



# Construction of a High-Density Genetic Map from RNA-Seq Data for an Arabidopsis Bay-0 × Shahdara RIL Population

Elise A. R. Serin<sup>1†</sup>, L. B. Snoek<sup>2,3†</sup>, Harm Nijveen<sup>1,4</sup>, Leo A. J. Willems<sup>1</sup>, Jose M. Jiménez-Gómez<sup>5,6</sup>, Henk W. M. Hilhorst<sup>1</sup> and Wilco Ligterink<sup>1\*</sup>

<sup>1</sup> Wageningen Seed Lab, Laboratory of Plant Physiology, Wageningen University, Wageningen, Netherlands, <sup>2</sup> Laboratory of Nematology, Wageningen University, Wageningen, Netherlands, <sup>3</sup> Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands, <sup>4</sup> Laboratory of Bioinformatics, Wageningen University, Wageningen, Netherlands, <sup>5</sup> Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany, <sup>6</sup> Institut Jean-Pierre Bourgin, Institut National de la Recherche Agronomique, AgroParisTech, Centre National de la Recherche Scientifique, Université Paris-Saclay, Versailles Cedex, France

## OPEN ACCESS

### Edited by:

Daniel Pinero,  
Universidad Nacional Autónoma  
de México, Mexico

### Reviewed by:

Rubén Alcázar,  
University of Barcelona, Spain  
Martin Mascher,  
Leibniz-Institut für Pflanzengenetik  
und Kulturpflanzenforschung (IPK),  
Germany

### \*Correspondence:

L. B. Snoek  
snoek.basten@gmail.com  
Wilco Ligterink  
wilco.ligterink@wur.nl

<sup>†</sup> These authors have contributed  
equally to this work.

### Specialty section:

This article was submitted to  
Plant Genetics and Genomics,  
a section of the journal  
Frontiers in Genetics

Received: 02 October 2017

Accepted: 21 November 2017

Published: 05 December 2017

### Citation:

Serin EAR, Snoek LB, Nijveen H,  
Willems LAJ, Jiménez-Gómez JM,  
Hilhorst HWM and Ligterink W (2017)  
Construction of a High-Density  
Genetic Map from RNA-Seq Data  
for an Arabidopsis Bay-0 × Shahdara  
RIL Population. *Front. Genet.* 8:201.  
doi: 10.3389/fgene.2017.00201

High-density genetic maps are essential for high resolution mapping of quantitative traits. Here, we present a new genetic map for an Arabidopsis Bayreuth × Shahdara recombinant inbred line (RIL) population, built on RNA-seq data. RNA-seq analysis on 160 RILs of this population identified 30,049 single-nucleotide polymorphisms (SNPs) covering the whole genome. Based on a 100-kbp window SNP binning method, 1059 bin-markers were identified, physically anchored on the genome. The total length of the RNA-seq genetic map spans 471.70 centimorgans (cM) with an average marker distance of 0.45 cM and a maximum marker distance of 4.81 cM. This high resolution genotyping revealed new recombination breakpoints in the population. To highlight the advantages of such high-density map, we compared it to two publicly available genetic maps for the same population, comprising 69 PCR-based markers and 497 gene expression markers derived from microarray data, respectively. In this study, we show that SNP markers can effectively be derived from RNA-seq data. The new RNA-seq map closes many existing gaps in marker coverage, saturating the previously available genetic maps. Quantitative trait locus (QTL) analysis for published phenotypes using the available genetic maps showed increased QTL mapping resolution and reduced QTL confidence interval using the RNA-seq map. The new high-density map is a valuable resource that facilitates the identification of candidate genes and map-based cloning approaches.

**Keywords:** Arabidopsis, genetic map, genotyping by sequencing, QTL mapping, RIL population, resolution, RNA-seq

## INTRODUCTION

Quantitative trait locus (QTL) analysis has successfully identified a large number of genetic loci that contribute to the regulation of quantitative phenotypes. The advent of -omics data has extended the range of usual mapping traits to molecular phenotypes offering new approaches for bridging the gap between genes and their function (Keurentjes et al., 2008). The idea that variation in gene

expression can be treated as a quantitative trait gave rise to the concept of genetical genomics (Jansen and Nap, 2001). In combination with a genetic map, quantitative variation in gene expression measured in a segregating population enables the identification of expression QTLs (eQTLs). Many eQTL studies have contributed to our understanding of the genetic architecture of regulatory variation of intricate traits in *Arabidopsis* West (Keurentjes et al., 2007; West et al., 2007; Terpstra et al., 2010; Snoek et al., 2012; Lowry et al., 2013; Cubillos et al., 2014) (for review see Joosen et al., 2009), poplar (Drost et al., 2015), tomato (Ranjan et al., 2016), as well as in other organisms (Li et al., 2006, 2010; Rockman et al., 2010; Vinuela et al., 2010; Aylor et al., 2011; King et al., 2014; Snoek et al., 2017; Sterken et al., 2017).

In essence, the success of QTL mapping is determined by the mapping resolution which mainly depends on the size of the population (and thus the number of recombination events), the complexity of the phenotype, and the number of available markers. High-density genetic maps are thus instrumental for accurate mapping of QTLs. Traditional methods used to obtain molecular markers were mainly PCR based (SSR, AFLP, RFLP). New methods to derive molecular markers have recently emerged, together with the advancement of high-throughput technologies. Particularly, single-nucleotide polymorphisms (SNPs) represent a rich source of potential markers due to their abundance (Alonso-Blanco et al., 2016). Differences in gene expression measured with microarrays as a result of probe hybridization sensitivity to underlying sequence polymorphisms have been used to derive SNP-based markers (West et al., 2006; Zych et al., 2015, 2017). More recently, next-generation sequencing technologies for transcriptome analysis (RNA-seq) have provided unprecedented opportunities for quantitative genetics in plants (Jimenez-Gomez, 2011). Becoming a standard for gene expression profiling, RNA-seq has also proven to be an efficient and cost-effective method to identify genome-wide SNPs (Piskol et al., 2013; Markelz et al., 2017). In the context of genetical genomics, RNA-seq on a segregating population can simultaneously provide the molecular phenotype and the sequence information for molecular markers that subsequently provide genotyping information for the population.

Segregating bi-parental populations such as recombinant inbred line (RIL) populations are powerful tools for QTL analysis (Koornneef et al., 2004). These immortal populations capture frequent recombination events in a relatively small sized population, thereby conveniently reducing the costs for genotyping. In this study, we utilized an *Arabidopsis thaliana* Bayreuth (Bay 0) × Shahdara (Sha) population that has been used extensively for genetic (Loudet et al., 2002; Jimenez-Gomez et al., 2010) and eQTL studies (Keurentjes et al., 2007; West et al., 2007). The original genetic map for this population consists of 69 markers segregating in 420 F6 RILs (Loudet et al., 2002). Further genotyping efforts on a subset of these RILs have introduced markers derived from gene expression data with microarrays, saturating the original map (West et al., 2006; Salathia et al., 2007; Zych et al., 2015). Here, we present the construction of a high-resolution genetic map from RNA-seq data of 160 RILs. We validate and show the improvements of this new map by

performing a QTL analysis with publicly available phenotypic data (Joosen et al., 2012).

## MATERIALS AND METHODS

### Plant Growth and Sample Preparation

Seeds from the *A. thaliana* accessions Bay-0 and Sha and a Bay-0 × Sha RIL population consisting of 165 lines were used. This population was initially developed by Loudet et al. (2002). As part of a larger experiment aiming to investigate genotype × environment interactions, the parental lines and the RILs were grown under standard and controlled mild stress conditions. In the standard condition, plants were grown under long day (16 h light/8 h dark) at 70% RH and 22°C/18°C (day/night) under artificial light (150  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ). The plants were watered with a standard nutritive solution (see Supplementary Table S1 in He et al., 2014) three times a week by flooding cycles. The same conditions were used for the stress environments, except for the varying parameter as indicated hereafter: high temperature (25°C day/23°C night), high light (300  $\mu\text{mol m}^{-2} \text{s}^{-1}$ ), and low phosphate (12.5  $\mu\text{M}$  phosphate instead of 0.5 mM in the standard nutritive solution).

The RILs and the parental lines were first grown with three to four plants per environment in a single climate cell under the control conditions mentioned above. When most of the plants flowered, the main stems of all plants were removed to increase the numbers of side branches and thereby seed production, and to ensure that all seeds would complete their development under the specific conditions. Subsequently the plants were transferred to different climate cells to continue their growth under the specific stress conditions. At the time all plants in a given condition produced a sufficient amount of fully matured seeds; the seeds were bulk harvested from the three to four plants per line. After drying, a fraction of the freshly harvested seeds were stored at  $-80^{\circ}\text{C}$  in sealed 2 ml tubes until RNA-seq library preparation.

### RNA Isolation and Sequencing

RNA was isolated from 4 to 5 mg of fresh harvested dry seeds that were stored at  $-80^{\circ}\text{C}$ . Each of the parents was measured in triplicate per condition, i.e.,  $4 \times 3 = 12$  replicates per parent. RNA was extracted from the seeds of 160 RILs selected in conformity to the generalized genetical genomics (GGG) strategy (Li et al., 2008; Supplementary Table S1). RNA was isolated using the NucleoSpin RNA Plant Isolation Kit (Macherey-Nagel 740949) but adding Plant RNA isolation Aid (Life Technologies) according to the manufacturer's protocol and instructions.

### RNA-Seq Reads Processing

Strand-specific RNA-seq libraries were prepared from each RNA sample using the TruSeq RNA kit from Illumina according to manufacturer's instructions. Poly-A-selected mRNA was sequenced using the Illumina HiSeq2500 sequencer, producing strand-specific single-end reads of 100 nucleotides. Reads were trimmed using Trimmomatic (version 0.33, Bolger et al.,

2014) to remove low quality nucleotides. Trimmed reads were subsequently mapped to the *A. thaliana* TAIR10 reference genome (Lamesch et al., 2012) using the HISAT2 software (version 2.0.1, Kim et al., 2015) with the “transcriptome mapping only” option. SNPs were called using the mpileup function of samtools (version 0.1.19, Li et al., 2009) and bcftools. The raw sequence data have been uploaded to the NCBI under the project identifier: PRJNA418075<sup>1</sup>.

## SNP Identification and RIL Genotyping

Variant call format (VCF) files were generated for each of the samples. Since not all SNPs are found in all genotypes, all vcf files were merged to generate a list with all variants present in at least one sample. From this unique list, information regarding the position in base pairs and the chromosome location of each SNP was retrieved and filtered for being consistent across the sequencing data of the parental lines. In order to get a more reliable genotypic score, canceling out any SNPs miscalls, and to reduce the overall number of markers, SNPs were grouped into bins. 1059 equal size artificial bins of 100 kbp were created along the whole genome. The scoring of the genotype was obtained based on the SNP information within each bin. For regions at the transition between two genotypic blocks, the bin score was rounded up and assigned to the closest genotypic score. The quality of the genotype scoring of the bins was assessed by correlation analysis.

## Nomenclature

The bins are ordered based on the genome sequence, thus the unit distance is not expressed in centimorgans (cM) but in bins of 100 kbp. Each bin is used as a marker and the midpoint position of the 100 kbp bin is used as the marker position. Markers were named RSM for RNA-seq markers, followed by the chromosome number of their location and their physical position in mega base pairs (Mbp). As an example RSM\_1\_0.05 corresponds to the marker at 0.05 Mbp on chromosome 1.

## Genetic Map Construction

The genetic distances in centimorgans of the 1059 markers for 160 RILs were estimated in order to describe and compare the new genetic map to previous maps. The genetic distances were estimated using the “est.map” function with “kosambi” distance from the R/qtl package (Broman et al., 2003; Arends et al., 2010). The correct order of the markers was verified by pairwise marker linkage analysis using the “est.rf” function. The recombination rate was determined based on the linear relation between the genetic and the physical positions of the marker. The segregation pattern was tested for all markers to identify markers that show significant distortion at the 5% level, after a Bonferroni correction for multiple testing. The statistical programming language R (version 3.3.2) (R Development Core Team, 2008) was used for all analyses. The genetic map and genotypic data are available in Supplementary Table S2.

<sup>1</sup><http://www.ncbi.nlm.nih.gov/bioproject/418075>

## QTL Comparison

To test the effect of increased marker coverage on QTL mapping, we re-mapped 510 published phenotypic traits using the RNA-seq (1059 markers), the pheno2geno (497 markers) (Zych et al., 2015) and the original map (69 markers) (Loudet et al., 2002). In order to compare the mapping resolution, the genetic distances were re-estimated for each map using 145 RILs common to the three studies (Supplementary Table S1). The scanone function in R/qtl was used with the default settings for the QTL mapping. LOD score peaks were called by chromosome for each trait, resulting in a total of 2550 (510\*5) peak LOD scores. The LOD threshold for the genome-wide significance at the level of 5% was determined after 1000 permutations using each map. The LOD thresholds obtained were 2.36, 2.64, and 2.76 using the original, pheno2geno, and RNA-seq map, respectively. The increased LOD thresholds for the Pheno2geno and the RNA-seq map can be explained by the larger number of markers which will result in a larger multiple testing correction. We used a stringent LOD threshold of 3 to identify and compare significant QTLs for all maps. The LOD score comparison was performed in a similar way as described in Zych et al. (2015). To be more confident about the comparison, QTLs were considered to have a higher or lower LOD score if the difference between the compared LOD scores was larger or equal to 0.5. The mapping resolution of the RNA-seq map was investigated by comparing the confidence intervals (CIs) of QTLs for the RNA-seq and the original map. LOD-1 CIs were determined for all significant QTLs (LOD > 3) for both maps. The genomic positions of the lower and upper limit of each CI were estimated from the equation of the linear relation between genetic and physical position of the markers. Subsequently, the CI width was determined for each QTL in Mbp. The analyses and figures were generated using Microsoft Excel, R/qtl, and the R ggplot2 package.

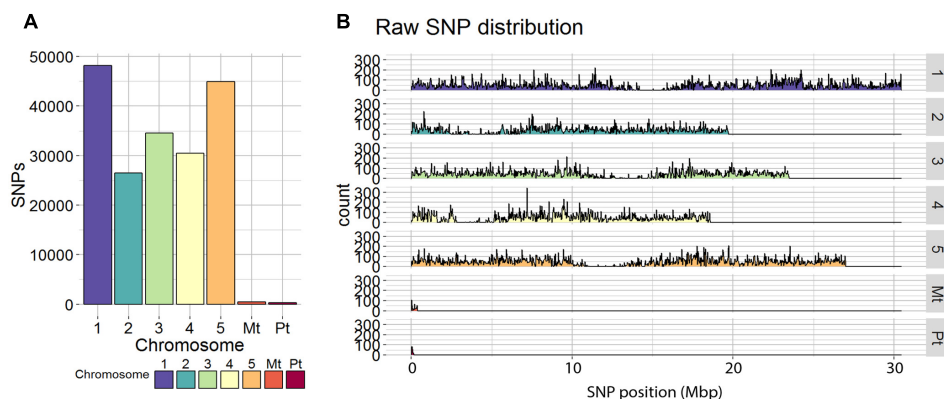
The cross object containing all data for the 510 phenotypes in the 160 RILs for the QTL analysis is available in Supplementary Table S3. The QTL results for the comparison of the LOD scores and CIs are provided in Supplementary Tables S4 and S5. QTL profiles of the re-mapped 510 traits are available for interactive analysis in AraQTL<sup>2</sup> (Nijveen et al., 2017).

## RESULTS

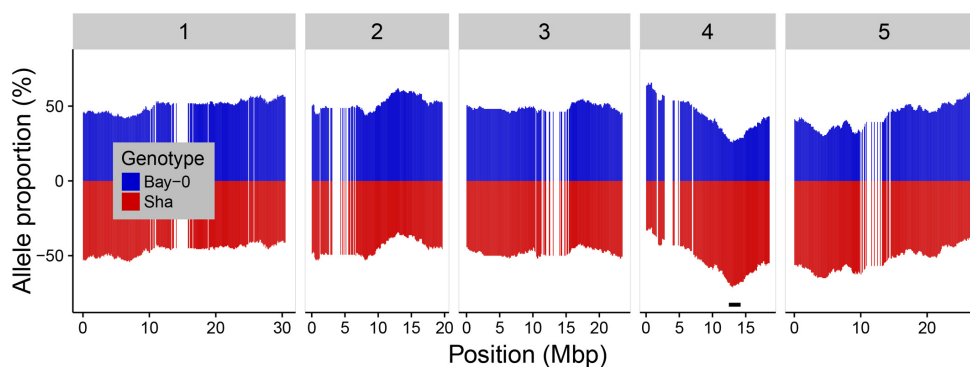
### Genotyping the RIL Population Using a SNP Binning Approach

Single-nucleotide polymorphisms calling resulted in 185,354 SNPs distributed over the five chromosomes, ranging from 26,514 SNPs for chromosome 2 to 48,151 SNPs for chromosome 1 (Figure 1). Regions with a few or no SNPs correspond to centromeric regions, known to have lower transcriptional density and expression activity (Schmid et al., 2005). Filtering and quality check of the SNPs (as described in the section “Materials and Methods”) resulted in a final number of 30,049 SNPs covering the whole genome.

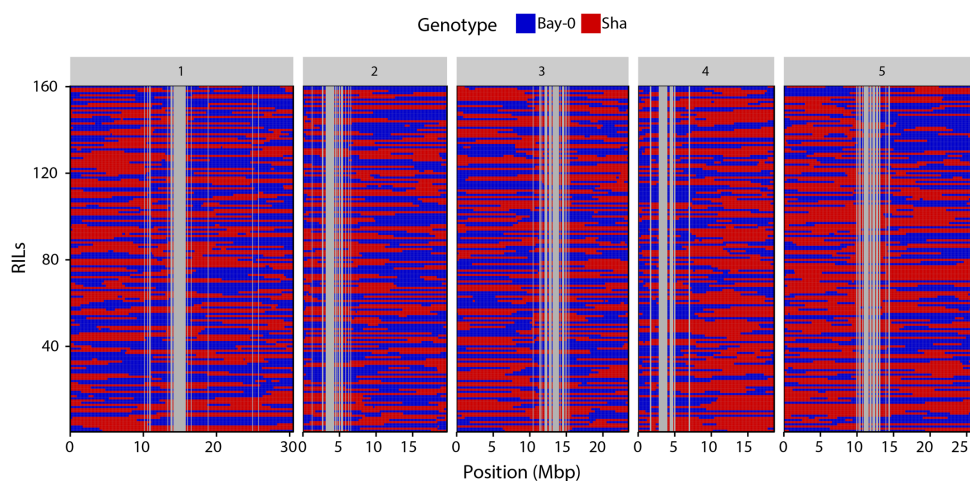
<sup>2</sup><http://www.bioinformatics.nl/AraQTL/>



**FIGURE 1 |** Raw SNP distribution from all genotyped RILs. **(A)** Total SNP count and **(B)** coverage counts of each SNP at each physical position on the chromosome in mega base pairs (Mbp) are displayed for each of the five chromosomes of *Arabidopsis thaliana* as well as the mitochondrial (Mt) and plastid (Pt) genomes.



**FIGURE 2 |** Allele distribution for the 1059 markers along the five chromosomes. Blue and red colors indicated the Bay-0 and the Sha allele percentages, respectively. The black horizontal bar indicates the region on chromosome 4 with 29 markers showing significant segregation distortion ( $p$ -value < 0.05 after Bonferroni correction).



**FIGURE 3 |** Haplotype representation of the 160 RILs. Each row corresponds to a RIL. Columns represent the 1059 genetic markers physically anchored on the five chromosomes. Blue boxes indicate Bay-0 genotype and red boxes indicate Sha genotypes.

**TABLE 1** | Characteristics of the 1059 marker genetic map using 160 RILs.

Chr	Markers	Total length (cM)	Average marker distance (cM)	Maximum gap (cM)	cross-overs	Recombination rate (kbp/cM)
chr 1	275	117.87	0.43	2.87	364	258.34
chr 2	171	76.29	0.45	3.23	236	257.58
chr 3	207	82.61	0.40	2.72	255	283.85
chr 4	163	92.12	0.57	4.81	281	201.36
chr 5	243	102.81	0.42	2.34	319	262.13
Total	Total 1059	Total 471.70	Average 0.45	Max 4.81	Total 1455	Average 252.65

The 100 kbp binning approach used, collapsed the 30,049 SNPs into 1059 bins distributed over the five chromosomes. Each bin contained on average  $\sim 24$  SNPs, with a minimum of 2 and a maximum of 130 SNPs per bin (Supplementary Figure S1). Overall, 96.7% of the bins could unambiguously be assigned to one of the parental genotypes.

Population-based SNPs segregated at the expected allele frequencies as global allelic equilibrium was observed with 49.3% Bay-0 alleles and 50.7% Sha alleles. Bias in the segregation ratio between the parental alleles was analyzed along the chromosomes (Figure 2). Statistically significant distortion of segregation was observed for 29 consecutive markers on chromosome 4, representing 2.78% of the total number of markers. These distorted markers correspond to the region comprised between the markers RSM\_4\_12.05 and RSM\_4\_14.85. The highest distortion was observed at the marker RSM\_4\_13.05 with 41 (25.6%) lines representing the Bay-0 allele versus 114 ( $114/160 = 71.25\%$ ) lines representing the Sha allele. This deviation from the allelic equilibrium at the chromosome 4 was also reported by Loudet et al. (2002).

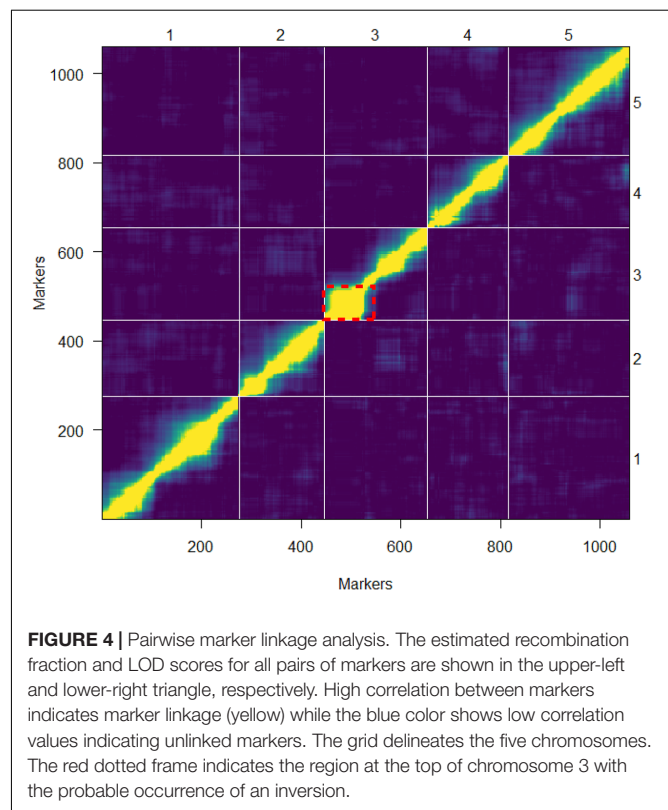
## RNA-Seq Genotyping Identifies New Introgressions

Visually, the binning method resulted in the identification of clear genotype blocks (Figure 3). Breakpoints were identified as the point of transition between two genotype blocks. In total, 1455 crossovers were identified with an average of 291 crossovers per chromosome (Table 1).

To identify introgressions that were previously not detected, the 1059 new markers together with the 69 “old” markers were first ordered based on their physical positions. New introgressions were then identified in the RILs as double recombination events occurring within a region spanned by two “old” flanking markers and of a minimum size of 200 kbp (two bins) (Supplementary Figure S2). We could identify 80 unambiguous introgressions with sizes ranging between 200 kbp and 3 Mbp, increasing the number of recombination events detected within the RIL population.

## High-Density Genetic Map

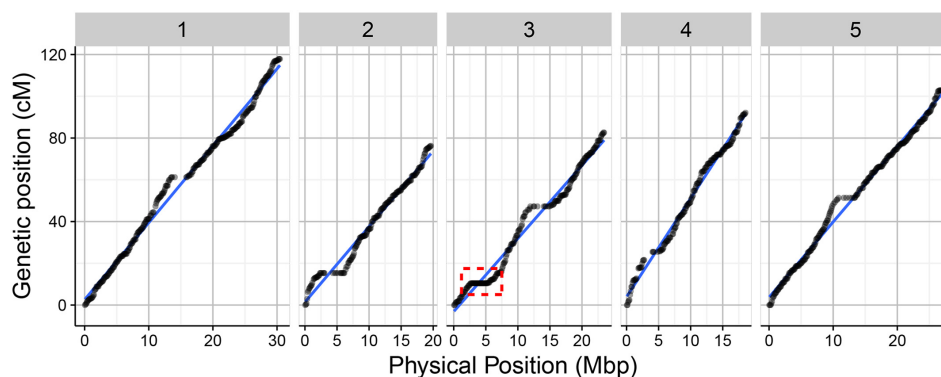
Using each bin as a marker, the linkage map was calculated in order to validate the order of the markers and evaluate the accuracy of the new map. The characteristics of the new map are reported in Table 1. The total length of the genetic map was 471.70 cM. The average genetic distance between two adjacent markers of 0.45 cM represents a great increase in marker density



as compared to the 6.1 cM of the 69 markers map for 420 RILs (Loudet et al., 2002). In the new map, the largest gap between two markers is 4.81 cM between the markers RSM\_4\_1.55 and RSM\_4\_1.85 on chromosome 4.

Overall, the order of the markers on the genetic map conforms to the physical position of the marker and is also supported by the pairwise marker linkage analysis (Figure 4). The recombination rate was calculated as the relation between the physical and genetic distances. Low recombination was observed at the centromeric regions where the physical distance was greater relative to the genetic distance. On the upper arm of chromosome 3, no recombination events occurred between the markers RSM\_3\_2.65 and RSM\_3\_5.25. This was also observed in the 69-markers map as well as in a Sha  $\times$  Col-0 RIL population<sup>3</sup>. A Sha-specific chromosomal inversion in this region was suggested (Figures 4, 5). The global recombination rate is

<sup>3</sup><http://publiclines.versailles.inra.fr/page/13>



**FIGURE 5 |** Relation between the genetic length in centimorgans (cM) and the physical length in Mbp for the 1059 markers along the five chromosome using 160 RILs of the Bay-0 × Sha RIL population. The red dotted frame indicates the region on the upper arm of chromosome 3 without recombination events.

**TABLE 2 |** Summary of genetic maps for the Bay-0 × Sha RIL population based on 145 RILs.

Genetic map parameters	Original	Pheno2Geno	RNA-seq
Number of markers	69	497	1059
Total length (cM)	480.1	499.1	464.4
Average marker distance (cM)	7.5	1	0.6
Maximum gap	22.9	11.6	4.9
Number of crossovers	1137	1366	1297
% genotyped	96.2	100	96.6
Global allele equilibrium	Bay 50.6% Sha 49.4%	Bay 49.7% Sha 50.3%	Bay 49.8% Sha 50.2%
Reference	Loudet et al., 2002	Zych et al., 2015	This study

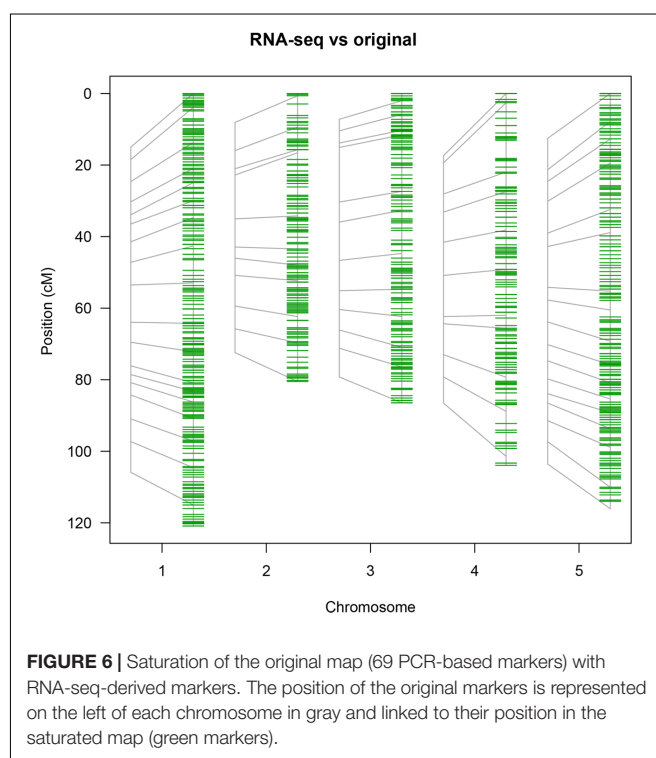
252.65 kbp/cM, i.e., 4.01 cM per 1 Mbp (Figure 5). This rate is consistent with previously reported recombination rate of 246 kbp/cM (Loudet et al., 2002).

## QTL Mapping Comparison

The original genetic map for the analyzed Bay × Sha population developed by Loudet et al. (2002) comprises 69 PCR-based markers. Recently, Zych et al. (2015) saturated the original map with 497 markers derived from microarray expression data (pheno2geno map). To compare the published maps to the RNA-seq map, the genetic distances were re-estimated using 145 RILs common to the three studies (Supplementary Table S1).

The RNA-seq map reduces the average distance between markers from 7.5 cM for the 69 marker map and 1 cM for the pheno2geno map to 0.6 cM (Table 2), closing many existing gaps in marker coverage (Figure 6). In addition, the RNA-seq map captures 1297 crossovers as compared to 1137 in the original map. The number of crossovers observed with the pheno2geno map (1366 cross-overs) is likely inflated due to the imputation of the genotypic data to 100% (% genotyped in Table 2).

Quantitative trait locus mapping was performed to evaluate the mapping resolution of the RNA-seq map as compared to the two other maps. Using a genome-scan single QTL model analysis, 510 published phenotypes were re-mapped using the three maps. The QTL analysis with the RNA-seq map resulted



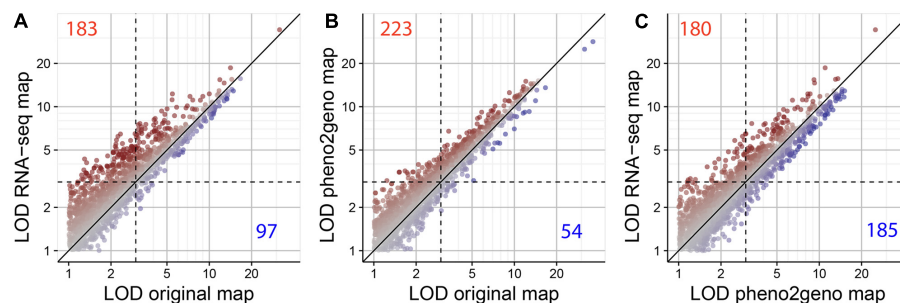
**FIGURE 6 |** Saturation of the original map (69 PCR-based markers) with RNA-seq-derived markers. The position of the original markers is represented on the left of each chromosome in gray and linked to their position in the saturated map (green markers).

in 754 significant QTLs (LOD > 3), while 684 and 568 significant QTLs were detected using the pheno2geno and the original map, respectively (Table 3 and Figure 7). QTLs were considered to have a higher or lower LOD score if the difference between the compared LOD scores was larger than or equal to 0.5. Respectively, 223 and 183 of the total number of significant QTLs in the original map did show an increased LOD score in the pheno2geno map and RNA-seq map (Figures 7A,B and Table 3). When compared to the pheno2geno map, the RNA-seq map resulted in 180 QTLs with a higher LOD score (Figure 7C and Table 3). The pheno2geno map identified 139 new QTLs compared to the original map, while the RNA-seq map added 208 new QTLs. One hundred and twenty-five new QTLs were

**TABLE 3** | Comparison of LOD scores using the different maps.

Genetic map <sup>1</sup> (/compared to)	Significant QTLs (LOD > 3)	"New" and "lost" QTLs <sup>2</sup>	Higher LOD QTLs <sup>3</sup>	Lower LOD QTLs <sup>4</sup>
Original	568	–	–	–
pheno2geno/original	684	<b>139/23 (24%/0.4%)</b>	223 (39%)	54 (9.5%)
RNA-seq/original	754	<b>208/22 (30%/0.4%)</b>	183 (32%)	97 (17%)
RNA-seq/pheno2geno		<b>125/55 (18%/8%)</b>	180 (26%)	185 (27%)

<sup>1</sup>The "new" maps used for the comparison are indicated in bold. <sup>2</sup>New QTLs are the number of QTLs with a LOD score above 3 in the new map and below 3 in the compared map (bold numbers). These numbers are compared to the number of significant QTLs in the compared map "lost" in the new map. <sup>3</sup>Higher LOD QTLs is the number of significant QTLs with a higher LOD score in the new map with a difference in LOD scores equal or larger than 0.5. <sup>4</sup>Lower LOD QTLs is the number of significant QTLs with a higher LOD score in the new map with a difference in LOD scores equal or larger than 0.5. Percentage of new, lost, higher, and lower LOD QTLs in relation to the total number of significant QTLs in the compared map are shown in brackets.



**FIGURE 7** | LOD score comparison of QTLs for 2550 QTL peaks of 510 published phenotypes using the original, the pheno2geno, and the RNA-seq map. **(A)** LOD scores with the RNA-seq map versus the original map, **(B)** LOD scores with the pheno2geno map versus the original map and **(C)** LOD scores with the RNA-seq map versus the pheno2geno map. The significance threshold is indicated by a dashed horizontal and vertical black line. "Stronger" LOD scores are plotted in red. Red and blue numbers correspond to the number of significant QTLs identified on the x-axis map with increased or decreased LOD scores in the y-axis map, respectively.

detected in the RNA-seq map as compared to the pheno2geno. In addition, an increase in the LOD scores was observed using the RNA-seq map as compared to the original map (average LOD score differences of 1.74) and the pheno2geno map (1.66) than for the pheno2geno compared to the original map (1.15) (**Table 4**). Together, these results indicate that the higher marker density of the RNA-seq map provides additional power to detect QTLs.

A main factor for the success of QTL experiments is the precision in the estimation of the position of the QTL. We assessed the RNA-seq map resolution by comparing the CI of QTLs detected in the original map and the RNA-seq map. The CI of 546 QTLs significant in both maps was calculated (LOD > 3). Four hundred and fifty-seven (84%) of the QTLs showed a reduced interval in the RNA-seq map (**Figure 8**). The

difference in interval width ranged from 0.08 to 25.58 Mbp. For example, the QTL for seed circularity at the top of chromosome 5 was delimited to a genomic region of less than 1.12 Mbp using the RNA-seq map compared to more than 26 Mbp using the original map (**Figure 9**). To verify the consistency of these results, the analysis was also conducted with a LOD threshold of 2 and for QTLs with higher LOD scores using the original map (Supplementary Figure S3). Eighty-one percent (770/952) of the QTLs showed a reduced CI using the RNA-seq map when the significance threshold was lowered to LOD > 2 (Supplementary Figure S3A). Analysis of 233 significant QTLs in both maps for which the LOD score was higher in the original map as compared to the RNA-seq map resulted in 72% (169/233) of these QTLs showing a reduced CI using the RNA-seq map (Supplementary Figure S3B). These results clearly show that the accuracy of the QTL mapping is improved by using the high-density SNP bin map.

**TABLE 4** | Average LOD score differences across the different maps.

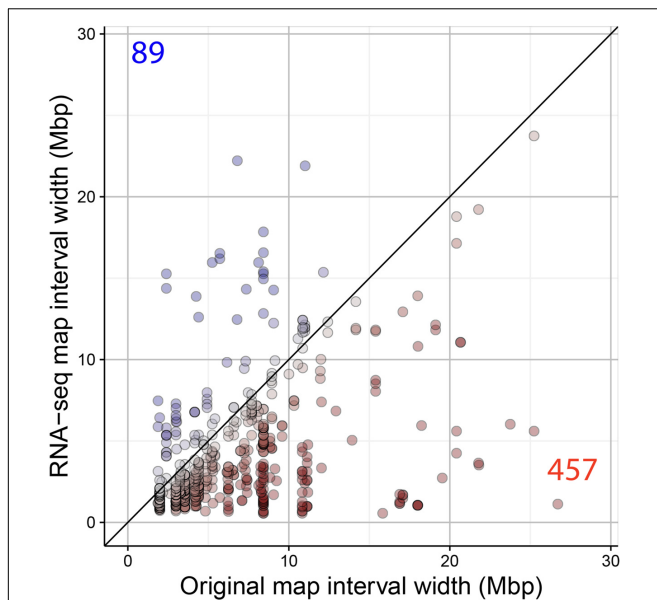
	A		
	Original	Pheno2geno	RNA-seq
<b>B</b>			
Original	–	1.15 (0.04)	1.74 (0.10)
Pheno2geno	1.45 (0.19)	–	1.66 (0.12)
RNA-seq	0.98 (0.04)	1.2 (0.04)	–

Numbers indicate the average LOD score difference for QTLs with a higher LOD score using map A as compared to map B. Standard errors are indicated in brackets. The numbers of QTLs used for the analysis are reported in **Table 3** (see higher and lower LOD QTLs).

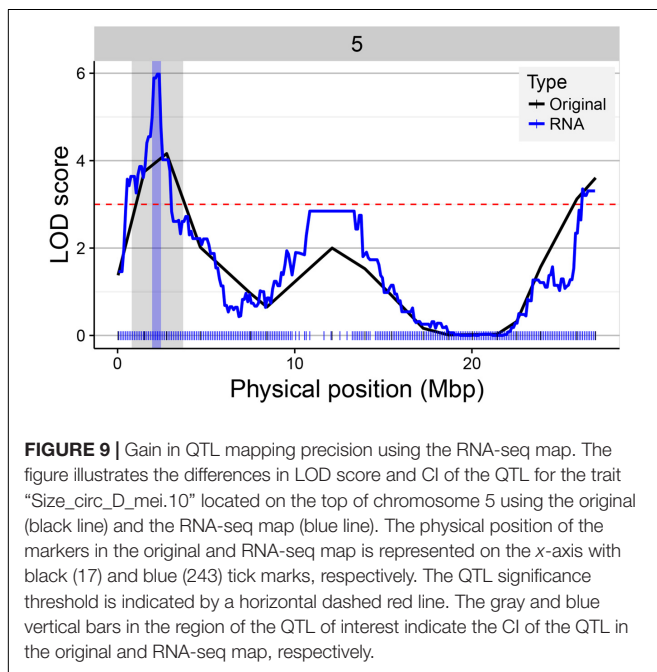
## DISCUSSION

### High-Density Genetic Map

In this study we showed that RNA-seq data can effectively be used for SNP calling, RIL genotyping, and the development of a high-density genetic linkage map. The used binning approach resulted in 1059 high-quality multi-SNP-based markers, providing a dense and equal coverage of markers physically anchored



**FIGURE 8 |** Comparison of the QTL mapping resolution using the original and the RNA-seq map. CIs (in Mbp) of QTLs detected in the original and the RNA-seq map are shown. Red and blue dots/values indicate the number of significant QTLs (LOD > 3) with reduced and increased CI in the RNA-seq map, respectively.



**FIGURE 9 |** Gain in QTL mapping precision using the RNA-seq map. The figure illustrates the differences in LOD score and CI of the QTL for the trait “Size\_circ\_D\_mel.10” located on the top of chromosome 5 using the original (black line) and the RNA-seq map (blue line). The physical position of the markers in the original and RNA-seq map is represented on the x-axis with black (17) and blue (243) tick marks, respectively. The QTL significance threshold is indicated by a horizontal dashed red line. The gray and blue vertical bars in the region of the QTL of interest indicate the CI of the QTL in the original and RNA-seq map, respectively.

to the genome. The high marker density enabled more precise identification of recombination breakpoints and revealed unknown recombination breakpoints within the RIL population (Table 2). As a result, the mapping resolution is no longer limited by the number of markers but rather depends on the number of recombination events captured by the mapping population. This means that the advantages of high-density genetic maps

in respect to mapping resolution will be considerably improved in combination with larger and/or more advanced designed populations (Balasubramanian et al., 2009; Kover et al., 2009; Liu et al., 2016). In comparison to the available genetic maps, the RNA-seq map could substantially increase QTLs linkage, eventually resulting in the identification of new QTLs (Table 3). Although the pheno2geno map showed a larger number of QTLs with higher LOD scores compared to the original map (Table 3), the RNA-seq map considerably increased the LOD scores of significant QTLs compared to both the original and the pheno2geno map (Table 4). Although we focussed in this study on the highest QTL per chromosome and per trait, we expect the RNA-seq map to also increase the overall number of QTLs after a more comprehensive analysis.

## Gain in QTL Mapping Resolution

The detection power and resolution of QTL mapping is significantly improved by high density genetic maps as compared to traditional markers (Yu et al., 2011). With the RNA-seq map, a major improvement was observed in the reduction of the LOD-1 CIs for 74% of the investigated QTLs. As a QTL CI in general encompasses a large number of genes, reduced CIs is of great benefit to narrow down the number of candidate genes for further investigation. In genetical genomics experiments, eQTLs can be identified as being either *cis*- or *trans*-regulated. Commonly, the distinction of both is made based on the distance, in cM or Mbp, between the gene and the eQTL peak or from the CI of the eQTL (Li et al., 2006, 2010; Keurentjes et al., 2007; West et al., 2007; Rockman et al., 2010; Terpstra et al., 2010; Vinuela et al., 2010; Aylor et al., 2011; Snoek et al., 2012, 2017; Lowry et al., 2013; Cubillos et al., 2014; King et al., 2014; Drost et al., 2015; Ranjan et al., 2016; Sterken et al., 2017). Therefore, gain in mapping precision is also likely to contribute to a more accurate identification of *cis*- versus *trans*-eQTLs.

## Advantages and Limitations of Using RNA-Seq Data

The use of RNA-seq presents several advantages over other methods. Our results show that RNA-seq data are a convenient and cost-effective source of SNP discovery, especially when a population is anyhow subjected to an eQTL analysis with the help of RNA-seq. RNA-seq can also overcome shortcomings identified from expression arrays based studies: while the effect of a SNP on the probe has enabled the identification of new sequence polymorphisms, weakened hybridization on microarrays based on expression studies can also cause the detection of false *cis*-eQTLs (Alberts et al., 2007; Chen et al., 2009). Furthermore, RNA-seq has the potential to study more complex levels of the genetic control of gene expression, for instance by quantification of alternative splicing (Filichkin et al., 2010; Yoo et al., 2016).

Single-nucleotide polymorphisms that are found with RNA-seq are inherently restricted to expressed exons, thus dependent on the developmental stage of the sequenced material and the experimental conditions. This restriction can also cause regions with low gene density or lowly expressed genes to be

under represented. However, these disadvantages will often not affect the mapping due to the high number of intermediate to highly expressed genes in any tissue and the SNPs present in those genes. Although our approach finds variants that affect protein-coding sequences, it is largely blind to SNPs in promoters, introns, and intergenic regions. However, SNPs that are causal for phenotypic variation will often be found in or close to genes and therefore, SNPs in large non-genic regions will hardly result in improvements of quantitative traits mapping (Li et al., 2012). In view of the abundance and saturation of SNPs that were discovered in this study, this does not cause a disadvantage, but might limit SNP detection for crosses from nearly identical parents.

## CONCLUSION

This study demonstrates that RNA-seq data can effectively be used for SNP discovery and the development of high-density genetic linkage maps. Here we provide a new SNP-based saturated genetic map for a Bay  $\times$  Sha RIL population. This saturated genetic map resulted in higher precision QTL mapping with more QTLs and considerably reducing the QTL CIs. Such improvements are of great benefit for the accurate mapping of more complex traits and the identification of causal genes.

## REFERENCES

- Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J. P., and Jansen, R. C. (2007). Sequence polymorphisms cause many false cis eQTLs. *PLOS ONE* 2:e622. doi: 10.1371/journal.pone.0000622
- Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., et al. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 481–491. doi: 10.1016/j.cell.2016.05.063
- Arends, D., Prins, P., Jansen, R. C., and Broman, K. W. (2010). R/qtl: high-throughput multiple QTL mapping. *Bioinformatics* 26, 2990–2992. doi: 10.1093/bioinformatics/btq565
- Aylor, D. L., Valdar, W., Foulds-Mathes, W., Buus, R. J., Verdugo, R. A., Baric, R. S., et al. (2011). Genetic analysis of complex traits in the emerging collaborative cross. *Genome Res.* 21, 1213–1222. doi: 10.1101/gr.111310.110
- Balasubramanian, S., Schwartz, C., Singh, A., Warthmann, N., Kim, M. C., Maloof, J. N., et al. (2009). QTL mapping in new *Arabidopsis thaliana* advanced intercross-recombinant inbred lines. *PLOS ONE* 4:e4318. doi: 10.1371/journal.pone.0004318
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890. doi: 10.1093/bioinformatics/btg112
- Chen, L., Page, G. P., Mehta, T., Feng, R., and Cui, X. (2009). Single nucleotide polymorphisms affect both cis- and trans-eQTLs. *Genomics* 93, 501–508. doi: 10.1016/j.ygeno.2009.01.011
- Cubillos, F. A., Stegle, O., Grondin, C., Canut, M., Tisne, S., Gy, I., et al. (2014). Extensive cis-regulatory variation robust to environmental perturbation in *Arabidopsis*. *Plant Cell* 26, 4298–4310. doi: 10.1105/tpc.114.130310
- Drost, D. R., Puranik, S., Novaes, E., Novaes, C. R. D. B., Dervinis, C., Gailing, O., et al. (2015). Genetical genomics of *Populus* leaf shape variation. *BMC Plant Biol.* 15:166. doi: 10.1186/s12870-015-0557-7
- Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., et al. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* 20, 45–58. doi: 10.1101/gr.093302.109

## AUTHOR CONTRIBUTIONS

WL designed the experiment. LW and ES grew the plants and provided the RNA samples. JJ-G processed the RNA samples and performed the RNA-seq analysis. HN processed the RNA-seq data and performed the SNP calling. LS developed and executed the binning and genotyping pipeline. ES analyzed the data and wrote the manuscript. HN, LS, JJ-G, HH, and WL read, revised, and agreed with the content of the manuscript.

## FUNDING

ES, HN, LW, and WL received financial support from the Dutch Technology Foundation (STW), which is part of the Netherlands Organization for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2017.00201/full#supplementary-material>

- He, H., De Souza Vidigal, D., Snoek, L. B., Schnabel, S., Nijveen, H., Hilhorst, H., et al. (2014). Interaction between parental environment and genotype affects plant and seed performance in *Arabidopsis*. *J. Exp. Bot.* 65, 6603–6615. doi: 10.1093/jxb/eru378
- Jansen, R. C., and Nap, J.-P. (2001). Genetical genomics: the added value from segregation. *Trends Genet.* 11, 388–391. doi: 10.1016/S0168-9525(01)02310-1
- Jimenez-Gomez, J. M. (2011). Next generation quantitative genetics in plants. *Front. Plant Sci.* 2:77. doi: 10.3389/fpls.2011.00077
- Jimenez-Gomez, J. M., Wallace, A. D., and Maloof, J. N. (2010). Network analysis identifies ELF3 as a QTL for the shade avoidance response in *Arabidopsis*. *PLOS Genet.* 6:e1001100. doi: 10.1371/journal.pgen.1001100
- Joosen, R. V., Arends, D., Willems, L. A., Ligterink, W., Jansen, R. C., and Hilhorst, H. W. (2012). Visualizing the genetic landscape of *Arabidopsis* seed performance. *Plant Physiol.* 158, 570–589. doi: 10.1104/pp.111.186676
- Joosen, R. V., Ligterink, W., Hilhorst, H. W., and Keurentjes, J. J. (2009). Advances in genetical genomics of plants. *Curr. Genomics* 10, 540–549. doi: 10.2174/138920209789503914
- Keurentjes, J. J., Fu, J., Terpstra, I. R., Garcia, J. M., Van Den Ackerveken, G., Snoek, L. B., et al. (2007). Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1708–1713. doi: 10.1073/pnas.0610429104
- Keurentjes, J. J., Koornneef, M., and Vreugdenhil, D. (2008). Quantitative genetics in the age of omics. *Curr. Opin. Plant Biol.* 11, 123–128. doi: 10.1016/j.pbi.2008.01.006
- Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317
- King, E. G., Sanderson, B. J., Mcneil, C. L., Long, A. D., and Macdonald, S. J. (2014). Genetic dissection of the *Drosophila melanogaster* female head transcriptome reveals widespread allelic heterogeneity. *PLOS Genet.* 10:e1004322. doi: 10.1371/journal.pgen.1004322
- Koornneef, M., Alonso-Blanco, C., and Vreugdenhil, D. (2004). Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu. Rev. Plant Biol.* 55, 141–172. doi: 10.1146/annurev.arplant.55.031903.141605

- Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., et al. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLOS Genet.* 5:e1000551. doi: 10.1371/journal.pgen.1000551
- Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The Arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210. doi: 10.1093/nar/gkr1090
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Li, S., Lu, Q., and Cui, Y. (2010). A systems biology approach for identifying novel pathway regulators in eQTL mapping. *J. Biopharm. Stat.* 20, 373–400. doi: 10.1080/10543400903572803
- Li, X., Zhu, C., Yeh, C. T., Wu, W., Takacs, E. M., Petsch, K. A., et al. (2012). Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genome Res.* 22, 2436–2444. doi: 10.1101/gr.140277.112
- Li, Y., Alvarez, O. A., Gutteling, E. W., Tijsterman, M., Fu, J., Riksen, J. A., et al. (2009). Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLOS Genet.* 2:e222. doi: 10.1371/journal.pgen.0020222
- Li, Y., Breitling, R., and Jansen, R. C. (2008). Generalizing genetical genomics: getting added value from environmental perturbation. *Trends Genet.* 24, 518–524. doi: 10.1016/j.tig.2008.08.001
- Liu, C., Zhou, Q., Dong, L., Wang, H., Liu, F., Weng, J., et al. (2016). Genetic architecture of the maize kernel row number revealed by combining QTL mapping using a high-density genetic map and bulked segregant RNA sequencing. *BMC Genomics* 17:915. doi: 10.1186/s12864-016-3240-y
- Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D., and Daniel-Vedele, F. (2002). Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theor. Appl. Genet.* 104, 1173–1184. doi: 10.1007/s00122-001-0825-9
- Lowry, D. B., Logan, T. L., Santuari, L., Hardtke, C. S., Richards, J. H., Derose-Wilson, L. J., et al. (2013). Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in Arabidopsis. *Plant Cell* 25, 3266–3279. doi: 10.1105/tpc.113.115352
- Markelz, R. J. C., Covington, M. F., Brock, M. T., Devisetty, U. K., Kliebenstein, D. J., Weinig, C., et al. (2017). Using RNA-seq for genomic scaffold placement, correcting assemblies, and genetic map creation in a common *Brassica rapa* mapping population. *G3* 7, 2259–2270. doi: 10.1534/g3.117.043000
- Nijveen, H., Ligterink, W., Keurentjes, J. J. B., Loudet, O., Long, J., Sterken, M. G., et al. (2017). AraQTL - workbench and archive for systems genetics in *Arabidopsis thaliana*. *Plant J.* 89, 1225–1235. doi: 10.1111/tpj.13457
- Piskol, R., Ramaswami, G., and Li, J. B. (2013). Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* 93, 641–651. doi: 10.1016/j.ajhg.2013.08.008
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Ranjan, A., Budke, J. M., Rowland, S. D., Chitwood, D. H., Kumar, R., Carriedo, L., et al. (2016). eQTL regulating transcript levels associated with diverse biological processes in tomato. *Plant Physiol.* 172, 328–340. doi: 10.1104/pp.16.00289
- Rockman, M. V., Skrovanek, S. S., and Kruglyak, L. (2010). Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* 330, 372–376. doi: 10.1126/science.1194208
- Salathia, N., Lee, H. N., Sangster, T. A., Morneau, K., Landry, C. R., Schellenberg, K., et al. (2007). Indel arrays: an affordable alternative for genotyping. *Plant J.* 51, 727–737. doi: 10.1111/j.1365-313X.2007.03194.x
- Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., et al. (2005). A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* 37, 501–506. doi: 10.1038/ng1543
- Snoek, L. B., Sterken, M. G., Bevers, R. P. J., Volkers, R. J. M., Van't Hof, A., Brenchley, R., et al. (2017). Contribution of trans regulatory eQTL to cryptic genetic variation in *C. elegans*. *BMC Genomics* 18:500. doi: 10.1186/s12864-017-3899-8
- Snoek, L. B., Terpstra, I. R., Dekter, R., Van Den Ackerveken, G., and Peeters, A. J. (2012). Genetical genomics reveals large scale genotype-by-environment interactions in *Arabidopsis thaliana*. *Front. Genet.* 3:317. doi: 10.3389/fgene.2012.00317
- Sterken, M. G., Van Bemmelen Van Der Laat, L., Riksen, J. A. G., Rodriguez, M., Schmid, T., Hajnal, A., et al. (2017). Ras/MAPK modifier loci revealed by eQTL in *C. elegans*. *G3* 7, 3185–3193. doi: 10.1534/g3.117.1120
- Terpstra, I. R., Snoek, L. B., Keurentjes, J. J., Peeters, A. J., and Van Den Ackerveken, G. (2010). Regulatory network identification by genetical genomics: signaling downstream of the Arabidopsis receptor-like kinase ERECTA. *Plant Physiol.* 154, 1067–1078. doi: 10.1104/pp.110.159996
- Vinuela, A., Snoek, L. B., Riksen, J. A., and Kammenga, J. E. (2010). Genome-wide gene expression regulation as a function of genotype and age in *C. elegans*. *Genome Res.* 20, 929–937. doi: 10.1101/gr.102160.109
- West, M. A., Kim, K., Kliebenstein, D. J., Van Leeuwen, H., Michelsmore, R. W., Doerge, R. W., et al. (2007). Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* 175, 1441–1450. doi: 10.1534/genetics.106.064972
- West, M. A., Van Leeuwen, H., Kozik, A., Kliebenstein, D. J., Doerge, R. W., St Clair, D. A., et al. (2006). High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. *Genome Res.* 16, 787–795. doi: 10.1101/gr.5011206
- Yoo, W., Kyung, S., Han, S., and Kim, S. (2016). Investigation of splicing quantitative trait loci in *Arabidopsis thaliana*. *Genomics Inform.* 14, 211–215. doi: 10.5808/GI.2016.14.4.211
- Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., et al. (2011). Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLOS ONE* 6:e17595. doi: 10.1371/journal.pone.0017595
- Zych, K., Li, Y., Van Der Velde, J. K., Joosen, R. V., Ligterink, W., Jansen, R. C., et al. (2015). Pheno2Geno - High-throughput generation of genetic markers and maps from molecular phenotypes for crosses between inbred strains. *BMC Bioinformatics* 16:51. doi: 10.1186/s12859-015-0475-6
- Zych, K., Snoek, L. B., Elvin, M., Rodriguez, M., Van Der Velde, K. J., Arends, D., et al. (2017). reGenotyper: detecting mislabeled samples in genetic data. *PLOS ONE* 12:e0171324. doi: 10.1371/journal.pone.0171324

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Serin, Snoek, Nijveen, Willems, Jiménez-Gómez, Hilhorst and Ligterink. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.