Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms

Andrew H. Paterson*, Eric S. Lander†‡, John D. Hewitt§, Susan Peterson*, Stephen E. Lincoln† & Steven E. Tanksley*

* Department of Plant Breeding and Biometry, Cornell University, Ithaca, New York 14853, USA
† Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA
‡ Harvard University, Cambridge, Massachusetts 02138, USA
§ Department of Vegetable Crops, University of California, Davis, California 95616, USA

The conflict between the Mendelian theory of particulate inheritance and the observation of continuous variation for most traits in nature was resolved in the early 1900s by the concept that quantitative traits can result from segregation of multiple genes, modified by environmental effects. Although pioneering experiments showed that linkage could occasionally be detected to such quantitative trait loci (QTLs), accurate and systematic mapping of QTLs has not been possible because the inheritance of an entire genome could not be studied with genetic markers. The use of restriction fragment length polymorphisms (RFLPs) has made such investigations possible, at least in principle. Here, we report the first use of a complete RFLP linkage map to resolve quantitative traits into discrete Mendelian factors, in an interspecific back-cross of tomato. Applying new analytical methods, we mapped at least six QTLs controlling fruit mass, four QTLs for the concentration of soluble solids and five QTLs for fruit pH. This approach is broadly applicable to the genetic dissection of quantitative inheritance of physiological, morphological and behavioural traits in any higher plant or animal.

The parents for the back-cross were the domestic tomato *Lycopersicon esculentum* cv. UC82B (donated E) and a wild South American green-fruited tomato *L. chmielewskii* accession LA1028 (donated CL). These strains have very different fruit masses (E vs. CL) and concentrations of soluble solids (E vs. CL) — traits of agricultural importance, because they jointly determine the yield of tomato paste. In addition, the strains are known to be polymorphic for genes affecting fruit pH, which is important for the optimal

preservation of tomato products; the difference in pH between the parental strains is, however, small.

A total of 237 back-cross plants, with E as the recurrent parent, were grown in the field at Davis, California. Between five and 20 fruit from each plant were assayed for fruit mass, soluble-solids concentration (*Brix*; see Fig. 1 legend for definition) and pH, each of which showed continuous variation (Fig. 1). Soluble-solids concentration correlated negatively with fruit mass (r = −0.42) and positively with pH (r = 0.33).

We had previously constructed a genetic linkage map of *tomato* with over 300 RFLPs and 20 isozyme markers, by analysing 46 F2 individuals derived from *L. esculentum* cv. VF36 x *L. pennellii* accession LA716 (E x P). The map is essentially complete: it has linkage groups covering all 12 tomato chromosomes with an average spacing of 5 cm between markers (1 cm is the distance along the chromosome which gives a recombination frequency of one percent). For QTL mapping, we selected a subset of markers spaced at approximately 20 cm intervals and displaying polymorphism between the E and CL strains. These included 63 RFLPs and five isozyme markers. In addition, the E and CL strains differ in two easily-scored, simply-inherited morphological traits: determinacy (described below) and uniform ripening, controlled by the *sp* and *u* genes, respectively. Although a few distal regions did not contain appropriate markers, we estimate that about 95% of the tomato genome was detectably linked to the markers used.

These 70 genetic markers were scored for each of the 237 E x CL back-cross progeny (as described in ref. 13), and a linkage map was constructed *de novo* using MAPMAKER. The map covers all 12 chromosomes with an average spacing of 14.3 cm. Although the linear order of markers inferred from the E x CL cross essentially agreed with that inferred from the E x P cross (but see Fig. 3 legend), genetic distances differed markedly in certain intervals (for example, 51 cm in E x P and 11 cm in E x CL, for the distance between the 45S ribosomal repeat and *TO1B* on chromosome 2). In total, the markers scored in both crossovers span 852 cm in the E x CL map versus 1103 cm in the E x P map, a highly significant (P < 0.01) difference. Skewed segregation (P < 0.05) was detected for 48 of the 70 markers, marking 21 distinct regions distributed over all 12 chromosomes. The heterozygote (E/CL) was over-abundant in 12 cases, whereas in nine cases the homozygote (E/E) was favoured. Overall, the effects of skewing approximately cancelled each other out: on average, the back-cross contained the expected 75% E genome (Fig. 2).

We then turned to the question of mapping the Mendelian factors that underly continuous variation in fruit mass, soluble-

---

**Fig. 1** Frequency distribution for fruit mass, soluble-solids concentration (*Brix*, a standard refractometric measure primarily detecting reducing sugars, but also affected by other soluble constituents; 1 °Brix is approximately 1% w/w) and pH in the E parental strain and in the back-cross (BC) progeny. The tomatoes were grown in the field at Davis, California, in a completely randomized design including 237 BC plants (with E as the recurrent parent), as well as E, CL and the F1 as controls. Neither CL nor the F1 progeny matured completely, as is typical in the central valley of California. Among the BC plants, six failed to mature and 12 produced too few fruit to reliably for quantitative traits. The absence of quantitative trait data for these few progeny should yield at most a slight bias in our analyses. Means and standard deviations for the distributions of the E parental strain (E filled bars) and the BC progeny (BC open bars) appear in the upper right of each histogram. The distributions for soluble-solids concentration and pH are approximately normal. The distribution of the BC progeny for fruit weight is clearly skewed; log _10_ (fruit mass) was studied throughout to achieve approximate normality (see ref. 5; E = 1.81 ± 0.07; BC = 1.20 ± 0.19). The proportion of variance due to environment was estimated to be the square of the ratio of the standard deviations (E/BC), for log-mass, solids and pH.
solids concentration and pH. The method of maximum likelihood and lod scores, commonly used in human linkage analysis, has recently been adapted to allow interval mapping of QTLs. At each position in the genome, one computes the 'most likely' phenotypic effect of a putative QTL affecting a trait (the effect which maximizes the likelihood of the observed data arising) and the odds ratio (the chance that the data would arise from a QTL with this effect divided by the chance that it would arise given no linked QTL). The lod score, defined as the $\log_{10}$ of the odds ratio, summarizes the strength of evidence in favour of the existence of a QTL with this effect at this position; if the lod score exceeds a pre-determined threshold, the presence of a QTL is inferred. The traditional approach to mapping QTLs involves standard linear regression, which accurately measures the effect of QTLs falling at marker loci only, underestimating the effects of other loci in proportion to the amount of recombination between marker and QTL. In contrast, interval mapping allows inference about points throughout the entire genome and avoids confounding phenotypic effects with recombination, by using information from flanking genetic markers. In the special case when a QTL falls exactly at a marker locus, interval mapping reduces to linear regression. A computer program, MAPMAKER-QTL, was written (S.E.L. and E.S.L., unpublished) to implement interval mapping.

Due to the large number of markers tested, an extremely high lod score threshold must be adopted to avoid false positives. Given the genetic length of the tomato genome and the density of markers used, a threshold of 2.4 gives a probability of less than 5% that even a single false positive will occur anywhere in the genome. This is approximately equivalent to requiring the significance level for any single test to be 0.001.

QTL likelihood maps, showing how lod scores for fruit mass, soluble-solids concentration and pH change as one moves along the genome, reveal multiple QTLs for each trait and estimate their location to within 20–30 cm (Fig. 3).

(1) Factors for fruit mass were found on six chromosomes (1, 4, 6, 7, 9 and 11). In each case, CL alleles decrease fruit mass (by 3.5 to 6.0 g), adding to a total reduction of 28.1 g for back-cross progeny carrying a CL allele at all six loci. This accounts for about half of the approximately 60 g difference between E and CL.

(2) Factors for soluble-solids concentration were found on four chromosomes (3, 4, 6 and 7). In each case, CL alleles elevate soluble-solids concentration (by 0.83 to 1.89 Brix), adding to a total of 4.57 Brix (versus a difference of ~3 Brix between the parental strains). This large effect in the back-cross is consistent with previous reports that high soluble-solids concentration exhibits dominance and overdominance. The QTL alleles for both fruit mass and soluble-solids concentration all produce effects in the direction predicted by the difference between the parental strains.

(3) Factors for pH were found on five chromosomes (3, 6, 7, 8 and 10). In addition, the lod score for a putative QTL on chromosome 9 fell just below our threshold. Because the parental strains do not differ greatly in pH, we suspected that CL alleles might not all produce effects in the same direction. In fact, pH was increased by four QTLs along the likely QTL on chromosome 9. This provides a genetic explanation for the observation that many back-cross progeny exhibited more extreme phenotypes than the parental strains (Fig. 1), a phenomenon known as transgression.

Together, the QTLs identified for fruit mass, soluble solids and pH account for 58%, 44% and 48%, respectively, of the phenotypic variance among the back-cross progeny, with another 13%, 9% and 11% attributable to environment.

The numbers of QTLs reported for each trait must be considered a minimum estimate. Because an extremely stringent threshold was used to avoid any false positives, some sub-threshold effects probably represent real QTLs. For example, the regions near TGI9 on chromosome 1, CDA1 on chromosome 5 and TGO8 on chromosome 12 may affect soluble-solids concentration and merit further attention in larger populations. Similarly, the region near the h locus on chromosome 10 may contain an additional QTL affecting pH (see Fig. 3 legend). Moreover, we cannot rule out the presence of many additional QTLs with tiny phenotypic effects—postulated in evolutionary theory and supported by some experimental evidence. Also, it is conceivable that some of our apparent QTLs actually represent several closely-linked QTLs, each with small phenotypic effects in the same direction—a phenomenon that might arise particularly in regions of genetic map compression. Finally, we should emphasize that the QTL mapping here strictly applies only to the specific environment tested and to heterozygosity for CL alleles. In principle, homozygosity for CL alleles could have been studied by using an F2 self between E and CL, but in practice too many of the progeny are sterile.

Some regions of the genome clearly exert effects on more than one trait (for example, chromosome 6; Fig. 3), providing a genetic explanation for at least some of the correlation between the traits. Although the present data are insufficient to distinguish between pleiotropic effects of a single gene and independent effects of tightly-linked loci, the frequent coincidence of QTL locations for different traits makes it likely that at least some of the effects are due to pleiotropy.

The region near sp on chromosome 6 has the largest effects on soluble solids and pH, as well as a substantial effect on fruit mass. The sp gene affects plant-growth habit: the dominant CL allele causes continuous apical growth (indeterminate habit), whereas the recessive E allele causes termination in an inflorescence ('determinate' or 'self-pruning' habit). Although indeterminacy has been reported previously to elevate both fruit mass and soluble-solids concentration within L. esculentum, we associated it with reduced fruit mass in both E X CL and another interspecific cross (E X L. cheesmani; A.H.P., S. Damon, J.D.H. and S.D.T., unpublished data). These differing results might be due to a second, tightly-linked locus or to unrelated modifier genes.

Overall, pairwise epistatic interactions between intervals were not common (about 5% of two-way analysis-of-variance tests were significant at 0.05). An interesting exception was the region near TGI6 on chromosome 8, at which the CL allele significantly enhanced the effect of three of the four QTLs for soluble-solids concentration. TGI6 also showed the most extreme segregation distortion of any marker scored (about 4:1 in favour of the E/E
Fig. 3 QTL likelihood maps indicating lod scores for fruit mass (solid lines and bars), soluble-solids concentration (dotted lines and bars) and pH (hatched lines and bars), throughout the 862 cM spanned by the 70 genetic markers. The RFLP linkage map used in the analysis is presented along the abscissa, in Kosambi cM. (The order of the markers agrees with the previously published map of the E × P cross, except for three inversions of adjacent markers: (TG24-CD15), (TG63-CD32B) and (TG30-TG36). In the first case, re-analysis of the E × P data with MAPMAKER indicates that the order shown here is the more likely in both E × P and E × CL. For the other two, the orders shown here are more likely in E × CL by odds of 10^3.1 and 10^3.1, but the inverse is more likely in E × P by 11:1 and 8:1 odds. These differences will be investigated in a larger E × P population.) Soluble-solids concentration and pH were analysed in "Brix" and pH units, respectively; allele effects on fruit mass are presented in g; log-transformation of fruit mass was used in all analyses to achieve approximate normality. The maximum likelihood effect of a putative QTL, as well as the lod score in favour of the existence of such a QTL, have been determined at points spaced every 1 cM throughout the genome, according to Lander and Botstein and a smooth curve plotted through the points. The height of the curve indicates the strength of the evidence (log_{10} of the odds ratio) for the presence of a QTL at each location—not the magnitude of the inferred allelic effect. The horizontal line at a height of 2.4 indicates the stringent threshold that the lod score must cross to allow the presence of a QTL to be inferred (see text). Information about the likely position of the QTL can also be inferred from the curve. The maximum likelihood position of the QTL is the highest point on the curve. Bars below each graph indicate a 10:1 likelihood support interval for the position of the QTL (the range outside which the likelihood falls by a lod score of 1.0), whereas the lines extending out from the bars indicate a 100:1 support interval. Phenotypic effects indicated beside the bars are the inferred effect of substituting a single CL allele for one of the two E alleles at the QTL. Several regions show sub-threshold effects on one or more traits (chromosome one near TG19, chromosome five near TG32 and chromosome 12 near TG68) which may represent QTLs but this requires additional testing. The region near TG68 may be particularly interesting, as it is the only instance found where the CL allele seems to decrease soluble-solids concentration (by about 0.7 "Brix").

In the case of chromosome 10, the lod score for pH crosses the significance threshold in two places. Controlling for the presence of a QTL near CD34A, we tested for the presence of a second QTL near u (by comparing the maximum lod scores assuming the presence of only the first QTL to the maximum lod score assuming the presence of two QTLs). Allowing for a QTL in the region of CD34A, the residual lod score near u falls below the required threshold. Thus, the evidence is not yet sufficient to support the presence of a QTL near u.

Methods. The lod score and the maximum likelihood estimate (MLE) of the phenotypic effect at any point in the genome is computed assuming that the distribution of phenotypes in the BC progeny represents a mixture of two normal distributions (of equal variance) with means depending on the genotype at a putative QTL at the given position. (Note that QTLs are considered individually and thus we did not assume that different QTL effects can be added—except in studying the possibility of two QTLs on chromosome 10 affecting pH.) Specifically, at a given position in the genome, the likelihood function for individual i with quantitative phenotype y_i is given by
\[ L(\alpha, \sigma^2) = \frac{\exp \left( -\frac{y_i - \alpha}{2\sigma^2} \right) + p_1 \exp \left( -\frac{y_i - \alpha - \sigma^2}{2\sigma^2} \right)}{2p_2}, \]
where \( \alpha \) is the effect of substituting a CL allele for an E allele at a putative QTL in the given position, \( \sigma^2 \) is the phenotypic variance not attributable to the QTL and \( p_1 \) and \( p_2 \) are the probabilities that individual i has genotype E/E and E/CL respectively, at the QTL (which can be computed on the basis of the genotypes at the flanking markers and the distance to the flanking markers). The likelihood function for the entire population is
\[ L = \prod L_i(\alpha, \sigma^2). \]
Also \( \alpha^* \) and \( \sigma^* \) denote the MLEs allowing the possibility of a QTL at the location (the values which maximize L) and \( \sigma^{**} \) denotes the MLE of \( \sigma \), subject to the constraint that no QTL is linked (\( \alpha = 0 \)). The lod score is then given by
\[ \log_{10} L(\alpha^*, \sigma^{**})/L(0, \sigma^{**}) \]. This method for QTL mapping is developed more fully in ref. 16.

The QTLs identified here may well differ from those that would be fixed by repeated back-crossing with continuing selection for a trait, a classical method for introgressing quantitative traits. Work on LA1563, a strain with increased soluble solids produced through back-crossing a different strain of E to CL has provided some suggestive evidence. By surveying RFLPs, Tanksley and Hewitt recently found that LA1563 has maintained three separate regions from CL: near CD56 on chromosome 10, near Got2 on chromosome 7 and near TG13 on chromosome 7. Here, we detected above-threshold effects in homozygotes and in a region known to exhibit skewed segregation in back-crosses to other green-fruited tomato species. The unusual properties of this region of CL clearly merit further study.
the last of these three regions only (which, interestingly, failed to show effects on soluble solids in a single-environment test by Tankley and Hewitt). Moreover, we detected QTLs affecting soluble-solids concentration in regions that did not seem to be retained. Unfortunately, the results of the two experiments are not directly comparable due to the use of a different E strain by Rick, possible environmental differences between the experiments, the possibility that small CL fragments containing QTLs went undetected in LA1563, the possibility that the region near TGI3 retained in LA1563 may not contain the QTL we detected here and the possibility that some of our sub-threshold effects are real. Although more detailed studies are clearly needed, it is interesting to speculate about why repeated back-crossing may fix a narrower class of QTLs than found by QTL mapping. Because such breeding programs demand horticultural acceptability, they are likely to select against otherwise-desirable QTLs which are closely linked to undesirable effects from the wild parent. If such QTLs can first be identified by mapping, it may be feasible to remove linked deleterious effects by recombination.

Having mapped several QTLs with relatively large effects, we are now making crosses to isolate them in near-isogenic lines. These lines will be used to characterize the QTLs in various dosages, genetic backgrounds, environments and combinations. By re-assembling selected CL alleles in an otherwise E genotype, we hope to engineer an agriculturally-useful tomato with a higher yield of soluble solids.

The general approach of QTL mapping is broadly applicable to a wide range of biological endeavours. In agriculture, it might be desirable to transfer to domestic strains many quantitative traits harboured in wild species, including resistance to diseases and pests, tolerance to drought, heat, cold and other adverse conditions, efficient use of resources and high nutritional quality. In mammalian physiology, selective breeding has generated rodent strains which differ greatly in quantitative traits such as hypertension, atherosclerosis, diabetes, predispositions to cancer, drug sensitivities and various behavioural patterns; information on the number, location and nature of these QTLs would be of value in medicine.

In evolutionary biology, the process of speciation can be investigated by studying the number and nature of genes underlying reproductive isolation. The availability of detailed RFLP linkage maps makes it possible to dissect quantitative traits into discrete genetic factors (QTLs): all regions of a genome can be assayed and accurate estimates of phenotypic effects and genetic position derived from interval analysis. Once QTLs are mapped, RFLP markers permit genetic manipulations such as rapid construction of near-isogenic lines: flanking manipulations may be used to retain the QTL and the study of the remaining markers may be used to speed progress by identifying individuals with a fortuitously high proportion of the desired genetic background (Fig. 2 legend). Using isogenic lines, the fundamental tools of genetics and molecular biology may be brought to bear on the study of QTLs— including testing of complementation, dominance and epistasis; characterization of physiological and biochemical differences between isogenic lines; isolation of additional alleles by mutagenesis (at least in favourable systems); and, eventually, physical mapping and molecular cloning of genetic factors underlying quantitative traits.

We thank Janice Chen, Mark Daly, Gerald Dickinson and Mitzi Aguirre for technical assistance. We thank our colleagues and the referees for comments on the manuscript. This work was supported in part by grants from the NSF (to E.S.L. and to S.D.T.), from the System Development Foundation (to E.S.L.), and from the US Department of Agriculture (to S.D.T.).

Received 8 July; accepted 9 September 1988.