Citation for published version:
Yang, S, Wang, L, Huang, J, Zhang, X, Yuan, Y, Chen, JQ, Hurst, LD & Tian, D 2015, 'Parent-progeny sequencing indicates higher mutation rates in heterozygotes', Nature, vol. 523, no. 7561, pp. 463-467. https://doi.org/10.1038/nature14649

DOI:
10.1038/nature14649

Publication date:
2015

Document Version
Peer reviewed version

Link to publication

Parent-progeny sequencing indicates higher mutation rates in heterozygotes.

Sihai Yang[1,3], Long Wang[1,3], Ju Huang[1,3], Xiaohui Zhang[1], Yang Yuan[1], Jian-Qun Chen,[1] Laurence D. Hurst* [2], Dacheng Tian*[1]

[1]State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, China; [2] Department of Biology and Biochemistry, University of Bath, Bath, U.K, BA2 7AY

[3] These authors contributed equally to this work.

*Correspondence should be addressed to D.T. (dtian@nju.edu.cn) or L.D.H. (bssldh@bath.ac.uk)

Mutation rates vary within genomes, but the causes of this remain unclear[1]. As many prior inferences rely on methods that assume an absence of selection, potentially leading to artifactual results[2], we call mutation events directly using a parent-offspring sequencing strategy focusing on *Arabidopsis* whilst using rice and honey bee for replication. Here we show that mutation rates are higher in heterozygotes and in proximity to crossover events. A correlation between recombination rate and intra-specific diversity is in part owing to a higher mutation rate in domains of high recombination/diversity. Implicating diversity *per se* as a cause, we find a ~3.5 fold higher mutation rate in heterozygotes than in homozygotes with mutation occurring in closer proximity to heterozygous sites than expected by chance. In a genome that is a patchwork of heterozygous and homozygous domains, mutations occur disproportionately more often in the heterozygous domains. If segregating mutations predispose to a higher local mutation rate, clusters of genes dominantly under purifying selection (more commonly homozygous) and under balancing selection (more commonly heterozygous), might have low and high mutation rates respectively. Results are consistent with this, there being a 10 times higher mutation rate in pathogen resistance genes, expected to be under positive/balancing selection. Consequently, we do not necessarily need to evoke extremely weak[1,2] selection on the mutation rate to explain why mutational hot and cold spots might correspond to regions under positive/balancing and purifying selection respectively[3,4].

To determine mutation rates we selected two purebred parents in both *Arabidopsis* (strains Col and Ler) and rice (strains 9311 and PA64s) (Figure 1). We selfed each and sequenced both parents (P_0) and progeny (P_1). In addition, we crossed to generate intra-specific F_1 heterozygotes. A single heterozygous F_1 seed in each species was selfed to generate multiple F_2 progeny. Comparing sequences between F_2s we could determine the F_1->F_2 mutation rate. While direct sequencing of genomes is the best way to detect *de novo* mutations[5,6], the error rate is high. We negate this by having multiple lines of quality control (Extended data 1a). First, we sequenced multiple independent DNA extractions from the same individual or inbred progeny of the individual, permitting a mutation to be called only if replicates agree. In practice, a mutation called in one extract from a given plant was always found in replicates. In addition, we use a consensus approach, comparing each focal individual against all other relevant samples[7]. For example, a presumptive mutation in an F_2 must be both called within a "mutated" sample and not called in both the sequenced parental genomes and all other F_2s. These criteria ensure that the mutation must have arisen sometime post a period late in the F_1, as all other F_2s share the same F_1 parent. To call a mutation we additionally require high sequence quality (score≥30; detail in Supplementary Table 1) and high coverage (>6000× for the sample cohorts and >40× for each sample) with at least 5 or more reads which must include both the forward and reverse reads. This approach is robust against sequencing or alignment errors in the reference genome[7]. False positive rates are negligible, while false negative rates are low (Methods).

In *Arabidopsis*, 237 base mutations and 67 small indels were detected in the 26 progeny of selfing *Arabidopsis* parents (P_0 to P_1) and 67 F_2 plants (F_1 to F_2) from the Col×Ler cross (Figure 1, Table 1 and Supplementary Table 2 & 3). To assess their reliability, several strategies were applied. First, Sanger sequencing confirmed 100% of 112 sampled base mutations and of 43 sampled indels present in F_2s. Confirmation requires that the mutation be present in the focal individual and absent in both parental genomes. Second, the sequenced 32 F_4 plants, derived from two F_2s (c52 and c64 with 10 mutations observed) (Fig. 1 and Supplementary Table 3), confirmed 100% of these F_2 mutations at a frequency of ~73% (slightly higher than the expected 62.5%). Third, we randomly sampled 4-8 F_3 plants from each of 21 sampled F_2s and subjected these F_3s to Sanger sequencing. This confirmed 99 out of 100 sampled base mutations and 24 out of 26 indels present in F_3s.

Comparison with prior estimates suggests that our method is robust. We estimate a rate of 7.4×10^{-9} base mutations per generation per site in homozygous individuals (i.e. P_0 -> P_1), similar to the prior estimate of 7.0×10^{-9} from mutation-accumulation Col lines[8]. As typically reported we observe more transitions than transversions (Supplementary Table 4, Extended data 2a) and for mutation to be disproportionately common at GC-rich nucleotide triplets (Supplementary Table 5). The ratio of point mutations to indels (3.9) is in line with prior estimates (3.11-5.8)[8,9]. Mutations in Col × Ler F_1 hybrids are as likely to occur on the Ler genome as on the Col genome (χ^2= 1.4, d.f. = 1, P=0.23).

We note one deviation from null expectation, this being a higher density of mutations in

Arabidopsis non-coding compared with coding regions, that cannot be accounted for in terms of differences in trinucleotide content (Supplementary Table 6). This suggests either underestimation of the mutation rate in coding sequences, possibly due to purifying selection, or a lower mutation rate in transcribed sequence, possibly owing to transcription coupled repair. A selectionist explanation predicts an increased relative frequency of indels that are multiples of three long in coding sequence. Even employing a one tailed test, we find no evidence for this. Of 81 indels, 62 and 12 are not multiples of three and outside and inside coding sequence respectively, while 5, outside and 2, inside, are multiples of three (Fisher's exact test, one tailed P=0.35).

Population-wide intragenomic diversity is commonly reported to be higher in genomic domains of high crossing-over[10], which we also see in *Arabidopsis* (Extended data 3a). This is typically ascribed to reduced selective interference between physically close alleles in domains of high recombination[10]. However it might also reflect a tendency for regions with high recombination rates to also be domains with high mutation rates, possibly because recombination is mutagenic[11–16]. Dissecting the chromosomes into 1 Mb non-overlapping regions we indeed find a positive correlation between mutation rates and the rates of crossover events in the 67 *Arabidopsis* F_2s and 32 F_4 plants (Fig. 2a). This is consistent with the possibility either that recombination is mutagenic or that both mutation and recombination preferentially occur in high diversity domains.

Given the very high intragenomic variation in crossover rates seen in honey bees, we examined the possibility that mutation happens more commonly in the vicinity of crossovers by examining *de novo* mutations in 46 honey bee genomes. In this species too intraspecific diversity is correlated with the crossing over rate[17]. Mutagenic effects of crossing over are thought to occur within 2kb of the break point[16]. Of 35 mutations, 2 mutations occurred within 2kb distance to a crossover breakpoint[17] (P=0.0012 with 10,000 randomizations; Extended data 2b). Thus in this genome too, new mutations occur in proximity to crossover events more often than expected by chance. We estimate the per genome mutation rate of a diploid queen to be 9.0×10^{-9} (6.8×10^{-9} for base substitution and 2.2×10^{-9} for indels).

While in the immediate vicinity of a double strand break (DSB) mutagenic repair may be acting[11–16], a higher rate of both recombination and mutation (mechanistically uncoupled) in domains of high diversity provides an alternative explanation for the correlation between mutation and recombination. That intraspecific diversity in *Arabidopsis* correlates with between-species divergence (Extended data 3b) is consistent with either possibility. A possible coupling between mutation and intragenomic diversity (i.e. heterozygosity) could be found if heterozygosity causes an increase in the mutation rate[18]. We test this by comparing progeny derived from heterozygous and homozygous parents in our two plant species. The point mutation rate (2.68×10^{-8}) as assayed from analysis of the F_2 progeny of heterozygous F_1 *Arabidopsis* is ~3.6-fold higher than that in the homozygous progeny of the selfed parents (two-tailed Brunner-Munzel (BM) test, P=3.64×10^{-8}). Similarly, the indel mutation rate in intergenic regions in heterozygote F_2s is ~2.8-fold higher than that in homozygotes (Table 1; two-tailed BM test, P=0.0012). The same pattern is seen in rice lines with 3.4-fold higher mutation rates in heterozygotes (3.2×10^{-9} and 1.1×10^{-8} per site per meiosis in homozygotes and heterozygotes respectively; Table 2; two-tailed BM test, P=0.0028). Analysis of 158 *Arabidopsis* point mutations in which Col, Ler and *A. lyrata* have the same state prior to mutation (and thus unlikely to hypermutagenic), reports that Col-Ler F_1 has a 5.02 fold larger mutation rate than the selfed Col or Ler parents (P_0->P_1) (BM test, P=1.02×10^{-7}).

The possibility that the degree of heterozygosity predicts the mutation rate can be further tested. Compared with F_2, a reduced mutation rate is expected in F_3 or F_4 selfed plants because the heterozygous regions will reduce by one half each generation. We identified 86 mutations detected in only one of the 32 F_4s, comprising 73 base and 13 indel mutations, giving a base mutation rate of 1.34×10^{-8} in F_4s, inherited from 18 F_3s of 2 F_2s (c52 and c64 in Fig. 1), and 1.60×10^{-8} in F_3s (for method see Ref. [8]; Fig. 2b and Extended data 1b for details). This ordering is as expected under the assumption that heterozygosity predicts mutation rates.

Were mutations easier to call in heterozygous regions the above may be artifactual. To address this we considered mutations from the F_1 to the F_2. In some genomic domains the F_2 preserves the heterozygosity of the F_1 (which is uniformly heterozygous during meiosis) while in some genomic locations the F_2 is homozygous. If artifact were to explain higher call rates in the heterozygous regions we expect more mutants called in the F_2 heterozygous domains. We do not observe this (153 mutations in the 54% heterozygous domains, 120 in homozygous domains, expected 146.5 and 126.5, respectively allowing for trinucleotide content; χ^2 with Yates correction =0.53, d.f. =1, P=0.47).

The above results may reflect either a) a tendency for heterozygotes to have genome-wide disruption of the mechanisms that prevent mutation (e.g. owing to disruption of coadapted heteromers), this being dependent on the proportion of the genome that is heterozygous or b) a genomically local effect of heterozygosity on the mutation rate. If the latter is the case, in genomes that are stratified into heterozygous and homozygous blocks the mutation rate should be higher in heterozygous domains. We observe this. There are 69 mutations in the regions of F_4s derived from heterozygous regions of both c64 (48% heterozygous blocks) and c52 (61% heterozygous blocks), compared with 27 in the regions of F_4s from homozygous regions of F_2s. Allowing both for the proportion of the genome covered and for differences in trinucleotide content, there is an excess in domains of heterozygosity (Expected 52.3 and 43.7; χ^2 with Yates correction =11.02, d.f.=1, P<0.001; Fig. 2c). Analysis of non-hypermutable sites confirms the same (χ^2 with Yates correction= 6.11, d.f. = 1, P=0.01).

A more conservative version of this test examines the 96 mutations in the F_4, accumulated since the F_2, in the 7% of the genome remaining heterozygous in the F_4. While such regions have a longer history of heterozygosis, many of the domains homozygous in the F_4 were heterozygous in the F_3. Nonetheless, we observe more mutations than expected in the heterozygous spans (heterozygous span: observe 13, expected 6.85; homozygous span 83 expected 89.15; χ^2 with Yates correction= 5.02, d.f. = 1, P=0.02). Analysis of non-hypermutable sites confirms this (χ^2 with Yates correction= 4.13, d.f. = 1, P=0.04). The above data support the notion that heterozygosity disposes to locally increased mutation rate.

If heterozygosity is causative we might expect mutational events to be close to heterozygous sites in the parents, while sites polymorphic in the population but not in the parents, need not be in especially close proximity to mutations. We find that parental heterozygous sites are significantly closer to mutational sites than expected (the red dots in Fig. 2d). There are a total of 273 mutations raised from F_1->F_2. The median distance of the *de novo* mutation to a heterozygous site is 167 bp (0 to 32694 bp), significantly smaller than the expected median distance with a random null (10,000 randomizations, Expected median = 207 bp; *P* =0.05). Of those mutations 113 are within 100bp of heterozygous sites, significantly more than expected by chance (10,000 randomizations, Expected number =93, *P*=0.005). As also expected, the level of diversity within the parents surrounding mutation sites is higher than the genome average (0.39% between two parents). By contrast, population polymorphism shows no such trend (Fig. 2d; Extended data 3c). The different patterns are consistent with local heterozygosity in the parent being causative, but a bias towards heterozygosity and mutation to both be intergenic might provide an alternative rationale.

On a broader scale, if we bin the genome into windows of 1Mb, we find a correlation between mutation rate and intra-specific diversity (Spearman's rho=0.76, *P*=0.0059), suggestive of broad scale mutational domains that impact on levels of polymorphism. If heterozygosity causes mutation such domains might be self-reinforcing, but the correlation alone is not evidence for this. Such an autocatalytic process suggests that both the highly polymorphic regions within a species and the species with higher rates of outcrossing could have higher mutation rates, compared with the conserved regions or self-crossing species, respectively. A number of studies indicate that the mating system (outcrossing or selfing), affects the mutation rate[19] and that mutations occur near pre-existing diversity[20], particularly near insertions/deletions[21]. However, as selfers and asexuals can retain linkage disequilibrium between mutator alleles and mutations, genome-wide selection on the mutation rate could confound between-species comparisons. More generally, understanding between-species variation is likely to be difficult owing to expected covariation with hard to know parameters, such as the effective population size.

It has been observed[3,4] that genomic hot and cold spots of indirectly inferred mutations accord with domains of genes putatively under strong purifying selection (mutational cold spots) and positive or balancing selection (mutational hot spots). The observation might, however, be an artifact of indirect methods to detect mutations: putatively neutral mutations in genes under strong purifying selection might be purged by selection if not neutral, causing an underestimation of mutation rates[2]. Our sequencing strategy largely avoids this problem. Nonetheless, we find evidence that genes expected to be under positive/balancing selection have high mutation rates. In total in *Arabidopsis*, 68 base mutations and 14 small indels occurred in coding sequences either as synonymous (21), non-synonymous or frame-shift mutations (59; Supplementary Table 6). Remarkably, 12 mutations are found in just a very few highly diversified gene families and hence prime targets of positive/balancing selection (Supplementary Table 7). Particular hotspots include 9 *LRR*-encoding (associated with pathogen resistance), and 3 F-box genes, where observed numbers greatly exceed the expected 0.89 (~10-fold higher) and 0.68 (~4.4-fold higher) mutations per family in these *Arabidopsis* F_2s respectively (Supplementary Table 7). Of the 17 coding mutations previously reported[8], one *NBS-LRR* gene (AT1G59780) and one *LRR*-pkinase were detected (Supplementary Table 7), suggesting that this result is repeatable. LRR-encoding and F-box genes have a lower GC content than the average (42.6%, 42.1% respectively versus mean of 44%), suggesting this is not owing to underlying nucleotide mutability.

While at first sight a higher mutation rate in genes associated with pathogen resistance (and positive/balancing selection more generally) makes sense in terms of selection acting on the mutation rate[3,18,22,23], such modifiers of the mutation rate acting locally will have such weak selection on them that such an explanation makes little theoretical sense[1,2], especially when population sizes are small. Our results suggest a resolution of this paradox: genes subject to balancing selection will have a higher chance of being heterozygous thus increasing the local mutation rate. That is to say, the selected variants could themselves be the modifiers of the mutation rate and hence their increase in frequency is attributed not to weak selection on the mutation rate, but strong selection on the direct phenotypic effects of some of the mutations.

We do not presume that heterozygosity is the only possible coupling between mutation and “non-essentiality”. Indeed, an explanation based on heterozygosity is not of obvious relevance to bacteria. The effect we observed suggesting correlation between DSB events and mutation might, however, be more general. Indeed, in bacteria DSB events can be mutagenic[24] and one need only hypothesise a coincidence between such recombination and non-essentiality, as seen in several eukaryotes[25], to provide an alternative explanation for hot and cold mutational spots. More immediate effects of transcription-coupled repair/mutation might also be of relevance.

While we make no attempt to investigate the underlying mechanism, we can speculate as to how heterozygosity might promote mutation. Several suggestions have been given[18], to which we add a possible coupling with poor-pairing during meiosis, as an immediate consequence of heterozygosity, especially for indels, may be poor-pairing quality or failure of homology search. Poor pairing might be mutagenic because physically exposed regions are more likely to proceed to Spo11 mediated DSB[26,27], repair of which is thought error-prone[28]. Similarly, the DNA damage response (DDR) protein MDC1 promotes accumulation of the sensor kinase ATR on unsynapsed chromosomes and chromatin loops in mammals[29]. Extended data 2c illustrates such a possible mechanism, where there are great differences in both length (47 versus 48 kb) and diversity (~10% between AT3G23110 and AT3G23120) between Col and Ler (or homologous chromosomes in the F_1).

If we have a major caveat about our results, it would be that the extent of size difference between Col and Ler is such that it may be unrepresentative of what normally happens in meiosis. Nonetheless, the poor-pairing model has the advantage that it might also explain the domains of higher mutation rate in homozygous Col[8]. During meiosis in homozygotes, repeating sequences (including clusters of homologous genes) can find homologous sequences at non-orthologous sites (ectopic recombination) and so force un-paired regions between homologous chromosomes. We analyzed the repeat sequences in and around our 145 and the previously found[8,9] 42 mutation bearing genes in homozygotes. Consistent with expectations 84.8% and 85.7% of these genes, including the gene AT1G59780, are located in repeat sequences or homologous gene clusters (Supplementary Table 7).

Acknowledgements This work was supported by National Natural Science Foundation of China (91331205, 91231102 and 31170210), PCSIRT (IRT_14R27) and JCIC-MCP to D.T. S.H. or J.Q.C.

Author Contributions D.T., L.D.H., S.Y. and J.Q.C. designed the experiments. S.Y., L.W., J.H., X.Z., L.D.H. and Y.Y. performed the experiments and analyzed the data. L.D.H. and D.T. wrote the paper.

Author Information All Illumina reads have been submitted to the short reads archive (http://www.ncbi.nlm.nih.gov/sra) under accession number PRJNA243018, PRJNA252997, PRJNA178613 and PRJNA232554. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.T. (dtian@nju.edu.cn) or L.D.H. (bssldh@bath.ac.uk)

References

1. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12,** 756–766 (2011).

2. Chen, X. & Zhang, J. No Gene-Specific Optimization of Mutation Rate in Escherichia coli. *Mol. Biol. Evol.* **30,** 1559–1562 (2013).

3. Chuang, J. H. & Li, H. Functional Bias and Spatial Organization of Genes in Mutational Hot and Cold Regions in the Human Genome. *PLoS Biol* **2,** e29 (2004).

4. Martincorena, I., Seshasayee, A. S. N. & Luscombe, N. M. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* **485,** 95–98 (2012).

5. Roach, J. C. *et al.* Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328,** 636–639 (2010).

6. Wang, J., Fan, H. C., Behr, B. & Quake, S. R. Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell* **150,** 402–412 (2012).

7. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci.* **109,** 18488–18492 (2012).

8. Ossowski, S. *et al.* The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis Thaliana. *Science* **327,** 92–94 (2010).

9. Jiang, C. *et al.* Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. *Genome Res.* **24,** 1821–1829 (2014).

10. Cutter, A. D. & Payseur, B. A. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* **14,** 262–274 (2013).

11. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18,** 337–340 (2002).

12. Magni, G. E. & Borstel, R. C. V. Different Rates of Spontaneous Mutation During Mitosis and Meiosis in Yeast. *Genetics* **47,** 1097–1108 (1962).

13. Perry, J. & Ashworth, A. Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **9,** 987–S3 (1999).

14. Pratto, F. *et al.* Recombination initiation maps of individual human genomes. *Science* **346,** 1256442 (2014).

15. Rattray, A., Santoyo, G., Shafer, B. & Strathern, J. N. Elevated Mutation Rate during Meiosis in Saccharomyces cerevisiae. *PLoS Genet* **11,** e1004910 (2015).

16. Arbeithuber, B., Betancourt, A. J., Ebner, T. & Tiemann-Boege, I. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl. Acad. Sci.* 201416622 (2015). doi:10.1073/pnas.1416622112

17. Liu, H. *et al.* Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. *Genome Biol.* **16,** 15 (2015).

18. Amos, W. Heterozygosity and mutation rate: evidence for an interaction and its implications. *BioEssays* **32,** 82–90 (2010).

19. Hollister, J. D., Ross-Ibarra, J. & Gaut, B. S. Indel-Associated Mutation Rate Varies with Mating System in Flowering Plants. *Mol. Biol. Evol.* **27,** 409–416 (2010).

20. Amos, W. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? *Proc. R. Soc. Lond. B Biol. Sci.* **277,** 1443–1449 (2010).

21. Tian, D. *et al.* Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455,** 105–108 (2008).

22. Pal, C., Maciá, M. D., Oliver, A., Schachar, I. & Buckling, A. Coevolution with viruses drives the

evolution of bacterial mutation rates. *Nature* **450,** 1079–1081 (2007).

23. Cox, E. C. On the Organization of Higher Chromosomes. *Nature* **239,** 133–134 (1972).
24. Shee, C., Gibson, J. L. & Rosenberg, S. M. Two Mechanisms Produce Mutation Hotspots at DNA Breaks in Escherichia coli. *Cell Rep.* **2,** 714–721 (2012).
25. Pál, C. & Hurst, L. D. Evidence for co-evolution of gene order and recombination rate. *Nat. Genet.* **33,** 392–395 (2003).
26. Gladyshev, E. & Kleckner, N. Direct recognition of homology between double helices of DNA in Neurospora crassa. *Nat. Commun.* **5,** (2014).
27. Boateng, K. A., Bellani, M. A., Gregoretti, I. V., Pratto, F. & Camerini-Otero, R. D. Homologous Pairing Preceding SPO11-Mediated Double-Strand Breaks in Mice. *Dev. Cell* **24,** 196–205 (2013).
28. Malkova, A. & Haber, J. E. Mutations Arising During Repair of Chromosome Breaks. *Annu. Rev. Genet.* **46,** 455–473 (2012).
29. Ichijima, Y. *et al.* MDC1 directs chromosome-wide silencing of the sex chromosomes in male germ cells. *Genes Dev.* **25,** 959–971 (2011).

Table 1. Number of spontaneous mutations per meiosis in the *Arabidopsis* genome.

Genotypes of the plants with meiosis	Sequenced samples		Base mutations (average mutations/sample)			Indel mutations (average mutations/sample)		
			Non-repeat regions	Repeat regions	Total (Average)	Non-repeat regions	Repeat regions	Total (Average)
Homozygotes $(P_0 \rightarrow P_1)$	26 P_1s		18 (0.69)	5 (0.19)	23 (0.88)	6 (0.29)	2 (0.08)	8 (0.31)
Heterozygotes $(F_1 \rightarrow F_2)$	67 F_2s		164 (2.45)	50 (0.75)	214 (3.19)	49 (0.673)	10 (0.15)	59 (0.88)
Heterozygotes $(F_2 \rightarrow F_4)$	32 F_4s	Specific	52 (1.62)	21 (0.66)	73 (2.28)	11 (0.34)	2 (0.06)	13 (0.41)
		Shared	4	5	9	1 (0.03)	0	1 (0.03)
	Average mutations/sample of $F_2 \rightarrow F_3$				(1.92)	ND		
	Average mutations/sample of $F_3 \rightarrow F_4$				(1.61)	ND		

The indel sizes range from 1 to 27 bp (2.91 on average; see Supplementary Table 8). The calculation of average mutations in the meiosis from $F_2 \rightarrow F_3$ and $F_3 \rightarrow F_4$ is described in Extended data 1b. ND, not determined because the number of indels is too small to calculate the average indel mutations per sample of $F_2 \rightarrow F_3$ or $F_3 \rightarrow F_4$.

Table 2. Numbers of spontaneous mutations (new SNPs) per meiosis in rice F_2s. Two rice samples, PA64s-3 and F2_29, were removed due to their low sequencing quality in one of the independent sequencings. The base substitution mutation rate is 3.2×10^{-9} and 1.1×10^{-8} per site per meiosis in homozygous and heterozygous rice genome, respectively.

Samples		SNPs	Indels	Samples	SNPs	Indels
Homozygotes	9311-1	0	1	PA64s-1	1	0
	9311-2	1	1	PA64s-2	1	2
	9311-3	3	0	**Average**	**1.2**	**0.8**
Heterozygotes (F_2s)	F2_22	5	2	F2_32	3	1
	F2_23	3	0	F2_56	1	1
	F2_24	4	2	F2_88	6	2
	F2_25	11	1	F2_89	6	0
	F2_27	2	1	F2_90	1	0
	F2_30	3	0	**Average**	**4.09**	**0.91**

Figure 1. Pedigree relationship of the sequenced *Arabidopsis* samples. The number of circles with solid line denotes how many samples from each generation are sequenced, e.g., the sequenced samples from c52 are equal to 2×1+10×2=22.

Figure 2. **Patterns of diversity, recombination and *de novo* mutation. a.** Relationship between the mutation and recombination rate. When the chromosomes are dissected into 1 Mb non-overlapping regions, recombination rate (cM/Mb) and mutation number per Mb can be obtained for each of them. When ranked then sorted by the recombination rates, the mean mutation rate per recombinational class was obtained. Line is standard regression (for relationship between recombination and diversity see Extended data 3a); **b** & **c**. variation in the mutation rates as a function of heterozygosity proportion during meiosis. Detailed calculation of mutation rates of F_3 to F_4, F_2 to F_3 is shown in Extended data 1b. The number of mutations was counted separately in the regions of F_4s derived from heterozygous or homozygous regions of F_2s, respectively. In 2c the numbers in brackets reflect the proportion of the genome that is hetero- or homo-zygous; **d.** Relationship between nucleotide diversity and the distance to the *de novo* mutations. Window 0 in x-axis is the 2×100 bp sequence surrounding the position of any given *de novo* mutation and 1-9 is 100-900 bp away from the mutation on both sides. For each window of 2×100 bp sequence, the average diversity is calculated. The red dots denote the diversity between Col and Ler, i.e. heterozygosity of parents, the green dots are the average diversity among 80 *Arabidopsis* populations at the same windows[40], and the blue dashes are the average genomic diversity (0.39%) between the two parental genomes (Col and Ler). Error bars, mean ± s.e.m. Test for difference in slope, *Z*=3.08, *P*=0.002.

Methods

Materials and sequencing

We selected two purebred parents in both *Arabidopsis* (strains Col and Ler) and rice (strains 9311 and PA64s) to cross to generate intra-specific F_1 heterozygotes. Col and Ler were female and male parent, respectively. In rice maternal PA64s and paternal 93-11 were crossed to generate their F_1, the super hybrid rice *LYP9*[30]. A single heterozygous F_1 seed in each species was used to generate F_2 progeny. In Arabidopsis two F_2s (lines c52 and c64) were used to generate F_3 and F_4 plants by self-crossing. A total of 67 *Arabidopsis* and 12 rice F_2s and 32 *Arabidopsis* F_4s were randomly selected for sequencing (Fig. 1 and Extended data 1a). In addition the self-crossed homozygous progeny from each pure parent (P_0->P_1) were sequenced (17 Col, 9 Ler, 3 9311 and 3 PA64s). Finally, one each of the four parents and 1 F_1 (in rice) were also sequenced, making a total of 148 plants. Of these, the F_2 and F_4 plants experienced one and three meiosis since F_1, respectively (Fig. 1 and Supplementary Table 1). Col and Ler seeds were gifts from Joy Bergelson at University of Chicago. *Oryza sativa* cultivars PA64s and 93-11 were obtained from Dr. Cailin Wang at the Institute of Food Crops, Jiangsu Academy of Agricultural Sciences, China.

Two DNA samples were extracted separately from two leaves using the CTAB method and sequenced independently for each of *Arabidopsis* parents, their 33 F_2s and all rice plants at BGI-Shenzhen. One DNA sample for the other 34 *Arabidopsis* plants was sequenced. For all, paired-end sequencing libraries with insert size of 500 bp were constructed for each DNA sample according to the manufacturer's instructions. Then, 2×100 bp paired-end reads were generated on Illumina HiSEq 2000.

For the analysis in honey bees (*Apis mellifera ligustica Spin*) three queens and 43 drones were collected from 3 colonies in a bee farm (details described in Liu et al.[17]).

Reads mapping and identification of candidate mutations

The Col genome (TAIR10) was downloaded from TAIR Web site (ftp://ftp.*Arabidopsis*.org/home/tair/Sequences/whole_chromosomes). The assembly Ler scaffolds, SNPs and indels were downloaded from 1001 Genomes (http://1001genomes.org/projects/assemblies.html). The repeat and non-repeat sequences in the genome were grouped by both annotated transposable elements, RepeatMasker regions for *Arabidopsis* (www.repeatmasker.org) and homologous fragments (identity >70%; alignment length >200 bp). Raw reads were cleaned by trimming adapter sequences and removing reads which contain more than 50% low quality bases (quality value≤5). All cleaned reads were mapped to TAIR10 reference genome after trimming and removing low-quality bases by using BWA-MEM (version 0.7.10) algorithm, which shows better performance than several other read aligners to date while mapping 100bp sequences[31]. Then the mapping results were processed using Picard MarkDuplicates to remove over-sequenced DNA molecules. Mapping artifacts introduced while aligning reads on the edges of indels were removed using the GATK package[32,33].

After that, the HaplotypeCaller in GATK package, which incorporates local re-assembly of haplotypes, was employed to call SNPs and indels. This heavily tested protocol, used in 1000 Genomes Project, was chosen as it provides the best reduction in false positives[34]. We joint-genotyped the relevant cohort with all *Arabidopsis* or rice samples and filtered out those sample-specific loci as the initial candidate sets. In these sets, the regions without reads in the parent samples or >8 other samples were excluded.

To ensure the accuracy of calling the *de novo* mutations, numerous stringent strategies were employed (Extended data 1a): (i) in each sample, the candidate "mutation" cannot be called in other non-descendent samples; (ii) the candidate mutation must be called in at least 5 reads and must include both the forward and reverse reads with high variant quality score (≥30 for indel and ≥50 for SNP); (iii) owing to alignment difficulties in the vicinity of indels, those base mutations located around indels (<10 bp each side) between the two parental genomes were removed; (iv) the called indels which have an ≤ 20 bp interval between them were discarded. All alignments were manually inspected in Integrative Genomics Viewer (IGV)[35]. For size distribution of indels see Supplementary Table 8.

Estimation of the possible false positives

The initial filtering may retain a number of false positives due to sequencing, mapping or genotyping errors. We employ a strategy that minimizes the false positive rate, but by necessity likely generates a higher false negative rate. While most of the errors are position-dependent, the mapping errors are less likely to show up in only a single individual in multi-independent samples[36]. Therefore, for any given focal mutation in a focal individual, we examined the reads from the same location in all other members of the cohort and removed those "mutations" where some reads carry the mutation allele in non-focal individuals (excepting descendents). This method becomes especially efficient with increasing sample size. For example with our >100 samples in *Arabidopsis* derived from a single source, all individuals should share the same error rate at the same position. Hence a mutation called in one and only one F_2 is likely to be real. This method is similar to the consensus approach[7], which is ideal with a large number of samples and is robust against sequencing or alignment errors presenting a very low false-positive rate[7].

In addition, we extracted all reads containing candidate mutation loci, and aligned them to the reference sequence in this region using Clustalw 2.0[37]. All alignments of each mutation-associated region were manually inspected by Integrative Genomics Viewer (IGV)[35] to minimize the risk of alignment artifacts and mapping errors. If a region has no companion in the reference genome it is ignored, possibly causing false negatives.

Further, in theory, all of these mutations detected in P_1 and F_2 samples should be heterozygous (the probability that the same mutation occurs in the same position of the genome in two independent meioses is negligible). As expected, only 17 (5.6%) out of the 304 mutations were reported as "homozygous". The residual homozygosity might be caused by biased library construction. In fact, as expected, most (15) of them have a total depth less than or around half of the sequencing depth. These mutations were all verified by PCR as present in the F_2 but absent in the parent (P_0).

Fourth, a true mutation must be heritable and segregating in its progeny but any sequencing error should be not. As expected we detected about half of the mutations called in F_2 generation in their offspring (21 F_3s were randomly sampled from 41 F_2s with seeds and 32 F_4s in Fig. 1). In addition we exclude the possibility of these mutations being present in their parents by PCR amplification and Sanger sequencing.

Finally, the errors could come from a time prior to the sequencing due to somatic mutation, library construction or DNA amplification at an earlier stage. These cases can be estimated by independent DNA extraction and sequencing for the same sample. The 51 plant individuals, each of which has been sequenced twice using DNA samples extracted separately from two leaves, provides an opportunity to test the false positives caused prior to sequencing. Based on those sequences, we found that all of mutations detected are present in both of the independent sequencing libraries.

Estimation of possible false negatives

While the NGS mapping-based method has good accuracy and a low false positive rate in detecting candidate mutations when applied with stringent filtering[6,8,9], the false negative rate remains difficult to estimate accurately, but given our stringency is likely to be considerably higher than our false positive rate. Some false negatives also appear because of technology limitations. For example, that we observe ~5.8% of F_2 mutations as being "homozygous", suggests that we could be missing mutations because they are appearing in the unsequenced component.

We took several approaches to attempt to estimate the false negative rate. In the first we applied the method of simulating mutations described by Keightley et al.[38]. In brief, 1,000 synthetic mutations were simulated by modifying sequencing reads for randomly selected sites in 20 *Arabidopsis* F_2s. Then, we realigned and analyzed the modified data using the same procedures as for the real data. Among these 1,000 synthetic mutations, 897 were considered as callable sites according to the criteria of Keightley et al.. Finally, 880 out of the 897 sites (~98.1%) were directly identified as mutations using our pipeline, suggesting a low frequency of false negatives amongst callable sites (1.9%). This does not however address the problem of mutations missing when the sequence is missing. Indeed, 12% of sites (120 of 1000) are missing.

A more direct way to estimate the false negative rate is to search for mutations found in more than one F_4 progeny, with these F_4s being derived from different F_3s but the same F_2. Such shared mutations most likely were in the F_2 but missed. We can then ask how many we missed in the F_2. In total, we identified 11 shared mutations of which 10 were correctly detected in F_2 ancestors. PCR and Sanger sequencing confirms that the newly identified mutation is really present, but not originally called, in the F_2. This suggests a 9.1%

(1/(10+1)=0.091) lower estimation of mutation rate due to false negatives.

The relationship between the divergence (*A. thaliana vs. A. lyrata*) and the average diversity of 80 *A. thaliana* ecotypes.

The whole-genome alignments between *A. lyrata* and *A. thaliana* -Col were downloaded from VISTA database[39]. Only alignments over 5000 bp were taken for further analysis. Non-unique alignments were discarded. First, the potential substitutions between *A. lyrata* and *A. thaliana* –Col were called. To this end, if the site of substitution was detected as a polymorphic site in the 80 *A. thaliana* ecotypes[40], it was removed (masked) prior to estimating the divergence between *A. thaliana vs. A. lyrata*. Thus only the remaining substitutions, which we presume to be fixed within the population of *A. thaliana,* were used to calculate the divergence between *A. thaliana vs. A. lyrata*. This was done to remove circularity in the divergence-diversity analysis. Only the single base changes at intergenic, intron and fourfold degenerate sites were used to estimate the divergence and diversity.

The intraspecific diversity in any pairwise between-strain comparison was defined as the proportion of relevant sites that are polymorphic per window (i.e. polymorphism density). The average diversity in the above regions among the 80 *A. thaliana* ecotypes was calculated in their corresponding regions of the alignments between *A. lyrata* and *A. thaliana* -Col. The between-ecotype diversity was then defined as the mean pairwise diversity comparing each of the 80 *A. thaliana* ecotypes to each other. The divergence between *A. thaliana* and *A. lyrata* was estimated using baseml with TN93 substitution model implemented in PAML[41].

Calculation of the mutation rate in the meiosis of F_2 to F_3 (EM3) and F_3 to F_4 (EM4).

In this study, 18 F_3s originated from 2 F_2s (c52 and c64) were selfed to produce 34 different F_4s. We define EM3 as the expectation of specific mutations in each F_3 and EM4 as the expectation of specific mutations in each F_4. The mutations shared in F_4s are deduced to be the meiosis mutations of F_2 to F_3 (Extended data 1b). However, some of the meiosis mutations of F_2 to F_3 have been lost due to the random drift or have been classified as F_4

specific mutations due to absence in other F_4s. Therefore, there may be an overestimate for these F_4 specific mutations generated in the meiosis of F_3 to F_4, and there may be an underestimate for these shared mutations in these F_4s which have been generated in the meiosis of F_2 to F_3.

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 estimate of new homozygous mutations per generation, and τ, the estimate of new heterozygous mutations per generation[8]:

$$\mu_3\text{: } \mu \text{ From } F_2 \text{ to } F_3$$

$$\mu_4\text{: } \mu \text{ From } F_3 \text{ to } F_4$$

$$\tau_3\text{: } \tau \text{ From } F_2 \text{ to } F_3$$

$$\tau_4\text{: } \tau \text{ From } F_3 \text{ to } F_4$$

For one generation, the estimate of mutations=Nμ+Nτ, N is the number of organisms in this generation.

For two generations, the estimate of mutations=$N_1\ \mu_1$+$N_1P\tau_1$+$N_2\mu_2$+$N_2\tau_2$. Here P is the probability that the heterozygous mutation in the first generation is inherited by its progenies, which depend on the number of its progeny.

In our study, as we have 6 F_3 with 1 progeny, 10 F_3 with the formula can be changed to:

All mutations observed in F_4=$18\mu_3 + (6P_1 + 10P_2 + 2P_3)\tau_3 + 32\mu_4 + 32\tau_4$

P_n is the likelihood that a mutation in a F_3 with n progenies is inherited.

For a heterozygous mutation in F_3 with n progeny, the probability that no progeny genotype is homozygous mutation (a/a) is 0.75^n, and at least one of the progeny carry homozygous mutation is 1-0.75^n.

For all the homozygous mutations in F_4:

$$18\mu_3 + 32\mu_4 + [6 \times (1 - 0.75) + 10 \times (1 - 0.75^2) + 2 \times (1 - 0.75^3)]\tau_3$$
$$= 19 + 3 \qquad (1)$$

For a heterozygous mutation in F_3, the probability that it is not inherited by the progenies is 0.25^n and the probability that the mutation appears as heterozygous in F_3s is $0.75^n - 0.25^n$. For F_3 with 1, 2 or 3 progenies, the likelihood is 0.5, 0.5 and 0.40625.

For all the heterozygous mutations in F_4:

$$(6 \times 0.5 + 10 \times 0.5 + 2 \times 0.40625)\tau_3 + 32\tau_4 = 54 + 6 \qquad (2)$$

The shared mutations in F_4 can be counted as the result of mutations in F_3. As shown (Extended data 1b) for the shared heterozygous mutations in F_4:

$$[10 \times 0.25 + 2 \times (0.5^3 + 3 \times 0.5^2 \times 0.25)]\tau_3 = 6 \qquad (3)$$

$$\tau_3 = 1.92$$

According to equation (2) and (3):

$$\tau_4 = 1.346$$

According to equation (1) and (3):

$$18\mu_3 + 32\mu_4 = 8.5 \qquad (4)$$

$$\mu_3 + 1.78\mu_4 = 0.472$$

$$\mu_3 \leq 0.472$$

$$\mu_4 \leq 0.266$$

If a homozygous mutation occurred in 10 F_3s with 2 progenies or 2 F_3s with 3 progenies (counted as μ_3), all of its progeny will carry homozygous mutations, which was not found in our result, so μ_3 was assumed to be 0.

$$\mu_3 = 0$$

$$\mu_4 = 0.266$$

$$EM3=\mu_3+\tau_3=1.92$$

$$EM4=\mu_4+\tau_4=1.612$$

Therefore, the mutation rates of F_2 to F_3 or F_3 to F_4 should be 1.60×10^{-8} or 1.34×10^{-8}, respectively.

Distribution of mutations and statistical analyses

To determine the distribution of mutations on chromosomes (Extended data 3d), the *de novo* mutations were used from our sequenced 26 P_1s, 67 F_2s, 32 F_4s, and two published data sets[8,9], all of which employed the ecotypes of Col, Ler or the offspring of Col×Ler. The recombination (crossover) data was collected from our 67 F_2s and 32 F_4s (Supplementary Table 9).

To know whether proximity to heterozygous sites could affect the mutation rate, we calculated the distance of the new mutations to heterozygous sites. To detect whether the observed mutations tend to arise on derived versus ancestral alleles, we make use of the alignments, described above, between *A. thaliana* (Col) and *A. lyrata*. If the same aligned nucleotide is seen in both *A. lyrata* and *A. thaliana* (prior to mutation) it was presumed to reflect the ancestral state. A total of 201 mutations (158 SNPs and 43 indels) have a clear ancestral state. Of the remaining 199, 93 have no alignments, 59 are in the gaps of *A. lyrata*, 15 are ambiguous due to non-unique alignments, and 32 have a different nucleotide compared to *A. thaliana* thus preventing ancestral state determination.

To estimate the expected number of mutations in heterozygous and homozygous compartments under a null expectation that heterozygosity per se is not a relevant parameter, we factor in both the absolute size of both compartments and, for point mutations, the trinucleotide content. The GC content of sequence flanking indels (35%) is almost identical to that of the genomic average (36%) so we make no nucleotide content correction for these. Given a total observed set of mutations, we calculate a mutation rate per given trinucleotide triplet, with the mutation centered with the triplet. We then for each compartment calculate the total number of each triplet to generate an expected number of point mutations per triplet. We then sum across all triplets to derive an expected total number of mutations in a given compartment. As an internal consistency check we calculate the sum across the two compartments, ensuring that this is the same as the observed total number of mutations. We thus have both observed and expected (allowing for nucleotide content and span length) number of mutations. For indels we just consider the proportion of all sequence in each compartment. We test for difference by chi squared test with Yate's correction.

Statistics were performed in R[42]. Brunner-Munzel (BM) test was implemented in lawstat package. When *P* values were derived from randomization, 10,000 randomizations were employed in which the data was randomly ascribed by shuffling of class (e.g. heterozygous or homozygous). The unbiased estimation of empirical *P*, meaning expected type I error rate, is n+1/m+1 where n is the number of observations as or more extreme than that observed in the real test reporting statistic and m is the number of randomization[43].

30. Gao, Z.-Y. *et al.* Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. *Proc. Natl. Acad. Sci.* (2013). doi:10.1073/pnas.1306579110

31. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013). at <http://arxiv.org/abs/1303.3997>

32. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

33. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

34. Ghoneim, D. H., Myers, J. R., Tuttle, E. & Paciorkowski, A. R. Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. *BMC Res. Notes* **7,** 864 (2014).

35. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).

36. Li, M. & Stoneking, M. A new approach for detecting low-level mutations in next-generation sequence data. *Genome Biol.* **13,** R34 (2012).

37. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).

38. Keightley, P. D., Ness, R. W., Halligan, D. L. & Haddrill, P. R. Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a Drosophila melanogaster Full-Sib Family. *Genetics* **196,** 313–320 (2014).

39. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32,** W273–W279 (2004).

40. Cao, J. *et al.* Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **43,**

956–963 (2011).

41. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).

42. R Development Core Team. R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. (2013).

43. North, B. V., Curtis, D. & Sham, P. C. A Note on the Calculation of Empirical P Values from Monte Carlo Procedures. *Am. J. Hum. Genet.* **72,** 498–499 (2003).

Extended data 1. Details of Materials and Methods. **a.** Schematic diagram of the detection of de novo mutations. **b.** The calculation of the expected mutations in the meiosis of F2 to F3 (EM3) and F3 to F4 (EM4). For further explanation see Methods.

Extended data 2. Mutational properties. a. Spectra of nucleotide substitutions in *Arabidopsis* and rice. **b.** Co-occurrence of mutations and crossover break points in bee. By using the sequence data of 43 honey bee drones and their 3 corresponding queens17, a total of 27 base and 8 indel mutations were detected. Noteworthy, 2 of 35 mutations are found in close proximity with crossover break points in the same sample (Distance<2 Kb; P =0.0012 with 10,000 randomizations), these ones being illustrated here. The crossover event is between the red and blue line with marker positions annotated. The positions of the mutations are annotated with arrows. **c.** The schematic diagram of the genomic structures and the possible pairings of two homologous chromosomes during the meiosis at two mutated *LRR-TM* genes (top panel) and one mutated *NBS-LRR* gene (bottom panel). The top panel shows the genomic structures between Col and Ler at the loci of AT3G23110, the receptor-like protein 37 with a non-synonymous mutation (Chr3:8224726, T→C) at sample of c74, and AT3G23120, the receptor like protein 38 with a deletion mutation (Chr3:8228194, Del:C, frameshift) at sample of c70. The bottom panel illustrates the genomic structures between two Col chromosomes at AT1G59780 and the mutations detected in a homozygous plant of Col[8]. Red arrows represent the position of mutation; the hatched areas indicate the identical sequences, the other regions being highly diversified; the dotted lines indicate the paired length of the homologs at the highly identical regions. During meiosis, possible pairings between parental chromosomes are illustrated, where the loops indicate the unpaired regions

Extended data 3. Correlation between mutations, recombination events, diversity and divergence. a. The relationship between nucleotide diversity (Col vs. Ler) and recombination rate. When the chromosomes were dissected into 100 kb non-overlapping windows, the diversity (polymorphism density) between Col and Ler and recombination rate in 67 F2s and 32 F4s were calculated for each window. When sorting the windows by the diversity and dividing them into 8 equal intervals (e.g., from 0 to 0.001, 0.001 to 0.002, 0.002 to 0.003……, and so on), the relationships between the average diversity and recombination rate is displayed. Error bars indicate s.e.m. **b.** The relationship between diversity and divergence. The black line represents standard linear regression line and is for illustrative purposes alone. The statistic is the result of Spearman's rank correlation. **c.** Relationship between mutation and distance to polymorphic sites. The mutation data were collected from our 67 F2s. Window 0 in x-axis is the 2×100 bp sequence surrounding the position of any given de novo mutation and 1-9 is 100-900 bp away from the mutation on both sides. For each window of 2×100 bp sequence, the average diversity is calculated. The black dots denote the average pairwise diversity among the published 80 Arabidopsis ecotypes; the red dots denote the average diversity between Col and the 80 ecotypes; the blue dots denote the average diversity between Ler and the 80 ecotypes. Error bars indicate s.e.m. **d.** Distribution of the mutations on the chromosomes**.** The gray vertical bars in the chromosomes denote the position of all collected mutations. When the chromosomes were dissected into 1Mb non-overlapping windows, the mutation numbers (blue shadow in the figure) were counted in each window. The red lines denote the average pairwise diversity among the published 80 Arabidopsis ecotypes