The rate and effect of *de novo* mutations in natural populations of *Arabidopsis thaliana*

Moises Exposito-Alonso[1,2†], Claude Becker[1†], Verena J. Schuenemann[3,4], Ella Reitter[3], Claudia Setzer[5], Radka Slovak[5], Benjamin Brachi[6§], Jörg Hagmann[1§], Dominik G. Grimm[1§], Chen Jiahui[6,7], Wolfgang Busch[5], Joy Bergelson[6], Rob W. Ness[8], Johannes Krause[3,4,9], Hernán A. Burbano[2,*], Detlef Weigel[1,*].

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[2]Research Group for Ancient Genomics and Evolution, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[3]Institute of Archaeological Sciences, University of Tübingen, 72070 Tübingen, Germany

[4]Senckenberg Center for Human Evolution and Paleoenvironment, University of Tübingen, 72070 Tübingen, Germany

[5]Gregor Mendel Institute, Austrian Academy of Sciences, 1030 Vienna, Austria

[6]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA

[7]Institute of Tibet Plateau Research, Chinese Academy of Sciences, Beijing 100101, China

[8]Department of Biology, University of Toronto Mississauga, Mississauga, Ontario L5L IC6, Canada.

[9]Max Planck Institute for the Science of Human History, 07743 Jena, Germany

[†]Co-first authors

[§]Current addresses: INRA, UMR 1202 Biodiversité Gènes & Communautés, 69 route d'Arcachon, 33610 CESTAS, France (B.B.); Computomics, 72072 Tübingen, Germany (J.H.); Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland (D.G.G.).

[*]Correspondence: weigel@weigelworld.org (D.W.), hernan.burbano@tuebingen.mpg.de (H.A.B.)

Short title: Mutation and selection in *Arabidopsis thaliana*

Character count: 68,031 (with spaces, without table & figures)

Keywords: colonization; mutation; herbarium; phylogenomics, population genomics; *Arabidopsis thaliana*

SUMMARY

Like many other species, the plant *Arabidopsis thaliana* has been introduced in recent history from its native Eurasian range to North America, with many individuals belonging to a single lineage. We have sequenced 100 genomes of present-day and herbarium specimens from this lineage, covering the time span from 1863 to 2006. Within-lineage recombination was nearly absent, greatly simplifying the genetic analysis, allowing direct estimation of the mutation rate and an introduction date in the early-17th century. The comparison of substitution rates at different sites throughout the genome reveals that genetic drift predominates, but that purifying selection in this rapidly expanding population is nevertheless evident even over short historical time scales. Furthermore, an association analysis identifies new mutations affecting root development, a trait important for adaptation in the wild. Our work illustrates how mutation and selection rates can be observed directly by combining modern genetic methods and historic samples.

HIGHLIGHTS

- A historically young colonizing lineage of *Arabidopsis thaliana* allows observation of contemporary evolutionary forces.
- Genomes from specimens collected over 150 years support direct calculation of mutation rates occurring in nature.
- Drift predominates, but purifying selection is evident genome-wide over historical time scales.
- New mutations with phenotypic effects can be identified and traced back in time and space.

INTRODUCTION

If we want to understand evolution and especially adaptation, we need to know rates of mutation and selection, which together determine the substitutions that can be observed in a population. Typically, one tries to infer evolutionary parameters from patterns of genetic diversity in extant individuals of a species. Unfortunately, demographic and genetic factors such as migration, fluctuating population sizes, recombination and gene conversion greatly complicate such inferences. Many scientists have therefore chosen to focus on mutations only, measuring their accumulation in artificial conditions, using mutation accumulation lines grown in the laboratory (Halligan and Keightley, 2009), or over very short time scales, for example in human parent-offspring trios (Roach et al., 2010).

An alternative approach is the use of older, but still simple lineages with limited genetic diversity, such as colonizing populations that have undergone a recent, strong genetic bottleneck. Such populations can be considered natural experiments in which one can test ecological or evolutionary hypotheses (Gauze, 1934; Maron et al., 2004; Sax et al., 2007). Recent colonization events permit the quantification of evolutionary forces related to adaptation – mutation, selection, genetic drift, recombination – that are still not well understood in invasion ecology (Barrett, 2014; Bock et al., 2015; Lee, 2002).

Humans have increasingly blurred biogeographical boundaries of species outside their native range by planned or serendipitous dissemination. While the exact reasons for success or failure in alien environments remain unclear, many species can become established in new areas, with North America being the continent with the highest number of naturalized plants (van Kleunen et al., 2015). Among these is the model plant *Arabidopsis thaliana*, which is native to Eurasia but has recently colonized and spread throughout much of North America (Platt et al., 2010). Although *A. thaliana* is not an invasive species, it has traits typical for successful colonizers, such as a high selfing rate, a durable seed bank and a short generation time (Baker, 1965).

Colonizing populations often start with very few individuals and therefore have low genetic diversity. The N. American *A. thaliana* population is much less diverse than what is seen in the native range, with one predominant lineage, named haplogroup-1 (HPG1), accounting for about half of all N. American individuals (Platt et al., 2010). The success of an isolated, selfing lineage that is genetically very uniform seems to contradict the common idea that such lineages are evolutionary dead-ends because they can adapt only through *de novo* mutations, a process predicted to be much slower than adaptation from standing variation (Barrett and Schluter, 2008), although we cannot know how long this lineage will last.

Ideally, to evaluate all evolutionary trajectories, including unsuccessful ones, one should have access not only to the evolved extant individuals, but also to their "unevolved" ancestors. The power of temporal transects has been aptly demonstrated with the genetic analysis of historical and archaeological samples of humans and microbes, relying on advances in the study of ancient DNA (aDNA) (Orlando et al., 2015; Shapiro and Hofreiter, 2014). Natural history collections that cover the past several hundreds of years offer an exciting, underused resource for such studies (Martin et al., 2013; Staats et al., 2013; Vandepitte et al., 2014; Weiß et al., 2015; Yoshida et al., 2013).

There is a rich history of sampling plants and storing them in herbaria. Importantly, herbaria do not merely house exotic, rare species collected in the more distant past, but also common plants that have been sampled for many decades over and over again, making them powerful tools for monitoring

89 recent colonization events in space and time (Crawford and Hoagland, 2009; Lankaua et al., 2009). Such
90 a resource exists for N. American *A. thaliana*. Here, we compare genomes from herbarium specimens,
91 collected between 1863 to 1993, and from live individuals, collected between 1993 and 2006, to date
92 the origin of this lineage, and infer mutation rates, selection, demography and migration routes. We also
93 identify *de novo* mutations in this lineage that are associated with phenotypes likely to be under selection
94 in the wild, which in turn correlate with climatic variables. Our analyses of a colonizing *A. thaliana* lineage
95 serve as a blueprint for future studies of similar colonizing or otherwise recently bottlenecked
96 populations, in order to understand mutation, selection and rapid adaptation in nature.

97 RESULTS AND DISCUSSION

98 Herbarium and modern HPG1 genomes

99 When analyzed with 149 genome-wide, intermediate-frequency SNP markers, about half of over 2,000
100 North American *A. thaliana* individuals collected between 1993 and 2006 were found to be very similar
101 (Platt et al., 2010). A recent study of 13 individuals from this collection confirmed that their genomes
102 were indeed almost identical (Hagmann et al., 2015). We selected 74 additional individuals for Illumina
103 whole-genome sequencing, aiming for broad geographic representation, and, where available, at least
104 two accessions per collection site (Fig. 1; Table S1).

105 We aimed to complement these data with genome information from 36 herbarium specimens
106 collected between 1863 and 1996 (Fig. 1; Table S1). To avoid contamination from exogenous sources,
107 DNA extraction and Illumina library preparation were carried out in a clean-room facility. Between 30%
108 and 86% of sequencing reads mapped to the *A. thaliana* reference genome (Fig. S1A), compared to ~90%
109 for the modern individuals. A number of biochemical features define aDNA and can be used to verify
110 authenticity (Krause et al., 2010; Prüfer and Meyer, 2015; Weiss et al., 2015). Typical for aDNA, most
111 DNA fragments were shorter than 100 bp (Fig. S1B). Deamination of cytosines to uracils at the end of
112 aDNA fragments (Hofreiter et al., 2001) is seen as cytosine to thymine (C-to-T) substitutions upon
113 sequencing (Briggs et al., 2007), and this rate of substitution at first sequenced base was between 1.3 to
114 4.4% in the different sequenced libraries (Fig. S1C). Moreover, aDNA breaks preferentially at purines
115 (Briggs et al., 2007), and purines were 1.5 – 1.8-fold enriched at fragment ends (Fig. S1D). Together this
116 indicated that DNA recovered from *A. thaliana* herbarium specimens was authentic.

117 Coverage of sequenced historic samples was 3- to 42-fold for herbarium and 22- to 105-fold for
118 modern samples. To identify within-lineage sequence differences, reads were mapped against an HPG1
119 pseudoreference (Hagmann et al., 2015). We focused on single nucleotide polymorphisms (SNPs)

120 because accurate identification of structural variants from short reads is difficult, particularly so in old
121 DNA molecules that have suffered from chemical breakage (Weiß et al., 2015). The herbarium genomes
122 subsequently confirmed as HPG1 had 96.8 to 107.2 Mb of the HPG1 pseudoreference covered by at
123 least three reads, compared with 108.0 to 108.3 Mb in the modern genomes. We found 109 to 222
124 SNPs relative to the HPG1 pseudoreference in the herbarium genomes, and 186 to 299 SNPs in the
125 modern genomes.

126 **Diversity and relationships within HPG1**

127 Among the 87 modern individuals, seven clearly did not belong to the HPG1 lineage, which could be due
128 to errors in the initial genotyping, or to lack of resolution based only on 149 SNPs. Four additional
129 individuals that were identical to the rest of the HPG1 lineage at the 149 genotyped SNPs (Fig. S2A)
130 appeared to have small stretches of introgression from other lineages and were therefore classified as
131 non-HPG1, as indicated by several methods (e.g., Fig. S2B). Of the 36 herbarium samples, nine turned
132 out to be non-HPG1 lines (Fig. S2A and S2B). In total, 76 modern and 27 herbarium samples were
133 identified as HPG1 by means of neighbor-joining trees and multidimensional scaling (MDS), including the
134 12 oldest herbarium specimens (Fig. S2C). The obvious homogeneity and abundance of HPG1 compared
135 to other N. American lineages greatly simplified its classification.

136 After removal of non-HPG1 lines, the HPG1 neighbor-joining tree reconstruction resulted in a
137 star-like phylogeny (Fig. 2A). MDS could not differentiate samples within the HPG1 group, with the first
138 and second dimensions each explaining only small amounts of variance, 8.8% and 8.0% (Fig. 2B). A
139 parsimony network identified a small fraction of reticulations indicative of intra-HPG1 recombination
140 (Fig. 2C). Removing three potential intra-HPG1 recombinants resolved the reticulations (Fig. 2D). The
141 remaining 73 modern and 27 herbarium samples (Table S1) appeared to constitute a clonal lineage
142 devoid of effective recombination and population structure, with no SNPs detected in chloroplasts nor
143 mitochondrial genomes, and with very low genome-wide nuclear diversity (π = 0.000002, θ_W = 0.00001,
144 4,368 segregating sites), which is two orders of magnitude lower than in the native range (θ_W = 0.008)
145 (Cao et al., 2011; Nordborg et al., 2005). The enrichment of low frequency variants (Tajima's *D* = -2.84)
146 and low levels of polymorphism in surveyed genomes is consistent with a recent bottleneck followed by
147 population expansion. We hypothesize that the bottleneck corresponds to a colonization founder event,
148 likely by one or only few very closely related individuals.

149 Although there was little evidence for intra-lineage recombination among the 100 remaining
150 individuals, a few isolated SNPs were shared between independent branches of the tree (Fig. 2A). We
151 therefore also formally estimated recombination within HPG1. The estimate was much lower ($4N_er = \rho$

152 = 3.0 x 10^{-6} cM bp^{-1}) than for a similar-sized collection of diverse *A. thaliana* individuals from the native
153 range (ρ = 7.5 x 10^{-2} cM bp^{-1}) (Choi et al., 2013). Linkage disequilibrium parameter *D*' did not decay with
154 physical distance (intercept = 0.99, slope = 0.00, $p < 0.0001$) among all SNP pairs. Furthermore, only
155 0.02% of SNP pairs were not in complete linkage disequilibrium (D'<1), indicating extensive linkage
156 between chromosomes. The four-gamete test, which determines whether all four possible gametes (ab,
157 aB, Ab, AB) are observed for two segregating loci, revealed that all configurations of SNPs could be
158 explained with as few as 38 recombination events for the 100 genomes. We argue that this number of
159 potential recombination events is sufficiently small that it does not invalidate the application of
160 phylogenetic methods to the HPG1 genomes, even though such methods are normally not appropriate
161 for genome-wide analyses. Indeed, other sources of failure of the four-gamete test and the violation of
162 phylogenetic assumptions could be sequencing errors, or lineage sorting of segregating sites from the
163 ancestral population.

164 To describe intra-HPG1 relationships in a more sophisticated manner than with a simple
165 neighbor-joining approach, we used Bayesian phylogenetic inference. We took advantage of the broad
166 distribution of collection dates of our herbarium samples (Fig. 1B) for tip calibration of phylogenetic
167 trees. The method that we used reconstructs a tree calibrated in time, based on genetic distance
168 between samples collected at different points in time. In this tree, the 76 modern individuals formed a
169 virtually monophyletic clade, with only four interspersed herbarium samples from the second half of the
170 20th century (Fig. 3A, B, Table S1). Geographic proximity did not explain the close genetic relationship of
171 these four herbarium and the modern individuals (Fig. 1, Table S1).

172 **Estimates of mutation rate and spectrum in the wild**

173 To estimate the substitution rate in the HPG1 lineage, we used a distance- and a phylogeny-based
174 method, both of which take advantage of the collection dates of our samples. It is necessary to
175 distinguish between substitutions and mutations. The substitution rate is the observed cumulative
176 change in DNA that results from several evolutionary forces, such as demography and natural selection.
177 These forces act in concert on the new mutations produced by DNA damage, repair and replication
178 errors, which are presumed to be constant over time (Barrick and Lenski, 2013).

179 In the distance method, the substitution rate is first calculated from the correlation of distances
180 of collecting dates with genetic distances, as measured in number of substitutions, then scaled to the size
181 of the genome accessible to Illumina sequencing (Fig. 3C). With this method, we estimated a rate of 3.3
182 x 10^{-9} substitutions $site^{-1}$ $year^{-1}$ (95% bootstrap Confidence Interval [CI]: 2.9 to 3.6 x 10^{-9}). If one
183 changes the thresholds for base calling, this affects both the number of called SNPs, and the fraction of

184 the genome that is interrogated for variants. We therefore explored how either more relaxed or more
185 stringent base calling methods affected our substitution rate estimates. We used three quality thresholds
186 of increasing stringency (see Experimental Procedures for details) and found that the impact was
187 negligible, with mean substitution rate estimates ranging from 3.0 to 4.0 x 10^{-9}, compared to our
188 standard threshold, which had given 3.3 x 10^{-9} substitutions $site^{-1}$ $year^{-1}$.

189 The Bayesian phylogenetic approach uses the collection years as tip-calibration points; its
190 application resulted in a very similar estimate, 4.0 x 10^{-9} substitutions $site^{-1}$ $year^{-1}$ (95% Highest Posterior
191 Probability Density [HPD]: 3.2 to 4.7 x 10^{-9}). We confirmed MCMC chain convergence on demographic
192 and tree topology parameters by repeating the analysis with this rate. The stability of all parameters
193 indicated that under a low complexity scenario with no population structure or recombination,
194 phylogenetic and population genetic methods generate congruent evolutionary rates.

195 Under neutral evolution, substitution and mutation rates should be the same, but typically
196 substitution rates are expressed per year, whereas mutation rates are expressed per generation, among
197 other conceptual differences (Barrick and Lenski, 2013; Kimura, 1967). Although *A. thaliana* has an
198 annual life cycle, the generation time in nature has been estimated to average 1.3 years (Lundemo et al.,
199 2009), because seeds could potentially survive 3 to 5 years in a seed bank (Montesinos et al., 2009). To
200 correctly compare the substitution rates from our study with mutation accumulation lines propagated in
201 the greenhouse (Ossowski et al., 2010), we re-scaled the estimated substitution rate by the 1.3 year
202 average, resulting in 4.2 x 10^{-9} substitutions $site^{-1}$ $generation^{-1}$ (95% CI 3.7 to 4.7 x 10^{-9}) (Fig. 3E, Table
203 S3).

204 To obtain the best possible estimate of short-term mutation rates for comparison, we
205 reanalyzed a recent re-sequencing dataset of mutation accumulation lines grown in the greenhouse
206 (Hagmann et al., 2015); from this, we confirmed a rate of 7.1 x 10^{-9} mutations $site^{-1}$ $generation^{-1}$ (95% CI
207 6.3 to 7.9 x 10^{-9}) (see Table S2 and Extended Experimental Procedures). In several species, including
208 *Escherichia coli* (Sniegowski et al., 1997) and *A. thaliana* (Jiang et al., 2014), growth under abiotic stress
209 can increase mutation rates. Although wild conditions can be considered moderately stressful
210 environments compared to standard greenhouse conditions, we found the generation-corrected
211 substitution rate in the HPG1 lineage to be lower than the mutation rate in greenhouse lines. The
212 mutation spectrum was, however, closer to that of greenhouse lines exposed to salt stress (Jiang et al.,
213 2014) than to the greenhouse lines grown under standard conditions (Ossowski et al., 2010) (Fig. 3D).
214 One possible contributor to a shift in mutation spectrum is DNA methylation, since methylated
215 cytosines are more likely to undergo substitutions than unmethylated cytosines, something that has been
216 observed in other natural accessions (Cao et al., 2011; Hagmann et al., 2015).

217 **Genome-wide inference of selection**

218 One likely explanation for the unexpected differences between the greenhouse mutation rate and our
219 estimate from the HPG1 population (Fig. 3E) is the effect of purifying selection, which should slow the
220 accumulation of mutations in the wild. In other organisms, including humans, estimates of short- and
221 long-term mutation rates differ considerably and have motivated a hot debate (Ho et al., 2005; Scally and
222 Durbin, 2012; Ségurel et al., 2014; Subramanian and Lambert, 2011). In humans, counterintuitively,
223 pedigree-based short-term estimates of nuclear mutation rates are lower (Kong et al., 2012; Roach et
224 al., 2010) than long-term estimates based on interspecific phylogenies (Nachman and Crowell, 2000).
225 Recently, the use of DNA retrieved from dated fossils (Fu et al., 2014) and new methods incorporating
226 recombination map scaling (Lipson et al., 2015) have produced more concordant, intermediate mutation
227 rates estimates. That long-term rates are lower is expected, since purifying selection would have had
228 more time to effectively remove deleterious mutations from the population. Indeed, older calibrating
229 points in human-great ape phylogenies have yielded lower substitution rate estimates (Subramanian and
230 Kumar, 2003). Alternatively, long- and short-term rates may really be different, because of changes in
231 generation times or fluctuating mutation rates (Green and Shapiro, 2013). Discrepancy could perhaps
232 also come from intra-specific variation in mutation rates (e.g. the effect of genetic background), reported
233 to be more than 7-fold across genotypes of *Chlamydomonas reinhardtii* (Ness et al., 2015). This, however,
234 does not seem to apply when comparing natural and greenhouse populations. Phylogenetic and
235 regression-based methods produced very similar estimates for the HPG1 population, and were similar
236 to mutation rate measurements in a greenhouse population with an exactly known number of
237 generations. We attribute the small differences between the *A. thaliana* populations to either the
238 efficiency of purifying selection over different temporal and environmental scales or to imperfect
239 knowledge of generation time.

240 To test the purifying selection hypothesis, we compared mutation rates in differently annotated
241 portions of the genome. Ideally, one would compare synonymous substitutions at four-fold degenerate
242 sites with non-synonymous substitutions, but there were too few of such substitutions in our data set to
243 achieve appropriate statistical power (on average 0.9 four-fold and 2.7 nonsynonymous mutations per 30
244 generations in mutation accumulation lines). We therefore used the net distances method to compare
245 rates in intergenic regions with whole-genome rates. The comparison of mutation rates across
246 annotations supported the hypothesis that purifying selection is the cause of different mutation rate
247 estimates in the HPG1 and greenhouse populations. The estimate for whole-genome rates was 33.59%
248 (95% CI 33.59 – 33.60) lower than the intergenic estimate in the HPG1 lineage, compared to 26.04%
249 (95% CI 21.44 - 29.31%) in the greenhouse population (Fig. 3E). In addition, medium-frequency variants

250 (4% ≤ allele frequency ≤ 50%) were more strongly depleted in the whole-genome set compared to
251 intergenic regions (Fisher's Exact test, p=0.03) in the HPG1 linage.

252 The observed rate at which new mutations accumulate in populations, the substitution rate,
253 depends on both the number of individual genomes in the population in which mutations occur, for
254 diploid species $2\ N_e$, and on the selection coefficient *s*, affecting the probability of fixation of a mutation.
255 When selection is negligible and only genetic drift operates, the probability of fixation of a new mutation
256 is equal to its frequency ($1/2\ N_e$). Under neutrality, the observed rate at which mutations accumulate
257 equals the rate at which mutations arise. If we assume that the behavior of intergenic substitutions is
258 close to neutrality, we can use it as the reference mutation rate, μ, and compare it with the genome-
259 wide substitution rate, *k*, to solve for the genome-wide selection coefficient of the fixation probability
260 equation from Kimura (1967). The coefficient responsible for the genome-wide deficit in substitutions
261 was $N_e\ s = -0.76$. Only a coefficient scaled by population size is meaningful in our context, since theory
262 predicts that selection is efficient when $N_e\ |s| > 1$, where $|s|$ is the absolute value of a hypothetical semi-
263 dominant genome-wide selection coefficient. Our estimate is negative, suggesting a net effect of purifying
264 selection, but its value is smaller than 1, indicating that the number of substitutions is largely determined
265 by population drift (Charlesworth and Charlesworth, 2010).

266 We were curious whether our result of net inefficient purifying selection is related to the
267 mating system, namely predominant selfing, or the recent genetic bottleneck of the HPG1 lineage. A
268 previous point estimate of the coefficient of selection $N_e\ s$ in *A. thaliana* was ~ -0.8, using an approach
269 based on polymorphism within *A. thaliana* and divergence between *A. thaliana* and its close relative *A.*
270 *lyrata* in 12 nuclear genes (Bustamante et al., 2002). The same study reported that in the genus
271 *Drosophila* $N_e\ s$ was positive and greater than one, indicative of widespread and effective selection. The
272 authors hypothesized that in highly selfing species, N_e decreases due to inbreeding, reducing the ability of
273 selection to purge slightly to moderately deleterious mutations, consistent with other studies
274 (Charlesworth and Wright, 2001; Ness et al., 2010; Wright et al., 2008).

275 We recognize that averaging selection coefficients across the entire genome may be
276 inappropriate if different genomic features are under very different selection regimes, resulting in a
277 highly dispersed or even bimodal distribution of selection coefficients. Point estimates should therefore
278 be treated with caution. Keightley and Eyre-Walker (2007) showed that this is the case in humans, by
279 estimating the distribution of purifying selection coefficients using the distribution of predicted fitness
280 effects of various polymorphisms. They found, however, that this did not apply to *Drosophila*
281 *melanogaster,* where almost the entire genome was under strong purifying selection, with $N_e\ s > 100$
282 (Keightley and Eyre-Walker, 2007). A case that may resemble more closely HPG1 evolution is that of

283 the plant *Eichhornia paniculata*, which experienced a recent intra-species transition to selfing. As a
284 consequence, purifying selection coefficients have become more broadly distributed, with the
285 proportion of almost neutral coefficients having increased due to low N_e, and the proportion of strongly
286 negative coefficients also having increased due to homozygosity, which uncovers recessive deleterious
287 sites (Arunkumar et al., 2015). Given these studies and our average selection coefficient estimate, we
288 hypothesize that a combination of brief evolutionary history and low N_e has reduced the efficiency of
289 natural selection, with only highly deleterious mutations being eliminated. More information could be
290 obtained by developing new models and performing simulations of site frequency spectra that include
291 different demographic scenarios in combination with selfing.

292 **Phenotypic effect and spatiotemporal context of *de novo* mutations**

293 In the HPG1 lineage, drift seems to determine genome-wide polymorphism patterns, but there is some
294 evidence for purifying selection. We wondered whether, in addition, we would be able to find signals of
295 adaptive, positive selection, expected to be much rarer and thus much more difficult to detect. Selection
296 scans based on population divergence or haplotype sharing decay are inappropriate when divergence
297 between samples is low and/or when there is high intra- and inter-chromosomal linkage disequilibrium.
298 We therefore adopted an association approach in an effort to link segregating mutations to climatic
299 variables as well as phenotypic variation in several traits of likely ecological relevance: flowering
300 phenology, fruit set (fecundity), seed size, root growth and morphology. Replicated measurements of
301 phenotypic traits in controlled conditions showed significant quantitative variation between lines as
302 described by broad sense heritability (Table S4). HPG1 individuals resemble near isogenic lines (NILs) in
303 that they share large segments of the genome. Formally, genetic mapping with NILs seeks to associate
304 phenotypes with large blocks of linked variants. It has been successfully used to examine the genetic
305 basis of many different traits in crop species (Brouwer and St Clair, 2004; Stec et al., 2013; Szalma et al.,
306 2007; Xie et al., 2006) and also in *A. thaliana* (Bentsink et al., 2010; Fletcher et al., 2013; Keurentjes et al.,
307 2007; Swarup et al., 1999; Weigel, 2012). Our approach has the advantage that it can discern the
308 phenotypic effects of a limited number of mutations free from confounding population structure (see
309 Extended Experimental Procedures). In association analyses, statistical power relies on variants with a
310 certain minimum frequency, hence we only considered ~400 variants with at least 5% allele frequency.
311 These are, however, not independent due to linkage disequilibrium, thus rather comprise haplotypes
312 (Templeton et al., 1988). Focusing on intermediate frequency variants not only increases statistical
313 power, but is also more likely to reveal adaptive mutations, because intermediate frequency variants will
314 be on average older and less likely to be deleterious.

315 With permutation tests to assess significance, we found several root phenotypes to be
316 significantly associated with 79 SNPs. Thirty-six of these were in protein coding genes and nine resulted
317 in non-synonymous substitutions. Nineteen other SNPs were associated with climate variables
318 (www.worldclim.org/bioclim) even after correction for latitude and longitude. Eight of these were
319 located in genes, and four resulted in non-synonymous substitutions (Table 1, Table S4, S5). We did not
320 find SNPs that were significantly associated with flowering, fecundity or seed size. In addition to
321 permutation testing, we applied a Bonferroni corrected significance threshold to account for multiple
322 traits tested. As an alternative to the permutation approach, we adjusted the significance threshold for
323 multiple traits and SNPs tested. Even with these two very conservative approaches, 13 and four genic
324 SNPs remained significant for root phenotypes and climate variables, respectively (Table 1).

325 The most common climate variable with significant SNP associations was precipitation during
326 the warmest quarter of the year, followed by mean temperature during the wettest quarter, and
327 precipitation during the wettest quarter and month. Some SNPs were associated with both climate
328 variables and root phenotypes, with the caveat that these traits can be correlated, for example, root
329 growth-related traits with precipitation-related variables and root gravitropism-related traits with
330 temperature-related variables. The non-independence of traits would have made our multiple testing
331 correction procedures even more stringent. SNPs associated with root variables alone and/or with
332 climate variables were first observed in older herbarium samples when compared with random SNPs
333 segregating at similar allele frequencies (Fig. 5A). This suggested an older origin for variants associated
334 with relevant phenotypes, which could point to positive selection having maintained them for over a
335 century.

336 Three SNPs in AT5G19330, AT1G54440 and AT2G16580 appeared particularly interesting (Fig.
337 S6 D-F). AT5G19330 overexpression increases salt stress tolerance (Kim et al., 2004). As proof of
338 concept and alternative corroboration of association analyses, we looked for very closely related
339 accessions (<<10 SNPs in other coding regions) that differed at AT5G19330. There were 20 such pairs
340 and they differed more in their gravitropic score phenotype than random pairs and almost-identical pairs
341 (Fig. S6, see Extended Experimental Procedures). AT1G54440, also associated with gravitropism,
342 encodes an epigenetic regulator, an RRP6-like protein (Zhang et al., 2014), while AT2G16580, associated
343 with root growth rate, encodes a member of the auxin-related SAUR family (Markakis et al., 2013;
344 Spartz et al., 2012). Together, these analyses suggest that root development is an ecologically relevant
345 trait in colonization of North America by HPG1, perhaps with a role in adaptation to climate-related
346 factors such as drought.

347 The SNP in AT5G19330 was not in perfect linkage disequilibrium with other significant SNPs
348 (r^2<0.6), but some of the other candidates were strongly linked (Fig. S6 E-F). Although linkage could
349 have a biological cause, e.g., simultaneous natural selection over different loci, we must point out that
350 estimated SNP effects may suffer from statistical confounding. Hence, the associated phenotypic effects
351 could correspond to one or several groups of linked mutations across chromosomes, maybe even
352 undetected causal variants, that arose simultaneously in the history of HPG1 population (Fig. S6 B-C).
353 Additional genetic analyses such as artificial crosses will help to disentangle the effects of individual SNPs.

354 **Population demography and migrations**

355 The substitution rate estimate immediately allows dating of the HPG1 colonization of North America.
356 We first inferred the root of the HPG1 phylogenetic tree using Bayesian methods. The mean estimate
357 was the year 1597 (Highest Posterior Probability Density 95%: 1519-1660) (Fig. 3A, B). We also used a
358 non-phylogenetic method that utilizes the relationship among the genetic distance of two individuals,
359 their average divergence time, and the mutation rate. The average divergence d between sequences can
360 be approximated by the mutation rate μ multiplied by twice the divergence time L, since mutations
361 accumulate on both branches of diverging sequences:

362 $$d = 2L \times \mu$$

363 We used our previously estimated substitution rate and the average pairwise genetic distance to
364 calculate a divergence time of 363 years. Subtracting this age from the average collecting date of our
365 samples gave a point estimate of 1615, very close to the Bayesian estimate of 1597. Both are in
366 agreement with a colonization in the post-Columbian era. We believe the substitution rate in the wild
367 reported here is more appropriate when dating evolutionary events in *Arabidopsis thaliana* that using the
368 higher greenhouse mutation rate, from which we had previously inferred a more recent colonization of
369 N. America by HPG1 (Hagmann et al., 2015).

370 Knowing both the mutation rate μ and average pairwise differences π, we can obtain an
371 approximate estimate of the effective population size (N_e), by solving the equation $\pi < 4N_e\mu < \theta_W$, from
372 which we can place N_e somewhere between 152 - 758. A single N_e value represents the harmonic mean
373 of N_e over time, and thus is much closer to the historic N_e minimum than to the arithmetic average over
374 time (Wright, 1940). That N_e is so small is consistent with the recent HPG1 founder bottleneck.
375 Pairwise genetic distances between samples within the same decade, an approximate measure of
376 diversity, increased over time (Fig. S5), which supports a trend of historic population growth. More
377 sophisticated inference of N_e through time came from our dated phylogeny and its coalescent model
378 (Fig. 3B). However, our model had no resolution at the root of the tree, where population size could be

379 N_e=1, since HPG1 may have been founded by a single individual, or a few almost identical individuals.
380 Until the early 19th century, the model suggested exponential population growth, followed by slight
381 shrinkage (Fig 3B). The shrinkage in population during the last century is reflected in time-calibrated
382 phylogenies (Fig. 3 A, B), which showed that modern samples descended from a very limited number of
383 historic sublineages, with only four 20th-century herbarium samples being closely related to modern
384 samples. Altogether, population size fluctuations and the disjoint distribution of *A. thaliana* today (Platt et
385 al., 2010) suggest that the N. American population passed through recurrent bottlenecks since the initial
386 colonization.

387 Since we knew both the collection years and origins of the HPG1 samples, we could also analyze
388 the migration dynamics of HPG1. The phylogeographic models suggested that HPG1 dispersed over
389 much of its modern range already soon after its introduction to N. America (Fig. S5 A, B). Based on the
390 collection dates and sites of the herbarium samples, we postulate that the oldest populations were
391 established in the Northeast, from where they migrated west in discrete long-distance dispersions, likely
392 helped by humans. Corroborating this hypothesis, we found a significant correlation between collection
393 date and either latitude (linear regression coefficient r = 0.32; p = 3.5 x 10^{-10}) or longitude (r = 0.20; p =
394 3.7 x 10^{-6}) (Fig. 4A), which we interpret as a net, yet highly dispersed, movement in a Northwestern
395 direction over time. Additional support comes from an isolation-by-distance signal, which is most
396 consistent with a historic westward dispersion and a more recent reverse eastward migration (Fig. 4 B,
397 C; see Extended Experimental Procedures). The Lake Michigan area, where major populations are found
398 today, was both the apparent source of new migrants and the region where most derived alleles of SNPs
399 associated with root and climate traits first appeared (Fig. 5B). The coincidence between these patterns
400 of HPG1 diversity and land use change for agricultural purposes in the last two centuries (Goldewijk and
401 Ramankutty, 2004) is striking, although historical sampling biases are unknown. We hypothesize that
402 agricultural changes could have driven the initial establishment of HPG1 in N. America, since most
403 current *A. thaliana* habitats are used agriculturally or are cultivated by humans in other ways.

404 CONCLUSIONS

405 We have exploited whole-genome information from historic and contemporary collections to
406 understand fine-scale genome evolutionary dynamics in the context of a recent colonization by
407 *Arabidopsis thaliana*. By deriving a rigorously supported estimate for the mutation rate in the wild, we
408 have answered the long-standing question of how rapidly diversity is generated in natural plant
409 populations. We have presented evidence that purifying selection explains the discrepancy between
410 short- and long-term mutation rate estimates. Finally, even though rapidly expanding populations such as

411 the one studied here are severely affected by drift, limited in diversity, and likely constrained by purifying
412 selection, we found *de novo* mutations with apparent phenotypic effects that could have been subject to
413 Darwinian, adaptive selection. Recent invasion and colonization events such as the *A. thaliana* HPG1
414 example are natural experiments ideally suited for analyzing adaptation to new environments. Finally,
415 our work should encourage others to unlock the potential of herbarium specimens for the study of
416 evolution in action.

417 EXPERIMENTAL PROCEDURES

418 Additional details are given in the Extended Experimental Procedures in Supplemental information.

419 Sample collection and DNA sequencing

420 Modern *A. thaliana* accessions were from the collection described by Platt and colleagues (2010); HPG1
421 candidates were identified based on 149 genome-wide SNPs (Table S1). Herbarium specimens
422 (collection dates 1863-1993) were directly sampled by our colleagues Jane Devos and Gautam Shirsekar,
423 or sent to us by collection curators from various herbaria (Table S1). DNA from herbarium specimens
424 was extracted as described (Yoshida et al., 2013) in a clean room facility at the University of Tübingen,.
425 Two sequencing libraries with sample-specific barcodes were prepared following established protocols,
426 with and without repair of deaminated sites using uracil-DNA glycosylase and endonuclease VIII (Briggs
427 et al., 2010; Kircher, 2012; Meyer and Kircher, 2010). DNA from modern individuals was extracted
428 from pools of eight siblings of each inbred line. Genomic DNA libraries were prepared using the TruSeq
429 DNA Sample prep kit or TruSeq Nano DNA sample prep kit (Illumina, San Diego, CA), and sequenced
430 on Illumina HiSeq and MiSeq instruments. Reads were mapped with GenomeMapper v0.4.5s
431 (Schneeberger et al., 2009) against an HPG1 pseudo-reference genome (Hagmann et al., 2015), and
432 against the Col-0 reference genome. Samples JK2509 to JK2531 were only mapped to the HPG1
433 pseudo-reference genome. Coverage, number of covered positions in the genome, and number of SNPs
434 identified per accession relative to HPG1 are reported in Table S1. We also re-sequenced the genomes
435 of twelve mutation accumulation (MA) lines (Becker et al., 2011; Shaw et al., 2000) (Table S2).

436 Phylogenetic methods and genome-wide statistics

437 We used four methods to estimate the relationships among modern accessions, and between modern
438 and herbarium samples: (i) multidimensional scaling (MDS) analysis; (ii) construction of a neighbor joining
439 tree with the adegenet package in R (Jombart, 2008), with branch support assessed with 1,000 bootstrap
440 iterations; (iii) construction of a parsimony network using SplitsTree v.4.12.3 (Huson and Bryant, 2006),

441 with confidence values calculated with 1,000 bootstrap iterations; (iv) performing a Bayesian
442 phylogenetic analysis using BEAST v.1.8 (Bouckaert et al., 2014; Drummond et al., 2012) (see below).

443 We estimated genetic diversity as Watterson´s θ (Watterson, 1975) and nucleotide diversity π,
444 and the difference between these two statistics as Tajimas's *D* (Tajima, 1989) using DnaSP v5 (Librado
445 and Rozas, 2009). We calculate the folded site frequency spectrum (SFS) as well as the unfolded SFS, for
446 which we assigned the ancestral state using the *Arabidopis lyrata* genome (Hu et al., 2011). We estimated
447 pairwise linkage disequilibrium (LD) between all possible combinations of informative sites, ignoring
448 singletons, by computing r^2, *D* and *D*' statistics. For the modern individuals, we calculated the
449 recombination parameter rho (*4Ner*) and performed the four-gamete-test (Hudson and Kaplan, 1985) to
450 identify the minimum number of recombination events. All LD and recombination related statistics were
451 determined using DnaSP v5 (Librado and Rozas, 2009).

452 **Substitution and mutation rate analyses**

453 We used genome-wide nuclear SNPs to calculate pairwise "net" genetic distances using the equation D'_{ij}
454 $= D_{ic}-D_{jc}$, where D'_{ij} is the net distance between a modern sample *i* and a herbarium sample *j*; D_{ic} the
455 distance between the modern sample *i* and the reference genome *c*; and D_{jc} is the distance between a
456 modern sample (j) and the reference genome (c). We calculated a pair-wise time distance in years, *T*'ij,
457 using the collection dates and linear regression: $D' = a+bT'$. The slope coefficient *b* describes the number
458 of substitution changes per year. We used either all SNPs or subsets of SNPs at different annotations
459 appropriately scaled by accessible genome length.

460 The second approach used Bayesian phylogenetics with the tip-calibration method implemented
461 in BEAST v1.8 software (Drummond et al., 2012). Our analysis optimized simultaneously and in an
462 iterative fashion using a Monte Carlo Markov Chain (MCMC) a tree topology, branch length,
463 substitution rate, and a demographic Skygrid model. The demographic model is a Bayesian
464 nonparametric one that is optimized for multiple loci and that allows for complex demographic
465 trajectories by estimating population sizes in time bins across the tree based on the number of
466 coalescent events per bin (Gill et al., 2012). We also performed a second analysis run using a fixed prior
467 for substitution rate of 3.3×10^{-9} substitutions site^{-1} year^{-1} that we had estimated empirically using the
468 net-distance method to confirm that the MCMC had the same parameter convergence, e.g. tree
469 topology, as the first "estimate-all-parameters" run.

470 **Inference of genome-wide selection parameters**

471 We separately analyzed sequences at different annotations, since some regions should be under a
472 different selection regime (less evolutionary constraint) than others. We estimated the average strength
473 of genome-wide selection by contrasting substitution rates in the entire genome and in intergenic
474 regions. We use the latter as a near-neutral contrast because it provides more statistical power in our
475 sample with limited diversity, than the more usual contrast between synonymous (or fourfold
476 degenerate) and non-synonymous sites. Selection was estimated based on the equation $k = \mu \times Q \times 2N_e$,
477 where Q is the fixation probability of a new mutation (Barrick and Lenski, 2013; Kimura, 1967), and the
478 equation $Q \approx s / 2N_e (1-e^{-2N_e s})$ (Charlesworth and Charlesworth, 2010).

479 **Association analyses and dating of new mutations**

480 We collected flowering, seed and root morphology phenotypes for 63 modern accessions. For
481 associations with climate parameters, we followed a similar rationale as described (Hancock et al., 2011).
482 We extracted information from the publicly available bioclim database
483 (http://www.worldclim.org/bioclim) at 2.5 degrees resolution raster and intersected it with geographic
484 locations of HPG1 samples (n = 100). We performed association analyses under several models and *p*-
485 value corrections using the R package GeneABEL (Aulchenko et al., 2007), with phenotypes and climatic
486 variables as response variables and SNPs as explanatory variables and appropriate correcting covariates.
487 Significance estimates were corrected with 1,000 permuted datasets, or with Bonferroni correction.

488 **Accession numbers**

489 Short reads have been deposited in the European Nucleotide Archive under the accession number TO
490 BE UPDATED UPON ACCEPTANCE.

491 **SUPPLEMENTAL INFORMATION**

492 Supplemental Information includes Extended Experimental Procedures, six supplemental figures and six
493 tables, and can be found online at TO BE UPDATED UPON ACCEPTANCE.

494 **ACKNOWLEDGEMENTS**

495 For providing and retrieving herbarium specimens, we thank Robert Capers (University of Connecticut),
496 Jane Devos and Gautam Shirsekar (MPI), Michael S. Dossmann (Arnold Arboretum), John Freudenstein
497 (Ohio State University), Cathy M. Herring (Agricultural Research Station, North Carolina State

University), Christine Niezgoda (Field Museum), Carol Ann McCormick (University of North Carolina), John Peter (New York Botanical Garden), and Marco Thines (Goethe University). We thank Xiaohui Zhao and Ian Henderson (SLCU) for recombination estimates in Eurasia, Christa Lanz (MPI) for support with sequencing, Christian Goeschl and Bettina Zierfuss (GMI) for assistance in root imaging, and Bonnie Wohlrab (GMI) for amplifying the seeds for root assays. We thank Magnus Nordborg for discussions and pointing us to the work of Templeton, Kay Pruefer for input on data analysis, Patricia Karlsson and Danelle Seymour for thorough proofreading and comments, and various present and past members of the Department of Molecular Biology of the Max Planck Institute for Developmental Biology for further comments on the manuscript. This work was supported by ERC Advanced Grant IMMUNEMESIS and the Max Planck Society.

AUTHOR CONTRIBUTIONS

H.A.B and D.W. conceived and supervised the project, and coordinated the collaborative effort. J.B. coordinated the collection of modern seed samples. C.J. B.B. and J.B. performed and analyzed flowering time and seed set greenhouse experiments. R.S. and W.B. conceived and analyzed root assays. C.S. and R.S. performed the root assays and seed size phenotyping. C.B. and J.H. sequenced and curated modern samples. H.A.B. coordinated the collection and analysis of herbarium samples. J.K. coordinated the extraction of DNA and library preparation of herbarium samples. V.J.S. and E.R. prepared sequencing libraries from herbarium specimens. C.B. called variants in HPG1. J.H. called variants in mutation accumulation lines. M.E.A. performed the population and quantitative genomic analyses with supervision of R.N., C.B. and H.A.B. The paper was written by M.E.A., C.B., H.A.B. and D.W. with comments from all coauthors.

REFERENCES

Arunkumar, R., Ness, R.W., Wright, S.I., and Barrett, S.C. (2015). The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. Genetics *199*, 817-829.

Aulchenko, Y.S., Ripke, S., Isaacs, A., and van Duijn, C.M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics *23*, 1294-1296.

Baker, H.G. (1965). Characteristics and modes of origin of weeds. In The Genetics of Colonizing Species, H.G. Baker, and G.L. Stebbins, eds. (New York: Academic Press), pp. 147-168.

Barrett, R.D.H., and Schluter, D. (2008). Adaptation from standing genetic variation. Trends Ecol Evol *23*, 38-44.

529 Barrett, S.C.H. (2014). Foundations of invasion genetics: the Baker and Stebbins legacy. Mol Ecol *24*,
530 1927-1941.
531 Barrick, J.E., and Lenski, R.E. (2013). Genome dynamics during experimental evolution. Nat Rev Genet
532 *14*, 827-839.
533 Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D. (2011).
534 Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature *480*, 245-249.
535 Bentsink, L., Hanson, J., Hanhart, C.J., Blankestijn-de Vries, H., Coltrane, C., Keizer, P., El-Lithy, M.,
536 Alonso-Blanco, C., de Andres, M.T., Reymond, M., *et al.* (2010). Natural variation for seed dormancy
537 in *Arabidopsis* is regulated by additive genetic and molecular pathways. Proc Natl Acad Sci USA *107*,
538 4264-4269.
539 Bock, D.G., Caseys, C., Cousens, R.D., Hahn, M.A., Heredia, S.M., Hübner, S., Turner, K.G., Whitney,
540 K.D., and Rieseberg, L.H. (2015). What we still don't know about invasion genetics. Mol Ecol *24*,
541 2277-2297.
542 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.a., Rambaut, A., and
543 Drummond, A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS
544 Comp Biol *10*, e1003537.
545 Briggs, A.W., Stenzel, U., Johnson, P.L., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan,
546 M.T., Lachmann, M., *et al.* (2007). Patterns of damage in genomic DNA sequences from a
547 Neandertal. Proc Natl Acad Sci USA *104*, 14616-14621.
548 Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S. (2010). Removal of deaminated
549 cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res *38*, e87.
550 Brouwer, D.J., and St Clair, D.A. (2004). Fine mapping of three quantitative trait loci for late blight
551 resistance in tomato using near isogenic lines (NILs) and sub-NILs. Theor Appl Genet *108*, 628-638.
552 Bustamante, C.D., Nielsen, R., Sawyer, S.A., Olsen, K.M., Purugganan, M.D., and Hartl, D.L. (2002). The
553 cost of inbreeding in Arabidopsis. Nature *416*, 531-534.
554 Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O.,
555 Lippert, C., *et al.* (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat
556 Genet *43*, 956-963.
557 Charlesworth, B., and Charlesworth, D. (2010). Elements of Evolutionary Genetics (Roberts and
558 Company: Greenwood Village, CO, 2010).
559 Charlesworth, D., and Wright, S.I. (2001). Breeding systems and genome evolution. Curr Opin Genet
560 Dev *11*, 685-690.

561 Choi, K., Zhao, X., Kelly, K.A., Venn, O., Higgins, J.D., Yelina, N.E., Hardcastle, T.J., Ziolkowski, P.A.,
562 Copenhaver, G.P., Franklin, F.C., *et al.* (2013). *Arabidopsis* meiotic crossover hot spots overlap with
563 H2A.Z nucleosomes at gene promoters. Nat Genet *45*, 1327-1336.
564 Crawford, P.H.C., and Hoagland, B.W. (2009). Can herbarium records be used to map alien species
565 invasion and native species expansion over the past 100 years? J Biogeography *36*, 651-661.
566 Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti
567 and the BEAST 1.7. Mol Biol Evol *29*, 1969-1973.
568 Fletcher, R.S., Mullen, J.L., Yoder, S., Bauerle, W.L., Reuning, G., Sen, S., Meyer, E., Juenger, T.E., and
569 McKay, J.K. (2013). Development of a next-generation NIL library in Arabidopsis thaliana for
570 dissecting complex traits. BMC Genomics *14*, 655.
571 Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L.F., Aximu-Petri, A.,
572 Prüfer, K., de Filippo, C., *et al.* (2014). Genome sequence of a 45,000-year-old modern human from
573 western Siberia. Nature *514*, 445-449.
574 Gauze, G.F. (1934). The Struggle for Existence (Baltimore: Williams & Wilkins).
575 Gill, M.S., Lemey, P., Faria, N.R., Rambaut, A., Shapiro, B., and Suchard, M.a. (2012). Improving Bayesian
576 population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol *30*, 713-
577 724.
578 Goldewijk, K.K., and Ramankutty, N. (2004). Land cover change over the last three centuries due to
579 human activities: The availability of new global data sets. GeoJournal *61*, 335-334.
580 Green, R.E., and Shapiro, B. (2013). Human evolution: turning back the clock. Curr Biol *23*, R286-288.
581 Hagmann, J., Becker, C., Müller, J., Stegle, O., Meyer, R.C., Wang, G., Schneeberger, K., Fitz, J., Altmann,
582 T., Bergelson, J., *et al.* (2015). Century-scale methylome stability in a recently diverged *Arabidopsis*
583 *thaliana* lineage. PLoS Genet *11*, e1004920.
584 Halligan, D.L., and Keightley, P.D. (2009). Spontaneous mutation accumulation studies in evolutionary
585 genetics. Annu Rev Ecol Evol S *40*, 151-172.
586 Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C.,
587 Roux, F., and Bergelson, J. (2011). Adaptation to climate across the *Arabidopsis thaliana* genome.
588 Science *334*, 83-86.
589 Ho, S.Y.W., Phillips, M.J., Cooper, A., and Drummond, A.J. (2005). Time dependency of molecular rate
590 estimates and systematic overestimation of recent divergence times. Mol Biol Evol *22*, 1561-1568.
591 Hofreiter, M., Jaenicke, V., Serre, D., Haeseler Av, A., and Pääbo, S. (2001). DNA sequences from
592 multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic
593 Acids Res *29*, 4793-4799.

594 Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood,
595 J., Gundlach, H., *et al.* (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome
596 size change. Nat Genet *43*, 476-481.
597 Hudson, R.R., and Kaplan, N.L. (1985). Statistical properties of the number of recombination events in
598 the history of a sample of DNA sequences. Genetics *111*, 147-164.
599 Huson, D.H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol
600 Biol Evol *23*, 254-267.
601 Jiang, C., Mithani, A., Belfield, E.J., Mott, R., Hurst, L.D., and Harberd, N.P. (2014). Environmentally
602 responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations.
603 Genome Res *24*, 1821-1829.
604 Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics
605 *24*, 1403-1405.
606 Keightley, P.D., and Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of
607 deleterious mutations and population demography based on nucleotide polymorphism frequencies.
608 Genetics *177*, 2251-2261.
609 Keurentjes, J.J., Bentsink, L., Alonso-Blanco, C., Hanhart, C.J., Blankestijn-De Vries, H., Effgen, S.,
610 Vreugdenhil, D., and Koornneef, M. (2007). Development of a near-isogenic line population of
611 *Arabidopsis thaliana* and comparison of mapping power with a recombinant inbred line population.
612 Genetics *175*, 891-905.
613 Kim, S., Choi, H.I., Ryu, H.J., Park, J.H., Kim, M.D., and Kim, S.Y. (2004). ARIA, an Arabidopsis arm
614 repeat protein interacting with a transcriptional regulator of abscisic acid-responsive gene
615 expression, is a novel abscisic acid signaling component. Plant Physiol *136*, 3639-3648.
616 Kimura, M. (1967). On the evolutionary adjustment of spontaneous mutation rates. Genet Res *9*, 23.
617 Kircher, M. (2012). Analysis of high-throughput ancient DNA sequencing data. Methods Mol Biol *840*,
618 197-228.
619 Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.a.,
620 Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., *et al.* (2012). Rate of de novo mutations and the
621 importance of father's age to disease risk. Nature *488*, 471-475.
622 Krause, J., Briggs, A.W., Kircher, M., Maricic, T., Zwyns, N., Derevianko, A., and Pääbo, S. (2010). A
623 complete mtDNA genome of an early modern human from Kostenki, Russia. Curr Biol *20*, 231-236.
624 Lankaua, R.A., Nuzzo, V., Spyreasa, G., and Davisc, A.S. (2009). Evolutionary limits ameliorate the
625 negative impact of an invasive plant. Proc Natl Acad Sci USA *107*, 1253.
626 Lee, C.E. (2002). Evolutionary genetics of invasive species. Trends Ecol Evol *17*, 386-391.

627 Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA
628 polymorphism data. Bioinformatics *25*, 1451-1452.
629 Lipson, M., Loh, P.R., Sankararaman, S., Patterson, N., Berger, B., and Reich, D. (2015). Calibrating the
630 Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes. PLoS Genet *11*,
631 e1005550.
632 Lundemo, S., Falahati-Anbaran, M., and Stenøien, H.K. (2009). Seed banks cause elevated generation
633 times and effective population sizes of *Arabidopsis thaliana* in northern Europe. Mol Ecol *18*, 2798-
634 2811.
635 Markakis, M.N., Boron, A.K., Van Loock, B., Saini, K., Cirera, S., Verbelen, J.P., and Vissenberg, K. (2013).
636 Characterization of a small auxin-up RNA (SAUR)-like gene involved in Arabidopsis thaliana
637 development. PLoS One *8*, e82596.
638 Maron, J.L., Vilà, M., Bommarco, R., Elmendorf, S., and Beardsley, P. (2004). Rapid evolution of an
639 invasive plant. Ecol Monogr *74*, 261-280.
640 Martin, M.D., Cappellini, E., Samaniego, J.A., Zepeda, M.L., Campos, P.F., Seguin-Orlando, A., Wales, N.,
641 Orlando, L., Ho, S.Y., Dietrich, F.S.*, et al.* (2013). Reconstructing genome evolution in historic
642 samples of the Irish potato famine pathogen. Nat Commun *4*, 2172.
643 Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target
644 capture and sequencing. Cold Spring Harb Protoc *2010*, pdb prot5448.
645 Montesinos, A., Tonsor, S.J., Alonso-Blanco, C., and Picó, F.X. (2009). Demographic and genetic patterns
646 of variation among populations of *Arabidopsis thaliana* from contrasting native environments. PLoS
647 One *4*, e7213.
648 Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans.
649 Genetics *156*, 297-304.
650 Ness, R.W., Morgan, A.D., Vasanthakrishnan, R.B., Colegrave, N., and Keightley, P.D. (2015). Extensive
651 de novo mutation rate variation between individuals and across the genome of *Chlamydomonas*
652 *reinhardtii*. Genome Res.
653 Ness, R.W., Wright, S.I., and Barrett, S.C. (2010). Mating-system variation, demographic history and
654 patterns of nucleotide diversity in the Tristylous plant Eichhornia paniculata. Genetics *184*, 381-392.
655 Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P.,
656 Gladstone, J., Goyal, R.*, et al.* (2005). The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol
657 *3*, e196.
658 Orlando, L., Gilbert, M.T., and Willerslev, E. (2015). Reconstructing ancient genomes and epigenomes.
659 Nat Rev Genet *16*, 395-408.

660 Ossowski, S., Schneeberger, K., Lucas-Lledo, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D.,
661 and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis*
662 *thaliana*. Science *327*, 92-94.
663 Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Agren, J., Bossdorf, O., Byers, D.,
664 Donohue, K., *et al.* (2010). The scale of population structure in *Arabidopsis thaliana*. PLoS Genet *6*,
665 e1000843.
666 Prüfer, K., and Meyer, M. (2015). Comment on "Late Pleistocene human skeleton and mtDNA link
667 Paleoamericans and modern Native Americans". Science *347*, 835.
668 Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P.,
669 Goodman, N., Bamshad, M., *et al.* (2010). Analysis of genetic inheritance in a family quartet by
670 whole-genome sequencing. Science *328*, 636-639.
671 Sax, D.F., Stachowicz, J.J., Brown, J.H., Bruno, J.F., Dawson, M.N., Gaines, S.D., Grosberg, R.K., Hastings,
672 A., Holt, R.D., Mayfield, M.M., *et al.* (2007). Ecological and evolutionary insights from species
673 invasions. Trends Ecol Evol *22*, 465-471.
674 Scally, A., and Durbin, R. (2012). Revising the human mutation rate: implications for understanding
675 human evolution. Nat Rev Genet *13*, 745-753.
676 Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., and Weigel, D.
677 (2009). Simultaneous alignment of short reads against multiple genomes. Genome Biol *10*, R98.
678 Ségurel, L., Wyman, M.J., and Przeworski, M. (2014). Determinants of mutation rate variation in the
679 human germline. Annu Rev Genomics Hum Genet *15*, 47-70.
680 Shapiro, B., and Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: new
681 insights from ancient DNA. Science *343*, 1236573.
682 Shaw, R.G., Byers, D.L., and Darmo, E. (2000). Spontaneous mutational effects on reproductive traits of
683 *Arabidopsis thaliana*. Genetics *155*, 369-378.
684 Sniegowski, P.D., Gerrish, P.J., and Lenski, R.E. (1997). Evolution of high mutation rates in experimental
685 populations of E. coli. Nature *387*, 703-705.
686 Spartz, A.K., Lee, S.H., Wenger, J.P., Gonzalez, N., Itoh, H., Inze, D., Peer, W.A., Murphy, A.S.,
687 Overvoorde, P.J., and Gray, W.M. (2012). The SAUR19 subfamily of SMALL AUXIN UP RNA genes
688 promote cell expansion. Plant J *70*, 978-990.
689 Staats, M., Erkens, R.H.J., van de Vossenberg, B., Wieringa, J.J., Kraaijeveld, K., Stielow, B., Geml, J.,
690 Richardson, J.E., and Bakker, F.T. (2013). Genomic treasure troves: complete genome sequencing of
691 herbarium and insect museum specimens. PLoS ONE *8*, e69189.

692 Stec, A.O., Bhaskar, P.B., Bolon, Y.T., Nolan, R., Shoemaker, R.C., Vance, C.P., and Stupar, R.M. (2013).
693 Genomic heterogeneity and structural variation in soybean near isogenic lines. Front Plant Sci *4*,
694 104.
695 Subramanian, S., and Kumar, S. (2003). Neutral substitutions occur at a faster rate in exons than in
696 noncoding DNA in primate genomes. Genome Res *13*, 838-844.
697 Subramanian, S., and Lambert, D.M. (2011). Time dependency of molecular evolutionary rates? Yes and
698 no. Genome Biology and Evolution *3*, 1324-1328.
699 Swarup, K., Alonso-Blanco, C., Lynn, J.R., Michaels, S.D., Amasino, R.M., Koornneef, M., and Millar, A.J.
700 (1999). Natural allelic variation identifies new genes in the *Arabidopsis* circadian system. Plant J *20*,
701 67-77.
702 Szalma, S.J., Hostert, B.M., Ledeaux, J.R., Stuber, C.W., and Holland, J.B. (2007). QTL mapping with near-
703 isogenic lines in maize. Theor Appl Genet *114*, 1211-1228.
704 Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.
705 Genetics *123*, 585-595.
706 Templeton, A.R., Sing, C.F., Kessling, A., and Humphries, S. (1988). A Cladistic-Analysis of Phenotype
707 Associations with Haplotypes Inferred from Restriction Endonuclease Mapping .2. The Analysis of
708 Natural-Populations. Genetics *120*, 1145-1154.
709 van Kleunen, M., Dawson, W., Essl, F., Pergl, J., Winter, M., Weber, E., Kreft, H., Weigelt, P., Kartesz, J.,
710 Nishino, M., *et al.* (2015). Global exchange and accumulation of non-native plants. Nature *525*, 100-
711 103.
712 Vandepitte, K., de Meyer, T., Helsen, K., van Acker, K., Roldán-Ruiz, I., Mergeay, J., and Honnay, O.
713 (2014). Rapid genetic adaptation precedes the spread of an exotic plant species. Mol Ecol *23*, 2157-
714 2164.
715 Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination.
716 Theor Pop Biol *7*, 256-276.
717 Weigel, D. (2012). Natural variation in Arabidopsis: from molecular genetics to ecological genomics.
718 Plant Physiol *158*, 2-22.
719 Weiss, C.L., Dannemann, M., Prufer, K., and Burbano, H.A. (2015). Contesting the presence of wheat in
720 the British Isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data. eLife
721 *4*.
722 Weiß, C.L., Schuenemann, V.J., Devos, J., Shirsekar, G., Reiter, E., Gould, B.A., Stinchcombe, J.R., Krause,
723 J., and Burbano, H.A. (2015). Temporal patterns of damage and decay kinetics of DNA retrieved
724 from plant herbarium specimens. bioRxiv http://dx.doi.org/10.1101/023135.

725 Wright, S.I., Ness, R.W., Foxe, J.P., and Barret, S.C.H. (2008). Genomic consequences of outcrossing and
726 selfing in plants. Int J Plant Sci *169*, 105-118.
727 Xie, X., Song, M.H., Jin, F., Ahn, S.N., Suh, J.P., Hwang, H.G., and McCouch, S.R. (2006). Fine mapping of
728 a grain weight quantitative trait locus on rice chromosome 8 using near-isogenic lines derived from a
729 cross between Oryza sativa and Oryza rufipogon. Theor Appl Genet *113*, 885-894.
730 Yoshida, K., Schuenemann, V.J., Cano, L.M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin, F.N.,
731 Kamoun, S., Krause, J., *et al.* (2013). The rise and fall of the *Phytophthora infestans* lineage that
732 triggered the Irish potato famine. eLife *2*, e00731.
733 Zhang, H., Tang, K., Qian, W., Duan, C.G., Wang, B., Zhang, H., Wang, P., Zhu, X., Lang, Z., Yang, Y., *et*
734 *al.* (2014). An Rrp6-like protein positively regulates noncoding RNA levels and DNA methylation in
735 Arabidopsis. Mol Cell *54*, 418-430.

736 **TABLE**

737 **Table 1. Genic SNPs associated with different traits.**

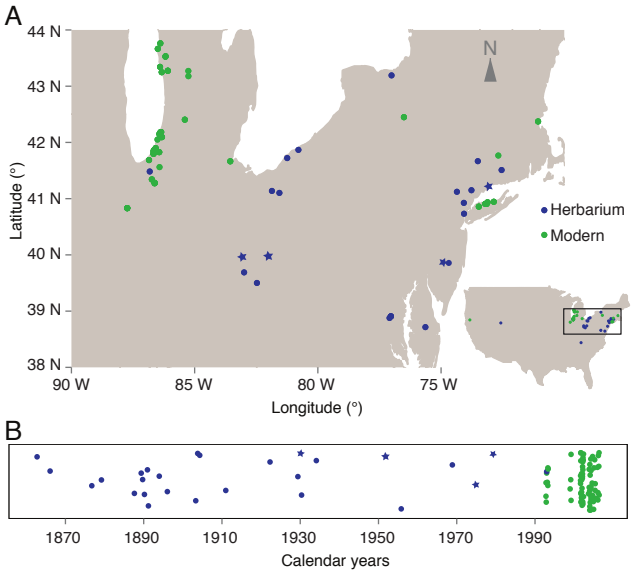
738 Most SNPs first appeared in sample JK2530 collected 1922 in Indiana. For non-synonymous SNPs, the amino acid
739 transition and the Grantham score (ranging from 0 to 215) are reported. All SNPs in the table were significant ($p < 0.05$) after raw p-values were permutation corrected. # highlights those whose permutation corrected p-values
741 were still significant when the threshold was corrected by multiple traits ($p<0.002$). * indicates SNPs when raw p-
742 values passed the threshold corrected by multiple SNP correction as well as multiple trait correction ($p<0.0001$).
743 See Table S4 for details on phenotypes and climatic variables, and Table S5 for information on all significant SNPs.

Trait†	Location (chr-bp)	Gene	Anno–tation	Protein	aa change	Multiple testing
G	1-958,948	AT1G03810	nonsyn	Oligonucleotide/ oligosaccharide binding fold	A>P, 27	
D	1-13,994,958	AT1G36933	transposon	Copia		
S	1-20,324,050	AT1G54440	intronic	RRP6-LIKE 1		#*
D	1-23,648,407	AT1G63740	nonsyn	TIR-NLR family	Y>S, 144	
G	2-358,395	AT2G01820	syn	RLK family		*
G	2-585,918	AT2G02220	syn	PSKR1		*
G	2-6,034,545	AT2G14247	syn	Expressed protein		*
G	2-7,047,529	AT2G16270	nonsyn	Unknown protein	P>A, 27	*
G	2-7,186,220	AT2G16580	intronic	SAUR8		*
G	2-10,495,275	AT2G24680	intronic	B3 family		*
G	2-12,415,084	AT2G28900	intronic	OEP16		
S	2-16,039,488	AT2G38290	3' UTR	AMT2		#*
S	2-16,247,290	AT2G38910	nonsyn	CPK20	A>G, 60	#*
G	2-16,333,662	AT2G39160	nonsyn	Unknown protein	A>G, 60	
G	3-2,500,258	AT3G07830	syn	PGA3		*
G	3-3,629,794	AT3G11530	intronic	VPS55		*
G	3-4,269,626	AT3G13229	5' UTR	DUF868 domain		*
D	3-11,873,293	AT3G30219	transposon	Gypsy		
G & D	4-4,228,138	AT4G07440	transposon	Oligonucleotide/ oligosaccharide binding fold		
G & D	4-9,046,942	AT4G15960	nonsyn	alpha/beta-hydrolase superfamily	A>Q, 24	
G & D	4-15,646,341	AT4G32410	syn	ANY1		
G	4-15,845,001	AT4G32840	3' UTR	PFK6		
D	5-4,245,213	AT5G13260	syn	Unknown protein		

D	5-4,500,202	AT5G13950	nonsyn	Unknown protein	A>G, 60	
G	5-4,797,923	AT5G14830	transposon	Retrotransposon		
G	5-6,508,329	AT5G19330	nonsyn	ARIA	C>W, 215	
G	5-11,090,365	AT5G29037	transposon	Gypsy		
G	5-12,312,975	AT5G32630	pseudogene	–		
G	5-12,358,159	AT5G32825	transposon	CACTA		
S	16024197	AT5G40020	intronic	thaumatin superfamily		#*

744 †Traits with significant associations were root gravitropism (G), root size (S), or summer precipitation,
745 related to drought conditions (D).

746 **FIGURES**



747

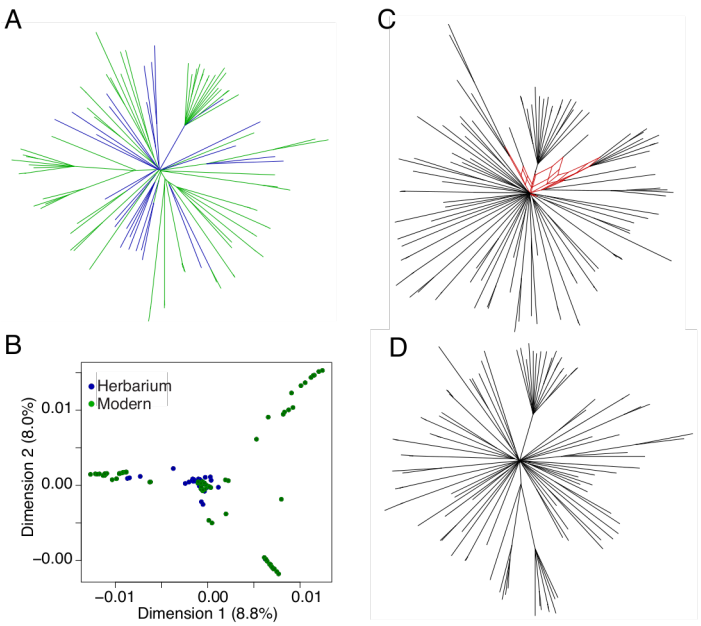
748 **Figure 1. Geographic location and temporal distribution of HPG1 samples.**

749 **(A)** Sampling location of herbarium specimens (blue) and modern individuals (green). **(B)** Temporal
750 distribution of samples (randomly jittered in a y axis for visualization). Stars indicate four herbarium
751 accessions that nest in the clade of modern accessions. See Fig. 3.

752 See also Figure S1.

753

754



755 **Figure 2. Relationship among herbarium and modern HPG1 samples.**

756 **(A)** Neighbor-joining tree. Consensus of 1,000 bootstrap replicates. Branch lengths indicate number of
757 base substitutions. **(B)** First two dimensions of a multidimensional scaling plot based on pairwise
758 identity-by-state distances. Fraction of variance explained given in parentheses. Phylogenetic network of
759 all samples using the parsimony splits algorithm, before **(C)** and after **(D)** removing intra-HPG1
760 recombinants.

761 See also Figure S2.

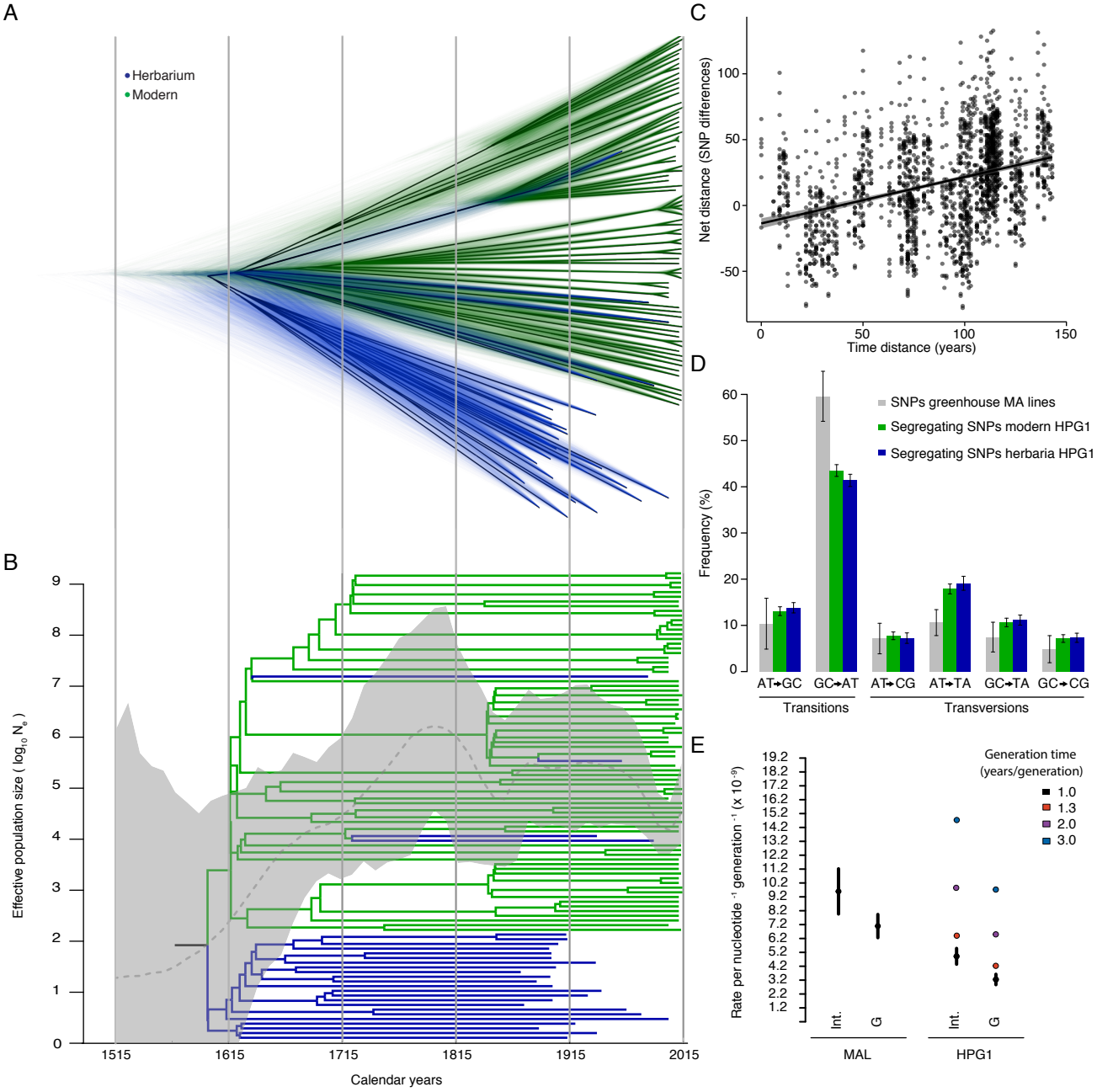


Figure 3. Substitution rates and demographic history.

(A) Bayesian phylogenetic analyses employing tip-calibration methodology. All 10,000 trees were superimposed as transparent lines, and the most common topology was plotted solid. Tree branches were calibrated with their corresponding collecting dates. **(B)** Maximum Clade Credibility (MCC) tree summarizing the trees in (A). The demographic model underlying the phylogenetic analysis is superimposed on the MCC tree. N_e was estimated by Bayesian Skygrid reconstruction; the mean N_e over time is shown as a dotted line and the 95% highest posterior density is shaded grey. **(C)**

770 Regression between pairwise net genetic and time distances. The slope of the linear regression line
771 corresponds to the whole-genome substitution rate per year. **(D)** Substitution spectra in HPG1
772 samples, compared to greenhouse-grown mutation accumulation (MA) lines. **(E)** Comparison of
773 mutation rates between greenhouse-grown MA Lines (MALs) and HPG1. 95% confidence intervals from
774 bootstrap resampling using regression approach from C are shown (see Table S3 for specific values).
775 See also Figures S3 and S5.

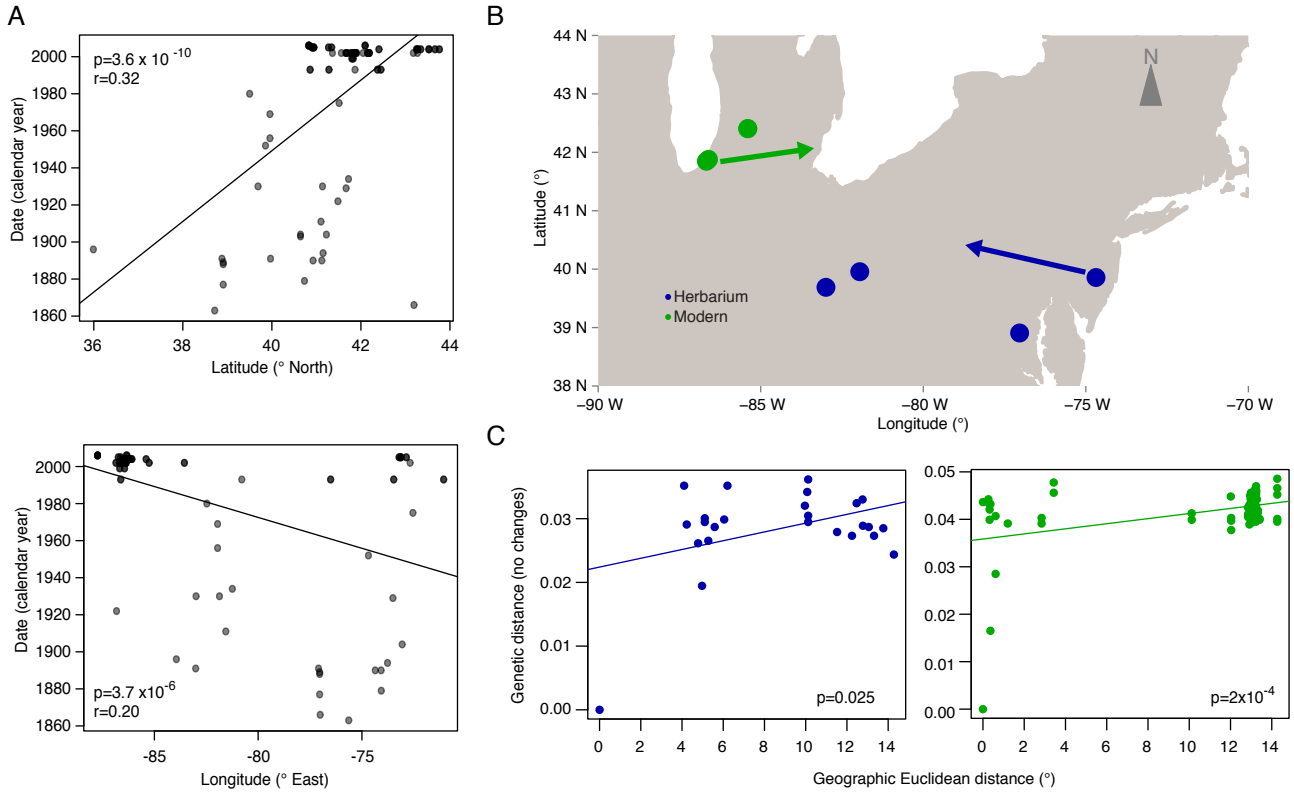


Figure 4. Migration dynamics of HPG1.

(A) Linear regression of longitude and latitude as a function of collection year. The p-value was obtained from the t-test of the slope. **(B)** Origin of herbarium and modern geographic spread, determined using separate heuristic searches of isolation-by-distance patterns. Three locations of modern samples and four of herbarium samples showed significant slope ($p<0.05$) in the isolation-by-distance pattern. That is, genetic distance increased when moving apart from those geographic locations. For one sample of each subset a likely migration trajectory is depicted by an arrow. **(C)** Isolation-by-distance patterns of the herbarium (left) and modern (right) samples from which the hypothetical trajectory in (C) was inferred.

See also Figure S5.

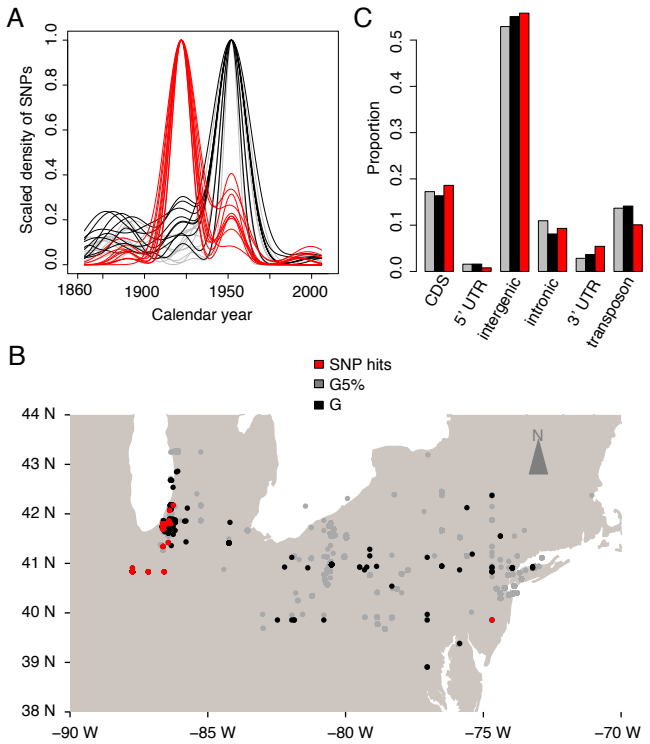


Figure 5. Spatial and temporal emergence of mutations associated with root morphology phenotypes and/or climate variables.

(A) Age distribution of the oldest herbarium sample with the derived allele of each SNP with a significant trait association, compared with genome-wide SNPs with at least 5% minor allele frequency (black), or without frequency cutoff (grey). **(B)** Spatial centroid of all samples carrying derived-allele SNPs shown in (A).

See also Figures S4 and S6.

SUPPLEMENTAL INFORMATION FOR

Exposito-Alonso, Becker et al.: **THE RATE AND EFFECT OF *DE NOVO* MUTATIONS IN NATURAL POPULATIONS OF *ARABIDOPSIS THALIANA***

Supplemental Tables

Tables S1 to S5 in file Exposito-Alonso_2016_TABLES_S1_to_S5.xlsx

Table S1. Sample information. Related to Figure 1.

Table S2. Sample information for greenhouse-grown mutation accumulation lines. Related to Figure 3.

Table S3. Mutation rate estimates for different annotations in HPG1 and greenhouse-grown mutation accumulation lines. Related to Figure 3.

Table S4. Description of phenotypic and climatic variables for association analyses. Related to Figure 5.

Table S5. SNP hits from association analyses. Related to Table 1.

Table S6. Trait distributions and QQ plots of association analyses. Related to Figure 5.
For each trait employed in association analyses, we report the histogram distribution and the QQ plot of p-values to ensure that no trait departs exaggeratedly from the normal distribution, and that no inflation of p-values is observed (when lambda <= 1, there is no inflation of false positives).

Supplemental Figures

Figure S1. Ancient-DNA-like characteristics of herbarium-derived libraries not treated with uracil glycosylase. Related to Figure 1.

Figure S2. Separation between HPG1 and other North American lineages. Related to Figure 2.

Figure S3. Substitution spectrum and relationship between methylation and substitutions. Related to Figure 3

Figure S4. Density of SNPs along all chromosomes and location of SNP hits. Related to Figure 5.

Figure S5. Bayesian phylogeographic inference using continuous trait models, and HPG1 genetic diversity in time and space. Related to Figure 4.

Figure S6. Linkage disequilibrium and SNPs with significant trait associations and correlations between SNP effects, frequency and age. Related to Figure 5.

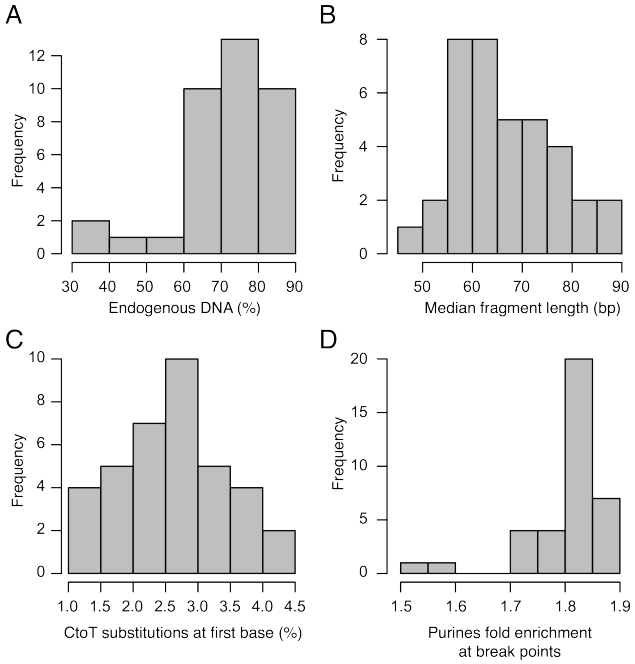
Supplemental Experimental Procedures

Supplemental References

SUPPLEMENTAL TABLES

See separate .xlsx file for Tables S1-5 and separate .pdf file for Graphic Table S6.

35 **SUPPLEMENTAL FIGURES**



36

37 **Figure S1. Ancient DNA-like characteristics of herbarium-derived libraries not**
38 **treated with uracil glycosylase.**
39 **(A)** Percentage of *Arabidopsis thaliana* endogenous DNA. **(B)** Median length of merged reads. **(C)**
40 Percentage of cytosine to thymine (C-to-T) substitutions at first base (5' end). **(D)** Relative enrichment
41 of purines (adenine and guanine) at 5' end breaking points. Position -1 is compared with position -5.
42 Numbers indicate genomic context before upstream reads' 5' end.
43 Related to Figure 1.

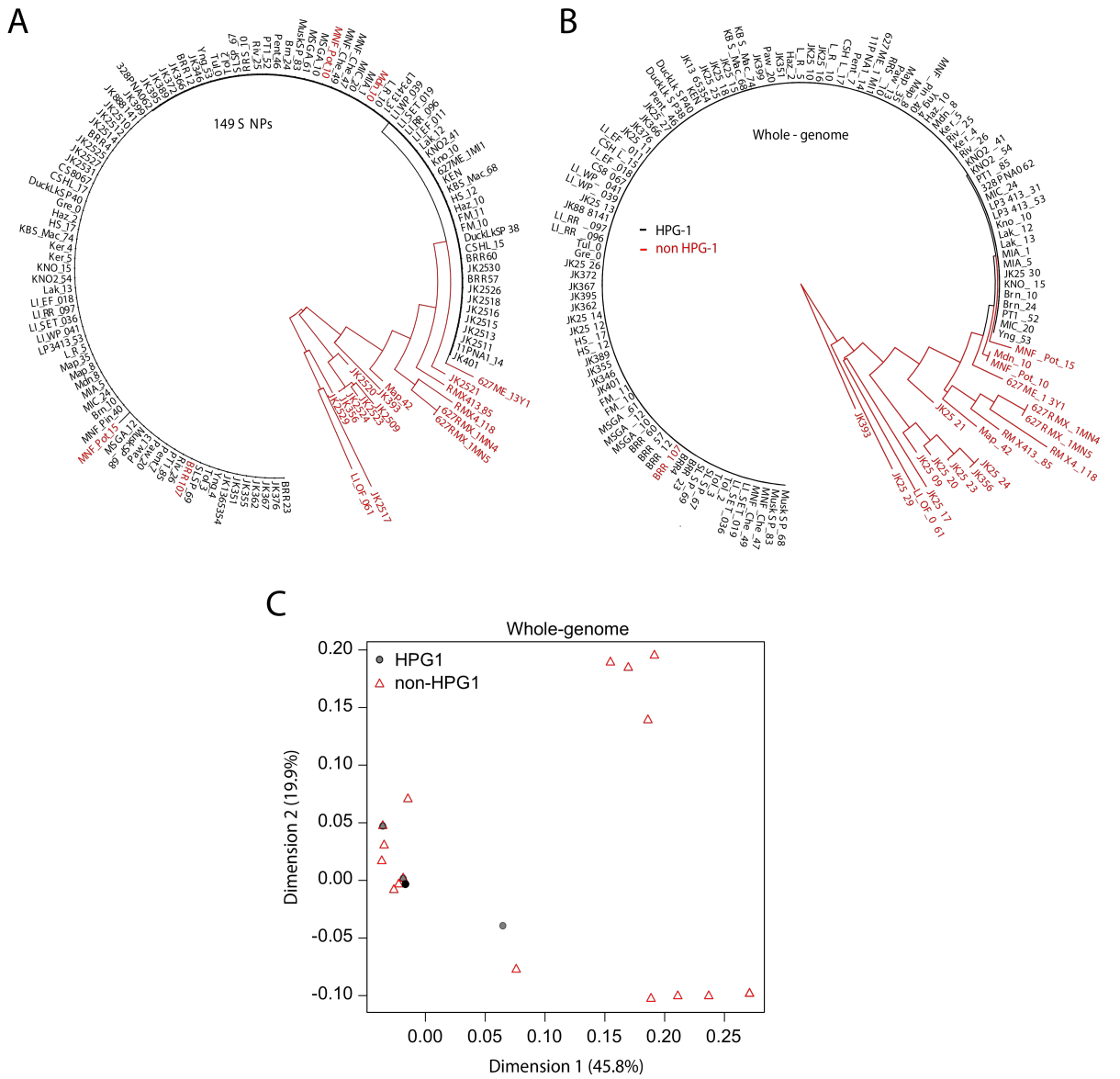


Figure S2. Separation between HPG1 and other North American lineages. **(A)** Neighbor-joining tree built using Illumina-based SNP calls at the 149 genotyping markers originally used to identify HPG1 candidates (consensus of 1,000 replicates). HPG1 accessions are shown in black, whereas other North American lineages are depicted in red. **(B)** Neighbor-joining tree based on genome-wide SNPs (Consensus of 1,000) replicates. Accessions colored as in (A). Note that three accessions originally classified as HPG1 based on 149 SNPs (A) are placed outside this clade. A further accession (BRRR7) within the HPG1 main branch turned out to be a recombinant that was removed from the analysis. **(C)** First two dimensions of a multidimensional scaling plot based on the identity by state pairwise distances. Notice that the black dot arises as a result of plotting multiple almost-identical

HPG1 grey dots. Numbers between parentheses indicate the percentage of the variance explained by each dimension.

Related to Figure 2.

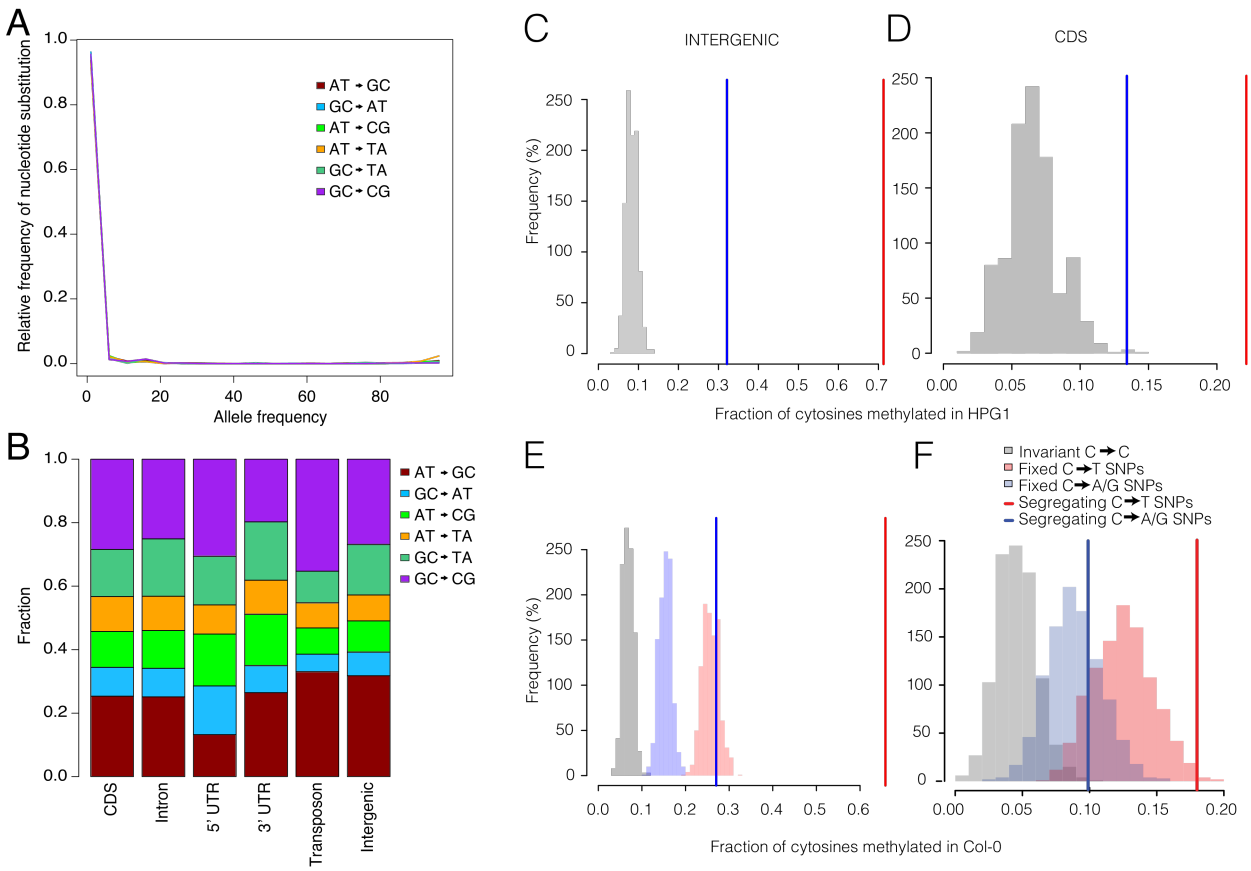


Figure S3. Substitution spectrum and relationship between methylation and substitutions.

(A) "Unfolded" site frequency spectrum using *Arabidopsis lyrata* as outgroup for all transitions and transversions. **(B)** Substitution spectrum for all transitions and transversions divided by genomic annotation. **(C, D)** Fraction of intergenic SNPs **(C)** and coding sequence (CDS) SNPs **(D)** that correspond to methylated cytosines in the HPG-1 pseudo-reference. Methylation data was taken from (Hagmann et al., 2015). **(E, F)** Fraction of intergenic **(E)** and CDS SNPs **(F)** that correspond to methylated cytosines in the Col-0 reference genome (methylation data from (Becker et al., 2011)). Blue and red lines indicate fractions for SNPs segregating within the HPG1 population. Red and blue histograms indicate fractions for subsets of SNPs fixed within the HPG1 population. Grey histograms indicate fractions for invariant positions, i.e., cytosines that have not undergone substitution. See Extended Experimental Procedures for details.

Related to Figure 3.

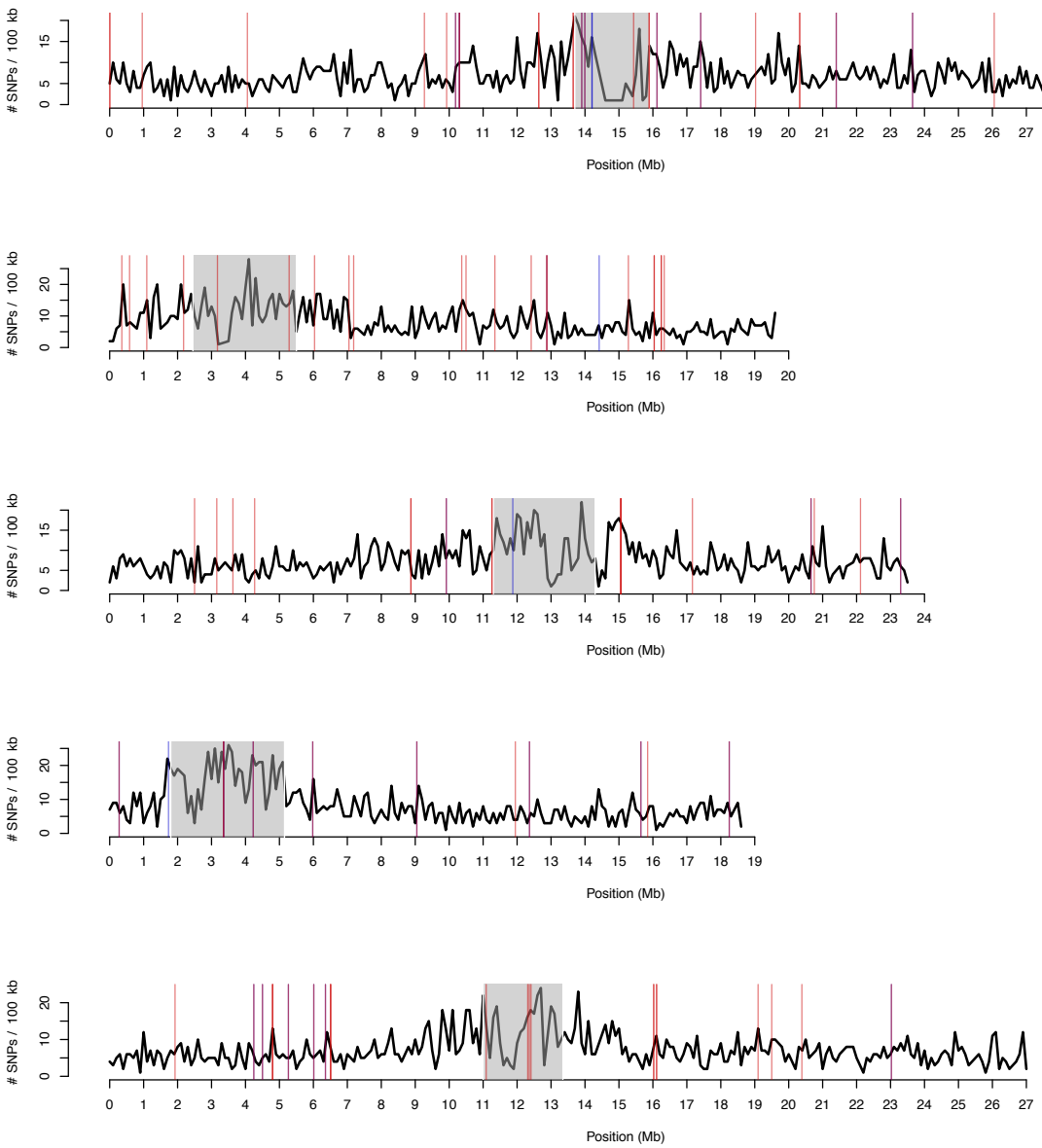


Figure S4. Density of SNPs along all chromosomes and location of SNP hits.

The line shows the number of SNPs per 100 kb window. Centromere locations are indicated by grey background. Vertical lines indicate SNPs associated with root phenotypes (red) and climatic variables (blue).

Related to Figure 5.

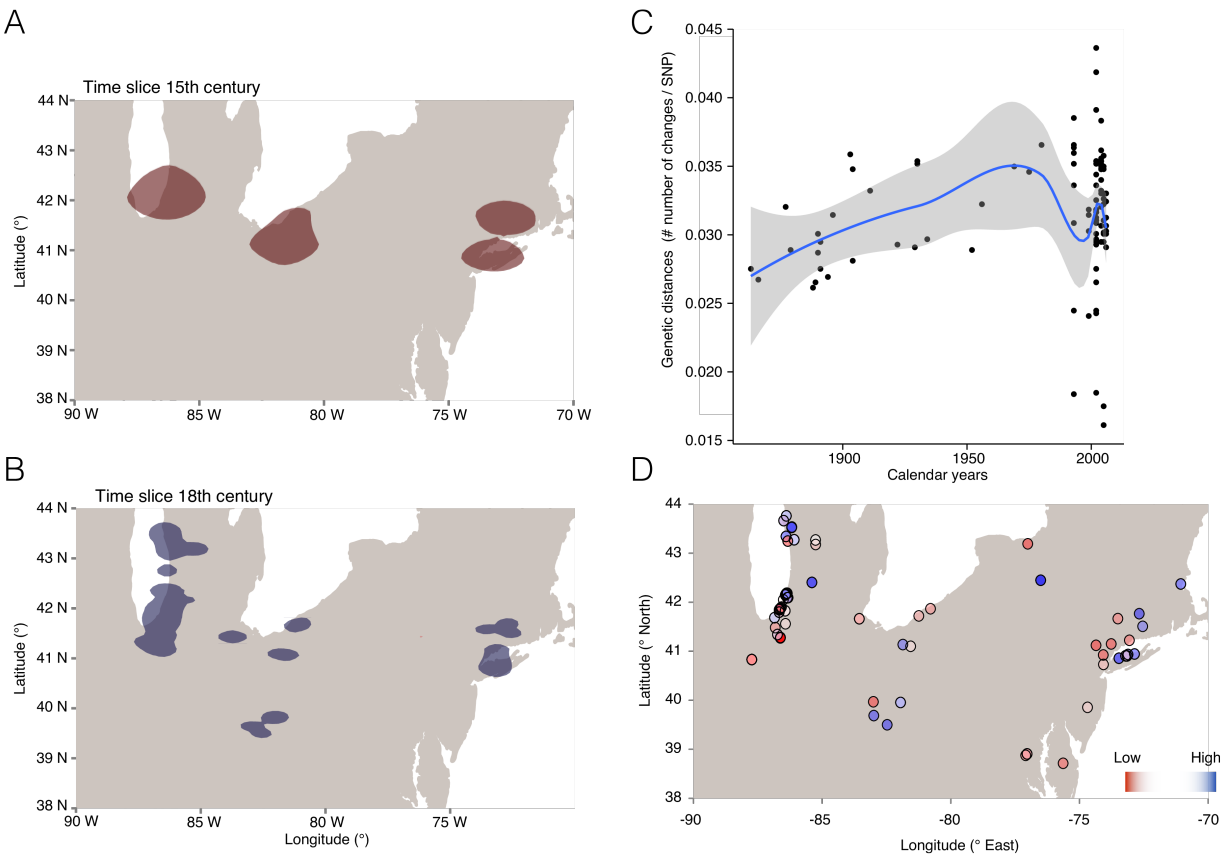
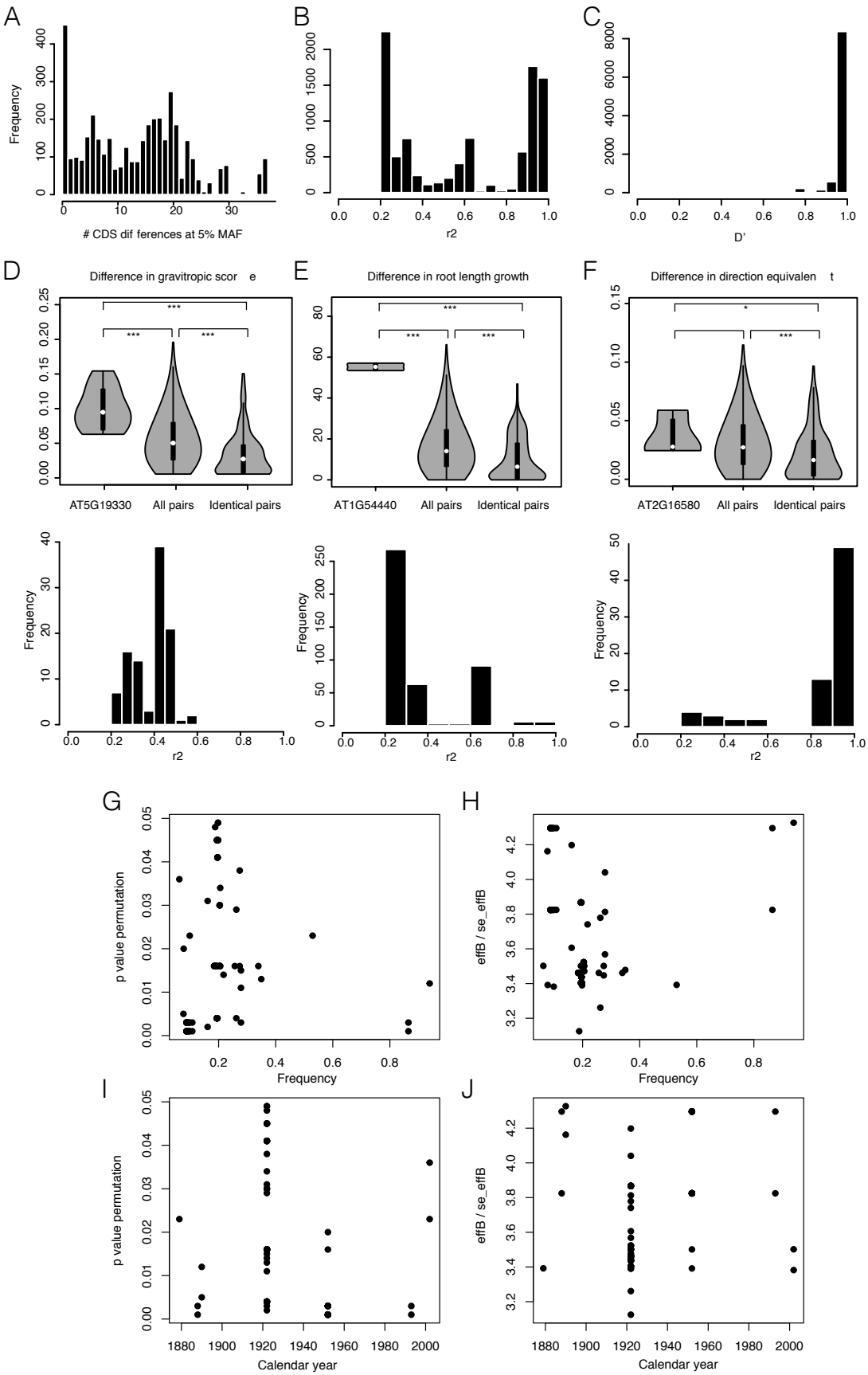


Figure S5. Bayesian phylogeographic inference using continuous trait models, and HPG1 genetic diversity in time and space.

(A, B) The model infers the most probable geographic location of each of the nodes of the phylogeny in Figure 3. **(A)** Ancestral distribution map summarizing the first ~100 years of the phylogenetic tree (green). The clouds represent the 95% interval of the Highest Posterior Probability Density of locations. **(B)** Current distribution map (blue) summarizing the last ~100 years. Clouds as in (A). **(C, D)** Diversity in time and space. **(C)** Diversity in time. Each point represents the average hamming genetic distance among samples within a decade. The black line shows the fit using a generalized additive model and the grey shaded area the 95% confidence interval. **(D)** Diversity in space. Each point represents the average hamming genetic distance among the 10 geographically closest neighbors. Genetic distances are represented as a scaled gradient from red (low) to blue (high) local genetic diversity.

Related to Figure 3 and 4.


A
Frequency
CDS dif ferences at 5% MAF
B
r2
C
D'
D
Difference in gravitropic scor e
AT5G19330
All pairs
Identical pairs
E
Difference in root length growth
AT1G54440
All pairs
Identical pairs
F
Difference in direction equivalen t
AT2G16580
All pairs
Identical pairs
G
p value permutation
Frequency
H
effB / se_effB
Frequency
I
p value permutation
Calendar year
J
effB / se_effB
Calendar year

92

93 **Figure S6. Linkage disequilibrium and SNPs with significant trait associations and**
94 **correlations between SNP effects, frequency and age.**
95 **(A-F)** Linkage disequilibrium and SNPs with significant trait associations. Histogram of genetic distances
96 **(A)** between samples when evaluating only coding regions at 5% minimum allele frequency. Linkage
97 disequilibrium between SNP hits measured as r^2 **(B)** and *D*' **(C)**. Three significant SNPs were further
98 studied to exemplify the power of association analyses with HPG1. For each, phenotypic differences
99 between accessions that differ in the focal SNP and that are otherwise virtually genetically identical are
100 compared both with all pairs of accession and with pairs of accessions completely identical for coding
101 regions. Below each violin plot is the histogram of linkage disequilibrium of the focal SNP with all other
102 SNP hits. The three focal SNPs evaluated are inside AT5G19330 **(D)**, AT1G54440 **(E)** and AT2G16580
103 **(F)** genes. **(G-J)** Correlation between SNP effects, frequency and age. Correlation between SNP
104 frequency and p-value **(G)**, frequency and effect **(H)**, age and p-value **(I)**, age and effect **(J)**.
105 Related to Figure 5.
106
107

108 # SUPPLEMENTAL EXPERIMENTAL PROCEDURES

109 ## Sample collection

110 Modern *A. thaliana* accessions were chosen from the collection described by Platt and colleagues (2010);
111 HPG1 candidates were identified based on 149 genome-wide SNPs (Table S1). Seeds were bulked at the
112 University of Chicago. Progeny for DNA extraction was grown at the Max Planck Institute for
113 Developmental Biology. Herbarium specimens (collection dates 1863-1993) were directly sampled by
114 our colleagues Jane Devos and Gautam Shirsekar, or sent to us by collection curators (Table S1). We
115 used 2 to 8 mm^2 of dried tissue for destructive sampling.

116 ## DNA extraction, library preparation and sequencing

117 DNA from herbarium specimens was extracted in a clean room facility at the University of Tübingen as
118 described (Yoshida et al., 2013). Two sequencing libraries were prepared for each specimen; without
119 and with repair of deaminated sites with uracil-DNA glycosylase and endonuclease VIII (Briggs et al.,
120 2010). DNA from modern, live samples was extracted from rosette leaves pooled from 8 individual
121 plants using the DNeasy plant mini kit (Qiagen, Hilgendorf, Germany). Genomic DNA libraries were
122 prepared using the TruSeq DNA Sample prep kit or TruSeq Nano DNA sample prep kit (Illumina, San
123 Diego, CA). Unrepaired herbarium libraries were screened for authenticity by sequencing at low
124 coverage on Illumina HiSeq 2500 or MiSeq instruments. Production sequencing (101 bp paired end) was
125 carried out on an Illumina HiSeq 2000 instrument.

126 ## Read processing

127 Paired-end reads from modern samples were trimmed and quality filtered before mapping using the
128 SHORE pipeline v0.9.0 (Hagmann et al., 2015; Ossowski et al., 2008). Because ancient DNA fragments
129 are short (Fig. S1B), forward and reverse reads for herbarium samples were merged after trimming,
130 requiring a minimum of 11 bp overlap (Yoshida et al., 2013), and were treated as single-end reads. Reads
131 were mapped with GenomeMapper v0.4.5s (Schneeberger et al., 2009) against an HPG1 pseudo-
132 reference genome (Hagmann et al., 2015), and against the Col-0 reference genome. Samples JK2509 to
133 JK2531 were only mapped to the HPG1 pseudo-reference genome. Coverage, number of covered
134 positions in the genome, and number of SNPs identified per accession relative to HPG1 are reported in
135 Table S1.

136 We also sequenced the genomes of twelve greenhouse-grown mutation accumulation (MA) lines
137 (Becker et al., 2011; Shaw et al., 2000) (Table S2). We called SNPs, indels and structural variants (SVs),
138 following the workflow and parameters described (Hagmann et al., 2015), but without repeated
139 iterations. This procedure resulted in 2,203 polymorphisms that were shared by all lines, indicating
140 errors in the reference sequence (12% of variants replaced N's in the TAIR9 genome) or genetic
141 differences in the founder plant of the MA population compared to the Col-0 individual that had been
142 used to generate the reference genome. In addition, we identified 388 segregating variants across the
143 twelve lines (Table S2), of which 350 were singletons. This analysis revealed on average 25.5 SNPs, 4.9
144 deletions and 3.2 insertions per 31st generation line (Table S2), compared to 19.6 SNPs, 2.4 deletions
145 and 1.0 insertions previously detected in the 30th generation with shorter read length and lower read
146 depth (Ossowski et al., 2010). The genome length accessed in this sequencing effort, 115,954,227 bp,
147 was used to scale the number of point mutations to a rate of 7.1×10^{-9} mutations site^{-1} generation^{-1}
148 (Table S3).

149 **Identification of *bona fide* HPG1 accessions and HPG1 phylogeny**

150 We established the relationships among samples at three levels of resolution: (i) the original 149 nuclear
151 SNP genotyping calls based on which the HPG1 haplogroup had been identified (Platt et al., 2010), (ii)
152 SNPs in the chloroplast genome (where we did not find any variants), (iii) and all nuclear genome SNPs.
153 At these three levels we performed a multidimensional scaling (MDS) analysis and built a neighbor-
154 joining tree using the adegenet package in R (Jombart et al., 2008).

155 We used four methods to estimate the relationships among modern accessions, and between
156 modern accessions and historic specimens: (i) multidimensional scaling (MDS) analysis; (ii) construction
157 of a neighbor joining tree with the adegenet package in R (Jombart, 2008), with branch support assessed
158 with 1,000 bootstrap iterations; (iii) construction of a parsimony network using SplitsTree v.4.12.3
159 (Huson and Bryant, 2006), with confidence values calculated with 1,000 bootstrap iterations; (iv)
160 performing a Bayesian phylogenetic analysis using BEAST v.1.8 (Bouckaert et al., 2014; Drummond et al.,
161 2012) (see below).

162 **Descriptive genome-wide statistics**

163 We estimated genetic diversity as Watterson´s θ and nucleotide diversity π, and the difference between
164 these two statistics as Tajimas's *D* using DnaSP v5 (Librado and Rozas, 2009), both for the entire dataset
165 and independently for modern and herbarium specimens. We calculated the folded and unfolded site
166 frequency spectrum (SFS) for the whole dataset. For the unfolded SFS, we assigned the ancestral state

using the *Arabidopsis lyrata* genome (Hu et al., 2011). We estimated pairwise linkage disequilibrium (LD) between all possible combinations of informative sites, ignoring singletons, by computing r^2, D and D' statistics. LD decay was estimated using a linear regression approach. For the modern individuals, we calculated the recombination parameter R and performed the four-gamete-test (Hudson and Kaplan, 1985) to identify the minimum number of recombination events. All LD and recombination related statistics were determined using DnaSP v5 (Librado and Rozas, 2009).

Substitution and mutation rate analyses

Greenhouse.grown mutation accumulation lines

Mutation rate estimated from greenhouse-grown mutation accumulation lines (Becker et al., 2011) was calculated per line, and the mean and confidence intervals are reported. For each 31st generation MA line, the number of point mutations detected was divided by 31 and by the total genome length. The genome length was determined as all base pairs with coverage higher or equal to 3, and a SHORE mapping quality score of at least 32 in one sample (Table S2).

Natural populations of HPG1

To estimate the number of nucleotide changes per year in natural populations of HPG1, we took advantage of the known collection years of the samples. We used genome-wide nuclear SNPs to calculate pairwise "net" genetic distances between historic and modern HPG1 samples using the equation $D'_{ij} = D_{ic}-D_{jc}$, where D'_{ij} is the net distance between a modern sample i and a historic sample j; D_{ic} the distance between the modern sample i and the reference genome c; and D_{jc} is the distance between a modern sample (j) and the reference genome (c). We calculated a pair-wise time distance in years, T'ij, between all modern and historic pairs using the collection dates and linear regression:

$$D' = a+bT'$$

The slope coefficient b describes the number of substitution changes per year. However, the points in the regression are not independent because different lines have some common evolutionary history, regression confident intervals would be "over-confident". We calculated more rigorous 95% confidence intervals using 1000 bootstrap resamples (Drummond et al., 2003). We used either all SNPs or SNPs at specific annotations. To scale the genome-wide substitution rate into a per-base rate, we used all positions that passed SNP or reference call quality thresholds, instead of using a single value of genome length.

196 The second approach to estimate a substitution rate was framed in Bayesian phylogenetics using
197 the tip-calibration approach implemented in BEAST v1.8 software (Drummond et al., 2012). After
198 systematic runs and chain convergence assessment of different demographic and molecular clock
199 models, we determined that the Skygrid demographic model and the lognormal relaxed molecular clock
200 were the most appropriate. Our analysis simultaneously optimized tree topology and length, substitution
201 rate, and the demographic model. Using the relationship between the time distance of two sequences
202 and the difference in branch length in the tree, BEAST estimates a molecular clock. Under a relaxed
203 molecular clock, the substitution rate is allowed to vary across branches with a lognormal distribution.
204 The prior used for molecular clock was a Continuous-Time Markov Chain (CTMC) (Ferreira and
205 Suchard, 2008). The demographic model is a Bayesian nonparametric demographic model that is
206 optimized for multiple loci, and which allows for complex demographic trajectories by estimating
207 population sizes in time bins (of 10 years in our case) across the tree, based on the number of
208 coalescent events per bin (Gill et al., 2012). In addition, to confirm that demography and root dating
209 converged on the same parameters, we performed a second estimate using a fixed substitution rate of
210 3.3×10^{-9} substitutions $site^{-1}$ $year^{-1}$ that we had estimated empirically using the net-distance method.

211 The analysis was carried out remotely at CIPRES PORTAL (v3.1 www.phylo.org) using
212 uninformative priors. The run took about 1,344 CPU hours and performed 1,000 million steps in a
213 Monte Carlo Markov Chain (MCMC), sampling every 100,000 steps. Burn-in was adjusted to 10% of
214 steps. To visualize the tree output we produced a Maximum Clade Credibility (MCC) tree with a
215 minimum posterior probability threshold of 0.8 and a 10% burn-in using TreeAnnotator (part of BEAST
216 package), and visualized the MCC tree using FigTree (tree.bio.ed.ac.uk/software/figtree/). Additionally,
217 we used DensiTree (Bouckaert, 2010) to draw simultaneously the 10,000 BEAST trees with the highest
218 posterior probability. Since all trees were drawn transparently, agreements in both topology and branch
219 lengths appear as densely colored regions (Fig. 3A), while areas with little agreement appear lighter.

220 **Demography and migration of HPG1**

221 From the Bayesian phylogenetic analyses described in previous sections, we studied the demographic
222 model estimated via Skygrid. We reconstructed a skyline plot that depicts changes in effective
223 population size, a measure of relative diversity, through time (Bouckaert et al., 2014; Drummond et al.,
224 2012). Implementation of non-phylogenetic methodologies for demographic inference exist, e.g. Multiple
225 sequentially Markovian coalescent (MSMC) (Schiffels and Durbin, 2014), but after exploring them we
226 concluded that their resolution is not sufficient for analyses of the last several centuries, as in our case.

227 We performed another Bayesian phylogenetic analysis incorporating a geographic location trait
228 (Lemey et al., 2010; Wilson and Barton, 1995). For this, Brownian diffusion parameters are estimated by
229 fitting a continuous gradient of geographic locations along tree branches, starting from the leaves of the
230 tree for which geographic locations are known, i.e. the collection sites of our samples. We excluded
231 three samples from the West coast of the United States, since propagation by Brownian diffusion along
232 large distances is an unrealistic model. We ran this analysis with the parameters described in the
233 previous sections and sliced the resulting 3D (temporal and geographical) phylogeny at the early 16th
234 century and late 18th century using SPREAD software (Bielejec et al., 2011).

235 We used a heuristic search using an isolation-by-distance pattern inspired by (Handley et al.,
236 2007) to find the origin of diffusion of HPG1 in North America, and compared it to the phylogeography
237 analyses. We performed pairwise tests of the relation between genetic and geographic distances using a
238 linear regression. Afterwards we decomposed for each sample the isolation-by-distance pattern (i.e.
239 each row of both distance matrices), and tested whether the slope of the regression still held, that is,
240 whether the remaining samples showed a gradual increase in genetic distance as they moved away from
241 the presumed origin. The sample locations that showed the steepest and most significant slopes were
242 assumed to have been closest to the origin of HPG1 diffusion. Because there are indications that more
243 than a single spread of the groups might have happened, we performed the isolation by distance analyses
244 for modern accessions and herbarium specimens separately. These two analyses allowed us to locate
245 the origin of the modern and historic diffusions of HPG1 in North America, respectively. The analysis
246 consisted of a heuristic search across all sampled locations, in which a regression between genetic
247 *distance ~ Euclidean geographic distances* was performed.

248 **Analysis of the methylation status of mutated sites**

249 As in many other species, the spectrum of *de novo* mutations in *A. thaliana* is biased towards G:C→A:T
250 transitions in greenhouse-grown mutation accumulation lines (Ossowski et al., 2010), leading to an
251 inflated transition-to-transversion ratio (Ts/Tv). This bias is less pronounced in recent mutations in a
252 Eurasian collection of natural accessions (Cao et al., 2011) and in HPG1 accessions (Fig. 3D). A recent
253 multigenerational salt stress experiment in the greenhouse also showed a more balanced Ts/Tv (Jiang et
254 al., 2014). These findings indicate that less benign conditions might promote a lower Ts/Tv.

255 The mechanisms underlying a high Ts/Tv ratio are unknown, but could include spontaneous deamination
256 of methylated cytosines (5-methyl-C → T). In agreement with this possibility, we found previously that
257 ancestral cytosines methylated in the *A. thaliana* reference strain had a more than two-fold higher

258 polymorphism rate than unmethylated cytosines, with the highest rate found in CHG sites (where H is
259 A, C or T) (Table 1, Cao et al., 2011).

260 We interrogated the putative evolutionary role of cytosine methylation in the mutability of
261 cytosine bases in the HPG1 accessions. For reference DNA methylation data, we used previously
262 generated bisulfite-sequencing data of HPG1 strains (Hagmann et al., 2015) and of Col-0 lines (Becker et
263 al., 2011), respectively. Our rationale was that if methylation affected mutability, this should reflect in the
264 proportion of mutated sites being methylated in the reference datasets, compared to that proportion
265 for non-mutated sites. To be able to determine the ancestral state of a given site, we only considered
266 positions for which we could determine that state by alignment to the *A. lyrata* genome (Hu et al., 2011).

267 The test set of genomic positions consisted of the *n* sites that were invariant cytosines in *A.*
268 *lyrata* and the *A. thaliana* Col-0 reference genome and whose derived allele was present in at least one
269 HPG1 accession (i.e., SNPs segregating within the HPG1 population). For these sites, we determined the
270 fraction of methylated cytosines as the number of corresponding sites classified as 'methylated' in the
271 HPG1 and Col-0 reference datasets, respectively, divided by *n*.

272 As a first control set of sites, hence called 'neutral', we selected cytosines that were invariant
273 between *A. lyrata*, Col-0, and all HPG1 accessions. A second control set, which we called 'fixed',
274 consisted of cytosines that were invariant between *A. lyrata* and Col-0, and that had mutated and had
275 been fixed in all HPG1 accessions. For both control sets we generated empirical distributions of the
276 fraction of sites that were methylated in the HPG1 and Col-0 reference datasets, respectively. To this
277 end, we randomly selected *n* positions with sequence information in the methylation datasets; this
278 process was repeated 1,000 times.

279 Ancestral cytosines with higher methylation proportion in both *A. thaliana* and HPG1 methylome
280 datasets were more likely to mutate to thymines (Fig. S3 C-F). Surprisingly, not only C→T but also
281 C→A/G segregating sites were more likely to have been methylated compared to the fixed and neutral
282 positions, which cannot be explained by higher deamination rates of methylated vs. unmethylated
283 cytosines.

284 There is an ongoing debate on how epigenetics, i.e. environmentally-induced modification with
285 non-Mendelian inheritance, could contribute to adaptation (Mirouze and Paszkowski, 2011; Nicotra et
286 al., 2010). This result could certainly constitute a genetically-based hypothesis of epigenetic roles in
287 adaptation, perhaps in favor of the "adaptive mutation" argument heavily evidenced in bacteria (Al
288 Mamun et al., 2012).

289 **Inference of genome-wide selection parameters**

290 We estimated the average strength of genome-wide selection using the non-equal relationship between
291 whole-genome and intergenic substitution rates. We selected intergenic regions as the neutral reference
292 because they should not involve any direct phenotypic or biochemical effect but have abundant sites to
293 compare with. This was based on the well-known relationship described by Kimura (1967):

294 $$k = \mu \times Q \times 2N_e,$$

295 where k is the substitution rate, μ the mutation rate, N_e the effective population size, and Q the fixation
296 probability of a new mutation. Under neutrality, substitution and mutation rate should be equal since
297 $Q = ½ N_e$ and the effective population size term, $2N_e$, cancels out in the equation. With a semidominant
298 genome-wide selection coefficient s acting on a new mutation, $Q \approx s / 2N_e (1-e^{-2N_e s})$ (Charlesworth and
299 Charlesworth, 2010). We used the intergenic substitution rate as proxy for the mutation rate μ and the
300 whole-genome substitution rate as proxy for the substitution rate k. We solved the equation for $2N_e s$,
301 known as the population selection parameter.

302 $$k / \mu = 2N_e \times (s / 2N_e (1-e^{-2N_e s}))$$

303 **Association analyses and dating of newly arisen mutations**

304 For 63 modern accessions, we measured time to bolting and flowering with four replicates, and
305 fecundity (as seed set) with one replicate in growth chambers at the University of Chicago. Additionally,
306 using ≥ 10 replicates we analyzed primary root phenotypes at the Gregor Mendel Institute in Vienna,
307 describing growth and morphological traits extracted from images as described (Slovak et al., 2014) (see
308 next section for details in phenotypic characterization). For associations with climate parameters, we
309 followed a similar rationale as described (Hancock et al., 2011). We extracted information from the
310 publicly available bioclim database (http://www.worldclim.org/bioclim) at 2.5 degrees resolution raster
311 and intersected it with geographic locations of HPG1 samples (n = 103).

312 We performed association analyses using the R package GenABEL (Aulchenko et al., 2007), with
313 measured phenotypes (p = 25) and climatic variables (c = 18) as response variables and SNPs as
314 explanatory variables. A Minimum Allele Frequency cutoff 5% was used. The number of assessed SNPs
315 was 391 in a dataset of only modern samples but imputed genotypes for missing data using Beagle v4.0
316 (Browning and Browning, 2009), and 456 SNPs with a dataset of modern and also historic samples,
317 although without imputation. For all associations, either phenotypic or climatic ones, minimum 63
318 individuals were genotyped for a SNP. All phenotypic variables were measured in common chamber or
319 common garden experiments. We first investigated broad sense heritability (*H2*) of each trait using

320 ANOVA partition of variance between and within lines using replicates (Table S4). Significance was
321 obtained by common F test in ANOVA. Secondly we used the *polygenic_hglm* function in GenABEL to fit
322 a genome wide kinship matrix in order to calculate a narrow sense heritability estimate (*h2*). Significance
323 was calculated employing a likelihood ratio test comparing with a null model. In principle, h2 is a
324 component of H2, then its values should always be h2 < H2. Our result cannot be interpreted in this
325 framework, since we employed genotype means for h2 calculation and replicate measurements for H2
326 calculation. This reduced the environmental and developmental noise and thus inflated h2 (Table S4). In
327 this framework, however, we could calculate h2 for climatic variables as well. Seed size had a particularly
328 high heritability, a pattern attributed to highly accurate and replicated measurements (see Phenotyping
329 section). For association analyses we first employed a linear mixed model that fitted the kinship matrix
330 using the *mmscore* function, and only three significant SNP hits were discovered using a 5% significance
331 threshold after False Discovery Rate correction (FDR). This was expected since we have very few
332 variants and these would have originated in an approximated phylogeny structure. We concluded that
333 fitting the kinship matrix in our model was not appropriate since there would be no leftover variation
334 for association with specific SNPs. With this rationale we employed a fixed effects linear model using the
335 function *qtscore* from GenABEL. To reduce the false-positive rate we took a conservative permutation
336 strategy that consisted in carrying out association analyses over 1,000 random datasets (permuting
337 phenotypes across individuals) and used the resulting p-value distribution to correct p-values estimated
338 with the original dataset. SNPs with p-values below 5% in the empirical p-value distribution were
339 considered significant (Table S5). In climatic models, we additionally included longitude and latitude as
340 covariates to correct any spurious association between SNPs and climate gradients created by the
341 migratory pattern of isolation by distance. Significant SNPs were interspersed throughout the genome
342 (Fig. S4) and their p-value and phenotypic effect did not correlate with the putative age of the SNPs
343 neither with the frequency, something that could have indicated that the significance was merely driven
344 by the higher statistical power of intermediate frequency variants (Fig. S5 G-J). Using QQ plots to assess
345 inflation or deflation of p-values, we observed generally that permutation corrected p-values were
346 deflated. Straight series of points in QQ plots indicate identical p-values for multiple SNPs, a pattern that
347 we attributed to long range LD, i.e. lack of independence (see Graphical Table S6 for trait distributions
348 and QQ plots from each association analysis). Due to this fact, we add two correction procedures
349 more: (1) Bonferroni-correcting the significance threshold for permutation corrected SNPs from 5% to
350 5% / number of traits, i.e. 0.2% for phenotype association and 0.27% for climatic associations. (2)
351 Bonferroni-correcting the significance threshold for raw p-values from 5% to 5% / (number of SNPs +
352 number of traits), i.e. 0.01% for phenotype and climatic associations (Table 1 and S5).

353 For each SNP in our dataset, we determined directionality of mutation, i.e. ancestral vs derived
354 alleles, by determining which state was found in the oldest herbarium samples. We compared the time
355 of emergence and the centroid of geographic distribution of the alternative alleles of SNP hits to random
356 draws of SNPs with the same minimum allele frequency filtering (5%).

357 On top of phenotypic and climatic associations of SNP hits, we also provide a putative protein
358 effect employing a commonly used amino acid matrix of biochemical effects (Grantham, 1974). Gene
359 name and ontology categorization of SNPs inside annotated transcriptional units was extracted from the
360 online tool www.arabidopsis.org/tools/bulk/go/.

361 **Association analyses – proof of concept examples**

362 We argue that the power of an association approach relies on the fact that HPG1 lines resemble Near
363 Isogenic Lines (NILs) produced by experimental crosses (Weigel, 2012). Similarly to genome-wide
364 association studies (GWA), power depends on a number of factors, namely the noise of phenotype
365 under study, architecture of phenotypic trait, quality of genotyping, population structure, sample
366 diversity, sample size, allele frequency, and recombination. On one hand, association analyses in NILs
367 suffers from large linkage blocks, but confident results can be achieved due to accurate measurement of
368 phenotypes, limited genetic differences between any two lines, and high quality genotypes. In common
369 GWA such as in humans, there are multiple confounding effects. Among the confounders are (1) that
370 any two samples differ in hundreds of thousands of SNP and (3) that historical and geographic
371 stratification produce non-random correlations among those SNP differences. This complicates
372 considerably the identification of phenotypic effects at specific genes, and power relies greatly on large
373 samples and frequent recombination between markers.

374 We exemplify the association analysis confidence with some examples. To provide support for
375 the nonsynonymous SNP on chromosome 5, at position 6,508,329 in AT5G19330, we looked for pairs
376 of lines that carry the ancestral and the derived allele, but that differ in few (or no other) SNP in the
377 genome. When considering all genic substitutions with a minimum allele frequency of 5% (Fig. S6A), we
378 identified 20 pairs of lines differing only in the AT5G19330 SNP and another linked SNP (which is
379 located on a different chromosome and had an association p-value > 0.4). The phenotypic differences in
380 mean gravitropic score of these almost-identical pairs were significantly higher than phenotypic
381 differences among all pairs of HPG1 lines, and genetically identical pairs attending to substitutions inside
382 genes (Fig. S6D). Furthermore this SNP was not in linkage disequilibrium with any other SNP hit (r^2 <
383 0.6) (Fig. S6D). A similar approach was used to examine the SNPs in AT2G02220 (Fig. S6E) and
384 AT2G16580 (Fig. S6F).

385 **Phenotyping**

386 Root phenotypes

387 Fifteen root phenotypes were scored for three replicates per genotype over a time-series experiment
388 via image analysis as described in detail in (Slovak et al., 2014). We used the mean and standard
389 deviation of the time series values for association analyses.

390 Seed size phenotype:

391 We dispersed the seeds of given genotypes on separate plastic square 12 x 12 cm Petri dishes. For
392 faster image acquisition we used cluster of eight Epson V600 scanners. The scanner cluster was
393 operated by the BRAT Multiscan image acquisition tool (https://www.gmi.oeaw.ac.at/research-
394 groups/wolfgang-busch/resources/brat/). The resulting 1600 dpi images were analyzed in Fiji software.
395 Scans were converted to 8-bit binary images, thresholded (parameters: setAutoThreshold("Default
396 dark"); setThreshold(20, 255)) and particles analyzed (inclusion parameters: size=0.04-0.25
397 circularity=0.70-1.00). The 2D seed size was measured in square millimeters (parameters: distance=1600
398 known=25.4 pixel=1 unit=mm) on ≥ 500 replicates per genotype.

399 Flowering time in growth chambers

400 We estimated the flowering time in growth chambers under 4 vernalization treatments. We grew 6
401 replicates per accession divided between two complete randomized blocks for each treatment. Seeds
402 were sown on a 1:1 mixture of Premier Pro-Mix and MetroMix and cold stratified for 6 days (6°C, no
403 light). We then let plants germinate and grow at 18°C, 14 hours light, 65% humidity. After 3 weeks, we
404 transferred the plants to the vernalization conditions (6°C, 8 hours light, and 65% humidity). The 4
405 treatments consisted of 0, 14, 28 and 63 days of vernalization, respectively. After the vernalization
406 treatment, plants were transferred back to the previous long day growth conditions. Trays were rotated
407 around the growth chambers every other day throughout the experiments, under both vernalization and
408 growth conditions. Germination, bolting and flowering dates were recorded every other day until all
409 plants had flowered. Days till flowering or bolting times were calculated from the germination date until
410 the first flower bud was developed and until the first flower opened, respectively. The average flowering
411 time and bolting time per genotype was used for association analyses.

412 Flowering time and fecundity in the field

413 To investigate variation in flowering time and fecundity in natural conditions, we grew 3 replicates for
414 each of the 78 accessions in a field experiment following a completely randomized block design. Seeds
415 were sowed between 09/20/2012 to 09/22/2012 in 66 well trays (well diameter=4cm) on soil from the

field site where plants were to be transplanted. The trays were cold stratified for seven days before being placed in a cold frame at the University of Chicago, IL, USA, (outdoors, no additional light or heat, but watered as needed and protected from precipitation). Seedlings were transplanted directly into tilled ground at the Warren Wood field station (41.84° North, 86.63° West), Michigan, USA on 10/13/2012 and 10/14/2012. Seedlings were watered-in and left to overwinter without further intervention. Plants were scored for bolting and flowering in Spring 2013. Upon maturation of all fruits, stems were harvested and stored between sheets of newsprint paper. To estimate the fecundity, stems were photographed on a black background and the size of each plant was estimated as the number of pixels occupied by the plant on the image. This measure correlates well with the total length of siliques produced, a classical estimator of fecundity in *A. thaliana* (Spearman's rho=0.84, *p*-value<0.001, data not shown).

427 **SUPPLEMENTAL REFERENCES**

428 Al Mamun, A.A., Lombardo, M.J., Shee, C., Lisewski, A.M., Gonzalez, C., Lin, D., Nehring, R.B., Saint-Ruf,
429 C., Gibson, J.L., Frisch, R.L., *et al.* (2012). Identity and function of a large gene network underlying
430 mutagenic repair of DNA breaks. Science *338*, 1344-1348.
431 Aulchenko, Y.S., Ripke, S., Isaacs, A., and van Duijn, C.M. (2007). GenABEL: an R library for genome-
432 wide association analysis. Bioinformatics *23*, 1294-1296.
433 Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D. (2011).
434 Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature *480*, 245-249.
435 Bielejec, F., Rambaut, A., Suchard, M.A., and Lemey, P. (2011). SPREAD: spatial phylogenetic
436 reconstruction of evolutionary dynamics. Bioinformatics *27*, 2910-2912.
437 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.a., Rambaut, A., and
438 Drummond, A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS
439 Comp Biol *10*, e1003537.
440 Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S. (2010). Removal of deaminated
441 cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res *38*, e87.
442 Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-
443 phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet *84*, 210-223.
444 Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O.,
445 Lippert, C., *et al.* (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat
446 Genet *43*, 956-963.
447 Charlesworth, B., and Charlesworth, D. (2010). Elements of Evolutionary Genetics (Roberts and
448 Company: Greenwood Village, CO, 2010).
449 Drummond, A., Pybus, O.G., and Rambaut, A. (2003). Inference of viral evolutionary rates from
450 molecular sequences. Adv Parasitol *54*, 331-358.
451 Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti
452 and the BEAST 1.7. Mol Biol Evol *29*, 1969-1973.
453 Ferreira, M.A.R., and Suchard, M.A. (2008). Bayesian analysis of elapsed times in continous-time
454 Markovchains. Can J Stat *36*, 355-368.
455 Gill, M.S., Lemey, P., Faria, N.R., Rambaut, A., Shapiro, B., and Suchard, M.a. (2012). Improving Bayesian
456 population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol *30*, 713-
457 724.
458 Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science *185*, 862-
459 864.

460 Hagmann, J., Becker, C., Müller, J., Stegle, O., Meyer, R.C., Wang, G., Schneeberger, K., Fitz, J., Altmann,
461 T., Bergelson, J., *et al.* (2015). Century-scale methylome stability in a recently diverged *Arabidopsis*
462 *thaliana* lineage. PLoS Genet *11*, e1004920.
463 Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C.,
464 Roux, F., and Bergelson, J. (2011). Adaptation to climate across the *Arabidopsis thaliana* genome.
465 Science *334*, 83-86.
466 Handley, L.J.L., Manica, A., Goudet, J., and Balloux, F. (2007). Going the distance: human population
467 genetics in a clinal world. Trends Genet *23*, 432-439.
468 Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood,
469 J., Gundlach, H., *et al.* (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome
470 size change. Nat Genet *43*, 476-481.
471 Hudson, R.R., and Kaplan, N.L. (1985). Statistical properties of the number of recombination events in
472 the history of a sample of DNA sequences. Genetics *111*, 147-164.
473 Huson, D.H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol
474 Biol Evol *23*, 254-267.
475 Jiang, C., Mithani, A., Belfield, E.J., Mott, R., Hurst, L.D., and Harberd, N.P. (2014). Environmentally
476 responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations.
477 Genome Res *24*, 1821-1829.
478 Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics
479 *24*, 1403-1405.
480 Lemey, P., Rambaut, A., Welch, J.J., and Suchard, M.a. (2010). Phylogeography takes a relaxed random
481 walk in continuous space and time. Mol Biol Evol *27*, 1877-1885.
482 Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA
483 polymorphism data. Bioinformatics *25*, 1451-1452.
484 Mirouze, M., and Paszkowski, J. (2011). Epigenetic contribution to stress adaptation in plants. Curr Opin
485 Plant Biol *14*, 267-274.
486 Nicotra, A.B., Atkin, O.K., Bonser, S.P., Davidson, A.M., Finnegan, E.J., Mathesius, U., Poot, P.,
487 Purugganan, M.D., Richards, C.L., Valladares, F., *et al.* (2010). Plant phenotypic plasticity in a
488 changing climate. Trends Plant Sci *15*, 684-692.
489 Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008).
490 Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res *18*, 2024-2033.
491 Ossowski, S., Schneeberger, K., Lucas-Lledo, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D.,
492 and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis*
493 *thaliana*. Science *327*, 92-94.

494 Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Agren, J., Bossdorf, O., Byers, D.,
495 Donohue, K., *et al.* (2010). The scale of population structure in *Arabidopsis thaliana*. PLoS Genet *6*,
496 e1000843.
497 Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple
498 genome sequences. Nat Genet.
499 Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., and Weigel, D.
500 (2009). Simultaneous alignment of short reads against multiple genomes. Genome Biol *10*, R98.
501 Shaw, R.G., Byers, D.L., and Darmo, E. (2000). Spontaneous mutational effects on reproductive traits of
502 *Arabidopsis thaliana*. Genetics *155*, 369-378.
503 Slovak, R., Goschl, C., Su, X., Shimotani, K., Shiina, T., and Busch, W. (2014). A scalable open-source
504 pipeline for large-scale root phenotyping of *Arabidopsis*. Plant Cell *26*, 2390-2403.
505 Weigel, D. (2012). Natural variation in Arabidopsis: from molecular genetics to ecological genomics.
506 Plant Physiol *158*, 2-22.
507 Wilson, I., and Barton, N.H. (1995). Genealogies and geography. Philosophical transactions of the Royal
508 Society of London Series B, Biological sciences *349*, 49-59.
509 Yoshida, K., Schuenemann, V.J., Cano, L.M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin, F.N.,
510 Kamoun, S., Krause, J., *et al.* (2013). The rise and fall of the *Phytophthora infestans* lineage that
511 triggered the Irish potato famine. eLife 2, e00731.

512

Table S1. Sample information, Related to Figure 1.

Latitude and longitude for historic samples were imputed from the geographic centroid of the most accurate toponym described in the herbarium specimen label.
Key of abbreviations of herbarium collections or seed sources:
UCONN = University of Connecticut herbarium; CFM = Chicago Field Museum;
NY = New York Botanical Garden; ABRC = Arabidopsis Biological Resources Center;
OSU = Ohio State University
H* indicate herbarium samples that cluster with the modern HPG1 clade rather than the historic HPG1 clade in Fig. 3.
Geographic location highlighted in Fig. 1.

Accession	Latitude (°N)	Longitude (° E)	State	Date collected	Alternative name	Collector/ Herbarium	Average coverage (x)
JK399	38.7155	-75.635591	DE	1863	888124	NY	9
JK366	43.1921	-77.0102	NY	1866	888144	NY	6.8
JK395	38.9068	-77.036667	DC	1877	888134	NY	10.3
JK888141	40.732007	-74.068455	NJ	1879	888141	NY	42
JK389	38.9068	-77.036667	DC	1888	1365363	NY	9.9
JK362	38.9068	-77.036667	DC	1889	1365364	NY	8.8
JK367	40.9249	-74.0755	NJ	1890	1365344	NY	16.7
JK372	41.1222	-74.3569	NJ	1890	1365332	NY	14.8
JK1365354	38.8782	-77.09048	VA	1891	1365354	NY	36.4
JK376	39.97	-83.01	NY	1891	1365337	NY	12.3
JK351	41.15	-73.766667	NY	1894	1365333	NY	16.1
JK355	35.99	-83.94	TE	1896	1365374	NY	14.3
JK356	n/a	n/a	GA	1897	1365375	NY	5.3
JK393	n/a	n/a	NC	1897	1365370	NY	30.4
JK346	40.643136	-111.95177	UT	1903	102365	NY	29.1
JK2525	41.224343	-73.06021	CT	1904	79391	UNCONN	12.5
JK2529	n/a	n/a	OH	1904	176849	CFM	11.4
JK401	40.643136	-111.95177	UT	1904	102364	NY	10.4
JK2513	41.102121	-81.560547	OH	1911	25	OSU	18.2
JK2509	n/a	n/a	CT	1917	11	OSU	15.1
JK2530	41.482862	-86.822602	IN	1922	531679	CFM	22.2
JK2526	41.666667	-73.508455	CT	1929	79409	UNCONN	16.3
JK2515	41.137296	-81.863779	OH	1930	30	OSU	21.3
JK2511	41.721618	-81.243317	OH	1934	14	OSU	5.6
JK2523	n/a	n/a	OH	1940	25707	UNC	13.1

JK2520	n/a	n/a	OH	1945	54051	UNC	20.3
JK2524	39.856783	-74.686954	NJ	1952	63978	UNC	13.8
JK2512	39.95607	-81.953309	OH	1956	21	OSU	16.7
JK2514	39.95607	-81.953309	OH	1969	27	OSU	28.4
JK2517	n/a	n/a	OH	1981	34	OSU	21.7
JK2521	n/a	n/a	OH	1992	565960	UNC	2.9
JK2518	41.867643	-80.789021	OH	1993	40	OSU	14.8
JK2531	39.856783	-74.686954	NJ	1952	1507461	CFM	15.1
JK2510	39.688861	-82.993218	OH	1930	13	OSU	21
JK2527	41.509059	-72.543694	CT	1975	79389	UNCONN	8.3
JK2516	39.500862	-82.472413	OH	1980	32	OSU	18.1
CSHL_15	40.8585	-73.4675	NY	1993	CSHL-15	ABRC	39.3
CSHL_17	40.8585	-73.4675	NY	1993	CSHL-17	ABRC	41.5
FM_10	42.4489	-76.5072	NY	1993	FM-10	ABRC	44.6
FM_11	42.4489	-76.5072	NY	1993	FM-11	ABRC	44.4
HS_12	42.373	-71.0627	MA	1993	HS-12	ABRC	48.8
HS_17	42.373	-71.0627	MA	1993	HS-17	ABRC	55.3
Kno_10	41.2816	-86.621	IN	1993	Kno-10	ABRC	39.4
KNO_15	41.2816	-86.621	IN	1993	KNO-15	ABRC	43.6
Gre_0	43.178	-85.2532	MI	1995	Gre-0	ABRC	44.6
Tul_0	43.2708	-85.2563	MI	1995	CS6877	ABRC	31.2
CS8067	41.3599	-122.755	CA	1996	uckhorn Pas	ABRC	66.4
Tol_2	41.6639	-83.5553	OH	1996	CS8022	ABRC	61
Tol_3	41.6639	-83.5553	OH	1996	CS8023	ABRC	40.2
MIA_1	41.7976	-86.6691	MI	1999	MIA-1	ABRC	73.1
MIA_5	41.7976	-86.6691	MI	1999	MIA-5	ABRC	62.9
MIC_20	41.8266	-86.4366	MI	1999	MIC-20	ABRC	39.9
MIC_24	41.8266	-86.4366	MI	1999	MIC-24	ABRC	33.8
Brn_10	41.9	-86.583	MI	2002	Brn-10	ABRC	33.3
Brn_24	41.9	-86.583	MI	2002	Brn-24	ABRC	38.4
Haz_10	41.879	-86.607	MI	2002	Haz-10	ABRC	33.8
Haz_2	41.879	-86.607	MI	2002	Haz-2	ABRC	39.7
Ker_4	42.184	-86.358	MI	2002	Ker-4	ABRC	32.1
Ker_5	42.184	-86.358	MI	2002	Ker-5	ABRC	62.9
L_R_10	41.847	-86.67	MI	2002	L-R-10	ABRC	22.4
L_R_5	41.847	-86.67	MI	2002	L-R-5	ABRC	60.6
Lak_12	41.8	-86.67	MI	2002	Lak-12	ABRC	37.8
Lak_13	41.8	-86.67	MI	2002	Lak-13	ABRC	28.5
Map_35	42.166	-86.412	MI	2002	Map-35	ABRC	64.7
Map_42	42.166	-86.412	MI	2002	Map-42	ABRC	46
Map_8	42.166	-86.412	MI	2002	Map-8	ABRC	33.4
Mdn_10	42.051	-86.509	MI	2002	Mdn-10	ABRC	34.9
Mdn_8	42.051	-86.509	MI	2002	Mdn-8	ABRC	37.4
Paw_13	42.148	-86.431	MI	2002	Paw-13	ABRC	43
Paw_20	42.148	-86.431	MI	2002	Paw-20	ABRC	41.3
Riv_25	42.184	-86.382	MI	2002	Riv-25	ABRC	36.8
Riv_26	42.184	-86.382	MI	2002	Riv-26	ABRC	35.7

Yng_4	41.865	-86.646	MI	2002	Yng-4	ABRC	41.3
Yng_53	41.865	-86.646	MI	2002	Yng-53	ABRC	46
RRS_10	41.5609	-86.4251	IN	2003	RRS-10	ABRC	41.8
DuckLkSP38	43.3431	-86.4045	MI	2004	DuckLkSP38	ABRC	37.1
DuckLkSP40	43.3431	-86.4045	MI	2004	DuckLkSP40	ABRC	39.6
KBS_Mac_68	42.405	-85.398	MI	2004	KBS-Mac-68	ABRC	41.3
KBS_Mac_74	42.405	-85.398	MI	2004	KBS-Mac-74	ABRC	37.7
MNF_Che_47	43.5251	-86.1843	MI	2004	MNF-Che-47	ABRC	27.6
MNF_Che_49	43.5251	-86.1843	MI	2004	MNF-Che-49	ABRC	28.5
MNF_Pin_40	43.5356	-86.1788	MI	2004	MNF-Pin-40	ABRC	47.9
MNF_Pot_10	43.595	-86.2657	MI	2004	MNF-Pot-10	ABRC	61.4
MNF_Pot_15	43.595	-86.2657	MI	2004	MNF-Pot-15	ABRC	25.2
MSGA_10	43.2749	-86.0891	MI	2004	MSGA-10	ABRC	41.9
MSGA_12	43.2749	-86.0891	MI	2004	MSGA-12	ABRC	42.8
MSGA_61	43.2749	-86.0891	MI	2004	MSGA-61	ABRC	45.5
MuskSP_68	43.2483	-86.3368	MI	2004	MuskSP-68	ABRC	25.8
MuskSP_83	43.2483	-86.3368	MI	2004	MuskSP-83	ABRC	29.9
Pent_46	43.7623	-86.3929	MI	2004	Pent-46	ABRC	48.3
Pent_7	43.7623	-86.3929	MI	2004	Pent-7	ABRC	55.7
SLSP_67	43.665	-86.496	MI	2004	SLSP-67	ABRC	53.5
SLSP_69	43.665	-86.496	MI	2004	SLSP-69	ABRC	35.5
KNO2_41	41.273	-86.625	IN	2005	KNO2.41	ABRC	44.7
KNO2_54	41.273	-86.625	IN	2005	KNO2.54	ABRC	44
LI_EF_011	40.9064	-73.1493	NY	2005	LI-EF-011	ABRC	68.6
LI_EF_018	40.9064	-73.1493	NY	2005	LI-EF-018	ABRC	39
LI_OF_061	40.7777	-72.9069	NY	2005	LI-OF-061	ABRC	58
LI_RR_096	40.9447	-72.8615	NY	2005	LI-RR-096	ABRC	63.5
LI_RR_097	40.9447	-72.8615	NY	2005	LI-RR-097	ABRC	40.8
LI_SET_019	40.9352	-73.114	NY	2005	LI-SET-019	ABRC	29.9
LI_SET_036	40.9352	-73.114	NY	2005	LI-SET-036	ABRC	41.5
LI_WP_039	40.9076	-73.2089	NY	2005	LI-WP-039	ABRC	104.8
LI_WP_041	40.9076	-73.2089	NY	2005	LI-WP-041	ABRC	76.5
PT1_52	41.3423	-86.7368	IN	2005	PT1.52	ABRC	50.6
PT1_85	41.3423	-86.7368	IN	2005	PT1.85	ABRC	46.1
RMX4_118	42.036	-86.511	MI	2005	RMX4.118	ABRC	41.8
11PNA1_14	42.0945	-86.3253	MI	2006	11PNA1.14	ABRC	47.5
328PNA062	42.0945	-86.3253	MI	2006	328PNA062	ABRC	47.3
627ME_13Y1	42.093	-86.359	MI	2006	n/a	ABRC	53.4
627ME_1MI1	42.093	-86.359	MI	2006	627ME-1MI1	ABRC	57.8
627RMX_1MN4	42.0333	-86.5128	MI	2006	n/a	ABRC	43.6
627RMX_1MN5	42.0333	-86.5128	MI	2006	n/a	ABRC	50.6
BRR107	40.8313	-87.735	IL	2006	BRR107	ABRC	28.5
BRR12	40.8313	-87.735	IL	2006	BRR12	ABRC	43.9
BRR23	40.8313	-87.735	IL	2006	BRR23	ABRC	30.7
BRR4	40.8313	-87.735	IL	2006	BRR4	ABRC	44.7
BRR57	40.8313	-87.735	IL	2006	BRR57	ABRC	28.4
BRR60	40.8313	-87.735	IL	2006	BRR60	ABRC	42.9
KEN	41.767	-72.677	CT	n/a	KEN	ABRC	55.2
LP3413_31	41.6862	-86.8513	IN	n/a	LP3413.31	ABRC	55.9

LP3413_53	41.6862	-86.8513	IN	n/a	LP3413.53	ABRC	51.2
RMX413_85	42.036	-86.511	MI	NA	RMX413.85	ABRC	38

Table S1 (cont'd)

Number of covered **positions (≥3x)** (mapped against HPG1 reference)	Number of covered **positions (≥3x)** (mapped against Col_0 reference)	SNPs vs_ HPG1 reference	Belongs to HPG1	Modern/ Herbarium	Column number in the available genome matrix
105,053,631	99,889,683	142	yes	H	101
100,379,839	95,118,236	123	yes	H	94
103,620,791	98,888,406	167	yes	H	100
107,211,409	102,634,255	161	yes	H	103
106,042,465	100,826,958	151	yes	H	98
103,997,716	98,876,320	153	yes	H	93
107,236,732	102,176,782	181	yes	H	95
106,285,178	101,480,369	163	yes	H	96
106,718,326	102,458,166	169	yes	H	88
105,962,154	100,840,125	145	yes	H	97
106,531,302	101,841,156	153	yes	H	90
106,391,637	101,455,311	192	yes	H	91
90,426,010	89,296,191	n/a	no	H	92
102,894,430	101,298,068	n/a	no	H	99
107,223,283	102,450,446	222	yes	H	89
105,025,845	n/a	138	yes	H	118
100,620,441	n/a	n/a	no	H	121
99,572,736	94,661,828	216	yes	H	102
106,309,854	n/a	176	yes	H	108
102,169,546	n/a	n/a	no	H	104
107,043,540	n/a	161	yes	H	122
107,026,827	n/a	161	yes	H	119
106,893,416	n/a	193	yes	H	110
95,822,372	n/a	109	yes	H	106
101,421,749	n/a	n/a	no	H	116

102,831,697	n/a	n/a	no	H	114
100,778,282	n/a	n/a	no	H	117
106,801,844	n/a	189	yes	H	107
107,044,415	n/a	219	yes	H	109
102,643,436	n/a	n/a	no	H	112
62,673,938	n/a	n/a	no	H	115
106,578,197	n/a	177	yes	H	113
106,158,181	n/a	177	yes	H*	123
106,305,970	n/a	178	yes	H*	105
104,089,205	n/a	200	yes	H*	120
106,464,569	n/a	198	yes	H*	111
108,189,771	105,955,885	243	yes	M	16
108,194,960	105,982,511	240	yes	M	17
108,203,215	106,052,866	269	yes	M	20
108,214,008	106,040,276	288	yes	M	21
108,230,030	106,124,249	251	yes	M	25
108,242,062	106,155,362	254	yes	M	26
108,198,601	105,985,288	226	yes	M	32
108,219,683	106,069,077	231	yes	M	33
108,209,345	106,032,827	207	yes	M	22
108,140,393	105,806,418	221	yes	M	85
108,260,489	106,243,277	294	yes	M	15
108,241,333	106,194,209	238	yes	M	83
108,184,749	105,953,559	232	yes	M	84
108,279,881	106,291,612	234	yes	M	56
108,263,557	106,250,560	235	yes	M	57
108,200,416	106,010,135	237	yes	M	58
108,176,527	105,728,326	237	yes	M	59
108,177,381	105,905,097	243	yes	M	7
108,208,482	105,951,803	228	yes	M	8
108,154,100	105,903,700	230	yes	M	23
108,201,103	106,004,251	288	yes	M	24
108,132,127	105,806,486	261	yes	M	30
108,259,905	106,246,278	259	yes	M	31
108,062,944	105,496,224	186	yes	M	49
108,255,795	106,209,826	299	yes	M	50
108,176,901	105,775,999	237	yes	M	36
107,955,559	105,553,559	226	yes	M	37
108,265,863	106,224,216	290	yes	M	51
107,303,032	106,093,945	n/a	no	M	52
108,155,999	105,921,907	287	yes	M	53
108,106,772	105,906,924	n/a	no	M	54
108,199,679	105,940,666	266	yes	M	55
108,159,739	105,980,721	267	yes	M	70
108,218,762	106,059,867	241	yes	M	71
108,186,632	105,779,717	273	yes	M	76
108,194,281	105,958,738	260	yes	M	77

108,182,789	106,000,003	289	yes	M	86
108,230,553	106,125,861	191	yes	M	87
108,208,144	106,033,465	274	yes	M	80
108,171,751	105,932,415	253	yes	M	18
108,204,654	105,969,244	257	yes	M	19
108,181,390	105,870,424	259	yes	M	27
108,160,645	105,801,702	265	yes	M	28
108,093,393	105,596,885	281	yes	M	60
108,082,202	105,661,610	274	yes	M	61
108,238,775	106,099,919	287	yes	M	62
108,189,553	106,228,588	n/a	no	M	63
108,543,185	107,022,924	n/a	no	M	64
108,191,659	106,019,404	233	yes	M	65
108,227,214	106,032,928	240	yes	M	66
108,210,152	106,077,183	247	yes	M	67
108,063,297	105,588,467	215	yes	M	68
108,099,368	105,721,042	222	yes	M	69
108,227,763	106,099,890	238	yes	M	72
108,220,625	106,144,167	240	yes	M	73
108,238,880	106,143,530	245	yes	M	81
108,160,835	105,899,252	249	yes	M	82
108,209,694	106,063,235	219	yes	M	34
108,212,430	105,903,373	218	yes	M	35
108,267,109	106,250,331	259	yes	M	38
108,244,306	105,898,497	230	yes	M	39
104,897,841	105,729,196	n/a	no	M	40
108,264,679	106,251,487	261	yes	M	41
108,211,310	105,992,095	249	yes	M	42
108,085,297	105,737,781	259	yes	M	43
108,216,592	106,006,605	238	yes	M	44
108,301,282	106,273,259	239	yes	M	45
108,287,248	106,322,146	235	yes	M	46
108,240,431	106,154,252	219	yes	M	74
108,220,150	106,097,633	233	yes	M	75
106,178,554	105,685,651	n/a	no	M	78
108,227,783	106,133,372	276	yes	M	1
108,221,709	106,127,272	223	yes	M	2
107,908,679	106,148,671	n/a	no	M	3
108,252,617	106,173,403	281	yes	M	4
106,799,549	105,789,469	n/a	no	M	5
106,885,430	105,897,441	n/a	no	M	6
108,896,513	107,320,745	n/a	no	M	9
108,190,572	106,031,493	232	yes	M	10
108,095,072	105,726,913	236	yes	M	11
108,180,840	106,033,507	219	yes	M	12
108,093,033	105,630,963	225	yes	M	13
108,281,285	106,199,572	229	yes	M	14
108,233,232	106,158,223	249	yes	M	29
108,244,332	106,190,596	227	yes	M	47

108,157,453	105,994,665	245	yes	M	48
106,816,221	105,483,632	n/a	no	M	79

Table S2. Sample information for Col-0 mutation accumulation lines. Related to Figure 3.

Line	Read depth	Gene-ration	Total	SNPs	Dele-tions	inser-tions	CDS	Non-syn	Syn	Intron	5' UTR	3' UTR
0-4-26	57	3	7	6	1	0	0	0	0	0	0	0
0-8-87	49	3	7	5	0	2	1	1	0	1	0	0
30-109	45	31	31	23	7	1	3	3	0	3	0	0
30-119	45	31	33	26	2	5	1	1	0	1	2	0
30-29	51	31	39	26	10	3	2	1	1	3	0	1
30-39	48	31	28	18	7	3	1	1	0	1	0	1
30-49	50	31	30	23	3	4	4	4	0	0	0	0
30-59	40	31	46	31	8	7	5	2	3	2	0	0
30-69	50	31	26	21	3	2	4	3	1	1	1	1
30-79	50	31	31	25	3	3	6	4	2	2	0	0
30-89	39	31	35	27	5	3	4	3	1	1	1	0
30-99	44	31	37	35	1	1	6	5	1	2	0	2
average (31st)			33.6	25.5	4.9	3.2	3.6	2.7	0.9	1.6	0.4	0.5
stdev (31st)			5.9	4.9	3.0	1.8	1.8	1.4	1.0	1.0	0.7	0.7
Total bases in genome			#########				30753966			17,446,837	4,289,789	2,508,199

Table S3. Mutation rate estimates for different annotations in HPG1 and mutation accumulation lines. Related to Figure 3.

Dataset	Parameter	CDS	Intron	5' UTR	3' UTR	TE	intergeni
MAL	mean	3.776	2.958	3.008	6.431	17.752	9.592
MAL	sem	1.928	1.786	5.258	9.094	7.420	2.628
MAL	lower 95CI.	4.971	4.065	6.267	12.067	22.351	11.221
MAL	upper 95CI.	2.581	1.851	-0.251	0.794	13.153	7.964
HPG1	mean	2.945	2.060	-2.782	-2.137	3.745	4.914
HPG1	sem	0.157	0.210	0.481	0.424	0.849	0.291
HPG1	lower 95CI.	2.637	1.648	-3.725	-2.968	2.080	4.343
HPG1	upper 95CI.	3.254	2.471	-1.838	-1.306	5.411	5.486
IPG1 1.3 years/generatio	mean	3.829	2.678	-3.616	-2.778	4.869	6.388
IPG1 1.3 years/generatio	sem	0.204	0.273	0.626	0.551	1.104	0.379
IPG1 1.3 years/generatio	lower 95CI.	3.428	2.143	-4.843	-3.858	2.704	5.645
IPG1 1.3 years/generatio	upper 95CI.	4.230	3.213	-2.389	-1.697	7.034	7.131

Table 45. Description of phenotypic and climatic variables for association mapping analyses. Related to Figure 5.
Mean and standard deviation across accessions for each phenotypic and climatic variable.
Broad sense heritabilities (H2) calculated from between line and within line (between replicate) variance in ANOVA framework.
Narrow sense heritability (h2) calculated employing linear mixed models and Kinship matrix from mean accession values.

Variable	Description	mean	s.d.	H2	p-value
FT_V0	Time from germination to the first flower opens (days) under 0 days of vernalization	101	4.53	0.009	7.28E-03
FT_V1	Time from germination to the first flower opens (days) under 14 days of vernalization	107	4.12	0.013	6.87E-04
FT_V2	Time from germination to the first flower opens (days) under 28 days of vernalization	102	3.22	0.012	1.04E-03
FT_V3	Time from germination to the first flower opens (days) under 63 days of vernalization	110	1.32	0.010	5.11E-03
B_V0	Time from germination to the first developed bud (days) under 0 days of vernalization	88.8	4	0.013	8.99E-04
B_V1	Time from germination to the first developed bud (days) under 14 days of vernalization	93.9	3.84	0.009	7.45E-03
B_V2	Time from germination to the first developed bud (days) under 28 days of vernalization	89.2	2.13	0.005	6.92E-02
B_V3	Time from germination to the first developed bud (days) under 63 days of vernalization	101	0.45	0.006	5.79E-02
Fecundity	Pixel area of inflorescence (correlation with number of fruits, rho=0.84)	0.0197	0.0042	0.001	3.56E-01
seed_size	Average seed size (mm2)	0.134	0.0053	0.998	3.48E-123
GR_rootLength	Average root growth rate	181	14.9	0.131	4.76E-77
GR_shootArea	Average of shoot area growth rate	2279	253	0.053	2.33E-24
rootLength	Average root length	467	35.8	0.048	2.01E-21
dirEquivalent	Average root direction index. Score for average pixel-by-pixel deviations from growth relative to vector of gravity	0.393	0.0277	0.059	2.62E-28
stdDevXY	Average root linearity coefficient of linear determination; R2 of linear regression line fitted to pixels of primary root skeleton	0.725	0.0429	0.018	4.54E-06
meanRootWidth	Average root width	5.27	0.177	0.038	5.30E-16
rootWidth20	Average width over first interval of the primary root length (0 to 20%) at hypocotyl/root junction	5.75	0.124	0.018	5.11E-06
rootWidth40	Average width over first interval of the primary root length (20 to 40%) at hypocotyl/root junction	5.35	0.19	0.033	3.87E-13

rootWidth60	Average width over first interval of the primary root length (40 to 60%) at hypocotyl/root junction	5.2	0.212	0.039	1.49E-16
rootWidth80	Average width over first interval of the primary root length (60 to 80%) at hypocotyl/root junction	5.11	0.241	0.045	4.67E-20
rootWidth100	Average width over first interval of the primary root length (80 to 100%) at hypocotyl/root junction	4.9	0.222	0.038	4.06E-16
gravitropicDir	Average root angle between root vector and the vertical axis of the picture (assumed vector of gravity) (°)	-7.22	2.56	0.024	7.69E-09
gravitropicScore	Average score for root angle intervals	0.1	0.0457	0.044	2.83E-19
TotLen.EucLen	Average root tortuosity: Total root length divided by Euclidian length	1.1	0.0097	0.009	6.83E-03
GR.TL	Average relative root growth rate: Root growth rate divided by total length at the earlier time point	0.673	0.0796	0.011	1.20E-03
BIO1	Annual Mean Temperature (ºC x 10)	98.1	12.8	NaN	NaN
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	107	7.65	NaN	NaN
BIO3	Isothermality (BIO2/BIO7) (* 100)	28.9	1.8	NaN	NaN
BIO4	Temperature Seasonality (standard deviation *100)	9169	483	NaN	NaN
BIO5	Max Temperature of Warmest Month (ºC x 10)	283	10.1	NaN	NaN
BIO6	Min Temperature of Coldest Month (ºC x 10)	-80.9	18	NaN	NaN
BIO7	Temperature Annual Range (BIO5-BIO6) (ºC x 10)	364	17.5	NaN	NaN
BIO8	Mean Temperature of Wettest Quarter (ºC x 10)	176	55.1	NaN	NaN
BIO9	Mean Temperature of Driest Quarter (ºC x 10)	-7.11	48.7	NaN	NaN
BIO10	Mean Temperature of Warmest Quarter (ºC x 10)	213	10.8	NaN	NaN
BIO11	Mean Temperature of Coldest Quarter (ºC x 10)	-24.1	18.2	NaN	NaN
BIO12	Annual Precipitation (mm)	990	109	NaN	NaN
BIO13	Precipitation of Wettest Month (mm)	103	6.72	NaN	NaN
BIO14	Precipitation of Driest Month (mm)	54.1	16.7	NaN	NaN
BIO15	Precipitation Seasonality (Coefficient of Variation)	17.8	5.51	NaN	NaN
BIO16	Precipitation of Wettest Quarter (mm)	291	19.7	NaN	NaN
BIO17	Precipitation of Driest Quarter (mm)	191	44.8	NaN	NaN
BIO18	Precipitation of Warmest Quarter (mm)	277	25.2	NaN	NaN
BIO19	Precipitation of Coldest Quarter (mm)	197	47	NaN	NaN

Table 56. SNP hits from association mapping. Related to Table 1.

Trait type	Trait	Chromosome	Position	Ancestral	Derived	Effect	Effect standard error	Sample size	p - value	p- value genome control corrected	p- value false discovery rate corrected	p- value permutation corrected	Ancestral allele count	Derived allele count	Allele frequency
pheno.	GR_rootLengtl	1	12638692	C	T	-12.100	3.164	63	######	0.014966	0.00373	0.003	94	9	0.0
pheno.	GR_shootArea	1	12638692	C	T	######	#####	63	######	0.018191	0.00049	0.000999	94	9	0.0
pheno.	GR_rootLengtl	1	13652509	C	A	-12.100	3.164	63	######	0.014966	0.00373	0.003	88	9	0.0
pheno.	GR_shootArea	1	13652509	C	A	######	#####	63	######	0.018191	0.00049	0.000999	88	9	0.0
pheno.	dirEquivalent	1	19024876	C	T	-0.014	0.004	63	######	0.040944	0.00518	0.018	81	19	0
pheno.	GR_shootArea	1	20324050	G	A	######	#####	63	######	0.018191	0.00049	0.000999	94	9	0.0
pheno.	GR_rootLengtl	1	20324050	G	A	-12.100	3.164	63	######	0.014966	0.00373	0.003	94	9	0.0
pheno.	dirEquivalent	1	26052913	A	T	-0.014	0.004	63	######	0.040944	0.00518	0.018	84	19	0.1
pheno.	GR_shootArea	1	29696198	G	A	######	#####	63	######	0.050091	0.00958	0.016	70	27	0.2
pheno.	GR_rootLengtl	1	30015381	T	A	-12.100	3.164	63	######	0.014966	0.00373	0.003	94	9	0.0
pheno.	GR_shootArea	1	30015381	T	A	######	#####	63	######	0.018191	0.00049	0.000999	94	9	0.0
pheno.	GR_rootLengtl	1	30143319	G	A	-12.100	3.164	63	######	0.014966	0.00373	0.003	93	9	0.0
pheno.	GR_shootArea	1	30143319	G	A	######	#####	63	######	0.018191	0.00049	0.000999	93	9	0.0
pheno.	dirEquivalent	1	958948	G	T	-0.014	0.004	63	######	0.040944	0.00518	0.018	83	19	0.1
pheno.	gravitropicScor	1	9925177	C	T	0.033	0.010	63	######	0.016078	0.06509	0.016	95	8	0.0
pheno.	dirEquivalent	2	10369545	T	C	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	dirEquivalent	2	10495275	A	C	-0.016	0.004	63	######	0.022915	0.00322	0.006	82	20	0.1
pheno.	dirEquivalent	2	1093203	C	T	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	dirEquivalent	2	11346211	C	A	-0.014	0.004	63	######	0.040944	0.00518	0.018	83	19	0.1
pheno.	dirEquivalent	2	12415084	T	A	-0.014	0.004	63	######	0.040944	0.00518	0.018	83	19	0.1
pheno.	dirEquivalent	2	12876361	A	C	-0.015	0.004	63	######	0.025674	0.00406	0.006	76	27	0.2
pheno.	gravitropicScor	2	12876361	A	C	-0.021	0.006	63	######	0.020114	0.06509	0.027	76	27	0.2

pheno.	dirEquivalent	2	15278350	A	G	-0.014	0.004	63	######	0.040944	0.00518	0.018	83	19	0.1
pheno.	GR_shootArea	2	16039488	T	G	######	#####	63	######	0.018191	0.00049	0.000999	94	9	0.0
pheno.	GR_rootLengtl	2	16039488	T	G	-12.100	3.164	63	######	0.014966	0.00373	0.003	94	9	0.0
pheno.	GR_rootLengtl	2	16247290	G	T	-12.100	3.164	63	######	0.014966	0.00373	0.003	93	9	0.0
pheno.	GR_shootArea	2	16247290	G	T	######	#####	63	######	0.018191	0.00049	0.000999	93	9	0.0
pheno.	dirEquivalent	2	16333662	G	A	-0.014	0.004	63	######	0.040944	0.00518	0.018	83	19	0.1
pheno.	dirEquivalent	2	2176891	T	C	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	GR_rootLengtl	2	3174832	T	A	6.340	1.869	63	######	0.030786	0.01703	0.017	48	54	0.5
pheno.	dirEquivalent	2	358395	C	T	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	TotLen.EucLer	2	5285907	C	A	-0.006	0.002	63	######	0.015312	0.0241	0.037	83	16	0.1
pheno.	dirEquivalent	2	5285907	C	A	-0.019	0.005	63	######	0.01317	0.00322	0.000999	83	16	0.1
pheno.	dirEquivalent	2	585918	C	T	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	dirEquivalent	2	6034545	C	T	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	dirEquivalent	2	7047529	G	T	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	dirEquivalent	2	7186220	C	T	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	GR_rootLengtl	3	11259214	A	T	-12.100	3.164	63	######	0.014966	0.00373	0.003	93	9	0.0
pheno.	GR_shootArea	3	11259214	A	T	######	#####	63	######	0.018191	0.00049	0.000999	93	9	0.0
pheno.	GR_rootLengtl	3	15050751	G	A	-12.100	3.164	63	######	0.014966	0.00373	0.003	91	11	0.1
pheno.	GR_shootArea	3	15050751	G	A	######	#####	63	######	0.018191	0.00049	0.000999	91	11	0.1
pheno.	dirEquivalent	3	17164638	C	A	-0.014	0.004	63	######	0.040944	0.00518	0.018	81	19	(
pheno.	dirEquivalent	3	2500258	C	A	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	dirEquivalent	3	3154804	C	T	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	dirEquivalent	3	3629794	C	T	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	dirEquivalent	3	4269626	G	T	-0.016	0.004	63	######	0.022915	0.00322	0.006	83	20	0.1
pheno.	GR_shootArea	3	8873116	C	T	######	#####	63	######	0.018191	0.00049	0.000999	93	10	0.0
pheno.	GR_rootLengtl	3	8873116	C	T	-12.100	3.164	63	######	0.014966	0.00373	0.003	93	10	0.0
pheno.	dirEquivalent	4	11948961	T	A	-0.014	0.004	63	######	0.038281	0.00518	0.016	81	20	0.
pheno.	dirEquivalent	4	12365323	C	T	-0.014	0.004	63	######	0.038281	0.00518	0.016	82	21	0.2
pheno.	dirEquivalent	4	15646341	C	A	-0.014	0.004	63	######	0.038281	0.00518	0.016	81	21	0.2
pheno.	dirEquivalent	4	15845001	A	T	-0.014	0.004	63	######	0.038281	0.00518	0.016	83	20	0.1
pheno.	dirEquivalent	4	18249171	T	A	-0.014	0.004	63	######	0.038281	0.00518	0.016	53	20	0.
pheno.	dirEquivalent	4	3355152	C	G	-0.014	0.004	63	######	0.038281	0.00518	0.016	82	21	0.2
pheno.	dirEquivalent	4	3355946	G	C	-0.014	0.004	63	######	0.038281	0.00518	0.016	82	21	0.2
pheno.	dirEquivalent	4	4228138	A	G	-0.014	0.004	63	######	0.038281	0.00518	0.016	82	20	0.1
pheno.	dirEquivalent	4	9046942	G	C	-0.014	0.004	63	######	0.038281	0.00518	0.016	82	21	0.2
pheno.	dirEquivalent	5	11090365	T	A	-0.014	0.004	63	######	0.040944	0.00518	0.018	83	19	0.1
pheno.	dirEquivalent	5	12312975	C	G	-0.014	0.004	63	######	0.040944	0.00518	0.018	84	19	0.1
pheno.	dirEquivalent	5	12358159	C	T	-0.014	0.004	63	######	0.040944	0.00518	0.018	83	19	0.1
pheno.	dirEquivalent	5	12409027	G	A	-0.014	0.004	63	######	0.040944	0.00518	0.018	84	19	0.1
pheno.	GR_rootLengtl	5	16024197	A	T	-12.100	3.164	63	######	0.014966	0.00373	0.003	92	10	0.

pheno.	GR_shootArea	5	16024197	A	T	######	#####	63	######	0.018191	0.00049	0.000999	92	10	0.
pheno.	GR_shootArea	5	16109431	G	A	######	#####	63	######	0.018191	0.00049	0.000999	10	64	0.8
pheno.	GR_rootLengtl	5	16109431	G	A	-12.100	3.164	63	######	0.014966	0.00373	0.003	10	64	0.8
pheno.	dirEquivalent	5	19099082	G	C	-0.014	0.004	63	######	0.040944	0.00518	0.018	83	19	0.1
pheno.	GR_rootLengtl	5	20388107	A	T	-10.700	3.164	63	######	0.030731	0.01703	0.017	82	9	0.0
pheno.	dirEquivalent	5	4797923	A	T	-0.014	0.004	63	######	0.040944	0.00518	0.018	82	19	0.1
pheno.	dirEquivalent	5	4797976	G	A	-0.014	0.004	63	######	0.040944	0.00518	0.018	55	19	0.2
pheno.	dirEquivalent	5	4798526	A	G	-0.014	0.004	63	######	0.040944	0.00518	0.018	37	19	0.3
pheno.	gravitropicScor	5	6508329	A	G	-0.020	0.006	63	######	0.013618	0.06509	0.008	67	36	0.3
climate	bio18	1	10187610	T	C	6.830	1.987	99	######	0.030738	0.01236	0.047	82	20	0.1
climate	bio18	1	13904611	C	T	6.570	1.756	90	######	0.018766	0.01236	0.016	72	20	0.2
climate	bio18	1	13994958	G	A	6.830	1.987	99	######	0.030738	0.01236	0.047	82	20	0.1
climate	bio18	1	17408807	C	T	6.830	1.987	99	######	0.030738	0.01236	0.047	82	20	0.1
climate	bio18	1	23648407	A	C	6.830	1.987	99	######	0.030738	0.01236	0.047	82	20	0.1
climate	bio16	1	29696198	G	A	5.250	1.377	94	######	0.004745	0.0632	0.016	70	27	0.2
climate	bio18	1	29696198	G	A	6.340	1.569	94	######	0.011173	0.01236	0.004	70	27	0.2
climate	bio13	2	14417366	A	G	3.990	0.959	64	######	0.000481	0.01466	0.004	60	5	0.0
climate	bio8	3	11873293	A	G	37.800	8.736	65	######	0.000113	0.00694	0.006	4	62	0.9
climate	bio18	4	12365323	C	T	6.850	1.944	100	######	0.026847	0.01236	0.035	82	21	0.2
climate	bio18	4	15646341	C	A	6.720	1.936	99	######	0.029136	0.01236	0.042	81	21	0.2
climate	bio11	4	1732480	T	A	-5.550	1.564	79	######	0.019162	0.01951	0.045	75	5	0.0
climate	bio4	4	1732480	T	A	224.000	#####	79	######	0.024548	0.01283	0.044	75	5	0.0
climate	bio18	4	18249171	T	A	6.910	2.005	71	######	0.030265	0.01236	0.047	53	20	0.
climate	bio18	4	279210	T	G	6.830	1.987	99	######	0.030738	0.01236	0.047	82	20	0.1
climate	bio18	4	3355152	C	G	6.850	1.944	100	######	0.026847	0.01236	0.035	82	21	0.2
climate	bio18	4	3355946	G	C	6.850	1.944	100	######	0.026847	0.01236	0.035	82	21	0.2
climate	bio18	4	4228138	A	G	6.830	1.987	99	######	0.030738	0.01236	0.047	82	20	0.1
climate	bio18	4	9046942	G	C	6.850	1.944	100	######	0.026847	0.01236	0.035	82	21	0.2
climate	bio18	5	4245213	A	T	6.830	1.987	99	######	0.030738	0.01236	0.047	82	20	0.1
climate	bio18	5	4500202	G	A	6.830	1.987	99	######	0.030738	0.01236	0.047	82	20	0.1

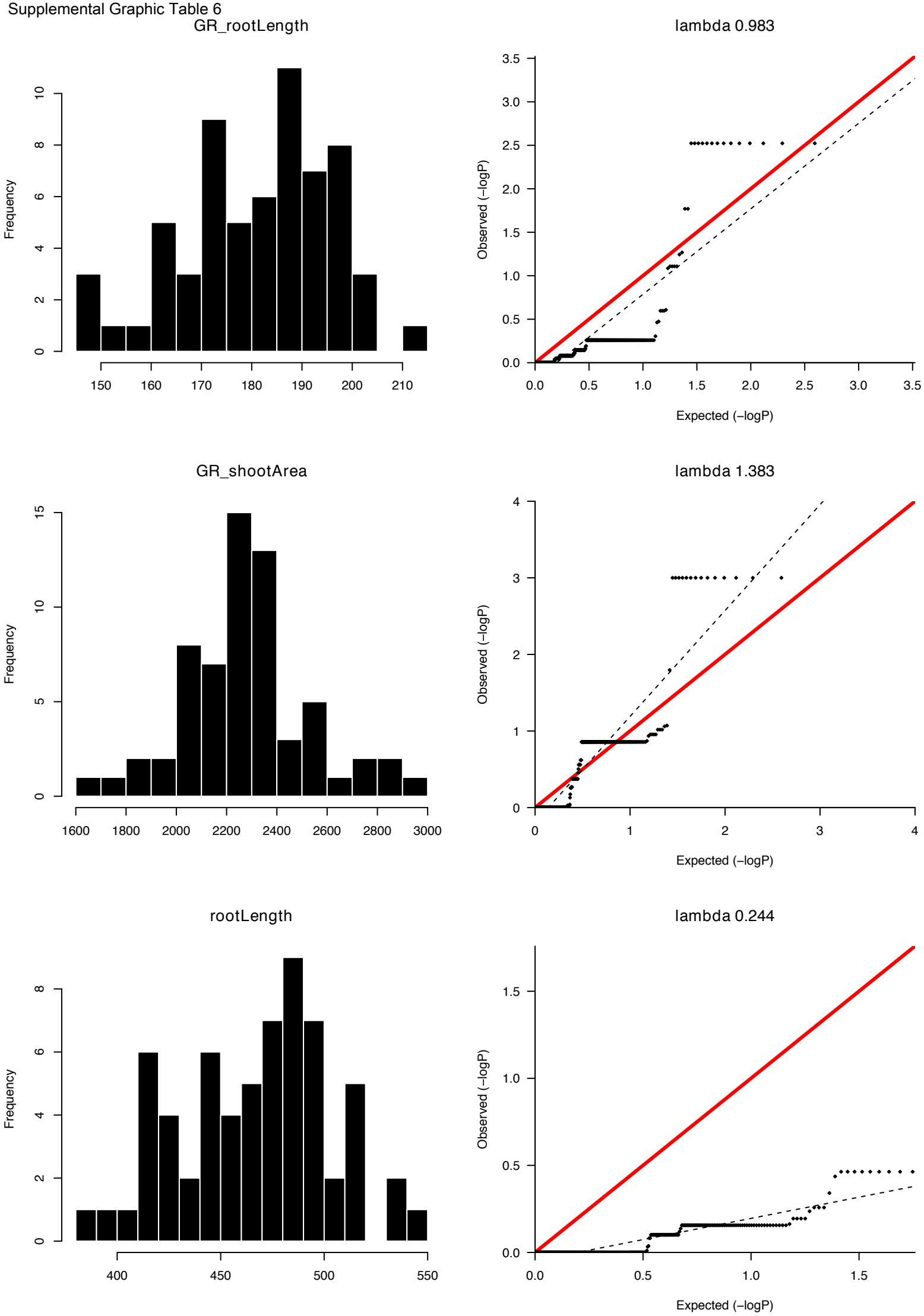
Table S6 (cont'd)

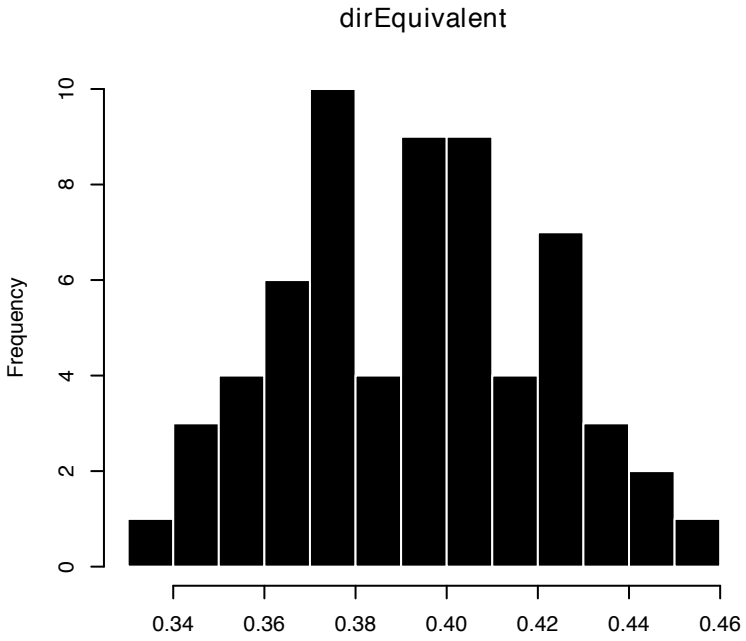
Longitude	Latitude	Substitution type	AA change	Gene	Biochemical effect (Grantham score)	Significant under permutation	Significant under permutation & Bonferroni trait correction	Significant under Bonferroni traits & SNPs correction
40.87013144	-81.29801711	interg.				✓	✓	✓
40.87013144	-81.29801711	interg.				✓	✓	✓
40.861787	-82.91865044	interg.				✓	✓	✓
40.861787	-82.91865044	interg.				✓	✓	✓
41.66797605	-85.25108189	interg.				✓		
40.861787	-82.91865044	intron		AT1G54440		✓	✓	✓
40.861787	-82.91865044	intron		AT1G54440		✓	✓	✓
41.66797605	-85.25108189	interg.				✓		
41.46632759	-84.86486874	interg.				✓		
40.861787	-82.91865044	interg.				✓	✓	✓
40.861787	-82.91865044	interg.				✓	✓	✓
40.861787	-82.91865044	interg.				✓	✓	✓
40.861787	-82.91865044	interg.				✓	✓	✓
41.66797605	-85.25108189	nonsyn	A->P	AT1G03810	27	✓		
40.86663538	-82.27759425	interg.				✓		
41.67782725	-85.3208278	interg.				✓		✓
41.67782725	-85.3208278	intron		AT2G24680		✓		✓
41.67782725	-85.3208278	interg.				✓		✓
41.66797605	-85.25108189	interg.				✓		
41.66797605	-85.25108189	intron		AT2G28900		✓		
41.67869644	-84.57000741	interg.				✓		
41.67869644	-84.57000741	interg.				✓		

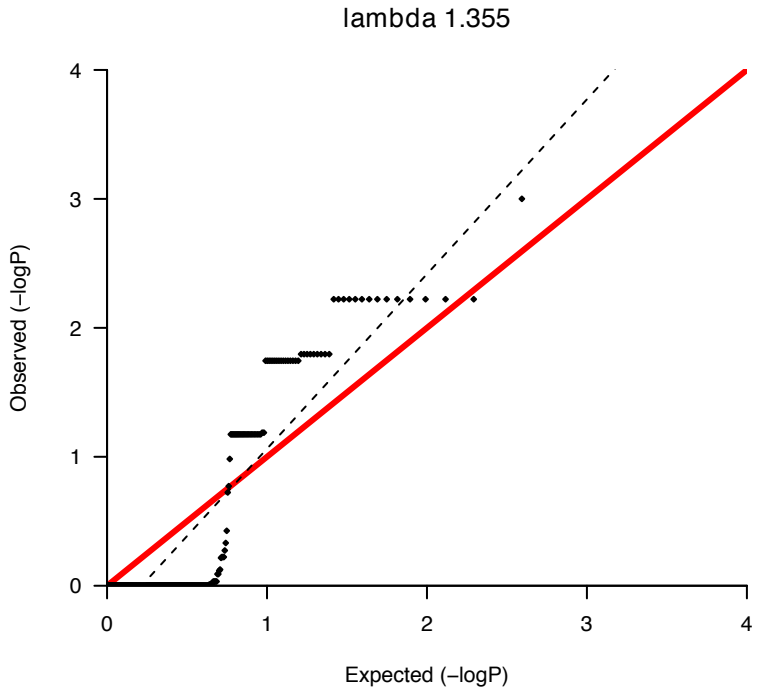
41.66797605	-85.25108189	interg.				✓		
40.861787	-82.91865044	3' UTR		AT2G38290		✓	✓	✓
40.861787	-82.91865044	3' UTR		AT2G38290		✓	✓	✓
40.861787	-82.91865044	nonsyn	A->G	AT2G3891(	60	✓	✓	✓
40.861787	-82.91865044	nonsyn	A->G	AT2G3891(	60	✓	✓	✓
41.66797605	-85.25108189	nonsyn	A->G	AT2G3916(	60	✓		
41.67782725	-85.3208278	interg.				✓		✓
41.32965606	-84.2807173	interg.				✓		
41.67782725	-85.3208278	syn	V->V	AT2G01820		✓		✓
41.54135281	-85.0139785	interg.				✓	✓	✓
41.54135281	-85.0139785	interg.				✓	✓	✓
41.67782725	-85.3208278	syn	G->G	AT2G02220		✓		✓
41.67782725	-85.3208278	syn	S->S	AT2G14247		✓		✓
41.67782725	-85.3208278	nonsyn	P->A	AT2G1627(	27	✓		✓
41.67782725	-85.3208278	intron		AT2G16580		✓		✓
40.861787	-82.91865044	interg.				✓	✓	✓
40.861787	-82.91865044	interg.				✓	✓	✓
40.24117118	-82.47677464	interg.				✓	✓	✓
40.24117118	-82.47677464	interg.				✓	✓	✓
41.66797605	-85.25108189	interg.				✓		
41.67782725	-85.3208278	syn	K->K	AT3G07830		✓		✓
41.67782725	-85.3208278	interg.				✓		✓
41.67782725	-85.3208278	intron		AT3G11530		✓		✓
41.67782725	-85.3208278	5' UTR		AT3G13229		✓		✓
40.8662483	-81.9417154	interg.				✓	✓	✓
40.8662483	-81.9417154	interg.				✓	✓	✓
41.68799415	-85.3222627	interg.				✓		
41.67822595	-85.39370743	interg.				✓		
41.67822595	-85.39370743	syn	E->E	AT4G32410		✓		
41.7692981	-85.9290451	3' UTR		AT4G32840		✓		
41.7692981	-85.9290451	interg.				✓		
41.67822595	-85.39370743	interg.				✓		
41.67822595	-85.39370743	interg.				✓		
41.66888725	-85.3310928	TE		AT4G07440		✓		
41.67822595	-85.39370743	nonsyn	H->Q	AT4G1596(	24	✓		
41.66797605	-85.25108189	TE		AT5G29037		✓		
41.66797605	-85.25108189	TE		AT5G32630		✓		
41.66797605	-85.25108189	TE		AT5G32825		✓		
41.66797605	-85.25108189	interg.				✓		
40.8662483	-81.9417154	intron		AT5G40020		✓	✓	✓

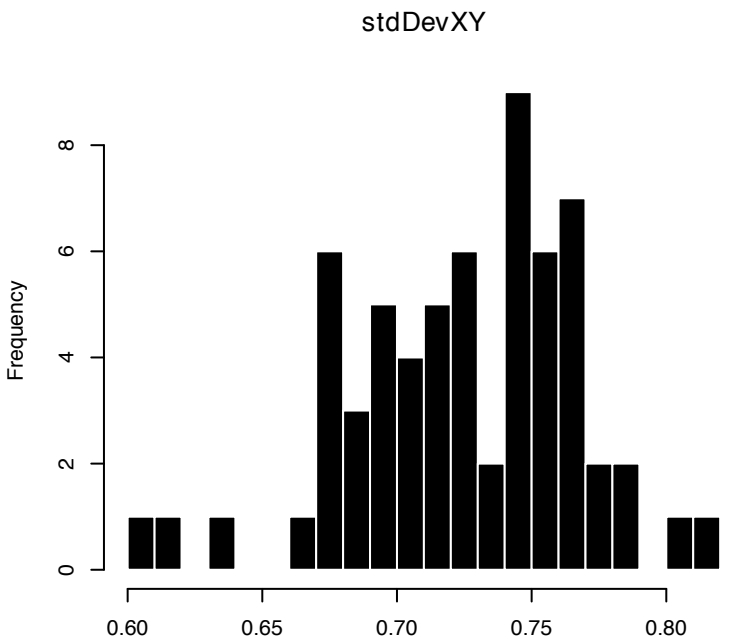
40.8662483	-81.9417154	intron		AT5G40020		✓	✓	✓
42.16498594	-84.42275313	interg.				✓	✓	✓
42.16498594	-84.42275313	interg.				✓	✓	✓
41.66797605	-85.25108189	interg.				✓		
41.03714444	-86.63891111	interg.				✓		
41.66797605	-85.25108189	TE		AT5G14830		✓		
41.66797605	-85.25108189	TE		AT5G14830		✓		
41.66797605	-85.25108189	interg.				✓		
41.97279292	-85.01427656	nonsyn	C->W	AT5G1933(	215	✓		
41.66888725	-85.3310928	interg.				✓		
41.66888725	-85.3310928	interg.				✓		
41.66888725	-85.3310928	TE		AT1G36933		✓		
41.66888725	-85.3310928	interg.				✓		
41.66888725	-85.3310928	nonsyn	Y->S	AT1G6374(	144	✓		
41.46632759	-84.86486874	interg.				✓		
41.46632759	-84.86486874	interg.				✓		
39.478949	-77.9063534	interg.				✓		
41.79308313	-83.68315273	TE		AT3G30219		✓		
41.67822595	-85.39370743	interg.				✓		
41.67822595	-85.39370743	syn	E->E	AT4G32410		✓		
41.04504	-87.5046	interg.				✓		
41.04504	-87.5046	interg.				✓		
41.7692981	-85.9290451	interg.				✓		
41.66888725	-85.3310928	interg.				✓		
41.67822595	-85.39370743	interg.				✓		
41.67822595	-85.39370743	interg.				✓		
41.66888725	-85.3310928	TE		AT4G07440		✓		
41.67822595	-85.39370743	nonsyn	H->Q	AT4G1596(	24	✓		
41.66888725	-85.3310928	syn	I->I	AT5G13260		✓		
41.66888725	-85.3310928	nonsyn	A->G	AT5G1395(	60	✓		

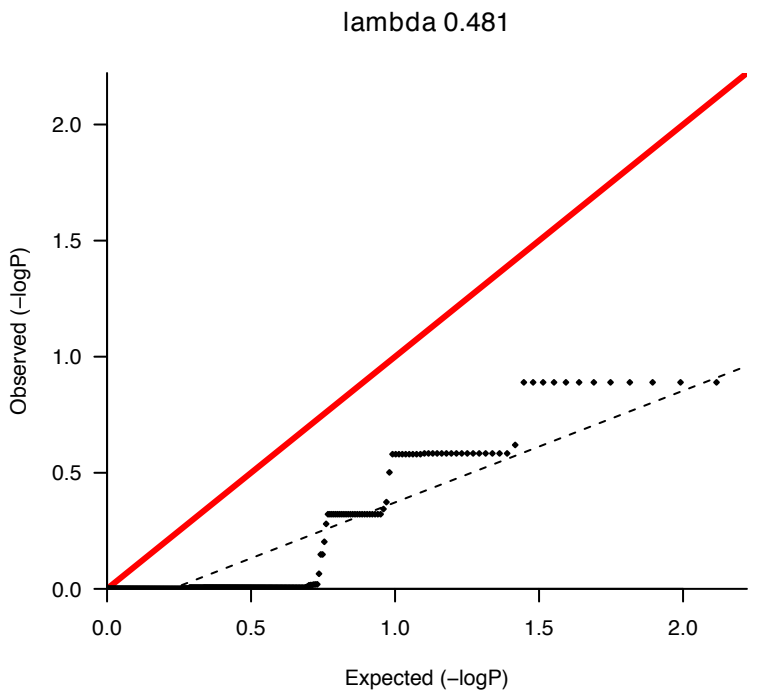
Supplemental Graphic Table 6

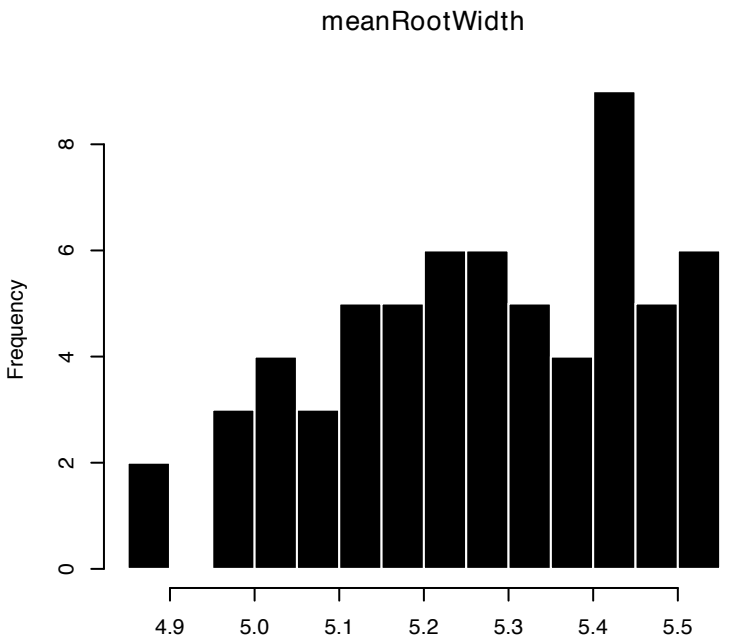


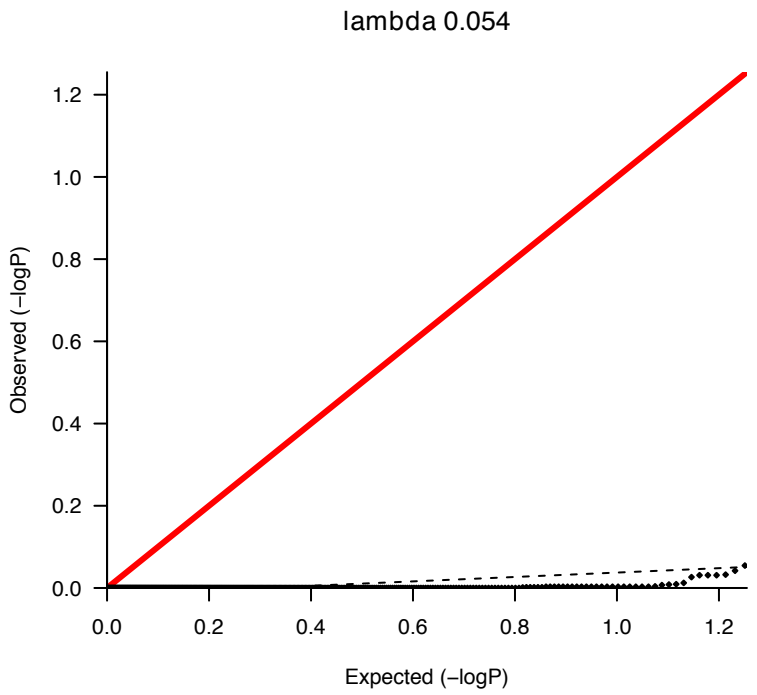

dirEquivalent
Frequency
0
2
4
6
8
10
0.34
0.36
0.38
0.40
0.42
0.44
0.46

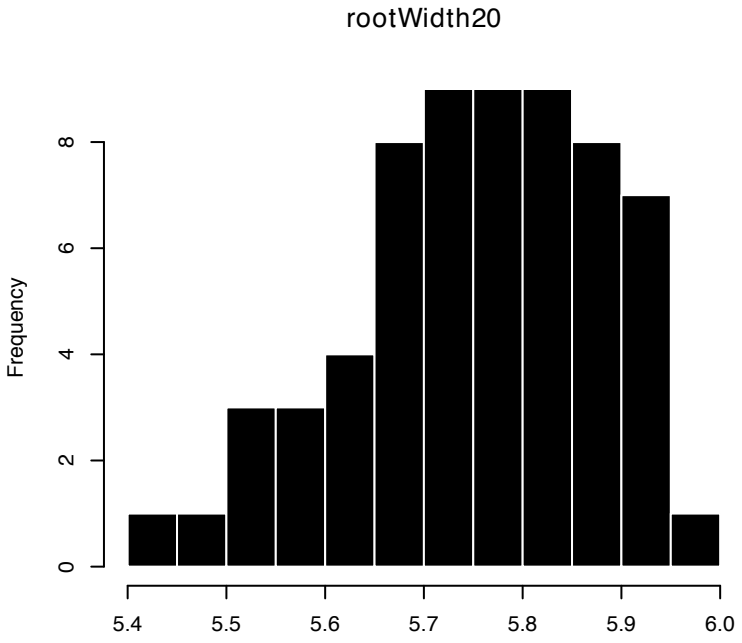

lambda 1.355
Observed (−logP)
Expected (−logP)
0
1
2
3
4

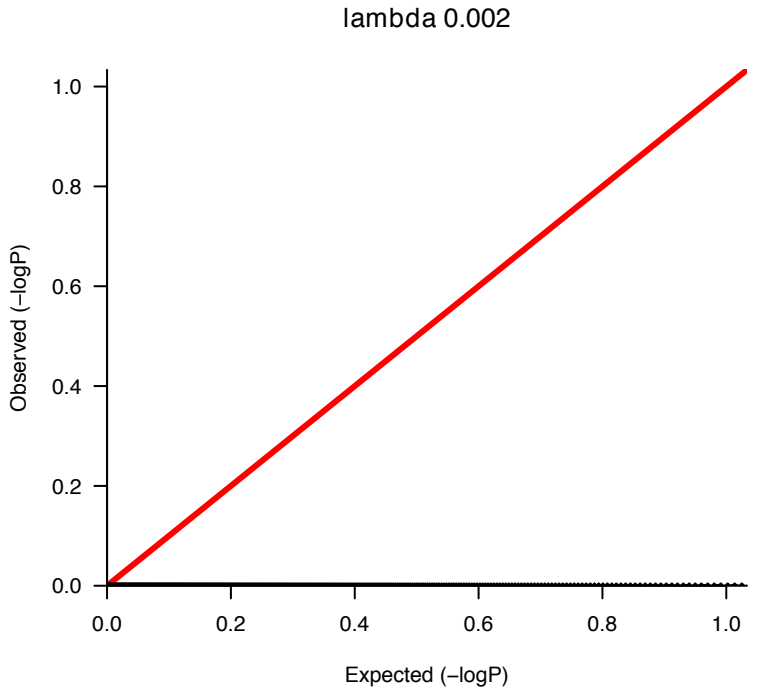

stdDevXY
Frequency
0
2
4
6
8
0.60
0.65
0.70
0.75
0.80

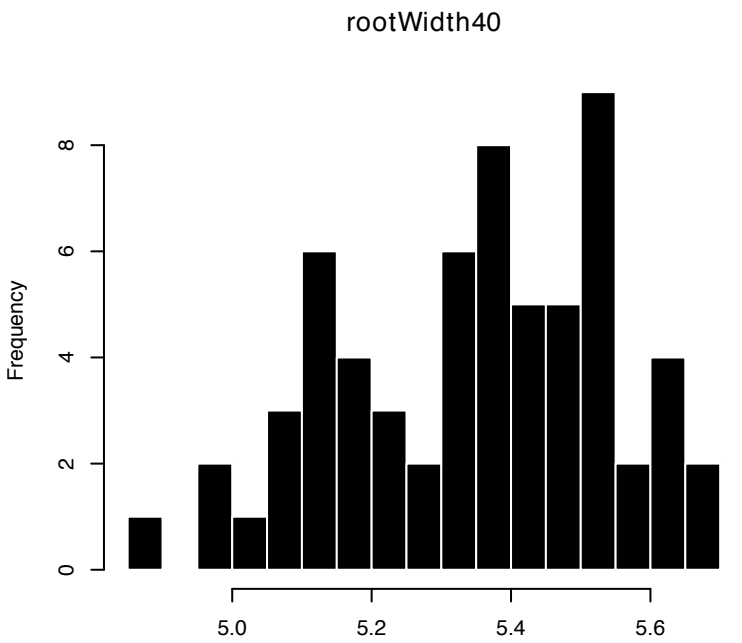

lambda 0.481
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

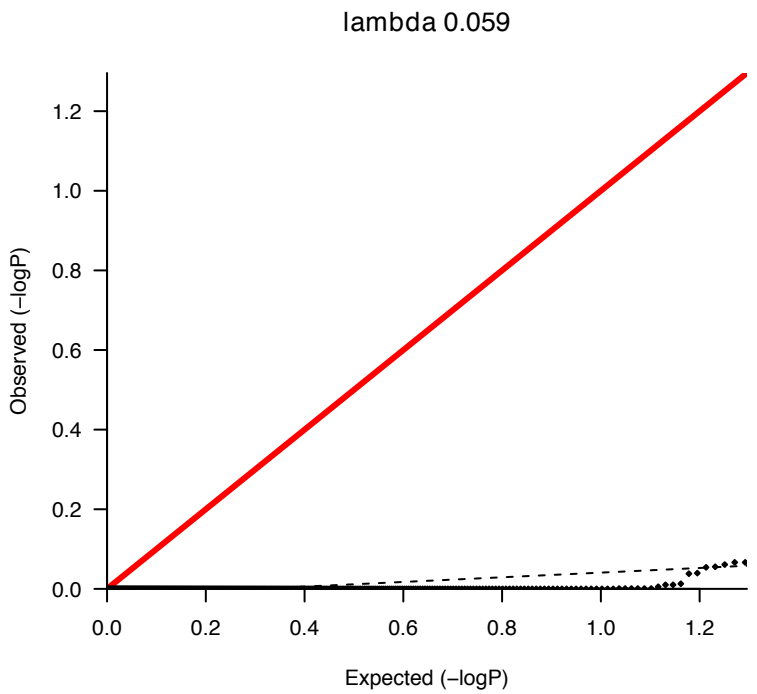

meanRootWidth
Frequency
0
2
4
6
8
4.9
5.0
5.1
5.2
5.3
5.4
5.5

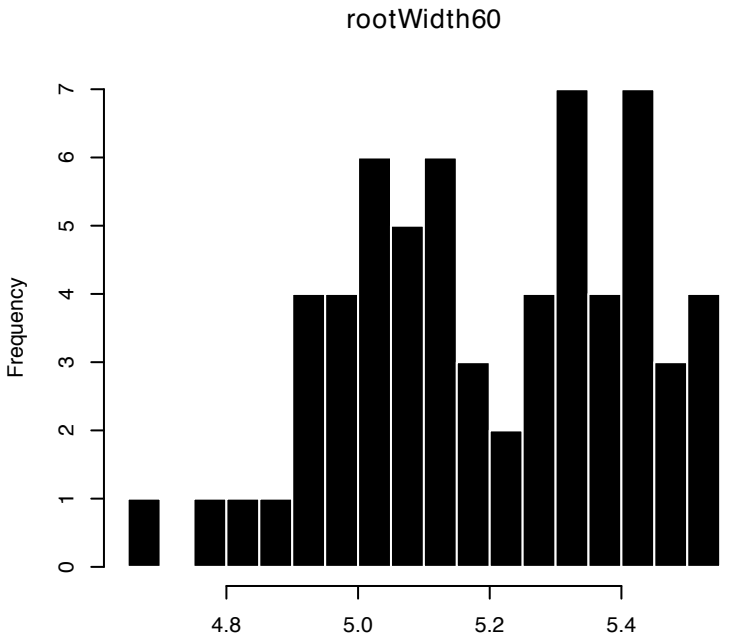

lambda 0.054
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2

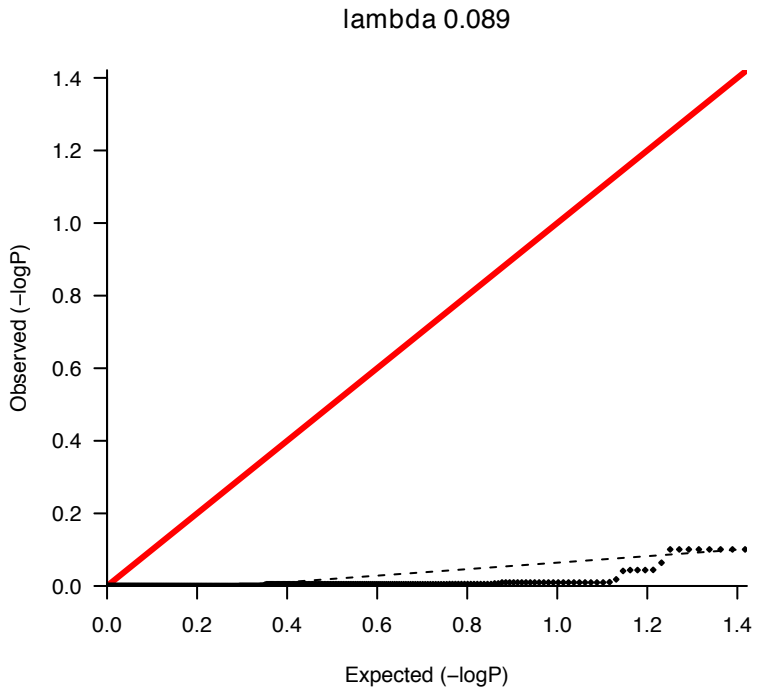

rootWidth20
Frequency
0
2
4
6
8
5.4
5.5
5.6
5.7
5.8
5.9
6.0

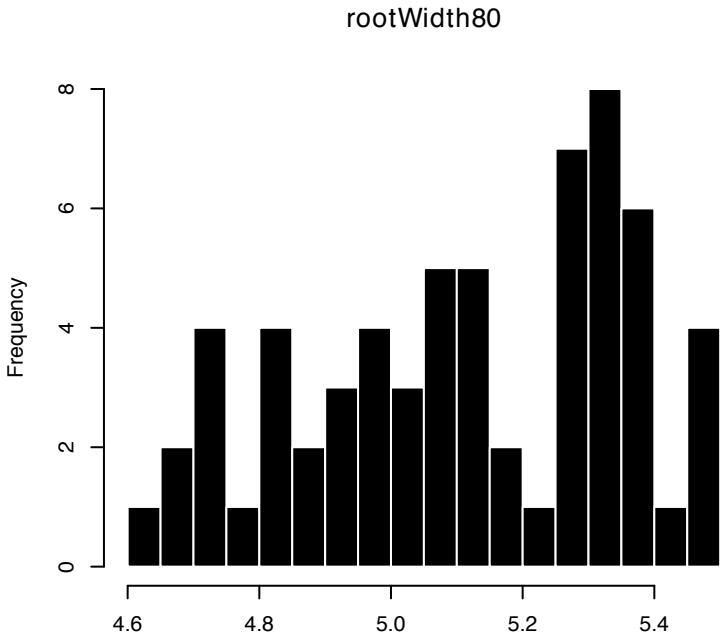

lambda 0.002
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0

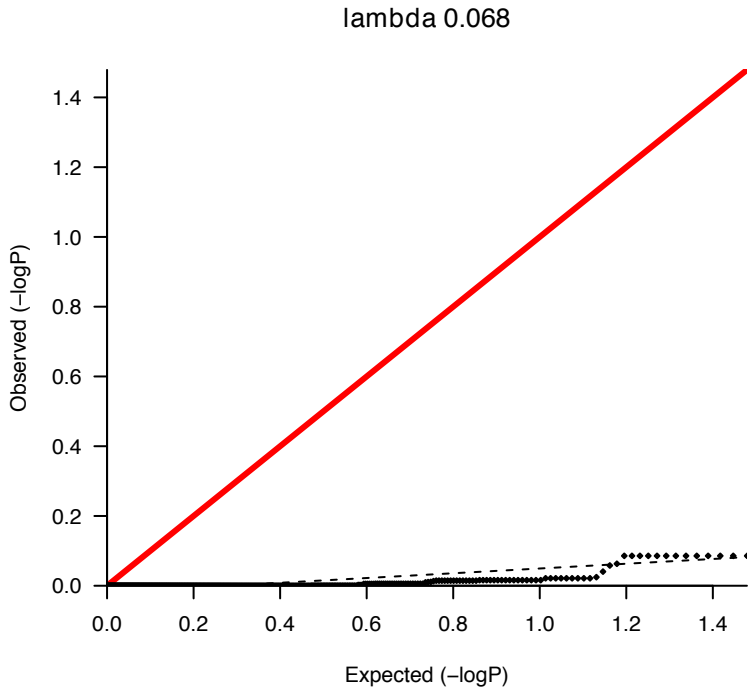

rootWidth40
Frequency
0
2
4
6
8
5.0
5.2
5.4
5.6

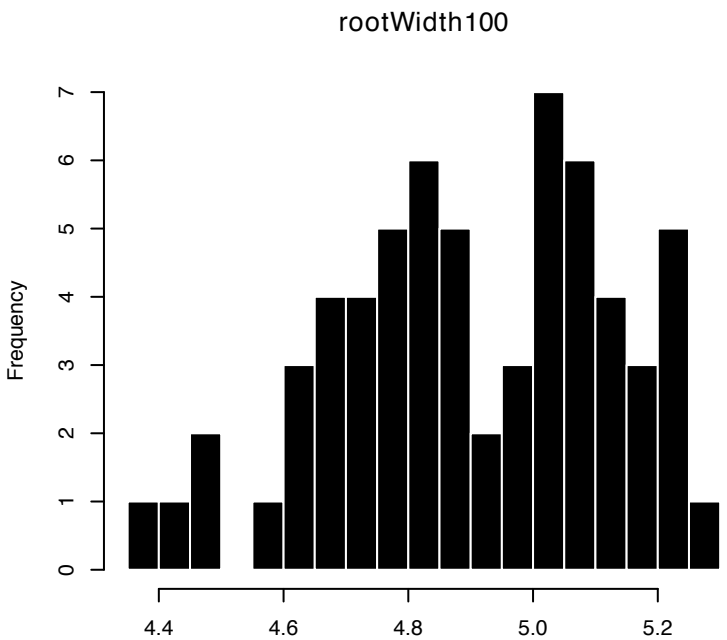

lambda 0.059
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2

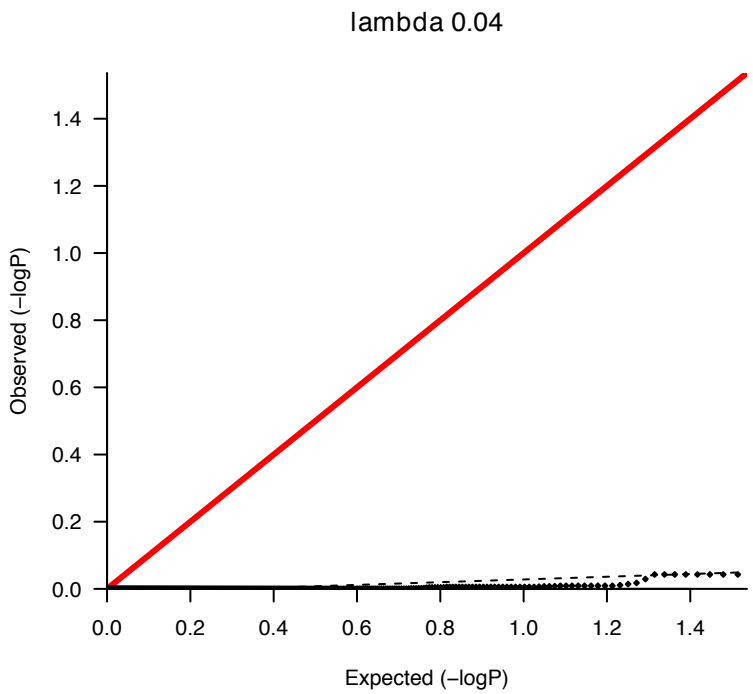

rootWidth60
Frequency
0
1
2
3
4
5
6
7
4.8
5.0
5.2
5.4

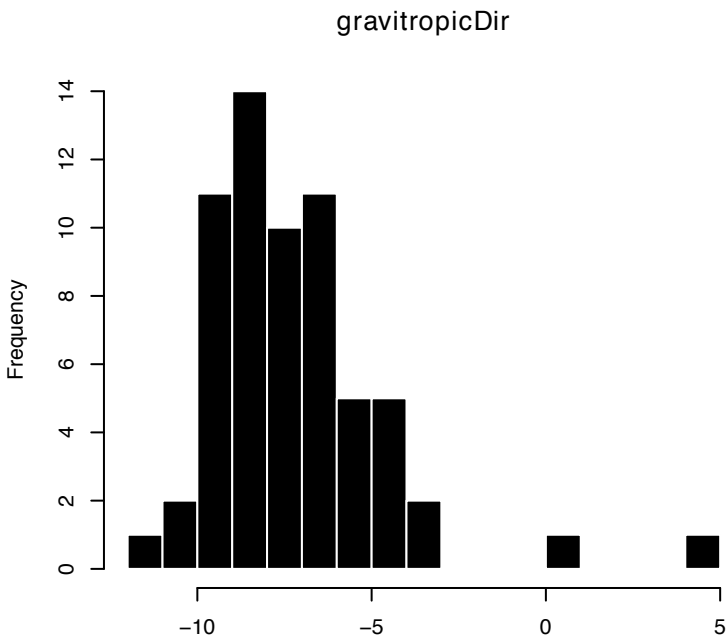

lambda 0.089
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

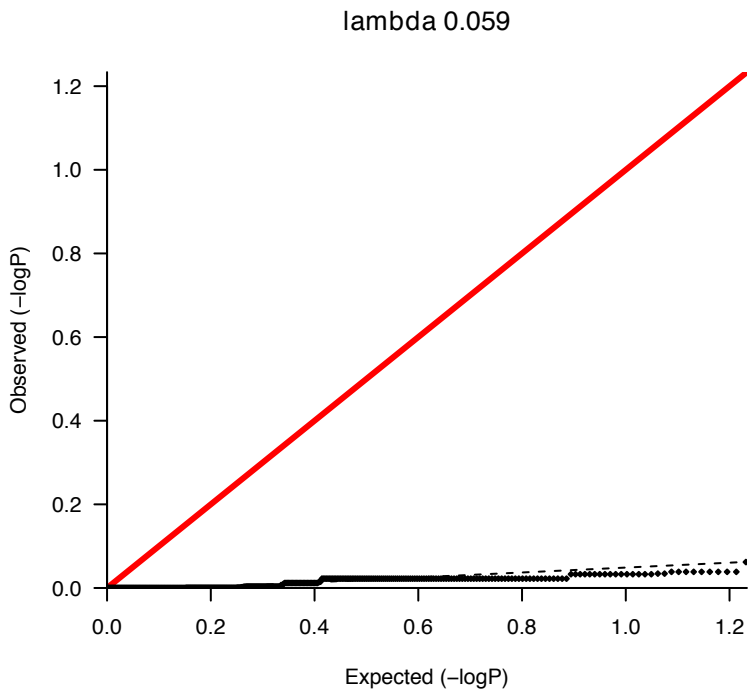

rootWidth80
Frequency
0
2
4
6
8
4.6
4.8
5.0
5.2
5.4

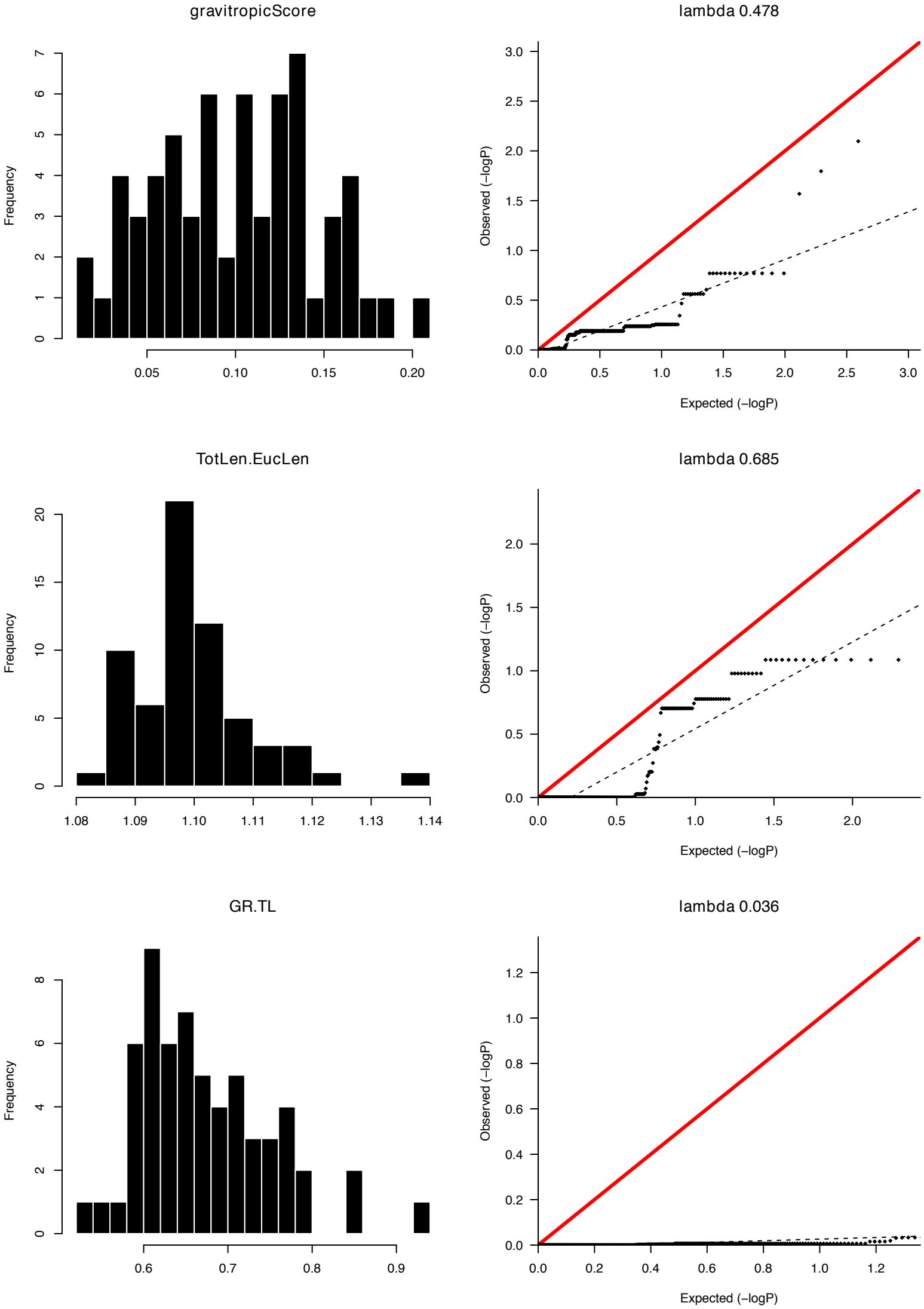

lambda 0.068
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

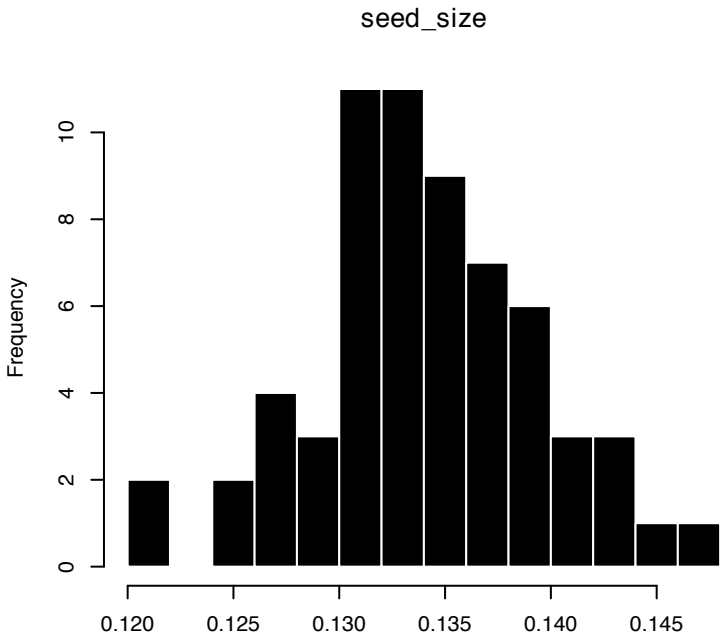

rootWidth100
Frequency
0
1
2
3
4
5
6
7
4.4
4.6
4.8
5.0
5.2

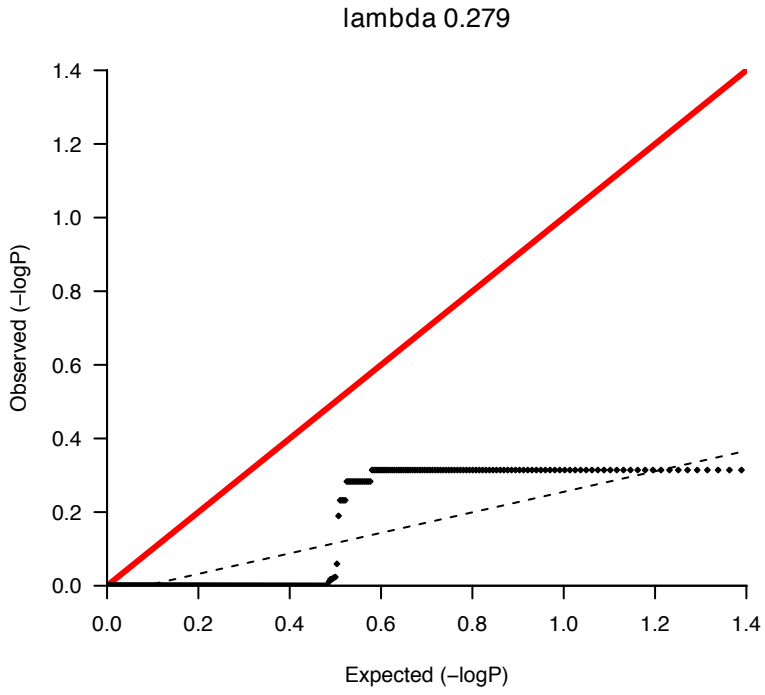

lambda 0.04
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

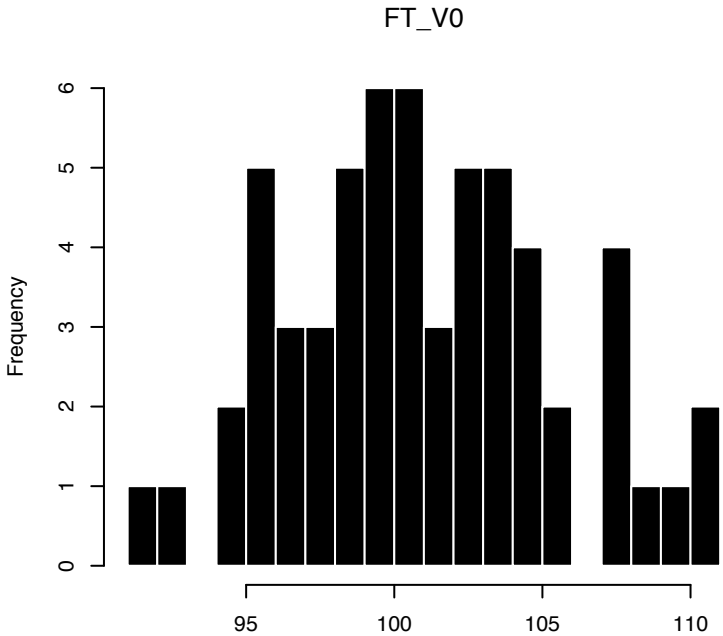

gravitropicDir
Frequency
0
2
4
6
8
10
12
14
−10
−5
0
5

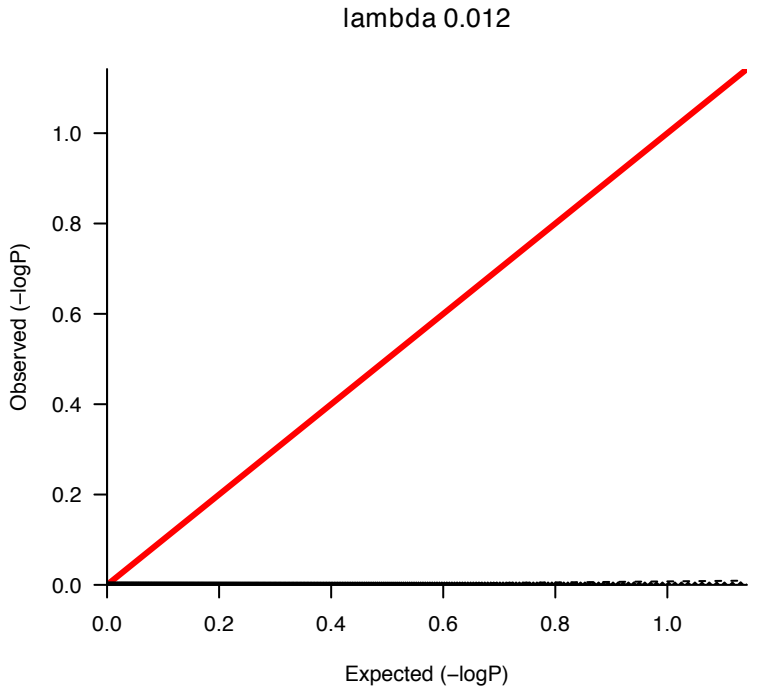

lambda 0.059
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2

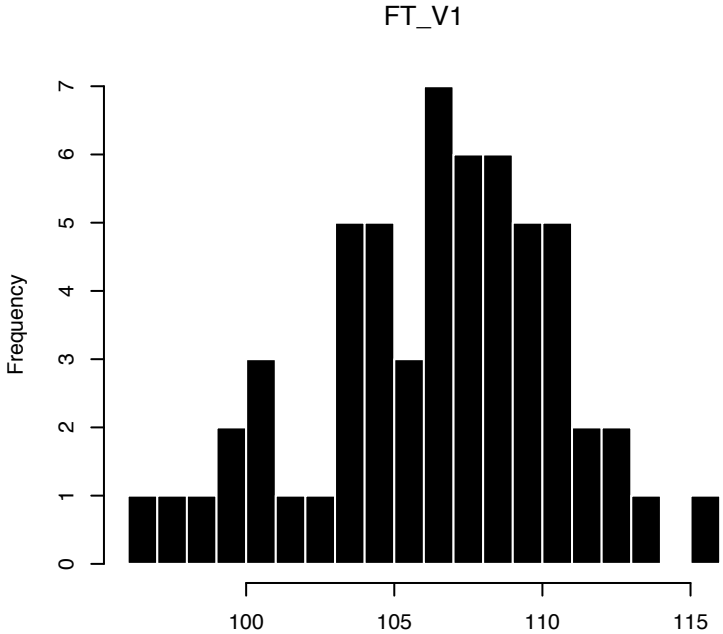

gravitropicScore
Frequency
0
1
2
3
4
5
6
7
0.05
0.10
0.15
0.20
lambda 0.478
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
TotLen.EucLen
Frequency
0
5
10
15
20
1.08
1.09
1.10
1.11
1.12
1.13
1.14
lambda 0.685
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
GR.TL
Frequency
0
2
4
6
8
0.6
0.7
0.8
0.9
lambda 0.036
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2

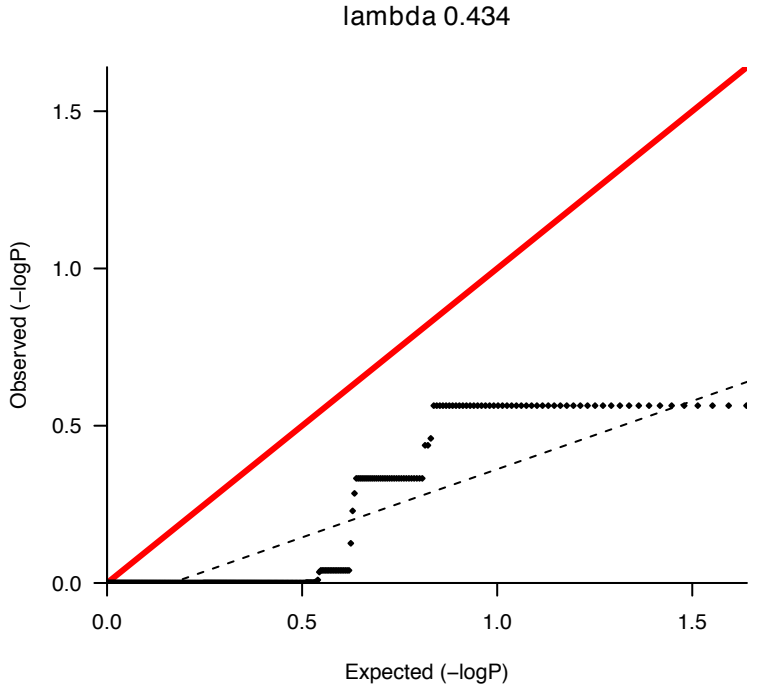

seed_size
Frequency
0
2
4
6
8
10
0.120
0.125
0.130
0.135
0.140
0.145

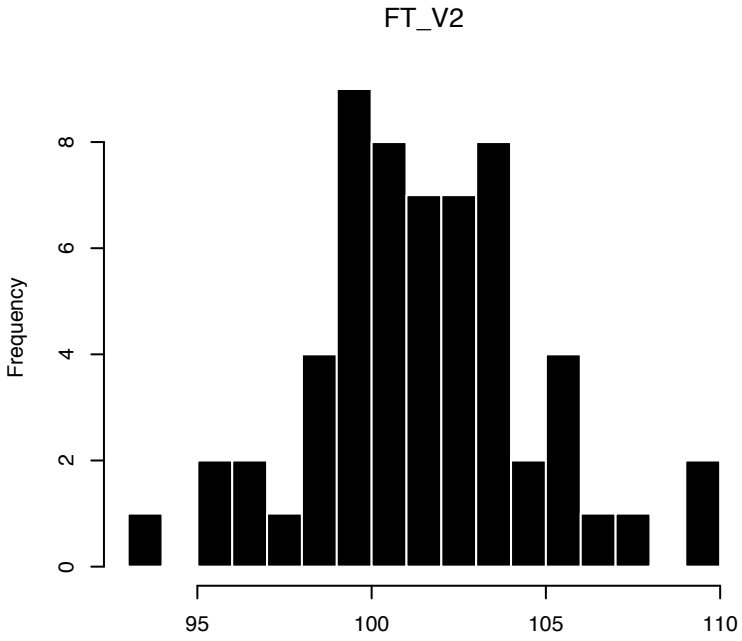

lambda 0.279
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

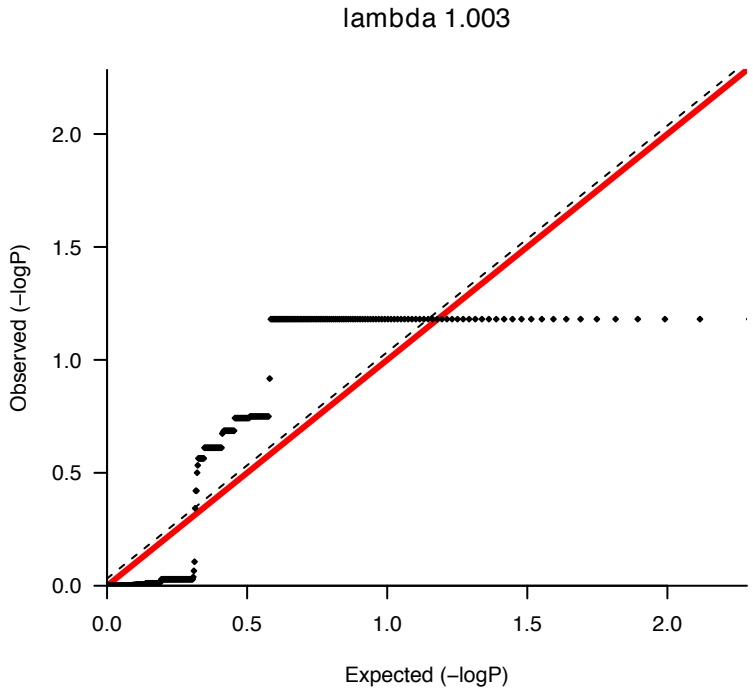

FT_V0
Frequency
0
1
2
3
4
5
6
95
100
105
110

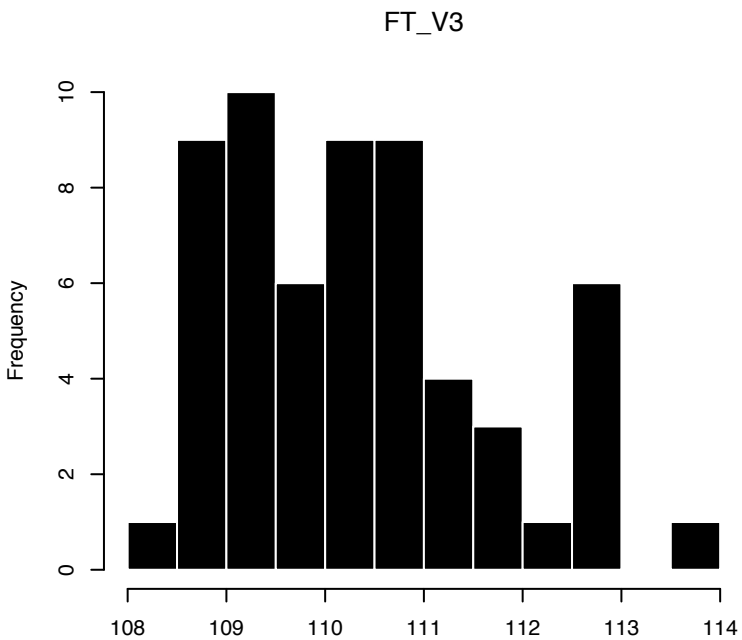

lambda 0.012
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0

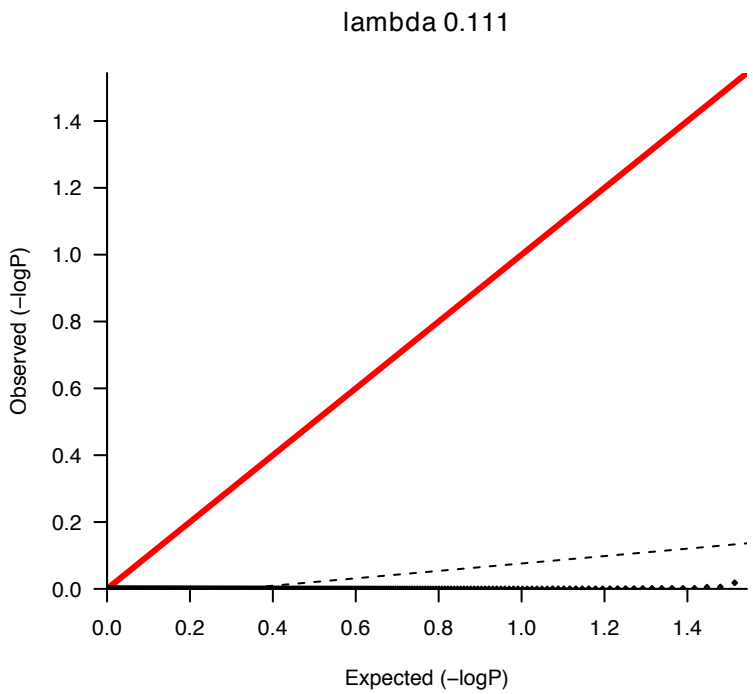

FT_V1
Frequency
0
1
2
3
4
5
6
7
100
105
110
115

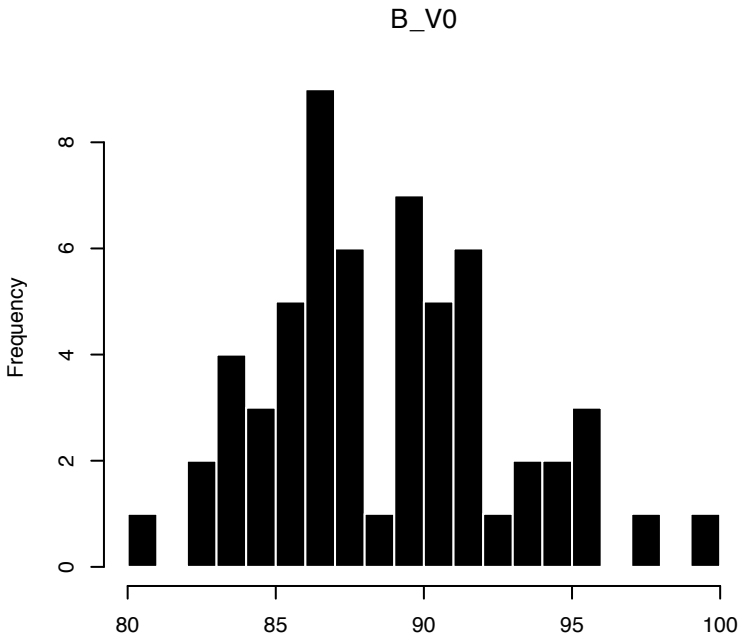

lambda 0.434
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

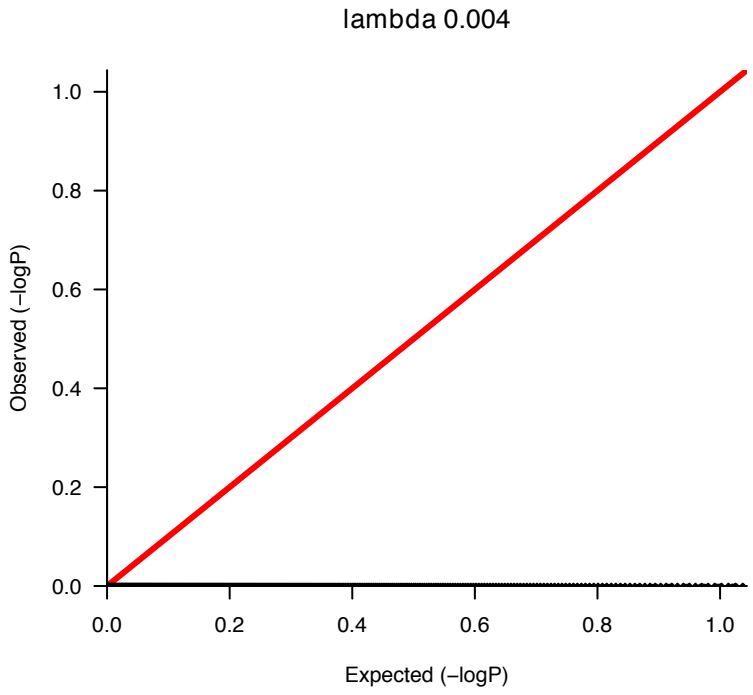

FT_V2
Frequency
0
2
4
6
8
95
100
105
110

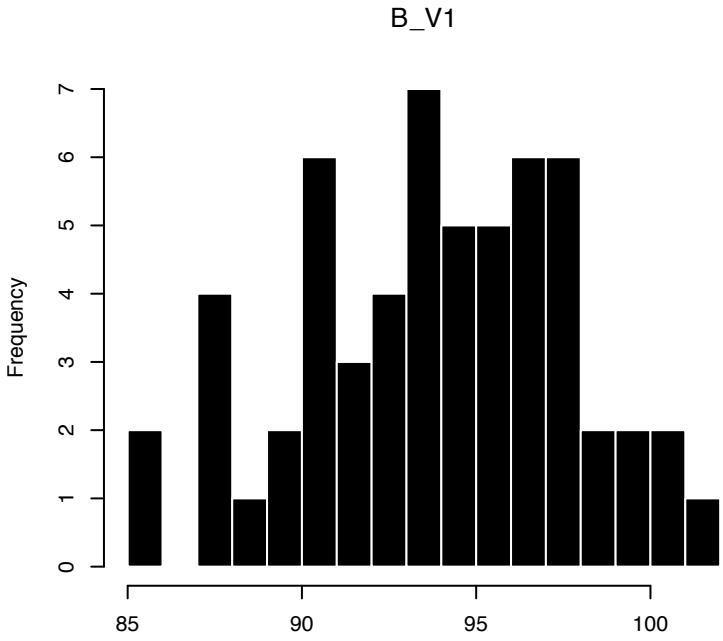

lambda 1.003
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

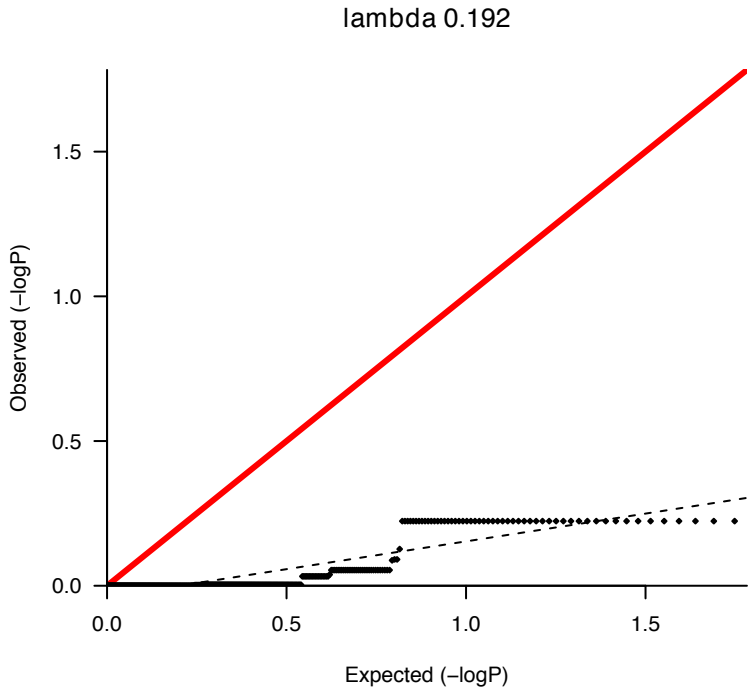

FT_V3
Frequency
0
2
4
6
8
10
108
109
110
111
112
113
114

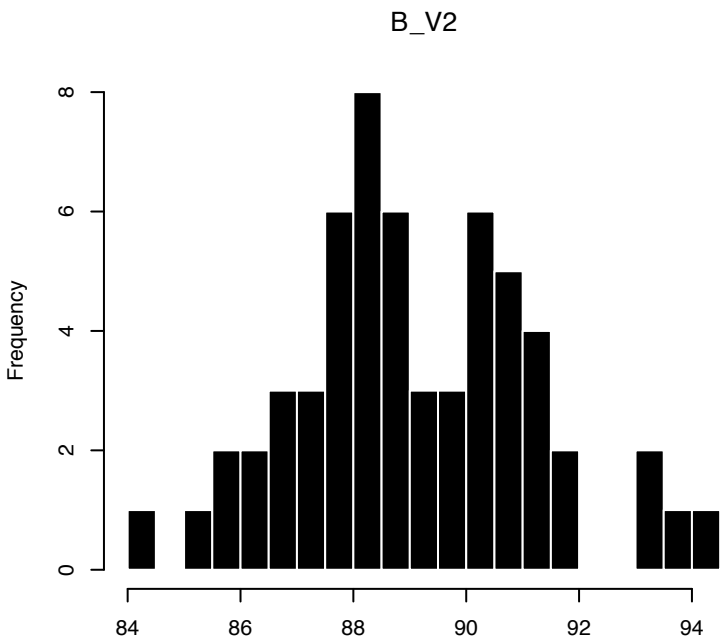

lambda 0.111
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

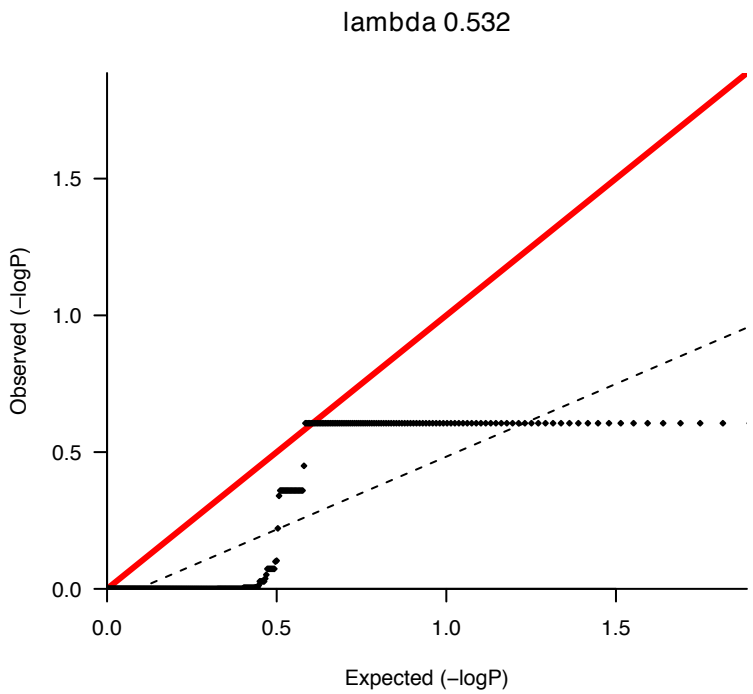

B_V0
Frequency
0
2
4
6
8
80
85
90
95
100

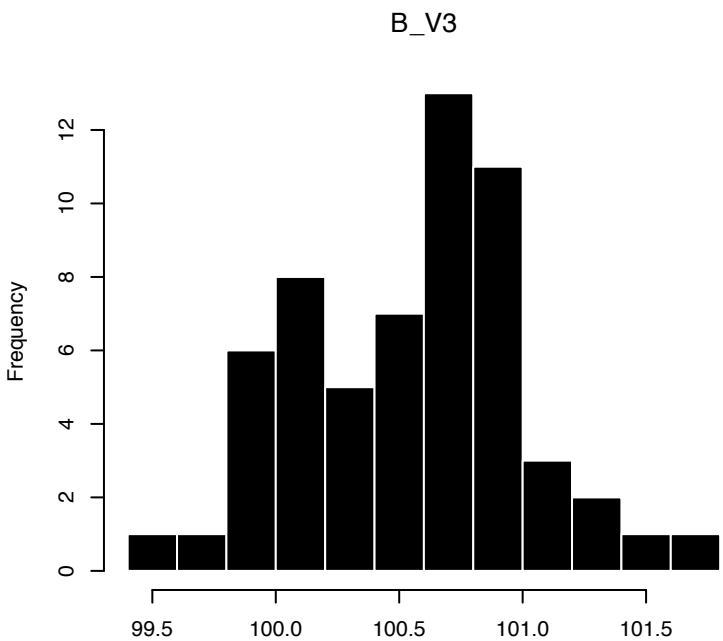

lambda 0.004
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0

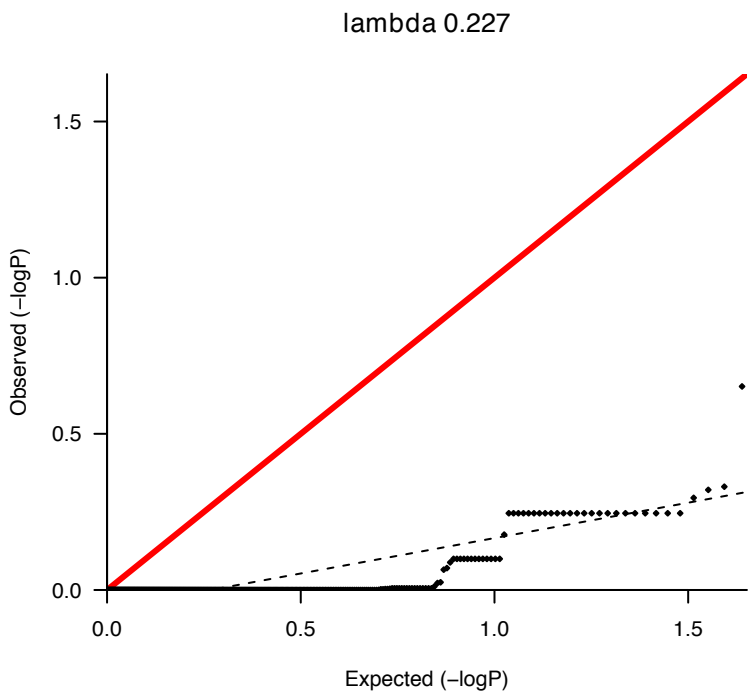

B_V1
Frequency
0
1
2
3
4
5
6
7
85
90
95
100

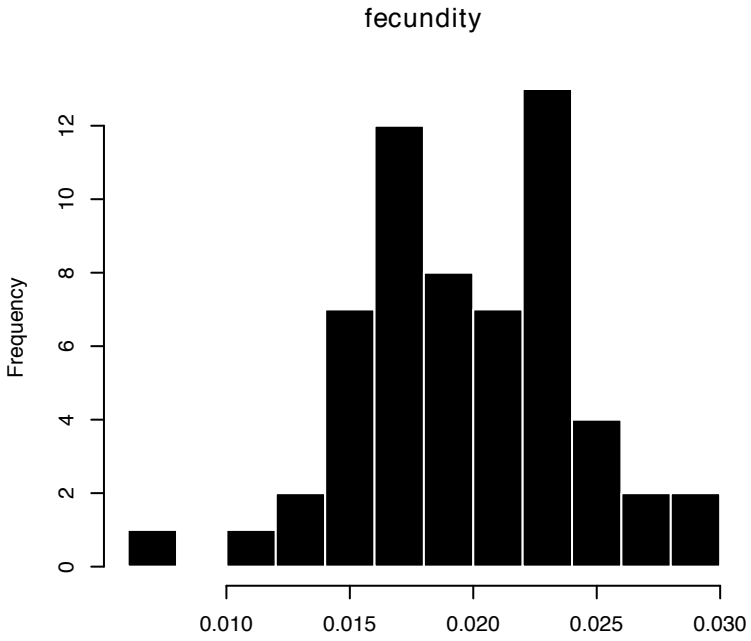

lambda 0.192
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

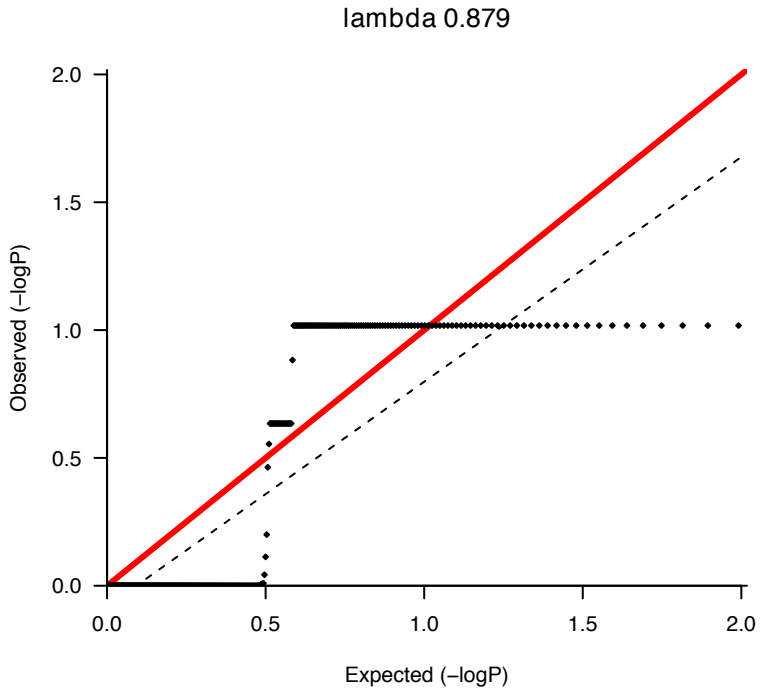

B_V2
Frequency
0
2
4
6
8
84
86
88
90
92
94

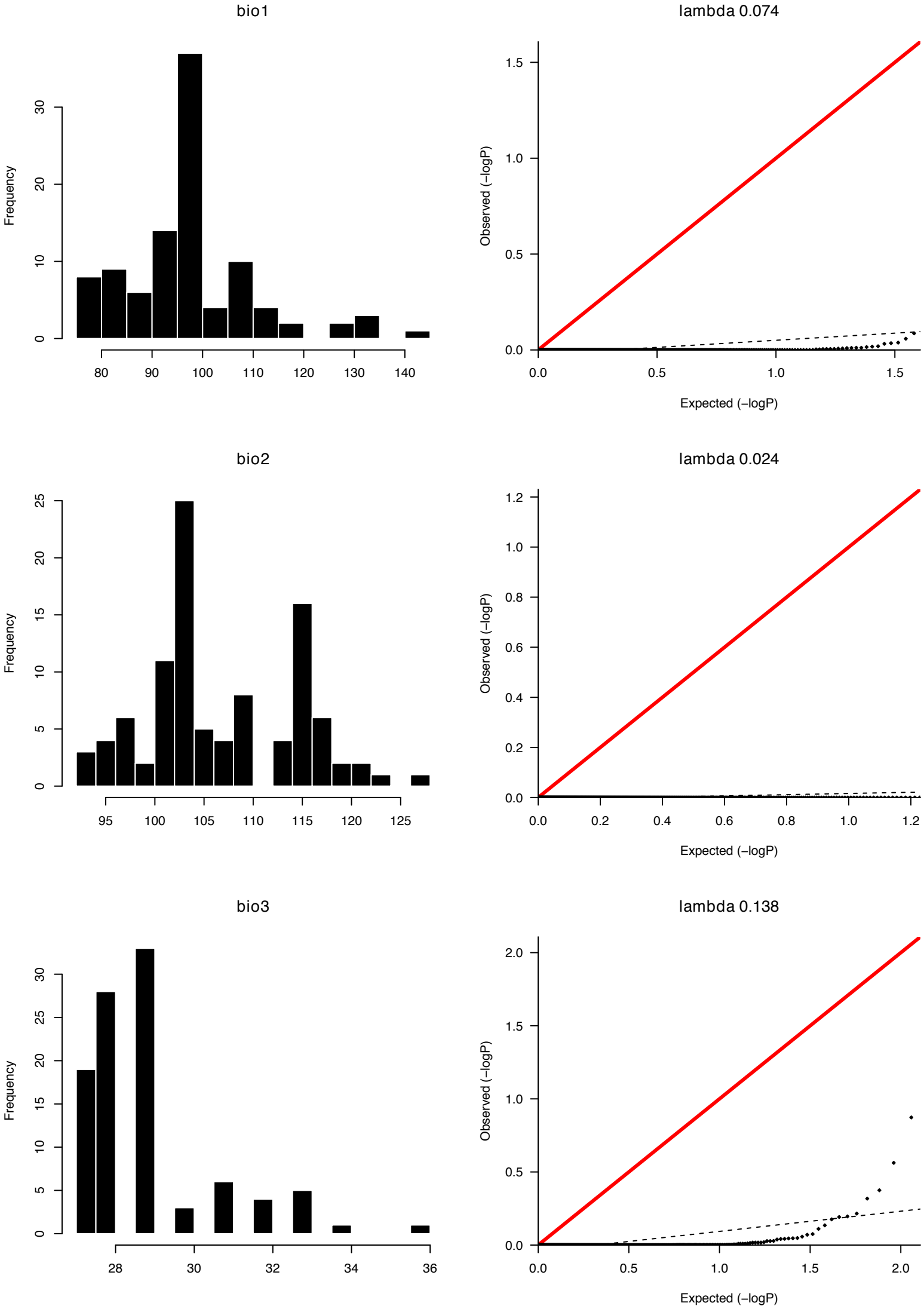

lambda 0.532
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

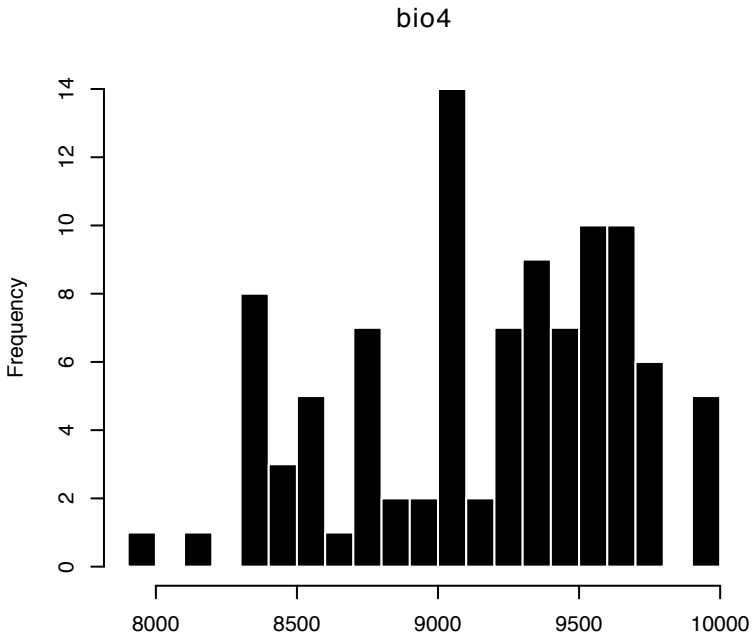

B_V3
Frequency
0
2
4
6
8
10
12
99.5
100.0
100.5
101.0
101.5

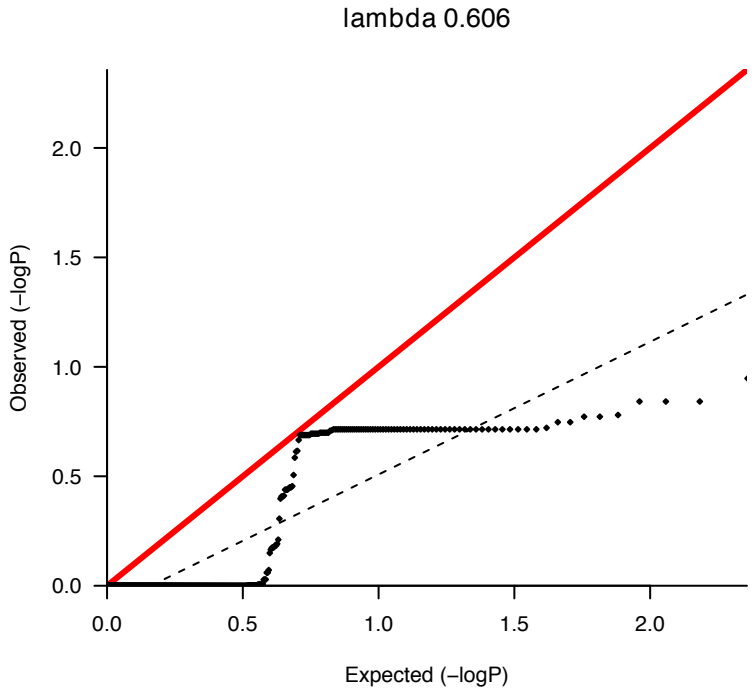

lambda 0.227
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

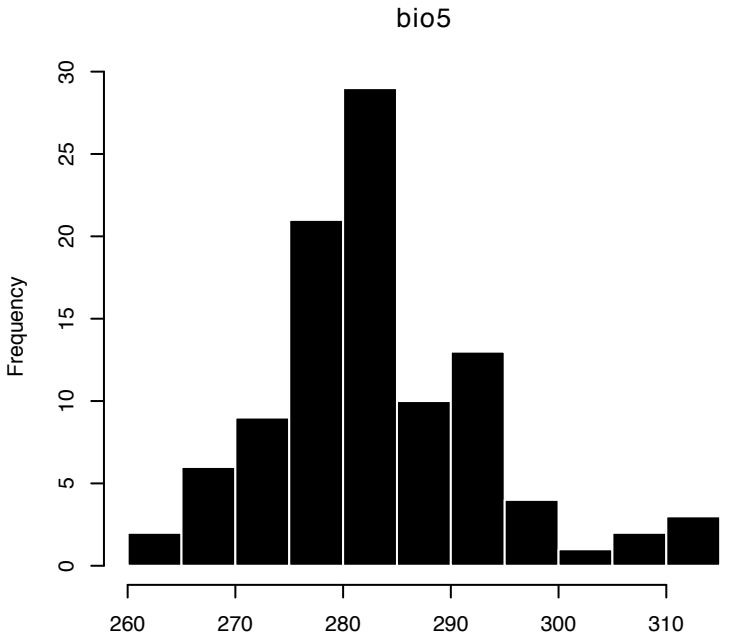

fecundity
Frequency
0
2
4
6
8
10
12
0.010
0.015
0.020
0.025
0.030

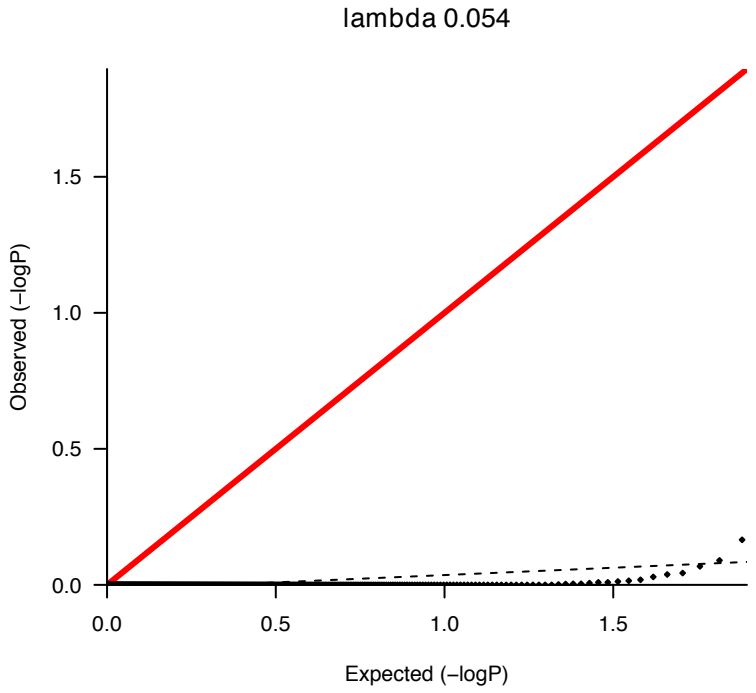

lambda 0.879
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

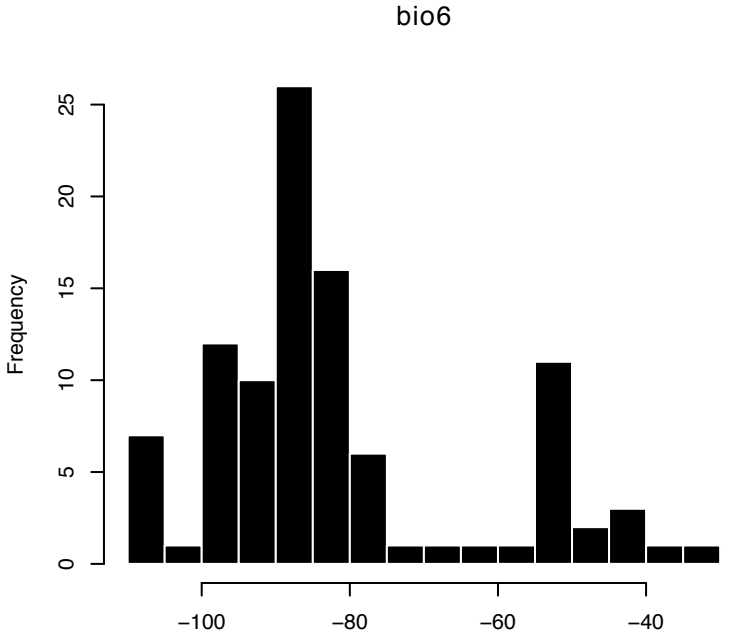

bio1
Frequency
0
10
20
30
80
90
100
110
120
130
140
lambda 0.074
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
bio2
Frequency
0
5
10
15
20
25
95
100
105
110
115
120
125
lambda 0.024
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
bio3
Frequency
0
5
10
15
20
25
30
28
30
32
34
36
lambda 0.138
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

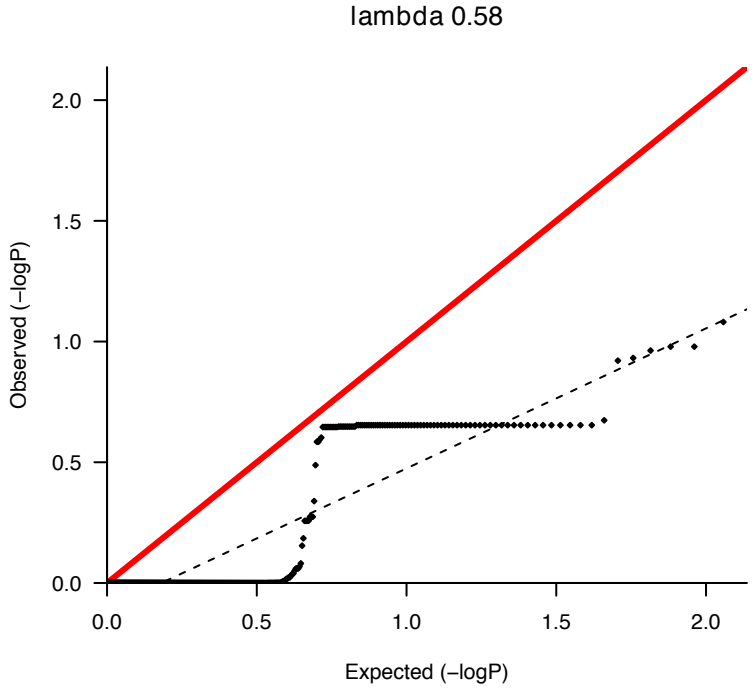

bio4
Frequency
0
2
4
6
8
10
12
14
8000
8500
9000
9500
10000

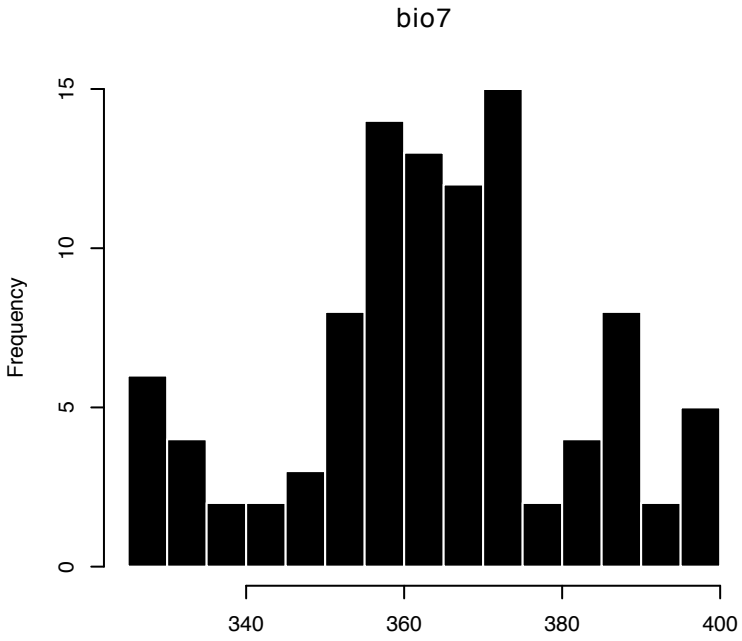

lambda 0.606
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

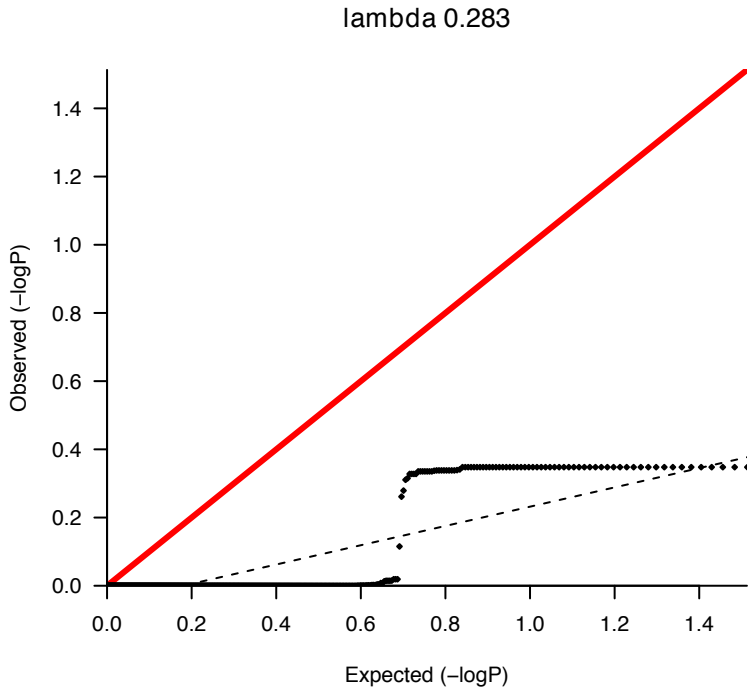

bio5
Frequency
0
5
10
15
20
25
30
260
270
280
290
300
310

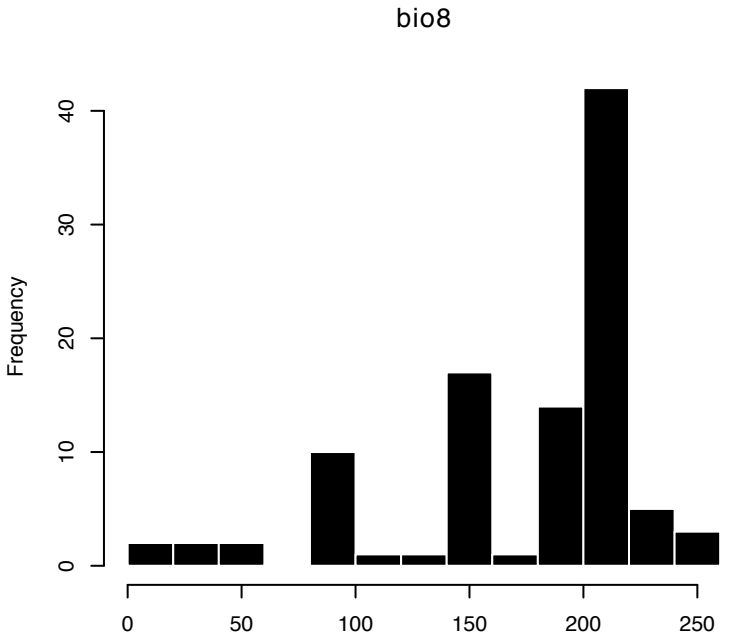

lambda 0.054
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

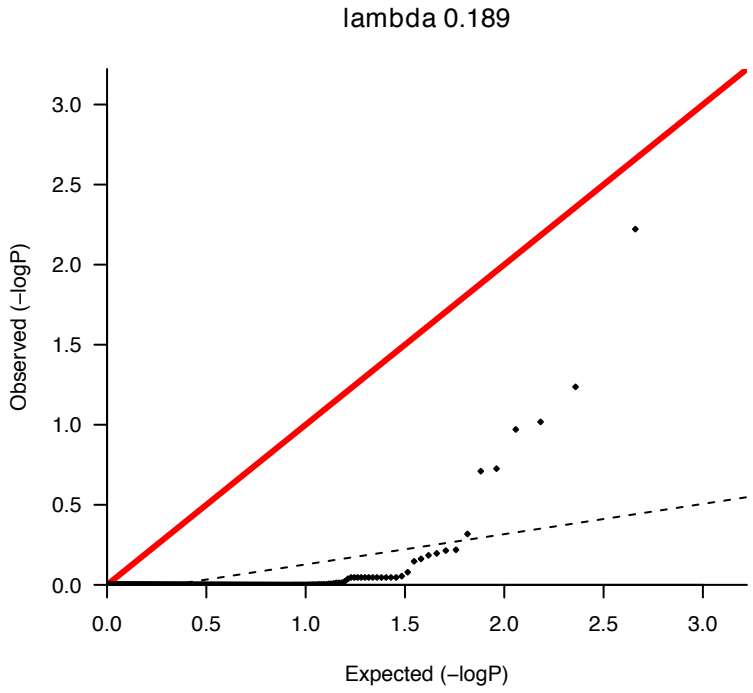

bio6
Frequency
0
5
10
15
20
25
−100
−80
−60
−40

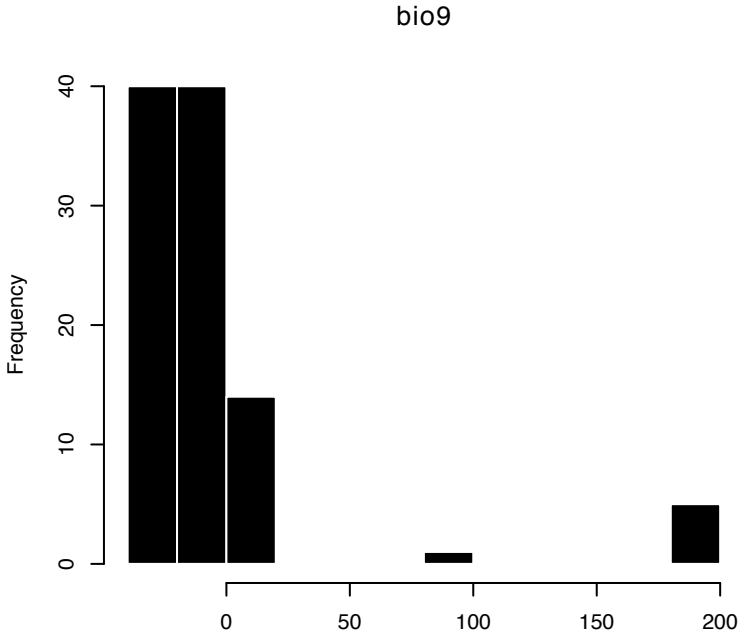

lambda 0.58
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

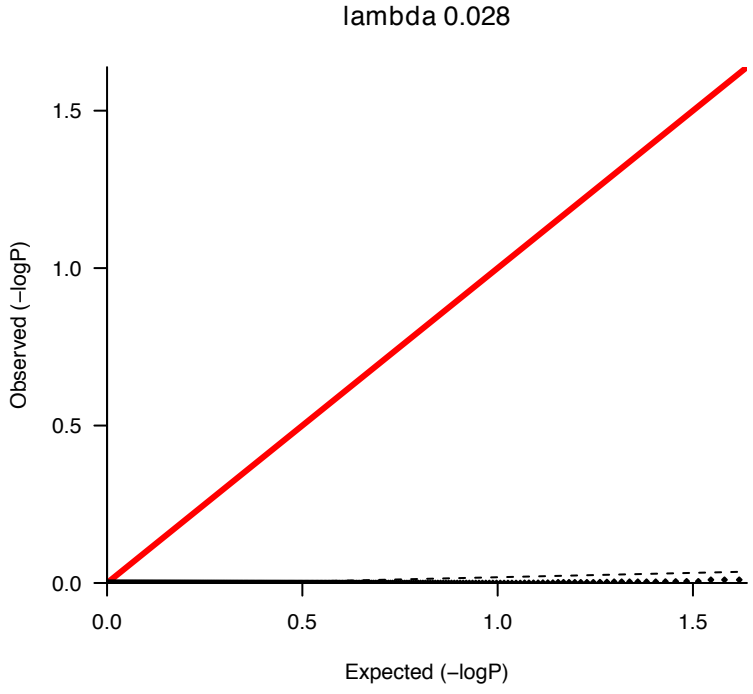

bio7
Frequency
0
5
10
15
340
360
380
400

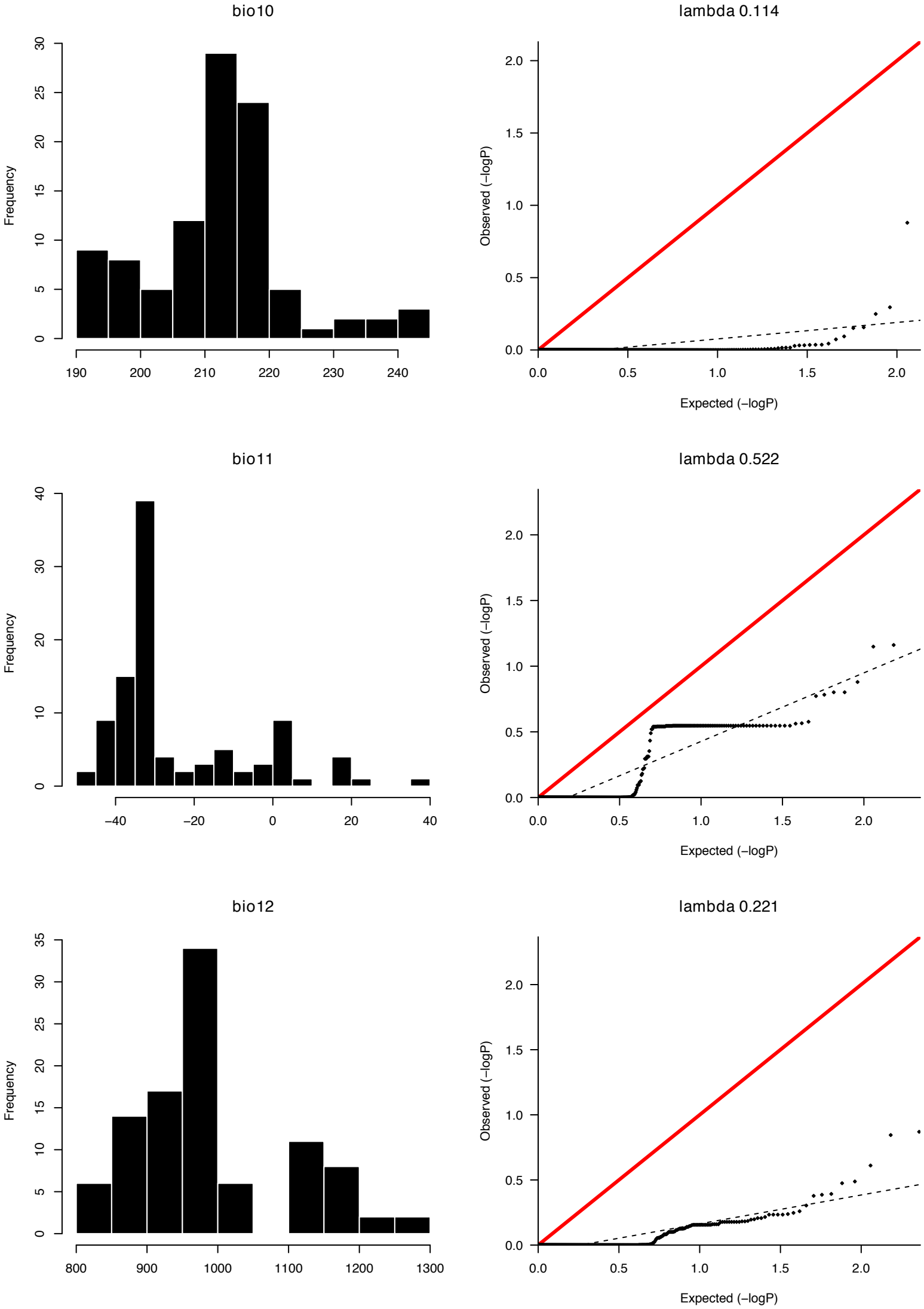

lambda 0.283
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

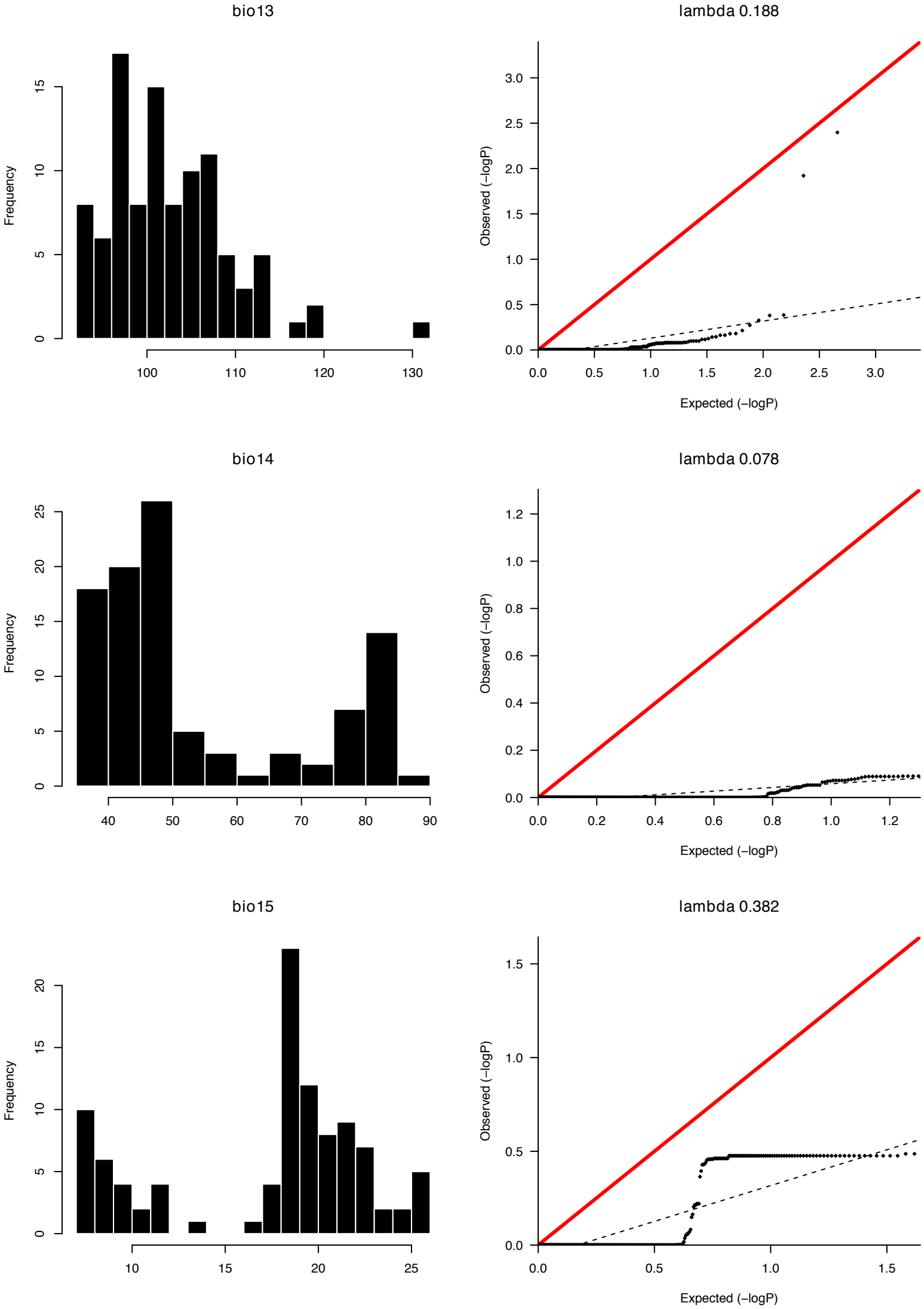

bio8
Frequency
0
10
20
30
40
0
50
100
150
200
250

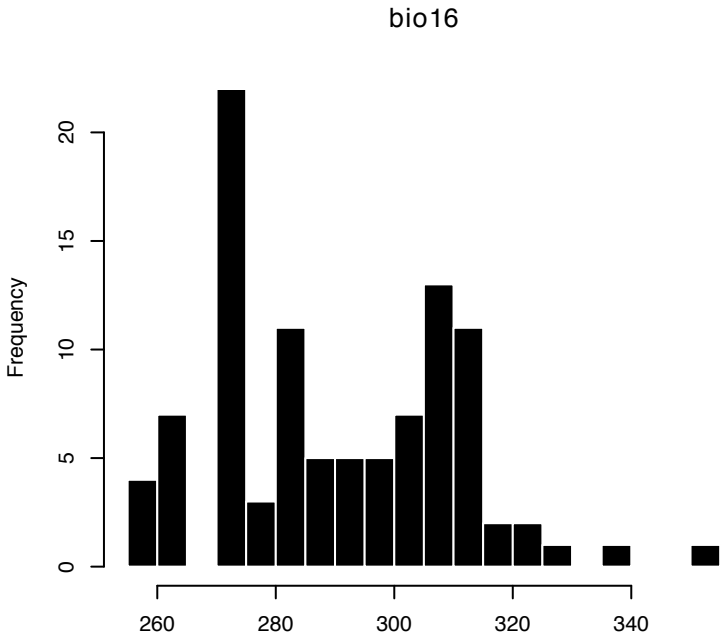

lambda 0.189
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

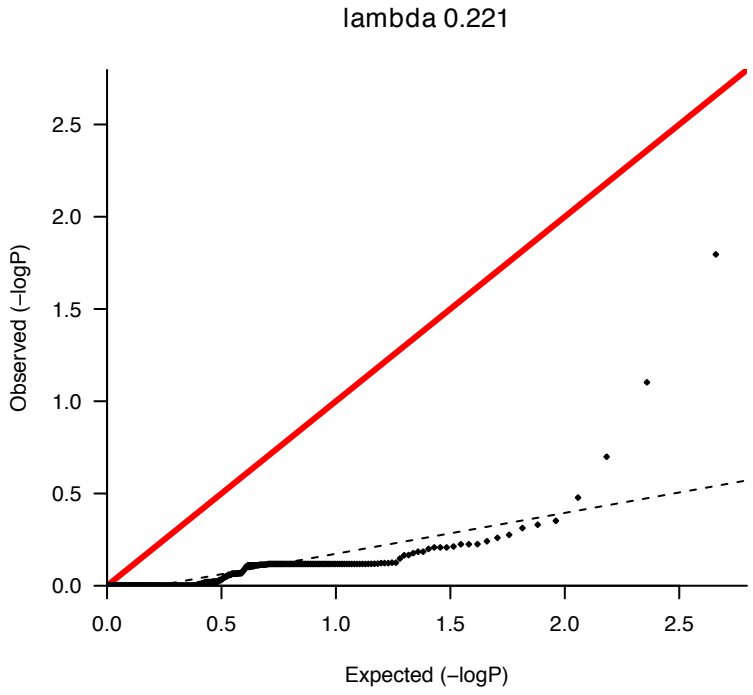

bio9
Frequency
0
10
20
30
40
0
50
100
150
200

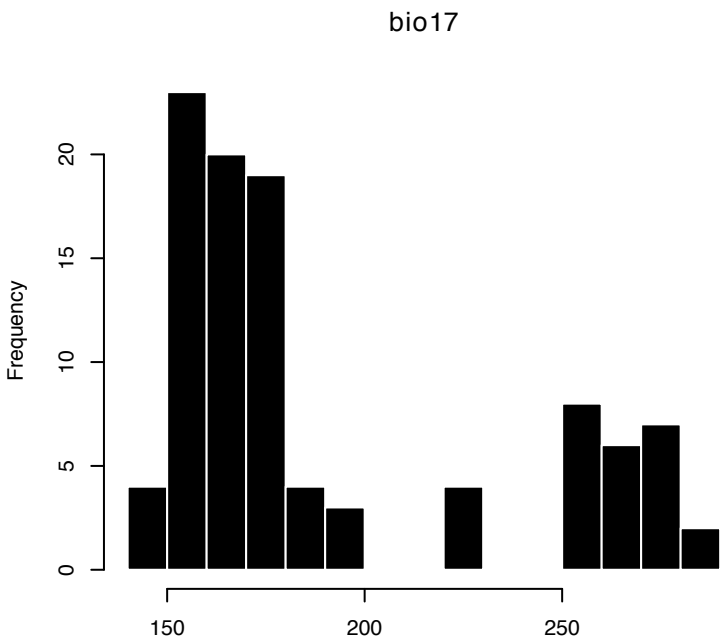

lambda 0.028
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

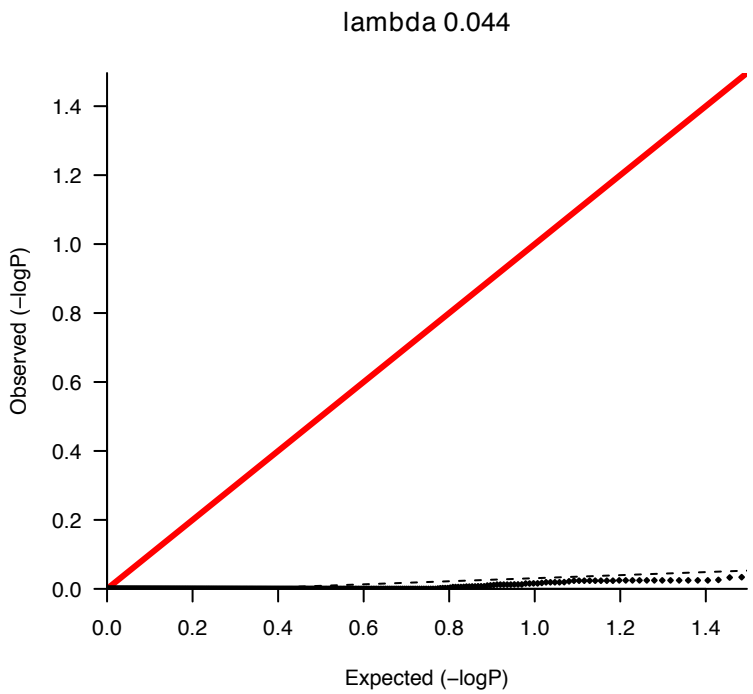

bio10
Frequency
0
5
10
15
20
25
30
190
200
210
220
230
240
lambda 0.114
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
bio11
Frequency
0
10
20
30
40
−40
−20
0
20
40
lambda 0.522
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
bio12
Frequency
0
5
10
15
20
25
30
35
800
900
1000
1100
1200
1300
lambda 0.221
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

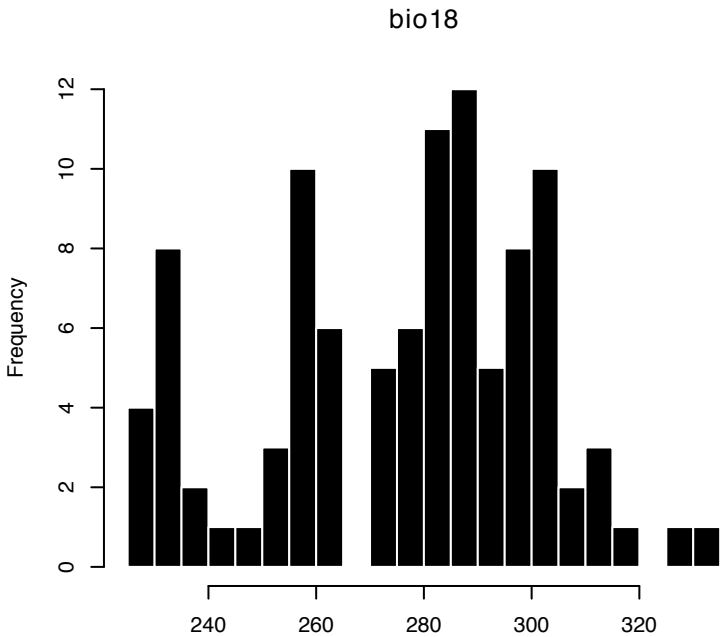

bio13
Frequency
0
5
10
15
100
110
120
130
lambda 0.188
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
bio14
Frequency
0
5
10
15
20
25
40
50
60
70
80
90
lambda 0.078
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
bio15
Frequency
0
5
10
15
20
10
15
20
25
lambda 0.382
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

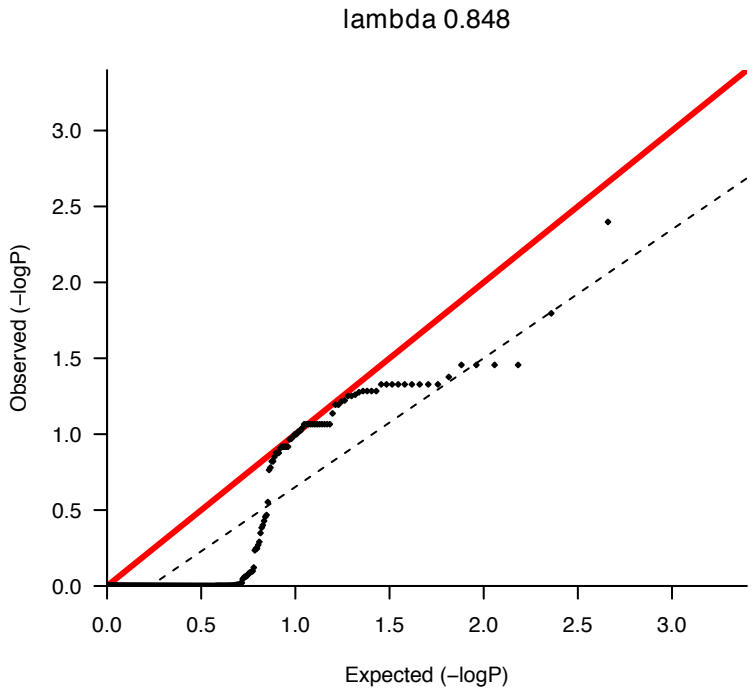

bio16
Frequency
0
5
10
15
20
260
280
300
320
340

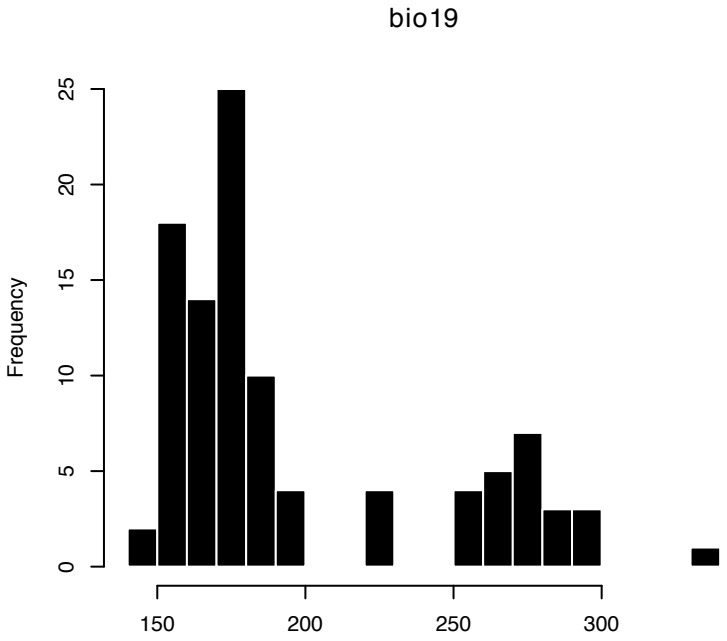

lambda 0.221
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
2.5

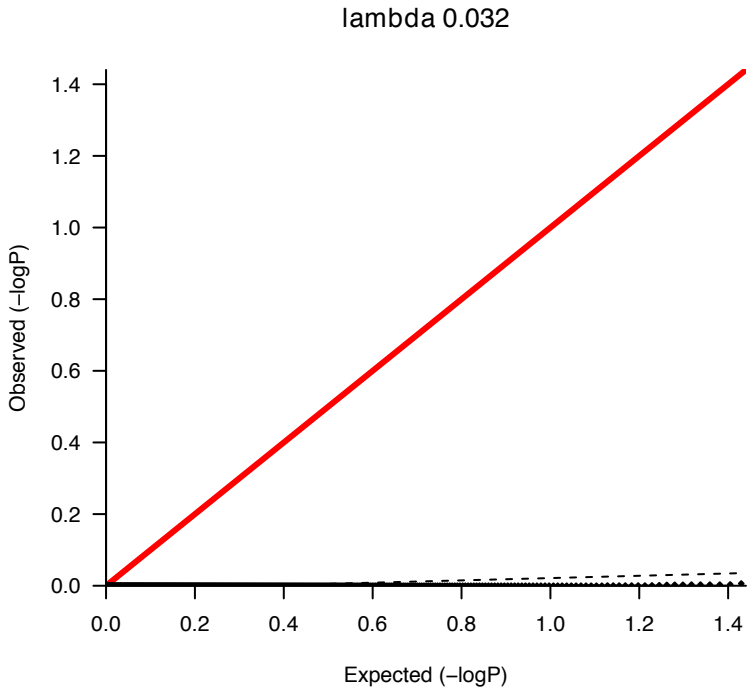

bio17
Frequency
0
5
10
15
20
150
200
250


lambda 0.044
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4


bio18
Frequency
0
2
4
6
8
10
12
240
260
280
300
320


lambda 0.848
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0


bio19
Frequency
0
5
10
15
20
25
150
200
250
300


lambda 0.032
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4