Title:

A complete chromosome substitution mapping panel reveals genome-wide epistasis in Arabidopsis

Authors:

Cris L. Wijnen[1†], Ramon Botet[1†], José van de Belt[1], Laurens Deurhof[1], Hans de Jong[1], C. Bastiaan de Snoo[2], Rob Dirks[2,3], Martin P. Boer[4], Fred A. van Eeuwijk[4], Erik Wijnker[1§], Joost J.B. Keurentjes[1§*]

[*] Correspondence to:

J.J.B. Keurentjes,

Droevendaalsesteeg 1,

6708 PB Wageningen,

The Netherlands,

+31317483149

joost.keurentjes@wur.nl

This PDF file includes:

Materials and Methods

Figs. S1 to S2

Tables S1 to S2

Material & Methods

Development of chromosome substitution lines

Chromosome substitution lines were obtained from crosses between inbred parental lines as previously described (*11*). In brief, semi-sterile Col-0 *RNAi:DMC1* transformed plants, that are impaired in crossover formation, were crossed with wild-type L*er* (CS20) plants to produce achiasmatic F_1 offspring. F_1 plants were then crossed to *GFP-TAILSWAP*, a haploid inducer line, to generate F_1-derived haploids and subsequently doubled haploids (*12*). A number of genotypes that were not obtained by the described approach were acquired by specific crosses between generated CSLs or between CSLs and parental lines, whether or not containing the *RNAi:DMC1* construct.

Confirmation of genotypes

Potential CSLs were genotyped with a set of 151 SNP markers using KASPar assays (Tables S3-S4). These markers covered about 120 Mbp of the total Arabidopsis genome. 95% of the marker intervals were smaller than 2.5 Mbp, which should be sufficient to detect incidental recombinant progeny. To exclude possible phenotypic effects of the *RNAi:DMC1* construct, the absence of the construct in the final selected CSLs was confirmed by additional PCR markers (*23*). During propagation we noted two CSLs ($Chr1^{Ler}/Chr2^{Col}/Chr3^{Ler}/Chr4^{Col}/Chr5^{Col}$ and $Chr1^{Ler}/Chr2^{Ler}/Chr3^{Ler}/Chr4^{Col}/Chr5^{Col}$) exhibiting high intra-line variation. Flow cytometry indicated occasional aneuploidy, suggesting that the plants still carried the RNAi-transgene. Data of these genotypes were excluded from further analyses and the CSLs in the panel were replaced by non-transgenic lines. Removal of these two lines during the data analyses caused non-significant allele frequency distortions of 3.3% at max. For all 32 genotypes of the complete CSL panel, construct-free CSLs are now available.

Development of near isogenic lines

Near isogenic lines were acquired by backcrossing the sCSLs to the recurrent parent and by crossing the resulting F_1 to the haploid inducer *GFP-TAILSWAP*. Since the F_1 were transgene-free, it allowed to obtain doubled haploid lines that recombined for a single chromosome. Subsequent genotyping was performed with part of the markers that were available for the confirmation of the CSL genotypes (Tables S3-S4). This allowed to fine-map regions with an approximate resolution of 5 cM (~2.5 Mbp).

Propagation

To avoid batch differences introduced by generating CSLs in different series of experiments all lines were first propagated simultaneously in a climate chamber. Seeds were sown on wet filter paper and placed at 4°C in the dark for four days to break residual dormancy and ensure uniform germination. After four days in the cold, plates were transferred to a climate cell at 20°C in the light. After two days, at radicle protrusion, germinating seeds were transferred to 4x4 cm Rockwool blocks in a climate cell set at long day conditions (16h/8h, 20°C/18°C, day/night). Relative humidity was set to 70% and watering was performed automatically with a Hyponex nutrient solution using a flooding system that bottom watered the Rockwool blocks. Five replicates per genotype were sown, and after germination, these were reduced to three well-established replicates. After two weeks of growth, single-leaf samples were taken for genotyping using three KASP-assays per chromosome (Table S4). In addition, a PCR for detecting the presence of the *RNAi:DMC1* construct was performed (*23*). Mature plants were dried and only a single plant was harvested per genotype, which served as the seed stock for the following mapping experiment or any further future experimentation.

Phenotyping experiment

The complete CSL panel was grown in twelve replicates in parallel with three replicates of 100 RILs (Tables S5-S6), obtained from the ABRC stock centre (https://abrc.osu.edu/). The handling of the seeds and growth conditions were similar to the propagation conditions, with the exception of short day growing conditions (8h/16h, 20°C/18°C, day/night). Plants were grown in a grid with equal distances between the positions of 12 rows x 60 columns. This grid was divided into three blocks of 12 x 20 each, which contained a single replicate of each RIL (1x100 lines) and four replicates for each CSL (4x32 lines) in a randomized complete block design. In a second separate experiment four replicates of each of the NILs (172 different genotypes in total; Tables S7-S8) segregating for chromosome 2 (37 genotypes with Col background and 39 with L*er* background) and chromosome 5 (45 genotypes with Col background and 51 with L*er* background) were grown in the same growth chamber under identical conditions. Here randomized complete blocks consisted of 12 x 30 positions that held two replicates of each genotype.

The number of days after planting at which the first flower opened was recorded (flowering time), at which time point the total length of the main inflorescence was measured (main stem length). Flowering time was corrected for germination date based on daily taken RGB-images by an automated camera system. The day at which the first green leaf could be detected was considered day zero. After three months, the experiment was terminated and plants not flowering by that time were considered outliers due to technical causes and removed from data analysis. Further outliers were determined by image analysis of individual plant growth performance and monitoring reports made during the experiment. Eventually, for most CSL genotypes at least ten replicates were analysed, with a few exceptions of which the CSL consisting of Chr1^{Ler}/Chr2^{Ler}/Chr3^{Col}/Chr4^{Ler}/Chr5^{Col} was most extreme with only four replicates (Table S9). For the NILs and RILs only genotypes for which at least two plants were available for each phenotype were included for data analyses (Tables S9-S10).

Statistical analyses

The phenotypic data of the RILs and the NILs was corrected for environmental effects using the R packages SpATS (*24*). The script was adapted to our experimental setup, where population and block were included as fixed terms in the model while genotype, row and column were in the random part of the model. The geno.decomp option of SpATS was used to allow for heterogeneous genetic variances for the different populations (respectively the CSLs and RILs in the first experiment and the four different NIL panels in the second experiment). With this model the best linear unbiased predictions (BLUPs) were obtained for the NILs and the RILs (Tables S11-S12).

The BLUPs of the NILs and RILs were used as input for the QTL analyses with linear mixed models in Genstat 18th edition. The 676 single feature polymorphism (SFP) markers for the RILs were obtained from previously published data (*25*). Markers with a physical distance of roughly 1 Mbp, corresponding to approximately 5 cM genetic distance in Arabidopsis, were selected (*26*). Genetic predictors between markers were calculated by interval mapping with a step size of 5 cM to bridge any large gaps. For the QTL analyses default settings were used, with minimum cofactor proximity of 50 cM, minimum separation for selected QTLs of 30 cM and Li and Ji threshold settings with genome wide significance levels of $\alpha = 0.05$ (*27*). Initially a single QTL model was fitted. QTLs of the initial analyses were included in the model as cofactors to test for additional QTLs. The QTLs detected and the –log10(p-values) of this composite interval mapping method are reported (Tables S13-S14). The support intervals were calculated as a drop of two units in the –log10(p-value) similar to a 2-LOD support interval.

The raw data of the CSLs was corrected for spatial trends with the SpATS R package, and the resulting spatial corrected raw data was used for further analyses (Table S15). Individual trait values were preferred over BLUPs for the analyses of the CSLs to increase the degrees of freedom. Either all (for the analyses of the complete CSL set) or a subset (all sCSLs or the sCSLs sharing a single recurrent parent) of the corrected raw data was analysed by applying a backward elimination approach in combination with a multiple linear regression model containing chromosome main effects and two- and three-way epistatic interactions (I).

$$(\mathrm{I}) \qquad y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \sum_{k=1}^{5} \sum_{l>k}^{5} b_{kl} x_{ik} x_{il} + \sum_{k=1}^{5} \sum_{l>k}^{5} \sum_{m>l}^{5} c_{klm} x_{ik} x_{il} x_{im} + \varepsilon_{ir}$$

where y_{ir} is the phenotype of genotype i in replicate r, μ is the overall mean, a_k is the additive effect for chromosome k, x_{ik} is an indicator variable, with $x_{ik} = 0$ $(x_{ik} = 1)$ if chromosome k for genotype i is L*er* (Col), b_{kl} are the effects for the two-way interactions between chromosomes k and l, c_{klm} are the effects of the three-way interactions between chromosomes k, l, and m, and ε_{ir} is the residual error for genotype i in replicate r.

To test three-way epistatic effects, the multiple linear regression model including all main, two- and three-way interactions (I) was compared with a model including main and two-way interactions (II) with backward selection of the AIC criterion using the stepAIC function of the MASS package (with α = 5.10^{-5} to correct for multiple testing) (*28*).

(II) $h_0\colon y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \sum_{k=1}^{5} \sum_{l>k}^{5} b_{kl} x_{ik} x_{il} + \varepsilon_{ir}$

A second step of parameter reduction was used to select the significant two-way interactions for the model with a similar significance threshold. Here, a model resulting from backward selection (IV) was compared to a model including only main effects (III):

(III) $h_0\colon y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \varepsilon_{ir}$

(IV) $h_1\colon y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \sum_{k=1}^{5} \sum_{l>k}^{5} b_{kl} x_{ik} x_{il} + \sum_{(k,l,m)\in S_3} c_{klm} x_{ik} x_{il} x_{im} + \varepsilon_{ir}$

Here S_3 represents the set of the earlier selected significant three-way interactions. Finally, the model including all significant two- and three-way interactions (VI) was tested versus a model consisting of only the mean and the residuals (V):

(V) $h_0\colon y_{ir} = \mu + \varepsilon_{ir}$

(VI) $h_1\colon y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + \sum_{(k,l)\in S_2} c_{kl} x_{ik} x_{il} + \sum_{(k,l,m)\in S_3} c_{klm} x_{ik} x_{il} x_{im} + \varepsilon_{ir}$

Here, S_2 in h_1 represents the significant two-way interaction terms that were selected in the previous round. This backward selection eventually resulted in a model that included all significant three- and two-way interactions and main effects and all terms underlying the significant interaction terms independent of their own significance according to the principal of marginality.

A similar approach was used for the analyses of the sCSLs were only the main effects model (III) was compared with a model consisting of only the mean and the residuals (V).

For the detection of interactions with the recurrent parental background (either Col or L*er*) all sCSLs were subjected to a similar backward selection procedure. Model (II) was adapted for chromosome x background interactions (VII) and compared with a model for main effects only (V) to test for significant interaction effects between the chromosomes and the background.

(VII) $h_1: y_{ir} = \mu + \sum_{k=1}^{5} a_k x_{ik} + b z_i + \sum_{k=1}^{5} c_k x_{ik} z_i + \varepsilon_{ir}$

Where b is the estimated background effect, z_i is an indicator variable, with $z_i = 0$ $(z_i = 1)$ if the background i is L*er* (Col), c_k are the effects for the interaction between chromosome k and the genetic background. Here significance thresholds were set to α = 1.10^{-3} to correct for multiple testing.

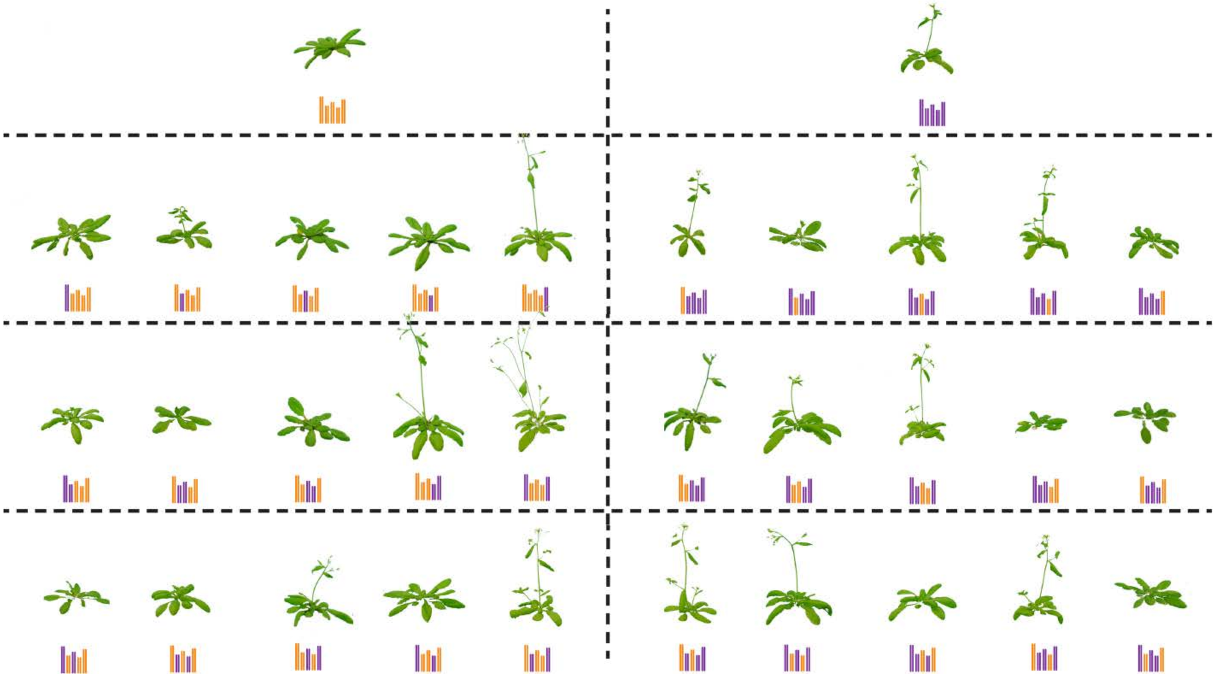

Fig. S1: Photographic presentation of phenotypic variation in a complete panel of CSLs. Each image depicts a representative phenotype of the genotype plotted below it. Arabidopsis genomes of each of the 32 CSLs are represented by five homozygous chromosomes derived from either the Col-0 (orange) or L*er* (purple) accession. Depicted plants are of identical age and images were taken at 23 days after sowing in long day (16h light) conditions.

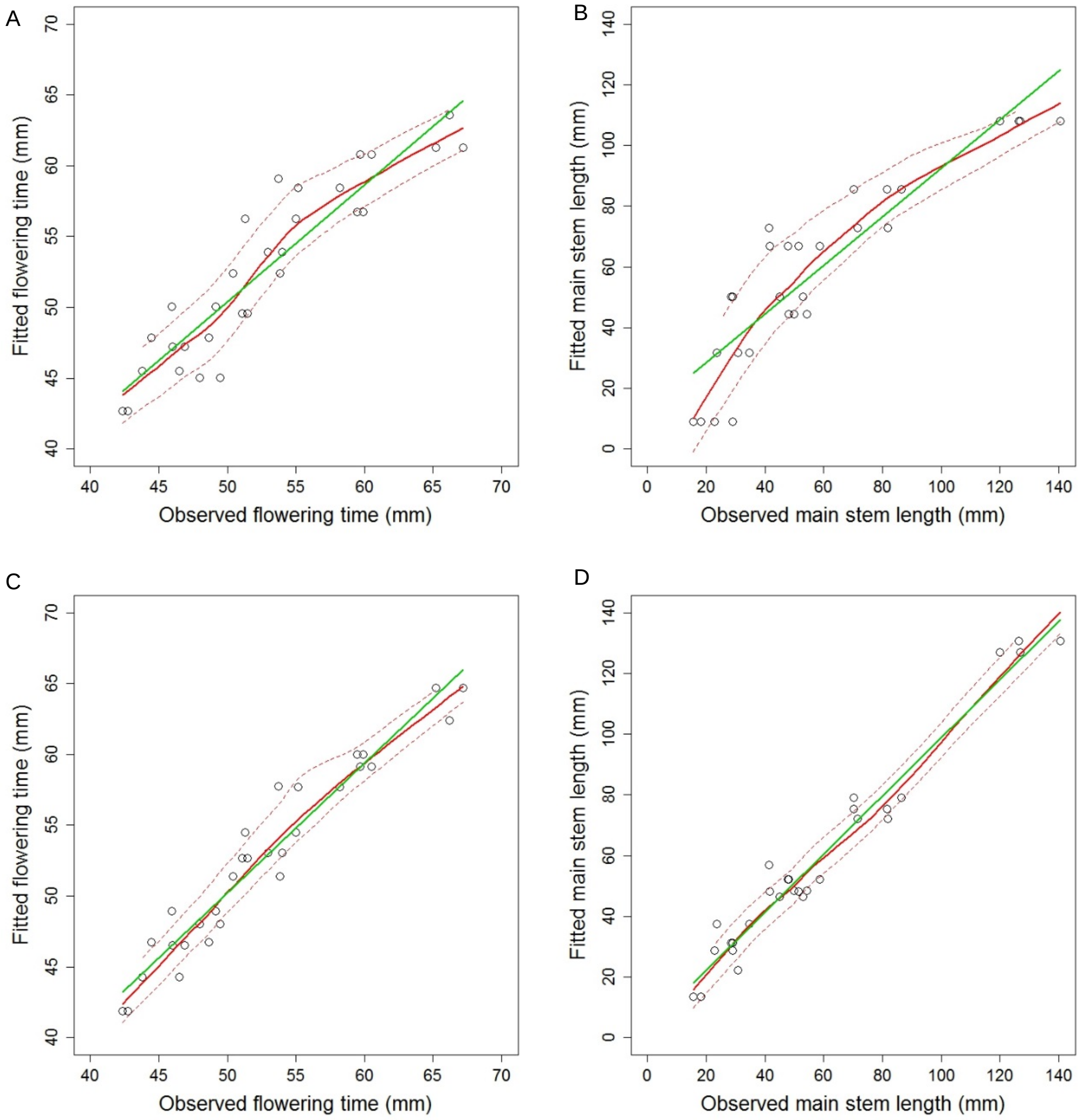

Fig. S2: Scatterplots of predicted trait values for models without and with interaction terms. The x-axis shows the observed mean phenotypic values of the CSls, and the y-axis the predicted values according to the corresponding model. A-B) Prediction of models without interaction terms for flowering time (FT ~ Chr1 + Chr2 + Chr3 + Chr5) and main stem length (MSL ~ Chr1 + Chr2 + Chr5), respectively. C-D) Prediction of the epistatic models for flowering time (FT ~ Chr1 + Chr2 + Chr3 + Chr5 + Chr1:Chr3 + Chr1:Chr5 + Chr3:Chr5) and main stem length (MSL ~ Chr1 + Chr2 + Chr3 + Chr5 + Chr1:Chr2 + Chr1:Chr5 + Chr2:Chr5 + Chr3:Chr5 + Chr1:Chr2:Chr5), respectively. For each plot, the linear regression is shown in green, while the red line represents a trend line of the data including a LOESS-confidence interval between the dashed lines.

Table S1: Detailed overview of main and interaction effects detected in CSL populations. For each significantly detected effect the trait for and population type in which it was detected is given. FT, flowering time; MSL, main stem length; Population type, sCSL: five sCSLs plus their indicated recurrent parent, sCSLs: all 10 sCSLs plus their recurrent parents, All CSLs: all 32 CSLs including the recurrent parental genotypes; Background genotype, the recurrent genotype for the sCSL populations; Chromosome number, the chromosomes for which main or interaction effects were detected, BG: effect of recurrent genotype in sCSL comparisons; Effect size, main effects: effect of background genotype or the substitution of a L*er* chromosome with a Col chromosome (ΔCol-L*er*), interaction effects: the average effect of the substitution of either one of the interacting chromosomes or background compared to the population mean, FT (days), MSL (mm); s.e., standard error of the effect size in the same units; Explained variance, proportion of the population variance explained by each effect; Significance, the significance of the effect.

Trait	Population type	Background genotype	Chromosome number	Effect size	s.e.	Explained variance (%)	Significance (P-value)
FT	sCSL	Col	II	5.4	0.9	2.33	8.7E-08
			III	-7.8	1.1	35.39	1.9E-09
			IV	4.3	0.9	0.05	9.7E-06
			V	12.7	0.9	49.79	< 2E-16
	sCSL	L*er*	I	-5.5	0.8	23.99	1.3E-08
			II	4.8	0.8	18.19	3.1E-08
			IV	-4.6	0.8	24.90	6.7E-08
			V	4.7	0.8	11.77	2.3E-07
	sCSLs		I	-3.6	0.7	1.83	1.3E-06
			II	4.7	0.7	27.66	2.8E-10
			III	-9.2	1.1	0.41	5.5E-14
			IV	2.8	0.9	0.80	2.2E-03
			V	11.2	0.9	41.02	< 2E-16
			BG	6.1	2.1	0.30	5.0E-03
			III:BG	10.3	1.2	11.55	3.2E-13
			IV:BG	6.8	1.1	2.51	7.0E-09
			V:BG	5.9	1.1	2.72	6.8E-07
	All CSLs		I	-1.5	0.6	0.76	9.5E-03
			II	4.7	0.3	10.81	< 2E-16
			III	-7.0	0.6	1.59	< 2E-16
			V	11.2	0.5	62.80	< 2E-16
			I:V	3.8	0.7	1.85	5.8E-08
			I:III	4.8	0.7	3.25	1.2E-11
			III:V	4.6	0.7	2.50	1.1E-10
MSL	sCSL	Col	I	-23.0	4.1	21.61	8.9E-07
			II	19.4	4.0	19.55	1.1E-05
			III	19.5	5.0	16.55	2.6E-04
			V	-23.0	4.0	16.24	4.5E-07
	sCSL	L*er*	II	88.7	4.0	86.48	< 2E-16
			V	-21.0	4.3	3.68	8.4E-06
	sCSLs		I	-14.5	3.3	10.67	2.5E-05
			II	21.6	4.3	34.27	1.6E-06
			III	21.7	5.4	3.32	1.0E-04
			V	-22.6	3.2	18.15	1.1E-10
			BG	-48.8	7.5	1.88	2.2E-09
			II:BG	63.5	5.6	16.88	< 2E-16
			III:BG	25.5	6.5	1.73	1.5E-04
	All CSLs		I	-25.8	3.5	13.18	2.5E-12
			II	17.5	3.2	33.85	7.3E-08
			III	15.2	2.6	0.00	6.7E-09
			V	-29.1	3.5	24.06	1.6E-15
			I:II	17.2	5.0	8.06	7.1E-04
			I:V	24.8	4.6	0.24	1.7E-07
			II:V	9.6	4.4	2.76	2.8E-02
			III:V	19.3	3.3	1.55	1.4E-08
			I:II:V	33.8	6.6	1.35	4.8E-07

Table S2: Detailed overview of the QTLs detected in RIL and NIL populations. For each significantly detected QTL the trait for and population type in which it was detected is given. FT, flowering time; MSL, main stem length; Background genotype, for the NILs the recurrent background is given, equal allele frequencies are assumed for RILs; Chromosome number, the chromosome on which the QTL was detected; Position, position on the chromosome where the strongest association was detected; Support interval, support intervals were calculated as a drop of two units in the –log10(p-value) surrounding the position of the most significant association; Effect size, effect of the homozygous substitution of a L*er* genotype with a Col genotype at the QTL (ΔCol-L*er*), FT (days), MSL (mm); s.e., standard error of the effect size in the same units; Explained variance, proportion of the total population variance explained by each QTL; Significance, the significance of the strongest association detected.

Trait	Population type	Background genotype	Chromosome number	Position (Mbp)	Support interval (Mbp)	Effect size	s.e.	Explained variance (%)	Significance (-log10(p))
FT	RILs	N.A.	I	23.8	22.2 - 24.1	-2.14	0.64	8.4	6.1
			II	11.2	10.1 - 12.4	-2.68	0.67	13.3	4.3
			II	18.3	15.3 - 19.5	4.18	0.63	32.4	4.2
	NILs	Col	V	8.0	7.3 - 8.8	7.38	0.65	78.2	28.8
	NILs	L*er*	V	8.8	8.0 - 9.7	4.47	0.88	39.7	6.4
MSL	RILs	N.A.	II	11.2	11.1 - 11.7	34.95	3.05	64.0	24.7
	NILs	Col	II	11.3	9.1 - 16.5	8.53	2.30	33.9	3.7
	NILs	L*er*	II	10.6	9.9 - 11.3	65.17	5.02	85.5	37.9