Supporting Online Material for

The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*

Stephan Ossowski, Korbinian Schneeberger, José Ignacio Lucas-Lledó,* Norman Warthmann, Richard M. Clark, Ruth G. Shaw, Detlef Weigel,* Michael Lynch

*To whom correspondence should be addressed. E-mail: joslucas@indiana.edu (J.I.L.-L.); weigel@weigelworld.org (D.W.)

Published 1 January 2010, *Science* **327**, 92 (2010)
DOI: 10.1126/science.1180677

This PDF file includes:

Materials and Methods
SOM Text
Figs. S1 to S8
Tables S1 to S4
References

The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*

Supporting Online Material

Stephan Ossowski[1], Korbinian Schneeberger[1], José Ignacio Lucas-Lledó[2],
Norman Warthmann[1], Richard M. Clark[3], Ruth G. Shaw[4],
Detlef Weigel[1] & Michael Lynch[2]

October 26, 2009

1. Dept. of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany.
2. Dept. of Biology, Indiana University, Bloomington, IN 47405, USA.
3. Dept. of Biology, University of Utah, Salt Lake City, UT 84112, USA.
4. Dept. of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN 55108, USA.

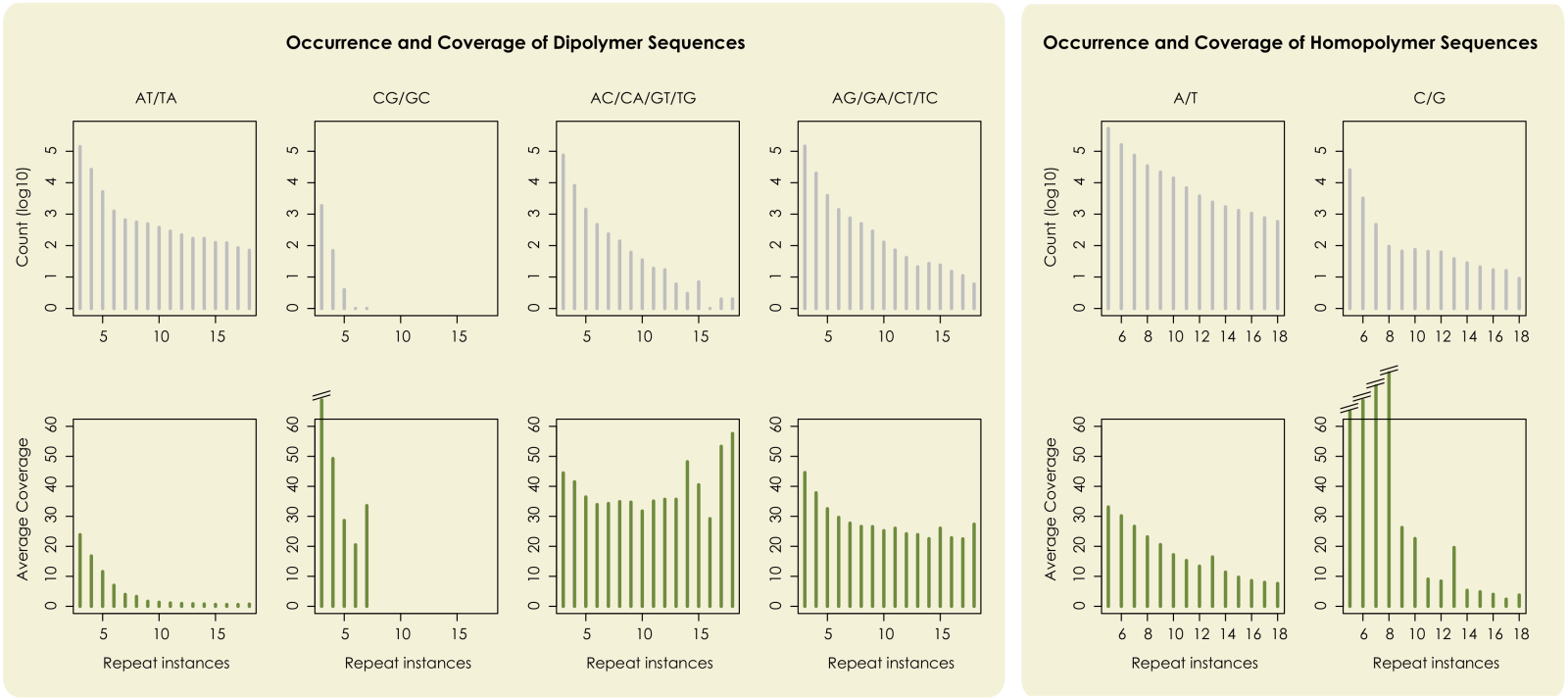

Figure S 1: Comparison of the frequency of different types of simple sequence repeats (SSR) in the genome (count) with their average coverage, as a function of the length of the repeatitive tract, in number of instances of the repeat unit. The number of occurrences of each type of SSR was obtained with a perl script, writen by K.S., and used before by Sureshkumar and colleagues (*1*). Data only shown for MA line 29.

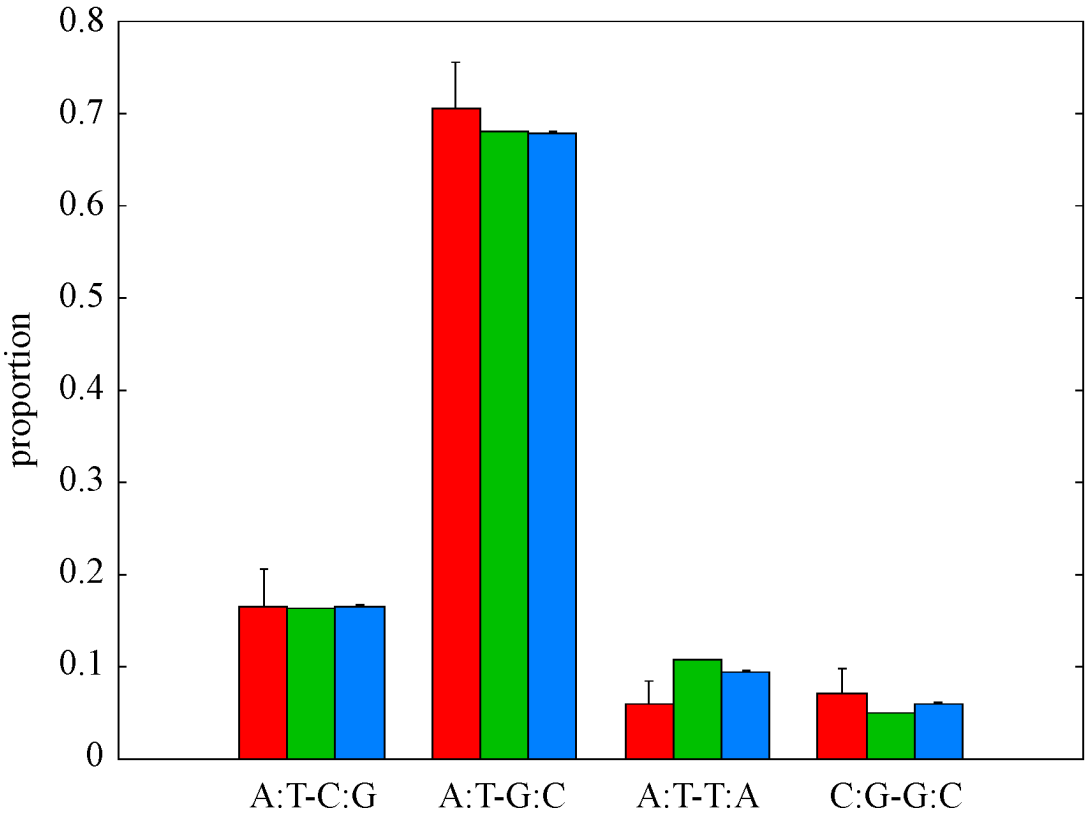

Figure S 2: Comparison of the proportions of the different types of base substitution mutations (red) with the respective proportions of SNP types in natural populations, surveyed by Clark and colleagues (blue) (*2*), or by Ossowski and colleagues (green) (*3*).

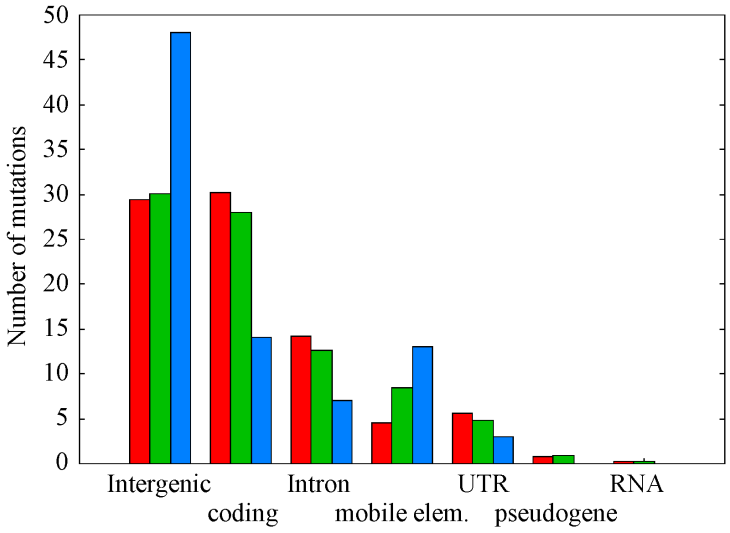

Figure S 3: Comparison of the observed (blue) and the expected mutations in different, exclusive functional contexts, according to the length of those functional components, their base composition (red), and their methylation levels (green) (*4*).

Materials and Methods

Mutation accumulation

The ancestor of the MA lines was maintained by selfing and single-seed descent for several generations, in order to reach mutation-drift equilibrium (*5*). Thus, the mutation-accumulation (MA) lines should have started in mutation-drift balance, and they were also propagated by selfing and single-seed descent as described before (*5*) during 30 generations. Because the majority of non-lethal mutations should behave in a neutral fashion under benign laboratory conditions in a line maintained by single seed descent (*6*), and segregation leads to the loss of 50% of new mutations under selfing, the number of fixed mutations per site per generation is essentially equal to the haploid mutation rate.

Sequencing and alignment

We purified genomic DNA from individuals of five 30 generations-old MA lines following standard procedures, and we prepared the DNA libraries for sequencing-by-synthesis following the manufacturer's protocols (Illumina, Inc.). Reads of between 36 and 43 bp were generated by the Genome Analyzer (cat. no. SY-301-1001; Illumina) and mapped to the reference genome of *Arabidopsis thaliana* using SHORE (available at http://1001genomes.org), as described elsewhere (*3*). A mean coverage depth between 23 and 31 was achieved per MA line. We expressed the coverage in every site and every MA line as a vector ('configuration') of four numbers, corresponding to the numbers of reads reporting each base (A, C, G, or T) in that genomic position and that MA line.

Consensus approach

To identify base-substitution mutations, we compared the base reported by most of the reads covering a site in one MA line (focus line) with the base inferred by the majority of reads covering the same site in all other MA lines (composite control). At least two other MA lines were required to compose the control. The use of a composite control, instead of the reference genome, is robust to the presence of sequencing errors in the reference genome and to the mutations fixed in the progenitor line before the mutation accumulation experiment. On the other side, this approach forced us to reduce the available sites for analysis to those sufficiently

covered in at least three MA lines. One additional difficulty of this approach is the identification of the mutant line in a mutant site. Under certain distributions of reads among MA lines, it is possible that a disagreement between the focal line and the composite control is observed in a mutant site even when the focal line is not the one harboring the mutation. To identify the mutant line, we sequencially used every MA line as focal at every site, and computed for every comparison the probability of obtaining that specific configuration of reads if the focal line was mutant, using equation 20 of a previous work (*7*) and an estimate of the sequencing error frequency explained below. Then, among the comparisons with a disagreement between the focal line and the composite control at a site, the most likely mutant line was chosen.

Sequencing error frequency

To estimate the MA line-specific frequency of sequencing errors, we used sites covered by at least 5 and at most 25 reads, in which the consensus base called in all 5 MA lines was the same, and no more than 20% of the reads called a discordant base in any MA line. Although the last condition seems to downwardly bias the estimate of the sequencing error frequency, the comparison with other threshold choices did not show evidence for such a bias, but suggested to be an effective measure to avoid the inclusion of other kinds of noise in the count of errors, such as the misalignment of reads from a paralogous origin. The average error frequency, over all MA lines, was 0.29%, and it ranged from 0.26 (MA 30_69) to 0.34% (MA 30_29).

The sequencing errors identified in this way showed a common spectrum in all MA lines. C:G→A:T errors were the most abundant ones, followed by A:T→C:G errors, while C:G→G:C errors were the less frequent ones (Figure 4).

Selection of sites

The probabilistic framework that we applied (*7*) accounts for sequencing errors when estimating the mutation rate. For that approach to be valid, sequencing errors must be the main source of false positives and false negatives during the count of putative mutations. Contrary to the assumed expectation of a constant error frequency across sites, a small proportion of sites was observed to hold a higher number of discordant base calls than expected, in agreement with what has been observed by others (*8*). In those sites, the discordant base calls are more likely originated by alignment errors than by sequencing errors. For example, if a site is duplicated

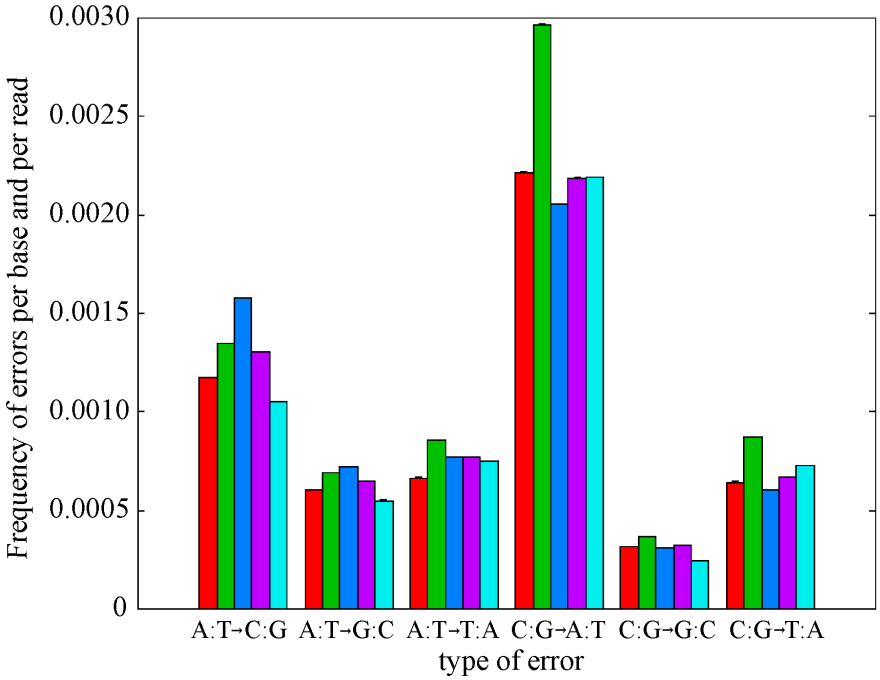

Figure S 4: Frequency of all types of sequencing errors in all five MA lines. Red, green, dark blue, purple, and light blue represent, respectively, mutation accumulation lines 119, 29, 49, 59, and 69.

in the target genome, and the duplication is not reported in the reference genome, the reads mapped to that site will have divergent paralogous origins and will frequently call discordant bases. Also, the few heterozygous sites that are expected to be present in the genomes of the MA lines will show the same pattern of ambiguous base calls. Keeping that in mind, we excluded sites suspected to be affected by these noise sources. Specifically, to be eligible for analysis, we required that sites fulfill the following conditions: 1) the site had to be covered by at least 5 and at most 60 reads in at least three lines. The upper threshold attempts to exclude erroneously high coverages associated with misalignments to paralogous DNA. 2) No more than one line was allowed to contain 'erroneous' reads, defined as a subset of the reads covering a site that report a nucleotide different from the nucleotide reported by most of the reads; and 3) if more than one base was read in one line, the consensus base had to be supported by at least 80% of the reads. The thresholds of the quality filters were adopted after we had explored over 100 alternative settings and attempted the confirmation of more than 500 putative mutations by Sanger sequencing (see below). We concluded that any relaxation of the current filters resulted in the inclusion of different proportions of false positives, while none of the mutations detected with

the final quality filters turned out to be false. Furthermore, assuming the estimated frequency of sequencing errors, we do not expect to have missed any mutations in the resulting set of sites analyzed. However, heterozygous mutations are expected to be excluded from analysis. Between 93 and 95 million sites out of the 120 million bp reference genome matched the quality requirements in each line.

The rationale behind this selection of sites is that the distribution of mutations among sites is independent of the criteria used to filter out untrusted sites. This seems to be true for homozygous mutations, which are the target of detection of this procedure. It has been argued that reads covering a mutant sites, which necessarily contain at least one mismatch to the reference genome, have a higher chance to be discarded by the sequencing quality filters than reads covering only wild-type sites (*8*). If this was the case, the requirement of a minimum coverage threshold would introduce a bias against the discovery of mutations. However, the coverage depth at mutant sites is not on average lower for the mutant line than for wild-type lines (paired t-test, p-value = 0.8; table 2). Our success in mapping reads containing mutations as efficiently as wild-type reads can be attributed to the closeness of the reference and the sequenced genomes, and to the performance of the Genome Mapper algorithm (http://1001genomes.org/downloads/genomemapper.html) (*9*).

To make the selection of sites compatible with the estimation of the probability of false consensus from the coverage depth and sequencing error frequency (*7*), a small modification of the original equations was needed. While the original computation of the probability of a false consensus adds up the probabilities of all possible configurations of reads that would lead to a false consensus, our implementation adds the probabilities of only those false configurations that would make the site elegible for analysis.

SHORE

In addition to the conservative selection of sites for the consensus approach, we used SHORE to detect single-base substitutions, short insertions and deletions (indels) of up to 3 base pairs (bp), and long deletions. The algorithms implemented in SHORE are more sensitive, because no threshold on read coverage is imposed, except that a homozygous call must be present in more than 50% of the overlapping reads. To identify false positive calls, SHORE calculates a quality score based on information from several features related to the quality of the read sequence and

the alignment. This approach increases the number of accessible positions and between 98.8 and 100.9 million sites per line were found to have sufficient information for calling either a mutation or the reference base. Single nucleotide substitutions and short indels were called if the variant had a quality greater than or equal to 25 (out of 40), no other line had a substitution or indel, and at least two lines had a high quality (≥ 25) reference call.

The high sensitivity of SHORE is illustrated by the fact that after attempted validation by Sanger sequencing of the top 500 candidate mutations, most of which were inferred with suboptimal settings of the consensus approach, only two mutations not detected by SHORE could be confirmed. Also, the detection of differentially fixed ancestral heterozygote sites, which must be extremely rare due to the inbreeding of the ancestral line, confirms the sensitivity of our approaches. We are, therefore, confident that false negatives are not biasing our estimates. Only one single-base substitution mutation detected by the consensus analysis was not detected by SHORE (see Table S1).

Estimation of the mutation rate

The expected number of putative mutations detected includes most of the true mutations present in the subset of sites analysed and some false positives, due to sequencing errors. Specifically, only one third of the true mutant sites with a false consensus base in either the focal line or the composite control are expected to be false negatives, since two thirds of those misleading consensus will still seem mutations, although of a different kind. On the other hand, all non-mutant sites with a false consensus in either the focal line or the composite control will be declared mutants. Given that the site-specific probability of a false consensus in either the focal line or the composite control can be estimated from the error frequency and the coverage depth (7), let $\bar{\pi}$ be its average across sites. And let m and m_1 be the number of sites analysed in a focal line and the true (unobservable) number of mutant sites among them, respectively. Then, the expected number of putative mutations detected, R, is: $E(R) = m_1(1 - \frac{1}{3}\bar{\pi}) + (m - m_1)\bar{\pi}$. Therefore, an estimate of m_1 for a specific MA line is:

$$\hat{m}_1 = \frac{R - \bar{\pi}m}{1 - \frac{4}{3}\bar{\pi}} \tag{1}$$

Given that conservative settings were used by the consensus approach, the expected number of mutations was equal to the observed one, indicating a very low probability of false positives

and false negatives.

Although all new mutations are heterozygous when they arise, a significant portion of them are inherited as homozygous by the zygote that will start the next generation (*10*). Specifically, one quarter of the mutations present in the germ line before the specialization of the reproductive tissues are expected to be homozygous at the beginning of the next generation. Let μ be the probability of a new homozygous mutation per site per generation, and τ, the probability of a new heterozygous mutation per site per generation. It can be shown that the accumulated probabilities of heterozygous and homozygous mutations over g generations are $2\tau(1-(\frac{1}{2})^g)$, and $g\mu+\frac{1}{2}\tau(g-2+(\frac{1}{2})^{g-1})$, respectively (*10*). These equations reflect the fact that half the mutations that are originally inherited as heterozygous will be lost on the long term. As a consequence, the count of mutations accumulated after g generations tends to $N \cdot g \cdot (\mu+\tau/2)$ as g increases, where N is the number of sites. This is so even if only homozygous mutations are counted. However, due to a finite number of generations, the estimate of the long term average mutation rate per site per generation, $\mu + \tau/2$, is downwardly biased if only homozygous mutations are counted (and upwardly biased if both kinds of mutations were counted). In the present case, $g = 30$, and the expectation of dividing the count of homozygous mutations by the product of the number of generations and the number of sites is $\mu + 14\tau/30$, from the expressions above. Put it another way, our raw estimate of the mutation rate is expected to be 1/30 of τ lower than the real mutation rate. Unfortunately, the values of μ and τ are hard to estimate, and depend on unknown details of plant development (*10*). In the worse scenario, if all mutations were originated after the specialization of the reproductive tissues and none of them was inherited in homozygosity by the following zygote, our estimate would be 28/30 times the true mutation rate, which is within the standard error of our estimates.

An even smaller bias in the opposite sense is expected from the ancestral heterozygote sites that were resolved with one of the alleles by one line and with the alternative allele by the other lines. These sites will be mistaken for mutations. After 30 generations, each allele has a 0.5 probability of being fixed. The resolution of each heterozygote site in the 5 MA lines is, therefore, a binomial experiment with $n = 5$ and $p = 0.5$. The number of ancestral single-base heterozygote sites can be estimated from the observed 2 single-base polymorphic sites with a common allele in two or three lines. If these sites are the result of differential resolution of ancestral heterozygote sites, they are expected to be 62.5% of the ancestral heterozygote sites.

Therefore, 3.2 sites are expected to have been single-base heterozygote polymorphisms in the ancestral line, and only in one of them (31.3%) one allele is expected to be fixed in one line, while the alternative allele was fixed by all other lines. One mutation should be substracted from the total count, although the effect of this correction is negligible.

Mutational spectrum

We assume that the mutational properties are symetric between strands, and therefore distinguish only two kinds of sites (A:T, and C:G) and three kinds of base-substitution mutations in each (A:T→C:G, A:T→G:C, A:T→T:A, C:G→A:T, C:G→G:C, and C:G→T:A). To estimate the conditional mutation rates, we applied a similar approach to that explained above for the unconditional mutation rate. The expected number of detected putative mutations of kind j, R_j, among all sites of kind i, m_i, is approximated by the number of true mutants of that kind with a true consensus in both the focal line and the composite control, $m_{1j}(1-\bar{\pi})$, plus the number of non-mutant sites of kind i with a false consensus of kind j in the focal line, $(m_i - m_{1j})\epsilon_j\bar{\pi}$, where ϵ_j is the proportion of errors of kind j among those possible in sites of kind i. Although other kinds of errors in the composite control of other sites not of kind i can contribute to this count of putative mutations, they are expected to be very few, since most of the false consensus must happen in the focal line, usually covered by a lower number of reads. Therefore, and estimation of the number of true mutations of kind j is given by:

$$\hat{m}_{1j} = \frac{R_j - m_i\epsilon_j\bar{\pi}}{1-\bar{\pi}(1+\epsilon_j)} \tag{2}$$

Which, divided by the number of generations times the number of sites of kind i analysed, gives the estimation of the conditional rate of mutations of kind j in sites of kind i.

Maximum likelihood estimation of the mutation rate

Given the number of reads calling each base at every site analysed and every MA line, the maximum likelihood estimates of the mutation rate and the sequencing error frequency, over all MA lines, were obtained by optimization of the log likelihood function (*7*). We obtained expressions of the derivatives of the log likelihood function with respect to those two parameters and implemented the Newton-Raphson's method of optimization in custom-made C++ scripts.

The initial values of the mutation rate and the error frequency, used to seed the optimization procedure, were taken from the consensus analysis.

Validation of mutations by Sanger sequencing

To choose the thresholds of the quality filters applied to select the valid sites, we first applied more than 100 different 'filters', that is, different thresholds for: the minimum coverage, the maximum coverge, the maximum number of MA lines with a discordant base, and the maximum proportion of discordant reads. Overall, 700 putative mutations were detected, most of which where inferred only under the most relaxed threshold settings. In agreement with our expectation of many of those putative mutations being false positives, the most permisive filters yielded higher estimates of the mutation rate (Figure 5). To identify the optimal filter settings, we attempted the validation of 596 putative mutations by Sanger sequencing. For each of them, the putative mutant line, and an additional MA line not supposed to carry the mutation were sequenced. 99 mutations were confirmed and 439, rejected, the rest remaining undetermined due to PCR difficulties. The results of the validations made it easy to choose the filter options that allowed the inference of most of the true mutations without including any of the false positives (Figure 6). Among the 85 mutations identified in this way by the consensus method, 83 were confirmed, and 2 were determined to be shared by two MA lines, which indicates that they are not true mutations, but ancestral heterozygous sites differentially fixed among the lines. Since the ancestral line was inbred, the number of heterozygous sites was expected to be very low. Our ability to detect those that resulted in shared polymorphisms among the MA lines confirms the sensitivity of our methods.

The two longest deletions, of 5445 and 610 bp, were validated by comparing the sizes of PCR products among lines in a gel.

Genomic deleterious mutation rate

Wright and collegues (*11*) estimated the divergence between *A. thaliana* and *A. lyrata* to be 0.126 synonymous substitutions per site in coding regions. If we substract the average within-species polymorphism level of 0.01 (*12, 13*), the net divergence becomes 0.116. Wright and collegues also estimated an average of 0.077 deleterious mutations per site between the two species (*11*). This implies that 66% of the mutations have been purged by natural selection in coding regions

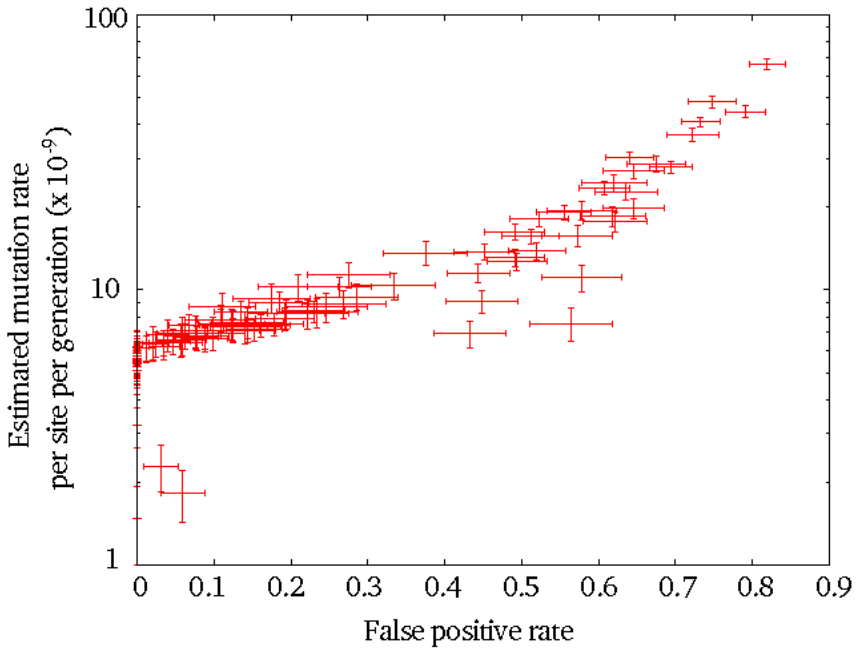

Figure S 5: The estimated mutation rate depends on the quality of the sites included in the analysis. For every selection of sites, its corresponding mutation rate is plotted against its rate of false positives. The proportion of false positives among all mutations is estimated as the proportion of mutations called and rejected by Sanger sequencing over all the mutations called for which Sanger sequencing provided any information.

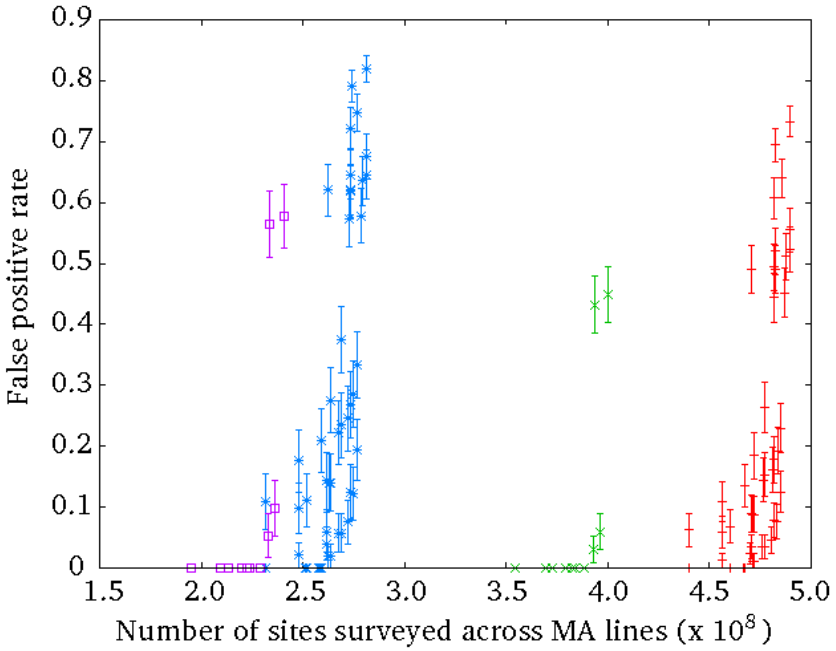

Figure S 6: For every subset of sites selected for analysis, its false positive rate (see Figure 5) is plotted against the total number of sites summed across MA lines included in that subset. In purple, subsets of sites with a maximum coverage of 25 and without any MA line containing discordant base calls; in blue, subsets of sites with maximum coverage of 25 and with at most one MA line containing discordant base calls; in green, subsets with maximum coverage of 50 and without MA lines containing discordant base calls; in red, subsets with maximum coverage of 50 allowing one MA line with discordant base calls.

during the divergence between the two species. If we multiply our estimate of the mutation rate per site per generation in coding regions, $3.15 \pm 0.8 \times 10^{-9}$, by the mean selective constraint of 0.66, and by twice the size of the coding genome (32915728 sites), we obtain an estimate of the genomic deleterious mutation rate of 0.14 ± 0.04 per site per generation.

Comparison between the spectrum of mutations and the spectrum of natural polymorphisms

To compare the spectrum of single-base mutations to the spectrum of synonymous polymorphisms among natural isolates of *A. thaliana*, we grouped mutant sites into the four possible classes without polarization: A:T↔C:G, A:T↔G:C, A:T↔T:A, and G:C↔C:G. The frequencies of these classes of mutations are what we call the 'unpolarized' spectrum of mutations, which can be compared to the spectrum of polymorphisms without having to infer their ancestral estate.

Supporting online text

Distribution of base-substitution mutations along the genome

A slight deficit of base substitutions is observed on chromosome 5 (G test p-value = 0.04; Figure 7), although there is no evidence of a lower level of intraspecific polymorphism or interspecific divergence on chromosome 5 (data not shown). Sequence coverage on chromosome 5 was not lower than for the other chromosomes.

Intergenic base substitutions happen more frequently near the centromere than far appart from it. Curiously, the effect of the distance to the centromere is not observed in genic regions (Figure 8). This could be due to a lack of power, or to differences in the processes involved in the effect of the centromere between genic and intergenic regions. We do not have an explanation of the variation of mutation rate along the chromosome in intergenic regions. Thus, it is difficult to speculate why genes should not be affected by the same processes. Natural selection is very unlikely to account for this pattern for several reasons: first, we do not see any significant deficit of nonsynonymous mutations, in agreement with a very low level of selection; second, we have no reason to believe that natural selection is stronger near the centromere than far apart from it; and third, if natural selection was removing genic mutations in pericentromeric regions, it would leave behind much more synonymous substitutions than what we see.

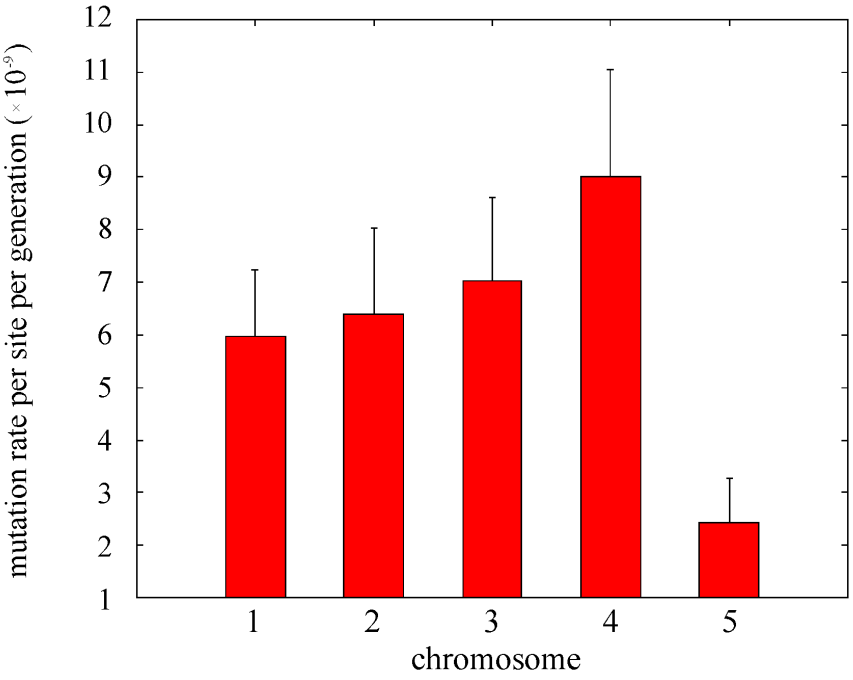

Figure S 7: Chromosome-specific mutation rate across MA lines. Error bars are standard errors.

Since methylation levels are higher near the centromere (*4*), only non-methylated sites have been used to test the effect of the distance to the centromere on the mutation rate. We used an heuristic definition of centromeres developed by Clark and colleagues (*2*). Table 4 show the lengths of the chromosomes and the positions of the centromeres according to that deffinition.

The effect of the distance to the centromere may be responsible for differences of mutation rate among chromosomes. Chromosome 5 is the second largest one, after chromosome 1. Therefore, it is possible that the small deficit of mutations observed in chromosome 5 is due to the fact that a large fraction of its length is far away from the centromere, in comparison to most other chromosomes. We, therefore, repeated the comparison of mutation rates among chromosomes using only sites at between 6 and 12 Mb from the centromere, where no effect of the distance to the centromere can be observed (figure 8). In those sites, no significant difference of mutation rate among chromosomes can be detected (G test, p-value = 0.36).

Comparison of the mutation rate between organisms

A scaling of the per-generation mutation rate with genome size has been observed in several species (*14*), and *A. thaliana* is consistent with this pattern. Our estimate of $6.5 \pm 0.7 \times 10^{-9}$

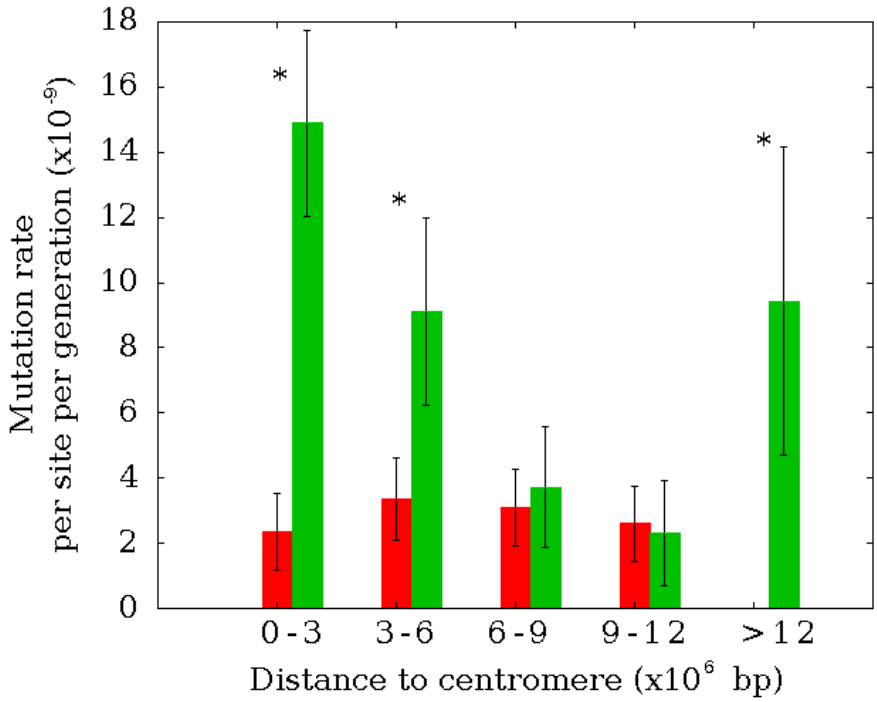

Figure S 8: Mutation rate per site per generation at different distances from the centromeres, for genic (red) and intergenic (green) regions. Only sites not known to be methylated have been used, to avoid any effect of methylation patterns along chromosomes. Sites from all 5 chromosomes have been classified by their distance to their respective centromeres along both arms. An asterisc indicates that mutation rates are significantly different between genes and intergenic regions at the 0.05 level (Fisher's exact test).

base-substitutions per site per generation for this species is also close to the lower bound of an indirect estimate based on the divergence between monocots and dicots (*15*). It translates into a per cell division mutation rate of about 0.2×10^{-9} base substitutions per site, assuming that there are around 30 to 40 cell divisions between a zygote and the gametes (*10*), which is close to the corresponding per cell division estimates for humans, *Drosophila melanogaster*, yeast, and bacteria (*16*). However, there is considerable uncertainty about the average number of cell divisions per generation in *A. thaliana*.

Time since divergence between *A. thaliana* and *A. lyrata*

Fossil evidence coupled with sequence analysis at two individual loci has previously led to an estimated synonymous substitution rate of 15×10^{-9} in the Brassicaceae (*17*), which in turn resulted in an estimated time of divergence since the separation of the two species of 4 to 5 MYA (*11, 17*). It has been noted that the rate estimated by Koch and colleagues (*17*) is high, and using a rate of 5.8×10^{-9} substitutions per year in each branch, as suggested by Wolfe and colleagues based on divergence between monocots and dicots (*15*), an upper bound of 10 MYA was proposed (*11*). Our directly measured rate of 6 to 7×10^{-9} base-substitution mutations per site per generation would yield an estimate of 8.7 ± 1.0 million years since the split between the two species. However, the divergence of 0.13 substitutions per site was calculated from protein coding regions (*11*), which we found to have a mutation rate of 3.1×10^{-9}. Thus, taking into account an average level of nucleotide diversity within species of 0.01 (*12, 13*), and assuming a generation time of one year, *A. thaliana* and *A. lyrata* separated around 17.9 ± 4.8 million years ago, much before what was previously thought.

Table S 1: List of validated mutations. Methylation levels are as reported by Cokus and collegues (*4*), unknown (UN), or not appliable (NA). UTR, untranslated region. Method indicates whether mutation was called by SHORE, consensus approach, or both.

MA line	chr.	position	mutation	context	methylation	method
29	1	936307	T→G	UTR	NA	SHORE
29	1	7068062	C→T	transposable	0.64	both
29	1	11046428	G→A	intergenic	1.00	both
29	1	22872635	A→G	intron	NA	both
29	2	4077436	A→T	intergenic	NA	both
29	2	5698741	deletion (1 bp)	intergenic	0.05	SHORE
29	2	14181788	T→C	nonsynonymous	NA	both
29	2	18980246	G→A	intergenic	0.00	SHORE
29	3	1207250	T→C	intron	NA	both
29	3	9510201	C→T	UTR	0.00	both
29	3	9708721	deletion (1 bp)	intergenic	NA	SHORE
29	3	11028301	A→G	intergenic	NA	both
29	3	15175453	G→A	intergenic	0.00	both
29	3	15506155	C→T	transposable	0.44	both
29	3	18081319	C→T	intergenic	0.00	both
29	4	2983508	C→T	transposable	0.14	both
29	4	5854619	C→T	intergenic	0.86	both
29	4	7627912	C→T	intron	0.00	both
29	5	10116709	C→T	intergenic	0.00	both
29	5	26151643	insertion (1 bp)	intergenic	NA	SHORE
49	1	13150205	C→T	intergenic	0.45	both
49	1	16359560	G→A	intergenic	0.18	both
49	1	17233704	G→T	intergenic	0.00	both
49	1	21998279	A→C	nonsynonymous	NA	both

Continued on next page

Table S 1 – continued from previous page

MA line	chr.	position	mutation	context	methylation	method
49	1	25338065	T→C	nonsynonymous	NA	both
49	1	29100212	deletion (5445 bp)	3 gens	NA	SHORE
49	2	3850915	A→T	transposable	NA	both
49	2	8833080	insertion (1 bp)	intergenic	NA	SHORE
49	2	10757312	C→T	intergenic	0.00	SHORE
49	3	11306735	T→C	intergenic	NA	SHORE
49	3	13455609	A→C	nonsynonymous	NA	both
49	3	15098177	C→T	intergenic	0.00	both
49	3	15591574	G→A	transposable	0.00	both
49	3	19431714	A→G	intergenic	NA	both
49	3	20167955	C→G	nonsynonymous	0.00	both
49	4	4105103	G→A	intergenic	0.00	both
49	4	4128739	C→A	transposable	0.00	both
49	4	5449854	T→C	intergenic	NA	both
49	4	15638148	C→T	intergenic	0.00	both
49	5	5513177	C→T	intergenic	0.00	both
49	5	8995983	deletion (1 bp)	intergenic	NA	SHORE
49	5	16113397	deletion (15 bp)	intron	NA	SHORE
59	1	1468806	C→G	intergenic	0.00	both
59	1	4450911	A→C	synonymous	NA	both
59	1	9552508	C→T	synonymous	0.75	both
59	1	11493432	G→A	intergenic	0.00	both
59	1	24572212	G→A	intergenic	0.00	both
59	1	27850165	G→A	intron	0.00	both
59	1	28700353	C→A	intergenic	0.00	both
59	2	3786851	G→T	intergenic	1.00	both
59	2	5887984	C→A	transposable	0.00	both
59	2	15896842	G→A	intergenic	0.00	both

Continued on next page

Table S 1 – continued from previous page

MA line	chr.	position	mutation	context	methylation	method
59	2	17502589	T→C	intergenic	NA	both
59	3	4316140	deletion (1 bp)	intron	NA	SHORE
59	3	10416525	C→T	intergenic	0.00	both
59	3	10811191	C→T	intergenic	0.00	both
59	3	12243310	T→G	intergenic	NA	both
59	3	18703173	C→G	intron	0.00	both
59	3	22082544	A→C	synonymous	NA	both
59	4	31291	deletion (2 bp)	nonsynonymous	NA	SHORE
59	4	1985679	G→A	transposable	0.84	both
59	4	2687693	deletion (2 bp)	intergenic	0.00	SHORE
59	4	4729650	A→C	transposable	NA	SHORE
59	4	5226368	G→T	intergenic	0.06	SHORE
59	4	5381530	C→T	intergenic	0.34	both
59	4	7844516	deletion (2 bp)	intron	0.00	SHORE
59	4	8362950	C→G	nonsynonymous	0.00	both
59	5	14687103	G→C	transposable	1.00	SHORE
59	5	20717113	deletion (1 bp)	nonsynonymous	NA	SHORE
59	5	20886692	G→C	nonsynonymous	0.00	SHORE
69	1	11947626	G→A	nonsynonymous	0.00	both
69	1	21686945	G→A	intergenic	0.00	both
69	1	28571633	C→T	intergenic	0.00	both
69	2	5805588	C→T	intergenic	UN	SHORE
69	2	6075002	A→T	transposable	NA	both
69	2	6445372	C→T	transposable	0.07	both
69	2	8592695	G→A	intergenic	0.00	both
69	2	11369734	G→A	intergenic	0.53	consensus
69	3	5081092	deletion (15 bp)	intergenic	NA	SHORE
69	3	16724903	G→A	intergenic	0.00	both

Continued on next page

Table S 1 – continued from previous page

MA line	chr.	position	mutation	context	methylation	method
69	4	2752261	G→T	nonsynonymous	0.00	SHORE
69	4	3685382	T→A	intergenic	NA	both
69	4	3688311	C→T	transposable	1.00	SHORE
69	4	6390518	A→G	intergenic	NA	SHORE
69	4	8208973	G→A	intergenic	0.00	both
69	5	6434118	G→A	transposable	0.43	both
69	5	15620702	C→T	intergenic	0.00	both
69	5	16973407	C→T	synonymous	0.57	both
69	5	22268356	G→A	intron	0.00	both
119	1	7065549	G→A	transposable	UN	SHORE
119	1	7591509	C→T	intron	0.00	both
119	1	9335711	C→T	intergenic	0.00	none
119	1	10978854	C→T	intergenic	0.00	both
119	1	12253505	A→G	intergenic	NA	both
119	1	16724283	T→A	intergenic	NA	both
119	1	18429314	insertion (1 bp)	intron	NA	SHORE
119	2	6946725	G→A	transposable	0.7	both
119	2	8680933	A→G	intergenic	NA	both
119	2	9528174	G→A	intergenic	0.00	both
119	3	47128	G→T	nonsynonymous	0.2	both
119	3	840906	G→A	intergenic	0.00	SHORE
119	3	11717753	G→C	intergenic	0.00	both
119	3	19123756	C→T	UTR	0.00	both
119	3	22524825	deletion (610 bp)	1 gene	NA	SHORE
119	4	1640531	insertion (1 bp)	intergenic	NA	SHORE
119	4	4280164	C→T	transposable	0.73	both
119	4	6418272	C→T	intergenic	0.00	both
119	4	8674353	C→T	nonsynonymous	0.00	both

Continued on next page

Table S 1 – continued from previous page

MA line	chr.	position	mutation	context	methylation	method
119	4	8799686	G→A	intergenic	0.00	both
119	4	13328344	G→A	intergenic	0.10	both
119	4	13514562	G→T	intergenic	0.00	both
119	4	13514563	C→G	intergenic	0.00	both
119	5	2650986	C→T	UTR	0.00	both
119	5	8925358	insertion (1 bp)	intergenic	NA	SHORE
119	5	13239247	C→T	intergenic	0.00	SHORE
119	5	14578077	C→T	intergenic	0.00	both

Table S 2: Average coverage depth and standard deviation at mutant sites in the mutant line and in the wild-type lines.

Mutant line	Mutant coverage	wild-type coverage
29	20.6 (8.8)	16.3 (7.8)
49	16.7 (6.7)	20.3 (7.5)
59	19.8 (9.0)	19.3 (10.2)
69	21.1 (13.4)	20.2 (11.1)
119	15.0 (9.7)	17.8 (11.7)
total	18.5 (9.8)	18.8 (9.8)

Table S 3: Proportions of types of base-substitution mutations in several organisms, total number of surveyed mutations (N), and transition to transversion ratio (Ts/Tv). Cytosine methylation is only present in *H. sapiens*, *A. thaliana*, *E. coli*, and, in very low levels, in *D. melanogaster*.

	Transitions		Transversions					
	AT→GC	GC→AT	AT→TA	GC→TA	AT→CG	GC→CG	N	Ts/Tv
H. sapiens [ref. (*18*)]	0.23	0.40	0.06	0.10	0.08	0.12	1925	1.73
D. melanogaster [ref. (*8*)]	0.19	0.30	0.12	0.16	0.11	0.11	174	0.98
C. elegans [ref. (*19*)]	0.10	0.20	0.22	0.28	0.11	0.08	391	0.45
A. thaliana [this work]	0.12	0.59	0.05	0.09	0.07	0.08	99	2.41
S. cerevisiae [ref. (*18*)]	0.10	0.29	0.06	0.31	0.06	0.18	1250	0.64
E. coli [ref. (*18*)]	0.16	0.32	0.19	0.11	0.16	0.06	1037	0.92

Table S 4: Lengths of chromosomes in base pairs and positions of the centromeres, according to Clark and colleagues (*2*)

Chromosome	length	start	end
1	30432563	13700000	15900000
2	19705359	2450000	5500000
3	23470805	11300000	14300000
4	18585042	1800000	5150000
5	26992728	11000000	13350000

References

1. S. Sureshkumar, *et al.*, *Science* **323**, 1060 (2009).
2. R. M. Clark, *et al.*, *Science* **317**, 338 (2007).
3. S. Ossowski, *et al.*, *Genome Res* **18**, 2024 (2008).
4. S. J. Cokus, *et al.*, *Nature* **452**, 215 (2008).
5. R. G. Shaw, D. L. Byers, E. Darmo, *Genetics* **155**, 369 (2000).
6. M. Kimura, *The neutral theory of molecular evolution* (Cambridge University Press, 1983).
7. M. Lynch, *Mol Biol Evol* **25**, 2409 (2008).
8. P. D. Keightley, *et al.*, *Genome Res* **19**, 1195 (2009).
9. K. Schneeberger, *et al.*, *Genome Biol* **10**, R98 (2009).
10. P. D. Hoffman, J. M. Leonard, G. E. Lindberg, S. R. Bollmann, J. B. Hays, *Genes Dev* **18**, 2676 (2004).
11. S. I. Wright, B. Lauga, D. Charlesworth, *Mol Biol Evol* **19**, 1407 (2002).
12. M. Nordborg, *et al.*, *PLoS Biol* **3**, e196 (2005).
13. A. Kawabe, A. Forrest, S. I. Wright, D. Charlesworth, *Genetics* **179**, 985 (2008).
14. M. Lynch, *Mol Biol Evol* **23**, 450 (2006).
15. K. H. Wolfe, W.-H. Li, P. M. Sharp, *Proc Natl Acad Sci U S A* **84**, 9054 (1987).
16. M. Lynch, *et al.*, *Proc Natl Acad Sci U S A* **105**, 9272 (2008).
17. M. A. Koch, B. Haubold, T. Mitchell-Olds, *Mol Biol Evol* **17**, 1483 (2000).
18. Data come from a literature review by M.L.
19. D. R. Denver, *et al.*, *Proc Natl Acad Sci U S A* **106**, 16310 (2009).