

# The extent of linkage disequilibrium in *Arabidopsis thaliana*

Magnus Nordborg<sup>1</sup>, Justin O. Borevitz<sup>2</sup>, Joy Bergelson<sup>3</sup>, Charles C. Berry<sup>4</sup>, Joanne Chory<sup>2,5</sup>, Jenny Hagenblad<sup>6</sup>, Martin Kreitman<sup>3</sup>, Julin N. Maloof<sup>2</sup>, Tina Noyes<sup>7</sup>, Peter J. Oefner<sup>7</sup>, Eli A. Stahl<sup>3</sup> & Detlef Weigel<sup>2,8</sup>

Published online: 7 January 2002, DOI: 10.1038/ng813

Linkage disequilibrium (LD), the nonrandom occurrence of alleles in haplotypes, has long been of interest to population geneticists. Recently, the rapidly increasing availability of genomic polymorphism data has fueled interest in LD as a tool for fine-scale mapping, in particular for human disease loci<sup>1</sup>. The chromosomal extent of LD is crucial in this context, because it determines how dense a map must be for associations to be detected and, conversely, limits how finely loci may be mapped<sup>2</sup>. *Arabidopsis thaliana* is expected to harbor unusually extensive LD because of its high degree of selfing<sup>3</sup>. Several polymorphism studies have found very strong LD within individual loci, but also evidence of some recombination<sup>4–6</sup>. Here we investigate the pattern of LD on a genomic scale and show that in global samples, LD decays within approximately 1 cM, or 250 kb. We also show that LD in local populations may be much stronger than that of global populations, presumably as a result of founder events. The combination of a relatively high level of polymorphism and extensive haplotype structure bodes well for developing a genome-wide LD map in *A. thaliana*.

Population genetics theory tells us that we should expect LD to vary greatly over the genome simply because of random history. Nevertheless, some general patterns are expected. Most notably, LD will tend to decay with distance, because historical recombination events will have made alleles at linked loci occur independently of each other. The average rate of decay will depend on the demographic history of the population as well as several other parameters. A high degree of inbreeding, such as results from selfing, can have a particularly strong effect. This is most easily seen by considering the extreme case of a completely selfing species in which recombination still occurs but no longer has any effect on LD. In partial-selfers, LD still decays with distance, although much slower than in comparable outcrossing species<sup>3</sup>.

*A. thaliana* is believed to be at least 99% selfing<sup>7–8</sup>. In addition, it is characterized by a patchy distribution of highly inbred populations<sup>7–9</sup>. LD should therefore be extensive in this species. Indeed, attempts have been made to estimate the 'phylogeny' of the various accessions (ecotypes) of *A. thaliana*<sup>10,11</sup>, the notion of which presumes little or no recombination between these accessions. As would be expected, however, given that polymorphism studies show that individual loci have experienced at least some recombination<sup>4–6</sup>, it is now clear that no such phylogeny exists<sup>12</sup>. What is not clear is the genomic scale over which recombination causes LD to decay.

To study the decay of LD, we sequenced 13 short segments from a 250-kb region surrounding the flowering time locus *FRI* in a global sample of 20 accessions. We found (Fig. 1) that LD decays strongly with distance in this region, which corresponds to roughly 1 cM (see Methods).

To investigate whether the rapid decay of LD observed in this region was representative of the genome, we analyzed 163 genome-wide single-nucleotide polymorphisms (SNPs)<sup>13</sup> typed in a global sample of 76 accessions. There is a very weak relationship between LD and distance in these data: the genome-wide markers are evidently sufficiently far apart for recombination to have broken up most associations (Fig. 2). This is in agreement with the data shown in Fig. 1, because only very few pairs of the SNP markers are as closely spaced as maximum spacing in the *FRI* region. The one exception is a region on chromosome 4 with an unusually dense marker distribution, but LD in this region seems to decay even more quickly than in the *FRI* region, perhaps within 50 kb (Fig. 2, bottom right plot). The difference between these regions could be due to many factors: chance (random history), different samples, differences in the local recombination rate, and selection. The

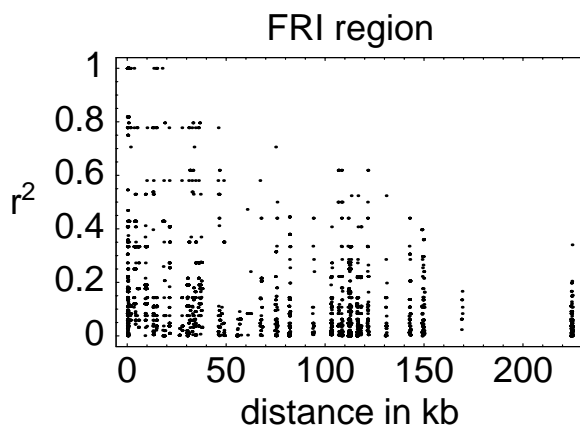
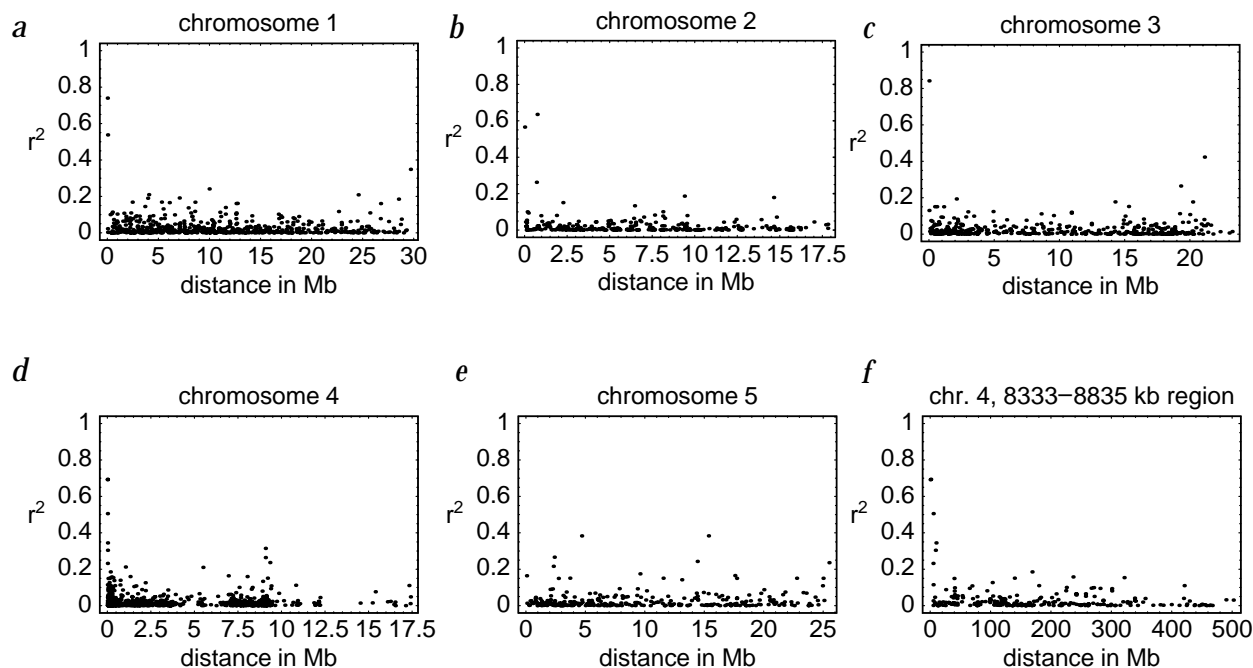


Fig. 1 LD as a function of distance. Plot shows LD in an approximately 250-kb (~1 cM) region on chromosome 4. The data consist of 83 polymorphic sites (from 20 individuals) in 13 widely separated segments of 0.5–1 kb; thus, the distribution of distances between pairs of markers is highly nonuniform.

<sup>1</sup>Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA. <sup>2</sup>Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, California, USA. <sup>3</sup>Department of Ecology & Evolution, University of Chicago, Chicago, Illinois, USA. <sup>4</sup>Department of Family & Preventive Medicine, University of California, San Diego, La Jolla, California, USA. <sup>5</sup>The Howard Hughes Medical Institute, La Jolla, California, USA.

<sup>6</sup>Department of Genetics, Lund University, Lund, Sweden. <sup>7</sup>Department of Biochemistry, Stanford University School of Medicine, Stanford, California, USA.

<sup>8</sup>Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany. Correspondence should be addressed to M.N. (e-mail: magnus@usc.edu).



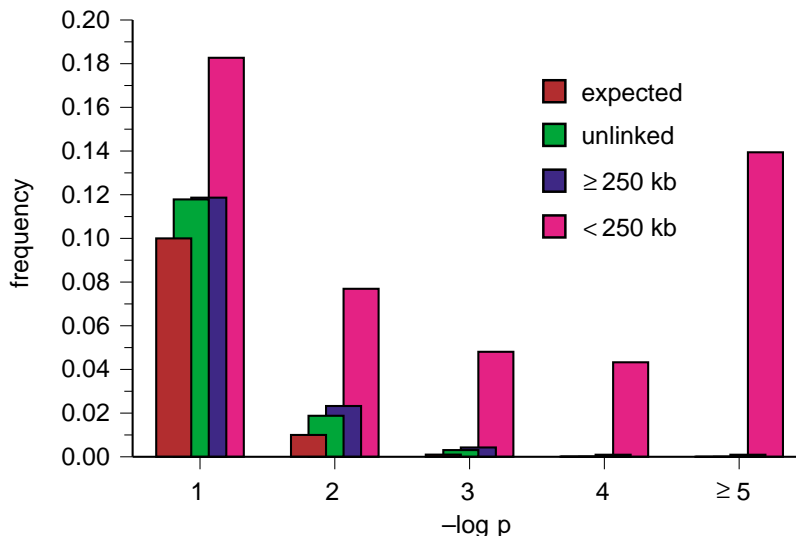
**Fig. 2** LD as a function of distance for 163 genome-wide SNP markers<sup>13</sup>. Markers were typed in 76 accessions. The number of pairs of markers separated by less than 250 kb (the maximum distance in Fig. 1) is 5, 11, 11, 179 and 2 for chromosomes 1–5, respectively. Chromosome 4 stands out because it contains a 500-kb region (not overlapping the *FRI* region) with 25 markers: the pattern of LD in this region is shown in the bottom right plot. Note that LD should reflect genetic rather than physical distance. Genetic positions are available for most of these markers, but because the estimates are highly unreliable for the short distances within which LD is strong, we show physical distances instead. Plots of LD against genetic distance look qualitatively very similar (data not shown).

influence of selection is supported by the observation that whereas *FRI* may be involved in local adaptation<sup>14</sup>, the other region contains the disease-resistance locus *RPP5*, which may be subject to frequency-dependent selection<sup>15</sup>. Local adaptation is expected to decrease the effective heterozygosity at a locus and thus increase LD; frequency-dependent selection is expected to have the opposite effect<sup>16</sup>. A potentially important difference between the samples concerns ascertainment: whereas the data shown in Fig. 1 include all polymorphic sites in the sample, the data in Fig. 2 include only polymorphisms that distinguish two particular accessions (Ler and Col)<sup>13</sup>. This could bias the results. However, analysis of subsamples of the data in Fig. 1 suggests that this difference between the samples does not explain the difference in the decay of LD (data not shown).

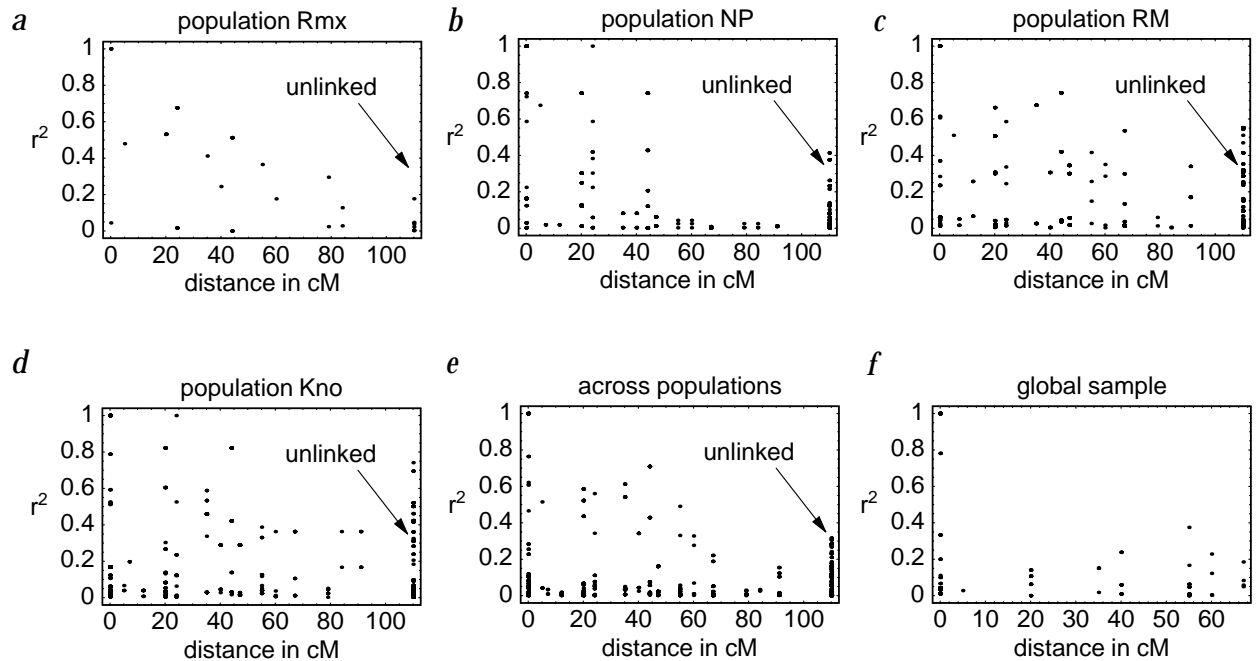
Notably, LD is highly dependent on the ages of the alleles involved. Even if LD in general decays within 250 kb, there will still be young alleles surrounded by much more extensive LD. Our results concern only the rate at which LD decays; there is little evidence for long-range LD in the present data. Indeed, the distribution of LD values for pairs of markers located more than 250 kb apart is similar to that for pairs of markers

located on different chromosomes (Fig. 3). The distribution for closely linked pairs is completely different. The absence of genome-wide LD suggests that no ‘phylogeny’ of accessions exists: as expected, trees built from these data are almost completely star-like, and there is little evidence of global population structure. There is a bias, however, towards rare alleles in the SNP data (data not shown). This is consistent with other studies and may reflect a population expansion<sup>12,17</sup>.

How does the pattern of LD reported here differ from that observed in other organisms? The rate of decay of LD in *A. thaliana* seems to be 1/50 of that in *Drosophila melanogaster*, where LD typically decays within 1 kb (ref. 18). Maize seems to be similar to



**Fig. 3** *P* values under Fisher's exact test for the genome-wide SNP markers. In the absence of LD, we would expect 10% of all comparisons to show association at the 0.1 level, 1% at the 0.01 level, and so on. The histogram shows these expectations along with the observed frequencies for comparisons between closely linked SNPs, distantly linked SNPs and SNPs on different chromosomes.



**Fig. 4** LD as a function of genetic distance in local Michigan populations. LD is much more extensive than in global samples. The plots show the decay of LD in four populations (samples of 71, 91, 117 and 80 individuals, respectively) as well as in the pooled sample of 359 individuals. The data consist of 39 segregating sites from seven loci, six of which are on chromosome 3, linked to the disease resistance locus *RPM1*. The bottom right plot shows a small subset of these markers typed in a global sample of accessions (samples of 12–28 individuals, depending on the marker). Consistent with the results presented in Figs 1 and 2, LD does not decay with distance in this plot.

*D. melanogaster* in this respect<sup>19,20</sup>. The extent of LD in humans is highly controversial<sup>21–23</sup>. Depending on the region studied, LD seems to decay over a scale of a few kilobases to several hundred kilobases, overlapping the range observed here. Note, however, that even if the extents of LD in humans and *A. thaliana* are comparable, the overall patterns of polymorphism are very different, because *A. thaliana* harbors ten times more polymorphism than *Homo sapiens*. In other words, there are many more polymorphic sites per recombination event in the *A. thaliana* genome, which leads to extensive haplotype structure. These differences are to some extent in accordance with theoretical predictions (see Methods).

A topic of much debate in human genetics is whether LD is more extensive in small, isolated populations than in larger ones<sup>21,22,24,25</sup>. This would be expected if the populations had been founded by very small numbers of people. It seems clear that such founder effects must be especially dramatic in *A. thaliana*, which is capable of founding a population with a single seed. To investigate this, we analyzed the pattern of LD in several local populations from Michigan, using markers surrounding the disease resistance locus *RPM1*. As shown in Fig. 4, there is extensive LD in these populations, with LD decaying on a genome-wide scale, over 50–100 cM. Indeed, as *A. thaliana* may have been recently introduced to North America, it is possible that the decay of LD observed in these populations is mainly due to outcrossing and recombination events that have taken place in the last 200 years. With respect to *RPP5*, it is possible that frequency-dependent selection on *RPM1* may have influenced the pattern<sup>26</sup>.

We have shown that LD in *A. thaliana* is extensive, but not too extensive to be useful for mapping. It is clear that a genome-wide map would require many markers per cM—that is, thousands for the entire genome. Compare this with maize and *D. melanogaster*, where markers every 100 bp seem to be needed (but mapping at a

very fine scale is, on the other hand, possible)<sup>18–20</sup>. It is also clear that the several hundred markers on the existing *A. thaliana* SNP chip<sup>13</sup> will not usually suffice. For example, hypocotyl elongation in response to light was measured for the accessions used in the SNP survey, but no significant associations with any marker were found (J.O.B., unpublished data). This is not unexpected, given the observed rate of decay of LD. Relevant data (flowering time) are also available for the accessions in the *FRI* study<sup>27</sup>. The density of markers in this region is clearly sufficient, but the sample size turns out to be too small, given the considerable genetic heterogeneity of the trait<sup>14</sup>, for any association to be found (M.N., unpublished data). Avoiding genetic heterogeneity may prove to be the main advantage of local population samples in humans<sup>1,25,28,29</sup>. As we have shown, however, local populations of *A. thaliana* may also have much more extensive LD, raising the possibility that it may be possible to ‘zoom in’ on loci by using appropriate population samples. Population structure increases the risk of false positives in LD mapping, but statistical methods to circumvent this problem are being developed<sup>19</sup>.

The advantages of selfing for LD mapping should be emphasized. First, inbreeding increases LD without greatly decreasing polymorphism<sup>3</sup>. Second, haplotype data can be obtained without extra effort, avoiding a major problem in human genetics. Both factors suggest that modern LD mapping methods based on haplotype sharing could be very powerful in this model species—and perhaps also in other selfers, such as rice and barley.

## Methods

**Sequencing survey in *FRI* region.** We sequenced 13 segments of 0.5–1 kb in a 250-kb region surrounding the flowering-time locus *FRI* in a sample of 20 previously described<sup>27</sup> accessions: Kent, Conlara, Rsch-4, Algtstrum, Tamm-46, Pu-2-3, Shakh dara, Lisse, Lund, Köln, Got-32, Tsu-0, Ler, NC-6, Mt-0, Vimmerby, Kz-9, Pu-2-8, Dem-4 and Col. Each segment was amplified

from genomic DNA using primers constructed from the *A. thaliana* genome sequence and then sequenced in both directions. We resolved almost all discrepancies by repeated sequencing; however, it should be noted that for our analyses only polymorphisms present in more than one individual were used. Details will be reported elsewhere (J.H. and M.N., manuscript submitted).

**SNP data.** We genotyped 76 accessions from a world-wide sample using the *Arabidopsis* AT412 SNP chip and used Intermap1 software<sup>13</sup> to evaluate raw values from hybridization intensities. We used 12 replicates of Ler and 5 replicates of Col to develop a predictive model for calling the genotypes of 163 reproducible markers (MNPs excluded). The model used logistical regression as implemented in the R statistical software. Using the coefficients from this model and the raw values from the accessions, we determined the probability of the Col genotype for each marker for each accession:  $P > 0.8$  was assigned 'Col' and  $P < 0.2$  was assigned 'Ler'; otherwise the genotype was not called. These cut-offs were found to have an error rate of  $8/2403 < 0.005$ .

**Sequencing survey in *Rpm1* region.** We typed *A. thaliana* populations in the Chicago area at *Rpm1* (chromosome 3, 13 cM) using allele-specific PCR<sup>26</sup>. Four polymorphic populations were sampled: Rmx and NP in Berrien County, Michigan; RM in LaPorte County, Indiana; and Kno in Starke County, Indiana. We either designed marker loci from genomic clone sequences near markers placed on both genetic and physical maps (*F4P13*, chr 3/5 cM; *MGH6*, 3/18; *F4P13*, 3/52; *F4F15*, 3/72), or chose them from known genetic markers (*Ap3*, 3/84; *Cal*, 1/46). We ascertained polymorphisms by first sequencing both strands in accessions previously studied at *Rpm1* (ref. 26) and then genotyped population samples using single-pass sequencing; new polymorphisms detected in the population samples were confirmed with additional sequencing. Details are reported elsewhere (E.A. Stahl *et al.*, manuscript submitted).

**LD analyses.** We plot LD as  $r^2$ , the squared correlation coefficient. We removed singleton polymorphisms in all data, as they contribute little information about LD and are often due to typing errors. For each pair of loci, we calculated LD using the maximum number of haplotypes—that is, individuals for which some data were missing were not removed.

**Theoretical predictions.** According to population genetics theory, the rate of decay of LD is mainly determined by the compound parameter  $4Nr(1-F)$ , in which  $N$  is a scaling factor (the effective population size),  $r$  is the recombination probability per generation per bp, and  $F$  is the inbreeding coefficient<sup>3</sup>. In outcrossers,  $F=0$ , whereas in a 99% selfer,  $F=0.98$ . Thus, 99% selfing should decelerate the rate of decay of LD—the rate of decay being approximately 2% of that predicted were no selfing to occur.

Direct estimates of the compound parameter  $Nr$  are unreliable. An alternative procedure is to obtain a rough estimate of  $r$  by comparing genetic and physical distances. For example, 1 cM corresponds, on average, to 1 Mb in humans, so  $r$  must be, on average,  $10^{-8}$ . In *D. melanogaster* and *A. thaliana*,  $r$  seems to be, on average, four times greater.

The effective population size,  $N$ , can be estimated through the level of nucleotide diversity, which according to standard population genetics theory is proportional to  $Nu$ , where  $u$  is the probability of a neutral mutation per base pair per generation. This parameter can be estimated using phylogenetic methods and is generally found to be  $10^{-9}$ – $10^{-8}$ . The effective population sizes of *D. melanogaster* and *A. thaliana* seem to be comparable; that of humans seems at least an order of magnitude smaller.

These kinds of rough calculations suggest that LD in *A. thaliana* should decay no more than 2% as quickly than in *D. melanogaster*, and that humans should have more extensive LD than *D. melanogaster* but less polymorphism. Note that these predictions ignore population structure.

**URLs.** For SNP data: R statistical software, <http://www.R-project.org>; list of accessions, marker data and the R analysis script, <http://www.arabidopsis.org>.

#### Acknowledgments

C. Dean, U. Johanson and J. Werner provided unpublished data. J. Spiegelman and R. Hufft provided technical assistance.

Received 17 September; accepted 30 November 2001.

- Cardon, L.R. & Bell, J.I. Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99 (2001).
- Altshuler, D., Daly, M. & Kruglyak, L. Guilt by association. *Nature Genet.* **26**, 135–137 (2000).
- Nordborg, M. Linkage disequilibrium, gene trees, and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929 (2000).
- Hanfstringl, U. *et al.* Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for balancing and directional selection. *Genetics* **138**, 811–828 (1994).
- Innan, H., Tajima, F., Terauchi, R. & Miyashita, N.T. Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. *Genetics* **143**, 1761–1770 (1996).
- Aguadé, M. Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the *FAH1* and *F3H* genes, in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **18**, 1–9 (2001).
- Abbot, R.J. & Gomes, M.F. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heyn. *Heredity* **62**, 411–418 (1989).
- Bergelson, J., Stahl, E., Dudek, S. & Kreitman, M. Genetic variation within and among populations of *Arabidopsis thaliana*. *Genetics* **148**, 1311–1323 (1998).
- Todokoro, S., Terauchi, R. & Kawano, S. Microsatellite polymorphisms in natural populations of *Arabidopsis thaliana* in Japan. *Jpn. J. Genet.* **70**, 543–554 (1995).
- King, G., Nienhuis, J. & Hussey, C. Genetic similarity among ecotypes of *Arabidopsis thaliana* estimated by analysis of restriction fragment length polymorphisms. *Theor. Appl. Genet.* **86**, 1028–1032 (1993).
- Breyne, P., Rombaut, D., Van Gysel, A., Van Montagu, M. & Gerats, T. AFLP analysis of genetic diversity within and between *Arabidopsis thaliana* ecotypes. *Mol. Gen. Genet.* **261**, 627–634 (1999).
- Sharbel, T.F., Haubold, B. & Mitchell-Olds, T. Genetics isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology* **9**, 2109–2118 (2000).
- Cho, R.J. *et al.* Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nature Genet.* **23**, 203–207 (1999).
- Johanson, U. *et al.* Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* **290**, 344–347 (2000).
- Noël, L. *et al.* Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. *Plant Cell* **11**, 2099–2111 (1999).
- Nordborg, M. Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514 (1997).
- Miyashita, N.T., Kawabe, A. & Innan, H. DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment length polymorphism analysis. *Genetics* **152**, 1723–1731 (1999).
- Long, A.D., Lyman, R.F., Langley, C.H. & Mackay, T.F.C. Two sites in the *Delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics* **149**, 999–1017 (1998).
- Thornsberry, J.M. *et al.* *Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genet.* **28**, 286–289 (2001).
- Tenaillon, M.I. *et al.* Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci. USA* **98**, 9161–9166 (2001).
- Pritchard, J.K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
- Reich, D.E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
- Stephens, J.C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
- Goddard, K.A.B., Hopkins, P.J., Hall, J.M. & Witte, J.S. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66**, 216–234 (2000).
- Dunning, A. *et al.* The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* **67**, 1544–1554 (2000).
- Stahl, E.A., Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J. Dynamics of disease resistance at the *Rpm1* locus of *Arabidopsis*. *Nature* **400**, 667–671 (1999).
- Nordborg, M. & Bergelson, J. The effect of seed and rosette cold treatment on germination and flowering time in some *Arabidopsis thaliana* (Brassicaceae) ecotypes. *Am. J. Bot.* **86**, 470–475 (1999).
- Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
- Wright, A.F., Carothers, A.D. & Pirastu, M. Population choice in mapping genes for complex diseases. *Nature Genet.* **23**, 397–404 (1999).