Supplemental Information

Sequencing of natural strains of *Arabidopsis thaliana* with short reads

Stephan Ossowski, Korbinian Schneeberger, Richard M. Clark,
Christa Lanz, Norman Warthmann, and Detlef Weigel

SUPPLEMENTAL METHODS

Biological material, library preparation, and SBS sequencing

Seed stocks for Col-0, Bur-0, and Tsu-1 were from sibling plants of those used for the resequencing array study (Arabidopsis Biological Resource Center stock numbers CS22681, CS22679, and CS22693) (Clark et al. 2007). Leaves from ~150 three-week old plants were ground in liquid N_2, and genomic DNA was purified from one gram of leaf powder with the DNeasy Plant Maxi Kit (cat. no. 68163; Qiagen Inc., Valencia, CA). From 500 ng of DNA, single-end SBS libraries were prepared with the 1 DNA Sample Kit (cat. no. FC-102-1001; Illumina Inc., San Diego, CA). Briefly, DNA was sheared by nebulization for 6 minutes with N_2 gas at a pressure of 32 psi. End repair of sheared fragments, addition of an A residue to the 3' end of blunted fragments, and ligation of adaptors was according to manufacturer's instructions. The entire adaptor-modified DNA was amplified by PCR for 18 cycles with the supplied PCR primers 1.1 and 1.2 and resolved on a 2% agarose gel (including 400 ng/ml ethidium bromide) run in TAE buffer for 60 minutes at 120 volts. Fragments of between ~120 to 170 bp were excised under illumination from a Dark Reader (Clare Chemical Research, Dolores, CO). DNA was isolated with a Gel Extraction Kit (Qiagen, Hilden, Germany) and quantified with a NanoDrop Spectrophotometer (Peqlab Biotechnologie GmbH, Germany). Diluted DNA was stored at -20˚C as a 10 nM stock in EB buffer (Qiagen) supplemented with 0.1% Tween-20.

Cluster preparation on flow cells and SBS sequencing

Clusters were produced with the Cluster Generation Kit (cat. no. FC-103-1002l; Illumina) on a Cluster Station (cat no. SY-301-2001; Illumina) using the "One_step_full_protocol_v3.0". Libraries were calibrated with dilution series. With an 18 pM solution of denatured DNA, ~30,000 to 35,000 raw clusters per tile were generated on the flow cell surface. Subsequently, we aimed for 35,000 clusters per tile for production sequencing. In total, we used 38, 27, and 16 lanes for Bur-0, Tsu-1, and Col-0, respectively, which includes pilot and production runs. Sequencing was performed on a Genome Analyzer (cat. no. SY-301-1001; Illumina) using the 36 Cycle Solexa Sequencing Kit (cat. no. FC-104-1003; Illumina). The .xml file for recipe "1GSequencing36cycles_v103.xml" was modified to allow extension and imaging of 40 cycles.

Oversampled regions

We analyzed the genome for regions with a 10-fold or higher increase of observed compared to expected coverage and identified between 239 and 262 kb per accession. To join *oversampled* regions in close proximity, we extended each seed until the observed coverage dropped below 5- fold the expected coverage. This identified between 345 to 385 kb of oversampled regions per accession. We found that 318 kb of this was shared between all samples.

Genome coverage estimation

We assessed genome coverage per accession as the read coverage averaged over all non-repetitive positions (Table 1). For Col-0, the reference accession, few sequence differences are present relative to the reference,

and the confounding effect of polymorphism in mapping reads does not complicate the coverage estimate. For Bur-0 and Tsu-1, the coverage estimates reported in Table 1 are underestimates, as some reads cannot be mapped to the reference sequence in polymorphic regions (e.g., Table 5). Employing the number of quality filtered reads and estimated genome coverage for Col-0, and extrapolating with the number of quality filtered reads for Bur-0 and Tsu-1, the estimated genome coverages for Bur-0 and Tsu-1 are 25.1 and 19.4, respectively. Uncertainties in read quality between runs potentially affect this extrapolation, as would any large-scale differences in genome size between accessions (e.g., for centromeric content). Nevertheless, these calculations suggest that the coverage estimates reported in Table 1 for Bur-0 and Tsu-1 only marginally underestimate the true coverage depths.

Performance of SHORE by sequence type

To assess SHORE performance for SNP predictions by sequence type, we classified each of the 2,896 SNPs in the Bur-0 dideoxy data as either coding or noncoding according to the TAIR7 *A. thaliana* genome annotation available at http://www.arabidopsis.org/. Specificity and sensitivity were calculated as for all SNPs (cf. Table S1).

SNP prediction with MAQ

We downloaded the Mapping and Assembly with Qualities (MAQ) software package from http://maq.sourceforge.net/. We converted the Illumina raw quality (*prb*) values to the FASTQ format that is defined by the Sanger Institute and recommended for MAQ (http://maq.wiki.sourceforge.net/FASTQ+Format). We employed the parameter sets presented in Table S3. For the SHORE, MAQ-unfiltered, and MAQ-filtered predictions, sensitivity and specificity were calculated as presented in Table S1.

Overlap of SBS and resequencing array PR predictions

To assess the quality of our SBS PR predictions, we examined overlap to resequencing array PR predictions generated in another study, but with the same accessions (Zeller et al. 2008). The overlap was measured as the position-wise percentage of SBS PR positions that had also been annotated as polymorphic in the resequencing array predictions. The overlap was 75% and 72% for Bur-0 and Tsu-1, respectively. To assess the likelihood of obtaining these overlap values by chance, for each accession we shifted the SBS PR predictions 1,000 times by a random offset, recalculating the overlap to the array PR predictions each time. The resulting distribution is shown in Figure S8 along with the observed values.

Validation of targeted assemblies by PCR and dideoxy sequencing

To assess the quality of *de novo* assemblies (Fig. 3), we amplified 192 respective regions from Bur-0 genomic DNA by PCR and dideoxy sequenced the resulting products. Primers for PCR were selected with Primer3 (Rozen and Skaletsky 2000) to be within the flanking (anchoring) 100 bp on either side of targeted indels. The annealing temperature thresholds were 59 and 62°C. Where primer design was successful, we classified the predicted PCR products as either short (≤ 250 bp) or long (> 250 bp). Within the short and long classes, we randomly selected 96 contigs each for evaluation. Our rationale for selecting a long class was that long contigs

are more challenging for assembly (e.g., long inserted sequences), and a completely random selection would have been biased towards smaller indels.

For validation attempts, we used the identical Bur-0 DNA sample as employed for SBS library construction (see Methods). We employed touchdown PCR as follows: 95°C for 2 min, followed by 1 cycle of 95°C for 15 sec, 63°C for 30 sec, and 72°C for 1.5 min, followed by 10 additional cycles as before but with decreasing annealing temperatures of 0.5°C per cycle, followed by 32 cycles of 95°C for 15 sec, 58°C for 30 sec, and 72°C for 1.5 min, followed by 72°C for 3 min. PCR-reactions included 1 µl of DNA at 1 ng/µl and 2 µl each of forward and reverse primers at 5 µM in 15 µl reactions with Pfu DNA polymerase (MBI Fermentas) according to manufacturer's instructions. The presence of PCR product was assessed by agarose gel electrophoresis. For the few cases where products were absent, another PCR amplification attempt was performed with the same touchdown PCR program, template, and primers, but using Extaq™ polymerase (TaKaRa). The PCR products were diluted 1:15 (short contigs) or 1:10 (long contigs) and analyzed by dideoxy-sequencing on a 3730XL capillary sequencer (Applied Biosystems). Forward and reverse reads were generated for each PCR product with the same primers used for amplification.

Of 192 PCR attempts (short and long), 188 were successful. The resulting sequence trace files were processed with the Staden package (Staden et al. 2000). Reads were quality clipped with *pregap4*, forward and reverse reads for each product were assembled using the *gap4* assembler, and resulting assemblies were manually curated. The sequences obtained by dideoxy-sequencing of all 188 PCR products were identical to the contigs predicted by targeted *de novo* assembly.

Expected coverage calculation

Apart from stochastic effects and sequence divergence, the expected coverage at a position given an overall sequence coverage is affected by (i) repeat content (reads in repeats cross map, increasing coverage) and by (ii) biases in the sequencing method (e.g., effects of GC content; Supplemental Figs. S3 and S4). To approximate the expected coverage at a position *p*, we first took the full set of 36-mers generated from the reference sequence, and aligned these to the genome with four Hamming distances (0 to 3). All mappings were recorded, and from this we calculated, for each position, four repeat content values rc_{p_h} defined as

$$rc_{p_h} = \max(0,(oc_{p_h} - 36)/36),$$

where oc_{p_h} is the observed coverage at position *p* applying Hamming distance *h* in the experiment. Second, as we observed that GC content was strongly and positively associated with per-position coverage, a bias in the method, we further defined the average GC content gc_p for each position *p* as the percentage GC content of all unambiguous bases in a window of 101 bp centered on *p*. From the experimental read mappings for each sample, we calculated the observed average coverage ac_{gc} for all occurring GC contents gc_p. Combining these three values, we estimated by sample the expected per-position coverage ec_p as

$$ec_p = \left\{d + (d \times rs_p)\right\} \times (ac_{gc_p} / 36),$$

where *d* is the sample-dependent sequencing depth, and where rs_p is the repeat scaling factor calculated as

$$rs_p = rc_{p_{\lfloor am \rfloor}} + \left\{(rc_{p_{\lceil am \rceil}} - rc_{p_{\lfloor am \rfloor}}) \times (am - \lfloor am \rfloor)\right\},$$

where am is the average number of mismatches of all reads overlapping *p* in the observed data.

Detection of duplicated regions and comparison to hybridization data

We used both coverage and sequence criteria to identify unique regions of the reference genome sequence that are present in approximately 2-3 copies in Bur-0 and Tsu-1. First, we identified all contiguous positions of at least 250 in length in non-repetitive regions for which the observed per position read coverage was between 1.2 to 3 fold the accession corrected expected coverage, ec_p. Second, and subject to the quality rules employed for base calling, we identified positions for which two bases were supported with relative frequencies each between 20 to 80%. These copy variable positions, or CVPs, are expected to result where similar but not identical sequences map to the same single copy region in the reference sequence. Heterozygosity would also produce a similar pattern; however, this is unlikely to have affected assignment of CVPs in our study as the accessions we used were highly homozygous as established by dideoxy sequencing at more than 1,200 loci spaced at regular intervals throughout the genome (Nordborg et al. 2005). We classified as duplicated regions those which met the elevated coverage criteria and that had at least one CVP.

To assess the quality of duplicated region predictions, we employed the resequencing array data (Clark et al. 2007). Unfortunately, resequencing microarray data is highly noisy, confounding use of quantitative intensity measures to directly detect additional copies of a sequence. We therefore assessed the quality of CVP predictions that lie within the predicted duplicated regions. Here, the relative intensities among probe sets at individual tiled positions corresponding to CVPs are expected to reflect copy number variation (i.e., probes querying two bases will hybridize at similar intensities if a region is duplicated). To remove the high fraction of positions for which the array data was not informative, we filtered positions for which, in the array data, the sum of the first and second most intense probes was 600 or more in both the accession with the CVP and in Col-0 on either the forward or reverse tiled strands. Where both strands in a sample met this criterion, data from one of the strands was randomly chosen. Further, we only considered those CVPs for which, from the aligned read data, the upstream and downstream 12 bp were predicted to be identical to Col-0. This requirement assured that the probe hybridization data was not affected by a polymorphism other than the CPV itself; the resequencing array probes were 25 bp in length, and off-center polymorphisms strongly inhibit hybridization (Clark et al. 2007; Zeller et al. 2008). For positions that met these criteria, we calculated *chastity* values corresponding to the intensity of maximally hybridizing probe divided by the intensities of the first and second most intensely hybridizing probes (a double peak therefore gives a value of ~0.5, while a single peak gives a higher value). With these selection criteria, we identified a set of 2042 CVPs in Bur-0 and 2172 CVPs in Tsu-1 for which

resequencing array data was available, including for Col-0. As assessed with the chastity values, double peaks were readily apparent at CVPs in the divergent accessions as compared to Col-0 (Fig. 4 and Supplemental Fig. S9; compare patterns of hybridization at the CVP positions shaded in yellow to adjacent positions).

Effects on genes

We assessed the effects of SNPs, 1-3 bp indels, and PRs of at least 100 bp (Tables 4 and 5) on coding (CDS) gene models from the TAIR7 annotation. For SNPs and PRs, annotation of polymorphisms was as described by Clark et al. (2007), and the method was extended for small indels (e.g., to identify frameshift changes). We also examined the distribution of major-effect changes for a subset of large gene families previously examined with the resequencing array data (Clark et al. 2007; see Supplemental Tables S7 and S8 therein). Major-effect changes are as defined in Supplemental Table S10, and genes were classified by family using gene descriptions in the TAIR7 release with the search terms in the first columns of Supplemental Tables S11 and S12.

SUPPLEMENTAL REFERENCES

Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T.T., Fu, G., Hinds, D.A. et al. 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis* thaliana. *Science* **317:** 338-342.

Dong, Q., Schlueter, S.D., and Brendel, V. 2004. PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res* **32:** D354-359.

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R. et al. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol*. **3:** e196.

Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol*. *Biol*. **132:** 365-386.

Staden, R., Beal, K., and Bonfield, J. 2000. The Staden package, 1998. *Methods Mol. Biol.* **132:** 115-130.

Zeller, G., Clark, R.M., Schneeberger, K., Bohlen, A., Weigel, D., and Rätsch, G. 2008. Detecting polymorphic regions in the *Arabidopsis thaliana* genome with resequencing microarrays. *Genome Res.* **18**: 918-929.

SUPPLEMENTAL TABLES

Table S1. Results from performance evaluations for SHORE and MAQ as assessed against 2,806 known SNPs in the nonrepetitive fraction of the Bur-0 dideoxy data.

SNPs by ascertainment method	Counts and performance calculations	Evaluation metric
Dideoxy sequenced SNPs	2,806	TP + FN
SHORE total predictions	2,373	TP + FP
MAQ-unfiltered total predictions	2,800	TP + FP
MAQ-filtered total predictions	1,911	TP + FP
SHORE correct SNPs	2,357	TP
MAQ-unfiltered correct SNPs	2,448	TP
MAQ-filtered correct SNPs	1,855	TP
SHORE *Specificity*	99.3%	TP / (TP + FP)
SHORE *Sensitivity*	84.0%	TP / (TP + FN)
MAQ-unfiltered *Specificity*	87.4%	TP / (TP + FP)
MAQ-unfiltered *Sensitivity*	87.2%	TP / (TP + FN)
MAQ-filtered *Specificity*	97.1%	TP / (TP + FP)
MAQ-filtered *Sensitivity*	66.1%	TP / (TP + FN)

SHORE predictions were generated as detailed in the Methods (see also Table S2), and the MAQ-unfiltered and MAQ-filtered predictions were generated as described in Supplemental Methods (see also Table S3). TP, true positive; FP, false positive; FN, false negative.

Table S2. Parameters used for base calling for genome-wide predictions.

Accession	Prediction type	Min. read support[a]	Min. core reads	Call %	Max. allowed coverage[b]
Bur-0	Reference	3	1	80	120
	SNP	3	1	80	120
	1-3 bp deletion	3	1	67	120
	1-3 bp insertion	3	1	67	120
Tsu-1	Reference	3	1	80	90
	SNP	3	1	80	90
	1-3 bp deletion	3	1	67	90
	1-3 bp insertion	3	1	67	90
Col-0	Reference	3	1	80	80
	SNP	3	1	80	80
	1-3 bp deletion	3	1	67	80
	1-3 bp insertion	3	1	67	80

[a]For Bur-0 subsampling, the minimum read support was varied from 1 to 6 with the other parameters as given (cf. Fig. 2).

[b]Coverage was selected to be at most 5 times the genome-wide coverage depth for which a given accession was sequenced (Table 1).

Note: Repetitive and/or oversampled positions, organellar repeats, and bases ambiguous in the reference sequence were excluded from predictions.

Table S3. Preprocessing, mapping, SNP detection and filtering for prediction with MAQ.

Preprocessing:

maq fasta2bfa reference.fa reference.bfa

maq fastq2bfq reads.fastq reads.bfq

Mapping and SNP detection:

maq map –n 4 –m 0.005 –C 513 out.aln.map reference.bfa reads.bfq 2> out.map.log

-n 4	Comparable number of allowed mismatched
-m 0.005	Polymorphism rate is roughly 1 SNP in 200 bp
-C 513	Infinite number of mappings allowed

maq mapmerge all.out.aln.map out.aln.map.1 out.aln.map.2

maq indelsoa reference.bfa all.out.aln.map > out.indel.soa

maq assemble –t 0.5 –r 0 –m 4 out.cns reference.bfq all.out.aln.map 2> out.cns.log

-t 0.5	Reduced to account for the not existing heterozygotes
-r 0	No heterozygote sites expected
-m 4	Comparable number of mismatches

maq cns2snp out.cns > out.snp

Filtering

maq.pl SNPfilter –D 125 –w 4 out.snp > filtered.snp

-D 125	Comparable maximum coverage criteria for oversampled regions
-w 4	Mask SNPs nearby potential indels, set to the number of bp defining the non-core region.

Parameters that were changed from the default settings are indicated with explanations/justifications (right). MAQ-unfiltered predictions were produced as described above but excluding the “Filtering” step indicated (bottom). The MAQ-filtered predictions included all steps.

Table S4. Concordance of Col-0 predictions with SBS data from Bur-0 and Tsu-1 and with Col-0 resequencing array data.

Prediction type and status		Agreement with Col-0 base calls from resequencing arrays[a]			Concordance with array data
	N	*n*	Yes	No	
Col-0 substitutions:					
Polymorphic in Bur-0 and/or Tsu-1[b]	963	408	404	4	99.0%
Reference in Bur-0 and/or Tsu-1[c]	74	30	22	8	73.3%
Other[d]	135	34	30	4	88.2%
Col-0 insertions:					
Polymorphic in Bur-0 and/or Tsu-1[b]	721	NA	NA	NA	NA
Reference in Bur-0 and/or Tsu-1[c]	NA	NA	NA	NA	NA
Other[d]	80	NA	NA	NA	NA
Col-0 deletions:					
Polymorphic in Bur-0 and/or Tsu-1[b]	451	NA	NA	NA	NA
Reference in Bur-0 and/or Tsu-1[c]	8	NA	NA	NA	NA
Other[d]	27	NA	NA	NA	NA

[a]Comparisons are to base calls for Col-0 from whole-genome resequencing microarrays using the identical Col-0 accession (stock number CS22681; Clark et al. 2007). Comparisons are only reported where base calls from the forward and reverse strand resequencing array probe quartets agreed and the mean quality score was greater than or equal to 15 (see Clark et al. 2007).

[b]Where both Bur-0 and Tsu-1 agreed with the Col-0 prediction, or where one agreed and the other was ambiguous ('N', or missing data).

[c]Where both Bur-0 and Tsu-1 agree with the reference sequence, or where one is reference and the other was ambiguous ('N', or missing data).

[d]Where data was missing from both Bur-0 and Tsu-1, or, e.g., where more than two sequence predictions were made across the accession trio (e.g., triallelic predictions).

NA, not applicable.

Table S5. SNPs, small indels, and reference base calls in moderately repetitive regions.

Accession(s)	SNPs	Indels				Ref. base calls (%)[b]
		1 bp	2 bp	3 bp	All	
Col-0	47	13	0	0	13	61.24
Bur-0[a]	20,795	1,081	107	21	1,209	46.27
Tsu-1[a]	18,047	941	69	16	1,026	43.07
Non-redundant[a]	35,610	1,866	162	31	2,056	NA

Table S6. Comparison of Col-0 PRs to known mis-assemblies, resequencing hybridization data, and the Bur-0 and Tsu-1 PR predictions.

Chr	Start	End	PRs	Included bp	Known mis-assembly[a]	Array hybridization quality for Col-0	Bur-0 and Tsu-1 PRs
1	14,511,399	14,866,522	31	3.350	No	4 long regions with low quality, highly repetitive in between	overlapping
1	15,438,047	15,438,687	1	641	No	low	equal
2	3,624,495	3,624,982	1	488	Vector	low	equal
3	1	104	1	104	Vector	low	equal
3	13,748,172	13,748,766	6	291	Vector	low	overlapping
3	13,754,115	13,760,800	2	67	Vector	low	equal
5	11,202,838	11,246,080	40	15,541	Vector	14 regions with low quality, highly repetitive in between	PRs of >50 equal & including vector contamination
5	12,846,176	12,851,904	6	1,670	No	3 long regions with low quality, highly repetitive in between	equal

[a]Defined as nearby or overlapping sequences with high sequence similarity to cloning vectors (Dong et al. 2004).

Table S7. Validation of targeted assemblies as assessed with PCR and dideoxy sequencing.

ID[a]	Validation type	Forward, reverse primers (5' to 3') for validation attempt	Amplified and sequenced?	Sequence identical to prediction?
LCR_71	short	ACTGATCTGGAAGAATGATGCAGT, GTCACTGTAATGACATCGGGTTT	yes	yes
LCR_77	short	ATCTCATGTTGTGCACTCAGTTGT, CTCTTCAAGGGATACACCAAAGAT	yes	yes
LCR_137	long	GCGAGAAGCTCGGAACAAA, ACTATTCTCCGCCTCTAAATAGTCTTG	yes	yes
LCR_164	long	GGACAAAGAAGGATTTGGGTATC, TCGATGAGTTCATTCTTGCTG	yes	yes
LCR_214	long	AATTTTGAGATCTACGTTAAGCCATC, GAAATTCGATTTGCGACCAC	yes	yes
LCR_229	long	ACTCTGGCTCTGCCTCTGAATA, GAAATGTTTCTGTAATCCAATCCA	yes	yes
LCR_331	long	TCTGAGAAATGGTGTCGCATAA, CCTAAACTAACTACAACATTAGGTCTCG	yes	yes
LCR_339	long	GCTCAACTCTCCCTAACCACCTA, TATTGTCAGCATAAGCTAAGAGTGAGT	yes	yes
LCR_349	short	ATACAGGAGCGCTTGCCAATA, TAGTATCATGAACCAACCATCTCTACA	yes	yes
LCR_373	long	AAAATTTGCTTAGCCGTTTCG, AGACACCACCACACTGGTCACT	yes	yes
LCR_377	long	ATCTTCTTATTCAAATGGGTCGAG, GCGATACTTGAGGAAAATGTGAA	yes	yes
LCR_455	short	AACTGGGAAAACGCGTGAA, CATTGATTCCTTCAAGCCAAG	yes	yes
LCR_461	short	CAGATACCATATGTGTCAAGCTTACTTT, TCTAGATGGAAGGTCGGGATAC	yes	yes
LCR_478	long	CAACTAATGCTTCCAAGTTAGCTTTT, CATTTTCACCTCCAGTAACAAAGA	yes	yes
LCR_536	long	GTATTTTGTTGACAATAGTTCGTCCAT, ACAACGTTCTTTTCACGTCTGTC	yes	yes
LCR_552	long	CCAATCAACGTGGTATATTTTCG, GTTTCTTATGCTTTGTGTTTCAAGTG	yes	yes
LCR_564	short	GCAAATTTACGAAGTTAGCCCAAT, GGAGGTGTATGAACCGACAAAT	yes	yes
LCR_611	short	ACATTGCTTTAGTTATAGCCTGGATT, AGCCGTGCAGAGCTTATCA	yes	yes
LCR_637	long	CTCCTTCACTCACACCATCAAA, TTAGGACCTTCAACCTAAGAACAAA	yes	yes
LCR_669	short	GTTAAACTACCGAGTGTATCCTTCTCC, TTTACGTACATGGAATGTTCTTGATG	yes	yes
LCR_688	long	GAGATAGTTATGTTTCATCCTTTGTCC, CTGTGACCACTTTCCTCTCTAGC	yes	yes
LCR_690	long	GATCTTTGCTGGGAAGACCA, TACTATCGACAAGAATTCATCCATAAG	yes	yes
LCR_739	long	TACCCGGCCCAGTGTATTT, AGTCCATGACTGATGCCAAGTAT	yes	yes
LCR_787	long	AATGCATCCCTCAGGCAAC, TTTTATGGGCTTCTCTATTGTAAAAGG	yes	yes
LCR_874	long	TCAGGTTGAGACAATATCACACG,	yes	yes

		GTAATCTAGTATTGAGAGATGGGCCTAA		
LCR_882	short	CCTCCACAGGCTTCCTCAAT, ATCTTGGGATCGAGCCTACAAC	yes	yes
LCR_922	long	AGTCAATAGTGAGTTGTTTGCTATCC, TGCTACAAAACCTTGGGAACTAC	yes	yes
LCR_929	long	ATGTGCGAGTACATACCTAAAACAGTA, AGATGCTCCTGCACCATTTG	yes	yes
LCR_973	short	ATTTTGTTCAGTCCATATGATTCG, CACGTGAACACATCGTTTTATG	yes	yes
LCR_976	long	GAGGCATACGAGCGCTTAAT, TTGTATGGGATTGACTTTACTTCTCC	yes	yes
LCR_1007	long	GTCTCCTTTTATCCTGTAATTCAAACA, AAAACCAGGCCAACATCTCAA	yes	yes
LCR_1013	long	GAAATTACGCGGAAGACGATG, TGCTATGTCGGCAAAGAGG	yes	yes
LCR_1066	short	TGACTAACCAAAGAGTGGATGC, GTGTTAGATAGAGCTTAGGAAAGACAAA	yes	yes
LCR_1080	short	AGCTTCTTCTCTCATTGTCTCTCCT, TTTGTTTCCCAGTTTGAGTGGA	yes	yes
LCR_1124	long	GGGTGAATGGAGCTCACAGA, CATGTCGATGCCTGCTCA	yes	yes
LCR_1138	long	AGCTTTTGTCTACTTTTCTTCTTCTCTG, TTGGTGCTTTTGTTTTGCTTG	yes	yes
LCR_1139	long	CGATTAGGTCAAGGAAGTAACCTTT, GTGGCACGTGAGGTTTTAGATT	yes	yes
LCR_1150	short	CTTGCTCAAGAAATAAACCCTAGAAC, TCCATAAACTCTTTGGTGGATCA	yes	yes
LCR_1159	long	TCGCAAGCCGTTCTTTTACT, AAATTGGATGAATGTTACGATGAC	yes	yes
LCR_1212	long	AAACAGAAATACCTGACCTTCCTACTC, TGGCTCGTGAAATCGAAGT	yes	yes
LCR_1268	short	AACTCCCGGAATGTGGCTTA, GCCTTAGCATTGGATTTAGATGAG	yes	yes
LCR_1284	short	AGCAGCTTCACAGGATACCTTTT, CACTATATAATTGTTGACGAACCGAAT	yes	yes
LCR_1329	long	TTCTCCACAGATTACTTGGCGTAG, TTTATGTGTCCCTACTAACTCATGGAT	yes	yes
LCR_1359	long	ACCTGGTGCTCGGATTGATT, TCCTCGACATCCTGCAACA	yes	yes
LCR_1366	long	GAGATCGTGACACTCATCACAATTA, TAAATAACTGGACTGAAAGAAGTGTGTC	yes	yes
LCR_1419	short	CACAGCATTAACACAACCCTCA, AGTAGCTGAATTTGATTTGCAGGA	yes	yes
LCR_1483	long	CTTCTGGTTTTCAACAATTGCAT, GATTGAGATAACAAAGTGTCCACCA	yes	yes
LCR_1487	short	ATTAATCTATTGATGGTGTCCATTGTC, AAGGGTTTTCTACAACTATATGGAATG	yes	yes
LCR_1493	short	ATTTATCTTCCCCAACACTGCTTT, TGGCTTCCTCTAAGACATGTACAAAC	yes	yes
LCR_1517	long	TTTTGTCTCTGCTTGTTTCTTAATCTC, GGTCAGTCTCTTTGTTTCAAGTTGTC	yes	yes
LCR_1581	long	TGTGACACCTTTTCTTGTATTCAGAT, TTTATGCTCTTCATTGACATTGCT	yes	yes
LCR_1604	long	ACCTTAGGAAACAATGTAACTATTCCAG, TCCATCGGTGTTGGATTGAT	yes	yes
LCR_1619	long	AGTATGGAGGAGACTCGTTGATTG, TTCATCTTCCCTAGCTGAACTCC	yes	yes

LCR_1656	short	CAACGCATGCAGTACCTTAAAA, TCTTTGAATCCTTGTGTTGTATCG	yes	yes
LCR_1808	short	AAAACTTAGTACCTCTCGAATGCTTTA, CTATTTCCCTTGCCGAAGATG	yes	yes
LCR_1855	long	TAGACCCAGTCAACTAACAGAACAGA, AGACGTTGGGGAATGTTAAAGAG	yes	yes
LCR_1921	long	CTAAACCAAACCAATCGCAAAAC, TATGTGTGAGTGTGTCTTGTAAGAGTGT	yes	yes
LCR_1949	short	AGATACTTCTTACTTACCCCACAAAGC, GGTAAATTATCCTGCCAAGAAAGG	yes	yes
LCR_1963	long	CTCATCCTCTCCGTCATCTTCT, CTAGGGTTCCAGTTAGAGCTGATT	yes	yes
LCR_1988	long	AGCTGGCCCTAGGAGACAAA, TTTGACTGTCTCTCTTCTTCTCACAAT	yes	yes
LCR_2011	short	AGTAGCTAATTCCCACAATATGACACC, CTGAGTGGATGGCTCCAGAA	yes	yes
LCR_2016	long	GCTAGCGTGTTCTGAAACCTGT, TCAAGCTTAGAAGGCAGTTCTTG	yes	yes
LCR_2024	short	ACAACAGGTTCATTAACTACAAGAGTCC, GGTTTGCGCTTTCTTTCTTGTT	yes	yes
LCR_2051	long	GAATTGTGGCGCATGTGTAG, CCACGAGGTATGTGCCAATC	yes	yes
LCR_2071	long	GTCTTGATGCTGTCGTCTTCC, AGGTCCTCACAAAGTAGAGCAGA	yes	yes
LCR_2102	long	CACCACATCTTGTTGCCATTTA, TTACAGGTGAGGAATGGAACTG	yes	yes
LCR_2111	short	CGCATGAAACCGTGCATT, TAAACGGTATTGAAATTGTAGGTTGG	yes	yes
LCR_2117	short	GCGATTGCGTTGATCTGC, TACCGCTTACTTCAATGTTACGCTAC	yes	yes
LCR_2161	long	CTTCCTTGGCGATTTCTGGTA, GTTGATGCTGAGAAGCGGTTA	yes	yes
LCR_2181	long	AGGTACGTTTCTTTGTTCCTTTTG, TATAACAGAAGTGAACTCTTTAGGAAGC	yes	yes
LCR_2290	long	CGCTGCCATGGGATAAAA, AGATCCTGATGTTACTACCAGTGCTT	yes	yes
LCR_2298	long	CCTGTTTCCGATGTACCTCCT, AACGTTGTTCTTCGCCAATC	yes	yes
LCR_2360	short	TGAGATAGAGAGAAACAAAGAAGATGA, GACCACAAAAGTTGAGTGAGTGG	yes	yes
LCR_2369	long	AGGTGGATCTTAAAATAAGGACATGC, TCAAAGGTGAATTTGGAAGCAA	yes	yes
LCR_2408	long	GGTCTCGCTGATGCTTTTGT, TGTATATAGAAAGACACAGGCTAACTCG	yes	yes
LCR_2433	long	ACGGGCATATCTTCCAGAAACT, GTAAATATGTCGAGACGAGATGTGA	yes	yes
LCR_2435	short	AGAATCTTCAACCACAAGTTACATAGAG, AAGAGTGTCACATGATGAAAGACCTAA	yes	yes
LCR_2459	long	TGCCTCAAGCCCAGCATA, CCGGCGTTGGATATGTCTAC	yes	yes
LCR_2506	short	GAGATTGCTATATCCATTACCCTTAGAT, TCACCTCCTCCATCTCCAA	yes	yes
LCR_2523	short	ACACCGTAAGAAGCGTGCAG, ACAAAATACCCCTTACGTCACTTC	yes	yes
LCR_2532	long	TTTGTGATTGGTTTGGTTTGC, TATCTTTGGATAAAGTTGAGGTTTCC	yes	yes
LCR_2556	long	TATCCCGGTACAGCCATACCTT,	yes	yes

		TGGCTAAGTGAAGTTTTGGTATTGTC		
LCR_2564	long	AGGCTGGTCTGATTCGTTGTA, TCTAAACTCAACGCCAAGAAGAA	yes	yes
LCR_2594	long	CTGTCAGAAGATGCAGTGGTAATG, GGGCTTGGAGAAGTCAAACTC	yes	yes
LCR_2599	long	TGAGCAAAGCGGAGGATG, AATGGGCTATTGGGCTTTATGA	yes	yes
LCR_2685	short	GGGCCCATACAATTGGTCTT, TTTTCCCAGAAGATAAGTAAATTAGGG	yes	yes
LCR_2691	long	TAACCCCTAATCACCATGAAGC, TCCCTTCATCGCCTTATTCAC	yes	yes
LCR_2705	long	TGATGCTGAGGTTTCTTCTTCA, CTTTACGAGACAAGAGAACATGGA	yes	yes
LCR_2709	short	TTGCTCTTACTGGACCAAACC, AGCCCAATCAAAACTAGATTCACT	yes	yes
LCR_2717	long	CTGACCAGTTTGAGGCCAAG, GAAACTTCAATCCCCAAAACC	yes	yes
LCR_2727	short	CATGCGCAACGAAGAAGC, CTCTCTGCAGTGGCTCTGACTA	yes	yes
LCR_2736	long	AGTCAACATTCACAAGCCCATA, CACATCGCGCTGAAGAGTT	yes	yes
LCR_2750	short	TCCGGTGTTCCTCCTCTCTT, TGTCAAACATAGAGTGTGAGAGTGAGAT	yes	yes
LCR_2768	long	ATCTGTTGAAGATAAGGATTAACGATG, GGTGGTCCTTGGACAATAGTATGA	yes	yes
LCR_2817	short	GGTATTTTCATGATGTTGTTGCTT, TGATGTGTGTGAAGGAAACTGC	yes	yes
LCR_2883	short	GAGCAGAAAATCGACGCTCAT, GATTCTCCCTTTGCAGCATAAAT	yes	yes
LCR_2968	short	CTCCTGAGCCCTGCAACTAA, TGTCACTCAATATCATCAATTACCG	yes	yes
LCR_2990	long	AAAGGTATCTCTAGACACTGCACCA, GATTCCACAATGAAGCGTTTG	yes	yes
LCR_3006	short	TTTTAAGTCATCACCAAATCATGG, TATAATTTTCAAGTGGAGGAAGCA	yes	yes
LCR_3150	long	TGATAGTGATAAAGACGGTGGTAGTT, GGTCACACACGGTTTCAAAGAT	yes	yes
LCR_3169	long	TTCATCTCCTCCAGCATCTTTC, GAAAACGTCAATTTCACTCTCTTTG	yes	yes
LCR_3233	short	GGCCCTGACTTTGAGGCTAT, ACTACTCTGTATTATAGCTGCTGCAAGG	yes	yes
LCR_3252	short	TAAACAATCGTAATTGCTTGATGG, ACAAGAAAATCTCACACAGATAACTAGC	yes	yes
LCR_3258	long	ATCCTGGTATTACAGAGGCAGACC, CTTACCTTAAGACACAGAGCAACCA	yes	yes
LCR_3268	long	CTGGTTTTCTTCCACCAACTCA, TAGTCAACATTGGAACTCAGGACA	yes	yes
LCR_3269	short	CTTAGCAAGAAGATCAGCATTATCG, AAGTTAGTGGAAATAGATGTGATGGTG	yes	yes
LCR_3271	long	TGTTGACTTTAAGCTTCTGTTGTTTAG, CGACGCATGAATAGGGATG	no	NA
LCR_3298	short	ATCTTTTCGATGTCCGCTTC, CTCATTATTCATTACATAGTCCACACG	yes	yes
LCR_3308	long	AGCTTCTAACTTCCTCACACCGTA, GAATTAGTAGAGACTTGGTTTGGAGAAT	yes	yes
LCR_3341	long	GACTCGACTCTTGCTCGTACTTG, CAAGCTAAAGCCTTGGAGCAC	yes	yes

LCR_3369	long	AGGACTTATCCTGCTCAAAGACG, CTTACGCAAGATGGTATGCTCTC	yes	yes
LCR_3400	long	CACATTCTTCTCTGCCAGGTTT, CTCACCGTCAGGAGGAGTTAGT	yes	yes
LCR_3417	short	TCACTAATACACTAGTAACCTCCAATCC, AGACGGCACCGTTTTCTGT	yes	yes
LCR_3427	long	AAATCCGTCTCTCGGCTAACTC, CACTTAAGATCTGTTACCCTGCTATG	yes	yes
LCR_3468	long	CAGTGTCTTGCTTTTGCAGAATTAG, CATTGTGGTTGACGGTACGTAAG	yes	yes
LCR_3482	long	AGACAGCAATTAAGCAGCATCG, TTTTAAGCTGGTCCTGGGTTG	yes	yes
LCR_3488	long	TCGCGAATCTCGCACCTAT, AGAAATTAACCTGAAAACGAACCA	yes	yes
LCR_3539	long	TTTTGGTCGACAGCTTATGTTACTG, GATCTGTATCTTGGCCACGTTCT	yes	yes
LCR_3542	long	TTTTGTATCAGATGGTGCACAAAG, CCAGTCTGACCGATTATCACAC	yes	yes
LCR_3552	long	AAGATGGTACTACTTGCACTACAACTCC, TTATCTTTGTGCTGCCAACATCT	yes	yes
LCR_3555	long	CTGAAATCTTGCCTTGGGATCT, CTGAGAGACGGCTTCTAGCAAA	yes	yes
LCR_3556	short	ATTTGTTTAAGGTATTGATGCCACTCT, AGCAAGCGCCATGGAAATA	yes	yes
LCR_3646	short	TGAGTAGTCTTTTGTTGACTGCTAGTTC, GGTGATCTTTCTTACATTTATGCAATC	yes	yes
LCR_3744	short	GTAATGGGTAAGGCTATTTCAGATGT, GCATGGCGTAAAACTCGCTAT	yes	yes
LCR_4003	long	ACTTATCATTGTCTCCCGCTTC, CAAGTGTTGCAGGATTGAGC	yes	yes
LCR_4017	long	AGGACACGAAATAGATGGGTATTTT, TGGGCGAGGCTCTCATTA	yes	yes
LCR_4048	short	ACTAAATAATCTACCCGGTCCAAAC, CTGAGGCTACAATCGGCATC	yes	yes
LCR_4078	short	CTCATTCCGATTTAGATTTCTAGGG, AAACGAATCTTCCTTCCTTCTTC	no	NA
LCR_4089	short	GAGAATTGGCTTTATATACTCACAATGG, TCAAGCATGTCCTTGTCCAAA	yes	yes
LCR_4102	short	TGAGTACATAGGTTTAAATGAGCAATC, ACAATGGTTTGAATCTCATACGTG	yes	yes
LCR_4163	short	ACTTAGTCAAGCTTCGGAATAATCA, GGCAGCCTCATCACTGAAA	yes	yes
LCR_4174	short	TTATCTCGAAACAAACGTTGCAG, GACTTGGAAATTTAAGCAAGAGACC	yes	yes
LCR_4189	short	TCGTATTCAATCTCAATTCAAGTCCT, CTTAAGAATTTGTGTTGGTTTCTGG	no	NA
LCR_4217	short	GACTTTATTACAACATCATGGCTTTTC, ATGAGCATCATTTGGTTTCTGG	yes	yes
LCR_4223	long	CCCTCAATCGCCCATTTT, TCCGAATTCAGTAAGGATGTTAGA	yes	yes
LCR_4233	short	TGGTTCCCGGTGTGAAGA, CCAGAATGAGGTTTTAGCTGATAGTT	yes	yes
LCR_4323	short	GAGTCAAAGGAGAATGGGGAAT, AAAACGGCCCTAGCTGGTT	yes	yes
LCR_4360	short	CTACACATGAGAACGTGGGTTAAG, ACAGCCAAGCAACGACCA	yes	yes
LCR_4383	short	CACATTTCTAAATTTCTGACCAACTTC,	yes	yes

		ATTCTCATATTTACTTCGGCACCA		
LCR_4425	short	CTGAGAATGGATATGCTTTAACACC, TACCATGGATCTCAGTAAGGCAAC	yes	yes
LCR_4512	long	TCTCCGTATCCCTCTTGACCTT, AGAAGATTAGGCTGCACACGAC	yes	yes
LCR_4577	short	GATTATCCACAACAATGGCTACG, TTGCACCAGCTCTCCTGTAAA	yes	yes
LCR_4605	short	GAGAAGCTGTGGGACGGAAT, TAGGTTGATACACTTAGCAATTGAACAC	yes	yes
LCR_4673	short	AGAGTCTGTAAACGAGATAAGCCAGA, ATCTGTTCAATGCGATGGAAGT	yes	yes
LCR_4705	short	CAAGTTCTTGAACCCCGTAGG, ACCTCCCACCTTGTCTTCGT	yes	yes
LCR_4789	short	AAGTATTAGCAGACGGACCTCCTTT, ATATTAAAACTCCCAACTTCGAGAGAG	yes	yes
LCR_4878	long	TTGCAGAAGAGTGGAACGAGT, GTTTCACCGGTTCAAGGATG	yes	yes
LCR_4879	long	GAGTGAGGCATCAATTCGGTTA, ATTGCTAATTGTATGTTTTGGTAGGTG	yes	yes
LCR_5043	long	GCATTGTCTTCTGCGTAGTAAAATC, GATCACTAATGTGTGTTCACCTCAA	yes	yes
LCR_5083	short	CATTGAGATTAGATTTGTGTGTGTCAG, ATGGATACCAGACACGATCATAGG	yes	yes
LCR_5152	short	AATTTGTCTGGGTATGAAAATACACTC, GTCCTTCATTTGTTAGGAACTTATGTC	yes	yes
LCR_5153	short	ATCTAAGTTCTGACAAGTTTCTGTTACG, TATGAGATCTGCACAGAACAATCA	yes	yes
LCR_5160	short	TCTCTGCAAGAAACCTACTGTGA, ATCTAAAAGCCAACTCAGCCAAC	yes	yes
LCR_5265	short	TGGAAATTGGAACAACAATGG, TTTGGTCGATCATCTTGCAG	yes	yes
LCR_5278	short	AGATGTGACTGAATCTAGAGCACGA, ACACTTGTCGAAATGGCTACTCA	yes	yes
LCR_5397	long	CTCGCTGCTTCTGCGATAA, ACGATCAATGCCAGGTTTGTA	yes	yes
LCR_5646	short	TAGAATATGACATTTGTGCCTACTTTC, CCGTCGAGAAAGTTAAGAGCTTC	yes	yes
LCR_5715	long	ACTAATTGGTGCCGCGTATCTA, TGGGAAGACGCAACCGTA	yes	yes
LCR_5894	long	CAGAGGGCGATTACGATGAA, TCGATTGAAGGTATTAGATAAGGAAAC	yes	yes
LCR_5903	short	GTTTACAGAATTGGGACATTGGA, CGTAAAATCAAAGTCGTCCTCTAAC	yes	yes
LCR_5918	short	TTTACAGGAGACAAAAGGGATCG, ACCTCCATCAAGAACTTCAAGAAA	yes	yes
LCR_5943	short	AACTCGACACACATCAACGTTAAATA, AGAGTCTTAGTGTTTTACATCAACTTCG	yes	yes
LCR_5980	short	TTCGAAACCCCAATCTGTATGT, GGTTTGTTGGGTTACTTAGCTTTT	yes	yes
LCR_5999	short	GAACCGAGCCAGTAAGGTCA, TTCTTGGTTTGGTAGCTTCAGTT	yes	yes
LCR_6108	short	CACTTCTCAAGTTATCGTTCTTTCATT, TTATGCAGCGAGTGCATGG	yes	yes
LCR_6394	long	AGGAGCAAATATTGCGTGTGAT, GATATCTCTCTCTACGCTCCTTCTGT	yes	yes
LCR_6442	long	TCAAATCCGTTCCTTTGTCTCT, CCCTGTCAATCTCTTTGCTTG	yes	yes

LCR_6476	short	TCCGCTTCAGCTCGTTTCT, ACCAACAAATCCTCGCCATTA	yes	yes
LCR_6490	short	GAAATAAATCTTTCGGAGAATACCC, AAAAGGGGCCCACCACTTA	yes	yes
LCR_6515	short	TATCTCCGTTCCTATTTGATGGAGT, TAAGAATACCGTTGTCAGTTAAAACCA	yes	yes
LCR_6586	short	TGGCAATGTTCCTAGTTTCCATAG, TCGACCTGAAAGAATGGAAGAA	yes	yes
LCR_6605	short	GTGGTTGCCAAACTCAAAGC, CACAAGTGCATTCACCGAAC	no	NA
LCR_6641	short	GCTTTCTCTGTCAACAATAAAATGAAC, AGTTTTACACTAACCAAAACCGAAC	yes	yes
LCR_6668	long	CCACCTTGGTTTCTCCATCC, AAAAGATTCATGATTGCTCTGCTT	yes	yes
LCR_6698	long	ATCCTCATATCATAAAGTTGAAGGAGA, TATAAGACATTCAAAGAGCGAAAACC	yes	yes
LCR_6713	short	CTCAAATCACACGCCTCGTT, GTATGAGGAGCCTCGAGATTGTT	yes	yes
LCR_6714	short	CAATTGTGGCAACACATATCCA, GATGATTGTGCTGTTTGTTTTGA	yes	yes
LCR_6748	short	GCCTACCACTAATGGCATAACA, CAACAAGGACCTGAAGATCCAC	yes	yes
LCR_6749	short	ATATCTCGTCGCAGACTCCATAA, ATTTCTCAGGTCATGATCAAACTCA	yes	yes
LCR_6768	long	TGTTGTTGTCTGGAGCTATTTGA, TAATGATTGTGAATGAACCAACG	yes	yes
LCR_6777	short	ACTGACCCAGCTGCTTGATTT, AACTTGAGAACTATATCATTCCATTGC	yes	yes
LCR_6906	short	CTTTGCTTTATTTCGACGACTTTT, TGGACAATACATACAATTCGAAGC	yes	yes
LCR_6964	short	GTGCTGTCATCTTCGCCTTC, ATTATGCTTACCACGAACAAAGC	yes	yes
LCR_7007	short	TGATTATGGTTTTGTCTCCACCT, GTACAAATGTGTTGAAATTCTATCGAG	yes	yes
LCR_7021	short	CCTCGTTGGCAGTAGTGGTT, TTTTCGGGATGCGAGGAT	yes	yes
LCR_7087	short	ACTGGTAGGTCCATTTTATCCTATACC, GGTGTCTCAAGGCTTCTTGTTC	yes	yes
LCR_7169	short	GGTGGACATGTACGGTGGTA, GTTGACTACAACCCTTCACCAAG	yes	yes
LCR_7221	short	AAACTAGCAGAACGTAAGCAAGGT, GGACTGCTATGGCTCATCCT	yes	yes
LCR_7236	short	TCGTCTTCTTCCCAGCCTTT, TGATCAGAGGTTTGGTTTCTTCT	yes	yes
LCR_7358	long	CTAGGATGGTAACCTTGTACCCTATG, GGGGAAACTATTTGTTGAAGAGC	yes	yes
LCR_7377	long	TCTTTATGCCATTGTAACTCAGGTAAT, CAACTTCTGTCTGAACCGGACT	yes	yes
LCR_7380	short	TGTGACGAGGCTGCAGAA, CATCTTGTTCTTAGATAACCCCTCTT	yes	yes

[a]See data release at http://1001genomes.org for the position of contig sequences in the reference genome.

NA, not applicable.

Table S8. Polymorphisms detected by accession in alignments of anchored contig sequences to the reference genome sequence.

Type	Length (bp)	Accession			
		Col-0	Bur-0	Tsu-1	Bur-0 & Tsu-1 (non-redundant)
SNPs	NA	28	13,599	6,997	19,283
Deletions	All	10	5,830	2,752	8,082
	1-3	10	2,233	1,062	3,122
	4-10	0	2,181	991	2,953
	11-50	0	1,212	594	1,710
	>50	0	204	105	297
Insertions	All	22	4,970	2,228	6,681
	1-3	17	1,981	971	2,772
	4-10	2	1,858	795	2,442
	10-50	3	1,009	434	1,326
	>50	0	122	28	144

NA, not applicable.

Table S9. Regions with an elevated coverage and their support by CVPs.

	Elevated coverage regions			**Elevated coverage regions with 1 or more CVPs**		
	n	**Total bp**	**Mean length (bp)**	***n***	**Total bp**	**Mean length (bp)**
Col-0	687	199,103	289.8	23	11,612	504.9
Bur-0	1,466	554,443	378.2	757	332,150	438.8
Tsu-1	1,278	498,614	390.2	872	363,575	416.9

Table S10. Major-effect changes by type in coding sequences with the number of distinct coding genes affected.

Type[a]	Bur-0		Tsu-1		Non-redundant	
	n	**Genes affected**	***n***	**Genes affected**	***n***	**Genes affected**
Premature stop	446	389	416	366	764	651
Stop codon to coding	159	153	135	129	225	216
Loss of initiation methionine	30	30	32	32	48	48
1-2 bp indel (frameshift)	605	515	573	477	1,027	833
1-2 bp insertion (frameshift)	523	441	460	407	812	682
PR of ≥100 bp	1,282	634	1,098	559	ND[b]	ND[b]
All large-effect changes	3,045	1,871	2,714	1,694	ND[b]	ND[b]

[a]Polymorphisms detected in Bur-0 and Tsu-1 from read alignments (cf. Tables 4 and 5).

[b]Non-redundant PRs could not be determined (ND). PRs identify intervals corresponding to polymorphic sequences, but whether PRs in different accessions at the same locations reflect identical underlying sequence difference cannot be directly assessed.

Table S11. Distribution of major-effect changes by gene family for Bur-0.

Gene Family	*n*	Total coding bases	All major-effect changes	Affected genes (count)	Affected genes (ratio)	Major-effect changes / CDS bases[a]
Ribosomal	411	218,805	9	6	1.5%	4.1×10^{-5}
bHLH	117	124,233	3	3	2.6%	2.4×10^{-5}
MYB domain	141	143,877	5	5	3.6%	3.5×10^{-5}
Cytochrome P450	246	362,802	54	31	12.6%	1.5×10^{-4}
F-box	679	802,338	175	128	18.9%	2.2×10^{-4}
NBS-LRR	121	378,939	146	52	43.0%	3.9×10^{-4}
All coding genes	26,819	33,156,064	3,045	1,871	7%	9.2×10^{-5}

[a]If an appreciable fraction of major-effect changes result from erroneous predictions, larger genes would be proportionally more likely to harbor such changes. Although gene size varies systematically by family, e.g., NBS-LRR genes are on average 5.9 fold larger than ribosomal protein family members, the incidence of major-effect changes per coding base per family nevertheless varies more than 10-fold.

Table S12. Distribution of major-effect changes by gene family for Tsu-1.

Gene Family	*n*	Total coding bases	All major-effect changes	Affected genes (count)	Affected genes (ratio)	Major-effect changes / CDS bases[a]
Ribosomal protein	411	218,805	4	4	1.0%	1.8×10^{-5}
bHLH	117	124,233	2	2	1.7%	1.6×10^{-5}
MYB domain	141	143,877	3	3	2.1%	2.1×10^{-5}
Cytochrome P450	246	362,802	39	23	9.4%	1.1×10^{-4}
F-box family	679	802,338	164	116	17.1%	2.0×10^{-4}
NBS-LRR	121	378,939	158	50	41.3%	4.2×10^{-4}
All coding genes	26,819	33,156,064	2732	1,694	6.3%	8.2×10^{-5}

[a]If an appreciable fraction of major-effect changes result from erroneous predictions, larger genes would be proportionally more likely to harbor such changes. Although gene size varies systematically by family, e.g., NBS-LRR genes are on average 5.9 fold larger than ribosomal protein family members, the incidence of major-effect changes per coding base per family nevertheless varies more than 10-fold.

SUPPLEMENTAL FIGURES

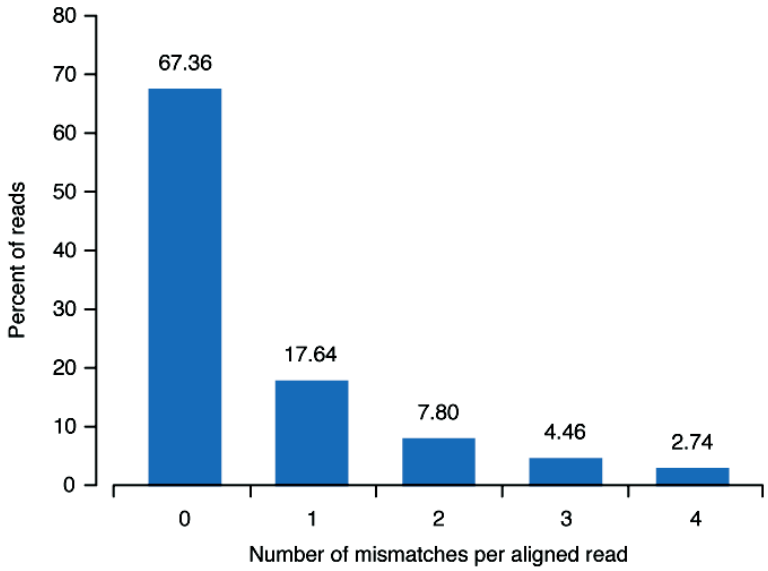

Figure S1. Mismatches per read as assessed with Col-0 alignments before applying base quality filtering. Only uniquely mapped reads were used in the analysis (n = 52,505,997).

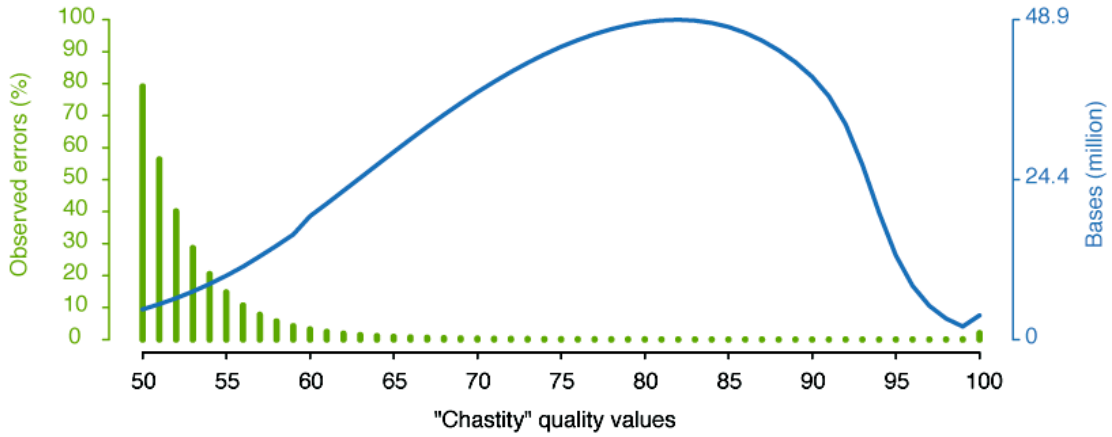

Figure S2. Observed errors at a given *chastity* value with the frequency spectrum for *chastity* values for all bases.

The data is from uniquely aligned Col-0 reads.

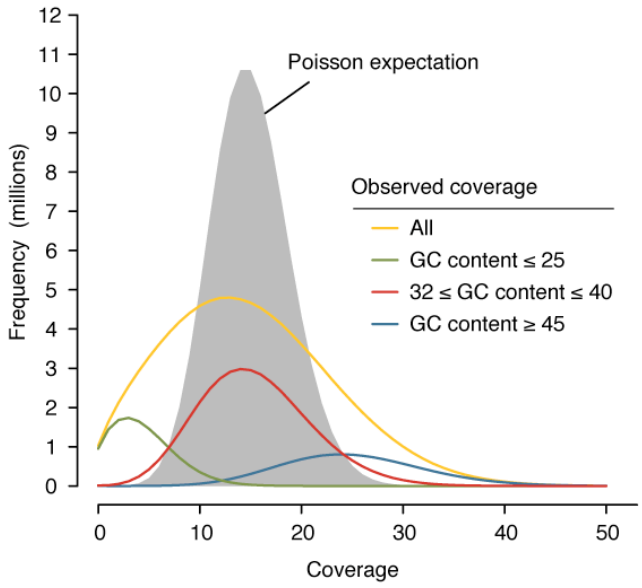

Figure S3. Random versus observed coverage in non-repetitive regions and the effect of GC composition. Data is from Col-0 with GC content assessed in 101 bp windows (see Fig. S4 for selection of window size).

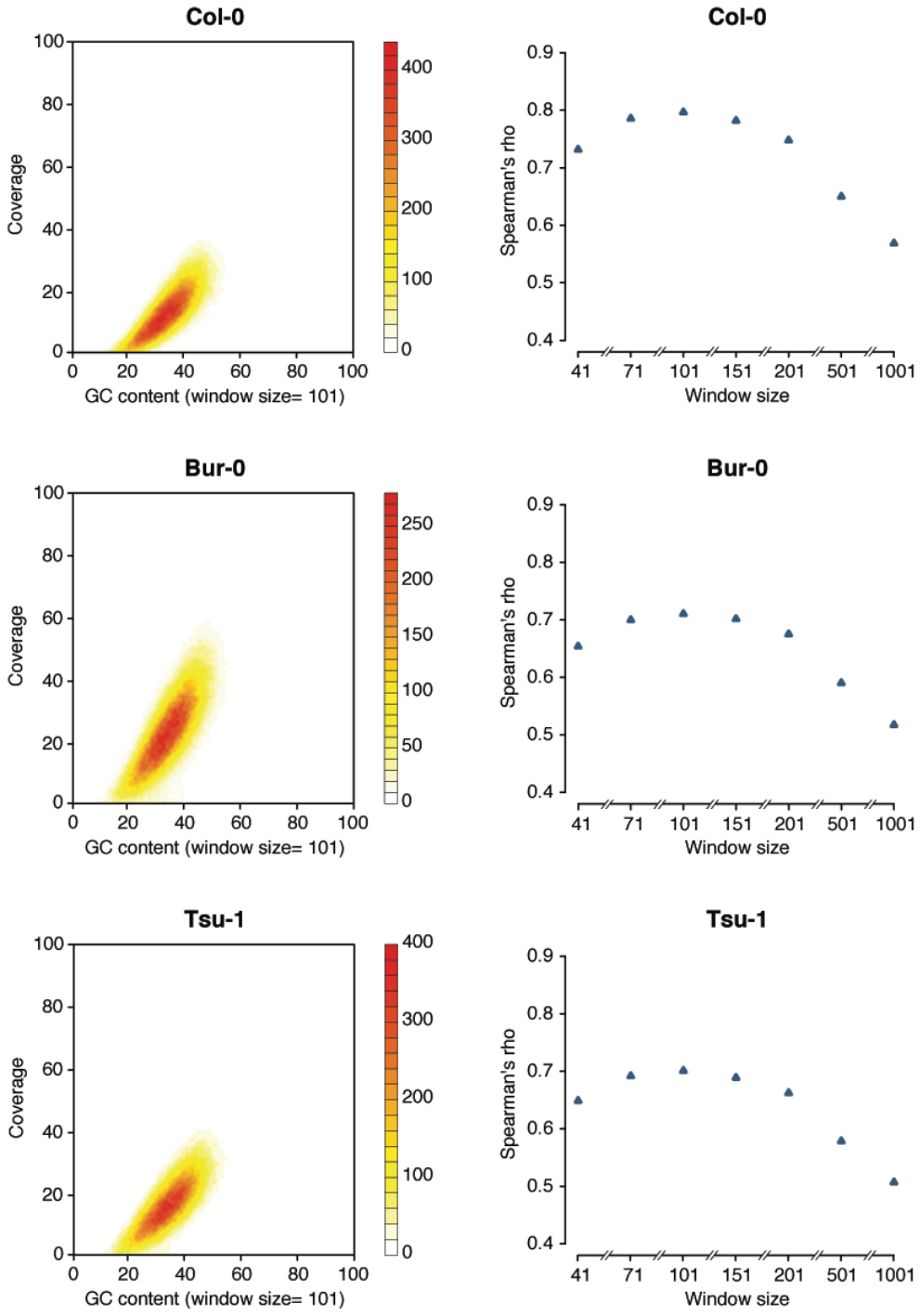

Figure S4. Correlation between GC content and coverage by accession at non-repetitive positions.
Contour plots (left) reveal the relationship between GC content in a 101 bp window and coverage. As assessed with Spearman's rank correlation coefficient (Spearman's ρ), the strongest correlation with local GC content was apparent with a 101 bp window (right; correlation coefficients were significant in all accessions with all window sizes at $P < 10^{-10}$).

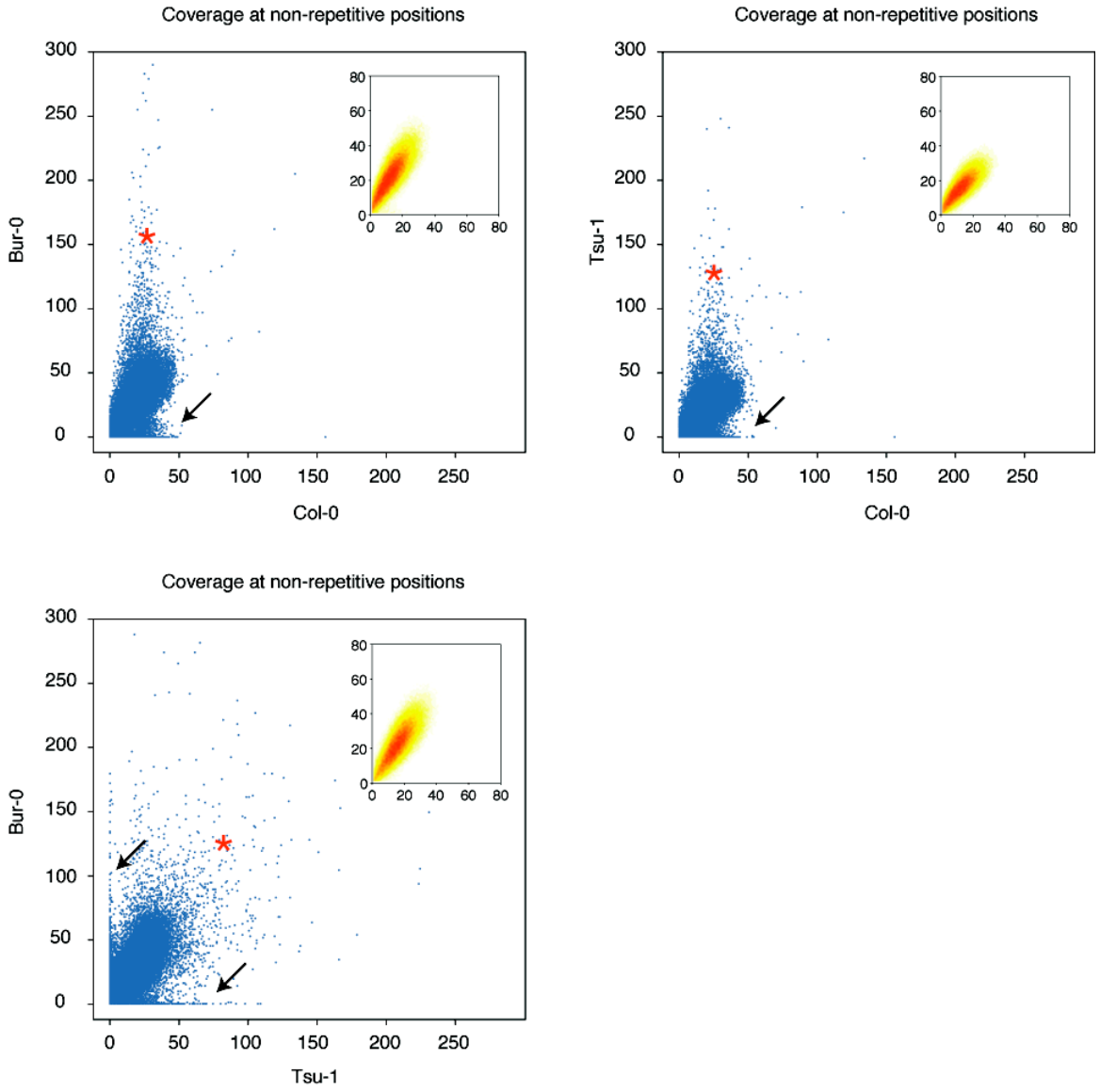

Figure S5. Pair-wise relationships for coverage among accessions.

Scatter plots are shown for each comparison, with contour plot insets for the regions of highest density (coverage less than 80). Although coverage is strongly correlated among data sets, regions of low or no coverage in a given accession identify potential polymorphic regions or deletions (black arrows). Conversely, potentially duplicated regions have proportionally higher coverage (red asterisks). The mean coverage depths per accession are given in Table 1.

Figure S6. Illustration of ambiguous alignments at read ends.

Alignments are shown for 15 reads that span or overlap a 2 bp insertion with red indicating mismatches or insertions in reads. For reads 1-11 that overlap the insertion at central positions, the alignments accurately reflect the polymorphism. However, reads 12-15, which terminate near the insertion, can be aligned with either mismatches (that suggest SNPs; B) or with gaps (C). Detection of polymorphisms from read mapping strategies that do now allow gaps, therefore, is likely to lead to a high false prediction rate. This is expected to be especially true when polymorphisms are inferred from low coverage data. As the potential for misalignments at central positions in reads is minimal, prioritizing central read alignment data is expected to drastically reduce the potential for false polymorphism predictions (see Methods). Note further that in polymorphic regions, few reads will map, and those that do are likely to have more mismatches owing to incorrect alignments; this motivated our approach to identify low coverage regions for targeted reassembly.

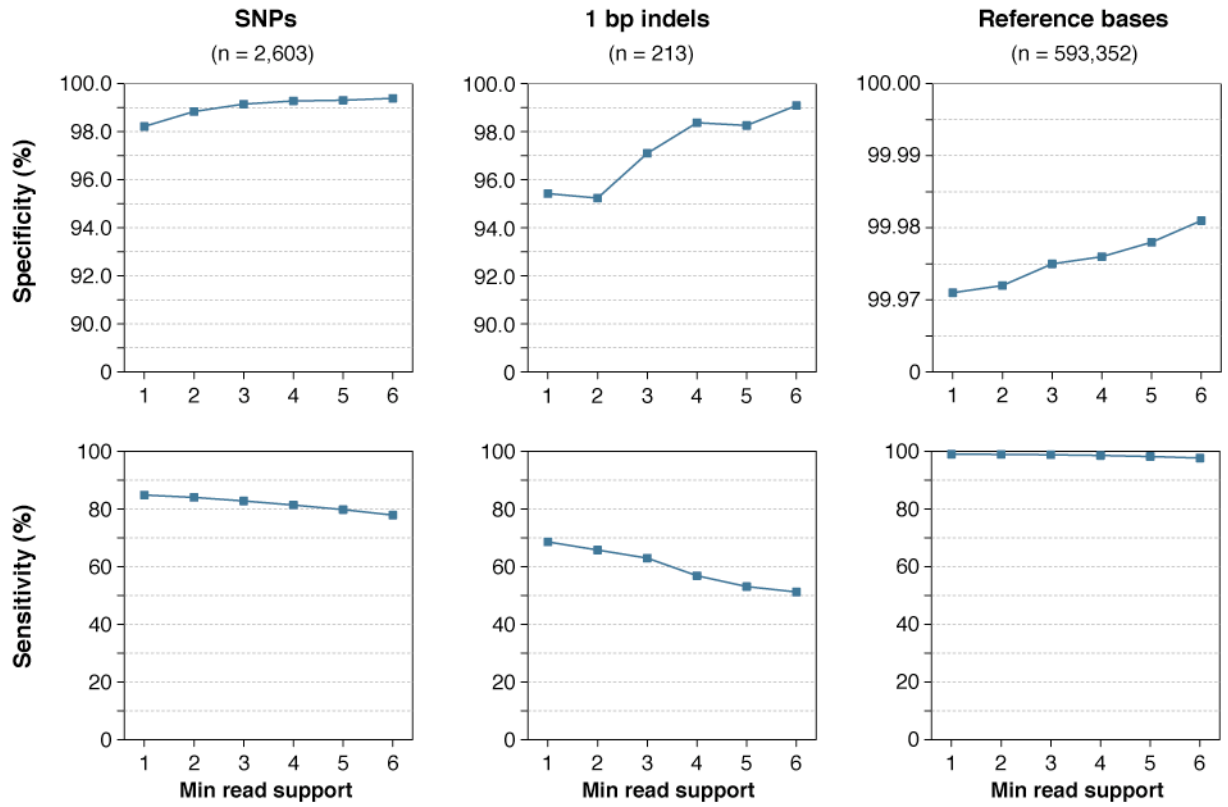

Figure S7. Performance evaluation for sequence predictions with all aligned reads for the Tsu-1 sequence data (cf. Fig. 2).

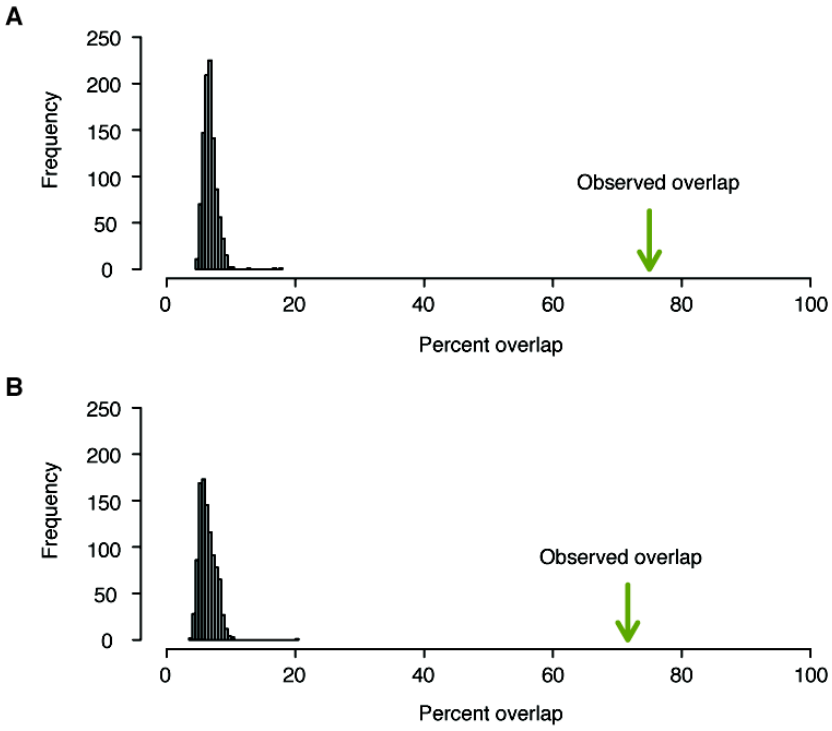

Figure S8. Expected distribution of per base overlaps of SBS PRs to resequencing PRs as assessed by permutation along with the observed values.

The observed overlap values for Bur-0 (A) and Tsu-1 (B) are indicated by green arrows (see Supplemental Methods).

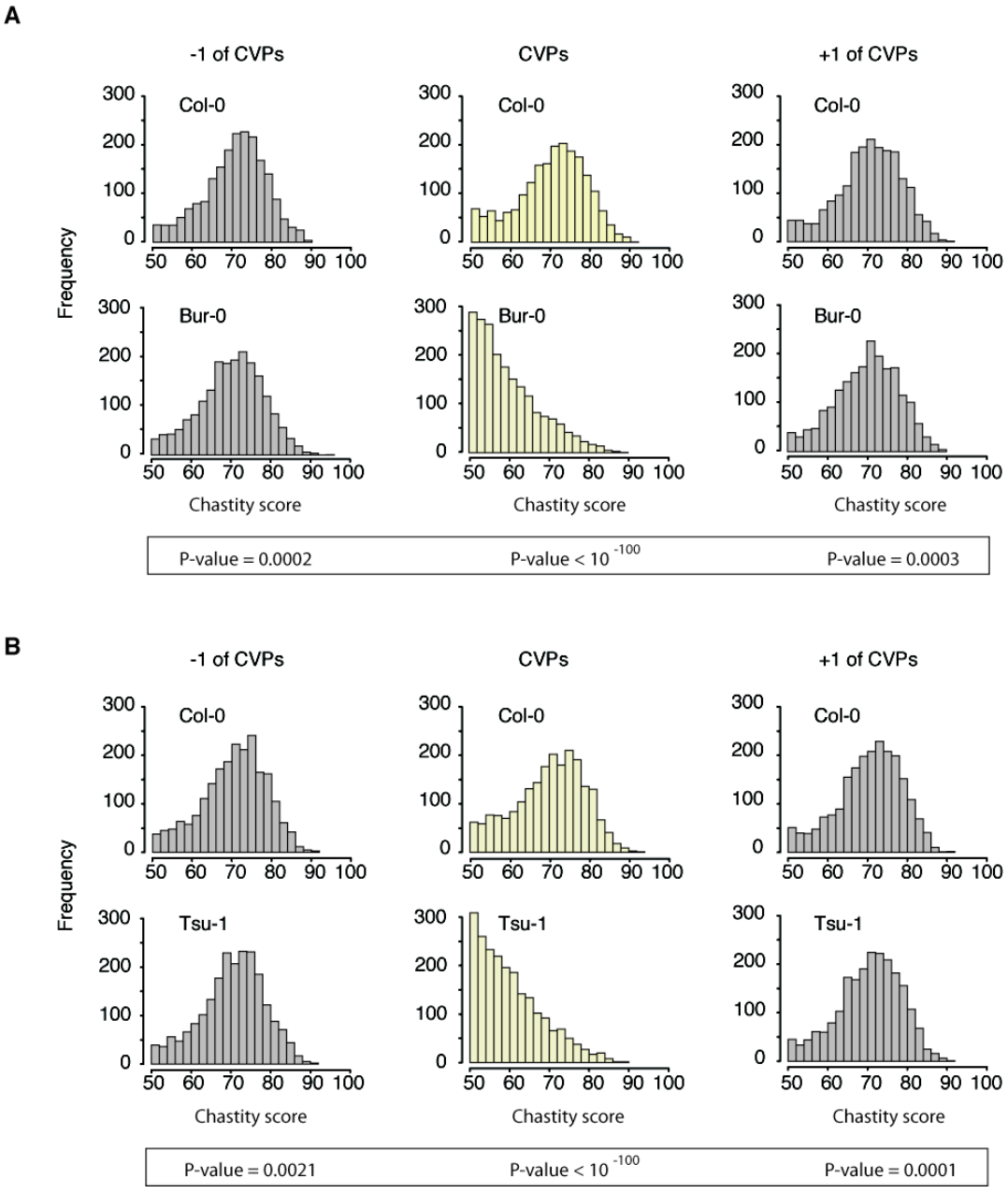

Figure S9. Resequencing array data *chastity* scores at and nearby CVPs.

CVPs in the resequencing array data for Col-0 compared to Bur-0 (A) and to Tsu-1 (B). Shown is the *chastity* score at CVP – 1, CVP (analogous to the yellow columns in Fig 4C), and CVP + 1. *Chastity* scores for both Bur-0 and Tsu-1 peak at 0.5 at CVPs while *chastity* scores for all positions in Col-0 as well as position -1 and +1 in Bur-0 and Tsu-1 peak at 0.75. The distributions for the -1 and +1 positions are significantly different as assessed by comparison of bin values with a χ^2- test, potentially reflecting differences in hybridization quality among accessions in the array experiments. In contrast, at CVP positions, the differences in distributions in the Col-0/Bur-0 and Col-0/Tsu-1 comparisons are highly significant.