 2). To allow for misaligned reads at the junction of regions of deletion or extreme polymorphism, we identified PRs as contiguous positions for which the read coverage was ≤ 3. Example PRs from accession Bur-0 are shown in Supplementary Fig. 2, and are characterized by precipitous drops and sudden recovery in read coverage, as expected for regions for which read alignment is not possible. By treating PRs as breaks (effectively missing information) in our assemblies, we calculated N50 values for each assembly (Supplementary Table 4).

A potential concern for our identification of PRs is that STAMPY, which we used for read alignment, prioritizes mapping of read pairs nearby each other when possible. As sequences were removed or added during assembly iterations, a concern was that the mapper pulled reads from one region to another region to maintain so-called happy read pairs, possibly leading to no-coverage regions as an artefact. To assess this possibility, we also generated PR ranges using the alignment data from the first iteration of IMR (the initial mapping of reads to the TAIR10 sequence). We found that PR bases in the final pseudochromosome builds were almost invariantly inclusive to PRs in the first iteration (Supplementary Fig. 1). At least in unique regions, this suggests that PRs ascertained from the final pseudochromsome builds are not an artefact of the iterative mapping procedure per se. Nevertheless, in repetitive regions, we cannot exclude the possibility that PRs could nevertheless be an artefact of the read mapping process.

By comparing PR locations to SNP positions in each accession, we found that ~500 SNPs per accession overlapped PRs (i.e., where coverage was ≤ 3 but nonzero and some SNPs were called). We removed these SNPs, which are suspect owing to extremely low read support and misalignment at the junction of complex sequences, from the curated SNP release (Supplementary Information section 1) and from most subsequent analyses.

3.7 Metrics for deleted and inserted bases, and total affected bases While assessing deletion and PR metrics relative to TAIR10 is straightforward, assessing insertion counts and lengths (either simple insertions or inserted sequences that are part of ISs) is more complicated where different sequences of the same length or different lengths are present among accessions at the same position relative to TAIR10. Here, the insertion events could have arisen as different mutational events (i.e., as expansion or contraction of microsatellite repeats), or alternatively the same event, but with subsequent divergence, or sequencing errors. Where multiple insertion alleles among accessions were present at a given location relative to TAIR10, for non-redundant bases affected calculations, only the length of the longest insertion was used (i.e., the lengths were not summed; see the "Non-redundant" entry in Supplementary Table 3). We applied these criteria as well in assessing non-redundant bases that differ relative to TAIR10 in unique regions (inserted bases were counted as unique if the insertion site – or sites if an IS – was at a unique position). As assessed with this measure, and combined with unique deleted and PR bases, and SNPs in unique regions, non-redundantly 13.9% of the length of the TAIR10 genome is variable relative to other accessions in unique sequence (see also main text).

4. Validation of Assemblies

4.1 Comparison to existing datasets We evaluated genome identity, and accuracy of the assemblies, against multiple, independent datasets as described below. (i) We used SNP genotypes for 1,090 SNPs across all 18 accessions that we previously collected[3]. These are a subset of the 1,260 SNPs reported in that study, the reduction caused by the requirement that the SNPs' flanking sequences map unambiguously to TAIR10 (so that we could identify the

SNP sequence position in the variant table) and that genotypes across all 18 accessions were called. We then compared the alleles called in the variant table to the SNPs, and ignored the positions where either a heterozygote or no base was called. The error rate was defined as the fraction of called alleles that were different between the two sets. We found that 99.02% of genotypes from 1,090 SNPs agreed with the sequenced bases, confirming the accessions' identities as progenitors of the MAGIC lines. (ii) To estimate the accuracy of the assemblies in more detail we examined two accessions Bur-0 and Ler-0 for which capillary sequence data was available. We compared the pseudochromosomes of our Bur-0 assembly to 1,442 single-copy fragments of total length 602 kb (mean length 417 bp)[6], and to 188 divergent fragments of total length 48 kb (mean length 255bp)[14], using BLAT with default parameters. We counted the numbers of mismatches (excluding all positions with heterozygous or ambiguous base calls), indel events or imbalanced substitutions, and divided it by the total length of aligned sequence to obtain error rates per 10 kb (thus we excluded a small amount of unaligned sequence). For Ler-0, two hand-finished regions (175kb from chromosome 5 and 339kb from chromosome 3[21]), were kindly provided to us by Dr. Paul Dikjwel, Massey University, New Zealand. These regions had been assembled independently using sequence data from a mixture of sources including our Ler-0 reads, as well as PCR products designed to resolve ambiguous regions. The 175kb fragment was contiguous, whilst the transposable element rich 339kb fragment (Genbank accession HQ698308) had been assembled into 6 contigs. We aligned the pseudochromosome of our Ler-0 assembly to these sequences using BLAT with default parameters, and counted the numbers and rates of mismatches and indels as above. The detailed analysis of errors at each stage of the assembly process is provided in Supplementary Table 2 (discussed above).

4.2 Validation of indels and imbalanced substitutions by PCR and sequencing For accession Ler-0 we selected 80 indels and imbalanced substitutions for experimental validation. We chose the 20 largest deletions, 20 largest insertions and 40 largest imbalanced substitutions for which it was possible to design PCR primers (we rejected 14 large indels and IS polymorphisms in divergent regions where it was not impossible to design primers that amplified both Col-0 and Ler-0). We amplified PCR fragments and compared the observed fragment sizes with the expectation for Ler-0. We then sequenced and confirmed the breakpoints using capillary sequencing. For 68/80 fragments we were able to design primers that gave a single product. Of those, PCR and sequencing confirmed the indel size and breakpoint sequence in 66 cases (97%). In the remaining cases, it was impossible to isolate a single product for sequencing from the PCR (see Supplementary Table 8 for details).

5. Generic Methods

We developed the approaches described in the following sections that were used, or generated key resources, for many subsequent analyses. Software is available upon request.

5.1 Identification of Repetitive Regions We aligned each 50-mer present in the TAIR10 genome sequence against TAIR10 using GenomeMapper, which maps fragments to all matching positions[22]. We scored the alignments using 5 different settings in increasing levels of similarity: 1) 4 mismatches and 2 gaps but at most 4 edit-operations, 2) 3 mismatches and 1 gap but at most 3 edit-operations, 3) 2 mismatches and 1 gap but at most 2 edit-operations, 4) 1 mismatches and 0 gaps, and 5) no mismatch and no gap. For each of the settings we recorded those positions p covered by more than the expected 50 alignments and assigned repetitiveness score R_p to each p as the maximum of the alignment scores at that position so

that it measures the degree of similarity of the position to other loci ($R_p=0$ at single-copy, or unique positions). The data are available from the supplementary website and are displayed in GBrowse (see Supplementary Information section 1).

5.2 Translating annotations between assemblies To map genomic annotations from reference to accession genome coordinates and back efficiently, we precomputed a table containing the position of a nucleotide in a genome of an accession in the reference genome, available from the supplementary website in HDF5 format (The HDF5 Group, http://www.hdfgroup.org/HDF5). To map a single exon, we first attempt to map the start and end positions. If these are not present in the target accession, then we map all nucleotides within the exon and use the first and last mapped positions as the mapped exon start and end. To map an entire transcript annotation, all exons of the transcript are mapped independently. An exon is removed from the transcript if no exonic nucleotide can be matched. Similarly, a transcript is removed if no exon is mappable and a gene is removed if no transcript is mappable. We transformed RNA-seq alignments in BAM format between accessions using the coordinate mapping HDF5 file to incorporate substitutions and indels into the RNA-seq alignments. The mapped alignments are represented as CIGAR strings in BAM format.

5.3 Generating non-redundant polymorphic regions When comparing PRs across all accessions one faces the problem that the boundaries of PRs shared between accessions are not strictly aligned. We therefore developed an approach based on dynamic programming that generates blocks with boundaries aligned among the accessions and a score reflecting the fraction of polymorphic bases per accession. The algorithm was designed in a way such that the boundaries were optimally chosen with respect to the original PRs, guaranteeing that the polymorphic degree of each accession changed as little as possible within a block. At every tenth nucleotide n in TAIR10, we calculated the minimal cumulative squared deviation L_n of being polymorphic p_{qa} and the averaged polymorphic degree μ_s for all positions q in the segment of length s preceding the nucleotide n in accession a, maximally scanning a region of 1,000 nt. The optimization was performed using the following recursive update formula for the current cost L_n:

$$L_n = \min_{s=1\ldots1000} \sum_{a=1}^{A} \sum_{q=n-s+1}^{s} (p_{q,a} - \mu_s)^2 + L_{n-s} + \text{switch cost}$$

The switch cost parameter allowed us to adjust the amount of generated blocks; a higher switch cost resulted in fewer blocks. Using a switch cost of 200 we merged 689,972 individual PRs to 27,242 non-redundant PRs with a median length of 1,506 nt. The resulting non-redundant polymorphic regions are well suited for use in association mapping.

6. Population Genetics

Population genetic analyses were performed with nuclear genome sequences and no results are reported for organellar genome sequences.

6.1 Measurement of nucleotide diversity for coding sequence and intergenic regions We created genome-wide FASTA alignments of all accessions (excluding Po-0 which was closely related to Oy-0, and extensively heterozygous) against TAIR10. Insertions relative to TAIR10 were ignored so that the alignment had the same coordinate system as the TAIR10 reference. Next, subalignments representing each TAIR10 annotated gene or intergenic region were made, and all repetitive bases ($R_p > 0$), PR regions, and ambiguous sites were

masked so that they would be treated as missing data and hence removed from the subsequent analysis. We used programs from the Analysis software package, based on the libsequence C++ library[23] to compute nucleotide diversity for different classes of sites. We computed nucleotide diversity for each intergenic region, and we used polydNdS to estimate nucleotide diversity for silent coding sites, replacement sites, and 4-fold degenerate sites.

To examine patterns of diversity as a function of chromosomal position, we took 400 kb long sliding windows along each TAIR10 chromosome where consecutive windows had a 50 kb offset. For each window, averages of nucleotide diversity across coding sequences or intergenic regions (weighted by the number of sites per region or silent sites per coding sequence, with data from two or more accessions, for example) were computed. A given coding sequence or intergenic region was only included when the midpoint of the region was within the 400 kb window.

Correlations between nucleotide diversity of different classes of sites and other genome features (e.g., amount of deleted sequence) were performed for nonoverlapping windows of 50 kb. Weighted averages for each region were computed as described above, except that only the portion of a region falling within each window was used in calculating the weighted average.

6.2 Decay of linkage disequilibrium SNP genotypes from the 18 accessions (Po-0 excluded) were filtered to remove ambiguous calls, those in repetitive regions, or those inclusive to PR regions. Since ambiguous (potentially heterozygous) calls were dropped, no haplotype phasing was necessary. Additionally, SNPs were excluded if not biallelic, if the minor allele frequency was less than or equal to 0.1, or if fewer than 16 of the 18 accessions exhibited a genotype at this position. These SNPs were used as input for the program PLINK (v1.07)[24] in order to compute the linkage disequilibrium (LD) as measured by r^2 (Fig. 1d). We also investigated the decay of LD for pairs of SNPs with matched allele frequencies [25]. We binned SNPs by minor allele frequency (0.11-0.2, 0.21-0.3, 0.31-0.4, and 0.41-0.5) and computed r^2 only for SNPs within the same frequency bin (Supplementary Fig. 16).

6.3 Haplotype sharing We took all SNPs identified in the 18 accessions (Po-0 excluded) and removed those that fell into repetitive regions ($R_p > 0$). We also dropped genotype calls inclusive to PRs by accession. From the remaining genotypes, we performed all pairwise comparisons of the 18 accessions for windows of 10,000 bases for TAIR10, counting the total number of differences and the total number of comparisons made for each pair. No comparison was made when one or both accessions in the pair had a genotype call other than A, C, G, or T (whether due to heterozygosity, a deletion, a PR region, or other missing data). Because each window of 10,000 bases can have a different number of repetitive bases, we also recorded the number of unique bases per window. Each pairwise accession comparison was then scanned for blocks of 5 or more consecutive windows where, for each window in the block, there are 1 or fewer SNP genotype differences per 1,000 unique bases, and the consecutive windows meeting this criterion were merged into a single pairwise block. For each block thus identified, those that involved less than 20 SNP comparisons were dropped. We then plotted the remaining blocks along each chromosome (Supplementary Figs 11-15, panels c), where each row represents a different pairwise comparison, and colour-coded each block based on the value of the ratio of SNP genotype differences per total unique sequence (darker shades represent lower ratios, i.e., higher levels of identity between the two accessions).

6.4 Ancestral recombination graphs To estimate the variation in ancestral relationships between the 18 accessions and Col-0 across the genome, we extracted local phylogenies from

ancestral recombination graphs (ARGs). Rather than attempting to partition the genome into a single ARG, we took account of the uncertainty in the placement of boundaries between neighbouring phylogenies by sampling up to five likely phylogenies at each locus. Previous studies[26] have shown that the embedded phylogenies generally provide good estimates of ancestry at each position, in particular for positions sufficiently far away from the boundaries of the region considered.

To do this, we removed all sites that were multiallelic or had heterozygous or private alleles. Accessions deleted at a site were treated as missing data. The filtering left 1.25 million sites for analysis, around each of which we inferred local phylogenies. We then identified a region centred on each site within which we could construct probable local phylogenies. The size of the region with the same phylogeny varied between sites, so each was determined dynamically to be sufficiently large that most accessions would be distinguishable: Starting from just the relevant position, the region was first symmetrically extended until at most four triplets of accessions that were identical across the region remained, to ensure sufficient sequence information to resolve most branching points. For regions smaller than 41 SNP positions (i.e., a window of 20 SNPs on either side) after the initial extension, we further extended the region until the Robinson-Fould distance[27] between phylogenies for consecutive region widths no longer showed a decreasing trend. Thus each local region size was determined by computing test phylogenies, to ensure that regions were sufficiently wide for the position of interest to be at a sufficient distance from the region ends. Whenever the final region for a position was a subset of the final region for another position, only the ARG from the larger region was inferred and local phylogenies were extracted from this ARG for both positions.

ARGs were inferred within each region using the kwarg program. This is a heuristic modification of beagle[28] that determines the ARG with the minimum number of recombinations required from a SNP data set. kwarg is initiated with the observed set of alleles across the region in the 19 accessions, and then moves back in time by performing mutation and recombination to coalesce the sequences into a single ancestor. In general there are multiple ancestries with the same numbers of events, so it is necessary to sample over them. kwarg exhibits strong similarities to the genealogy inference of Margarita[29], but aims to make local choices keeping the number of recombinations low rather than to choose recombinations allowing coalescence of long tracts. After experimentation, we decided to take the first five ARGs inferred for each region, and all statistics reported are averages over them. A time limit was imposed on all ARG inferences. This resulted in 11,001 positions for which no local phylogeny was estimated, and a further 8,452 ARG inferences that were prematurely terminated. Of the original 1,255,082 sites, we sampled one or more phylogenies for the 1,244,081 sites, with an average of 4.99 phylogenies sampled for each. We counted the frequency of a tree as 0.2 multiplied by the number of times it was sampled. In total, at the 1.25 million positions, 2.55 million different trees were sampled, of which 0.28 million had frequency 1 or greater (i.e., were sampled 5 or more times). The most frequently observed tree occurred only 189.2 times, corresponding to about 0.015% of all positions. The total number of possible binary trees with 19 leaves is approximately 6.3×10^{18} so only a small fraction (4.0×10^{-13}) was observed. In order to ensure trees with the same topology were represented in the same way, in the supplementary data available from our website, the trees are always rooted at Col-0.

To assess the relatedness between pairs of accessions, for each tree we determined the smallest leaf set containing both accessions taken over all bipartitions of the tree. We will refer to this as the minimum clade size distance (MCS) between the accessions – indeed, when the root is not in the same partitions as the accessions, this measure is the size of the smallest clade containing both accessions. Though the local phylogenies are extracted as

rooted trees and the recombination model does impose some time ordering constraints, the true location of the root may be poorly determined. Hence, we do not rely on the reported location of the root for defining the MCS. We decided against using the standard path length distance between accessions, as branch lengths in the local phylogenies cannot be reliably estimated from the small amount of mutational information in a region.

Pairwise MCS were analysed both in terms variation with distance across the genome (measuring how fast the phylogeny changes as a function of physical distance) and as genome wide distributions over phylogenetic distances (measuring the fraction of loci where a pair of accessions are a given phylogenetic distance apart). The former plots show rapid variation, though there are exceptions where two accessions remain at small distance for extended regions of several hundred kilobases or even megabases. We used R^2 correlation between phylogenies (described below) to measure their similarity. This quantity decays to near background levels within about 25kb, consistent with the linkage disequilibrium analysis (Fig 1d). The observed MCS were compared to the expected MCS for random tree topologies with 19 leaves, computed using a recursion based on standard methods for counting the number of different topologies. For a given tree topology, the signature vector **n** is defined to be the list of sizes of the subtrees branching off the path connecting the two accessions of interest. The number of topologies with a specific signature $\mathbf{n} = n_1,...,n_{|\mathbf{n}|}$ is defined recursively by

$$N(\mathbf{n}) = \sum_{i=1}^{|\mathbf{n}|} \begin{cases} N(n_1,...,n_{i-1},n_{i+1},...,n_{|\mathbf{n}|}) & \text{if } n_i = 1 \\ (2n_i - 3)N(n_1,...,n_{i-1},n_{i-1},n_{i+1},...,n_{|\mathbf{n}|}) & \text{otherwise} \end{cases},$$

with $N() = 1$. If $n = 2 + \sum_{i=1}^{|\mathbf{n}|} n_i$ is the total number of accessions, then the minimum clade size distance for trees with signature n is $n - \max\{n_i \,|\, 1 \le i \le |\mathbf{n}|\}$. Generally the observed distribution matches the random distribution quite well, though it is slightly more concentrated at the mode (minimum clade size of 13). However, for most accessions we observe pairwise comparisons with one or more accessions for which the distribution is bimodal with a minor mode at minimum clade sizes of 2 or 3, reflecting the extended regions of close relationships also observed in the plots of variation across the genome.

The decay in correlation between phylogenies as a function of physical distance apart was estimated from the MCS distances between all pairs of accessions. Thus, for each pair of accessions *a, b* and physical distance *d* we counted the empirical probability distribution $f(x,y,d)$, the fraction of sites *i* such that the MCS between *a* and *b* is *x* at *i* and *y* at *i*+*d*, and also the marginal distributions $F(x,d) = \sum_y f(x,y,d)$ and $G(y,d) = \sum_x f(x,y,d)$. We computed the correlation coefficient R^2 between phylogenies from the quantities $X(d) = \sum_x N_d F(x,d)x$ $XY(d) = \sum_{xy} N_d f(x,y,d)xy$, $XX(d) = \sum_x N_d F(x,d)x^2$, where N_d is the number of pairs of sites a distance *d* apart. We also computed analogous quantities *Y(d)*, *YY(d)* with *G* substituted for *F*. The Pearson correlation coefficient for sites a distance *d* apart is the average over all accessions *a, b* is $r_{ab}(d)^2 = \dfrac{N_d XY(d) - X(d)Y(d)}{\sqrt{N_d XX(d) - X(d)^2}\sqrt{N_d YY(d) - Y(d)^2}}$.

7. Transcriptome Sequencing

7.1 Collection of seedling tissue for RNA-seq studies We collected seedling tissue, with biological replication, for all 19 founder accessions for the production of standard Illumina mRNA sequencing (RNA-seq) libraries (Supplementary Information section 7.4). In addition, we further collected seedling, root, and floral bud RNA samples for accessions Col-0 and Can-0 for production of strand-specific RNA-seq libraries (Supplementary Information section 7.5).

For the seedling stage, seeds were sterilized with chlorine gas for 3 hours, and stratified for 5 days in 0.1% agarose at 4°C. Plants were germinated and grown on Sunshine Mix 4 Aggregate Plus soil (cat. no. LA4, Sun Gro Horticulture, Bellevue, WA, USA) supplemented with ~0.03 g of Miracle-Grow® all purpose plant food in 10 cm square pots in a Percival AR-66L environment growth chamber. Germination and subsequent growth was at 20°C under standard long day growth conditions (16 hours light and 8 hours dark). Approximately 16 seedlings were grown per pot, and 9 pots were planted per accession. To minimize environmental effects on development resulting from position within the growth chamber, plants were rotated through the chamber daily. After either 11 or 12 days, depending on per accession germination and growth rates, the aerial portions of seedlings were harvested at a uniform developmental stage (the time the fourth true leaf appeared, Supplementary Fig. 3; the developing aerial rosettes were detached just below the cotyledons). For each biological replicate per accession, rosettes from 20 seedlings were collected and combined. To further minimize environmental effects, seedlings for a given replicate were collected in approximately equal numbers from each pot within the growth chamber for a given accession. Biological replicates for each accession were isolated on the same day, and to minimize circadian and light dependent effects on gene expression[30,31], tissue collections were carried out at 8 ± 0.5 hours into the light cycle. All seedling samples were collected over a three-week period.

Root tissue was collected from 100 10-day old seedlings grown under the same conditions as for seedlings. Seeds were stratified for 5 days at 4°C and arranged in horizontal rows on vertically stacked agar plates (1% sucrose, 1% agar, 1x Murashige and Skoog salt, and 2.3 mM MES free acid monohydrate). To minimize environmental effects on root growth all seedlings were grown simultaneously and rotated through the chamber daily. Root tissue collections were carried out at 8 ± 0.25 hour into the light cycle. Roots were excised approximately 2 mm below the end of the hypocotyl.

For collection of floral tissue, plants were grown on Sunshine Mix 4 Aggregate Plus soil supplemented with ~0.03 g of Miracle-Grow® all purpose plant food in 10 cm square pots. Plants were germinated at 20°C under standard long day growth conditions for 6 days then vernalized at 4°C for 6 weeks under short day conditions (8 hours light and 16 hours of dark). After six weeks at 4°C, plants were grown under the same environmental conditions used for seedling and root tissue collection. Roughly 120 stage-12 floral buds were harvested from 10 Can-0 and Col-0 plants over a period of 1 week. The first five floral buds on each floral stem were excluded from collection. Floral buds were collected into liquid nitrogen at 8 ± 0.5 hours into the light cycle.

Tissue from all stages was collected directly into liquid nitrogen and subsequently stored at -80°C.

7.2 Preparation of RNA for library construction In general, total RNA isolation was as described by Filichkin *et al.*[32] with modification. Briefly, frozen seedling tissue was pulverized in liquid nitrogen with mortar and pestle (cat. no. 60313, CoorsTek®, Golden, CO, USA), and RNA was extracted from pulverized tissue with PureLink Plant RNA

Reagent (cat. no. 12322-012, Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. The precipitated RNA was resuspended in RNAsecure® Resuspension Solution (cat. no. AM7010, Ambion, Austin, TX, USA), according to the manufacturer's instructions to protect against RNA degradation. To remove potential DNA contamination, samples were incubated with Turbo DNase (cat. no. AM2238, Ambion) at 37°C for 15 minutes. Following DNase treatment, RNA was precipitated with 2.5 volumes of ethanol, 0.1 volumes of ammonium acetate, and 1 µl of glycogen (5 mg/ml) and resuspended in RNase free water. Total RNA concentrations were estimated with a NanoDrop 2000 (Thermo Scientific, Waltham, MA, USA).

From each sample, messenger RNA (mRNA) was purified from 35 µg of total RNA with two rounds of selection with DynalBeads® Oligo $(dT)_{25}$ beads (cat. no. 610.05, Invitrogen) as describe by the manufacturer. Isolated mRNA was suspended in RNase free water.

7.3 Barcoding of Illumina RNA-seq libraries To facilitate analysis of multiple RNA samples, with replication, we constructed barcoded RNA-seq libraries with adapter sequences after[33] suitable for sequencing using the Illumina SBS method[34]. Briefly, barcode sequences ACGT, CATT, GTAT, and TGCT extend the standard Illumina single-end library adapters, allowing four samples to be run within a single Illumina flowcell lane. To generate adapters, PAGE purified oligos obtained from Integrated DNA Technologies (Coralville, IA, USA) were diluted to a concentration of 200 mM in 10 mM NaCl, equal volumes of 200 mM oligo pairs were mixed and incubated at 90°C for two minutes, and the mixture was cooled to 30°C at a rate of -2°C/min in a thermal cycler as described in[33]. The annealed adapters were then snap cooled on ice, diluted to a final working concentration of 15mM with nuclease free water, and stored at -20°C.

7.4 Preparation of standard Illumina mRNA sequencing libraries RNA sequencing (RNA-seq) libraries were generated with methods adapted from the Illumina mRNA sample preparation protocol (cat. no 1004894 Rev.A; see also [34]). Briefly, ~150 ng of mRNA were fragmented at 70°C for two minutes with Ambion RNA fragmentation reagent (cat. no. AM8740, Ambion), precipitated with ethanol as described above, and resuspended in 10 µl of RNase free water. The quantity and quality of fragmented mRNA was evaluated on a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA) with RNA 6000 Nano Kit reagents (cat. no. 5067-1511, Agilent).

Following fragmentation, first-strand cDNA synthesis starting with 50-100 ng of mRNA was performed in a 20 ul reaction with the Invitrogen Superscript First Strand Synthesis System (cat. no. 11904-018, Invitrogen; 250 ng of random hexamers were used for first-strand cDNA synthesis). The entire first-strand cDNA reaction was then brought to 100 µl with second strand synthesis reagents consisting of 1X (final) reaction buffer (500mM TRIS-HCl pH 7.8, 50 mM $MgCl_2$, and 10 mM DTT as a 10X concentrate), 50 units of DNA polymerase I (cat. no. 18010-025, Invitrogen), and 2 units of RNase H (cat. no. 18021-014, Invitrogen). Second-strand cDNA synthesis was carried out at 16°C for 2.5 hours, and double-stranded cDNA was purified using a Qiaquick PCR purification Kit according to the manufacturer's instructions (cat. no. 28106, Qiagen).

End repair, dA-tailing, and the ligation of barcoded adapters to target cDNA inserts was carried out with New England BioLabs NEBNExt Sample Preparation Modules according to manufacturer's instructions except that half reaction volumes were used (End Repair Module cat. no. E6050L, dA-Tailing Module cat. no. E6053L, Quick Ligation Module cat. no. E6056L, New England Biolabs, Ipswich, MA, USA). One µl of a 15 mM barcode adapter stock was used in a given quick ligation reaction. In between the above steps in

library construction, samples were purified with MiniElute PCR purification kits (cat. no. 28006, Qiagen), and resuspended in 20 µl (end repair and ligation) or 15 µl (dA-tailing) of elution buffer (cat. no. 19086, Qiagen)

Prior to amplification of the libraries by PCR, adapter-ligated cDNAs were size selected in the range of 230-330 bp with gel electrophoresis on 2% agarose gels run for one hour at 120 Volts (to ensure uniform size selection between samples, DNA markers of 230 and 330 bp were PCR amplified from Lambda genomic DNA and run adjacent to each library; primer sequences used for Lambda amplification of the 230 and 330 bp fragments were 5'-TGCAATGCCACAAAGAAGAG-3' and 5'-AGACGATACGTCGAAGTGAC-3', and 5'-GGGAAAATCCCCTAAAACGA-3' and 5'-TCACGTTCACGCATCAGGCG-3', respectively). DNA was purified from gel fragments using a QIAquick Gel Extraction Kit following the manufacturers instructions (Qiagen, cat#28704).

For library amplification, adapter-ligated cDNA was PCR amplified using Phusion Polymerase and Phusion HF buffer (cat. no. F-530L, New England Biolabs). Each barcoded library was amplified (independently) for 15 cycles of 98°C for 10 seconds, 65°C for 30 seconds, and 72°C for 30 seconds, followed by a five minute incubation at 72°C. PCR products were purified with QIAquick PCR Purification Kits (cat. no. 28106).

Following amplification, the quantity and quality (concentration and size distribution) of each library was evaluated on a Bioanalyzer 2100 with DNA 1000 Kits (Agilent, cat. no. 5067-1504, Agilent; in all cases, the estimated library sizes agreed closely with that expected from the size selection). Template concentrations were adjusted to 10 nM using a solution of 10 mM Tris-HCl, pH 8.5, and containing 0.1% Tween 20 to maintain library titer (loss of diluted product to tube walls) according to Illumina sequencing protocols (see below).

7.5 Preparation of Illumina Strand-specific RNA sequencing libraries Libraries were generated with methods adapted from the Illumina Directional mRNA-Seq Prep Pre-Release Protocol (cat. no FC-102-1010 Rev.A). RNA isolation, DNase treatment, mRNA purification, and RNA fragmentation procedures are identical to those listed above.

Following fragmentation, 50-100 ng of RNA were treated with Antarctic phosphatase (cat no MO289S, New England Biolabs) in a 20 µl reaction for 30 minutes at 37°C followed by 5 minutes at 65°C. Phosphatase treatment of fragmented RNA was carried out in the presence of the RNase inhibitor RNaseOUT (Illumina Digital Gene Expression for Small RNA Sample Prep Kit p/n 1002398). Phosphatase-treated fragmented RNA was then treated with T4 Polynucleotide Kinase (PNK) (cat no M0201L, New England Biolabs). 20 µl of phosphatase-treated RNA was added to a 50 µl reaction containing 1x PNK buffer, 1 mM ATP (cat no RA02825, Epicentre, Madison WI, USA), 1 µl RNaseOUT, and 20 units of T4 PNK. PNK treatment was carried out at 37°C for 60 minutes. PNK-treated RNA was column purified using RNeasy Mini Kit (Qiagen, cat no 74104) following the manufacturer's instructions. RNA was eluted off the RNeasy column using 20 µl of water and concentrated to 6 µl using a speedvac.

Ligation of RNA adapters was carried out following end repair of phosphatase and PNK treated RNA. 1 µl of the Illumina v1.5 sRNA 3' adapter (10 fold dilution) was added to 6 µl of RNA and heated to 70°C for two minutes. Following incubation at 70°C, 3' adapter ligation was carried out using T4 Truncated RNA Ligase 2 (cat no MO242L, New England Biolabs) in a 10 µl reaction containing 1x T4 RLN2 reaction buffer, 8 mM MgCl2, and 1 µl of RNaseOUT. The 3' adapter ligation reaction was carried out at 22°C for 1 hour. Following 3' RNA adapter ligation 1 µl of 10mM ATP, 1 µl of the SRA 5' adapter, and 1 µl of T4 RNA ligase was added to the 3' ligation reaction and incubated at 20°C for one hour (SRA 5' adapter and T4 RNA ligase included in the Illumina Digital Gene Expression for Small RNA Sample Prep Kit p/n 1002398).

First-strand cDNA synthesis was performed starting with 8 µl of adapter-ligated mRNA in a 20 µl reaction with the Invitrogen Superscript First Strand Synthesis System (cat. no. 11904-018, Invitrogen; 0.4 µl of the SRA-RT primer provided in the Illumina Digital Gene Expression for Small RNA Sample Prep Kit was used for first-strand cDNA synthesis). Prior to size selection, cDNA was amplified using Phusion Polymerase and Phusion HF buffer (cat. no. F-530L, New England Biolabs). Each strand specific library was amplified for 3 cycles of 98°C for 10 seconds, 65°C for 30 seconds, and 72°C for 30 seconds, followed by a five minute incubation at 72°C using primers GX1 and GX2 (primers included in the Illumina Digital Gene Expression for Small RNA Sample Prep Kit).

Prior to amplification of the libraries by PCR, adapter-ligated cDNAs were size selected in the range of 175-275 bp with gel electrophoresis on 2% agarose gels run for one hour at 120 Volts. To ensure uniform size selection between samples, DNA markers of 175 and 275 bp were PCR amplified from Lambda genomic DNA and run adjacent to each library; primer sequences used for Lambda amplification of the 175 and 275 bp fragments are 5'-TGCAATGCCACAAAGAAGAG-3' and 5'-TTCATCTCACTACCACAACGAG-3', and 5'-TGCAATGCCACAAAGAAGAG-3' and 5'-GGATTGCATTTTGCAGACCT-3', respectively. DNA was purified from gel fragments using a QIAquick Gel Extraction Kit following the manufacturers instructions (Qiagen, cat#28704).

For library amplification, adapter-ligated cDNA was PCR amplified using Phusion Polymerase and Phusion HF buffer (cat. no. F-530L, New England Biolabs). Each library was amplified for 14 cycles of 98°C for 10 seconds, 65°C for 30 seconds, and 72°C for 30 seconds, followed by a five minute incubation at 72°C. PCR products were purified with QIAquick PCR Purification Kits (cat. no. 28106). Library size and yield was assessed as described above.

7.6 Cluster Generation and Single-Read Sequencing Illumina single-read cluster generation and sequencing was performed according to the manufacterer's instructions. Briefly, flowcell preparation on a Cluster Station was performed with v4 Illumina Single Read Cluster Generation Kits (cat. no. 15003972) using the "SR_Amplification_Linearization_Blocking_PrimerHyb_v7" Cluster Station workflow. For cluster generation, equal volumes of four barcoded 10 nM libraries were combined and diluted to 8 pM for flowcell hybridization. All sequencing was performed on a Genome Analyzer IIx (GAIIx) instrument using v4 Sequencing Kits (cat. no. 15003925) and with an Illumina supported modification of the "GA2_76Cycle_SR_v7.xml" sequencing program to perform 82 cycles of imaging. We used version SCS2.6 of the GAIIx control software with Real Time Analysis (RTA) enabled and set to use "per lane" parameters for basecalling.

7.7 Data Filtering and Processing of RNA-seq reads Read data was processed with the Illumina analysis pipeline (version CASAVA 1.6.0) without alignment (USE_BASES set to 'all', ANAYSIS set to 'sequence') to give sequence data with associated quality scores. Subsequently, barcode sequences at the beginning of each read were identified and trimmed. For a read to be assigned to a library, the first four bases of a sequence were required to match (exactly) a barcode sequence. To eliminate matches reflecting low quality sequence (basecalling errors), we further required that two of the first three bases have a quality score $\geq$ 30, and that all of the first three bases have quality scores of $\geq$ 28. The 4th base of each barcode sequence, a 'T', is needed for the Illumina library construction method, and we required the fourth base to have a quality score $\geq$ 20 (the fourth position does not discriminate between barcodes *per se*, it does indicate a correctly cloned product). After barcode assignment and trimming, we removed reads for which the beginning of the trimmed sequence matched Illumina adapter sequence, reflecting known adapter-adapter artefacts of

the library construction method. With these criteria, on average 88% of RNA-seq reads were usable (values ranged from 75%-96% among libraries).

7.8 Assessing adaptor ligation biases In a pilot study, we tested for potential biases in adapter ligation among barcodes. For instance, the barcode sequences might vary in ligation efficiency (possibly in a target sequence specific manner). To do this, we constructed libraries with each barcode from an identical sample of fragmented Col-0 mRNA. The four libraries, which were constructed as described above, were run in a single Illumina flowcell lane (in equal proportions). After barcode assignment and trimming, reads were aligned to the TAIR9 reference using CLC's Genomics Workbench (CLC bio, Muehltal, Germany), and RPKM values[35] were generated for each gene in the nuclear genome. Uniquely aligning reads with three or fewer mismatches were included in this analysis. To examine correlation among libraries, R^2 values were calculated for all possible pair-wise comparisons using the R statistical computing environment[36] for moderately or highly expressed genes, and the relative proportion of genes with 2-fold or greater differences in expression was assessed (Supplementary Fig. 4). R^2 values were all ≥ 0.997, and less than 0.1 % of genes differed in expression by more than 2-fold. This analysis indicated that inter-barcode ligation biases, a potential source of error in quantifying gene expression, are minimal using our library construction methods.

7.9 Assessing SNP predictions with RNA-seq alignments To further assess genome assemblies, as well as our methods for RNA-seq read alignments, we determined the agreement between SNP calls from the assemblies and base calls at SNP positions as assessed from aligned RNA-seq reads (Supplementary Information section 9). Although our RNA-seq data is also Illumina based, the reads are longer than those used for genome assembly (78 bp versus a maximum of 51 bp, respectively), and base calls from RNA-seq reads are independent per se of potential assembly errors. We limited our use of RNA-seq reads to assessing SNP concordance, as assessing indels from RNA-seq read alignments is confounded by the splitting of reads at (known and potentially unknown) splice sites.

Prior to base calling, we removed RNA-seq reads or alignment segments of reads that were likely misaligned. Initially, we excluded all reads that were split into more than 3 segments, reads with alignment codes other than 'M' and 'N' according to the .SAM alignment nomenclature v1.4-r962, and reads with adjoining matching segments spaced more than 7 kb apart (few introns in *A. thaliana* are more than 7 kb). Across all accessions, these filters removed 2.17% of aligned RNA-seq reads. Additionally, within reads we excluded from analyses aligned segments of 5 bp or less, and we excluded the 5 bp at the beginning and ending of reads (this was done to remove potential misalignments of read ends nearby exon-intron junctions). With the remaining alignments, we generated base calls at SNP positions in given accessions (relative to Col-0). To make a base call, we required coverage of at least one read for which an overlying base had a quality score ≥ 25. Where multiple RNA-seq reads covered a position, overlying base calls were required to agree at least 85% of the time.

On a per accession level, we calculated concordance between SNP calls and RNA-seq base calls for two sets of SNPs. First, we assessed concordance with all SNPs predicted in a given accession (~99.7%, Supplementary Table 5). Second, we note that SNPs predicted to disrupt gene models (and that are less common than other SNP types) are expected to have higher prediction error rates than for all SNPs (see Clark *et al.*[18] for detailed discussion). A high error rate at such "major-effect SNPs", i.e., those predicted to introduce premature stop codons, would confound *de novo* gene prediction as described in Supplementary Information section 10. Therefore, we calculated concordance at major-effect SNPs separately

(Supplementary Table 6). Across all major effect SNP types and accessions, the concordance dropped but very slightly to 99.03%; this small drop contrasts with that observed from noisier microarray resequencing data[18].

8. Analysis of Polymorphisms with Respect to TAIR10 Genes

In order to analyse the sequences of the 18 accessions, we first extracted and prepared annotations from the TAIR10 genome.

8.1 Classification of the genome by sequence type and gene family We classified bases in TAIR10 as coding, untranslated region, intron, transposable element, pseudogene, or intergenic. Where gene models or features overlapped, a given base was assigned to one sequence type in the above order (descending). Bases in annotated non-coding genes were classified as untranslated, and transposable element classifications were based on TAIR10 annotated transposable element genes as well as transposable element fragments. Gene family classifications were primarily from the "gene_families_sep_29_09_update" list hosted on TAIR. Additional resources were used for some gene family classifications: defensin-like[37], F-box[18], NB-LRR[38], receptor-like kinase[39], and plant transcription factor genes[40]. Owing to updates to the *Arabidopsis* genome annotation (from earlier versions to the TAIR10 annotation), some genes from the above sources are no longer present in the TAIR10 annotation, and these were removed from analyses. Initially, we limited our analysis of gene families to those having at least 50 members after resolving gene lists with the current TAIR10 annotation. Subsequently, we removed a small set of additional genes that were assigned to multiple gene families.

8.2 Disruptions in TAIR10 protein-coding genes TAIR10 contains 33,602 gene annotations classified as "gene", "pseudogene", and "transposable_element_gene". We first verified the TAIR10 sequence for expected splice sites (consensus GT/GC for donor splice sites and AG for acceptor splice site), translation start sites (TIS; consensus ATG) and translation stop sites (stop; consensus TAA/TAG/TGA). The consensus was missing for 350 splice sites, 87 TIS and 4 stop sites. The annotated coding sequences had open reading frames in all but 81 cases, and there were 66 frame-shifted coding sequences whose lengths were not a multiple of 3. We mapped the TAIR10 annotations to each of the 18 accessions' genomes and repeated these verification steps, ignoring instances present in TAIR10. On average, per accession, there were 709 splice site disruptions (56% in coding regions), 723 disrupted TISs, 444 stop site disruptions, 4,325 newly introduced premature in-frame stop codons and 2,791 frame shifts in transcripts of any of the 33,602 genes. Supplementary Table 13 summarises all consensus and reading frame disruptions, and detailed information is available on the supplementary website. We found that disruptions occur 3-8 times more frequently in genes with lowest median expression among the accessions compared to the highest expressed ones (see Supplementary Fig. 22).

8.3 TAIR10 miRNA Disruptions We mapped the 177 TAIR10 miRNA genes to the accessions' genomes and determined which of the known 243 miRNA stems[41] matched the RNA transcript. The miRNA stem sequences differed from the reference accession in 37 miRNA genes (on average 11 per accession). See the last column of Supplementary Table 13 for a summary.

9. RNA-seq Alignments We aligned 189 million 78-bp-long single-end Illumina RNA-seq reads (Supplementary Information section 7.4) with PALMapper[42] with at most 3 mismatches and 1 gap, to chromosomes 1-5 of the 18 assembled genomes and TAIR10. Matches to chloroplast and mitochondrial sequences were ignored. We recorded an average of 1.22 alignments per read. Initially, spliced alignments were required to have an intron length of less than 25 kb, to be split at splice site consensus sequences GT/GC and AG, and to perfectly match 5 bases adjacent to splice sites. For scoring spliced alignments, we used computational splice site predictions[43] from the mGeneToolbox[44], which were optimized using QPalma[45] based on artificially generated reads sampled from annotated TAIR10 genes[42]. If the read could not be mapped within given tolerances and the a tail contained a large fraction of A's or T's, the read is trimmed (minimum length 40 nt) and realigned. If no spliced alignment with a consensus splice junction GC-AG or GT-AG was found, a remapping phase permitted spliced alignments with all possible dinucleotidic sequences as splice sites. In this case, non- consensus spliced alignments had to perfectly match 7 bases adjacent to splice sites and were always reported on the positive strand. Introns predicted from at least two spliced alignments (four for non-consensus junctions) and at least 12 nt (30 nt for non-consensus junctions) from the read boundary were used in a second round of alignments to create a library of intron junctions with sequence context 78 nt on each end. Then, in a second round of alignments, each read was aligned to the intron junction library generated in the first step, in order to map reads with previously predicted introns close to the read boundary. (We used version 0.5 of Palmapper with parameters "-M 3 -G 1 -E 4 -l 15 -L 25 -K 8 -C 35 -I 25000 -NI 2 -SA 100 -CT 10 -a -S -seed-hit-cancel-threshold 10000 -report-map-read -report-spliced-read -report-map-region -report-splice-sites 0.9 -filter-max-mismatches 0 -filter-max-gaps 0 -filter-splice-region 5 -qpalma-use-map-max-len 2000 -f sam -threads 2 -polytrim 40 -qpalma-prb-offset-fix -include-unmapped-reads -min-spliced-segment-len 12 -junction-remapping-coverage 2 -junction-remapping <junction-file>"). In total we aligned 180 million reads (95%), including 146 million uniquely aligned reads and 35 million reads where the best alignment was spliced. Details for each strain and biological replicate are given in Supplementary Table 14.

We repeated these steps for the additional 178 million strand-specific RNA-seq reads from the two accessions Col-0 and Can-0 (Supplementary Information section 7.5). We were able to align 152 million reads (88%), out of which 127 million reads aligned uniquely and for 37 million reads the best alignment was spliced. The smaller fraction of aligned reads can, among other factors, be attributed to a larger number of reads with a polyA-tail. See Supplementary Table 15 for details.

We previously also generated other versions of these alignments, were we only considered canonical splice sites (introns with GT-AG or GC-AG), allowed 4 mismatches and only recorded the 10 best matches per alignment (Palmapper parameters "-M 4 -G 1 -E 4 -l 15 -L 35 -K 12 -C 45 -I 25000 -NI 2 -SA 10 -CT 10 -z 10 -S -seed-hit-cancel-threshold 10000 -report-map-read -report-spliced-read -report-map-region -report-splice-sites 0.9 -filter-max-mismatches 1 -filter-max-gaps 0 -filter-splice-min-edit 1 -filter-splice-region 5 -qpalma-use-map-max-len 10000 -f sam -threads 1 -polytrim 40 -qpalma-prb-offset-fix"). These alignments were used for the first step of *de novo* genome annotation (see Supplementary section 10). A very similar set of alignments was generated for expression estimates (cf. Supplementary section 11; with Palmapper parameters "-M 4 -G 1 -E 4 -l 15 -L 35 -K 12 -C 45 -I 25000 -NI 2 -SA 10 -CT 10 -z 10 -S -seed-hit-cancel-threshold 10000 -report-map-read -report-spliced-read -report-map-region -report-splice-sites 0.9 -filter-max-mismatches 1 -filter-max-gaps 0 -filter-splice-region 5 -qpalma-use-map-max-len 10000 -f sam -threads 4 -polytrim 40 -qpalma-prb-offset-fix"). Higher quality alignments became necessary in downstream analyses leading to refined sets of alignments used at different

stages of the project. For instance, we used the highest quality alignments for consolidation and detection of alternative splicing. These alignments are available in BAM format from the supplementary website in both strain and reference coordinates.

10. Genome Annotation

10.1 Overview We annotated each of the 18 genomes in three stages, including (1) using solely the genome sequences (assemblies) of each accession to make *ab initio* gene predictions, (2) incorporating RNA-seq reads for *de novo* predictions, and finally (3) by consolidation with the TAIR10 annotation (Supplementary Fig. 20 shows an overview). Each step produced a different set of annotations, which were compared using the F-score metric with the TAIR10 annotation mapped to the accessions' coordinate systems (see Supplementary Table 16). The F-score is calculated as the product of the recall (the true positive rate) and the precision (the fraction of true positives in all positives) of a method divided by the mean of recall and precision.

10.2 Ab initio annotation We used mGene[44] to train *ab initio* predictors that recognize genomic signals (e.g., splice sites) and genomic segment types (e.g., introns and exons) based solely on the genomic DNA sequences, using the annotated (TAIR9) protein-coding genes in a five fold cross-validation scheme[44]. We then predicted these features in each of the 18 genomes *de novo*. We mapped the cross-validation split information for all signals to the accession coordinates to obtain unbiased predictions. We used 7,500 annotated protein-coding genes (TAIR9) to train the mGene system on the reference genome (Col-0). The trained system was then used to annotate the 18 accessions (and re-annotate Col-0). On average 26,292 genes were predicted per accession. The average transcript-level F-scores were high (58%). We further used this annotation-centred approach as a baseline to assess the RNA-seq-based annotation methods developed in the context of this study (Supplementary Information section 10.3).

10.3 De novo annotation using RNA-seq alignments We produced a second set of predictions by combining the *ab initio* predictions with RNA-seq read alignments, in particular, we predicted introns from spliced alignments and regions with overlying RNA-seq read alignments (see Supplementary Fig. 19 for an outline of the method)[46]. We used 7,500 annotated protein-coding genes (TAIR9) and ≈12 million aligned RNA-seq reads from Col-0 mapped to TAIR10 to train the system, which was then used to annotate the 18 accessions and Col-0 (see above and Supplementary Table 14). On average, 24,681 genes where predicted per accession. The average exon-level and transcript-level F-scores were 86% and 63%, respectively. The increase in the transcript-level F-score by 5% was due to both the incorporation of RNA-seq read alignments, and that fewer lowly expressed genes were predicted.

Supplementary Table 17 compares the Col-0 *de novo* predictions to TAIR10; mGene.ngs is more accurate (transcript F-score 65.2%) than predictions based on genome sequence alone (mGene[44]; 59.6%) or RNA-seq alignments alone (Cufflinks[47]; 37.5%).

The RNA-seq-based gene predictions are available from the supplementary website and the GBrowse track "De novo gene annotations by mGene.ngs".

10.4 Consolidated annotation We first identified "units" of orthologous annotated genes across the accessions and TAIR10. We use the word unit rather than gene because it is possible that a TAIR10 gene might be split into several transcriptionally independent units

(but no unit contains more than one TAIR10 gene). The coordinates of each accession's predicted transcripts were translated to TAIR10. All overlapping transcripts were treated as a unit, and duplicates (with identical exon/intron structure) were removed. The TAIR10 gene classification (protein-coding, transposable element gene, pseudogene, etc.) was passed onto the unit. We then back-transformed each unit to the original accession coordinates and restored lost accession-specific features (e.g., exons absent in TAIR10). We removed transcripts from a unit if splice site consensus sequences in an accession's genome were absent, or if they overlapped with more than one TAIR10 gene. We incorporated any TAIR10 annotated transcripts into their corresponding units and also mapped these back to each accession, removing invalid transcripts where a splice site consensus was missing in the accession. If the resulting unit contained at least one valid TAIR10 annotated transcript/isoform, it was merged with transcripts in the unit provided that all introns were confirmed by RNA-seq alignment or were part of an annotated TAIR10 transcript, and the transcript contained at least one intron different from the annotated introns. If the unit did not contain a TAIR10 transcript then the predicted transcript with the highest number of confirmed introns was chosen. In rare cases there was no transcript in the unit (either they were all removed or the gene was missed in *de novo* gene prediction). However, in most cases (≈99%) there was at least one consolidated transcript for each TAIR10 gene in each accession's genome.

We noticed a number of remaining RNA-seq-confirmed introns that were not part of any annotated transcript. With the aim to integrate the introns with highest confidence, we constructed a strictly filtered set of introns from spliced RNA-seq read alignments. We required the intron to be confirmed by at least two reads with the split position at least 12nt from the read boundary and matching with at most 1 mismatch. For each such intron, we identified exons in transcripts with boundaries at most 50 nt away from the intron boundary. For each case a new transcript including this intron (or otherwise the previous transcript structure) was generated. The coding sequences of protein-coding transcripts were determined as follows: The longest region without stop codon in the mRNA was identified. We then checked whether the 5' and 3' ends of this region coincide with the transcript boundaries. If this condition was violated, we used the first in-frame occurrence of the translation initiation signal (TIS) consensus ATG as start and the first stop codon (TAA/TAG/TGA) as end of the coding sequence. If the 3' end of the region coincided with the transcript end, we extended the 3' end of the last exon by at most 300nt in order to terminate the open reading frame. This modification of the transcript was necessary, as the transcript ends were often predicted or annotated too short to include a suitable in-frame stop codon. If the 5' end of the region coincided with the transcript start, we checked the region 300 nt upstream of the transcript for a suitable alternative TIS consensus signal and used it, if the length of the implied coding region increased by at least 100 nt. For consistency, we also applied this strategy to the annotations of the reference accession Col-0. The statistics of the resulting consolidated accession-specific gene annotation are given in Supplementary Table 18. The supplementary website contains files in GFF3 and the GBrowse genome browser shows tracks with the consolidated gene annotations (track name "Consolidated gene predictions").

10.5 Assessment of Consolidated Gene Annotations We used additional high-coverage 82 bp strand-specific RNA-seq data for accessions Col-0 and Can-0 to assess the quality of the consolidated annotations. This additional RNA-seq data was solely used for validation purposes and not for prediction (see Supplementary Table 15). We refer to this data as strand-specific validation read data (SS) and to the other RNA-seq data as non-strand-specific annotation read data (NSS). We considered five different annotations: a) the *ab initio* gene

predictions by mGene, b) the *de novo* gene prediction using RNA-seq data, c) the TAIR10 annotation mapped to the accession's coordinate system, where we removed transcripts without valid intron junctions, d) the consolidated gene prediction based on the non-strand-specific annotation read data and e) for the purpose of comparison, consolidated gene prediction based on the validation read data. We include annotation e) for comparison purposes only and to estimate how much better the annotation might be if we had used more RNA-seq data during the consolidation step.

Based on the validation data, we could verify transcript structures, in particular intron junctions. We defined two sets of introns based on the validation data: a strictly filtered set SS-S, only containing introns confirmed by at least 10 reads with no mismatches and with splice positions at least 20 nt from the read boundary, and a set SS-R with relaxed filtering (only one confirmation, at most one mismatch, split position at least 8 nt from read boundary). Supplementary Table 19 reports, for each annotation step (a)-(e), the numbers of introns confirmed by RNA-seq read alignments (set SS-R). Between 83-88% of all introns of any of the annotations were confirmed by RNA-seq. We also determined which predicted introns do not appear in the TAIR10 annotation ("novel introns"), and which had RNA-seq support (set SS-R). For *ab initio* and *de novo* predictions many new introns were suggested, which, however, cannot be confirmed. Introns in consolidated gene predictions (NSS) can largely be confirmed, 99% for Col-0 and 82% for Can-0. The smaller fraction for Can-0 can be explained by new transcripts that replace TAIR10 transcripts and were not valid after mapping to the Can-0 genome. The large number of confirmed novel introns shows the advantage of the consolidation strategy of combining the existing TAIR10 annotation with *de novo* gene predictions.

We further identified which TAIR10 introns were absent in the annotations (a)-(e). For *ab initio* and *de novo* predictions, about 20,000 introns were erroneously missing on average per accession. Many were confirmed by RNA-seq, indicating incompleteness of these annotations. For Col-0, the consolidated gene predictions always contained the TAIR10 transcripts as a subset, and hence, there were no introns missing. For the consolidated annotations of Can-0, ≈1,850 introns were missing. However, only a small fraction was confirmed by RNA-seq (≈170 introns, ≈9%). This indicates that the consolidated annotations do not miss many transcripts/introns compared to the mapped TAIR10 annotation.

Finally, for the strictly filtered intron sets SS-S and NSS-S, we identified those introns that were confirmed with high confidence from the RNA-seq read alignments in each of the annotations. For the NSS set, most of the 70,000 and 76,000 introns (for Col-0 and Can-0, respectively) were present, with the fewest number of introns not represented in the consolidated gene annotations and the TAIR10 annotation. For the SS set, however, thousands of high-confidence introns were not part of any annotated transcript, particularly for *ab initio* and *de novo* predictions. Most of these introns were within the boundaries of known genes and provide evidence of extensive alternative splicing. This indicates that, based on the NSS read data alone, which was generated only from seedling (the deeper RNA-seq validation data set was from three tissue types), it is very challenging to provide complete annotation of alternative isoforms.

10.6 Analyses of the predicted RNA transcript and amino-acid sequences For each consolidated gene annotation in each accession, we determined the nucleotide sequences of the whole transcript as well as of the coding region and the predicted amino-acid sequence (FASTA files with all sequences are available from the supplementary website). *Amino acid sequence changes:* We employed the Myers-Hirschberg algorithm implemented in the Sean toolbox[48] to align the I_{gA} amino-acid sequences S_{gAi} $(i = 1,...,I_{gA})$ of each gene g in the consolidated annotation of accession A with all I_{gB} sequences S_{gBi} $(i = 1,...,I_{gB})$ of the

corresponding gene of another accession *B*. The distance between the sequences in gene *g* in the two accessions *A* and *B* is computed as

$$D_{g,A,B} = \sum_i \min_j \{Id(S_{g,A_i}, S_{g,B_j}) / \max(|S_{g,A_i}|, |S_{g,B_j}|)\}$$

where *Id(X,Y)* denotes the number of identities in the global alignment of the two sequences *X* and *Y*, and $|X|$ is the length of X. *Nucleotide sequence changes:* We repeated the outlined strategy for the RNA transcript sequences (including 5' UTRs, CDS regions and 3' UTRs for protein-coding genes).

For Figs 2c and 4a and Supplementary Figs 24 and 26 we computed the average AA-distance between sequences from pairs of accessions for every gene. For protein-coding and pseudogenes we considered all possible pairs. For the *A. lyrata* comparison, we computed the AA-distance between an *A. lyrata* AA sequence and the best matching Col-0 AA sequence. When considering disruptions (Fig. 2c & Supplementary Fig. 24), we only considered pairs involving at least one accession with the considered disruption. The figures show the fraction of genes with average distances in different intervals.

10.7 Clustering of amino-acid sequences For each protein-coding gene *g* we used the distances D_g between the accessions' amino acid sequences to cluster accessions with similar sequences. We used single-linkage clustering as implemented in the MATLAB statistics toolbox to determine the hierarchical similarity between the sequences (see Fig. 2b for an example). We identified groups of almost identical AA sequences (termed *isoform* in the main text) by grouping accessions with AA-distances smaller than 2.2% on the same cluster. Accessions from within the group differ by at most 2.2% of their corresponding AA sequence. The threshold was determined by considering the 5% largest AA-distances of accession pairs of protein-coding genes without large effect disruptions (i.e., among the genes and accession pairs with expected small changes, we consider 95% as not sufficiently different). We determined the group size for each gene and accession they are part of and computed the relative frequency of the group sizes, which are shown in Fig. 2d and Supplementary Fig. 25.

10.8 Annotation of novel genes To identify genes not annotated in TAIR10, we combined the *de novo* gene predictions from mGene.ngs with transcript predictions from Cufflinks[47] using default parameters. For each gene unit, we mapped the transcript predictions to the reference genome and excluded genes overlapping any TAIR10 annotated gene, pseudogene or transposable element gene. In addition, only transcripts for which at least 50% of nucleotides were covered with RNA-seq reads were retained. We merged overlapping transcripts on the same strand into a single unit and removed duplicates. This led to 496 novel units/genes consisting of 1,352 transcripts with an average transcript length of 443 nt (max 2,988 nt). We then transformed the TAIR10 coordinates back to the accessions, removed transcripts with missing splice consensus and determined the longest open reading frame, obtaining on average 459 new genes per accession with an average transcript length of 443 nt. We determined the longest transcript variant over all accessions and isoforms (average length of 531 nt) and aligned the transcript sequences to the NCBI EST database (excluding human and mouse ESTs) using BLAST[49], and obtained 293 matches (221 sequences match to *Arabidopsis thaliana* ESTs best). This adds evidence of expression for 293 transcripts not annotated in TAIR10. When excluding matches to ESTs from *Arabidopsis thaliana*, 72 matches remain, most of which against ESTs from *Brassica napus* (16 times best hit), *Raphanus raphanistrum* (12), *Brassica rapa* (7), and *Brassica oleracea* (7), and *Raphanus sativus* (5). We aligned the 203 transcripts without EST matches (average length 456 nt) against the non-redundant protein database using BLASTX[50] with e-value cut-off 0.01. We

found matches for 70 sequences, most of which matched to known or hypothetical proteins in *Arabidopsis lyrata* (35 times best hit) and *Arabidopsis thaliana* (28).

We generated a consolidated set of novel genes by reducing redundancy among the predicted transcripts. For this we generated a splicing graph (see Supplementary Information section 10.10) and then generated all possible transcripts from the graph (typically leading to a significantly smaller number of transcripts). For each transcript we inferred the maximal open reading frame if it was found to be longer than 100 bp (non-redundantly leading to 226 ORFs). For each novel gene and accession we then verified that there is sufficient NSS RNA-seq evidence for expression (see Supplementary Information section 10.7). On average per accession we found 228 novel genes with evidence for expression (284 transcripts and 77 introns). A large fraction (≈94.3%) of the novel genes could be validated using the independent SS RNA-seq data.

10.9 Novel Genes in Sequence Missing from TAIR10 We also searched for additional novel genes that were entirely absent from the TAIR10 genome sequence. For each of the 18 genomes, we took the contigs we had previously obtained by *de novo* assembly and aligned them to TAIR10. We worked with the *de novo* contigs to include sequences that might not have been assembled into the 18 genomes. All sequences longer than 50 bp within these contigs that did not match TAIR10 were extracted. We aligned each novel sequence to all the other novel sequences using BLAT[51].

Since many novel sequences are shared by several accessions, but with variable lengths, we created a non-redundant set of novel expressed sequence sets by using the RNA-seq data as a guide. First, all RNA-seq reads for the 18 accessions were mapped to TAIR10 reference and annotated using TopHat[52]. All remaining unmapped reads were then aligned to the novel sequences above using TopHat; each novel sequence was given a score based on the number of RNA-seq reads that map to it. Each of the 18 genomes was aligned independently, so that the same RNA-seq read could match to multiple novel sequences from different accessions, but could only match one novel sequence within an accession. We then sorted all novel sequences across all 18 genomes by RNA-seq score. The sequences with the highest scores were then selected in decreasing order, applying a filter that a sequence was omitted if over 90% of the reads matching a given sequence already matched to other novel sequences higher up the list. The procedure continued until all novel sequences were processed.

Next, all RNA-seq reads from the 18 accessions that were not previously mapped to TAIR10 were mapped to the novel sequence dataset using TopHat. Subsequently, Cufflinks[47] was used to predict all possible transcripts. The novel genes were then assembled using the predicted exon positions. We identified 221 novel genes, with an average length of 640 nt (available from our supplementary website). We then aligned these genes to the NCBI EST database using BLAST[49], and obtained 83 matches (average length of matching sequences was 596 nt). Of these, 62 best match ESTs from other *Arabidopsis* accessions, such as Ler, Ws, Wassilewskija, and Ei-2. 26 sequences matched ESTs from other plants or algae (but not ESTs of *Arabidopsis thaliana*) such as *Raphanus sativus* (3 times best hit), *Chlamydomonas reinhardtii* (3), *Brassica oleracea* (7), *Quercus robur* (1), *Saccharum officinarum* (1), *Eutrema halophilum* (1) and *Brassica napus* (1). We aligned the 138 transcripts without EST matches (average length 665 nt) against the non-redundant protein database using BLASTX[50] with e-value cut-off 0.01. We found matches for 91 sequences, most of which matched to known or hypothetical proteins in *Arabidopsis lyrata* (43 times best hit), *Arabidopsis thaliana* (40), *Vitis vinifera* (2), *Brassica rapa* (2), *Dictyostelium discoideum* (1), and *Arabidopsis arenosa* (1).

10.10 Construction of splicing graphs based on RNA-seq alignments We identified intron retention and exon skip events for each accession using splicing graphs, as implemented in mGene-Toolbox[44] with extensions described in[53]. For each gene we constructed a splicing graph[54] representing the set of all possible transcripts, initialised with the exons of annotated transcripts as nodes, and edges connecting nodes whenever an annotated intron connects two exons. For each accession, we generated a splicing graph that was then modified to reflect any accession-splicing events supported by the RNA-seq alignments (we only used uniquely mapped reads for this analysis). We discriminated the following events: *Intron retention:* For each annotated intron, we checked whether at least 75% of the intron was covered with RNA-seq reads and that the average alignment coverage was between 10% and 120% of the coverage of flanking exons, which must have less than 4-fold difference in their alignment coverage. These parameters were chosen to exclude biologically implausible cases. If all conditions were satisfied, the corresponding edge was deleted and the nodes merged accordingly. *Intron retention within annotated exons:* If spliced alignments of at least two RNA-seq reads confirmed an intron that was completely contained in an exon represented in the splicing graph, we extended the splicing graph by introducing two new exons connected by this intron. The 5' end of the generated 5' exon and the 3' end of the 3' generated exon inherit the 5' and 3' edges of the original exon, respectively. *Exon skipping:* For each intron confirmed by at least two spliced RNA-seq alignments, we checked whether it connected two exons in the splicing graph that are not connected by an edge. If so, the graph was extended by a new edge connecting the two exons. *Alternative 5' and 3' splicing:* If only one end of the intron coincided with an exon in the splicing graph, we identified the nearest exon in the vicinity of the other intron end. If the intron and exon end were less than 40 bp apart, we created a new exon in the splicing graph with one shifted end connected to the confirmed intron. The other end inherited all edges from the original exon.

Using the modified splicing graphs, we then identified candidate intron retention and exon skip events. For intron retentions we searched for edges connected to a node that completely covered the intron. For exon skips, we found all triplets of exons where the first was connected to the second and third and the second connected to the third (i.e., the second exon can be spliced in and out). We then combined all candidate events from all accessions by mapping the introns' accession coordinates to TAIR10 and then merging overlapping candidates. *Intron Retentions:* For each candidate in each accession, we checked whether (a) over 75% of the intron is covered with RNA-seq reads and (b) the intron is spliced out, as above. The minimum of the sum of strains for which these conditions hold respectively was used as a confidence level for this candidate. This resulted in 3,425 candidates with confidence of at least 1 (i.e., the event was private to a single accession). For higher confidences, the number was significantly reduced (see Supplementary Table 20). *Exon Skipping:* For each resulting exon skip candidate, we checked for RNA-seq read alignment evidence for (a) the intron connecting exon 1 with 2, (b) exon 2 with 3, and (c) exon 1 with 3. The minimum of the three numbers defined a confidence measure for this candidate. This resulted in 205 candidates with confidence of at least 1. For higher confidences, the number is significantly reduced (see Supplementary Table 21). The detected alternative splicing events with confidence 1 and higher are available as data files with supporting quantitative information from the supplementary website.

10.11 Reproducibility of alternative splicing event detection We repeated our procedure for detecting alternative splicing events separately on each of the two biological replicates of RNA-seq data from each accession, and measured how the sets of retained introns and skipped exons overlap between the two replicates and with the set derived from both replicates. First, the number of confirmed intron retention events drops significantly to 47%

(more for higher confidences) due to the lower read coverage, which crucially influences the detection strategy (see Supplementary Table 20). For exon skips, the drop is less severe (59-60%; see Supplementary Table 21). The overlap of detected retained introns and skipped exons between replicates is between 62% and 66% at confidence level 1 or higher. At higher confidence levels the overlap ratio increased, but the number of events was significantly reduced. Therefore, for future analyses, we suggest using intron retention and exon skip candidates with confidence of at least 3 (2,545 retained introns and 155 exon skips, both with 69% reproducibility). These overlap ratios are a good estimate of how well the current procedure on both replicates will be reproducible when repeating the experiments and analyses.

10.12 Comparison of annotated alternative splicing events with TAIR10 We also computed splice graphs for the annotated genes in the TAIR10 annotation and found 2,791 intron retention and 621 exon skip events, of which 515 and 53 overlapped with intron retention and exon skip events, respectively, that we detected from RNA-seq data. A similarly small overlap of events predicted from tiling arrays and RNA-seq data using a broad panel of tissues with annotated events was observed previously in[53]. We scored each annotation event as before and confirmed 1,111 and 197 events, respectively, using RNA-seq read alignments from both replicates. Supplementary Tables 20 and 21 show the overlap between the events at different confidence levels. At low confidence (1 and higher) we find overlap ratios of 41% and 25%, which increases to 60% and 38% for confidence 3 and higher and reaches 95% and 80% for the highest confidences for intron retentions and exon skips, respectively.

10.13 Consolidated set of alternative splicing events For subsequent analysis, we created a consolidated set of alternative splicing events containing newly detected events as well as events annotated in TAIR10. We chose a confidence cut-off of 3, i.e., required that each event had RNA-seq evidence in at least three accessions. In total we obtained 2,819 intron retention events and 229 exon skip events.

11. Quantification of gene expression

11.1 RNA-seq read counts We quantified expression for 65,238 annotated features (i.e., genes, pseudogenes, transposons and others), including those obtained from the *de novo* annotation (Supplementary Section 10), separately for each strain and biological replicate. We removed all RNA-seq reads with ambiguous mappings within their respective accessions, that is, if any alternative match had a Hamming distance within at most two units (i.e., at most 2 mismatches or indels) from the best alignment's score. We then removed any suboptimal alignments from the remaining reads, such that each mapped read had a single placement. On average per accession, 7.7 million reads could be aligned uniquely (81% of aligned reads; see Supplementary Table 14). We counted the number of reads mapping within each exon of the TAIR10 annotation mapped to the accession's genome coordinates, excluding exonic regions that overlapped with another annotated gene. In total, between 1,241,437 and 4,920,935 reads were used to estimate gene expression in each strain and replicate. Following[55], we used the raw mapped RNA-seq counts to estimate individual library sizes, and summed across all protein-coding genes, excluding pseudogenes, transposable element genes, and non-coding RNAs (in particular, tRNAs and rRNAs). For final expression estimates, we assessed the impact of dropping multiple mapping reads, which overall was minor. Expression estimates only changed for 631 genes (<1%) in at least

one accession (FDR 5%, see Supplementary Fig. 39), when not discarding multiple mapping reads.

11.2 Numbers of expressed genes per accession To estimate the number of expressed genes, we fitted a model to the observed counts that allows for a small number of reads that map to unexpressed genes, e.g., due to leaky transcription. For this background model of transcription, we assumed that the number of reads observed by chance in any one gene follows a Poisson distribution with a fixed rate λ, irrespective of the gene length. Under this model, the probability that no read maps to a single gene is proportional to $e^{-\lambda}$. We fitted the rate parameter λ from the observed distribution of unexpressed annotated features, setting λ to the negative log of the fraction of features with no mapped reads (estimates ranged from λ =0.33 to 0.39, in individual strains). For each annotated feature and in each strain, we used this fitted model to calculate the p-values of observing the measured gene expression count from this background model, i.e., the p-value that the feature was not expressed. Genome-wide significance was estimated using q-values[56]. In total, between 18,598 and 19,593 protein-coding genes were detected as being expressed (see Supplementary Table 22); overall, there were 21,381 protein-coding genes with significant evidence for expression in at least one strain (q-value<0.05).

11.3 Variance function estimation Several analyses rely on an estimate of the technical and biological variability of the expression levels. Due to the count nature of RNA-Seq expression estimates, this variance is not constant but varies as a function of the mean level of the expression level itself (for a detailed discussion see for example[55]). For each gene in every strain, we estimated the expression level as the mean expression count across replicates and the variance from the difference between the replicates (see Supplementary Information section 11.1). Based on all these expression/variance pairs, we then fitted a global variance function, mapping the expression estimates (the mean fitted between replicates) to the empirical variance (variance between the replicates) using the DESeq package[55]. To improve the stability of this fit, we only considered coding genes. Importantly, the resulting variance function accounts for over-dispersion, combining technical variability and biological variability. This estimated variance function was used for testing differential gene expression (see Supplementary Information section 11.7) and for the variance stabilizing transformation of the raw counts. The latter was carried out using the DESeq package[55].

11.4 Gene structural variation filter To understand the impact of gene structural variation on expression and to distinguish it from overall level of expression, we also estimated expression after excluding genomic segments that were unreliable or polymorphic (e.g., an exon that was deleted in some accessions). Thus, we mapped all uncovered or deleted regions to the reference genome, calculated the union across accessions and mapped the resulting segments back onto the respective accessions (affecting a total of 20.425 Mb overlapping exonic regions, including exons of annotated genes, pseudogenes, transposable element genes and transposable elements). Reads mapping onto those regions were excluded for expression estimations, reducing the total length of analysed exonic regions from 75.476 Mb to 55.052 Mb. Similarly, to account for structural variation between gene models, we restricted exonic regions to the intersection of all overlapping exons when mapped onto reference coordinates and then mapped back to the respective accession coordinates (based on de novo gene predictions produced by mGene.ngs). Therefore, only regions annotated as exonic in all accessions were considered for expression estimates, retaining a total of 54,870 Mb of exonic regions.

11.5 Effects of gene structural variation on gene expression We compared gene expression estimates with and without using the structural variation filter. After filtering for structural variation, 457 protein-coding genes (3,550 when including transposable elements) had a resulting quantification region of length zero, for example, because the gene had been deleted in at least one strain. Next, we investigated to what extent the expression estimates changed for genes that remained measurable after applying the more stringent read filter. Because the filter systematically reduces the resulting counts, we rescaled the counts in the filtered dataset by the ratio of the number of positions used for the unfiltered versus the filtered expression estimates. A scatter plot of the uncorrected versus the corrected estimates of gene expression as raw counts is shown in Supplementary Fig. 33. We then tested for significant differences between expression estimates with and without the structural variation filter, using a negative binomial model. Genome-wide significance was estimated using q-values[56]. This test used the variance function estimated earlier and was implemented using the DESeq package[55]. In total, 425 annotated features (264 protein-coding genes) had an expression estimate that deviated significantly when applying the gene structural variation filter, suggesting that structural variation did play a causal role for expression variability in these genes (see Supplementary Table 23).

11.6 Heritability We used the variance-stabilized expression levels to estimate the variability between accessions σ_B^2 and between biological replicates (within accessions) σ_W^2. Following the approach taken in Keurentjes *et al*.[57], we calculated broad sense heritability as the ratio of the empirical estimate of the variance between accessions divided by the total variance $\sigma_B^2 / (\sigma_B^2 + 2\sigma_W^2)$. Here, the factor 2 accounts for the number duplicate measurements per accession. To allow for a meaningful comparison of the within variability (from biological replicates) and the between variability (variation across strains), we used variance-stabilized estimates of the expression counts where the noise variance is approximately constant as a function of the expression estimate (see Supplementary Information section 11.3).

Next, we investigated differences of the distribution of heritability across gene types, distinguishing between protein-coding genes, newly predicted genes, pseudogenes and non-coding RNAs. We estimated cumulative distribution functions (CDF) of the heritability for each respective gene type from expressed genes. To account for intrinsic uncertainty due to finite sampling, we performed 10-fold cross validation, repeatedly estimating the CDF from 90% of the genes. The estimates were combined in a mean CDF and empirical uncertainty of the CDF as plus or minus one standard deviation error bars (Supplementary Fig. 27). We found protein-coding genes to be most heritable, followed by newly predicted genes, pseudogenes and non-coding RNAs. Reassuringly, the heritability of newly predicted genes was similar to those from established gene models, suggesting they are genuine. The remaining differences in heritability between these genes can be explained by the shorter length of newly predicted genes; very short genes were generally found to be less heritable, see Supplementary Fig. 28.

11.7 Testing for differentially expressed genes We used DESeq[55] to test for differential expression of all 65,238 annotated features between each pair of accessions. We used the previously estimated variance function, thereby accounting for technical and biological variation. We conservatively accounted for multiple testing ($19 \cdot 18/2 = 171$ tests per gene), employing a Bonferroni correction, yielding a p-value for differential expression for each gene. Genome-wide significance was estimated using q-values[56]. At 5% genome-wide FDR, we found 9,786 annotated features (9,360 protein-coding genes) that were significantly differentially expressed (Supplementary Fig. 31). For a breakdown of differential expression

across gene categories and families, see Supplementary Table 23, Fig. 4, and Supplementary Fig. 39).

For comparison, we also considered an ANOVA test on the variance-stabilized expression estimates to test for differential expression. Overall, we found good agreement for top ranked genes (Supplementary Fig. 29). In the downstream analysis, we considered estimates of differential expression from the pairwise testing approach, because it yields additional information, such as the number of strains that participate in a particular expression pattern, allowing for differentiation of private vs. common patterns of differential expression.

11.8 Private vs. common differential gene expression We extended the concept of minor allele frequency to gene expression levels in order to classify genes according to whether their expression differences were common or private to a few accessions. For each gene, we iteratively removed the accession that reduced the significance of differential expression the most when discarded. This yielded a series of significance estimates for increasing minor allele frequency cut-off values. We found genes with rare differential expression patterns to be over-represented in some gene families. See main text, Fig. 4d, and Supplementary Fig. 41 for detailed discussion.

11.9 Effect of copy number variation on gene expression For each gene in each strain, we estimated the copy number from the coverage of genomic reads, accounting for strain-specific deletions and insertions. Thus we counted the number of read starts for each gene, discarding regions that were deleted in at least one strain (similar to Supplementary Information section 11.1). To estimate a gene-specific copy number, we normalized counts by the median of the total number of reads across genes for each strain, which accounts for differences in library size. We used DESeq on the raw read counts to test for differential copy number variation between strains. In total, we found 6,121/65,238 (9%) genes at a FDR <5%, to exhibit differential copy numbers between at least one pair of strains. Newly predicted genes and pseudogenes were among the gene categories with the greatest extent of copy number variation (Supplementary Fig. 34). It is notable that a minority of gene families carries the majority of genes with variable copy numbers, with NB-LRR genes and B3 transcription factors standing out the most. Despite the considerable overall level of copy number variation, the impact on gene expression was low. Among the 6,121 genes with copy number variation, 388 genes were expressed and 237 were differentially expressed (Supplementary Figs 35 and Fig. 36). Among the 237 variable genes, we could attribute copy number variation as a cause of expression variation for 54 genes (see Supplementary Information section 12.4). Therefore overall, our characterization of gene expression variation is only marginally affected by copy number variation.

11.10 Testing for differential alternative splicing Analogously to the analysis of gene expression levels, we analysed the predicted intron retention and exon skip events to assess the significance of differences between accessions. We scored intron retention events by the number of RNA-seq reads supporting the intron retention relative to the number of reads that only map into the flanking exons, For this test, read counts of both biological replicates were additively combined. Similarly, exon skips were scored as the number of reads supporting the exon skip relative to the number of reads that only map into the flanking. We tested for differential intron usage and exon skips between accessions using Fisher's exact test. For each alternative splicing event, the pairwise tests between strains were combined into a single p-value for differential alternative splicing using the Bonferroni correction. Genome-wide significance levels of the events were estimated using q-values. At significance level 5%

FDR, we found 142 out of 2,819 (5%) intron retentions to show different degrees of prevalence between strains. Similarly, 17/229 (7%) of the exon skips occurred with differing rates between the strains. The set of differentially expressed splicing signatures is available from the supporting website.

11.11 GO analyses Gene Ontology (GO) categories and their mapping onto genes were obtained from http://www.geneontology.org/. Statistically significant enriched or depleted categories were determined using GOrilla[58]. Enrichment of GO terms in a set of 9,360 differentially expressed coding genes (DESeq, FDR < 0.05) was carried out using the Gene Ontology Enrichment Analysis and Visualization Tool[58,59]. Of the differentially expressed genes, 5,760 were recognized and associated with a GO term, with 180 differentially expressed genes present in enriched GO term (biological process) lists as assessed with a P-value < 10^{-3} (Supplementary Table 24**)**.

12. Genetic Association of Gene Expression

12.1 Overview Association tests between genotype and gene expression variables were based on a reduced representation of the genotype variants table. Owing to the small sample size, genetic variants between strains were binarized, differentiating between the most frequent allele (major) and the second most frequent allele (minor). Any additional alleles were marked as missing and excluded. We further deleted variants with minor allele frequency less than four. Genome-wide, this left 989,999 variants for analysis. Genome-wide scans for polymorphisms with a *trans* regulatory role on gene expression were not possible in a sample of only 19 accession, thus, we restricted analysis to proximal *cis*-acting variants associated with expression variation.

12.2 **Nucleotide variant *eQTLs*** We considered each of the 21,837 expressed genes (Section ***11.2)*** as candidates for possible *cis* regulation by nucleotide variants. For each gene, we restricted the analysis to variants within a local window of 30 kb up- and downstream of the gene. This resulted in between 7 and 2,110 genetic variants (average 495) per gene, each of which we tested for association with the expression profile of the corresponding gene. The expression trait was defined at the level of whole gene expression (not individual exons) using the variance-stabilized estimate of gene expression (see Supplementary Information section 11.1). Biological replicates were explicitly included in the analysis. For each individual test, we fitted a standard linear model and alternatively a mixed linear model (see Section 13), where the latter accounts for confounding variation (see Section 13), between the genetic variant and the expression levels, and calculated the likelihood ratio score (LR score). For downstream analyses, we used results from the mixed model, although likelihood ratios were generally in good agreement, suggesting little impact of confounders (see Section 13). We employed the Westfall-Young correction to account for multiple testing, resulting in an overall p-value of *cis* regulation for each gene. In this permutation-based correction approach, we used the maximum LR across the set of tested markers and estimated corrected p-values from the relative ranking position in the tail of an empirical null distribution. This empirical distribution was obtained from the maximum LR from 10,000 permuted datasets, shuffling the order of the accessions in the phenotype relative to the genotype. Permutations were stratified to respect the structure of biological replicates. Based on this set of p-values, genome-wide significance across genes was estimated using q-values.

Genome-wide, we found 889 *cis*-eQTLs (5% FDR) between nucleotide variants and expression traits. Most associations were found for protein-coding genes (818) (see

Supplementary Table 23 for a breakdown across gene types). We investigated the distribution of positions of the most strongly associated variant within each eQTL relative to the gene start, discarding any associations that fell into the coding region of a second gene. This left 647 associations to protein-coding genes. Associations tended to cluster near the gene start, with the majority of them found in the upstream core promoter region or in the beginning of the gene (Fig. 3c). We used the genome annotations to dissect the location of the strongest associations relative to the corresponding gene model by tabulating the frequency of associations occurring in specific regions of the gene model (Fig. 3d). Estimates were obtained as raw association counts, normalized by the fractional length of any one region relative to the total length of all considered regions. The density of associations is shown. The full set of *cis*-acting nucleotide variant eQTLs, including information on their position and alleles, is available from the supplementary website.

12.3 Large effect eQTLs In addition to individual polymorphisms, we also tested for associations between large effect changes in genes specific to accessions and expression differences. We used the large effect changes previously identified (Supplementary Information section 8.2**)** and created binary phenotypes that represented the presence or absence for each of the 7 large-effect categories (TIS site, stop site, frame shift, premature stop codon, splice disruption, ORF splice disruption, CDS_len disruption) in each accession and for each gene. We removed genes without at least one large effect change, leaving 3,789 genes for analysis. Testing was implemented analogously to nucleotide variant eQTLs (Section 12.2). Out of 3,789 genes we found 220 significant associations to the large effect indicator (FDR 5%). Half of these large effect eQTLs were found in genes that also had a gene expression nucleotide variant *cis*-eQTL. This non-perfect overlap is expected, because the large effect variants used for testing represent a composite genotype. Because of the dominating overlap with nucleotide variants, large effect eQTLs were not investigated further. The set of large effect QTLs is available on the supplementary website.

12.4 Copy number variation eQTLs Genetic association tests between copy number variation and gene expression were carried out similarly as the approach taken for large effect eQTLs. The linear association test was done between the copy number estimate in each strain (Section 11.9) and the corresponding gene expression profiles. Out of 9,786 differentially expressed genes, we found 67 significant associations between the copy number variation and the expression profile (FDR 5%). This analysis supports the conclusions drawn in Section 11.9, i.e., that copy number variation has scarce relevance for gene expression variability. Notably, the few genes that were associated with copy number variation tended to exhibit high fold change variation (main document Fig. 3).

12.5 cis eQTLs We define *cis* eQTLs as the composite of possible direct proximal genetic effects. These include nucleotide variant associations (Section 12.4), copy number variation eQTLs (Section 12.5) as well as genetic variation due to gene structural variation (Section 11.5). At FDR 5%, there were 1,207 genes with a *cis* eQTL association (1,016 protein-coding genes). A complete breakdown across association types and gene categories is provided in Supplementary Table 23.

12.6 Splicing QTLs Genetic association tests for intron retention and exon skips were carried out similarly to the approach taken for nucleotide variant eQTLs. As for differential testing (Supplementary Information section 11.10)**,** we estimated splicing phenotypes as the ratio of reads that support the alternative splicing events versus reads in flanking exons. We tested for association with variants in a region of 10 kb up- and downstream of the intron. Again,

significance was assessed using a Westfall-Young correction followed by q-value-based approach for genome-wide significance levels. In total, we found 34 associations with intron retention variation at significance level FDR 5%. We repeated this analysis for the 229 exon skip events in the consolidated set. Here, we found only 4 *cis*-associations at FDR 5%. Due to their small number, we did not analyse the exon skip QTLs further. The genetic causes for intron retention were predominantly found within the intron itself (Supplementary Fig. 32). We also checked, that these associations were true splicing events and not confounded by expression variation. For this, we investigated the distribution of splicing QTLs relative to the start of the corresponding genes. Reassuringly, the clustering of associations near the gene starts we observed for expression QTLs (Fig. 3c,d) did not carry over for splicing events. This implies that the identified QTLs are likely true regulators of the splicing event itself. The set of cis-splicing QTLs is available from the supplementary website.

13. Impact of Confounding Factors on Gene Expression Variability

We attempted to minimize confounding sources of variation that can have a significant effect on the observed gene expression variation. First, we controlled for possible environmental perturbations. Biological replicates of individual plant accessions were grown simultaneously in carefully controlled environments. Second, the genetic design was chosen as to minimize possible confounding population structure. The 18 accessions were selected to be diverse (Supplementary Information section 2). Here, we provide additional analyses to assess the effectiveness of both of these measures.

13.1 Impact of population structure Analysis of association studies have been shown to be susceptible to genetic relatedness between individuals in the study population. This confounding effect can lead to inflated results of an association scan; see for example Atwell *et al.*[60] for a discussion of the relevance of population structure in the context of genome-wide QTL analyses in *Arabidopsis*. Here, we checked the magnitude of population influences on gene expression variation. We estimated a kinship matrix that represented the pair-wise genetic similarity between accessions. From the complete set of 989,999 binarized filtered variants (Supplementary material Section 12.1), we estimated kindship (K_{pop}, Supplementary Fig. 37), using identity by descent as similarity measure. Overall relationships were weak, with the exception being Oy-0 and Po-0, which appeared to be genetically related, as is apparent from ancestry analysis (see genome-wide patterns of ancestry, main document).

Next, we investigated to what extent population structure affected gene expression variation on a genome-wide scale. For each of 21,837 expressed genes, we fitted the expression profile using a random effect model, employing the estimated kinship as the random effect and evaluated the marginal likelihood of the resulting Gaussian covariance model. We repeated the analysis for 10,000 permuted datasets, with the phenotypes shuffled with respect to the kinship matrix. We estimated p-values from the position of the likelihood on the non-permuted test relative to the empirical distribution from permuted tests. Genome-wide significance was assessed using q-values. Reassuringly, there were no strong effects. Only 190 genes had an uncorrected p-value smaller than 1%, and after correction for genome-wide significance, we found 11 genes that suggested some regulation by population effects (FDR 50%). Furthermore, genes with greater evidence for population structure regulation, tended to exhibit lower fold change variation (see Supplementary Fig. 38).

13.2 Confounding expression heterogeneity In addition to population structure, subtle environmental perturbations, sample handling or sample history have all been shown to

significantly contribute to expression variability. These factors induce expression heterogeneity (e.g., [61]), i.e., a confounding correlation structure between the samples. If not accounted for, expression heterogeneity can substantially alter the outcome of downstream analysis, including eQTL scans (see for example [61,62,63]). We used Bayesian Factor analysis (implemented in PEER)[63] to estimate expression heterogeneity in the expression profiles. This approach is based on fitting a bilinear model, slightly generalizing methods like principle component analysis, to identify hidden factors that account for gene expression heterogeneity. We used the expression profiles of all expressed genes (21,837) from the 19 accessions and biological replicates to fit PEER. We set the maximum number of factors to 19, the number of accessions, and otherwise used the default settings. After training, PEER retained 4 variable hidden factors $\boldsymbol{X}$ (as determined by automatic relevance determination[63]). Taken together, these factors explained 21% of the total gene expression variance. Similar to kindship, the inner product of the estimated confounders implies a confounding similarity structure between strains: $K_{\mathrm{expr}} = XX^T$.

13.3 Accounting for confounders in eQTL scans We used mixed linear models (as in EMMAX[64]) to estimate confounding structure within the eQTL analysis. We compared eQTL scans without confounder correction to eQTL scans when accounting for population structure using the kindship matrix (Section 13.1). Finally, we combined the hidden factors of expression heterogeneity (Section 13.2) and the kindship, jointly accounting for population structure and unknown expression confounders. Given the estimated expression heterogeneity confounders (K_{expr}, see Section 13.2) and the kindship matrix (K_{pop}, see Section 13.1), we fitted a random effect model, jointly to all gene expression levels, determining appropriate weighting parameters of the population structure (α) and expression heterogeneity (β): $K_{\mathrm{conf}} = \alpha K_{\mathrm{pop}} + \beta K_{\mathrm{expr}}$. Details of covariance parameter learning for joint correction can be found here[65]; Similar approaches, combining correction for population structure with expression confounders, have also been proposed by Listgarten et al[66]. We used the resulting confounding covariance structure for eQTL testing. We compared QTL testing without accounting for confounders, correction for population structure and joint correction for both types of confounding variation. Comparative association results are provided in Supplementary Fig. 38. All reported downstream analyses are based on the most stringent correction, including population structure and expression confounders.

References

1. Chen, Y. *et al.* Ensembl variation resources. *BMC Genomics* **11**, 293 (2010).
2. Weigel, D. & Mott, R. The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol* **10**, 107 (2009).
3. Kover, P. X. *et al.* A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet* **5**, e1000551 (2009).
4. McMullen, M. D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737-740 (2009).
5. Aylor, D. L. *et al.* Genetic analysis of complex traits in the emerging collaborative cross. *Genome Res*, doi: 10.1101/gr.111310.110 (2011).
6. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* **3**, e196 (2005).
7. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858 (2008).
8. Lunter, G. & Goodson, M. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*, doi: 10.1101/gr.111120.110 (2011).
9. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
10. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**, 265-272 (2010).
11. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117-1123 (2009).
12. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).
13. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
14. Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* **18**, 2024-2033 (2008).
15. Price, T. S. *et al.* SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res* **33**, 3455-3464 (2005).
16. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).
17. Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-317 (2010).
18. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* **317**, 338-342 (2007).
19. Zeller, G. *et al.* Detecting polymorphic regions in *Arabidopsis thaliana* with resequencing microarrays. *Genome Res* **18**, 918-929 (2008).
20. Santuari, L. *et al.* Substantial deletion overlap among divergent Arabidopsis genomes revealed by intersection of short reads and tiling arrays. *Genome Biol* **11**, R4 (2010).
21. Lai, A. G., Denton-Giles, M., Mueller-Roeber, B., Schippers, J. H. & Dijkwel, P. P. Positional information resolves structural variations and uncovers an evolutionarily divergent genetic locus in accessions of *Arabidopsis thaliana*. *Genome Biol Evol*, doi:10.1093/gbe/evr038 (2011).
22. Schneeberger, K. *et al.* Simultaneous alignment of short reads against multiple genomes. *Genome Biol* **10**, R98 (2009).

23. Thornton, K. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325-2327 (2003).
24. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
25. Eberle, M. A., Rieder, M. J., Kruglyak, L. & Nickerson, D. A. Allele frequency matching between SNPs reveals an excess of linkage disequilibrium in genic regions of the human genome. *PLoS Genet* **2**, e142 (2006).
26. Song, Y. S. & Hein, J. Constructing minimal ancestral recombination graphs. *J Comput Biol* **12**, 147-169 (2005).
27. Robinson, D. R. & Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131-147 (1981).
28. Lyngsø, R. B., Song, Y. S. & Hein, J. J. in *Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI)* Vol. 3692 eds R. Casadio & G. Myers) 239–250 (Springer, 2005).
29. Minichiello, M. J. & Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet* **79**, 910-922 (2006).
30. Michael, T. P. *et al.* Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genet* **4**, e14 (2008).
31. Covington, M. F., Maloof, J. N., Straume, M., Kay, S. A. & Harmer, S. L. Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol* **9**, R130 (2008).
32. Filichkin, S. A. *et al.* Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res* **20**, 45-58 (2010).
33. Cronn, R. *et al.* Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res* **36**, e122 (2008).
34. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
35. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).
36. Ihaka, R. & Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299-314 (1996).
37. Silverstein, K. A., Graham, M. A., Paape, T. D. & VandenBosch, K. A. Genome organization of more than 300 defensin-like genes in *Arabidopsis*. *Plant Physiol* **138**, 600-610 (2005).
38. Meyers, B. C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R. W. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* **15**, 809-834 (2003).
39. Shiu, S. H. & Bleecker, A. B. Plant receptor-like kinase gene family: diversity, function, and signaling. *Sci STKE* **2001**, re22 (2001).
40. Guo, A. Y. *et al.* PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res* **36**, D966-969 (2008).
41. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152-157 (2011).
42. Jean, G., Kahles, A., Sreedharan, V. T., De Bona, F. & Ratsch, G. RNA-Seq read alignments with PALMapper. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 16.1 (2010).
43. Sonnenburg, S., Schweikert, G., Philips, P., Behr, J. & Ratsch, G. Accurate splice site prediction using support vector machines. *BMC Bioinformatics* **8 Suppl 10**, S7 (2007).
44. Schweikert, G. *et al.* mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res* **19**, 2133-2143 (2009).

45. De Bona, F., Ossowski, S., Schneeberger, K. & Ratsch, G. Optimal spliced alignments of short sequence reads. *Bioinformatics* **24**, i174-180 (2008).
46. Behr, R. & Raetsch, G. De novo RNA-seq-based Genome Annotation. *In Preparation* (2011).
47. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).
48. Doring, A., Weese, D., Rausch, T. & Reinert, K. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics* **9**, 11 (2008).
49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
50. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
51. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).
52. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
53. Eichner, J., Zeller, G., Laubinger, S. & Ratsch, G. Support vector machines-based identification of alternative splicing in Arabidopsis thaliana from whole-genome tiling arrays. *BMC Bioinformatics* **12**, 55 (2011).
54. Heber, S., Alekseyev, M., Sze, S. H., Tang, H. & Pevzner, P. A. Splicing graphs and EST assembly problem. *Bioinformatics* **18 Suppl 1**, S181-188 (2002).
55. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
56. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445 (2003).
57. Keurentjes, J. J. *et al.* Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc Natl Acad Sci U S A* **104**, 1708-1713 (2007).
58. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
59. Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* **3**, e39 (2007).
60. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627-631 (2010).
61. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724-1735 (2007).
62. Kang, H. M., Ye, C. & Eskin, E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* **180**, 1909-1925 (2008).
63. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* **6**, e1000770 (2010).
64. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-354 (2010).
65. Fusi, N., Stegle, O. & Larwence, N. D. Accurate modeling of confounding variation in eQTL studies leads to a great increase in power to detect trans-regulatory effects. *Nature Precedings* **In press.** (2011).

66. Listgarten, J., Kadie, C., Schadt, E. E. & Heckerman, D. Correction for hidden confounders in the genetic analysis of gene expression. *Proc Natl Acad Sci U S A* **107**, 16465-16470 (2010).
67. Nicol, J. W., Helt, G. A., Blanchard, S. G., Jr., Raja, A. & Loraine, A. E. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* **25**, 2730-2731 (2009).

Supplementary Table 1. Sequenced Accession Statistics: Details of libraries sequenced for 18 genomes of *Arabidopsis thaliana*.

Accession	Origin	AIMS Stock #	Library	Read Length	Library insert size/bp[b]			Sequence Produced /Gb[c]			Coverage[d]			*de novo* contig size/bp[e]			# *de novo* contigs	total contig length /Mb
					25%	50%	75%	total	map	rm dup	initial/ library	initial/ total	final	longest	N50	N90		
Bur-0	Ireland	CS6643	download[a]	35				1.6	1.16	0.8	6.69	27.39	25	24654	1984	279	260365	111.29
			bur PhaseII	51	385	406	416	2.97	2.77	2.48	20.69							
Can-0	Canary Isles	CS6660	can Phase I	36	178	188	198	4.67	4.43	3.21	26.86	54.25	47	22495	1736	266	273359	111.31
			can PhaseII	51	462	483	499	4.76	4.37	3.28	27.39							
Ct-1	Italy	CS6674	ct Phase I	36	196	206	217	5.18	4.91	3.78	31.6	57.4	50	21849	1854	268	270434	111.41
			ct PhaseII	51	407	422	438	4.35	4.07	3.09	25.81							
Edi-0	Scotland	CS6688	edi Phase I	36	160	171	183	4.97	4.44	3.46	28.95	57.92	52	22926	1906	221	322787	117.04
			edi PhaseII	51	409	420	429	5.06	4.76	3.47	28.97							
Hi-0	Netherlands	CS6736	hi Amplified	36	358	375	391	1.02	0.94	0.86	7.23	41.29	33	19083	1441	236	306478	112.09
			hi Nonamp	36	380	393	406	1.78	1.65	1.48	12.38							
			hi Phase I	36	198	208	218	4.06	3.89	2.6	21.69							
Kn-0	Lithuania	CS6762	kn Phase I	36	209	224	239	4.83	4.05	2.59	21.63	41.35	28	18886	1524	186	385673	127.67
			kn PhaseII	51	354	370	385	2.94	2.57	2.36	19.72							
Ler-0	Germany (now Poland)	CS20	ler Phase I	36	213	230	249	5.14	4.31	2.7	22.6	38.85	27	17955	1372	228	315154	113.15
			ler PhaseII	51	403	410	417	2.5	2.3	1.95	16.26							
Mt-0	Libya	CS1380	mt Phase I	36	154	181	211	4.12	3.88	3.42	28.54	39.13	30	7478	646	182	404940	113.46
			mt PhaseII	51	1394	1504	1613	2.33	1.88	1.27	10.59							
No-0	Germany	CS6805	no Phase I	36	178	187	196	5.2	4.84	2.92	24.43	45.56	38	20011	1901	253	284666	112.8
			no PhaseII	51	304	315	327	3.17	2.96	2.53	21.13							
Oy-0	Norway	CS6824	oy Phase I	36	186	194	203	4.94	4.69	3.64	30.41	60.98	54	28033	1756	266	272178	111.35
			oy PhaseII	51	420	431	440	5.14	4.83	3.66	30.57							
Po-0	Germany	CS6839	po Phase I	36	164	170	176	3.64	3.06	2.31	19.29	48.38	41	19320	1258	190	414957	119.92
			po PhaseII	51	400	413	421	5.08	4.79	3.48	29.09							
Rsch-4	Russia	CS6850	rsch Phase I	36	187	205	225	5.04	4.42	2.62	21.91	50.41	38	22767	1780	256	281651	112.17
			rsch PhaseII	51	398	409	418	5.01	4.64	3.41	28.5							
Sf-2	Spain	CS6857	sf Phase I	36	174	179	183	2.95	2.8	1.94	16.25	45.39	40	22295	1886	266	272602	111.46
			sf PhaseII	51	375	383	390	5.11	4.78	3.49	29.14							
Tsu-0	Japan	CS6874	tsu Phase I	36	178	189	199	4.91	4.72	3.41	28.5	57.61	48	20007	1761	263	274548	112.15
			tsu PhaseII	51	390	416	426	4.97	4.63	3.48	29.11							

Wil-2	Russia	CS6889	wil2 Phase I	36	165	170	174	3.42	3.21	2.21	18.48	46.06	40	19784	1826	269	273385	111.33
			wil2 PhaseII	51	452	460	467	4.81	4.52	3.3	27.58							
Ws-0	Russia	CS6891	ws Phase I	36	381	406	429	4.66	3.49	1.72	14.39	44.57	33	28032	1732	231	329115	124.97
			ws_tsl	36	376	403	426	1.6	1.28	0.93	7.78							
			ws PhaseII	51	210	215	220	3.58	3.22	2.68	22.4							
Wu-0	Germany	CS6897	wu Phase I	36	191	204	216	3.14	2.99	2.07	17.27	35.04	26	23074	1955	273	264048	111.26
			wu PhaseII	51	354	393	410	2.69	2.5	2.13	17.77							
Zu-0	Germany	CS6902	zu Phase I	36	138.5	191	207	3.93	3.41	2.57	21.51	42.08	31	23671	1807	219	317803	116.42
			zu PhaseII	51	233.5	346	411	3.04	2.87	2.46	20.56							

[a] Publicly available data downloaded from the Weigel lab. All other data were generated specifically for this project
[b] Library insert size estimated from alignment of paired-end reads to TAIR10; 25%, 75% are quartiles.
[c] Total: amount in Gb of sequence generated after standard quality control filtering; mapped: amount mapped to TAIR10; rmdup: amount mapped after removing duplicate reads
[d] Initial coverage is the total length of rmdup sequence divided by 0.11966775 Gb (TAIR10 genome length). Final coverage is the median coverage after reads were re-mapped to the final assembly.
[e] *De novo* contigs were generated by SOAPdenovo. Lengths are for contigs, ignoring scaffolds. Only contigs at least 50 bp long included. N50, N90 are threshold lengths such that 50% and 90% of contigs are longer, respectively.

Supplementary Table 2. Counts of sequence differences to Bur-0 and Ler-0 test data as a function of assembly steps. Numbers in brackets are the error rates per 10kb (assuming that all the differences are due to errors in the assembly), obtained by dividing by the total length of the respective test datasets and multiplying by 10,000.

	Assembly step with differences to test data[a]								
test datasets &	Iteration								
polymorphism types	0	1	2	3	4	5	IMR	DENOM	Final
Bur-0 Survey[b]									
SNPs[c]	2756	549	357	292	265	251	252	156	79
(per 10kb)	(45.8)	(9.1)	(5.9)	(4.8)	(4.4)	(4.2)	(4.2)	(2.6)	(1.3)
Indels[d]	347	91	52	26	20	20	20	33	12
(per 10kb)	(5.8)	(1.5)	(0.9)	(0.4)	(0.3)	(0.3)	(0.3)	(0.5)	(0.2)
ISs	4	2	0	0	0	0	0	1	0
(per 10kb)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Bur-0 Divergent[e]									
SNPs	1539	849	540	436	394	387	387	15	4
(per 10kb)	(529)	(292)	(185)	(150)	(136)	(133)	(133)	(5.2)	(1.4)
Indels	320	131	57	31	20	14	14	3	1
(per 10kb)	(110)	(45.1)	(19.6)	(10.7)	(6.9)	(4.8)	(4.8)	(1.0)	(0.3)
ISs	10	8	6	4	4	4	4	0	0
(per 10kb)	(0.3)	(0.3)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)
Ler-0 chr3 339kb[f]									
(TE rich)									
SNPs[c]	3314	1165	834	726	668	653	655	675	448
(per 10kb)	(108)	(37.7)	(26.9)	(23.4)	(21.5)	(21.0)	(21.1)	(21.8)	(14.4)
Indels[d]	936	376	206	145	130	123	123	298	120
(per 10kb)	(30.5)	(12.2)	(6.7)	(4.7)	(4.2)	(4.0)	(4.0)	(9.6)	(3.9)
Ler-0 chr5 175kb[g]									
(TE poor)									
SNPs	603	139	85	74	67	66	66	36	21
(per 10kb)	(34.6)	(8.0)	(4.9)	(4.2)	(3.8)	(3.8)	(3.8)	(2.1)	(1.2)
Indels	244	58	40	34	34	37	36	50	32
(per 10kb)	(14.0)	(3.3)	(2.3)	(1.9)	(1.9)	(2.1)	(2.1)	(2.9)	(1.8)

[a] Differences between the Bur-0 test data and Bur-0 pseudochromosome sequences as a function of iteration were assessed as described in Supplementary Information section 4.1. The iteration 0 comparison reflects differences between the Bur-0 Sanger test data and TAIR10 prior to assembly.
[b] Sanger sequence data consisting of 1442 sequences total length 602207 bp[6,18].
[c] In addition to SNPs, 104 sequence differences corresponding to ambiguous positions (bases "K", "M", "R","S", "W", "Y") were observed in alignments of the Bur-0 Survey data to the final assembly.
[d] Refers to the number of insertion/deletion events, irrespective of length
[e] Sanger sequence data consisting of 188 sequences total length 29076 bp[14].
[f,g] Ler-0 manually assembled regions provided by Dr. Paul Dikjwel, Massey University, New Zealand. The 339kb region from Chr3 contains approx 48% TEs (transposable elements), the 175kb region on Chr5 contains 8% TEs.

Supplementary Table 3. Indel variation and polymorphic regions (PRs) by accession at all bases.

			Imbalanced substitutions				
Accession	Deletion no. (kb included)	Insertion no. (kb included)	No.	Deleted bases (kb)	Inserted bases (kb)	PRs (kb included)	Total bases (Mb)
Bur-0	78,216 (1,953)	91,402 (659)	11,595	801	194	40,314 (3,700)	7.307
Can-0	94,820 (2,387)	114,945 (801)	14,578	948	222	36,754 (3,196)	7.555
Ct-1	69,901 (1,864)	73,926 (334)	10,104	688	178	34,256 (2,692)	5.756
Edi-0	75,258 (1,893)	92,872 (685)	11,507	771	183	33,145 (3,053)	6.585
Hi-0	69,157 (1,309)	76,225 (388)	9,528	474	119	58,187 (2,359)	4.649
Kn-0	77,351 (1,845)	92,120 (653)	11,192	882	174	53,005 (3,522)	7.077
Ler-0	76,609 (1,950)	92,008 (650)	11,326	834	173	43,551 (3,355)	6.962
Mt-0	69,330 (2,028)	79,835 (482)	10,046	640	107	40,298 (3,110)	6.367
No-0	72,035 (2,035)	86,487 (617)	11,228	767	183	32,422 (2,515)	6.117
Oy-0	68,149 (1,707)	72,856 (332)	9,787	818	160	29,559 (2,671)	5.689
Po-0	71,720 (1,327)	79,515 (515)	8,194	455	110	36,715 (2,124)	4.530
Rsch-4	68,258 (1,825)	82,870 (613)	10,425	768	161	28,012 (2,618)	5.986
Sf-2	81,225 (2,030)	98562 (734)	12,560	831	182	38,277 (3,529)	7.306
Tsu-0	73,958 (1,972)	89,870 (652)	11,435	903	184	29,844 (2,776)	6.487
Wil-2	77,646 (2,009)	94,185 (681)	12,243	1,026	191	37,710 (2,948)	6.586
Ws-0	75,751 (1,930)	91,536 (665)	11,853	789	181	43,002 (3,495)	7.060
Wu-0	69,475 (1,949)	82,045 (561)	10,395	674	170	36,459 (3,262)	6.617
Zu-0	73,983 (1,907)	89,150 (639)	11,620	826	172	38,480 (3,344)	6.888
Non-redundant	492,896 (12,220)	707,725 (5,327)	104,090	9,602	1,291	13,727 [a]	28.83

[a] A value for non-redundant PRs is not given because allelic relationships cannot be assigned unambiguously (the actual sequence is unknown).

Supplementary Table 4. Polymorphic region (PR) counts, inclusive positions, and N50 contig sizes by accession for iteration 1 and final assemblies.

	Iteration 1		Final	
Accession	PRs[a] (Mb inclusive)	Contig N50[b] (kb)	PRs[a] (Mb inclusive)	Contig N50[b] (kb)
Bur-0	61,213 (5.51)	25.04	40,314 (3.89)	73.00
Can-0	60,188 (5.12)	24.97	36,754 (3.39)	75.46
Ct-1	50,245 (4.27)	30.90	34,256 (2.88)	78.57
Edi-0	50,815 (4.61)	29.30	33,145 (3.24)	69.42
Hi-0	74,074 (3.32)	34.80	58,187 (2.52)	76.95
Kn-0	74,186 (5.29)	15.81	53,005 (3.72)	26.69
Ler-0	64,270 (5.02)	24.14	43,551 (3.54)	53.36
Mt-0	60,495 (4.63)	23.63	40,298 (3.28)	43.37
No-0	51,694 (4.09)	31.31	32,422 (2.70)	84.37
Oy-0	45,236 (4.06)	39.09	29,559 (2.78)	107.53
Po-0	49,042 (3.04)	57.73	36,715 (2.29)	126.21
Rsch-4	45,047 (4.08)	37.82	28,012 (2.78)	113.38
Sf-2	58,758 (5.14)	31.94	38,277 (3.72)	98.22
Tsu-0	48,234 (4.49)	41.03	29,844 (2.96)	135.59
Wil-2	57,913 (4.63)	29.22	37,710 (3.14)	84.32
Ws-0	62,334 (5.19)	21.86	43,002 (3.69)	41.06
Wu-0	55,363 (4.86)	30.67	36,459 (3.43)	80.96
Zu-0	56,891 (5.04)	29.70	38,480 (3.53)	86.67

[a] PRs denote regions of no (or low) read coverage as defined in Supplementary Information section 3.6 (these are, effectively, breaks in the assemblies).
[b] Contigs are defined as regions of contiguous coverage between PRs.

Supplementary Table 5. Variable bases by accession and RNA-seq support for identified SNPs.

	Nucleotide differences		RNA-seq support (SNPs)		
Accession	SNPs	Ambiguous positions	Agree	Disagree	% concordance[a]
Bur-0	673,965	51,113	101,981	248	99.7574
Can-0	789,187	54,148	118,965	343	99.7125
Ct-1	650,332	44,342	100,175	271	99.73
Edi-0	630,728	35,210	93,151	356	99.62
Hi-0	497,688	157,526	78,702	164	99.79
Kn-0	637,034	43,211	98,204	265	99.73
Ler-0	647,094	42,652	98,421	256	99.74
Mt-0	588,481	41,237	84,438	266	99.69
No-0	611,346	45,886	87,517	233	99.73
Oy-0	639,949	42,316	83,055	287	99.65
Po-0[b]	446,422	267,439	57,604	688	98.82
Rsch-4	584,081	44,225	76,089	222	99.71
Sf-2	671,638	72,661	87,741	234	99.73
Tsu-0	615,062	43,030	85,738	243	99.72
Wil-2	661,673	48,184	89,111	316	99.65
Ws-0	652,654	47,181	84,478	207	99.76
Wu-0	592,611	47,585	82,864	227	99.73
Zu-0	631,624	44,391	92,757	219	99.76
Nonredundant	SNPs = 3,071,117		SNPs with RNA-seq support = 503,825		

[a] Excluding Po-0, for which our sample is extensively heterozygous (see Supplementary Fig. 6), the concordance as summed across all accessions is 99.72%.
[b] The reported ranges (by accession) for SNPs with RNA-seq read support, as reported in the main text, excludes the heterozygous Po-0 sample.

Supplementary Table 6. Concordance between base calls at major effect SNPs in assemblies and RNA-seq read data.

	Major effect SNP no. (% concordance) [a]				
Accession	PTC [b]	Met-Dis [c]	Ter-Dis [d]	5' SS [e]	3' SS [f]
Bur-0	181 (98.34)	31 (100.00)	39 (97.44)	24 (100.00)	39 (100.00)
Can-0	192 (98.96)	27 (100.00)	39 (100.00)	38 (100.00)	41 (100.00)
Ct-1	157 (100.00)	29 (100.00)	28 (100.00)	31 (100.00)	39 (100.00)
Edi-0	153 (99.35)	31 (100.00)	34 (97.06)	26 (100.00)	44 (97.73)
Hi-0	117 (100.00)	21 (100.00)	33 (100.00)	22 (100.00)	36 (97.22)
Kn-0	131 (99.24)	21 (100.00)	41 (100.00)	18 (94.44)	28 (100.00)
Ler-0	157 (98.73)	27 (96.30)	32 (96.88)	34 (97.06)	44 (100.00)
Mt-0	137 (97.81)	16 (100.00)	32 (100.00)	22 (100.00)	44 (97.73)
No-0	132 (100.00)	27 (100.00)	37 (97.30)	28 (100.00)	31 (100.00)
Oy-0	125 (98.40)	25 (96.00)	37 (100.00)	19 (100.00)	30 (100.00)
Po-0[b]	95 (96.84)	15 (100.00)	19 (94.74)	14 (100.00)	27 (100.00)
Rsch-4	120 (100.00)	23 (95.65)	36 (100.00)	28 (100.00)	36 (97.22)
Sf-2	128 (99.22)	32 (100.00)	32 (100.00)	36 (100.00)	37 (100.00)
Tsu-0	140 (98.57)	27 (96.30)	31 (100.00)	21 (100.00)	42 (100.00)
Wil-2	152 (96.05)	18 (94.44)	31 (100.00)	27 (100.00)	30 (100.00)
Ws-0	179 (98.88)	19 (100.00)	30 (100.00)	30 (100.00)	34 (100.00)
Wu-0	123 (100.00)	19 (100.00)	31 (100.00)	24 (100.00)	38 (100.00)
Zu-0	146 (98.63)	28 (100.00)	30 (100.00)	36 (100.00)	39 (100.00)

[a] Sample sizes are for the 20.2 to 24.7% of major effect SNPs (by accession) for which base calls could be assessed from aligned RNA-seq reads (i.e., they were in expressed regions). As an example concordance calculation, for a sample size of 153 with a concordance of 99.35%, RNA-seq read data supported 152 SNPs and disagreed with 1 SNP.

[b] PTC: premature termination codon

[c] Met-Dis: Initiation codon (ATG) in Col-0 codes for another amino acid.

[d] Ter-Dis: Termination codon in Col-0 codes for an amino acid.

[e] 5' SS: 5' splice site (changes between GT and GC 5' splice sites, which can maintain functionality, were excluded).

[f] 3' SS: 3' splice site

Supplementary Table 7. Length distribution of non-redundant deletions, insertions, and imbalanced substitutions in the 18 genomes.

Indel size	Deletions no. (% of total)	Insertion no. (% of total)	IS no. (% of total)[a]
1 bp	204,297 (41.45)	192,126 (27.15)	Not applicable
2-5 bp	136,838 (27.76)	179,258 (25.33)	37,708 (36.23)
6-20 bp	100,811 (20.45)	246,690 (34.86)	27,585 (26.50)
21-50 bp	30,772 (6.24)	83,879 (11.85)	20,047 (19.26)
51-100 bp	8,149 (1.65)	4,456 (0.63)	7,550 (7.25)
> 0.1 to 0.5 kb	6,571 (1.33)	1,070 (0.15)	6,796 (6.53)
> 0.5 to 1 kb	2,366 (0.48)	180 (0.03)	2,224 (2.14)
> 1 to 5 kb	2,405 (0.49)	64 (0.01)	1,754 (1.69)
> 5 to 10 kb	526 (0.11)	2 (0.00)	340 (0.33)
> 10 kb	161 (0.03)	0 (0.00)	86 (0.08)

[a] Lengths of imbalanced substitutions are the sum of deleted plus inserted bases.

Supplementary Table 8. Details of attempts to verify large indels and imbalanced substitutions predicted to segregate between Col-0 and Ler-0. The two insertions where the verification failed are in **bold**. [a] Type; D: Deletion (relative to TAIR10); I: Insertion; IS: Imbalanced Substitution. [b] Lengths are negative for deletions or for imbalanced substitutions when the Ler-0 sequence is shorter than the Col-0 sequence it replaces. [c] Length confirmed by PCR and breakpoint confirmed by Sanger sequencing: Y=Yes, N=No, R=repetitive, i.e. primers did not give single product in PCR of Col-0 or Ler-0; D=divergent, i.e., impossible to design primers that would work for both Col-0 and Ler-0.

chr	TAIR10 coord	type[a]	length / bp[b]	OK[c]
5	12949163	D	-22446	Y
1	16942828	D	-19369	Y
3	14766073	D	-17033	Y
2	566748	D	-15824	Y
4	4294663	D	-14522	Y
2	5565097	D	-13899	Y
2	4949288	D	-12418	Y
4	2151822	D	-12370	Y
2	12566146	D	-12277	Y
2	6620446	D	-11107	Y
1	9004833	D	-10880	Y
2	1366863	D	-10205	Y
1	13658291	D	-9437	Y
3	15002232	D	-9179	Y
3	7855162	D	-9095	Y
2	13959050	D	1538	Y
3	14250632	I	1001	Y
1	16509682	I	938	Y
2	17676960	I	869	Y
1	16925571	I	738	Y
5	12985545	I	695	Y
2	17203824	I	673	Y
1	26740765	I	661	Y
5	20365490	I	634	Y
2	18439568	I	608	Y
5	18220129	I	498	Y
4	13293999	I	487	Y
3	23071793	I	449	Y
3	10411642	I	440	Y
4	2336369	I	420	Y
4	2451670	I	414	Y
3	16085433	IS	-22121	Y
5	13776963	IS	-13743	Y
3	11935618	IS	-9963	Y
1	11502899	IS	-9757	Y
1	19336258	IS	-9370	Y
5	20393070	IS	-8339	Y
2	5040997	IS	-7430	Y
5	6072493	IS	-7428	Y
1	22950147	IS	-6985	Y
5	15334062	IS	-5977	Y
3	13573362	IS	-5946	Y
1	16311071	IS	-5833	Y
3	10339646	IS	-5810	Y
5	16259102	IS	-5648	Y
1	16551723	IS	-5352	Y
4	4705788	IS	-5352	Y
1	21700834	IS	-5285	Y
3	14956994	IS	-5279	Y
4	5751636	IS	3407	Y
5	17978762	IS	1238	Y
1	23058462	IS	1236	Y
4	8785919	IS	1164	Y
3	9001259	IS	1074	Y
5	17298747	IS	858	Y
4	15698653	IS	821	Y
2	4646275	IS	789	Y
2	10404106	IS	771	Y
5	6402431	IS	733	Y
1	5176622	IS	708	Y
5	17023406	IS	705	Y
3	11735446	IS	688	Y
2	2689814	IS	683	Y
1	4301458	IS	667	Y
1	13862467	IS	930	Y
3	15250473	D	-13325	R
5	12177079	D	-11665	R
5	11320511	D	-11351	R
5	11810259	D	-11333	R
2	5367810	D	-9728	R
1	16729004	D	1367	R
5	22579951	I	507	R
4	5683898	IS	-13956	R
5	12876869	IS	-9376	R
1	14824865	IS	-6568	R
3	16704459	IS	737	R
5	17402244	IS	687	R
2	12557617	IS	960	R
5	**15359063**	**I**	**527**	**N**
3	**10501233**	**I**	**445**	**N**
3	10970029	IS	-5992	D
1	16729139	IS	1104	D
1	16652062	IS	873	D
4	4235382	D	-16295	D
2	9293052	I	875	D
1	11090919	I	511	D
5	7819697	I	505	D
4	4957627	IS	-37730	D
2	6592458	IS	-15859	D
4	5912796	IS	-9329	D
3	14356532	IS	-5800	D
2	2610988	IS	-5358	D
5	10071025	IS	1190	D
4	16256272	IS	670	D

Supplementary Table 9. Correlations of nucleotide diversity with deletions, PRs in 50 kb windows.

	π (intergenic)		π (synonymous)		π (nonsynonymous)	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Repetitive bases	0.297	0.381	0.389	0.404	0.439	0.484
Unique intergenic bases	-0.119	-0.097	-0.094	-0.008	-0.146	-0.065
Deleted intergenic bases	0.580	0.687	0.546	0.649	0.494	0.652
PR intergenic bases	0.434	0.590	0.549	0.587	0.584	0.630
Unique coding bases	-0.310	-0.338	-0.363	-0.392	-0.360	-0.410
Deleted coding bases	0.267	0.426	0.399	0.468	0.409	0.575
PR coding bases	0.240	0.397	0.420	0.446	0.461	0.529
Bases of column type [a]	0.063	0.102	-0.397	-0.407	-0.393	-0.424
Distance to centromere	-0.484	-0.535	-0.394	-0.482	-0.336	-0.515
Protein-coding gene count	-0.231	-0.238	-0.275	-0.249	-0.278	-0.222
NBS-LRR gene count	0.215	0.192	0.301	0.212	0.331	0.233
GC content	-0.306	-0.333	-0.261	-0.371	-0.183	-0.358

[a] Number of intergenic bases in window used in computing π (intergenic), number of synonymous bases used in computing π (synonymous), and number of nonsynonymous bases used in computing π (nonsynonymous)

Supplementary Table 10. Correlations among nucleotide diversity measures in 50 kb windows.

	π (nonsynonymous)		π (intergenic)	
	Pearson	Spearman	Pearson	Spearman
π (synonymous)	0.821	0.865	0.674	0.797
π (nonsynonymous)			0.490	0.751

Supplementary Table 11. Correlations among deletions, PRs, and other genome features in 50 kb windows.

	Unique intergenic bases	Deleted intergenic bases	PR intergenic bases	Unique coding bases	Deleted coding bases	PR coding bases	Nonsynonymous bases	Synonymous bases	Intergenic bases	Distance to centromere	Protein-coding gene count	NBS-LRR gene count	GC content
Repetitive bases	-0.430	0.327	0.446	-0.736	0.165	0.222	-0.732	-0.734	-0.228	-0.535	-0.539	0.053	-0.284
Unique intergenic bases		0.102	-0.099	-0.104	-0.089	-0.113	-0.074	-0.072	0.904	0.233	0.056	-0.058	-0.428
Deleted intergenic bases			0.621	-0.427	0.408	0.305	-0.438	-0.445	0.260	-0.400	-0.269	0.154	-0.383
PR intergenic bases				-0.432	0.320	0.496	-0.458	-0.462	0.064	-0.372	-0.330	0.188	-0.260
Unique coding bases					-0.048	-0.096	0.969	0.968	-0.355	0.539	0.654	0.036	0.694
Deleted coding bases						0.684	-0.122	-0.130	-0.050	-0.136	-0.021	0.362	-0.004
PR coding bases							-0.176	-0.183	-0.071	-0.139	-0.088	0.404	-0.004
Nonsynonymous bases								0.998	-0.320	0.537	0.651	0.001	0.662
Synonymous bases									-0.320	0.541	0.648	-0.007	0.673
Intergenic bases										-0.042	-0.172	-0.054	-0.668
Distance to centromere											0.475	-0.039	0.387
Protein-coding gene count												-0.022	0.448
NBS-LRR gene count													0.013

Only Pearson's correlation coefficient displayed.

Supplementary Table 12. Multiple regression analysis results of the genomic features that best account for variability in nucleotide diversity (**in bold**) among 50 kb windows [a].

	Estimate	Std. Error	*t*-value	*p*-value
π (synonymous) [b]				
PR intergenic bases (mean)[c]	5.72×10^{-6}	4.93×10^{-7}	11.602	$< 2\times10^{-16}$
NBS-LRR gene count	3.60×10^{-3}	5.04×10^{-4}	7.146	1.23×10^{-12}
Distance to centromere	-3.03×10^{-10}	4.75×10^{-11}	-6.364	2.42×10^{-10}
Deleted intergenic bases (mean)	6.14×10^{-6}	1.15×10^{-6}	5.324	1.12×10^{-7}
Deleted coding bases	2.30×10^{-6}	4.48×10^{-7}	5.14	3.01×10^{-7}
Repetitive bases	1.30×10^{-7}	2.68×10^{-8}	4.868	1.21×10^{-6}
PR coding bases	1.44×10^{-6}	4.07×10^{-7}	3.54	4.08×10^{-4}
Deleted intergenic bases	8.93×10^{-7}	2.72×10^{-7}	3.279	1.06×10^{-3}
Deleted coding bases (mean)	-4.98×10^{-6}	1.52×10^{-6}	-3.269	1.10×10^{-3}
GC content	-2.67×10^{-2}	9.95×10^{-3}	-2.685	7.32×10^{-3}
π (intergenic) [d]				
Distance to centromere	-2.32×10^{-10}	1.94×10^{-11}	-12	$< 2\times10^{-16}$
Deleted intergenic bases (mean)	4.28×10^{-6}	4.37×10^{-7}	9.783	$< 2\times10^{-16}$
Unique intergenic bases	-1.60×10^{-7}	1.75×10^{-8}	-9.138	$< 2\times10^{-16}$
GC content	-3.84×10^{-2}	5.01×10^{-3}	-7.671	2.62×10^{-14}
NBS-LRR gene count	1.41×10^{-3}	1.92×10^{-4}	7.353	2.78×10^{-13}
Deleted intergenic bases	6.73×10^{-7}	9.82×10^{-8}	6.846	1.00×10^{-11}
Repetitive bases	-6.08×10^{-8}	1.27×10^{-8}	-4.792	1.77×10^{-6}
π (nonsynonymous) [e]				
π (synonymous)	2.36×10^{-1}	5.63×10^{-3}	41.881	$< 2\times10^{-16}$
PR intergenic bases (mean)	1.03×10^{-6}	1.91×10^{-7}	5.395	7.64×10^{-8}
NBS-LRR gene count	6.37×10^{-4}	1.32×10^{-4}	4.816	1.57×10^{-6}
GC content	1.61×10^{-2}	3.48×10^{-3}	4.621	4.05×10^{-6}
Repetitive bases	4.30×10^{-8}	9.43×10^{-9}	4.558	5.46×10^{-6}
Deleted intergenic bases (mean)	-1.07×10^{-6}	2.79×10^{-7}	-3.843	1.25×10^{-4}
Distance to centromere	3.82×10^{-11}	1.23×10^{-11}	3.117	1.85×10^{-3}
PR coding bases	3.22×10^{-7}	1.16×10^{-7}	2.767	5.70×10^{-3}
PR coding bases (mean)	-8.25×10^{-7}	3.08×10^{-7}	-2.678	7.47×10^{-3}
Deleted intergenic bases	1.81×10^{-7}	6.93×10^{-8}	2.609	9.14×10^{-3}

[a] Multiple regression analyses with stepwise model selection were performed with the statistical package R[36].
[b] Using the following predictor variables: repetitive bases, unique, deleted, and PR bases for intergenic and coding sites, mean across accessions of deleted bases and PR bases for intergenic and coding sites, number of synonymous sites, number of nonsynonymous sites, distance from the centromere, count of protein-coding genes, count of NBS-LRR genes, and GC content.
[c] All features are measured with respect to positions (regardless of accession, so that a deletion in a single accession leads to a position being classified as deleted) except those features where "mean" is indicated, where the feature corresponds to the mean number of bases of a certain type across all accessions.
[d] Using the same predictor variables as for π (synonymous).
[e] Using the same predictor variables as for π (synonymous) in addition to using π (synonymous) itself as a predictor of π (nonsynonymous).

Supplementary Table 13. Numbers of large-effect disruptions observed in annotated genes (TAIR10) on the accessions' genomes. We count the total number of disruptions of translation start and stop sites, introductions of premature stop codons, splice site disruptions (separately for UTR and CDS splice sites and for acceptor (acc) and donor (don) splice site). We also count how often more than one disruption type occurs (column "multiple disruptions") and additionally report the number of disruptions when no other type of disruption was determined. In the second last column we report the number of genes per accession with more than 50% in deletions or polymorphic regions. The last column shows the number of disrupted miRNA gene stem sequences.

Access-ion	translation start consensus disruption (single/mult)	translation stop consensus disruption (single/mult)	premature stop introduced (single/mult)	frame shift introduced (single/mult)	Splice site consensus disruption UTR [Acc (sngl/mlt) Don (sngl/mlt)]	Splice site consensus disruption CDS [Acc (sngl/mlt) Don (sngl/mlt)]	multi dis-rupt-ions	>50% CDS in deletion or PRs (union 1,675)	mi-RNA disr.
Bur-0	307/766	173/487	3552/4285	2182/2794	186/187 142/143	245/251 172/176	770	780	11
Can-0	323/896	157/515	5054/6161	2587/3308	214/215 142/142	284/286 197/198	907	726	18
Ct-1	280/697	169/428	3094/3805	2026/2559	173/175 103/105	220/226 142/143	666	768	13
Edi-0	263/675	146/426	3410/4292	2230/2765	175/175 126/126	236/237 145/146	676	768	12
Hi-0	276/632	137/361	2644/3372	2105/2567	161/166 115/115	218/220 156/158	568	832	14
Kn-0	254/688	161/467	3683/4304	2354/2962	174/174 116/117	250/253 161/162	724	766	15
Ler-0	264/716	155/467	4022/4858	2324/2907	204/204 138/139	274/276 157/160	744	774	10
Mt-0	256/614	131/402	2928/3685	2132/2646	171/171 115/116	223/226 140/143	624	765	15
No-0	296/741	150/436	3704/4441	2139/2704	198/200 123/124	230/234 158/158	716	778	11
Oy-0	263/685	162/435	2512/3102	2064/2607	157/159 109/109	214/216 132/135	679	759	10
Po-0	329/729	161/402	2621/3373	2217/2732	168/170 107/107	233/235 142/146	629	827	14
Rsch-4	257/694	139/413	3455/4341	2117/2668	192/192 125/125	222/224 145/146	688	794	9
Sf-2	305/759	172/494	3886/5063	2390/2984	219/219 150/150	246/249 151/154	766	765	14
Tsu-0	249/677	152/464	4135/5298	2124/2685	180/181 115/115	214/215 141/143	716	776	17
Wil-2	263/771	155/497	3804/4748	2222/2892	178/178 131/132	244/246 187/189	830	762	15
Ws-0	294/788	149/467	3715/4711	2316/2958	221/221 138/139	240/245 156/158	798	770	13
Wu-0	269/727	166/424	3109/3879	2127/2706	178/179 126/126	234/235 133/137	705	776	11
Zu-0	286/765	148/410	3352/4139	2217/2793	200/201 129/130	235/237 156/159	716	776	13

Supplementary Table 14. Statistics of the total number of reads and their alignments for each strain for the non-strand-specific RNA-seq experiments. For unaligned reads, no alignment with at most 3 mismatches and/or 1 indel was found. "Spliced aligned" is the number and percentage of aligned reads for which the best reported alignment is spliced. "Uniquely aligned" is the number and percentage of aligned reads where the best alignment is at least 2 mismatches/indels better than the second best alignment. R1, R2 are the respective biological replicates.

		Total reads	Number of Alignments	Unaligned reads (%)	Aligned reads (%)	Spliced aligned (%)	Uniquely aligned (%)
Bur-0	R1	5,153,116	6,097,055	339,405 (6.6)	4,813,711 (93.4)	925,190 (19.2)	3,963,122 (82.3)
	R2	5,170,977	6,257,330	286,383 (5.5)	4,884,594 (94.5)	926,228 (19.0)	3,960,515 (81.1)
Can-0	R1	5,218,603	6,181,791	350,073 (6.7)	4,868,530 (93.3)	938,500 (19.3)	4,003,726 (82.2)
	R2	4,362,707	5,361,524	197,832 (4.5)	4,164,875 (95.5)	789,361 (19.0)	3,362,441 (80.7)
Col-0	R1	6,541,414	7,888,342	366,724 (5.6)	6,174,690 (94.4)	1,198,754 (19.4)	5,043,583 (81.7)
	R2	5,967,344	7,576,524	253,944 (4.3)	5,713,400 (95.7)	1,104,667 (19.3)	4,463,852 (78.1)
Ct-1	R1	6,017,053	7,225,764	347,390 (5.8)	5,669,663 (94.2)	1,114,906 (19.7)	4,694,957 (82.8)
	R2	7,149,920	9,206,659	293,327 (4.1)	6,856,593 (95.9)	1,251,579 (18.3)	5,276,032 (76.9)
Edi-0	R1	4,310,843	5,176,241	191,432 (4.4)	4,119,411 (95.6)	814,702 (19.8)	3,440,143 (83.5)
	R2	5,885,013	7,206,750	255,871 (4.3)	5,629,142 (95.7)	1,074,982 (19.1)	4,575,156 (81.3)
Hi-0	R1	5,649,429	7,024,185	236,071 (4.2)	5,413,358 (95.8)	1,065,806 (19.7)	4,310,861 (79.6)
	R2	5,852,316	7,046,913	255,590 (4.4)	5,596,726 (95.6)	1,131,671 (20.2)	4,651,407 (83.1)
Kn-0	R1	5,630,359	7,039,811	230,623 (4.1)	5,399,736 (95.9)	1,033,974 (19.1)	4,291,442 (79.5)
	R2	6,374,743	7,753,201	285,979 (4.5)	6,088,764 (95.5)	1,209,261 (19.9)	5,013,952 (82.3)
Ler-0	R1	5,592,414	6,769,184	231,647 (4.1)	5,360,767 (95.9)	1,066,887 (19.9)	4,425,663 (82.6)
	R2	6,221,913	7,534,106	242,166 (3.9)	5,979,747 (96.1)	1,194,336 (20.0)	4,956,873 (82.9)
Mt-0	R1	4,516,070	5,462,924	271,323 (6.0)	4,244,747 (94.0)	829,325 (19.5)	3,450,777 (81.3)
	R2	5,129,286	6,314,371	276,296 (5.4)	4,852,990 (94.6)	916,175 (18.9)	3,853,041 (79.4)
No-0	R1	4,228,272	5,176,011	202,750 (4.8)	4,025,522 (95.2)	784,807 (19.5)	3,270,358 (81.2)
	R2	4,209,813	5,054,272	210,334 (5.0)	3,999,479 (95.0)	772,819 (19.3)	3,317,571 (83.0)
Oy-0	R1	3,709,991	4,465,143	195,761 (5.3)	3,514,231 (94.7)	713,047 (20.3)	2,927,967 (83.3)
	R2	4,849,414	6,090,750	261,023 (5.4)	4,588,391 (94.6)	850,087 (18.5)	3,580,228 (78.0)
Po-0	R1	3,964,327	4,859,267	189,304 (4.8)	3,775,023 (95.2)	740,769 (19.6)	3,068,128 (81.3)
	R2	5,218,228	6,298,353	252,229 (4.8)	4,965,999 (95.2)	952,550 (19.2)	4,092,647 (82.4)
Rsch-4	R1	3,164,983	3,773,710	155,070 (4.9)	3,009,913 (95.1)	578,985 (19.2)	2,505,114 (83.2)
	R2	4,511,326	5,353,195	272,648 (6.0)	4,238,678 (94.0)	817,046 (19.3)	3,501,832 (82.6)
Sf-2	R1	4,846,948	5,815,992	231,630 (4.8)	4,615,318 (95.2)	900,452 (19.5)	3,832,581 (83.0)
	R2	3,156,357	3,676,616	219,420 (7.0)	2,936,937 (93.0)	554,628 (18.9)	2,454,330 (83.6)
Tsu-0	R1	4,395,567	5,439,346	195,196 (4.4)	4,200,371 (95.6)	790,327 (18.8)	3,350,296 (79.8)
	R2	4,576,605	5,587,998	235,940 (5.2)	4,340,665 (94.8)	806,787 (18.6)	3,491,180 (80.4)
Wil-2	R1	5,587,286	6,839,714	237,593 (4.3)	5,349,693 (95.7)	971,695 (18.2)	4,316,038 (80.7)
	R2	3,851,926	4,617,481	198,189 (5.1)	3,653,737 (94.9)	697,615 (19.1)	3,020,722 (82.7)
Ws-0	R1	5,608,605	7,067,629	233,456 (4.2)	5,375,149 (95.8)	1,056,007 (19.6)	4,264,839 (79.3)
	R2	1,742,446	2,107,741	117,188 (6.7)	1,625,258 (93.3)	325,735 (20.0)	1,319,774 (81.2)
Wu-0	R1	4,565,786	5,983,649	201,828 (4.4)	4,363,958 (95.6)	781,294 (17.9)	3,214,379 (73.7)
	R2	5,514,096	6,675,738	270,920 (4.9)	5,243,176 (95.1)	1,043,699 (19.9)	4,337,521 (82.7)
Zu-0	R1	6,080,728	7,515,645	259,824 (4.3)	5,820,904 (95.7)	1,164,280 (20.0)	4,693,104 (80.6)
	R2	4,796,035	5,967,344	217,854 (4.5)	4,578,181 (95.5)	882,601 (19.3)	3,641,523 (79.5)
Average		**4,982,165**	**6,091,797**	**243,954 (5.0)**	**4,738,211** (95.0)	**915,040** (19.3)	**3,840,465** (81.1)
Sum		**189,322,259**	**231,488,269**	**9,270,233**	**180,052,027**	**34,771,534**	**145,937,675**

Supplementary Table 15. Statistics of number of strand-specific reads and their alignments for accessions Can-0 and Col-0 used for validation of gene annotations. See Supplementary Table 14 for details.

	Total reads	Number of Alignments	Unaligned reads (%)	Aligned reads (%)	Spliced aligned (%)	Uniquely aligned (%)
Can-0 (SS)	89,241,368	131,652,987	12,747,542 (9.7)	76,493,842 (90.3)	18,571,427 (24.3)	62,963,878 (82.3)
Col-0 (SS)	89,133,500	123,542,552	13,246,892 (14.9)	75,886,624 (85.1)	18,560,359 (24.4)	63,767,073 (84.0)
Mean	**89,187,434**	**127,597,769**	**12,997,217 (12.3)**	**76,190,233 (87.7)**	**18,565,893 (24.4)**	63,365,475 (83.2)
Sum	**178,374,868**	**255,195,539**	**25,994,434**	**152,380,466**	**37,131,786**	**126,730,951**

Supplementary Table 16. Agreement of *de novo* predictions of protein-coding genes based on RNA-seq-based read alignments on different strains with the TAIR10 genome annotation. Reported are the coding (CDS) exon and transcript level sensitivity (SN), specificity (SP) and F-score (F).

	Exon				Transcript			
Accession	# predicted	SN (%)	SP (%)	F (%)	# predicted	SN (%)	SP (%)	F (%)
Col-0	136,391	84.3	91.2	87.6	25,077	57.6	75.5	65.2
Bur-0	135,177	83.9	88.8	86.3	24,711	56.1	70.6	62.5
Can-0	134,871	83.6	88.8	86.1	24,550	55.6	70.4	62.2
Ct-1	135,422	84.3	89.1	86.6	24,774	56.9	71.4	63.3
Edi-0	135,351	84.2	89.0	86.5	24,747	56.8	71.3	63.2
Hi-0	135,244	83.9	88.8	86.2	24,684	56.1	70.6	62.5
Kn-0	135,598	84.2	88.8	86.4	24,757	56.5	71.0	62.9
Ler-0	135,286	84.1	89.0	86.5	24,731	56.8	71.4	63.3
Mt-0	135,184	84.0	88.9	86.4	24,692	56.6	71.2	63.1
No-0	135,183	83.9	88.8	86.3	24,696	56.2	70.7	62.6
Oy-0	135,175	84.0	88.9	86.4	24,664	56.4	71.0	62.9
Po-0	134,765	83.1	88.2	85.6	24,402	54.0	68.8	60.5
Rsch-4	134,967	83.8	88.9	86.3	24,633	56.3	71.0	62.8
Sf-2	134,745	83.6	88.8	86.1	24,523	55.5	70.4	62.1
Tsu-0	134,952	84.0	89.1	86.4	24,625	56.5	71.3	63.0
Wil-2	134,938	83.8	88.9	86.3	24,657	56.0	70.6	62.5
Ws-0	134,842	83.7	88.8	86.2	24,551	55.8	70.6	62.3
Wu-0	135,305	84.0	88.9	86.4	24,705	56.5	71.1	63.0
Zu-0	135,434	84.2	89.0	86.5	24,765	56.9	71.5	63.4
Average	**135,202**	**83.9**	**89.0**	**86.4**	**24,681**	**56.3**	**71.1**	**62.8**

Supplementary Table 17. Comparison of three annotation strategies on the reference accession Col-0: a) mGene, which predicts protein-coding genes *ab initio* and uses the genome sequence only, b) mGene.ngs, which uses the genome sequence as well as RNA-seq read alignments to predict protein-coding genes *de novo*, and c) cufflinks, which only uses the RNA-seq read alignments to predict transcripts to which we assigned open reading frames (ORF) of length at least 100nt and 300nt (the 590 and 2,884 transcripts, respectively, for which we could not identify an open reading frame were omitted from this analysis). Reported are the coding (CDS) exon and transcript level sensitivity (SN), specificity (SP) and F-score (F).

	CDS Exon				CDS Transcript			
	# predicted	**SN (%)**	**SP (%)**	**F (%)**	**# predicted**	**SN (%)**	**SP (%)**	**F (%)**
mGene	146,241	84.8	85.5	85.2	26,649	56.5	63.0	59.6
mGene.ngs	136,391	84.3	91.2	87.6	25,077	57.6	75.5	65.2
Cufflinks ORF≥100	94,921	53.6	88.1	66.7	16,811	28.6	52.0	36.9
Cufflinks ORF≥300	90,176	52.5	90.7	66.5	14,517	27.5	58.8	37.5

Supplementary Table 18. Features of the consolidated annotation of the 18 genomes based on RNA-seq-based gene predictions and the TAIR 10 reference annotation (excluding novel genes). The upper part describes the consolidated annotations used for most analyses (NSS). The lower part is based on additional, independent, strand-specific (SS) validation RNA-seq data only available for Col-0 and Can-0, leading to a larger number of predicted novel transcripts and introns.

Acces-sion	Genes	Protein-coding transcript	Non-coding transcript	Novel transcript	Introns	Novel introns	TIS sites	Novel TIS sites	Stop codon sites	Novel stop codon sites	genes with modifications (union=8,757)
Col-0	33,295	41,303	1,395	1,687	129,368	1,143	33,710	323	34,204	351	1,604
Bur-0	32,842	40,526	1,368	2,152	127,344	1,262	33,201	406	33,635	405	1,977
Can-0	32,741	40,332	1,362	2,149	127,038	1,289	33,090	423	33,555	467	2,100
Ct-1	32,858	40,765	1,368	2,384	127,813	1,402	33,271	425	33,776	501	2,232
Edi-0	32,847	40,623	1,366	2,209	127,570	1,317	33,252	443	33,705	455	2,105
Hi-0	32,968	40,857	1,381	2,325	127,934	1,421	33,376	450	33,840	474	2,199
Kn-0	32,833	40,696	1,372	2,307	127,601	1,441	33,235	441	33,723	477	2,191
Ler-0	32,852	40,839	1,376	2,559	127,585	1,592	33,289	518	33,756	542	2,406
Mt-0	32,849	40,513	1,372	2,080	127,469	1,205	33,193	370	33,661	399	1,911
No-0	32,817	40,415	1,366	1,975	127,475	1,209	33,162	389	33,630	422	1,920
Oy-0	32,866	40,428	1,367	1,841	127,383	1,112	33,217	386	33,677	399	1,835
Po-0	32,987	40,648	1,377	2,021	127,605	1,251	33,326	376	33,817	413	1,945
Rsch-4	32,867	40,346	1,364	1,802	127,237	1,050	33,194	341	33,620	358	1,742
Sf-2	32,806	40,319	1,383	1,980	127,109	1,115	33,138	377	33,569	399	1,875
Tsu-0	32,832	40,506	1,372	2,030	127,443	1,176	33,211	410	33,658	407	1,941
Wil-2	32,776	40,363	1,369	1,987	127,197	1,208	33,109	383	33,584	439	1,948
Ws-0	32,844	40,217	1,365	1,727	127,097	1,030	33,160	355	33,586	357	1,725
Wu-0	32,879	40,545	1,377	2,041	127,605	1,249	33,241	410	33,670	392	1,984
Zu-0	32,883	40,745	1,383	2,300	127,699	1,456	33,303	482	33,770	489	2,217
Col-0 SS	33,295	43,546	1,402	4,420	131,341	3,101	33,656	269	34,066	213	2,959
Can-0 SS	32,741	42,617	1,369	4,937	129,045	3,276	33,107	437	33,477	386	3,566

Supplementary Table 19. Independent RNA-seq data was used to validate features of different annotations. See Supplementary Information section 10.5 for additional details.

Strain/Annotation	Annotated introns			Novel introns			Missed TAIR 10 introns			Missed RNA-seq introns	
	total	conf.	%	total	conf.	%	total	conf.	%	NSS	SS
Col-0											
Ab initio	118,964	98,467	82.8	11,789	1,574	13.4	20,679	9,656	46.7	1,204	8,883
De novo	113,691	99,861	87.8	6,008	1,216	20.2	20,171	7,904	39.2	685	7,133
TAIR10	127,854	106,549	83.3	0	0		0	0		256	4,094
Consolidated NSS	128,995	107,682	83.5	1,143	1,134	99.2	0	0		177	3,638
Consolidated SS	130,953	109,596	83.7	3,101	3,048	98.3	0	0		198	1,631
Can-0											
Ab initio	118,031	98,114	83.1	11,883	1,831	15.4	21,103	9,289	44.0	869	8,723
De novo	112,407	99,003	88.1	6,281	1,405	22.4	21,125	7,974	37.7	499	7,170
TAIR10	127,251	105,572	83.0	0	0		0	0		203	4,137
Consolidated NSS	126,689	106,453	84.0	1,289	1,055	81.8	1,851	174	9.4	156	3,718
Consolidated SS	128,690	108,424	84.3	3,276	3,016	92.1	1,837	164	8.9	164	1,797

Supplementary Table 20. Numbers of detected and annotated intron retention events. R1/R2 denote the RNA-seq reads obtained from biological replicates 1 and 2. We report the number of detected events for the union of the reads and for each replicate separately. We also compute the overlap (agreement/size of smaller set) between the detected and annotated events as well as between the detected events when using the replicates separately.

Confidence	Detected R1+R2	TAIR 10 Annotation	Overlap Detected & Annotation	Detected R1	Detected R2	Overlap R1 & R2
0	4,962	2,791	515 (18%)	2,834 (57%)	2,909 (59%)	2,052 (72%)
1	3,425	1,111	459 (41%)	1,622 (47%)	1,634 (48%)	1,010 (62%)
2	2,906	848	443 (52%)	1,286 (44%)	1,325 (46%)	846 (66%)
3	2,545	688	414 (60%)	1,081 (42%)	1,097 (43%)	744 (69%)
4	2,233	576	381 (66%)	918 (41%)	909 (41%)	646 (71%)
5	1,946	499	355 (71%)	784 (40%)	781 (40%)	582 (75%)
6	1,703	440	335 (76%)	666 (39%)	656 (39%)	487 (74%)
7	1,459	379	298 (79%)	564 (39%)	558 (38%)	421 (75%)
8	1,254	329	270 (82%)	476 (38%)	461 (37%)	351 (76%)
9	1,093	281	241 (86%)	381 (35%)	375 (34%)	284 (76%)
10	923	241	214 (89%)	317 (34%)	297 (32%)	228 (77%)
11	772	210	191 (91%)	256 (33%)	242 (31%)	181 (75%)
12	654	178	166 (93%)	208 (32%)	193 (30%)	145 (75%)
13	517	152	143 (94%)	168 (32%)	154 (30%)	119 (77%)
14	413	124	118 (95%)	131 (32%)	113 (27%)	86 (76%)
15	331	106	103 (97%)	98 (30%)	81 (24%)	68 (84%)
16	237	78	75 (96%)	60 (25%)	61 (26%)	43 (72%)
17	160	53	53 (100%)	38 (24%)	32 (20%)	19 (59%)
18	89	26	26 (100%)	17 (19%)	8 (9%)	5 (63%)
19	0	0	0 (n/a)	0 (n/a)	0 (n/a)	0 (n/a)

Supplementary Table 21. Numbers of detected and annotated exon skip events. R1/R2 denote the RNA-seq reads obtained from biological replicates 1 and 2. We report the number of detected events for the union of the reads and for each replicated separately. We also compute the overlap (agreement/size of smaller set) between the detected and annotated events as well as between the detected events when using the replicates separately.

Confidence	Detected R1+R2	TAIR 10 Annotation	Overlap Detected & Annotation	Detected R1	Detected R2	Overlap R1 & R2
0	336	621	53 (16%)	200 (60%)	199 (59%)	129 (65%)
1	205	197	50 (25%)	126 (61%)	131 (64%)	83 (66%)
2	179	151	48 (32%)	104 (58%)	105 (59%)	70 (67%)
3	155	120	46 (38%)	87 (56%)	87 (56%)	60 (69%)
4	131	103	45 (44%)	77 (59%)	75 (57%)	54 (72%)
5	116	90	41 (46%)	63 (54%)	64 (55%)	46 (73%)
6	101	77	38 (49%)	52 (51%)	52 (51%)	37 (71%)
7	93	65	37 (57%)	45 (48%)	45 (48%)	32 (71%)
8	83	57	36 (63%)	41 (49%)	39 (47%)	28 (72%)
9	72	50	33 (66%)	33 (46%)	33 (46%)	23 (70%)
10	62	45	30 (67%)	26 (42%)	23 (37%)	15 (65%)
11	57	41	27 (66%)	18 (32%)	15 (26%)	7 (47%)
12	43	28	20 (71%)	9 (21%)	10 (23%)	5 (56%)
13	33	23	15 (65%)	6 (18%)	9 (27%)	5 (83%)
14	27	16	13 (81%)	5 (19%)	7 (26%)	4 (80%)
15	21	12	9 (75%)	5 (24%)	4 (19%)	3 (75%)
16	13	10	7 (70%)	3 (23%)	3 (23%)	2 (67%)
17	5	5	4 (80%)	3 (60%)	1 (20%)	1 (100%)
18	4	3	3 (100%)	2 (50%)	0 (0%)	0 (n/a)
19	0	0	0 (n/a)	0 (n/a)	0 (n/a)	0 (n/a)

Supplementary Table 22. Number of expressed protein-coding genes as determined by comparison with a backround model of leeky transcription.

Accession	Number of expressed protein-coding genes at q-value ≤ 0.05
Bur-0	19,238
Can-0	19,119
Col-0	19,571
Ct-1	19,383
Edi-0	19,336
Hi-0	19,370
Kn-0	19,593
Ler-0	19,536
Mt-0	19,198
No-0	19,035
Oy-0	18,996
Po-0	19,314
Rsch-4	18,877
Sf-2	18,827
Tsu-0	18,999
Wil-2	19,013
Ws-0	18,598
Wu-0	19,085
Zu-0	19,333

Supplementary Table 23. Gene expression summarized by gene types and certain gene properties.

	Protein-coding genes	ncRNA genes	Novel genes	Trans-posable elements	Trans-posable element genes	Pseudo-genes
Total number	27,416	1167	447	31,189	3,903	924
Expressed (q-value≤0.05)	20,550	253	314	408	88	196
Heritability (>30 %)	14,378	120	176	225	39	104
Differentially expressed (q-value≤0.05)	9,360	84	121	85	47	81
cis eQTLS (q-value≤0.05)	1016	37	52	49	24	29
Structural variation (q-value≤0.05)	236	28	21	59	23	13
Large effect eQTLs (q-value≤0.05)	204	0	0	6	0	10
Copy number variation eQTLs (q-value≤0.05)	49	5	5	3	1	4
cis nucleotide variant eQTLS (q-value≤0.05)	818	14	31	8	3	15

Supplementary Table 24. Enriched gene ontology (GO) process terms for differentially expressed genes.

GO Term	P-value	Gene no.[a]
Defense response [b]	8.05E-15	68
Response to stress [b]	5.50E-13	87
Response to stimulus [b]	8.69E-11	168
Apoptosis [b]	9.53E-09	34
Programmed cell death [b]	1.54E-08	35
Cell death [b]	2.16E-08	35
Death [b]	2.16E-08	35
Immune system process [b]	1.04E-06	26
Immune response [b]	3.00E-06	25
Innate immune response [b]	3.00E-06	25
Response to biotic stimulus [b]	5.68E-06	29
Multi-organism process	6.30E-06	27
Response to other organism [b]	1.14E-05	26
S-glycoside metabolic process [b]	8.58E-05	5
Glucosinolate metabolic process [b]	8.58E-05	5
Glycosinolate metabolic process [b]	8.58E-05	5
Defense response to fungus [b]	5.74E-04	7
Response to fungus [b]	5.74E-04	7

[a]: Number of Genes Associate With GO Term
[b]: GO process including genes associated with response to biotic factors such as disease resistance (R) proteins, avirulence-responsive protein, pathogenesis-related thaumatin family proteins, or proteins associated with glucosinolate biosynthesis (GO terms with a single gene excluded).

Supplementary Table 25. Nucleotide diversity for nuclear genes as a function of expression and differential expression status.

	Gene classification							
	All genes		Not expressed		Expressed, not differentially		Differentially expressed	
	Unfiltered	Filtered[a]	Unfiltered	Filtered	Unfiltered	Filtered	Unfiltered	Filtered
Mean π_N weighted per gene[b] ($\times 10^{-3}$)	2.52	1.48	4.60	2.24	1.77	1.28	2.39	1.52
Median π_N per gene ($\times 10^{-3}$)	1.15	0.87	2.62	1.45	0.89	0.76	1.11	0.88
Mean π_S weighted per gene[c] ($\times 10^{-3}$)	9.94	7.97	11.70	8.17	7.90	6.80	11.54	9.46
Median π_S per gene ($\times 10^{-3}$)	5.08	4.11	6.95	4.94	3.82	3.40	6.25	5.14
n[d]	23369	16461	4307	2038	10482	8251	8580	6172
Median π_N/π_S per gene	0.22	0.18	0.36	0.26	0.21	0.19	0.18	0.16
n[e]	21520	14987	3907	1783	9580	7483	8033	5721
Mean π_N weighted per base[f] ($\times 10^{-3}$)	2.22	1.38	4.29	2.08	1.59	1.21	2.30	1.44
Mean π_S weighted per base[g] ($\times 10^{-3}$)	9.27	7.59	11.60	8.45	7.23	6.37	11.08	9.09
Ratio of above per base weighted means	0.24	0.18	0.37	0.25	0.22	0.19	0.21	0.16
n[h]	26424	17725	6095	2631	11100	8576	9229	6518

[a] genes with evidence of gene model disruptions or where over 10% of amino acids are predicted to differ between any pair of accessions are dropped

[b] nonsynonymous nucleotide diversity (mean of per base value across genes, i.e., genes weighted equally)

[c] synonymous nucleotide diversity (mean of per base value across genes, i.e., genes weighted equally)

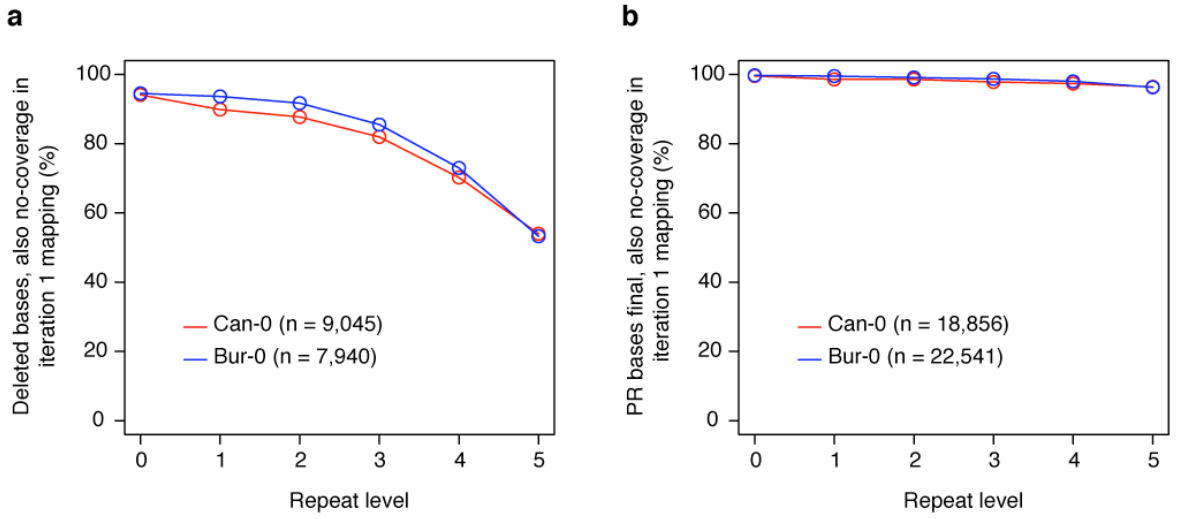
[d] number of genes used in each calculation from section above (genes where nucleotide diversity was measured at fewer than 50 synonymous sites were dropped)

[e] number of genes used in each calculation from section above (genes where nucleotide diversity was measured at fewer than 50 synonymous sites were dropped and genes with no silent polymorphism were dropped)

[f] nonsynonymous nucleotide diversity (weighted per base rather than per gene)

[g] synonymous nucleotide diversity (weighted per base rather than per gene)

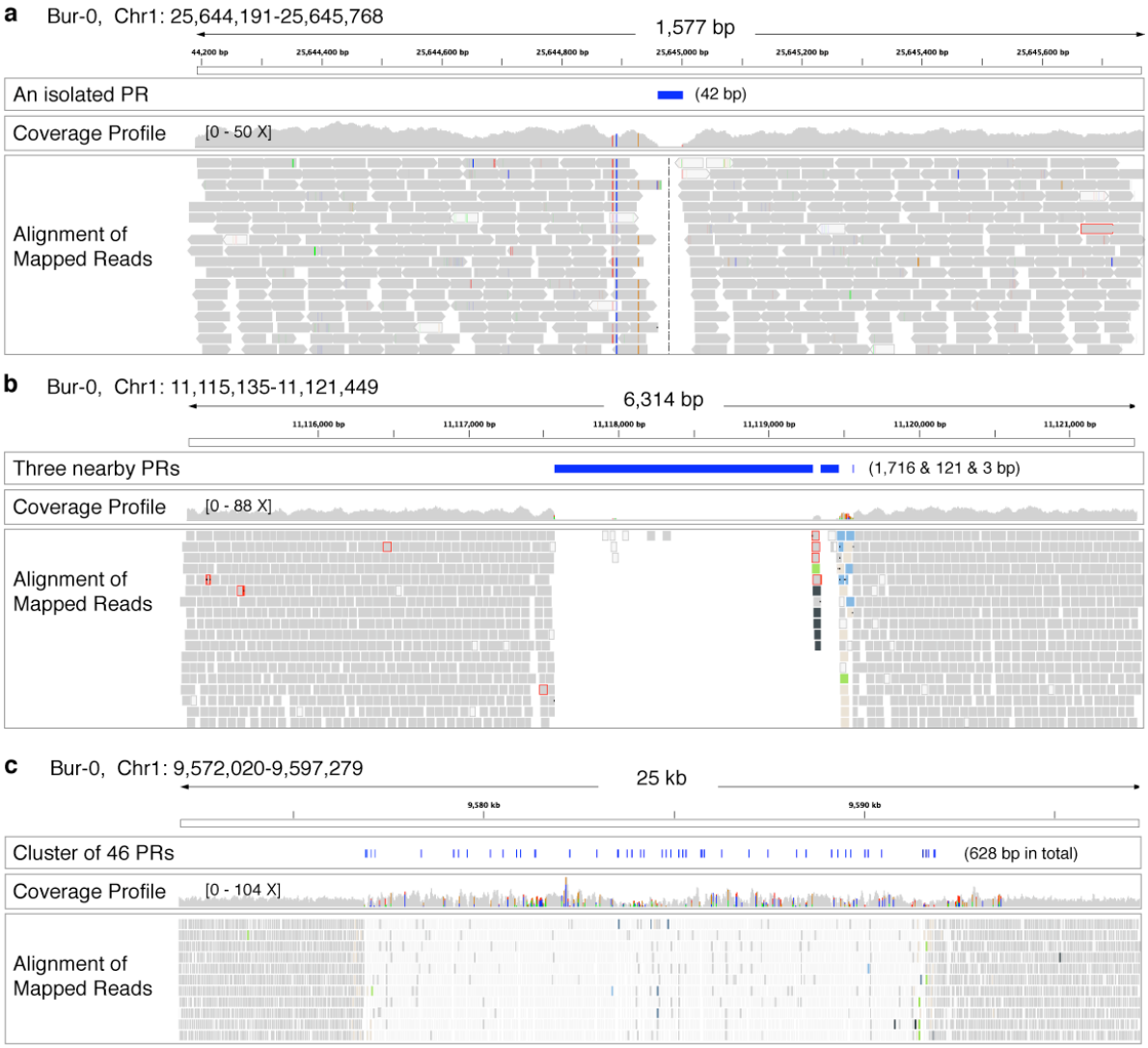
[h] number of genes used in each calculation from section above (genes where nucleotide diversity was measured at fewer than 0.5 synonymous sites and fewer than 0.5 nonsynonymous sites were dropped)



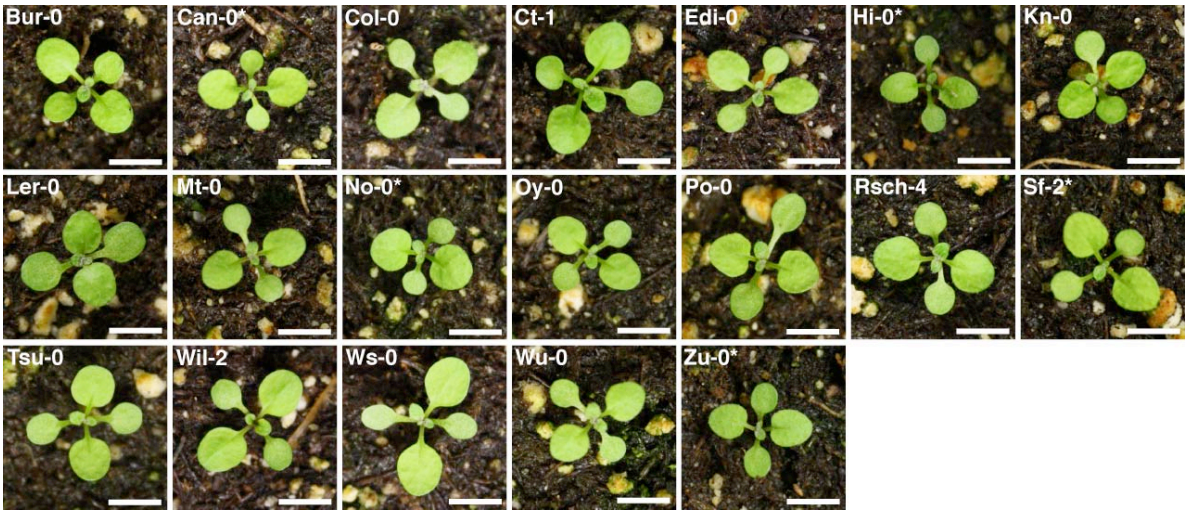
Repeat level	0	1	2	3	4	5
Can-0, kb	1259	61	101	184	330	901
Bur-0, kb	939	46	78	150	297	830

Repeat level	0	1	2	3	4	5
Can-0, kb	1533	95	157	260	308	541
Bur-0, kb	1428	102	179	331	471	822

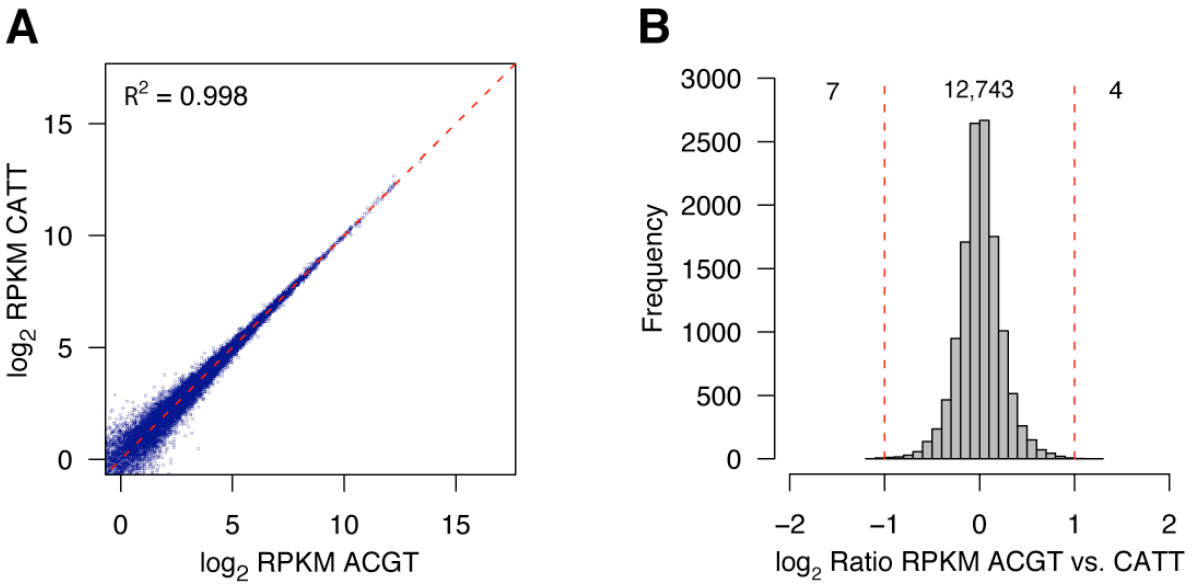
Supplementary Figure 1. Deleted and polymorphic region (PR) bases in the final Bur-0 and Can-0 assemblies that were not covered by reads in iteration 1. We examined the fraction of deleted (a) or PR (no-coverage) bases (b) in the final assemblies for Bur-0 and Can-0 – as a function of repeat level [0 (unique) to 5 (highly repetitive)] – that were no-coverage in the iteration 1 read alignments (the coverage criteria for assignment as no-coverage in iteration 1 was the same as for assigning PR regions in the final assemblies). The analysis was based on deletions and PRs of >20 bp, and the 10 bp at the end of deletion or PR variants was excluded from analysis as junctions often have overlapping (misaligned) read bases. The bases included in predictions as a function of repeat level are as indicated (bottom). At unique bases for both deleted and PR regions, if bases were not covered in the final assembly, they also tended to not be covered in the iteration 1 mapping. This pattern was observed for all repeat levels for PRs. This suggests that the iterative mapping approach, in which changes to the mapping target (genome) were introduced at each step, was not leading to false no-coverage regions at later iteration steps (i.e., reads were not being pulled from one region of the genome to another by the read mapper in an attempt to keep read pairs together, which could lead to false prediction of PRs which are assessed from coverage criteria). Moreover, while deletion predictions used information from discordantly mapped paired end reads ("stretched" pairs), deleted sequences in unique regions in the final assemblies would be expected to be no-coverage in the first iteration (in a deletion of unique bases, if predicted correctly, no reads could possibly align). For unique sequences, this expectation is observed (e.g., see panel a repeat level 0). Bur-0 and Can-0 were chosen for analysis because they reflect the range of read coverage among the sequenced accessions (Supplementary Table 1), and because they have no extended tracks of residual heterozygosity (Supplementary Fig. 5).



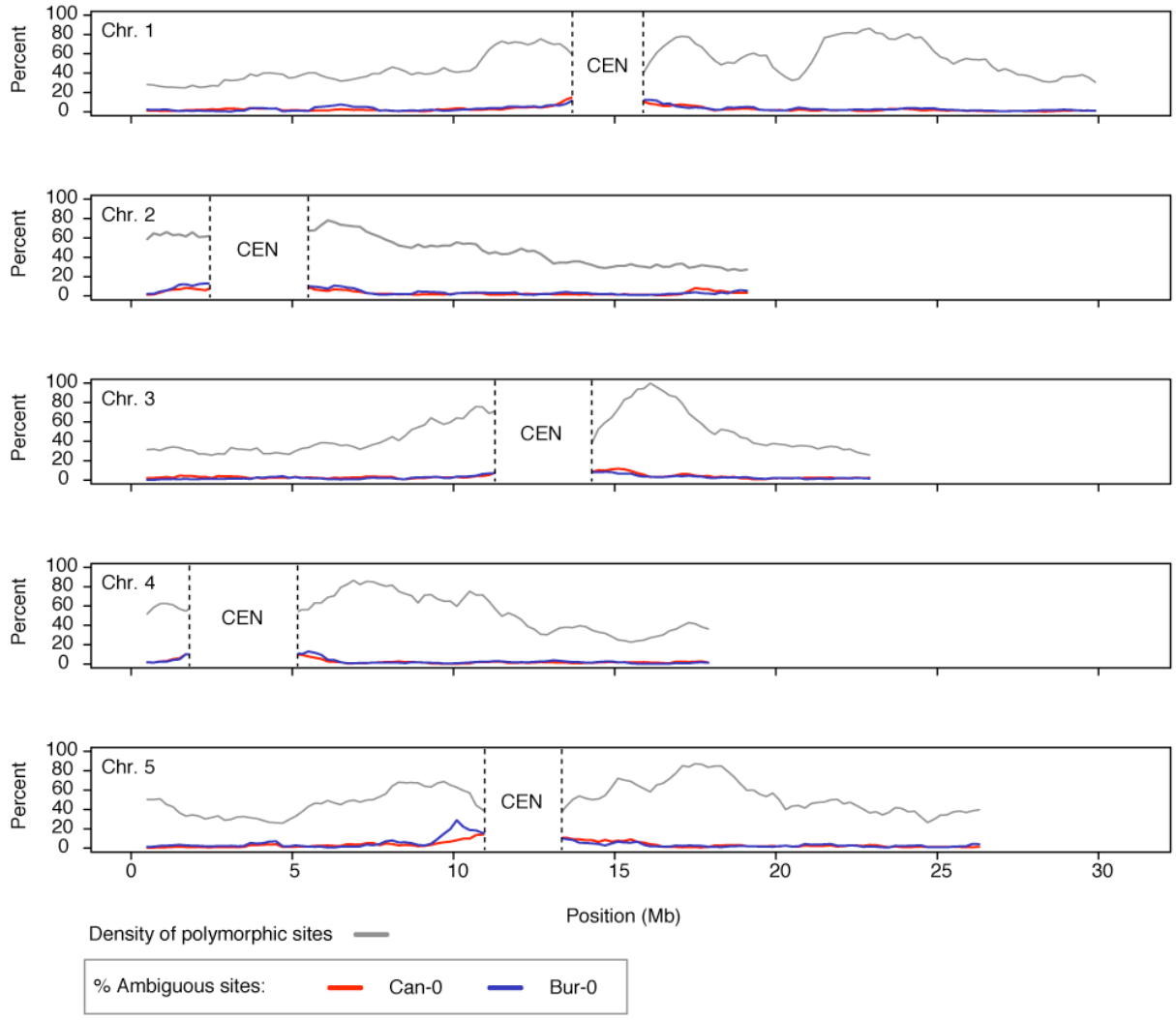
Supplementary Figure 2. Examples of polymorphic regions in accession Bur-0. Read alignments (shades of grey) and PRs (blue bars, with length or summed lengths shown in parentheses) for three regions in Bur-0 (shown are Illumina genomic DNA sequencing reads aligned to the final Bur-0 assembly). Small to large PRs (a and b, respectively) in regions of well supported read mappings (dark grey). In regions of repeats (c; reads with poor mapping qualities reflecting repetitive mappings are shaded light grey), clusters of PRs were frequently observed in the assemblies, many of which are likely to be spurious. Plots are modified from Integrated Genome Browser outputs (IGV)[67].



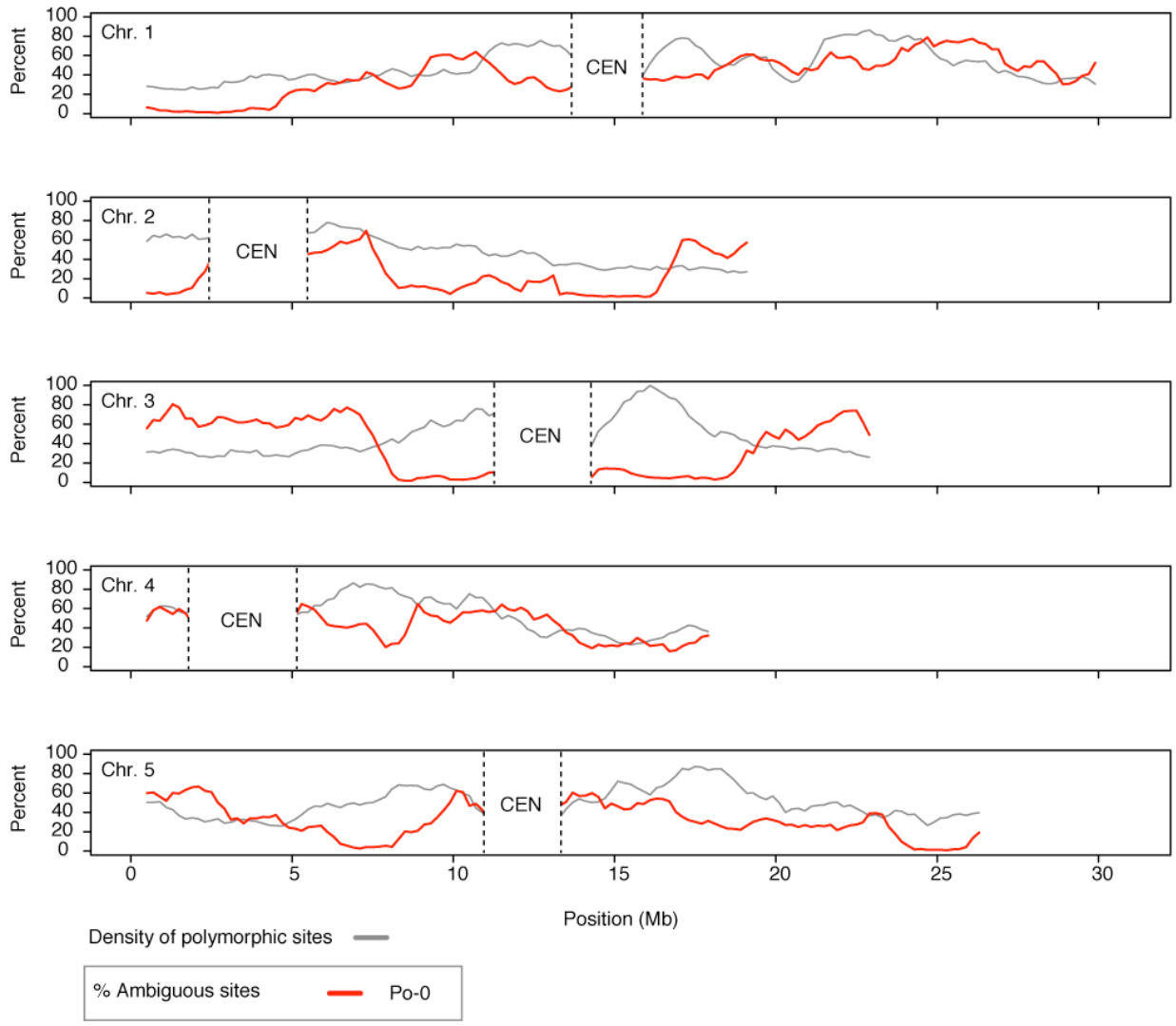
Supplementary Figure 3. Representative images of seedling developmental stages collected for RNA isolation. Accession images are representatives of the developmental stage 11 days after placing stratified seeds on soil. Accessions with an "*" were slower growing/germinating and therefore photos were taken 12 days after placing stratified seeds on soil. Scale bars 5 mm.



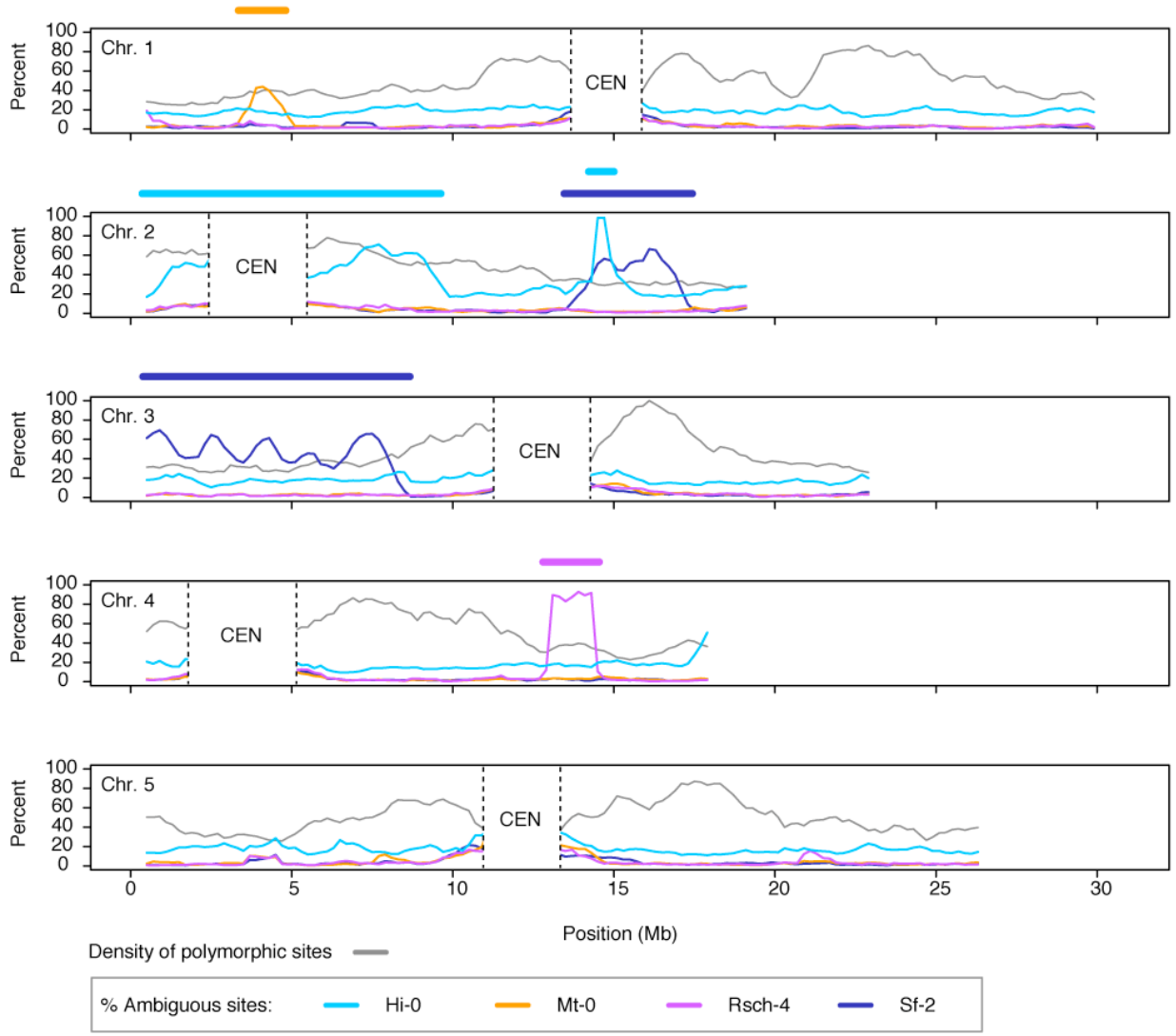
Supplementary Figure 4. Example comparison of expression values and the distribution of expression variation between barcoded RNA-seq samples. **A** Log_2 transformed normalized expression produced in CLC Genomics Workbench for identical Col-0 RNA sequencing libraries differing only in the added barcode. R^2 values were calculated for each of the possible barcoded-library comparisons and in all cases exceeded 0.993. **B** Example histogram showing the number of genes with greater that two-fold increased (4 genes) or decreased (7 genes) relative to the number with less than two-fold increased or decreased expression (12,743 genes) between replicated RNA sequencing experiments differing only in barcode sequence. Red dashed lines indicate two-fold differential expression thresholds. All gene expression levels compared in the histogram had on average greater than 20 reads mapped. Differential expression was analysed for each of the possible barcoded-library comparisons and the maximum number of genes with >2-fold difference in expression was 20.



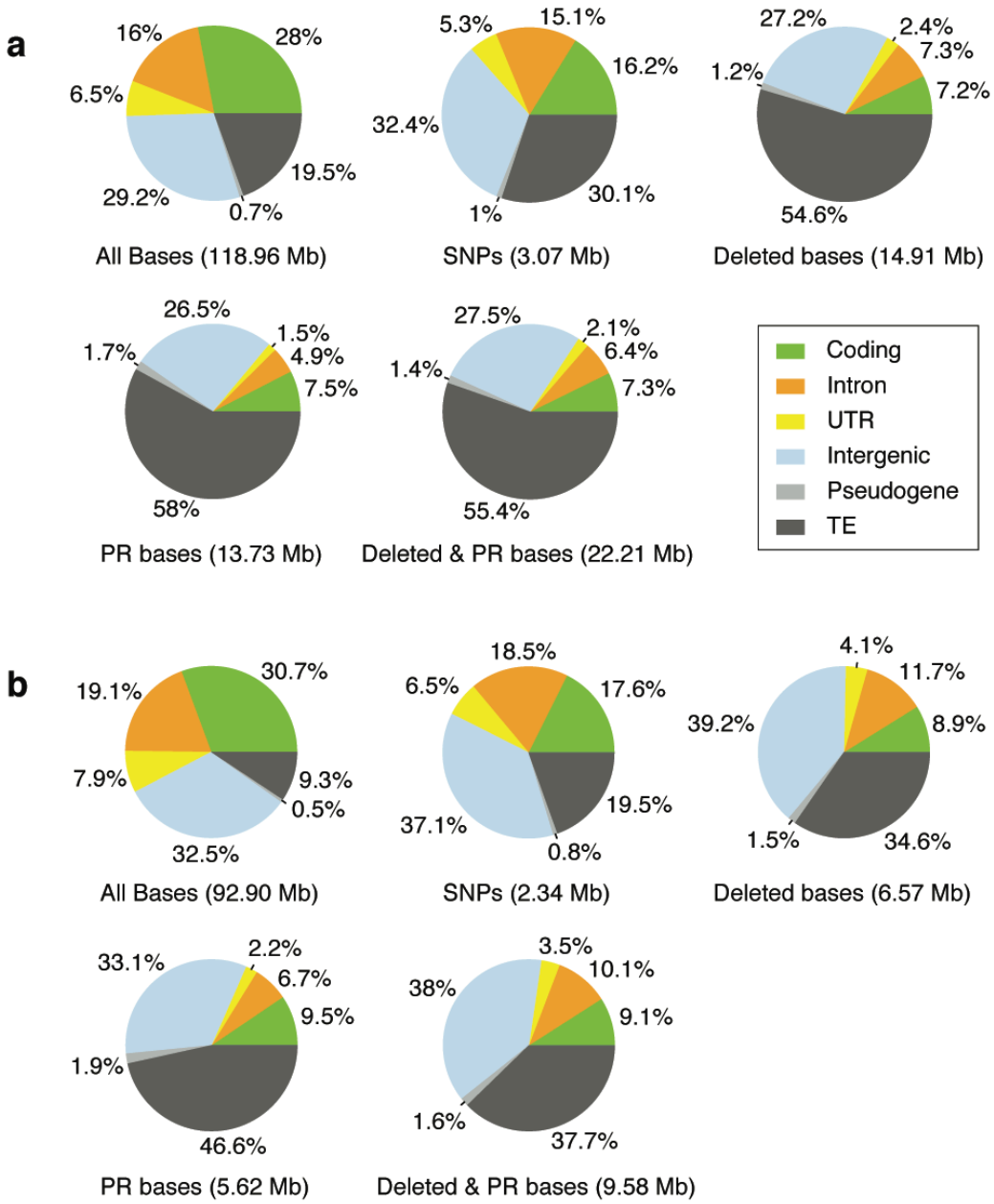
Supplementary Figure 5. Distribution of ambiguous bases relative to all non-reference base calls along chromosomes 1-5 for accessions Can-0 and Bur-0. The percentage of ambiguous bases ("K", "M", "R","S", "W", "Y") as a function of all non-reference base calls (ambiguous bases and called SNP bases) in Bur-0 and Can-0 was calculated in overlapping windows of length 1.0 Mb (midpoints of windows are plotted, with 200 kb offsets). The grey line denotes the density of non-reference base calls calculated with the same window size and offset using all accessions, and with normalization to the maximum value in the genome. Approximate positions of centromeres (CEN) are after Clark *et al.*[18]. Because of assembly uncertainties in highly repetitive regions, centromeric sequences were excluded from the analysis. Residual regions of heterozygosity in the founder accessions, which would be expected to give rise to extended tracts of high % ambiguity, are not apparent in either Bur-0 or Can-0. The pattern of apparent uniform homozygosity observed for Bur-0 and Can-0 is shared with other accessions except for Hi-0, Mt-0, Po-0, Rsch-4 and Sf-2 (see Supplementary Figs 6 and 7).



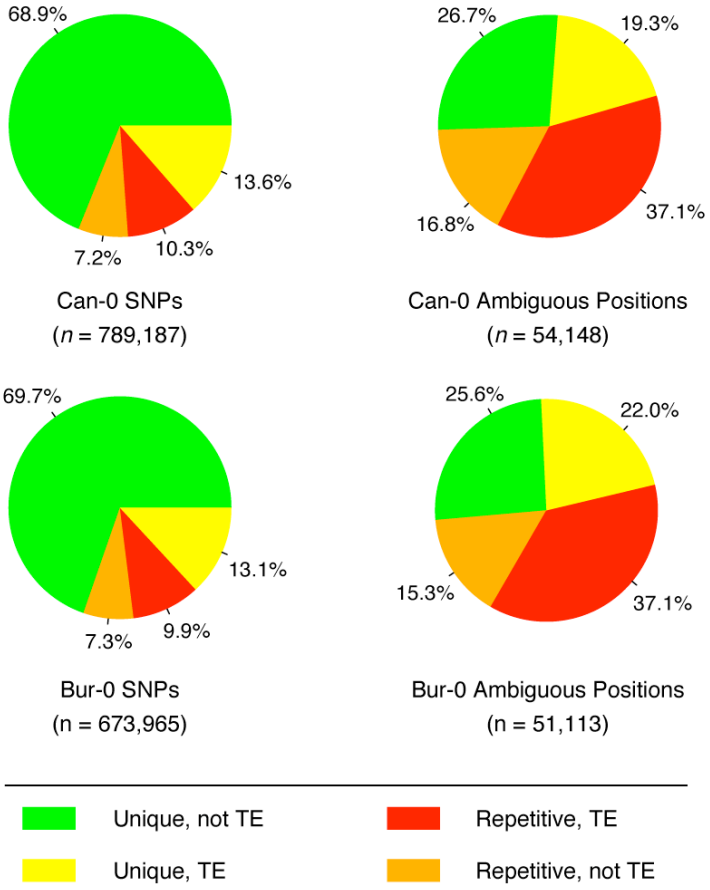
Supplementary Figure 6. The Po-0 accession used for genome sequencing and assembly was extensively heterozygous. Plotted values and features are as for Supplementary Fig. 5 but for accession Po-0.



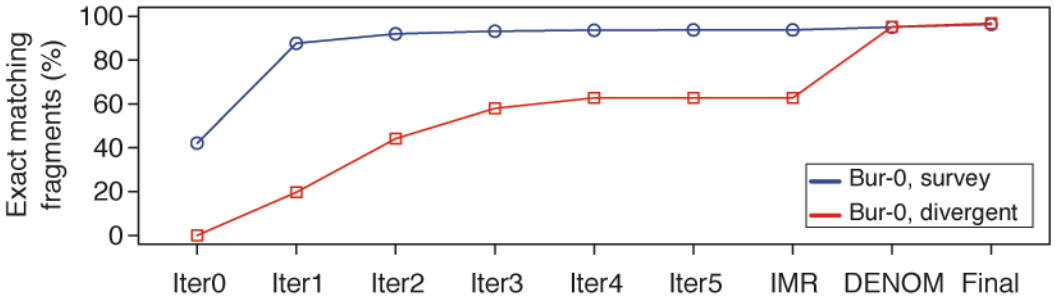
Supplementary Figure 7. Regions of apparent heterozygosity in accessions Hi-0, Mt-0, Rsch-4 and Sf-2. Plotted values and features are as for Supplementary Fig. 5 but for accessions Hi-0, Mt-0, Rsch-4 and Sf-2 (colour coded as at bottom). Visual inspection revealed apparent tracts of heterozygosity in each of these accessions [thick bars above each chromosome denote the regions of apparent heterozygosity: Hi-0 (Chr 2: 0-10 Mb, 13-16), Mt-0 (Chr 1: 3.5-5 Mb), Rsch-4 (Chr 4: 13-15 Mb), and Sf-2 (Chr 2: 13-17.5 Mb and Chr 3: 0-8 Mb)]. Excluding Po-0, which was extensively heterozygous (see Supplementary Fig. 6), only ~1.4% of genomic sequences across all of the accessions were observed to be obviously heterozygous.



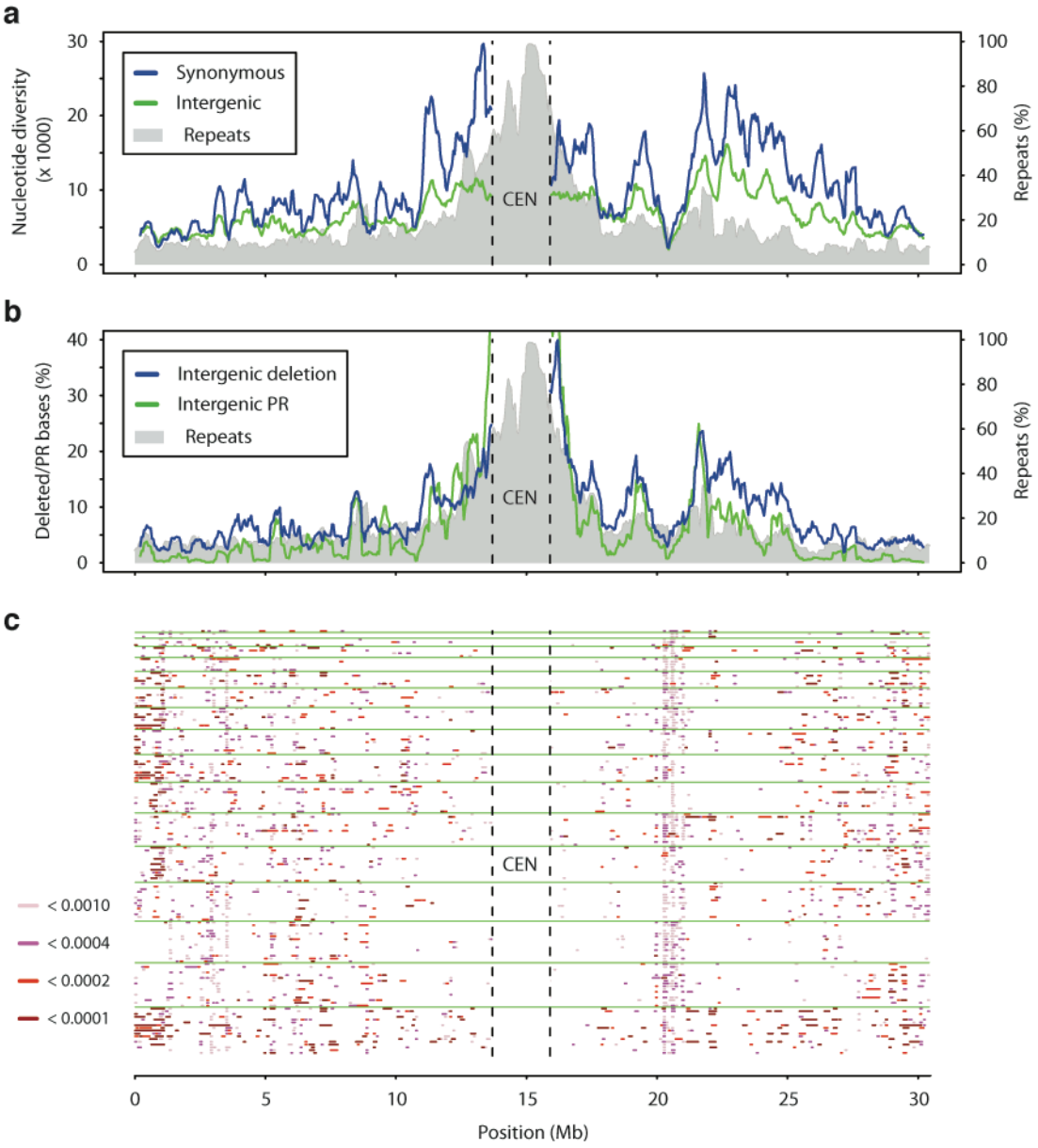
Supplementary Figure 8. Distribution of polymorphisms by sequence type. Percentage of (a) all bases and (b) bases in unique regions by annotation and polymorphism types. Classifications for all genomic positions (top left in panels a and b) are from the TAIR10 annotation (Supplementary Information section 5.1). Unique positions are for repeat level 0 as defined in Supplementary Information section 8.1 Annotation classifications are as indicated at bottom right, and bases inclusive to a polymorphism category are as indicated below each pie chart. Values reflect non-redundant counts [for instance, 6.57 Mb (top right chart) is deleted in at least one accession in unique regions].



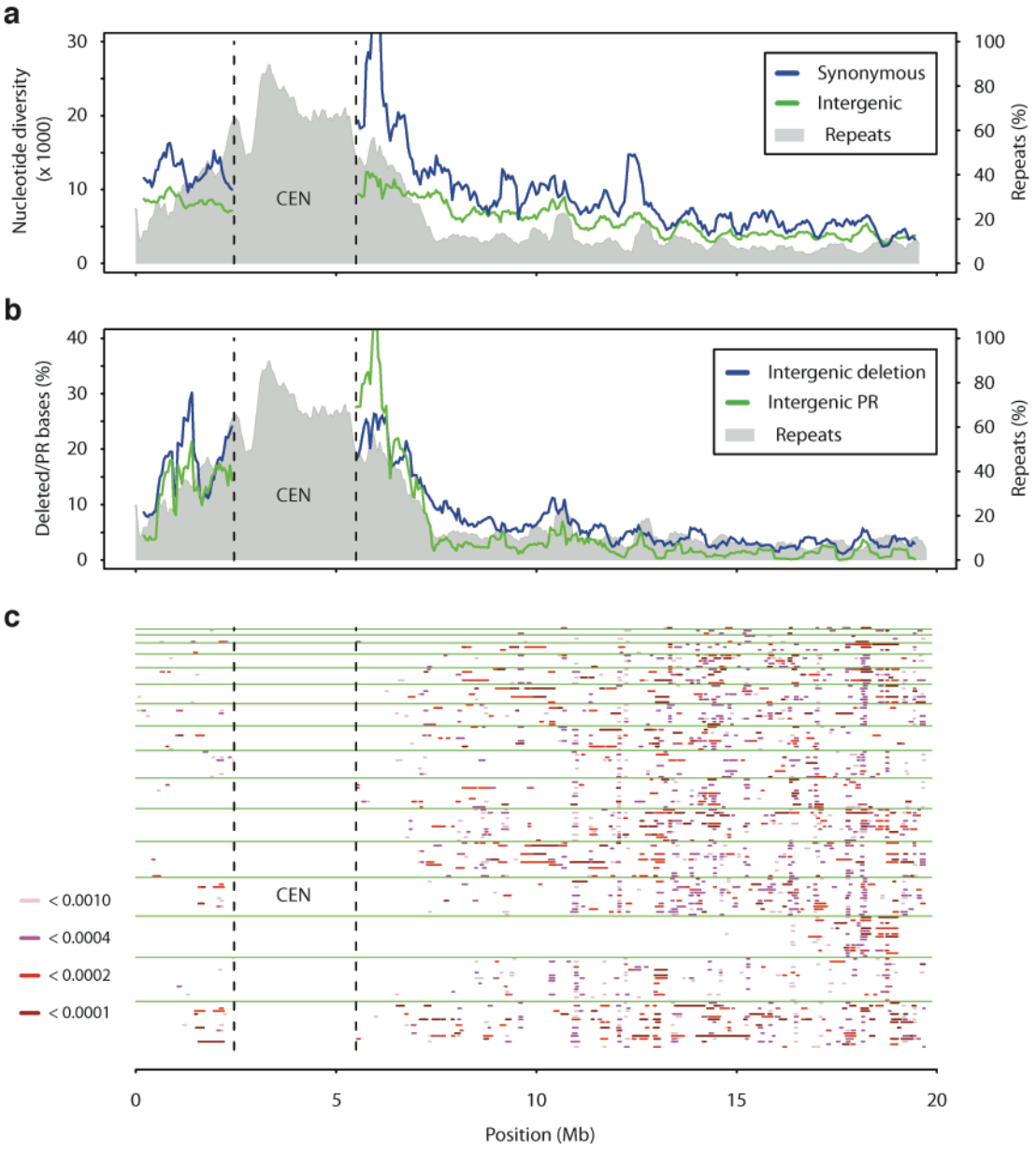
Supplementary Figure 9. SNP and ambiguous positions by repeat level and transposable element content. For Bur-0 and Can-0, two accessions for which no regions of residual heterozygosity were apparent (see Supplementary Fig. 5), the relationships between SNP positions and ambiguous positions (bases "K", "M", "R", "S", "W", "Y") as a function of repeat and transposable element (TE) content is shown. Approximate 83% of SNPs are at unique bases (left), with ~13% of SNP positions in annotated TEs that are nonetheless in unique regions as assessed with whether a 50mer maps uniquely in the reference genome sequence (see legend at bottom, and Supplementary Information section 5.1). Compared to SNPs, ambiguous positions (right) are more frequent in repetitive regions, as well as in TE sequences regardless of whether the TE regions are unique or repetitive.



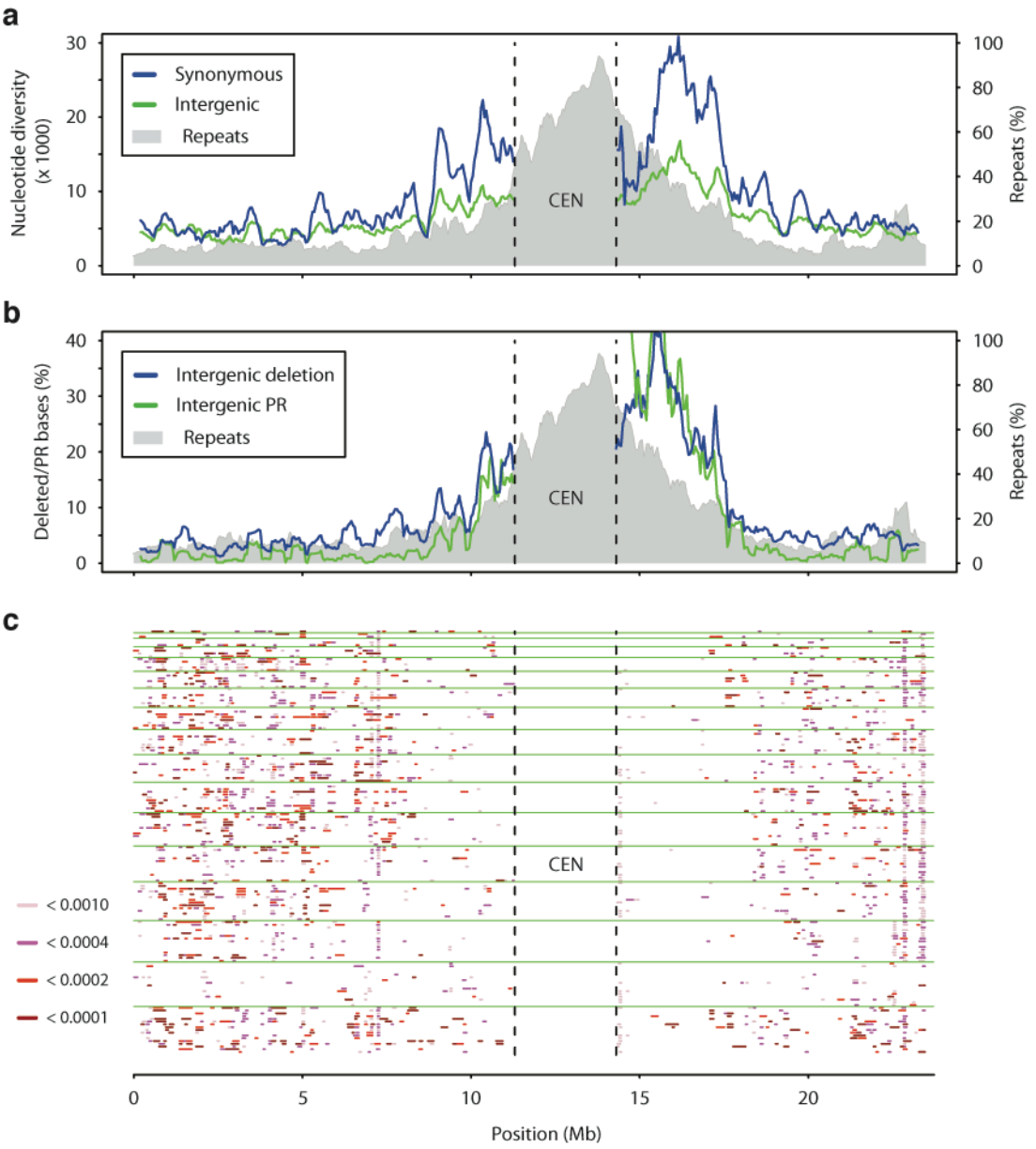
Supplementary Figure 10. Validation sequence matches to the Bur-0 assembly as a function of assembly step. Shown are the percent of perfectly matching fragments from two datasets of PCR amplified and sequenced regions for accession Bur-0 (the Bur-0 survey[6] and divergent[14] validation sets are as described in the legend for Fig. 1b). The *de novo* assembly step (DENOM) markedly improved accuracy as assessed with the divergent dataset (red), for which sequence differences to the reference genome are largest (i.e., small to moderate sized indels, SNP clusters, or both).



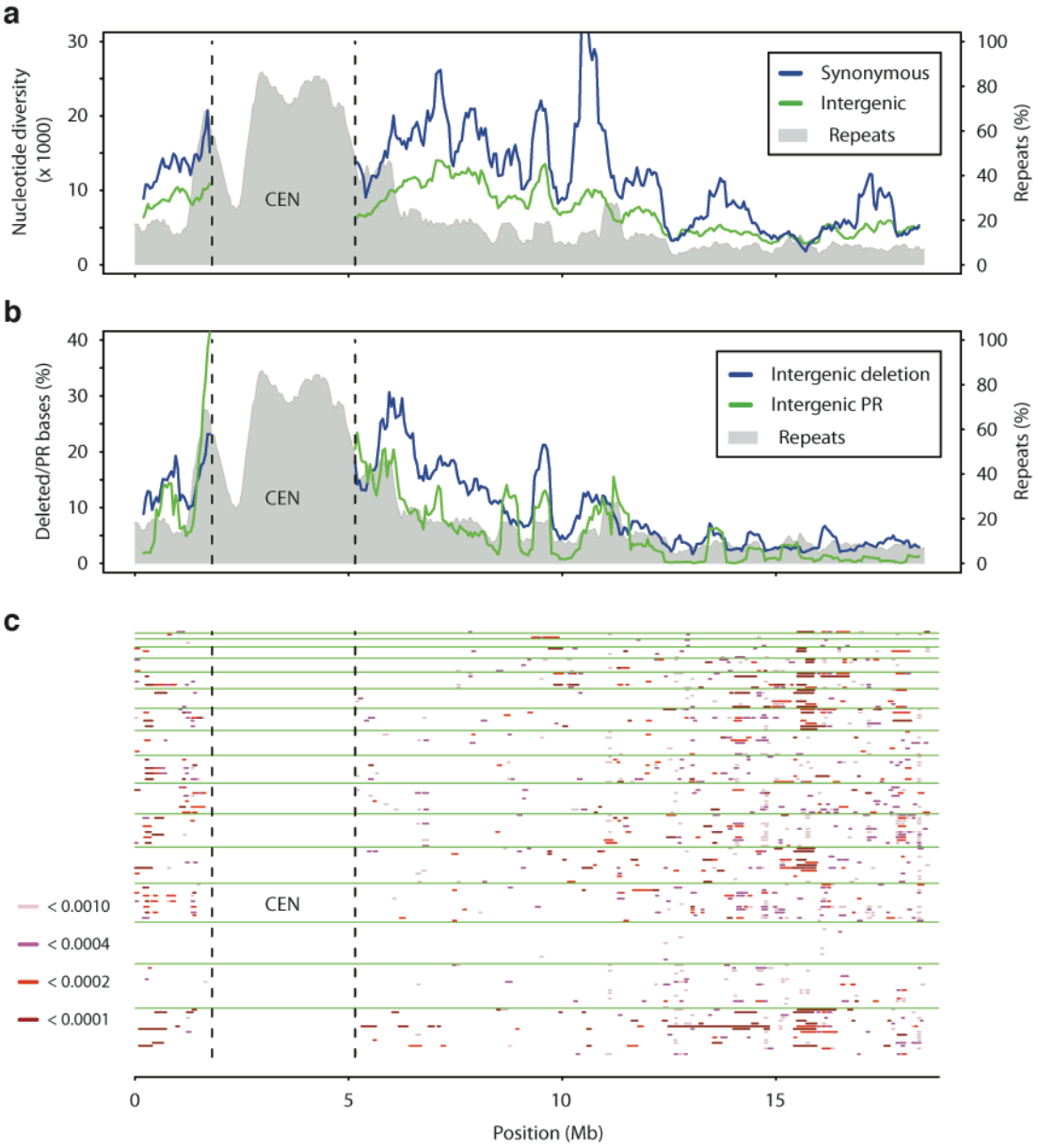
Supplementary Figure 11. Genome-wide patterns of diversity, chromosome 1. (a) Average pairwise nucleotide diversity for the sample of 18 accessions (Po-0 excluded) is plotted for synonymous coding and intergenic sites along the chromosome with sliding windows of 400 kb (counted from all sites, as measured against the TAIR10 sequence) with a 50 kb offset. The value for each window is plotted at its midpoint. Percent of each window masked as repetitive is also displayed. (b) Percent of unique (nonrepetitive) intergenic sites in each window that are either deleted or involved in a PR region in at least one accession. (c) Regions of extensive pairwise haplotype sharing, with colour representing level of similarity measured as differences at SNP positions divided by total number of unique sites compared. Comparisons between pairs of accessions are represented as rows sorted along the *y* axis. Horizontal green lines demarcate comparisons using one accession. Each possible pairwise comparison is shown only once. For each panel, the portion of the chromosome with the highest repeat content corresponding to the approximate location of the centromere (defined as in Clark *et al.*[18]) is indicated, and only repeat content is plotted here.



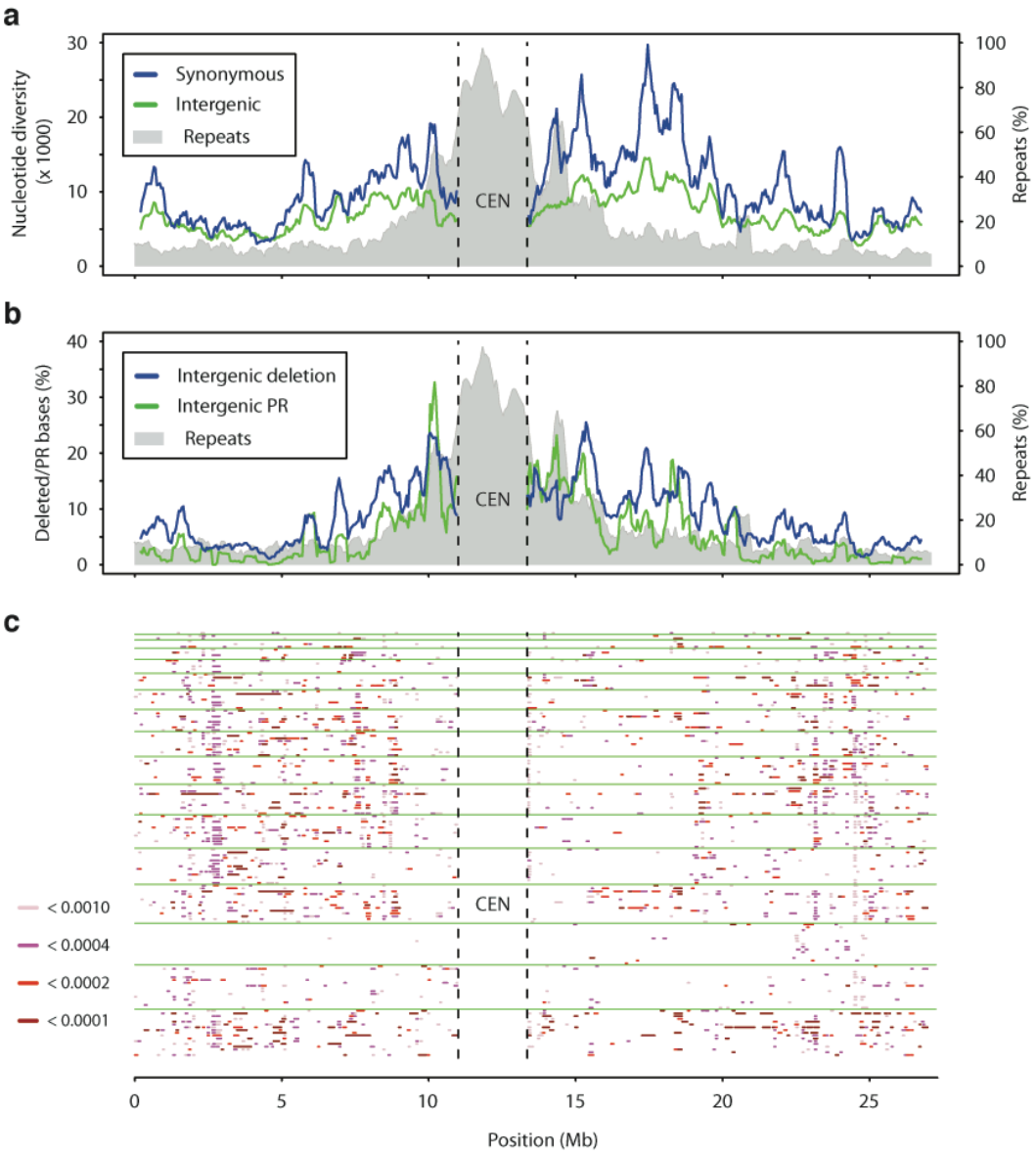
Supplementary Figure 12. Genome-wide patterns of diversity, chromosome 2. (a) Average pairwise nucleotide diversity for the sample of 18 accessions (Po-0 excluded) is plotted for synonymous coding and intergenic sites along the chromosome with sliding windows of 400 kb (counted from all sites, as measured against TAIR10 sequence) with a 50 kb offset. The value for each window is plotted at its midpoint. Percent of each window masked as repetitive is also displayed. (b) Percent of unique (nonrepetitive) intergenic sites in each window that are either deleted or involved in a PR region in at least one accession. (c) Regions of extensive pairwise haplotype sharing, with colour representing level of similarity measured as differences at SNP positions divided by total number of unique sites compared. Accession pairs are plotted as in Supplementary Fig. 11. For each panel, the portion of the chromosome with the highest repeat content corresponding to the approximate location of the centromere (defined as in Clark *et al.*[18]) is indicated, and only repeat content is plotted here.



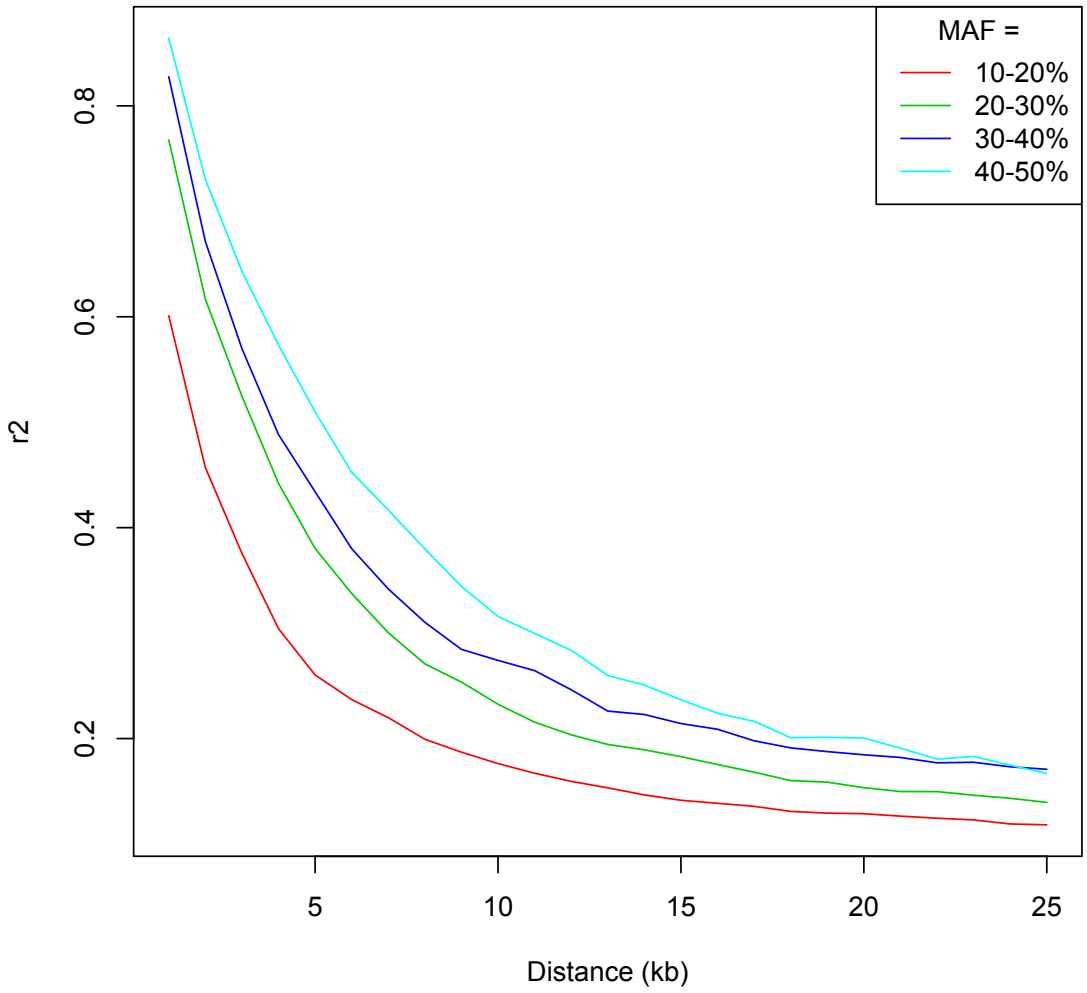
Supplementary Figure 13. Genome-wide patterns of diversity, chromosome 3. (a) Average pairwise nucleotide diversity for the sample of 18 accessions (Po-0 excluded) is plotted for synonymous coding and intergenic sites along the chromosome with sliding windows of 400 kb (counted from all sites, as measured against the TAIR10 sequence) with a 50 kb offset. The value for each window is plotted at its midpoint. Percent of each window masked as repetitive is also displayed. (b) Percent of unique (nonrepetitive) intergenic sites in each window that are either deleted or involved in a PR region in at least one accession. (c) Regions of extensive pairwise haplotype sharing, with colour representing level of similarity measured as differences at SNP positions divided by total number of unique sites compared. Accession pairs are plotted as in Supplementary Fig. 11. For each panel, the portion of the chromosome with the highest repeat content corresponding to the approximate location of the centromere (defined as in Clark *et al.*[18]) is indicated, and only repeat content is plotted here.



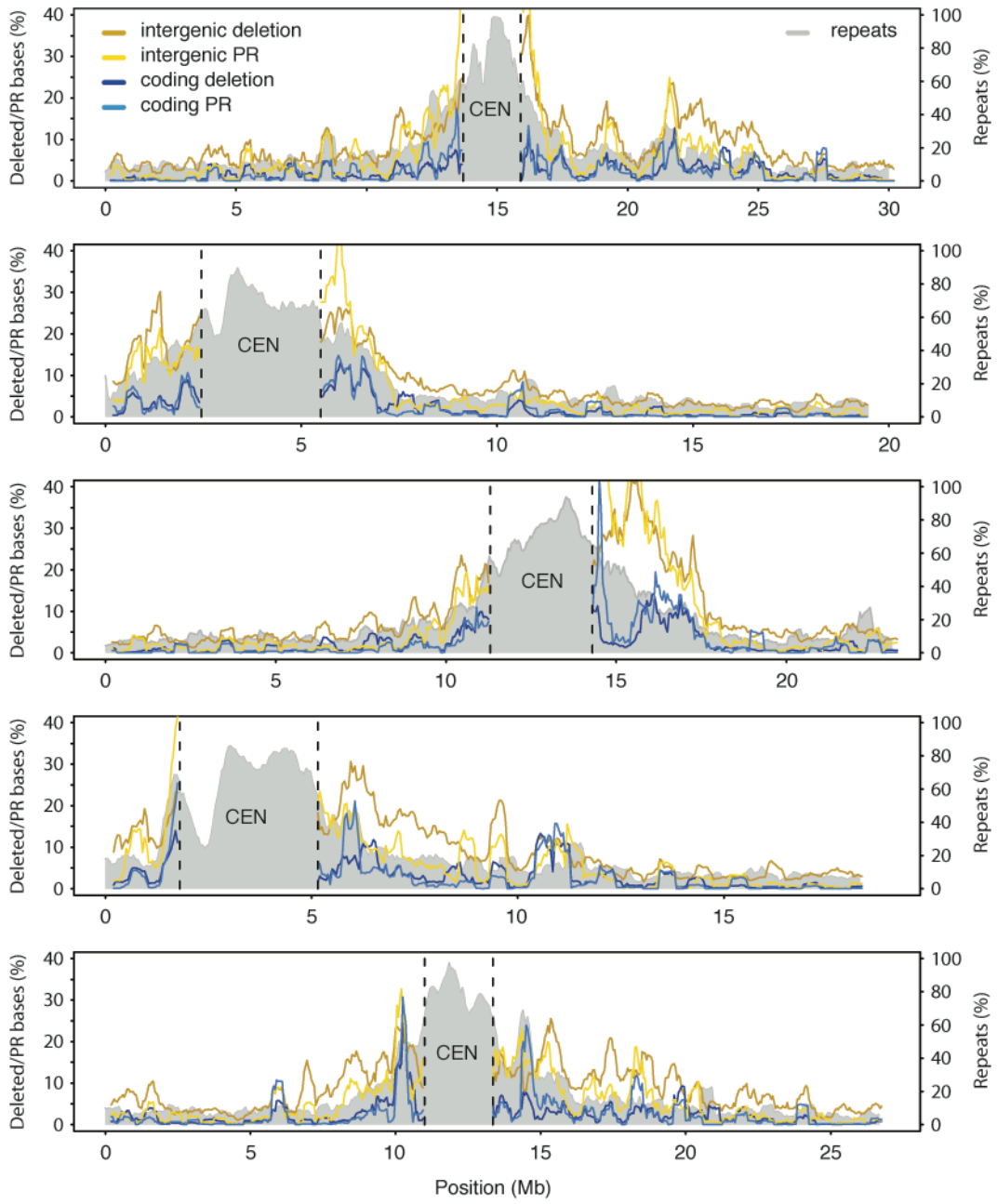
Supplementary Figure 14. Genome-wide patterns of diversity, chromosome 4. (a) Average pairwise nucleotide diversity for the sample of 18 accessions (Po-0 excluded) is plotted for synonymous coding and intergenic sites along the chromosome with sliding windows of 400 kb (counted from all sites, as measured against the TAIR10 sequence) with a 50 kb offset. The value for each window is plotted at its midpoint. Percent of each window masked as repetitive is also displayed. (b) Percent of unique (nonrepetitive) intergenic sites in each window that are either deleted or involved in a PR region in at least one accession. (c) Regions of extensive pairwise haplotype sharing, with colour representing level of similarity measured as differences at SNP positions divided by total number of unique sites compared. Accession pairs are plotted as in Supplementary Fig. 11. For each panel, the portion of the chromosome with the highest repeat content corresponding to the approximate location of the centromere (defined as in Clark *et al.*[18]) is indicated, and only repeat content is plotted here.



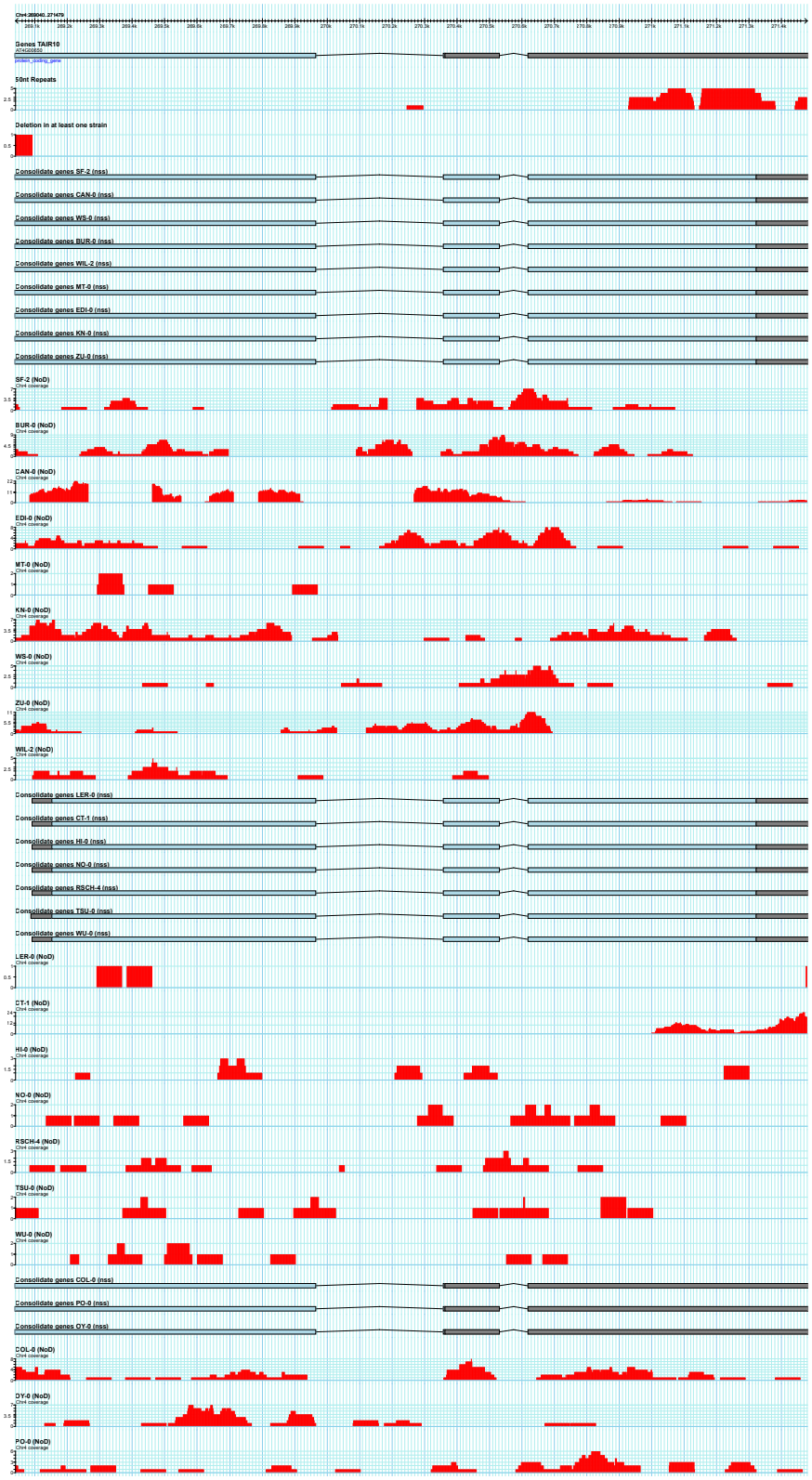
Supplementary Figure 15. Genome-wide patterns of diversity, chromosome 5. (a) Average pairwise nucleotide diversity for the sample of 18 accessions (Po-0 excluded) is plotted for synonymous coding and intergenic sites along the chromosome with sliding windows of 400 kb (counted from all sites, as measured against the TAIR10 sequence) with a 50 kb offset. The value for each window is plotted at its midpoint. Percent of each window masked as repetitive is also displayed. (b) Percent of unique (nonrepetitive) intergenic sites in each window that are either deleted or involved in a PR region in at least one accession. (c) Regions of extensive pairwise haplotype sharing, with colour representing level of similarity measured as differences at SNP positions divided by total number of unique sites compared. Accession pairs are plotted as in Supplementary Fig. 11. For each panel, the portion of the chromosome with the highest repeat content corresponding to the approximate location of the centromere (defined as in Clark et al. 2007[18]) is indicated, and only repeat content is plotted here.



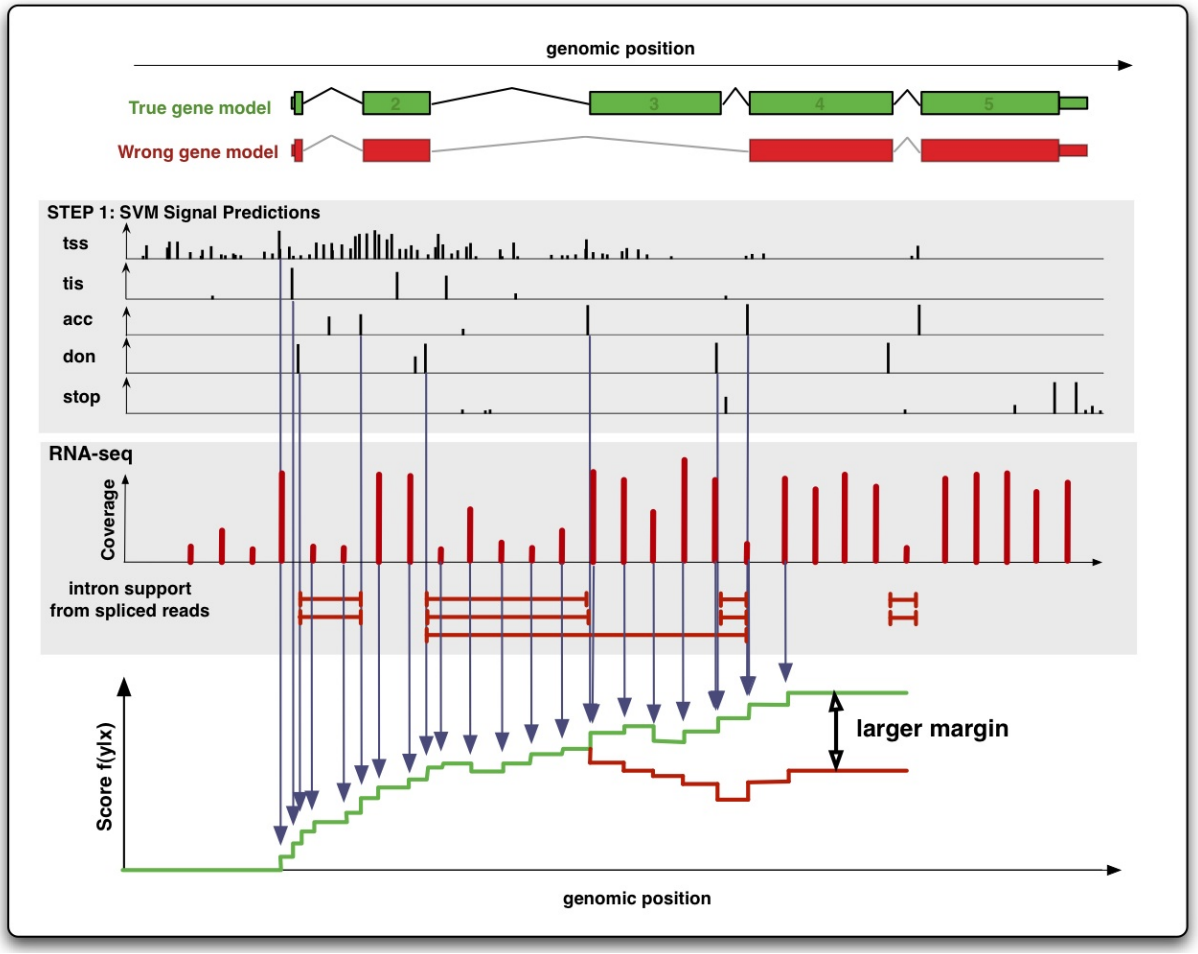
Supplementary Figure 16. Decay of linkage disequilibrium for pairs of SNPs with matched allele frequencies. Decay of r^2 is plotted as a function of distance between SNPs. SNPs were binned by minor allele frequency (MAF), then the LD measure r^2 was calculated for pairs of SNPs matched for MAF bin.



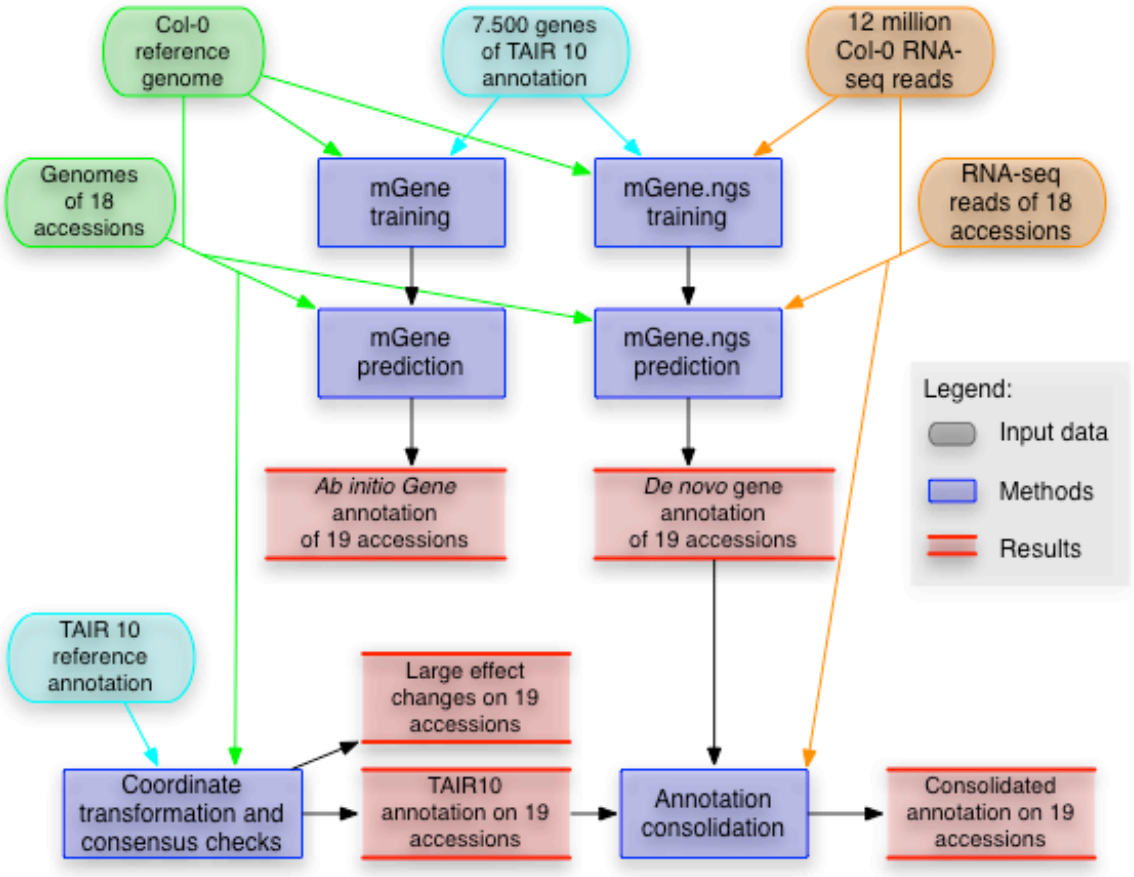
Supplementary Figure 17. Genome-wide occurrence of deletions and PR regions in intergenic and coding sequence. Proportion of unique (nonrepetitive) sequence from either intergenic or coding regions that were either deleted or defined as a PR region in at least one of the 17 non-Col-0 accessions (Po-0 excluded) is plotted along the chromosome with sliding windows of 400 kb (counted from all sites, as measured against the TAIR10 sequence) with a 50 kb offset. The value for each window is plotted at its midpoint. Percent of each window masked as repetitive is also displayed. The portion of the chromosome with the highest repeat content corresponding to the approximate location of the centromere (defined as in Clark et al. 2007[18]) is indicated, and only repeat content is plotted here.



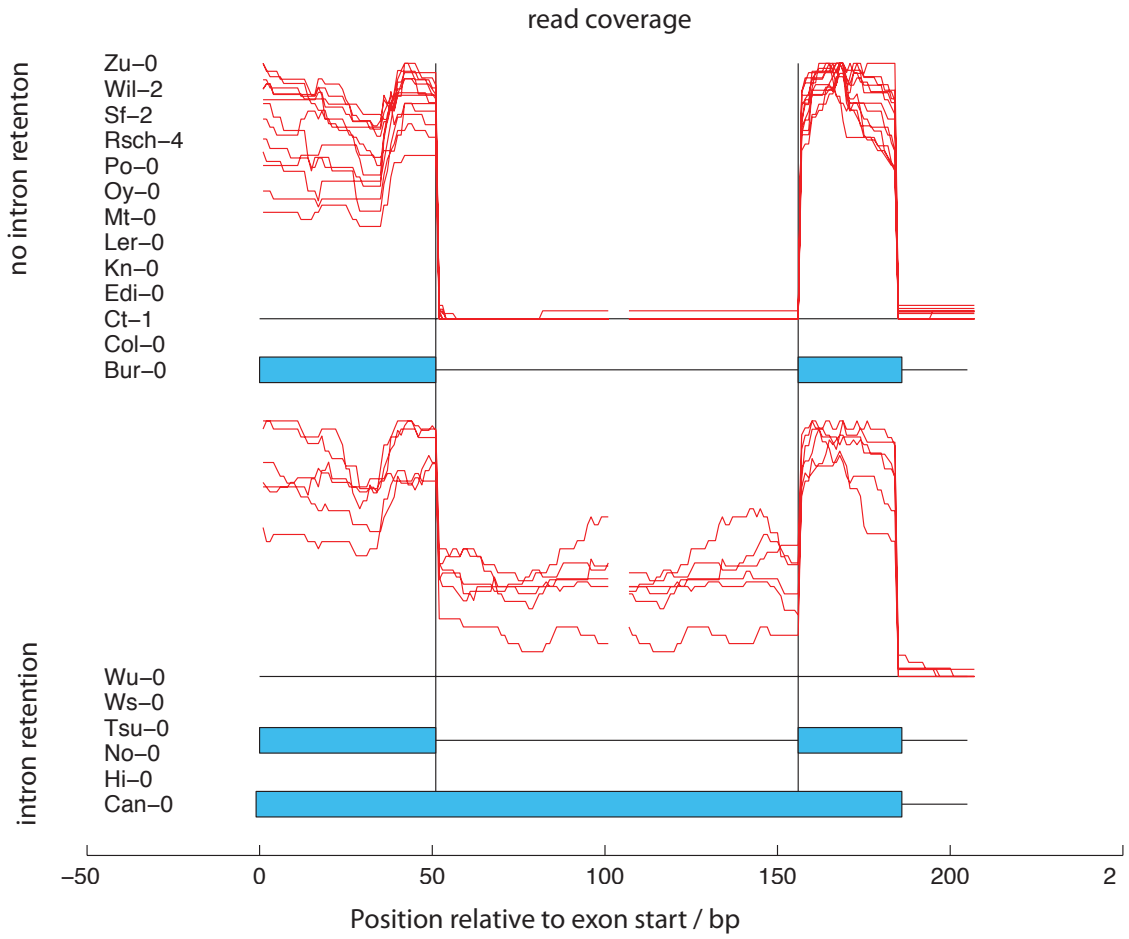
Supplementary Figure 18. Gene annotations and RNA-seq evidence for the *FRIGIDA* locus. Our annotation and amino-acid sequence analysis pipeline detected three distinct isoforms shared among the 19 accessions (confirming existing knowledge). We find RNA-seq evidence for expression in most accessions (only considering unique matches, see Supplementary Information section 9), except for Ler-0 and only weakly for Mt-0 and CT-1.



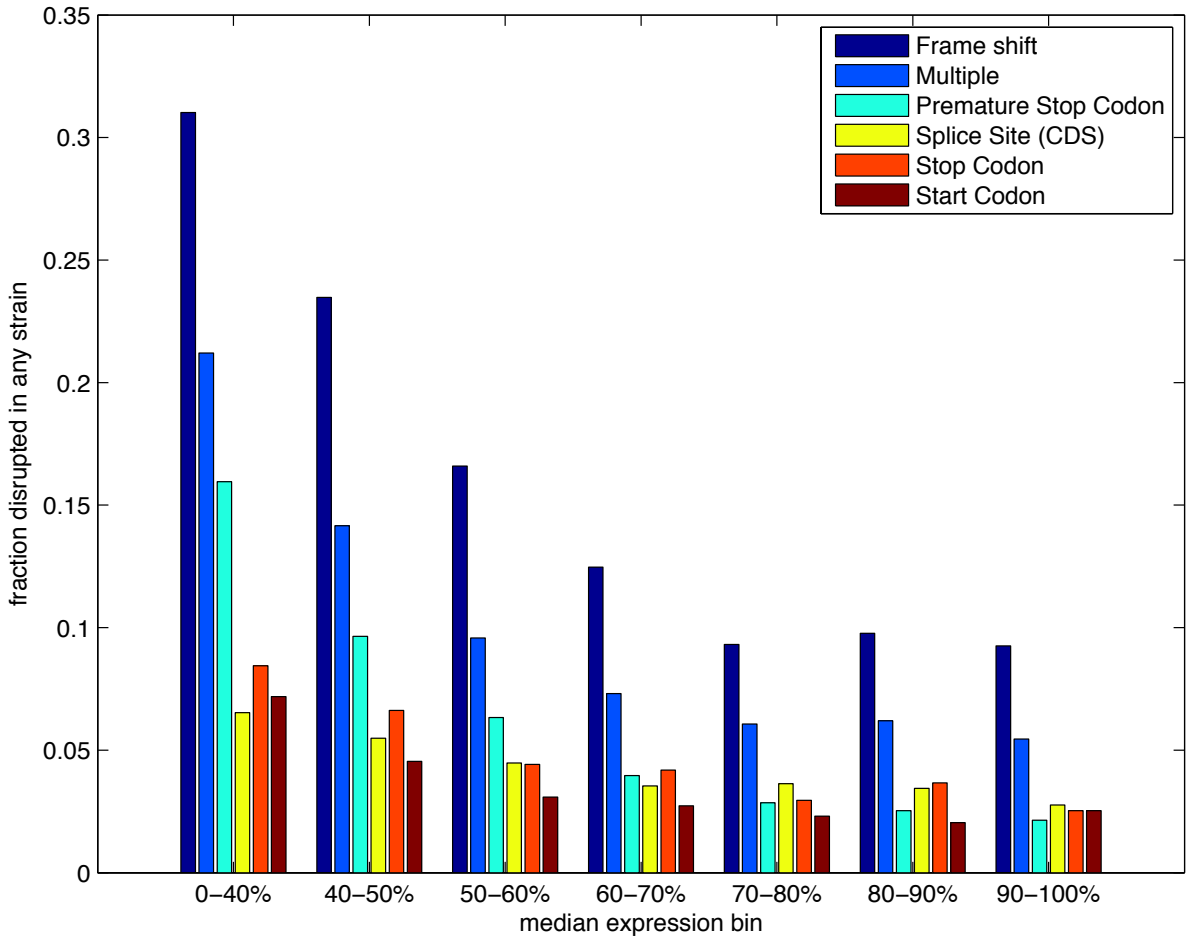
Supplementary Figure 19. Outline of the mGene.ngs approach. We use SVM-based signal predictions of transcription start, translation start, splice sites and translation stop and transcriptions stop sites. Combined with evidence from RNA-seq read alignments including the read-coverage and spliced alignments, we compute a score for possible gene structures. The parameters of the scoring function are tuned such that known gene models in a training set score significantly higher than wrong gene structures. Using the optimized scoring function we can then predict gene structures *de novo*.



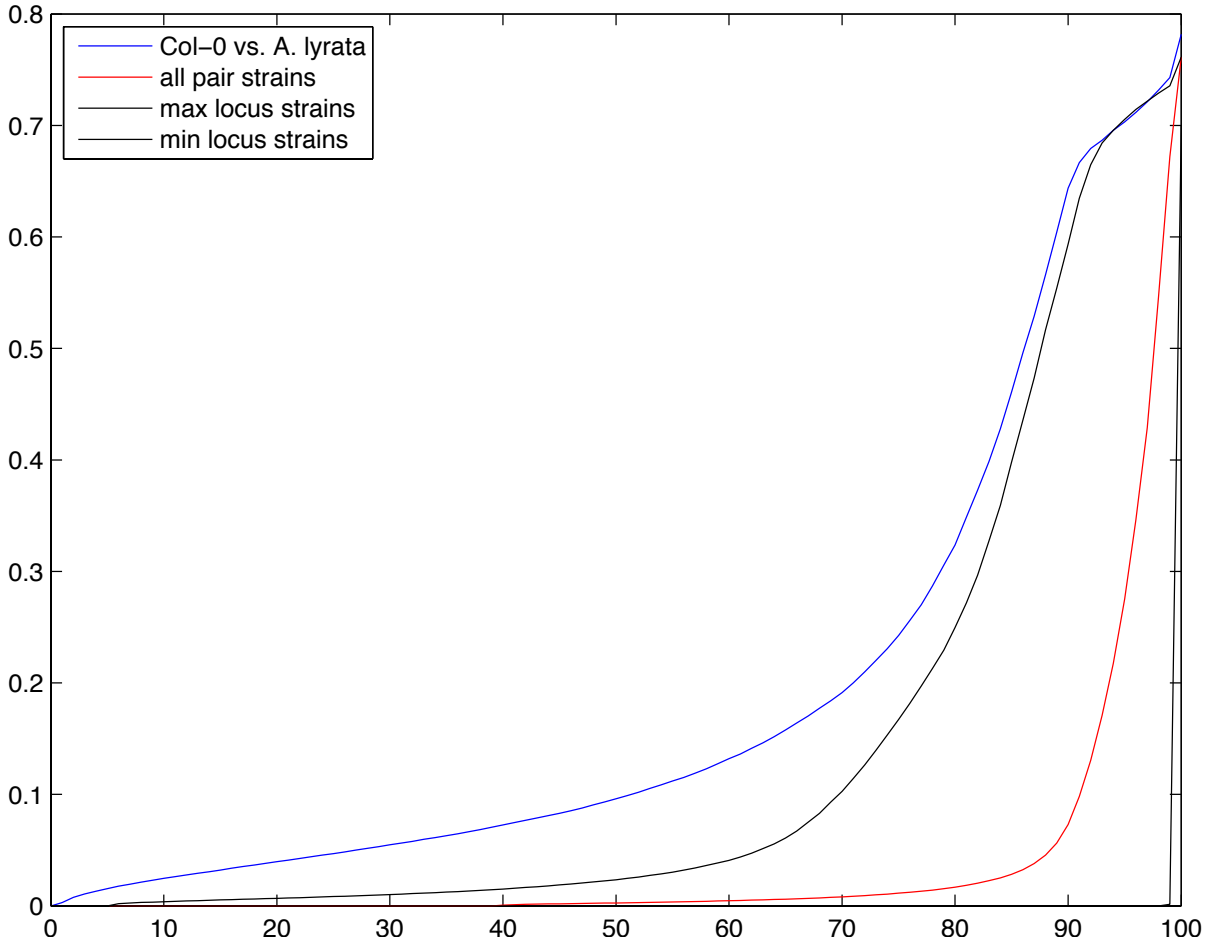
Supplementary Figure 20. Shown is the workflow of the re-annotation of the 19 accessions without and with using RNA-seq read alignments. The mGene gene predictor is trained using the reference genome and 7.500 protein-coding genes of the reference annotation. After training the trained system is applied to the reference genome as well as the 18 other accession's genomes leading to *ab initio* gene predictions. mGene.ngs additionally uses RNA-seq read alignments during training and prediction and produces *de novo* gene predictions. In a consolidation step, the TAIR 10 annotation mapped to the accession genomes is merged with the de novo gene predictions to obtain consolidated gene predictions for each accession.



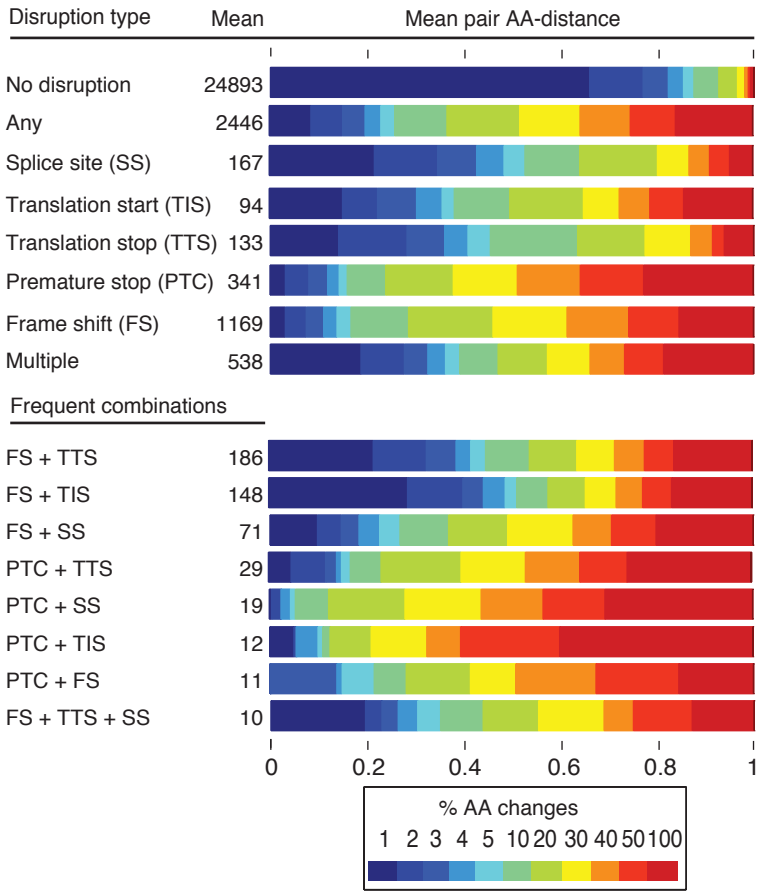
Supplementary Figure 21. Example for differential intron retention that is only present within a subset of the strains.



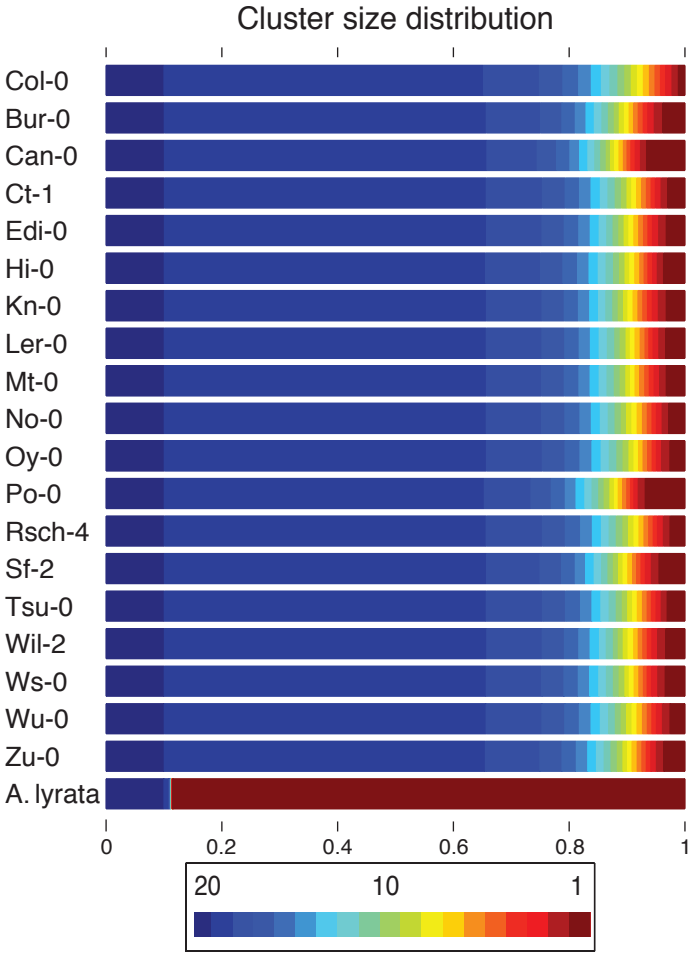
Supplementary Figure 22. Frequency of occurrence of different types of disruptions in at least one accession for groups of genes with varying median expression level among the accessions. The lowest expressed genes (median RPKM ≈0) show a 3-8 times higher frequency of disruption than the highest expressed genes.



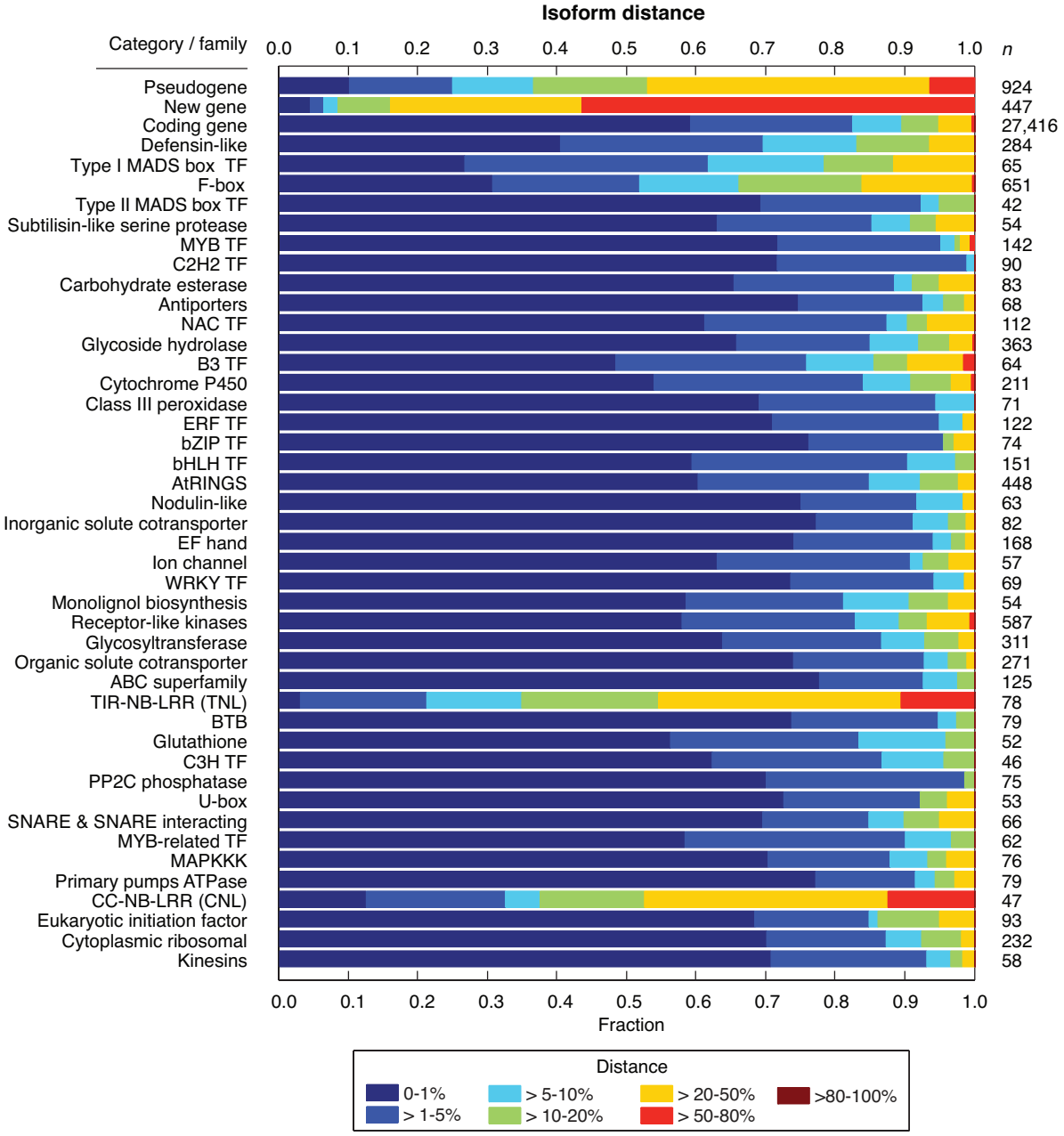
Supplementary Figure 23. Comparison between the AA sequences of between different species and accessions: For each AA sequence in one species/accession, we determine the sequence with smallest AA-distance in the other species/accession by globally aligning the sequences to plausible candidates. Shown is the relation of the fraction of protein sequences (x-axis) with AA-distance smaller than a threshold (y-axis). We compare the proteome differences between *A. thaliana* (Col-0) and *A. lyrata*, the average difference between any pair of the 19 accessions (pair average), the minimal and the maximal change among all pairs of accessions (pair min/max).



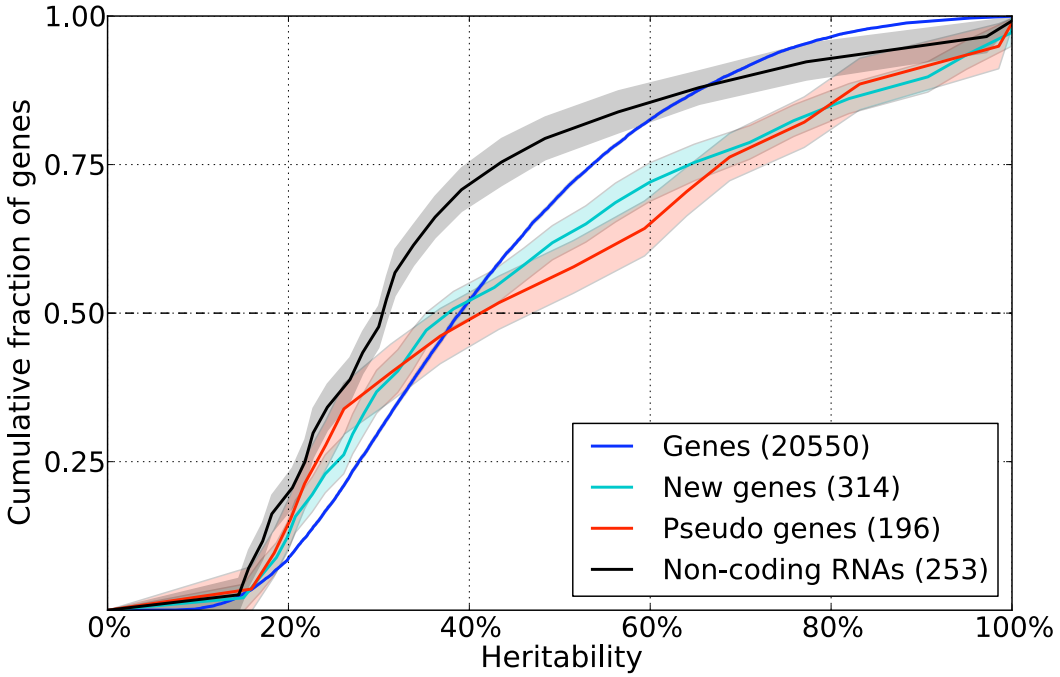
Supplementary Figure 24. Influence of large effect disruptions on AA sequence difference: Displayed is the fraction of genes with different degrees of AA sequence change between accessions with different combinations of disruptions and other accessions. We observe that splice site disruptions lead to least and FCs/PTCs lead to most severe AA-sequence changes. Among genes with multiple disruptions, combinations of FCs/PTCs with other disruptions, in particular TTS and TIS, are frequent. Surprisingly, the AA sequence differences for multiple disruptions are often smaller than one would expect, in particular smaller than the individual disruptions' sequence changes.



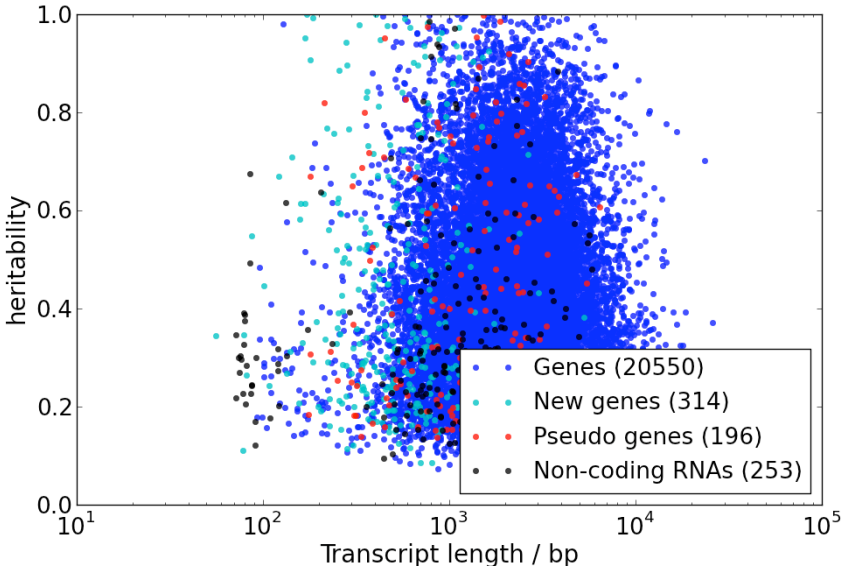
Supplementary Figure 25. Isoform frequencies for the 19 accessions and *A. lyrata*. For each gene and accession/species we determined the size of the cluster of which the gene's protein variant is part.



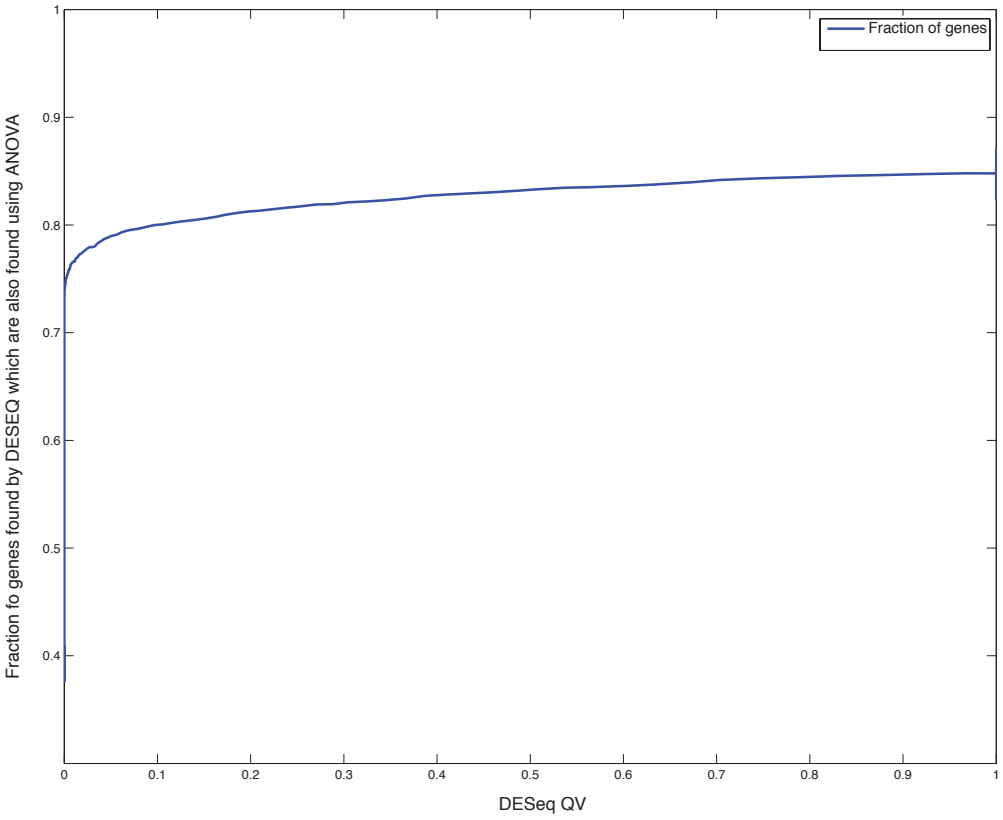
Supplementary Figure 26. Proteome diversity for gene categories and families. Reported is the fraction of genes with mean AA-distance to other accessions in the given interval. Gene categories and families shown are a superset of those shown in Fig. 4a. TF: transcription factor.



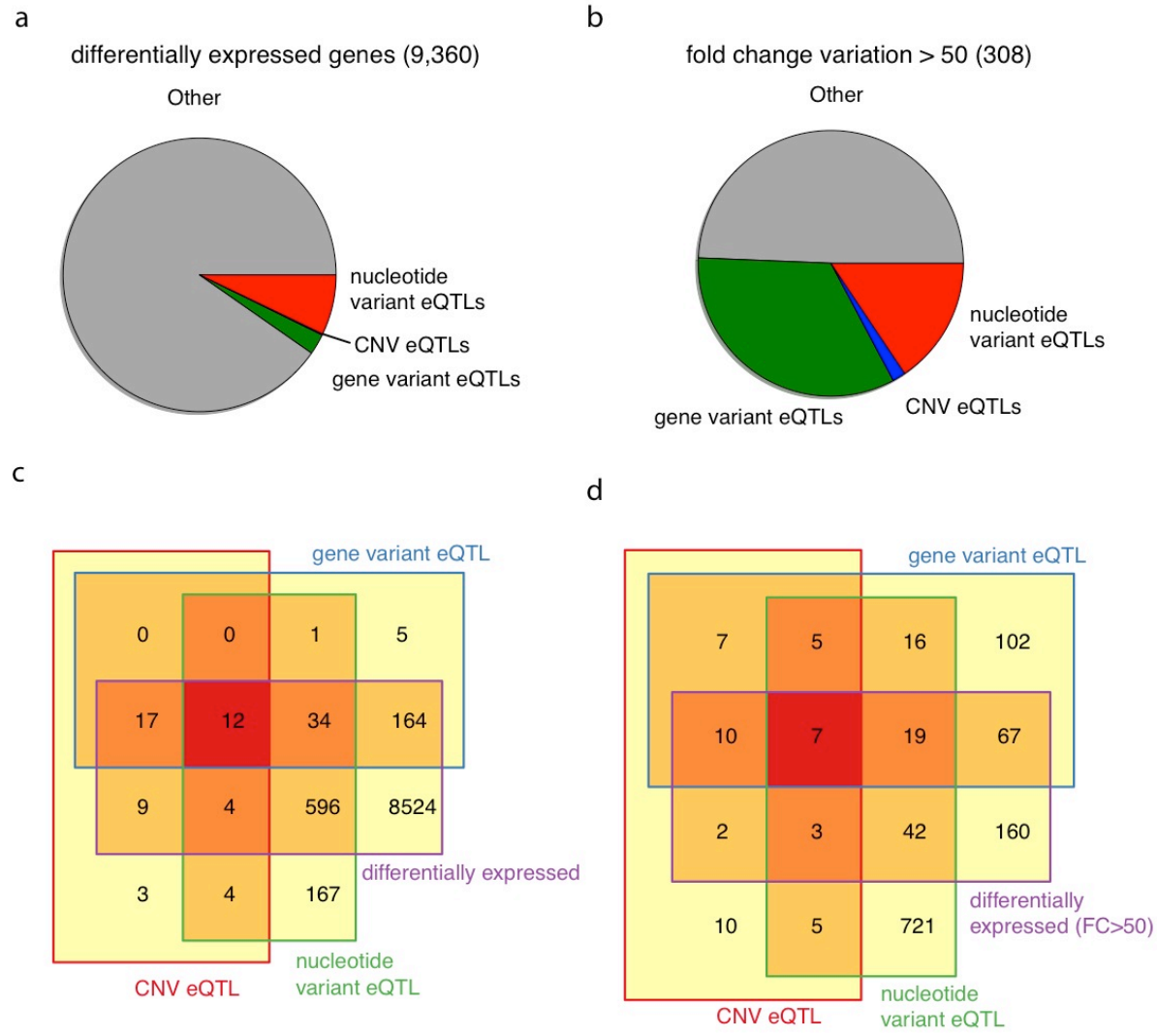
Supplementary Figure 27. Heritability of gene expression by gene type. Cumulative distribution functions of RNA-seq expression heritability in seedlings for different gene categories. Estimates are from expressed genes (q-value<0.05), where shaded areas indicate one standard deviation error bars, reflecting different sample sizes. Error bars were estimated from density estimation subset on 10-fold cross validation.



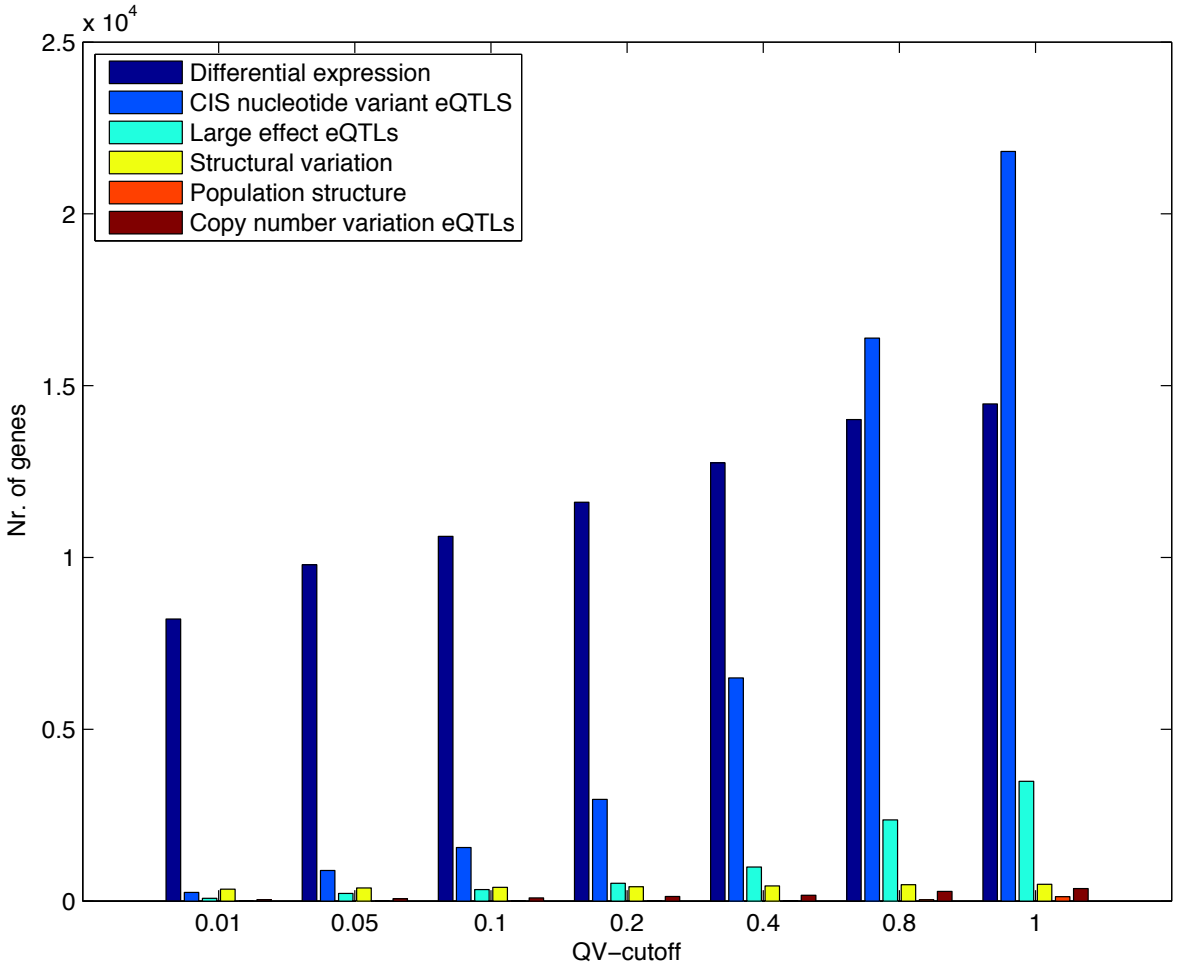
Supplementary Figure 28. Empirical heritability of individual genes from distinct gene types as a function of the gene length. There is a general trend that shorter genes are less heritable.



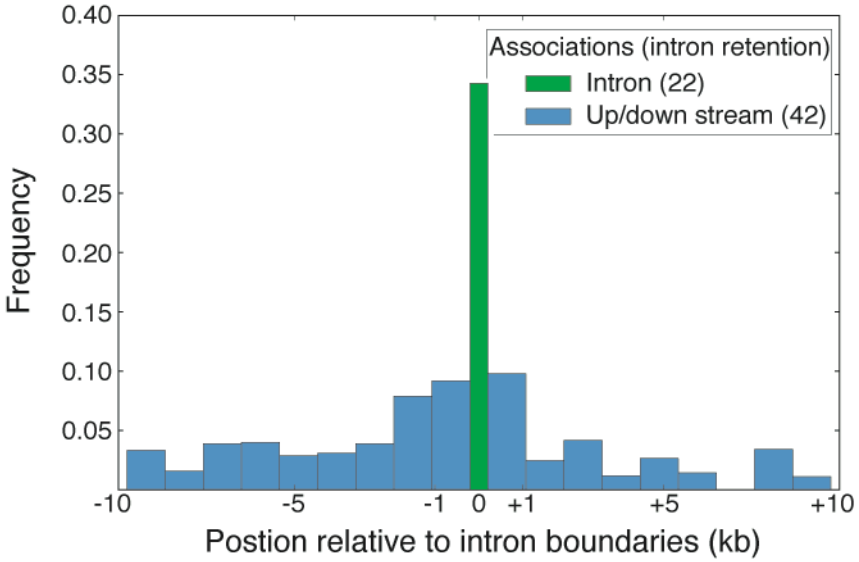
Supplementary Figure 29. Agreement of the ranking of genes detected as differentially expressed by DESeq with a ranking produced by ANOVA on variance stabilised counts.



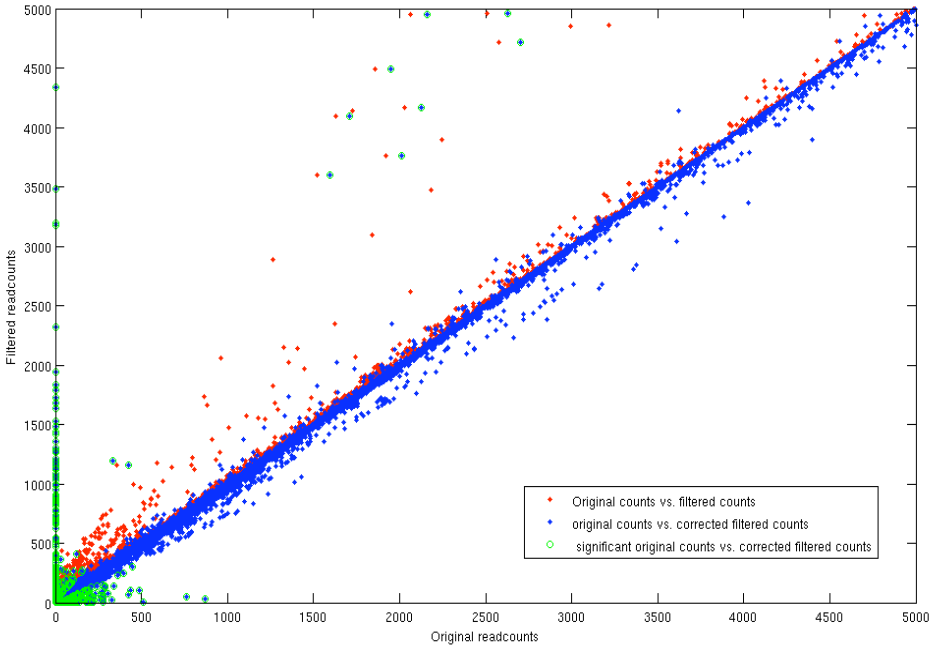
Supplementary Figure 30. Overview of different factors contributing to expression variation for all genes (left panel) and highly variables genes with fold change > 50 (right panel). **(a,b)** breakdown of differentially expressed genes into identified sources of variation. Overlaps resolved by priorisation in the order: gene variant eQTLs, CNV eQTLs, nucleotide variant eQTLs. **(c,d)** illustration of overlaps for the respective categories.



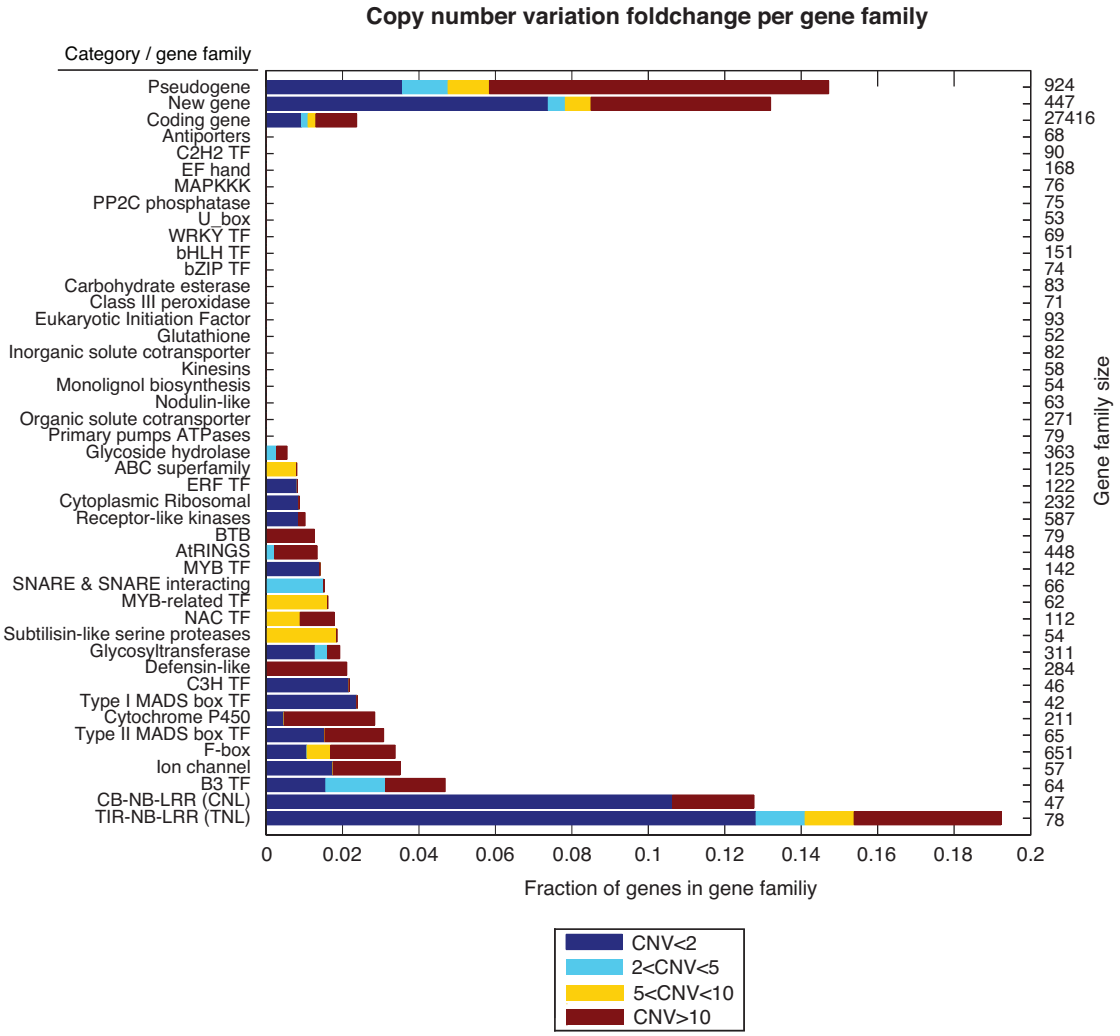
Supplementary Figure 31. Figure illustrating the number of differentially expressed genes (DESeq), genes with a *cis*-eQTL (*cis* eQTL), a large effect eQTL (le eQTL) or that exhibit significant expression changes because of structured variation (Structured variation). Shown are absolute numbers in each category for increasing FDR cut-off values.



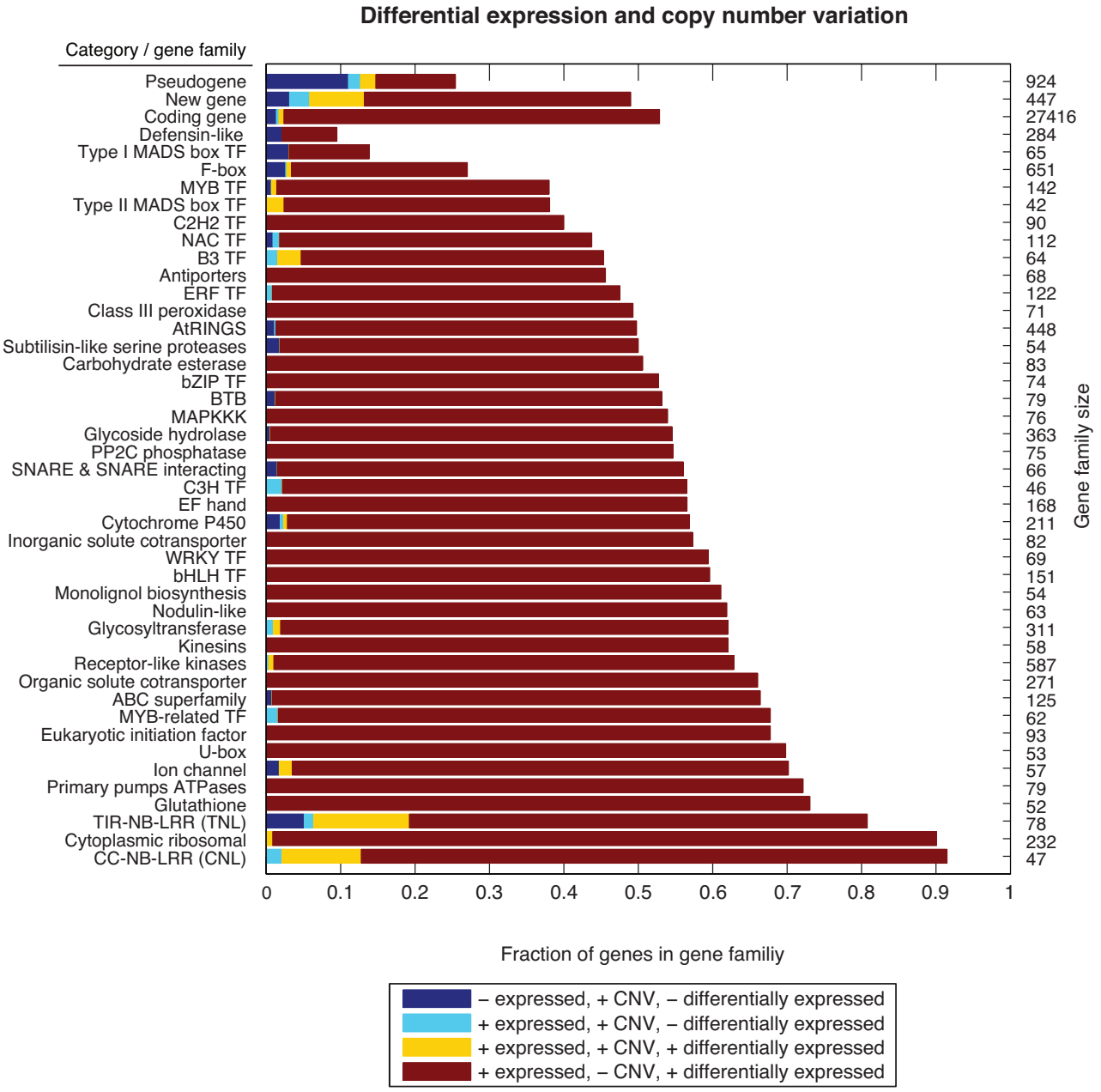
Supplementary Figure 32. Spatial distribution of genetic associations to intron retention. Shown is the position of the most associated SNP for 64 genetic associations of relative intron retention rates (FDR 10%), relative to the intron start.



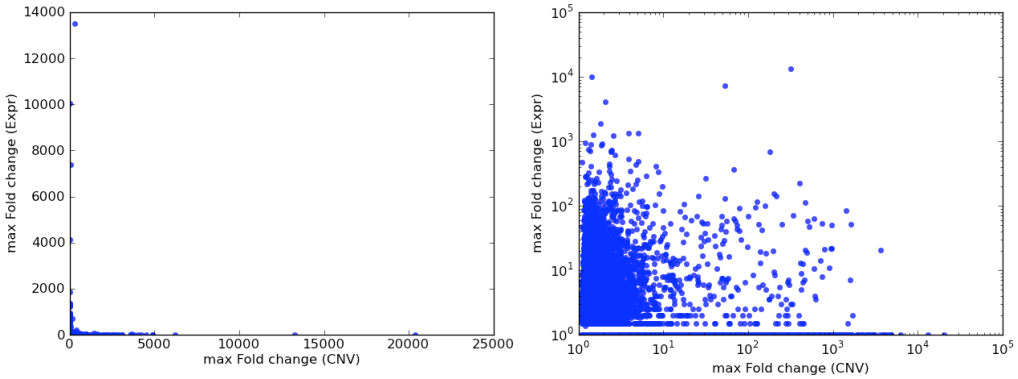
Supplementary Figure 33. Visualization of the impact of structural variation on gene expression estimates. Shown are read counts for expressed genes without filtering (X-axes) versus read counts when applying the structural variation filter. Because of a reduction of the gene length used for quantification, counts on the filtered set were rescaled (corrected filtered counts) and tested for significant differences of gene expression (5% FDR in green).



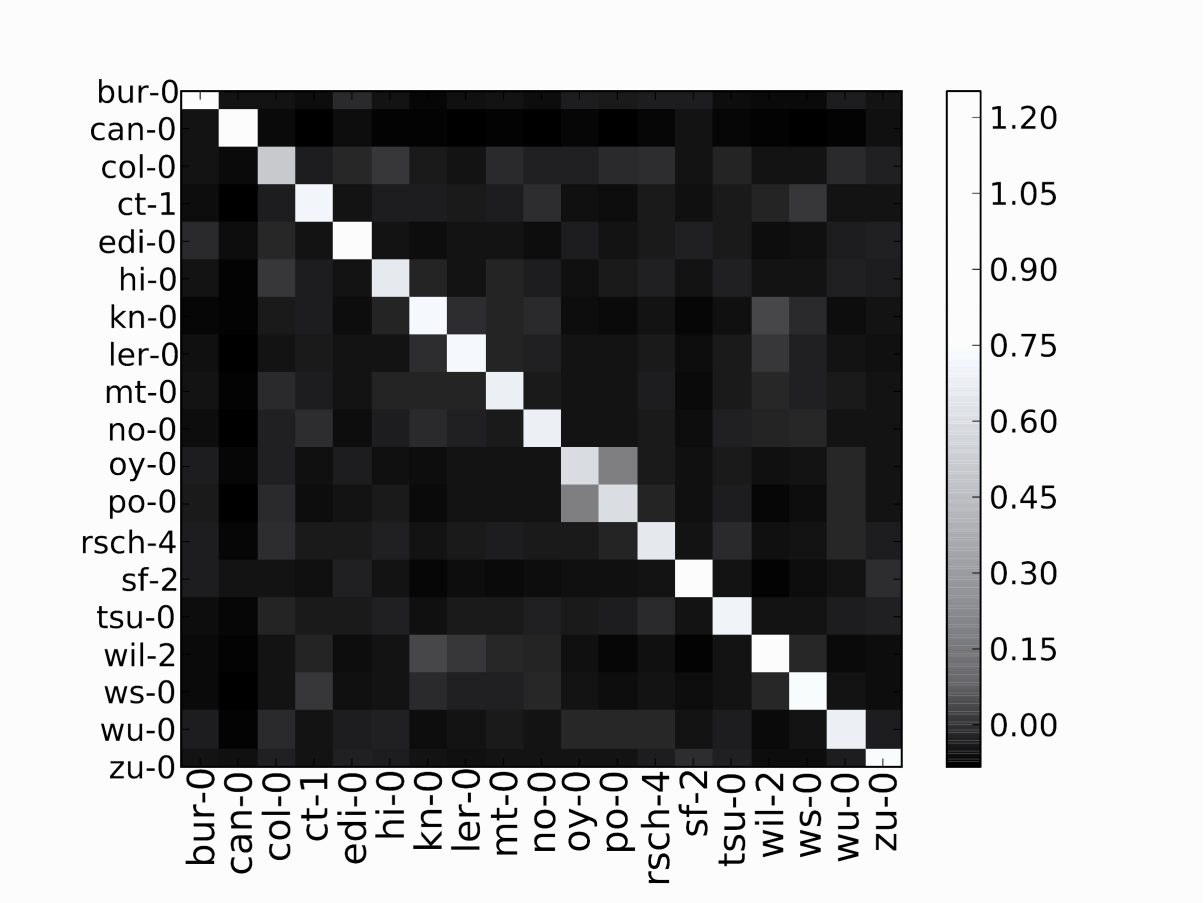
Supplementary Figure 34. Significant copy number variation between strains, broken down into gene categories and gene families. Fold changes denote ratio of the maximum copy number and the minimum copy number between the strains for each gene.



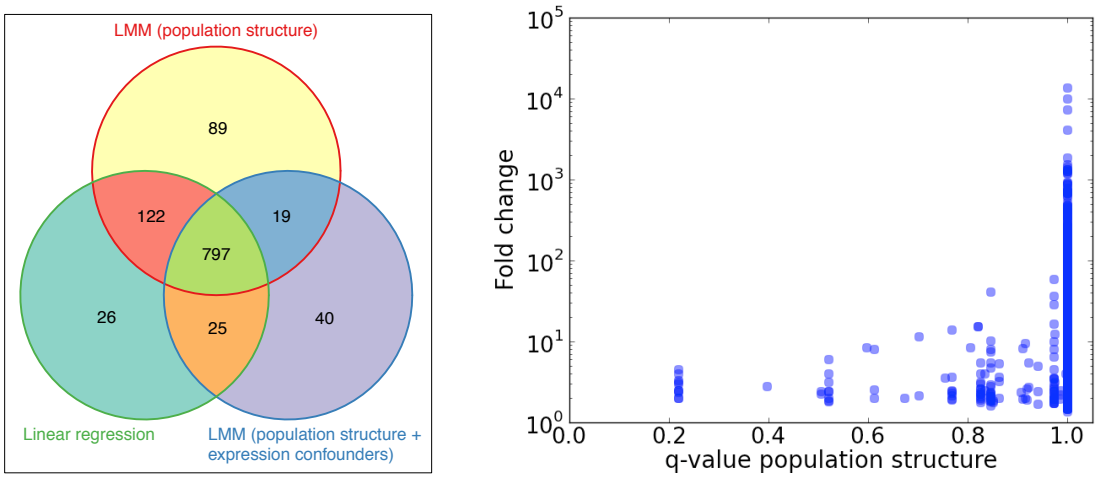
Supplementary Figure 35. Relationship between gene expression variation and copy number differences. Shown are expressed genes, differentially expressed genes and genes with copy number variation, broken down into gene categories and gene families. Overall, the role of copy number variation on gene expression is low as most variable genes are not expressed or not variable in expression.



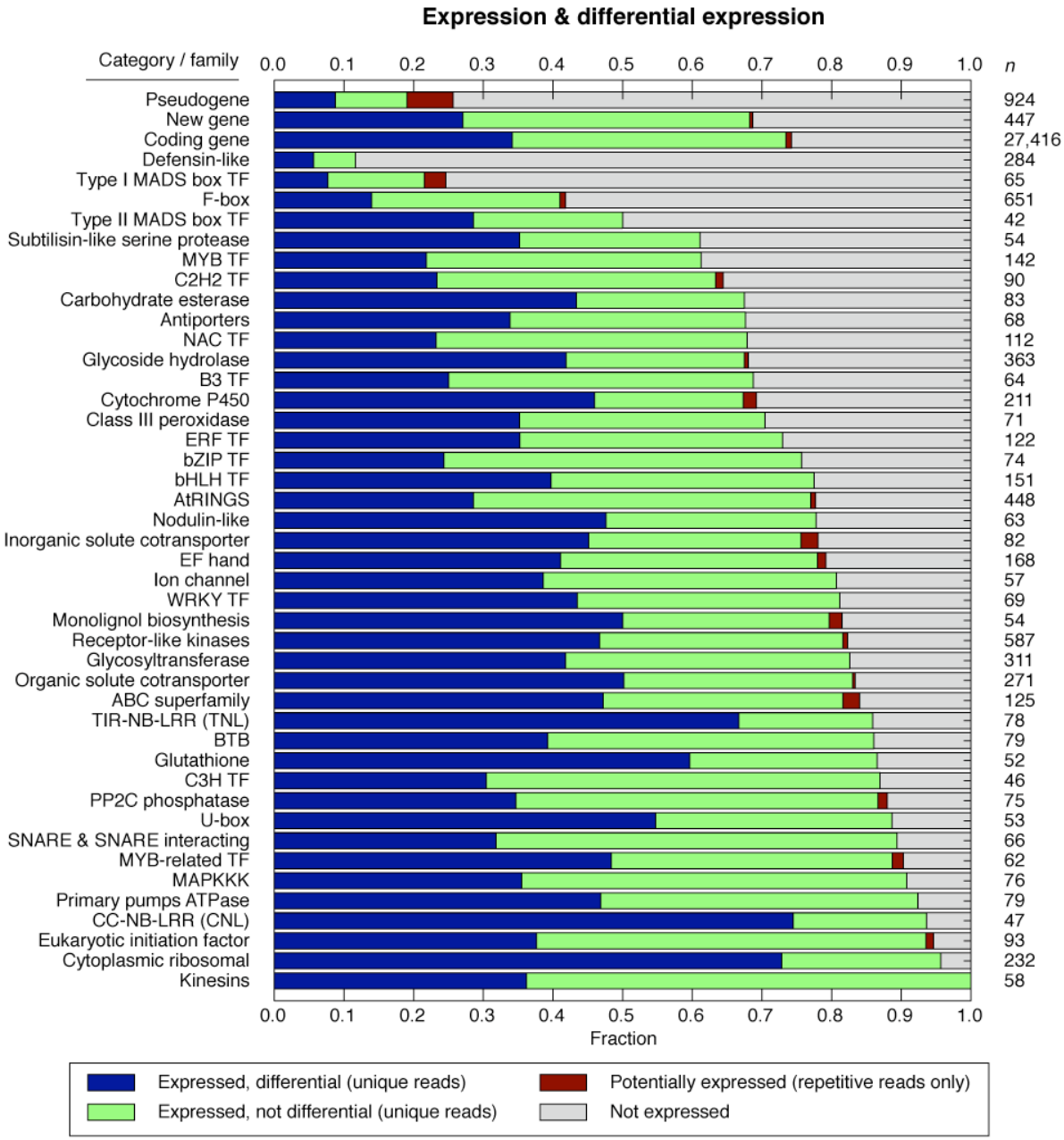
Supplementary Figure 36. Scatter plot of expression fold changes versus corresponding fold change differences in copy number. Left: plotted on a linear scale, right: plotted on logarithmic scale.



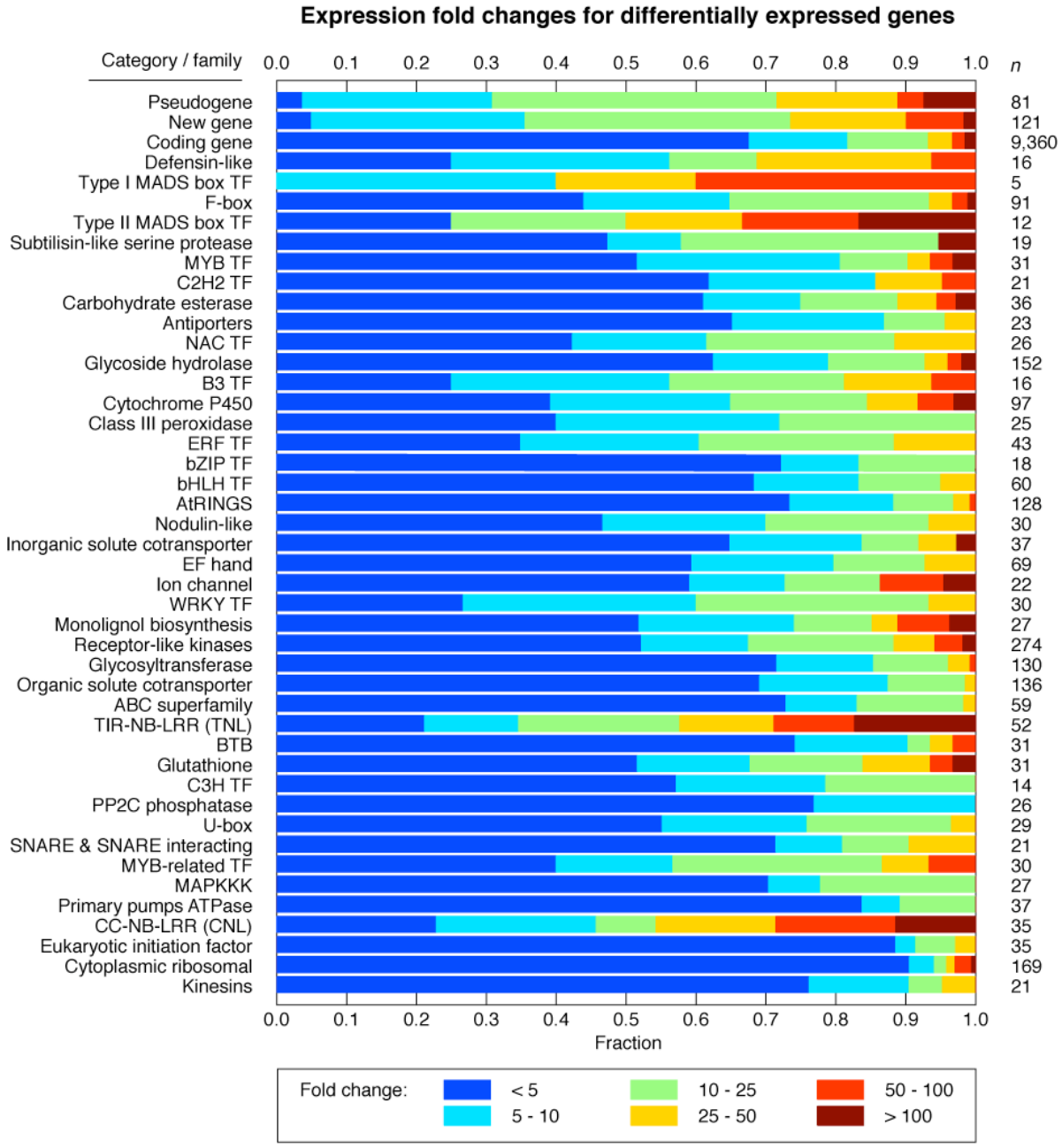
Supplementary Figure 37. Pairwise genetic similarity between accessions, estimated from the empirical covariance of a binarized variants table. Results suggest little populational relatedness with the exception of oy-0 and po-0, which are genetically closely related.



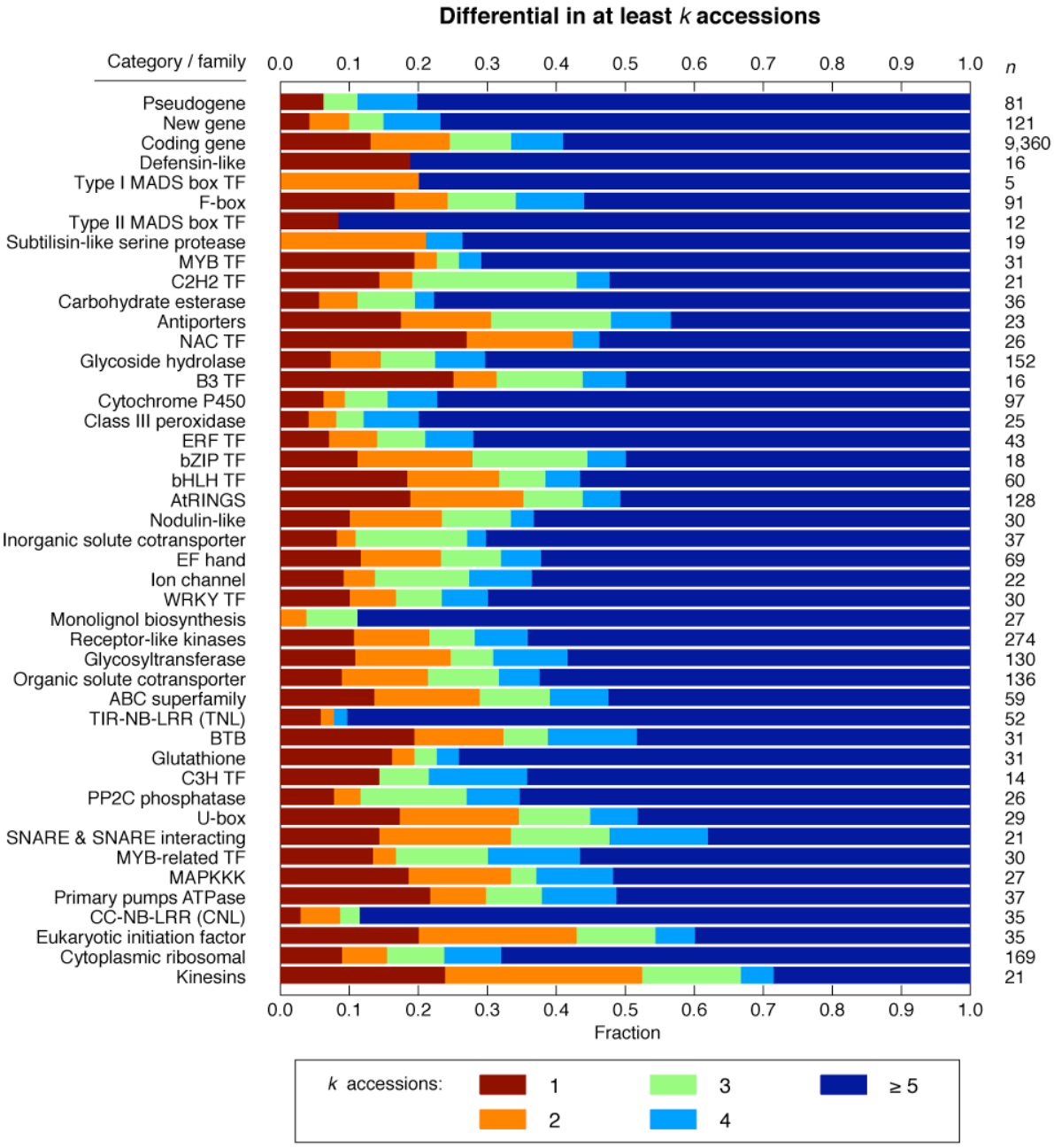
Supplementary Figure 38. Assessment of impact of confounding variation. Left: Overlap of *cis* variant eQTLs for alternative calling methods (FDR 5%). Compared are no confounder correction (Linear regression), correction for population structure (LMM population structure) and correction for population structure and expression confounders (LMM population structure & Expression confounders). Right: Scatter plot of the significance of population effects (q-value=0 significant, q-value=1 insignificant) as a function of log fold change of gene expression variation. Genes that are more significantly regulated by population structure tent to have lower log fold change differences, suggesting little populational regulation of gene expression.



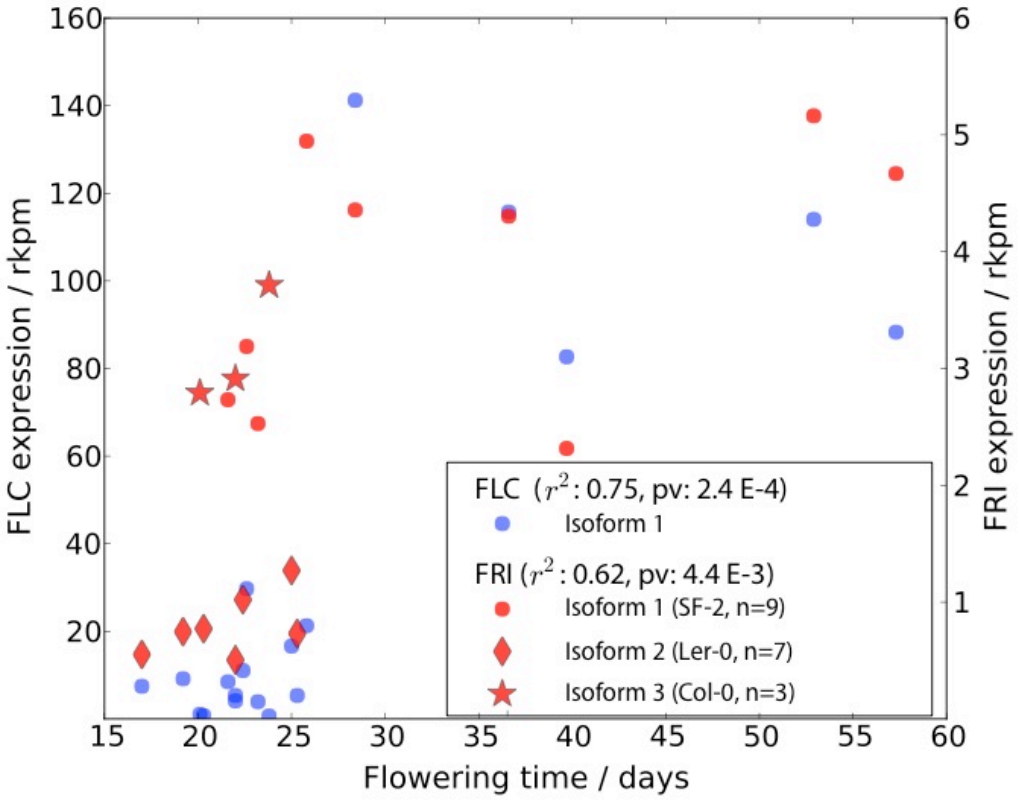
Supplementary Figure 39. Fraction of differentially expressed, expressed (includes differentially expressed), potentially expressed, and genes that are not expressed by category or gene family. Expression status is as assessed across all accessions. Potentially expressed genes are supported solely by RNA-seq reads that map to multiple genes or genomic locations (i.e., repetitively mapping reads). Gene categories and families shown are a superset of those shown in Fig. 4b. TF: transcription factor.



Supplementary Figure 40. Distribution of genes by category or family classified by fold-change. Gene categories and families shown are a superset of those shown in Fig. 4c. Fold-change is as assessed between the lowest and highest across 19 accessions. TF: transcription factor.



Supplementary Figure 41. Fraction of genes by category or family contributing to differential expression at a given frequency (as assessed by differential expression in *k* accessions). Gene categories and families shown are a superset of those shown in Fig. 4d. TF: transcription factor.



Supplementary Figure 42. Scatter plot of the mean flowering time of the 18 accessions and Col-0 versus the expression level of *FLC* and *FRI*.