Genetic mapping of male pheromone response in the European corn borer identifies candidate genes regulating neurogenesis

Koutroumpa, F.A.; Groot, A.T.; Dekker, T.; Heckel, D.G.

DOI
10.1073/pnas.1610515113

Publication date
2016

Document Version
Other version

Published in
Proceedings of the National Academy of Sciences of the United States of America

Supporting Information

Koutroumpa et al. 10.1073/pnas.1610515113

SI Materials and Methods

Strategies for QTL Analysis of the Male Response. For the purpose of fine-scale genetic mapping, there are disadvantages in using any male response as a dichotomous trait. A single error in scoring that trait will produce a spurious recombinant in the linkage map, resulting in an inaccurate location of the *Resp* locus. Multiple errors in scoring the trait will inflate the estimated recombination rate in the region of the *Resp* locus, increase the inaccuracy of its location, and tend to bias the location of the *Resp* gene toward the end of a chromosome. When the behavioral trait is assumed to be scored without error, there is no way to estimate statistically the effects of possible scoring errors. The likelihoods of different gene orders can be compared, but the spurious recombinants will decrease the likelihood of the most likely order and will inflate the estimate of recombination rates. Therefore, we took great care to devise a quantitative score and used a discriminant function approach to discriminate more precisely between homozygote and heterozygote responses, as explained in detail below. Two behavioral characters were measured as a response to each of the pheromone blends (Z, H, and E): wing-fanning and hair pencil extrusion. For each blend, a behavioral score was calculated by weighing the behaviors as shown in Fig. S1. Each male was held within a tube with screening at both ends. The pheromone plume could pass through the tube, and the male could be observed at close range. This method allowed us to test many males in succession, greatly increasing the sample size in each generation. If males had been allowed to fly to the lure, recapture of the male within the wind tunnel would have been required, possibly resulting in damage and likely influencing the response to the next blend to be tested.

Optimization of Preference Score. The LOD function resulting from the preliminary QTL analysis exhibited a broad peak at 84.2 cM in the interval Q between *bgi00672* and *bgi12356*, with a maximum LOD score of 17.7, explaining 22% of the phenotypic variance in the behavioral score (Fig. S3*A*). We used this entire interval Q to optimize the weightings in the behavioral score. We identified four groups: 149 males homozygous for the Z allele at all marker loci in the interval (group A); 149 heterozygous males with one Z allele and one E allele at all marker loci in interval Q (group B); 73 males with at least one crossover in the interval, producing some homozygous and some heterozygous markers within interval Q (group C); and 99 males with missing data, such that none of the other conditions could be established with complete certainly (group D).

Unless there is a double crossover within one of the subintervals Q1, Q2, or Q3 (Fig. S3), group A males are homozygous for the entire interval Q of the Z-strain chromosome and therefore must be homozygous for the Z allele of the *Resp* locus. Double crossovers within one of the three subintervals would not be experimentally detectable, unless additional markers were scored within them, but we can estimate the probability of a double crossover as ≤0.05 (see below); thus overall at least 95% of group A males are expected not to show a double crossover and thus to be homozygous for the Z allele of the *Resp* locus. By the same reasoning, at least 95% of group B males are expected to be heterozygotes, with one Z allele and one E allele at the *Resp* locus. Therefore, independent of the actual position of the *Resp* gene within interval Q, we can take A and B as representative groups of *Resp* ZZ homozygotes and ZE heterozygotes, respectively, and can use these two groups in a discriminant function analysis. This analysis determines which combination of parameters in the formula for the behavioral score best discriminates between the two genotypes at the *Resp* locus.

The goal of discriminant analysis is to find the set of predictor variables that best discriminates between two types of observations described by a known class variable. Here the class variable is the genotype at the *Resp* locus (assumed to be known for groups A and B), and the predictor variables are the parameters used to calculate the behavioral score. We ask the following: Is the way we initially chose to measure the phenotypic score the best way to characterize the behavioral differences between the genotypes, and, if not, can we find a better way? Although classical discriminant function analysis assumes that the predictor variables are normally distributed, we make no such assumptions because we return to the original concept of the discriminant rule as defined by Fisher (84), namely to maximize the between-groups sum of squares, $SS_{between}$, relative to the within-groups sum of squares, SS_{within}. $SS_{between}$ is equivalent to the squared difference of the behavioral score means of groups A and B, and SS_{within} is related to the weighted average of the variance of the behavioral score within the two groups. We searched the space of parameters used to define the behavioral score over a regular grid and calculated the $SS_{between}/SS_{within}$ ratio (hereafter the "SS ratio") for each parameter set. We defined the optimized parameter set (Par13.3) as that which produced the largest value of this ratio. We found that the optimized parameter set increased the SS ratio from 84.2 to 137.5. QTL analysis using the same markers with the new behavioral score calculated for all the males (groups A–D) shifted the LOD peak 2 cM to the right, to 86.2 cM, increased the maximum LOD value from 17.7 to 26.1, and increased the fraction of phenotypic variance explained from 22 to 31%. We subsequently used the optimized parameter set to calculate the behavioral scores used for the final QTL analysis using more markers within interval Q.

We estimated the probability of an undetected double crossover within intervals Q1, Q2, or Q3 by taking each interval length (3.9, 5.9, or 26.3 cM), converting them to crossover probabilities (0.037, 0.055, or 0.204), squaring these, and adding them to produce a sum of 0.046 ~5%. Squaring the crossover probabilities assumes no interference; typically there is interference, and the frequency of double crossovers will be much smaller than 5%. Double crossovers within interval Q involving single crossovers in Q1 and Q2, or in Q1 and Q3, or in Q2 and Q3 will be more common, but these are detectable, and such males are classified into group C. The occurrence of undetected double crossovers in interval Q will affect only the assumption that group A males have no E allele at the *Resp* locus and that group B males have no Z allele at the *Resp* locus. If these assumptions are violated up to 5% of the time, the discriminant analysis will be only slightly less discriminating than optimal. We have seen that, even when suboptimal, the discriminant analysis approach greatly improved the peak LOD score and the fraction of variance explained.

Our subsequent analysis (see below) eliminated the OR cluster from serious consideration as a candidate for the *Resp* locus. However, the OR cluster had been included in the interval Q along with the *Resp* locus, so no bias against the OR cluster was introduced by excluding it from the discriminant analysis.

Evaluation of Peaks in the Likelihood Function. In addition to the major peak, the log-likelihood function resulting from the final QTL analysis showed a second peak at 108 cM with a maximum

LOD score of 20.5, flanked by the two markers *paraplegin* and *pdp1*. This peak is approximately 10^7 times less likely to be the *Resp* locus than is any position within the 2-LOD confidence interval, but it could represent a second QTL with a smaller effect. To examine this possibility, we conducted two-way fixed-effects ANOVA, using the genotype at *kon-tiki* (the marker nearest the major peak in the largest number of males genotyped) as one factor and the genotype at *paraplegin* or *pdp1* as the other factor. The dependent variable was the behavioral score for each individual. In each ANOVA, *kon-tiki* had a highly significant main effect on the behavioral score ($P < 0.0001$), but the main effect of the other marker and the interaction effect were not significant (Fig. S4). Thus, when the large effect of the main peak is accounted for, the overall contributions of markers near the minor peak are not significant. The presence of a peak rising above two nonsignificant markers may be caused by the contrasting marker effects within a given genotype class for *kon-tiki* (Fig. S4). For *kon-tiki* EZ heterozygotes the *pdp1* ZZ homozygote has a higher average score than the *pdp1* EZ heterozygote, but for *kon-tiki* ZZ homozygotes the *pdp1* ZZ and EZ scores are the same. For *kon-tiki* EZ heterozygotes the *paraplegin* ZZ and EZ are the same, but for *kon-tiki* ZZ homozygotes *paraplegin* ZZ is higher than *paraplegin* EZ. Therefore, possessing two copies of the Z allele at markers around the minor peak increases the tendency to prefer the Z blend in some but not all genetic backgrounds. If there is a second minor QTL under this peak, its effect depends on the genotype at the major QTL, and its contribution to the overall variance of the behavioral score is less than 10% of the contribution of the major QTL.

Evaluation of ORs as Candidate Genes. To test the possibility that the ORs represent a third, minor QTL, we conducted a two-way, fixed-effect ANOVA using the genotypes at *kon-tiki* and the ORs as factors (Fig. S5). The dependent variable was the behavioral score for each individual. As before, the main effect of *kon-tiki* was highly significant ($P < 0.0001$), but the main effect of the OR genotype and the interaction effect were both nonsignificant. The OR genotype had contrasting effects on the behavioral score, depending on the *kon-tiki* genotype. Within *kon-tiki* EZ genotypes, OR ZZ homozygotes had higher scores than OR ZE heterozygotes, but for *kon-tiki* ZZ homozygotes, OR ZZ homozygotes had lower scores than OR EZ heterozygotes (Fig. S5). Thus, sequence variation at the OR cluster makes no consistent contribution to male pheromone preference.

A. Starting parameters

	Day 1			Day 2 - data not used		
	Z	H	E	Z	H	E
A	2	1	1			
B	100	100	100			
C	60	60	60			
D	40	40	40			
F	220	220	220			
G			2			

TotalScoreDay1 = Zscore - Hscore - G*Escore (Day 1 parameters)

Distribution of scores, training set

B. Optimized parameters

	Day 1			Day 2		
	Z	H	E	Z	H	E
A	2	1	1	2	1	1
B	100	100	100	160	160	160
C	60	60	60	40	40	40
D	40	40	40	100	100	100
F	220	220	220	140	140	140
G			1.3			1

TotalScoreDay1 = Zscore - Hscore - G*Escore (Day 1 parameters)
TotalScoreDay2 = Zscore - Hscore - G*Escore (Day 2 parameters)

TotalScore = (TotalScoreDay1 + TotalScoreDay2)/2 if scored on both days
TotalScore = TotalScoreDay(day scored) if scored on only one day

Distribution of scores, training set

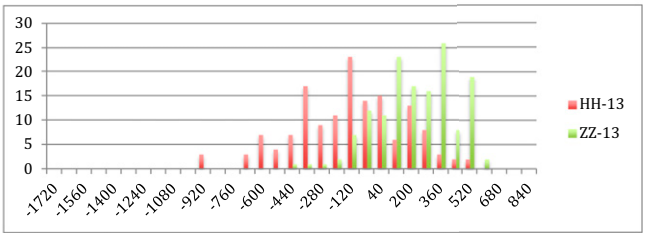


Fig. S1. Formula optimization for behavioral phenotype scoring. (*A*) Results for the starting parameter set. (*B*) Results for the optimized parameter set.

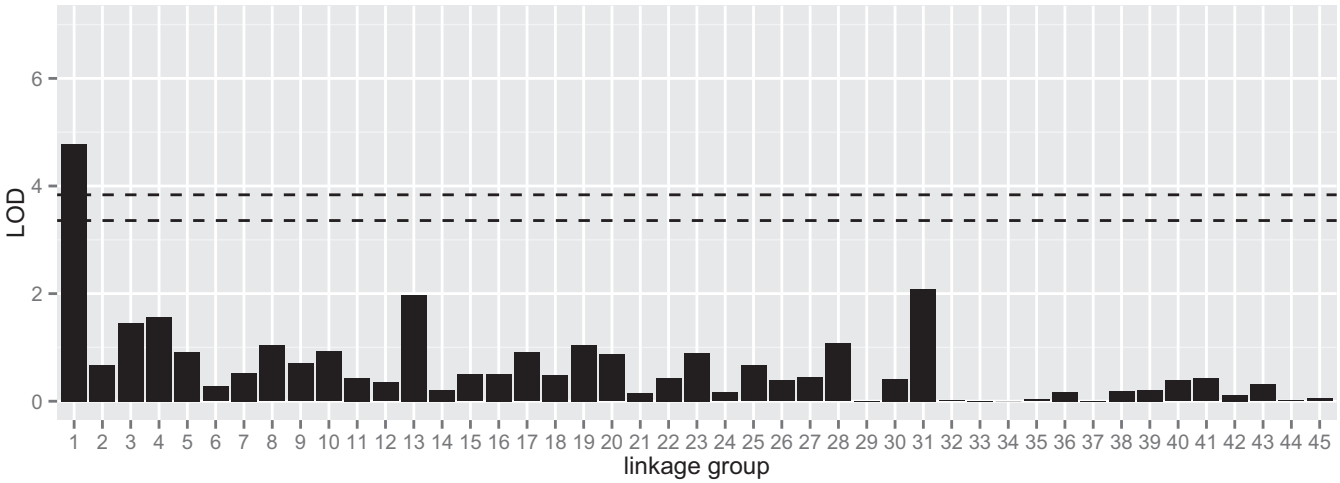


Fig. S2. Genome-wide QTL analysis with AFLPs to identify major linkage groups contributing to male behavioral response.

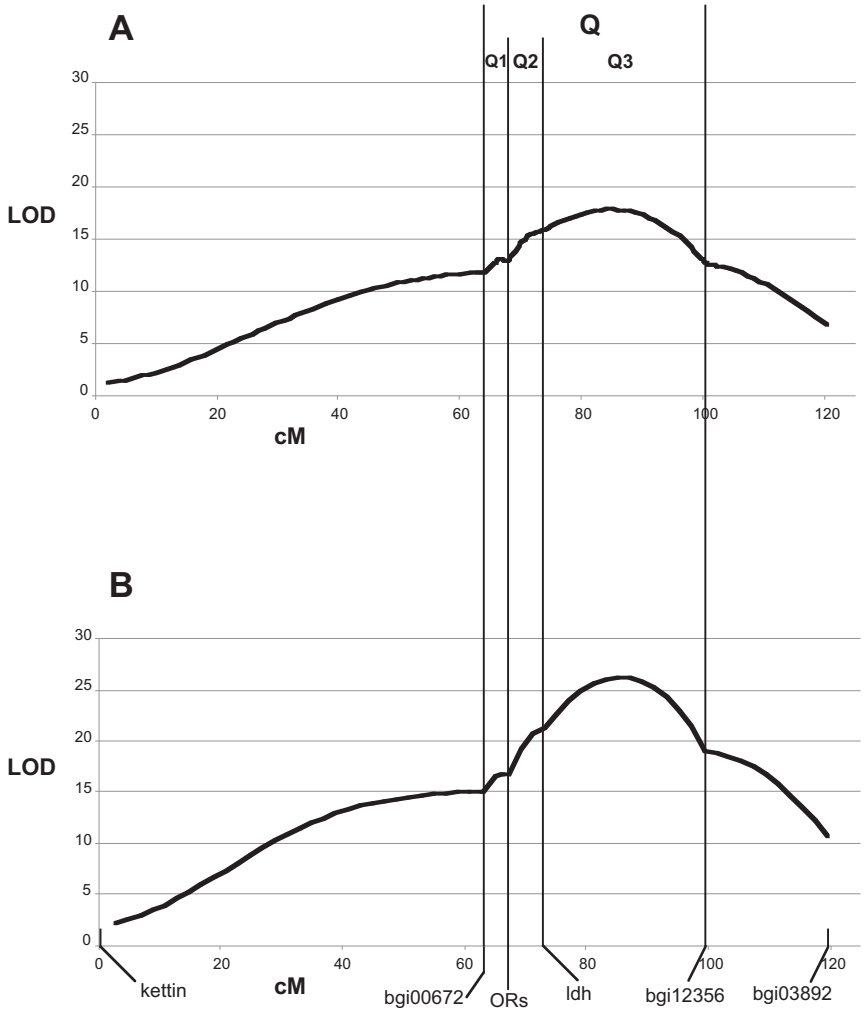


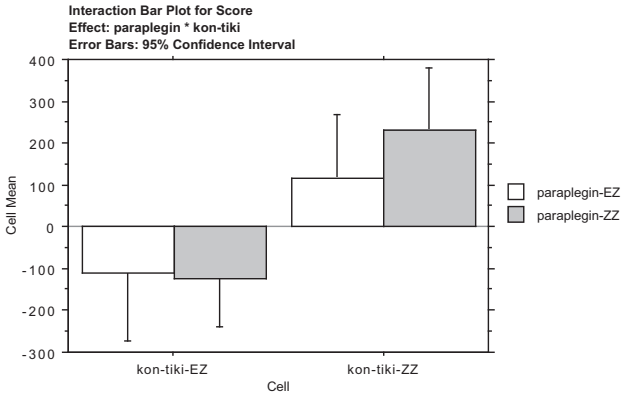
Fig. S3. LOD scores for the starting and optimized parameter sets. (*A*) The starting parameter set. (*B*) The optimized parameter set.

ANOVA Table for Score

	DF	Sum of Squares	Mean Square	F-Value	P-Value
paraplegin	1	35656.532	35656.532	.550	.4617
kon-tiki	1	1124206.621	1124206.621	17.339	.0001
paraplegin * kon-tiki	1	53108.625	53108.625	.819	.3697
Residual	51	3306599.282	64835.280		

Means Table for Score
Effect: paraplegin * kon-tiki

	Count	Mean	Std. Dev.	Std. Err.
paraplegin-EZ, kon-tiki-EZ	11	-111.845	242.281	73.050
paraplegin-EZ, kon-tiki-ZZ	18	115.878	305.304	71.961
paraplegin-ZZ, kon-tiki-EZ	14	-123.268	200.710	53.642
paraplegin-ZZ, kon-tiki-ZZ	12	230.938	235.742	68.053



ANOVA Table for Score

	DF	Sum of Squares	Mean Square	F-Value	P-Value
pdp1	1	27248.756	27248.756	.469	.4962
kon-tiki	1	948217.505	948217.505	16.305	.0001
pdp1 * kon-tiki	1	104903.980	104903.980	1.804	.1841
Residual	63	3663858.194	58156.479		

Means Table for Score
Effect: pdp1 * kon-tiki

	Count	Mean	Std. Dev.	Std. Err.
pdp1-EZ, kon-tiki-EZ	20	-141.268	205.313	45.909
pdp1-EZ, kon-tiki-ZZ	25	200.852	257.272	51.454
pdp1-ZZ, kon-tiki-EZ	9	-12.356	195.104	65.035
pdp1-ZZ, kon-tiki-ZZ	13	158.981	284.296	78.849

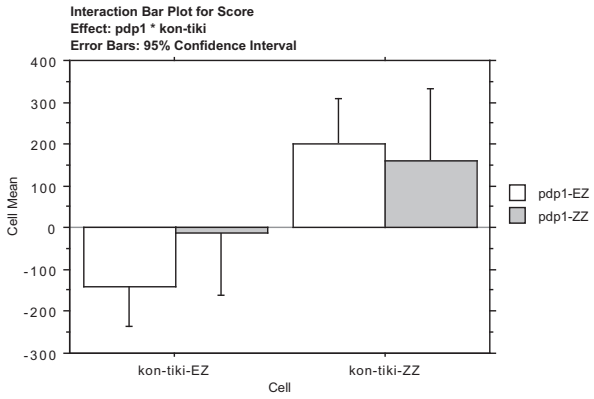


Fig. S4. Two-way fixed-effects ANOVAs on the preference score comparing the marker closest to the main QTL peak (*kon-tiki*) with two markers flanking the minor QTL peak (*paraplegin* and *pdp1*). Preference is strongly associated with the *kon-tiki* phenotype but not with the genotypes of the other markers or their interactions.

ANOVA Table for Score

	DF	Sum of Squares	Mean Square	F-Value	P-Value
OR	1	2533.877	2533.877	.039	.8440
kon-tiki	1	2703793.311	2703793.311	41.396	<.0001
OR * kon-tiki	1	150098.653	150098.653	2.298	.1304
Residual	360	23513510.772	65315.308		

Means Table for Score
Effect: OR * kon-tiki

	Count	Mean	Std. Dev.	Std. Err.
OR-EZ, kon-tiki-EZ	168	-160.664	284.489	21.949
OR-EZ, kon-tiki-ZZ	14	274.143	184.703	49.364
OR-ZZ, kon-tiki-EZ	10	-88.525	239.899	75.863
OR-ZZ, kon-tiki-ZZ	172	180.459	229.875	17.528

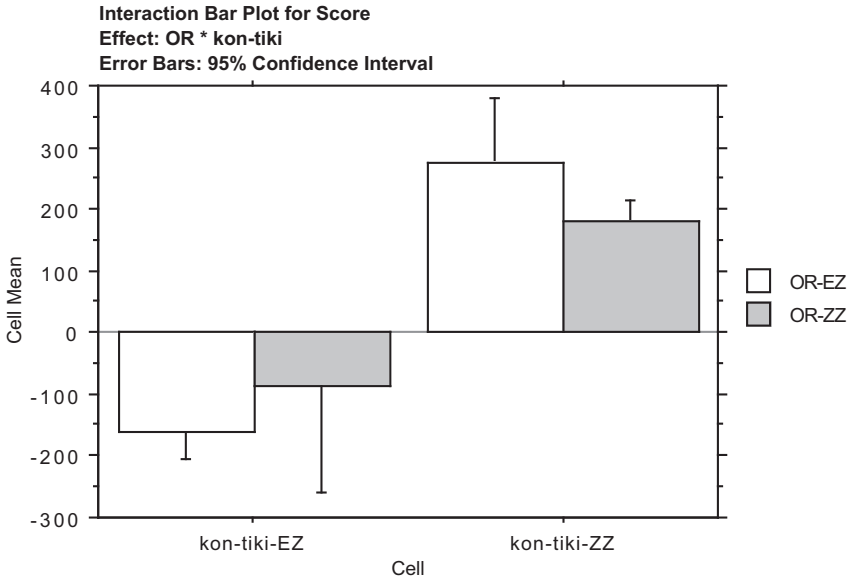


Fig. S5. Two-way fixed-effects ANOVA on the preference score comparing the marker closest to the major QTL peak (*kon-tiki*) with the cluster of OR genes. Preference is strongly affected by the *kon-tiki* genotype but not by the OR genotype or interaction.

Table S1. Primer sequences designed from multispecies sequence alignments

Gene	Forward primer	Reverse primer
Tpi	TGTGAACACTCTGAAAAAGGGTCC	TCCCAGTTGCTGCCAATAGC
	GCTATTGGCAGCAACTGGGA	ATGTCAACGAAYTCTGGCTTGAG
kettin	TGAAATCCCGGAACCAGTAACA	CAAAACCTGAGTAAAAATGGGCTT
	CCAGAAGGCATGAAGAAGATCCA	CAAGCTCCTATTTTCACTACTCACCTC
bgi00672	CTGAAAGAGTTCATCTGGAAGACG	CACTTTGTAGACGGCAGCCAC
G04	CAGCTACGGAGCTCCGTCCGC	GTATACGAGGGTCTTCTCTTC
G02	CCGAGCGTGCCGTTGC	GCTTCCGTGTGCTGGCG
	GCTCTTCCTTTCTCCGTGTGC	GCTTCCGTGTGCTGGCG
OnubOR4	CTGTTCCTGCTGCATAACG	TTCGCAGTAGCATGAAGTAG
OnubOR5	GTGCTGTTCCTGCTCTAC	CGTTCGCAAGAACATGAAG
OnubOR6	CGATACGGACCTTTGACTATGATCG	AAGCAGTTCGCTGGTTGCTGTTG
ldh	ACCTCCACCCACGGCTACAT	TCTCTTTCCAGTTCTCGGGGTC
highwire	GGCCAGGCNATGGTNATHAAYGA	CTTCAGGGTGGGNGTNGGRAARAANGC
	CATCAACGACGAGGTGGGCTTC	CAGGATGGAGTTGGTCAGCTCCAG
pentraxin	ACCATGCTGCARCTNTAYCAYGTNGC	TCCTTCTCGGGRTTRTCNCKYTTRTC
	ACCATGCTGCARCTNTAYCAYGTNGC	TGGCCGTTNACRTAYTCYTGYTG
	GCACACAAAGACCACAAACATCA	GACCGTTCACGTACTCTTGTTG
bgi12244	TTCCTGATCTCNCCNAARAARGARGC	CAGCTTCTGCCARTARAANGCYTCNAC
	TAGGGCTCGCAGCTTATCACT	TGACTTATTTCGTTGACTAGATCG
trol	CACAACGCCGGNGTNAAYATHAC	GGGGGGGCANGTRCAYTCYTCNAC
	ATCACGGACTTCGTCATGGAGTC	TCTCTGTTGGTNAGNGGRCANGGRCA
	CACGGACTTCGTCATGGAG	TAGATGGGTCGCCGTAGGT
kon-tiki	TGCGAGAAYATHACNGGNGTNATG	GATGAAGGGGGGNGANGCYTGDATYTC
	CAGACTATTCATCGGGTCGAGT	TCCTGAGAAAACTGAGCGATCT
bric-a-brac	TTCTGCCTGCGNTGGAAYAAYTAYCA	CTTCAAGTCGCAGTACTTTAT
	ATGTGGCARAARTGYTGGAAYAC	GCGCTGGGGCCNAGCATRTTRTGNAC
	ATGTGGCAGAAGTGTTGGAACAC	CTGGGGCCGAGCATGTTGTGGAC
CCR4-not	ATCTACCCCGAYGAYACNCARACNCC	TCCACCAGRTANAGRCANCGCCANGT
	GGCGGCCCNAAYTCNCCNTTYGCNATG	CGCACGTGCTCRCARTAYTTRTGRTA
	CCTTACAACACAACACGAACG	GGTAGTCTTTCAACCGCGACT
	ATCGCCAACGCNACNAAYATHGAYAC	TCGCTGAAGTTNGCDATNCCYTTRTC
	GCACTGGACCGCTTCAAGTC	GTTCGATGGACGCTCTCTTC
	GCCGTGCTNGCNACNGARAAYATGGA	GGCTTGGGDATRAACATNCCNGCRTC
	GCCGTGCTNGCNACNGARAAYATGGA	ARGCACTCNAGNAGNGTRTGCATRTT
	AGAGCCAAGGGNCARACNCCNAAYATG	TGGGTTCTTGATGAGTTCGAT
	AGAGCCAAGGGNCARACNCCNAAYATG	CGATGAATGTGATGAGAAGTCC
	CGACTTCGACTATGAGGGAAG	GAGGCAGCACGAGAAATAGTG
On-bgi012356	TGAGCATATCGTTCCCGACTG	TTCAAGACCAAGAGGGTGCTTC
paraplegin	GGCCCCCCCCGGNTGYGGNAARAC	CGGCTCTTCATNCCRTCCATYTCNAC
	CGCCCTGCATCATCTACATA	CGCGTACGATGGTCTGTTT
	CATCTACATAGACGAGATGGATGC	GTACGATGGTCTGTTTGTACCG
	CCCGGCMGNTTCGACMGNCACATCYT	TGTGATCTGTGAGCTGAGTTTGA
bgi03892	TCGTACCTCCCGTCGCAAAC	TGATGGCTTCCTCCATTTCCTT
	TGTTGACGCTGGAACTCCTCC	CTTCTAAGGCTAGGTCGTCGATCTC
pdp1	GAGGAGCTGAARCCNCARCCNATGAT	TCTCAGGGCGATYTGRTTYTCYTTCAT
	TAAACGCCGAGGAAGTAGTCG	TCCGATAGATTTGTGAACCTTCC