The rate and effect of *de novo* mutations in a colonizing lineage of *Arabidopsis thaliana*

Moises Exposito-Alonso[1,2†], Claude Becker[1†], Verena J. Schuenemann[3,4], Ella Reiter[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 to: hernan.burbano@tuebingen.mpg.de, weigel@weigelworld.org

Keywords: colonization, mutation, selection, herbarium genomes, aDNA, phylogenomics, population genomics, association mapping, *Arabidopsis thaliana*

Because colonizations and invasions are often associated with genetic bottlenecks, they offer an opportunity to directly observe de novo mutations and their subsequent fate. North America has recently been colonized by *Arabidopsis thaliana*, and many of the individuals found today belong to a single lineage, HPG1. To determine substitution rates under natural conditions in this lineage, we have sequenced 100 HPG1 genomes from plants collected between 1863 and 2006. We infer that the last common HPG1 ancestor lived in the early 17th century, most likely the time when HPG1 began to colonize N. America. Demographic reconstructions infer substantial population size fluctuations during the past four centuries. Even though changing demographics can undermine the effect of natural selection, we observed that mutations at coding sites were at lower frequency than mutations at other sites, consistent with the effect of purifying selection. Exceptionally, some mutations rose to high frequency and some had measurable effects in root development, consistent with positive selection acting over mutations with an adaptive value. Our work showcases how by applying genomics methods to a combination of modern and historic samples we can learn about plant colonisations and invasions and observe "evolution in action".

Knowledge of mutation rates and efficacy of selection, which together determine the substitutions that can be observed in a population, is essential for understanding evolution[1]. Mutation rates are required to estimate the effective diversity of populations[2], to date historic population splits[3], and to predict what opportunities there are for rapid novel adaptations, for instance, to drugs or pesticides[4,5]. Two extreme approaches to discover these parameters are either short-term mutation accumulation experiments in the laboratory[6] and pedigree-based estimates[7], or interspecific phylogenomic comparisons over millions of years[8]. Since mutation rates are generally studied independently of natural selection and population dynamics contexts, the two approaches often yield estimates that do not coincide, which has generated a heated controversy[9,10]. An alternative not exploited so far is the analysis of naturally occurring "evolutionary experiments" such as colonizing or invasive populations associated with a recent and strong genetic bottleneck[11,12]. Both colonizations and invasions are increasingly common due to human movement[13,14]. The study of these natural experiments is especially powerful when time-stamped samples from historic specimens can be used as internal calibration points[15,16]. Colonizing populations often start with very few individuals and therefore have low genetic diversity. The N. American population of the self-fertilizing plant *Arabidopsis thaliana* is no exception, with many individuals belonging to a genetically very similar lineage, haplogroup-1 (HPG1), that spread over large geographic areas[17]. The study of the origin along with demographic and selective dynamics of mutations in HPG1 can further help us solve the current paradox generated by two evolutionary conjectures: "Baker's law", inspired by the observation that selfing species are more successfully colonizing new environments than outcrossing ones, and "Muller's ratchet", which posits that selfing populations are evolutionary dead ends, since low diversity and accumulation of more deleterious mutations should hinder their ability to adapt to new environments[18–21].

To better understand the evolution of HPG1, we sequenced 27 herbarium specimens, from 1863–1993, and 76 live isolates of this lineage collected 1993–2006 (Fig. 1; Table S1). DNA retrieved from herbarium specimens showed biochemical features typical of ancient DNA[22], which indicates that the DNA recovered from historic samples is authentic (Fig. S1, see Supplementary Online Material [SOM] for details). We mapped reads against an HPG1 pseudo-reference genome[23], focusing on single nucleotide

polymorphisms (SNPs) because the short sequence reads of herbarium samples preclude accurate calling of structural variants. Genome sequences were of high quality, with herbarium samples covering 96.8–107.2 Mb of the 119 Mb reference, and modern samples covering 108.0–108.3 Mb (Table S1). Pairs of herbarium genomes differed on average by 109-222 SNPs, and pairs of modern genomes by 186-299 SNPs, that is, they ranged from 99.9997 to 99.9999 % identity .

A neighbor joining tree (Fig. 2A), multi-dimensional scaling (MDS) (Fig. 2B), and a parsimony network (Fig. 2C) confirmed very close relatedness of the HPG1 genomes, with only three apparent intra-HPG1 recombinants (Fig. 2C). Removing these resolved the reticulations in the parsimony network (Fig. 2D). The remaining 100 samples (Table S1) constitute a quasi-clonal lineage mostly devoid of effective recombination and population structure, and without SNPs in organellar genomes. The very low genome-wide nuclear diversity (π = 0.000002, θ_W = 0.00001, 5,013 segregating sites) is two orders of magnitude lower than in the native range of the species (θ_W = 0.007) (ref. [24]) (Table S1). The enrichment of low frequency variants in the site frequency spectrum (Tajima's *D* = -2.84; global= -2.04) (ref. [24]) and low levels of polymorphism are consistent with a recent bottleneck followed by population expansion (Fig. 3). The obvious explanation is that the bottleneck corresponds to a colonization founder event, likely by very few closely related individuals, or perhaps only a single plant. To describe intra-HPG1 relationships in a more sophisticated manner, we used Bayesian phylogenetic inference, exploiting collection dates for tip calibration of phylogenetic trees. The 76 modern individuals formed a largely monophyletic clade, with only four interspersed herbarium samples from the second half of the 20th century (Fig. 3A, B). Long branches reflected an abundance of singletons, typical of expanding populations after bottlenecks.

To estimate the substitution rate in the HPG1 lineage, we used distance- and phylogeny-based methods that take advantage of the known collection dates. One has to distinguish between the *mutation rate*, which is the rate at which genomes change due to DNA damage, faulty repair, gene conversion and replication errors, and *substitution rate*, which is the rate at which mutations survive and accumulate after demographic and natural selective processes[25]. Under neutral evolution, mutation and substitution rates should be equal[26]. The simple evolutionary history of the HPG1 natural population enables direct estimates of substitution rates, and by comparing these with mutation rates calculated in controlled conditions, we can learn about demographic and selective forces. In the distance method, the substitution rate is calculated from correlation between differences in collection time in historic-modern sample pairs, and the number of changes between those pairs relative to a reference (Fig. 3C), scaled to the size of the genome accessible to Illumina sequencing. This method resulted in an estimated rate of $2.11*10^{-9}$ substitutions site^{-1} year^{-1} (95% bootstrap Confidence Interval [CI]: 1.88–$2.33*10^{-9}$) using rigorous SNP calling quality thresholds. Relaxing the quality thresholds for base calling and minimum genotyped rate affects both the number of called SNPs and the length of the interrogated reference sequence[27]. These largely cancelled each other out, and our estimates were relatively stable, between 2.1–$3.2*10^{-9}$ substitutions site^{-1} year^{-1} (Table S3). The Bayesian phylogenetic approach, which uses the collection years for tip calibration and assumes a relaxed molecular clock, yielded a similar estimate, $4.0*10^{-9}$, with confidence ranges overlapping the above estimates (95% Highest Posterior Probability Density [HPPD]: 3.2–$4.7*10^{-9}$). Based on the results obtained with different methods, we can confidently say that the substitution rate in the wild should be between 2 to 5 $*10^{-9}$ site^{-1} year^{-1}. We recommend

that these rates be used to date temporal splits between populations. To be able to compare our substitution rate with the mutation rate, both need to be expressed per generation. While *A. thaliana* is an annual plant, seed bank dynamics generate a delay of average generation time at the population scale. A comprehensive study of multiple *A. thaliana* populations reports an average generation time of 1.3 years[28], with a notable variance across populations. Re-scaling with the mean generation time led to an adjusted substitution rate of $2.7*10^{-9}$ substitutions site^{-1} generation^{-1} (95% CI $2.4\text{-}3.0*10^{-9}$) (Fig. 3E). This is much lower than the rate of $7.1*10^{-9}$ mutations site^{-1} generation^{-1} (95% CI $6.3–7.9*10^{-9}$) (Tables S2, S3) that one can calculate from resequencing data for mutation accumulation (MA) lines in the Col-0 reference background grown in the greenhouse[29].

Differences in "per generation" rates could be caused by several factors, such as an imperfect knowledge of the generation time in the wild (for rates using different generation times, see Fig. 3E). In addition, mutagenic environmental factors, genome background, mutation spectrum, or methodological idiosyncrasies can affect the estimates. For example, transposons, which comprise ~8% of the genome and ~19% of the SNPs in greenhouse MA lines, had fewer SNPs called than expected in HPG1 (~13%). This is likely due to difficulties when mapping reads to genomic areas with extensive structural variation[30],[30]a[30]d[30] could have contributed to the lower substitution rate estimates for HPG1 (Fig. 3E, Fig S3C, Table S3). In addition, the substitution spectrum in HPG1 is shifted to a lower transition/transversion ratio compared to the MA lines (Fig. 3C and Fig. S3), which could be caused by methylated cytosines (see Fig. S4 and SOM). Finally, an alternative evolutionary explanation for the rate differences is that purifying selection slows the accumulation of mutations in the wild by removing deleterious mutations (Fig. 3E). To find evidence of negative selection independently of dataset comparisons, we looked at the site frequency spectrum of different annotations within the HPG1 dataset. Medium-frequency variants, which are more exposed to purifying selection[31], were more sharply depleted in genomic regions expected to be under greater selection constraint (genic and nonsynonymous sites) than in putatively more neutral ones (intergenic or synonymous sites) (Fisher's Exact test, p-value <0.05 for both comparisons, see Fig. S5). Therefore, even if we cannot say with certainty that purifying selection drives the differences between HPG1 and MA rates, it must be responsible for the differences between different types of sites within HPG1.

The substitution rate allows dating of HPG1's origin. The mean estimate from Bayesian methods was the year 1597 (HPPD 95%: 1519–1660) (Fig. 3A, B). We also used a non-phylogenetic method that utilizes the relationship between the average genetic distance between any two individuals, with the substitution rate multiplied by twice the divergence time and the genome size; solving by the divergence time in the equation we obtained 353 years. When subtracted from the average collection date of our samples, the corresponding point estimate is 1625, within the confidence interval of the Bayesian estimate. This corresponds to the date of the last common ancestor of HPG1, and should thus be close to the time of introduction of HPG1 to N. America. (The date is older than our previous estimate, for which we had naively applied the higher greenhouse mutation rate[23]). Inference of N_e through time suggested exponential population growth until the early 19th century (Fig 3B, Fig S6C). During the 20th century the N_e trajectory showed oscillating patterns between growth and bottlenecks, which are typical of selfing organisms [32], and which likely led to a replacement of most HPG1 sublineages, as the modern samples are all very closely related (Fig. 3 A, B).

155 Since we knew both the collection years and locations of origin of the HPG1 samples, we could
156 also analyze the migration dynamics of HPG1. Although unknown sources of sampling bias could affect
157 our analyses[16], the phylogeographic models suggested that HPG1 came to cover much of its modern
158 range soon after its introduction to N. America (Fig. S6 A,B). We found a significant correlation between
159 collection date and both latitude and longitude (Fig. 1C) , which we interpret as a net, highly dispersed,
160 movement in a northwestern direction over time. Additional support for this hypothesis comes from an
161 isolation-by-distance signal, which is most consistent with a historic westward migration and a more
162 recent reverse eastward migration (Fig. S6 E,F). The apparent source of those new migrants now
163 persisting along the East coast was the Lake Michigan area .

164 Finally, while we did not expect to easily find mutations that have helped the HPG1 lineage to
165 adapt to its new N. American environment, we wanted to determine whether any of the mutations have
166 measurable phenotypic effects. Focusing on flowering-related, reproductive and root traits of likely
167 ecological relevance, we detected significant quantitative heritable variation (Table S4). We used an
168 approach borrowed from GWAS to find SNPs that had increased in frequency (>5%) and were associated
169 with these phenotypes. Because conventional GWAS relies on recombination, which is almost absent
170 from our population, our approach could not identify individual SNPs, but only SNP cohorts distributed
171 across the genome[33]. We found 79 SNPs associated with root traits, of which nine resulted in
172 nonsynonymous changes. We did not find any SNPs associated with flowering time, even though it is
173 thought to be a key player of rapid adaptation in many annual species[34,35]. Nineteen other SNPs, of
174 which four were nonsynonymous, were associated with climate variables (www.worldclim.org) even
175 after correction for latitude and longitude, and some of the hits overlapped between root traits and
176 climate variables (Table 1, Table S5, Fig. S7). Although a good number of SNPs was associated with
177 phenotypes and/or climate variables, it is not possible to confidently pinpoint individual candidate SNPs,
178 since the extent of whole-genome linkage disequilibrium (LD) in HPG1 is high (Fig. S9 B,C). However,
179 there is a gradient in the extent of LD between SNPs associated with root architecture, which could help
180 to determine particularly promising candidates for molecular characterization and quantification of
181 fitness in natural conditions (Table 1, Fig. S9 D-F). For example, the gene AT5G19330, overexpression of
182 which confers salt tolerance[36], contains a SNP that was unlinked to other hits and that has risen in the
183 last four centuries to a frequency of 40%. This SNP is likely to change protein function, as it leads to a
184 substitution of cysteine for tryptophan. Another nonsynonymous SNP is located in AT2G38910, encoding
185 a calcium dependent kinase that belongs to a family of factors involved in root hydraulic conductivity and
186 phytohormone response[37,38]. Remarkably, most derived root-associated alleles, when compared with
187 equally frequent neutral alleles, were first seen in older herbarium samples, most of which were
188 collected near Lake Michigan, the apparent source of modern populations (Fig. S8). Altogether, the rise
189 in frequency, older age of some *de novo* mutations, and their quantifiable phenotypic effects and
190 climatic correlations strengthen the hypothesis that they might have an adaptive value and were under
191 positive selection. These results favor Baker's law, which alleges that selfing species can often adapt to
192 new environments, over the evolutionary dead-end hypothesis of Muller's ratchet. Furthermore, these
193 suggestive signals of rapid adaptation via *de novo* mutations could change the current paradigm that
194 invasive species adapt most often from sources of standing variation, either because an incomplete

bottleneck left residual variation or because there has been subsequent admixture with native species or secondary colonizers[39,40].

In summary, we have exploited whole-genome information from historic and contemporary collections of a herbaceous plant to empirically characterize the effect of evolutionary forces during a recent colonization. With this natural time series experiment, we could directly estimate the nuclear substitution rate in wild *A. thaliana* populations. This parameter, which provides immediate ability to date key events of populations, is only known for one other species: *Homo sapiens* [41]. We have presented evidence that purifying selection is perceptible already over time scales spanning only a few centuries. Although the colonizing population we have investigated has limited diversity and suffered rapid fluctuations in population size, there appear to be *de novo* mutations with phenotypic effects that contributed to rapid adaptation. While *A. thaliana* HPG1 is not an invasive species, it can teach us about fundamental evolutionary processes behind successful colonizations and adaptation to new environments. Our work should encourage others to search for similar natural experiments and to unlock the potential of herbarium specimens to study "evolution in action".

Online Content Methods, along with any additional Extended Data display items, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Supplementary Information is available in the online version of the paper.

Acknowledgments For providing and retrieving herbarium specimens, we thank R. Capers, J. Devos, G. Shirsekar, M. S. Dossmann, J. Freudenstein, C. M. Herring, C. Niezgoda, C. A. McCormick, J. Peter and M. Thines. We thank X. Zhao and I. Henderson for recombination estimates, C. Lanz for sequencing support, C. Goeschl, B. Zierfuss and B. Wohlrab for help with root analyses, and P. Lang, D. Seymour, and D. Koenig for thorough proofreading and comments on the manuscript. We thank M. Nordborg for discussions and pointing us to the work of A.R. Templeton, K. Pruefer for input on data analysis, and the Weigel and Burbano labs for comments. Supported by ERC Advanced Grant IMMUNEMESIS and the President's Fund of the Max Planck Society, project "Darwin".

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. C.S. and R.S. performed and analyzed root assays and seed size measurements under the supervision of W.B.; C.B. and J.H. sequenced and curated modern samples, coordinated by D.W.; 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.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.A.B. (hernan.burbano@tue.mpg.de) or D.W. (weigel@weigelworld.org).

235

REFERENCES

236 1. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17,** 704–714
237 (2016).
238 2. Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.*
239 **10,** e1001388 (2012).
240 3. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution.
241 *Nat. Rev. Genet.* **13,** 745–753 (2012).
242 4. Pennings, P. S. & Hermisson, J. Soft Sweeps II—Molecular Population Genetics of Adaptation from Recurrent
243 Mutation or Migration. *Mol. Biol. Evol.* **23,** 1076–1084 (2006).
244 5. Karasov, T., Messer, P. W. & Petrov, D. A. Evidence that adaptation in Drosophila is not limited by mutation at
245 single sites. *PLoS Genet.* **6,** e1000924 (2010).
246 6. Halligan, D. L. & Keightley, P. D. Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. *Annu.*
247 *Rev. Ecol. Evol. Syst.* **40,** 151–172 (2009).
248 7. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*
249 **328,** 636–639 (2010).
250 8. Wolfe, K. H., Sharp, P. M. & Li, W.-H. Rates of synonymous substitution in plant nuclear genes. *J. Mol. Evol.* **29,**
251 208–211 (1989).
252 9. Subramanian, S. & Lambert, D. M. Selective constraints determine the time dependency of molecular rates for
253 human nuclear genomes. *Genome Biol. Evol.* **4,** 1127–1132 (2012).
254 10. Gao, Z., Wyman, M. J., Sella, G. & Przeworski, M. Interpreting the Dependence of Mutation Rates on Age and
255 Time. *PLoS Biol.* **14,** e1002355 (2016).
256 11. Sax, D. F. *et al.* Ecological and evolutionary insights from species invasions. *Trends Ecol. Evol.* **22,** 465–471
257 (2007).
258 12. G. F. Gauze. *The struggle for existence*. (The Williams & Wilkins company, 1934).
259 13. van Kleunen, M. *et al.* Global exchange and accumulation of non-native plants. *Nature* **525,** 100–103 (2015).
260 14. Razanajatovo, M. *et al.* Plants capable of selfing are more likely to become naturalized. *Nat. Commun.* **7,**
261 13313 (2016).
262 15. Green, R. E. & Shapiro, B. Human evolution: turning back the clock. *Curr. Biol.* **23,** R286–8 (2013).
263 16. Crawford, P. H. C. & Hoagland, B. W. Can herbarium records be used to map alien species invasion and native
264 species expansion over the past 100 years? *J. Biogeogr.* **36,** 651–661 (2009).
265 17. Platt, A. *et al.* The scale of population structure in Arabidopsis thaliana. *PLoS Genet.* **6,** e1000843 (2010).
266 18. Barrett, S. C. H. Foundations of invasion genetics: the Baker and Stebbins legacy. *Mol. Ecol.* **24,** 1927–1941
267 (2015).
268 19. Stebbins, G. L. Self Fertilization and Population Variability in the Higher Plants. *Am. Nat.* **91,** 337–354 (1957).
269 20. Muller, H. J. THE RELATION OF RECOMBINATION TO MUTATIONAL ADVANCE. *Mutat. Res.* **106,** 2–9 (1964).
270 21. Lynch, M., Conery, J. & Burger, R. Mutation Accumulation and the Extinction of Small Populations. *Am. Nat.*
271 **146,** 489–518 (1995).
272 22. Weiß, C. L. *et al.* Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium
273 specimens. *Royal Society Open Science* **3,** 160239 (2016).
274 23. Hagmann, J. *et al.* Century-scale Methylome Stability in a Recently Diverged Arabidopsis thaliana Lineage. *PLoS*
275 *Genet.* **11,** e1004920–e1004920 (2015).
276 24. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis
277 thaliana. *Cell* **0,** (2016).
278 25. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14,** 827–839
279 (2013).

26. Kimura, M. On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.* **9,** 23–23 (1967).
27. Ness, R. W., Morgan, A. D., Colegrave, N. & Keightley, P. D. Estimate of the spontaneous mutation rate in Chlamydomonas reinhardtii. *Genetics* **192,** 1447–1454 (2012).
28. Falahati-Anbaran, M., Lundemo, S. & Stenøien, H. K. Seed dispersal in time can counteract the effect of gene flow between natural populations of Arabidopsis thaliana. *New Phytol.* **202,** 1043–1054 (2014).
29. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* **327,** 92–94 (2010).
30. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13,** 36–46 (2012).
31. Charlesworth, B. & Charlesworth, D. *Elements of Evolutionary Genetics*. (Roberts and Company Publishers, 2010).
32. Arunkumar, R., Ness, R. W., Wright, S. I. & Barrett, S. C. H. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics* **199,** 817–829 (2015).
33. Templeton, A. R., Sing, C. F., Kessling, A. & Humphries, S. A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics* **120,** 1145–1154 (1988).
34. Franks, S. J., Sim, S. & Weis, A. E. Rapid evolution of flowering time by an annual plant in response to a climate fluctuation. *Proceedings of the National Academy of Sciences* **104,** 1278–1282 (2007).
35. Bradshaw, W. E. & Holzapfel, C. M. Genetic response to rapid climate change: it's seasonal timing that matters. *Mol. Ecol.* **17,** 157–166 (2008).
36. Kim, S. *et al.* ARIA, an Arabidopsis arm repeat protein interacting with a transcriptional regulator of abscisic acid-responsive gene expression, is a novel abscisic acid signaling component. *Plant Physiol.* **136,** 3639–3648 (2004).
37. Li, G. *et al.* The calcium-dependent protein kinase CPK7 acts on root hydraulic conductivity. *Plant Cell Environ.* **38,** 1312–1320 (2015).
38. Choi, H.-I. *et al.* Arabidopsis calcium-dependent protein kinase AtCPK32 interacts with ABF4, a transcriptional regulator of abscisic acid-responsive gene expression, and modulates its activity. *Plant Physiol.* **139,** 1750–1761 (2005).
39. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23,** 38–44 (2008).
40. Dlugosch, K. M. & Parker, I. M. Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Mol. Ecol.* **17,** 431–449 (2008).
41. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–449 (2014).

314

METHODS

315 **Sample collection and DNA sequencing.** Modern *A. thaliana* accessions were from the collection
316 described by Platt and colleagues[17] which identified HPG1 candidates based on 149 genome-wide SNPs
317 (Table S1). Herbarium specimens were directly sampled by Max Planck colleagues Jane Devos and
318 Gautam Shirsekar, or sent to us by collection curators from various herbaria (Table S1). DNA from
319 herbarium specimens was extracted as described[42] in a clean room facility at the University of Tübingen.
320 Two sequencing libraries with sample-specific barcodes were prepared following established protocols,
321 with and without repair of deaminated sites using uracil-DNA glycosylase and endonuclease VIII (refs.
322 [43–45]) (see Supplementary Online Material [SOM]). DNA from modern individuals was extracted from
323 pools of eight siblings using the DNeasy plant mini kit (Qiagen, Hilgendorf, Germany). Genomic DNA
324 libraries were prepared using the TruSeq DNA Sample or TruSeq Nano DNA sample prep kits (Illumina,
325 San Diego, CA), and sequenced on Illumina HiSeq 2000, HiSeq 2500 or MiSeq instruments. Paired-end
326 reads from modern samples were trimmed and quality filtered before mapping using the SHORE pipeline
327 v0.9.0 (ref. [46,47]). Because ancient DNA fragments are short (Fig. S1B), we merged forward and reverse
328 reads for herbarium samples after trimming, requiring a minimum of 11 bp overlap[48], and treated the
329 resulting as single-end reads. Reads were mapped with GenomeMapper v0.4.5s (ref. [49]) against an HPG1
330 pseudo-reference genome[23], and against the Col-0 reference genome, and SNPs were called with
331 SHORE[23,50] using different thresholds. Samples JK2509 to JK2531 were only mapped to the HPG1
332 pseudo-reference genome. Average coverage depth, number of covered genome positions, and number
333 of SNPs identified per accession relative to HPG1 are reported in Table S1. We also re-sequenced the
334 genomes of twelve Col-0 MA lines[50,51] (Table S2).

335 **Phylogenetic methods and genome-wide statistics**. We used four methods to estimate the relationships
336 among modern accessions, and between modern and herbarium HPG1 samples: (i) multidimensional
337 scaling (MDS); (ii) construction of a neighbor joining tree with the adegenet package in R (ref. [52]), with
338 branch support assessed with 1,000 bootstrap iterations and consensus reported; (iii) construction of a
339 parsimony network using SplitsTree v.4.12.3 (ref. [53]), with confidence values calculated with 1,000
340 bootstrap iterations; (iv) performing a Bayesian phylogenetic analysis using BEAST v.1.8 (ref. [54]) (see
341 below).

342 We estimated genetic diversity as Watterson´s θ (ref. [55]) and nucleotide diversity π, and the
343 difference between these two statistics as Tajimas's *D* (ref. [56]) using DnaSP v5 (ref. [57]). We calculated the
344 folded site frequency spectrum (SFS) as well as the unfolded SFS, for which we assigned the ancestral
345 state using the *A. lyrata* genome[58]. We estimated pairwise linkage disequilibrium (LD) between all
346 possible combinations of informative sites, ignoring singletons, by computing r^2, *D* and *D*' statistics. For
347 the modern individuals, we calculated the recombination parameter rho ($4N_er$) and performed the
348 four-gamete-test[59] to identify the minimum number of recombination events. All LD and recombination
349 related statistics were determined using DnaSP v5 (ref. [57]) (see SOM).

350 **Substitution and mutation rate analyses.** We used genome-wide nuclear SNPs to calculate pairwise
351 "net" genetic distances using the equation $D'_{ij} = D_{ic}-D_{jc}$, where D'_{ij} is the net distance between a modern
352 sample *i* and a herbarium 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 pairwise time distance in years between the collection times, *T*'ij, and calculated the linear regression: $D' = a+bT'$. The slope coefficient *b* describes the number of substitution changes per year. We used either all SNPs or subsets of SNPs at different annotations (genic, intergenic etc.) appropriately scaled by accessible genome length and with confidence intervals determined by bootstrap (see SOM and Fig. S3).

The second approach used Bayesian phylogenetics with the tip-calibration method implemented in BEAST v1.8 (ref. [54]). Our analysis optimized simultaneously and in an iterative fashion using a Monte Carlo Markov Chain (MCMC) a tree topology, branch length, substitution rate, and a demographic Skygrid model (Fig. 3 A,B; see SOM). The demographic model is a Bayesian nonparametric one that is optimized for multiple loci and that allows for complex demographic trajectories by estimating population sizes in time bins across the tree based on the number of coalescent - branching - events per bin (ref. [60]). We also performed a second analysis run using a fixed prior for substitution rate of $3*10^{-9}$ substitutions site^{-1} year^{-1} based on our previous net distance estimate to confirm that the MCMC had the same parameter convergence, e.g. tree topology, as in the first "estimate-all-parameters" run.

Inference of genome-wide selection. We separately analyzed sequences at different annotations, since certain regions regions should be under a different selection regime (less evolutionary constraint) than others. We compared the means and confidence intervals of substitution rates in the entire genome and in intergenic regions for both datasets, HPG1 population and laboratory Col-0 MA lines. Only within the HPG1 population, we also tested for an interaction between low and common allele frequency polymorphisms and putatively selected and putatively neutral annotations (comparisons: entire genome - intergenic, genic - intergenic, nonsynonymous - synonymous). The formal test was Fisher exact test and low and common frequency SNPs were defined in all possible cutoffs from 1 to 45% allele frequency (Fig. S5). The signal captured is based on the assumption that purifying selection is more efficient at intermediate frequencies, pushing deleterious variants towards lower frequency in the spectrum.

Association analyses and dating of new mutations. We collected flowering, seed and root morphology phenotypes for 63 accessions. For associations with climate parameters, we followed a similar rationale as described[61]. We extracted information from the bioclim database (http://www.worldclim.org/bioclim) at a 2.5 degrees resolution raster and intersected it with geographic locations of HPG1 samples (n = 100). We performed association analyses under several models and *p*-value corrections using the R package GeneABEL (ref. [62]), with phenotypes and climatic variables as response variables and SNPs as explanatory variables; appropriately correcting for covariates. Resulting p-values were adjusted with an empirical p-value distribution generated from 1,000 permuted datasets, or with a double Bonferroni correction: 5% / (number of SNPs + number of phenotypes tested).

Accession numbers. Short reads have been deposited in the European Nucleotide Archive under the accession number XXXXX.

METHODS REFERENCES

42. Yoshida, K. *et al.* The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine.

Elife **2,** e00731 (2013).
43. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** db.prot5448 (2010).
44. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38,** e87 (2010).
45. Kircher, M. in *Ancient DNA* (eds. Shapiro, B. & Hofreiter, M.) 197–228 (Humana Press, 2011).
46. Hagmann, J. *et al.* Century-scale Methylome Stability in a Recently Diverged Arabidopsis thaliana Lineage. *PLoS Genet.* **11,** e1004920–e1004920 (2015).
47. Ossowski, S. *et al.* Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.* **18,** 2024–2033 (2008).
48. Yoshida, K. *et al.* The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *eLife* **2,** e00731 (2013).
49. Schneeberger, K. *et al.* Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10,** R98 (2009).
50. Becker, C. *et al.* Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature* **480,** 245–249 (2011).
51. Shaw, R. G., Byers, D. L. & Darmo, E. Spontaneous mutational effects on reproductive traits of arabidopsis thaliana. *Genetics* **155,** 369–378 (2000).
52. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24,** 1403–1405 (2008).
53. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23,** 254–267 (2006).
54. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–1973 (2012).
55. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7,** 256–276 (1975).
56. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).
57. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25,** 1451–1452 (2009).
58. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43,** 476–481 (2011).
59. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111,** 147–164 (1985).
60. Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30,** 713–724 (2012).
61. Hancock, A. M. *et al.* Adaptation to climate across the Arabidopsis thaliana genome. *Science* **334,** 83–86 (2011).
62. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23,** 1294–1296 (2007).

434 # FIGURE LEGENDS

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

436 **(A)** Sampling location of herbarium (blue) and modern individuals (green). **(B)** Temporal distribution of
437 samples (random vertical jitter for visualization purposes). Stars indicate four herbarium accessions that
438 nest within the clade of modern accessions (see Fig. 3). **(C)** Linear regression of latitude and longitude as
439 a function of collection year (p-value of the slope and Pearson correlation coefficient are indicated)

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

441 **(A)** Neighbor joining tree, consensus of 1,000 bootstrap replicates. Branch lengths indicate number of
442 base substitutions (colors represent herbarium (blue) and modern individuals (green)). Scale line shows
443 the equivalent branch length of 80 nucleotide changes. Note that no outgroup was included. **(B)** First
444 two dimensions of a multidimensional scaling plot based on pairwise identity-by-state distances. Fraction
445 of variance explained given in parentheses. **(C, D)** Network of all samples using the parsimony splits
446 algorithm, before **(C)** and after **(D)** removing intra-HPG1 recombinants (in red).

447 **Figure 3. Substitution rates and demographic history.**

448 **(A)** Bayesian phylogenetic analyses employing the tip calibration methodology. A total of 10,000 trees
449 were superimposed as transparent lines, and the most common topology was plotted solidly. Tree
450 branches were calibrated with their corresponding collection dates. **(B)** Maximum Clade Credibility
451 (MCC) tree summarizing the trees in (A). The demographic model underlying the phylogenetic analysis,
452 Bayesian Skygrid reconstruction, is superimposed; the mean N_e over time is shown as a dotted line and
453 the 95% HPD is shaded grey. Note the scale line shows the equivalent branch length of 50 nucleotide
454 changes. **(C)** Regression between pairwise net genetic and time distances. The slope of the linear
455 regression line corresponds to the genome substitution rate per year. **(D)** Substitution spectra in HPG1
456 samples, compared to greenhouse-grown mutation accumulation (MA) lines. **(E)** Comparison of
457 genome-wide, intergenic, intronic, and genic substitution rates in HPG1 and mutation rates in
458 greenhouse-grown MA lines. Substitution rates for HPG1 were re-scaled to a per generation basis
459 assuming different generation times. Confidence intervals in HPG1 substitution rates were obtained from
460 95% confidence intervals of the slope from 1,000 bootstrap (see Table S4 for actual values).

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

462 Most SNPs first appeared in sample JK2530 collected 1922 in Indiana. For non-synonymous SNPs, the
463 amino acid change and the Grantham score (ranging from 0 to 215), which measures the
464 physico-chemical properties of the amino acids, are reported. All SNPs in the table were significant ($p <$
465 0.05) after raw p-values were corrected by an empirical p-value distribution from a permutation
466 procedure. * highlights those that also passed a double Bonferroni threshold, correcting by number of
467 SNPs and number of phenotypes ($p < 0.0001$). LD corresponds to how many other SNP hits are in high

468 linkage (r^2>0.5). See Table S5 for information on all significant SNPs and Table S4 for details on
469 phenotypes and climatic variables.

470

Figure 1

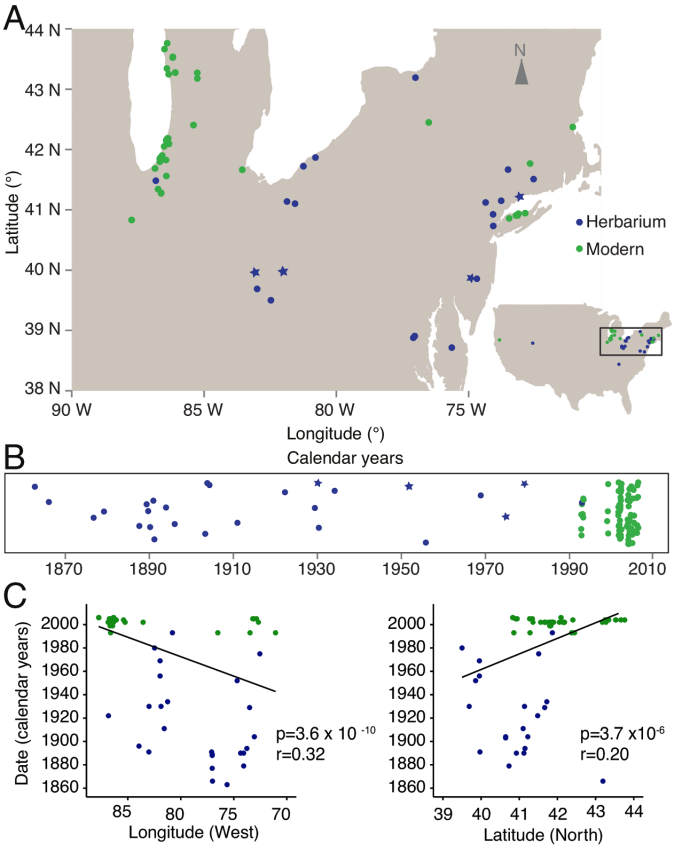

A
44 N
43 N
42 N
41 N
40 N
39 N
38 N
Latitude (°)
90 W
85 W
80 W
75 W
Longitude (°)
N
Herbarium
Modern
B
Calendar years
1870
1890
1910
1930
1950
1970
1990
2010
C
Date (calendar years)
2000
1980
1960
1940
1920
1900
1880
1860
p=3.6 x 10 -10
r=0.32
85
80
75
70
Longitude (West)
p=3.7 x10-6
r=0.20
39
40
41
42
43
44
Latitude (North)

Figure 2

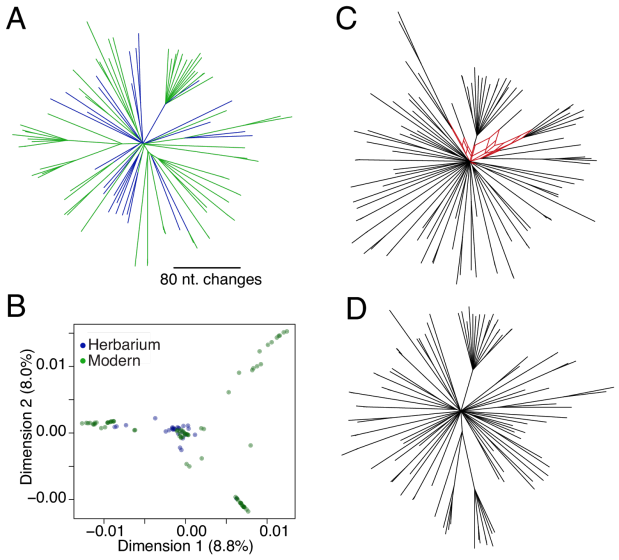
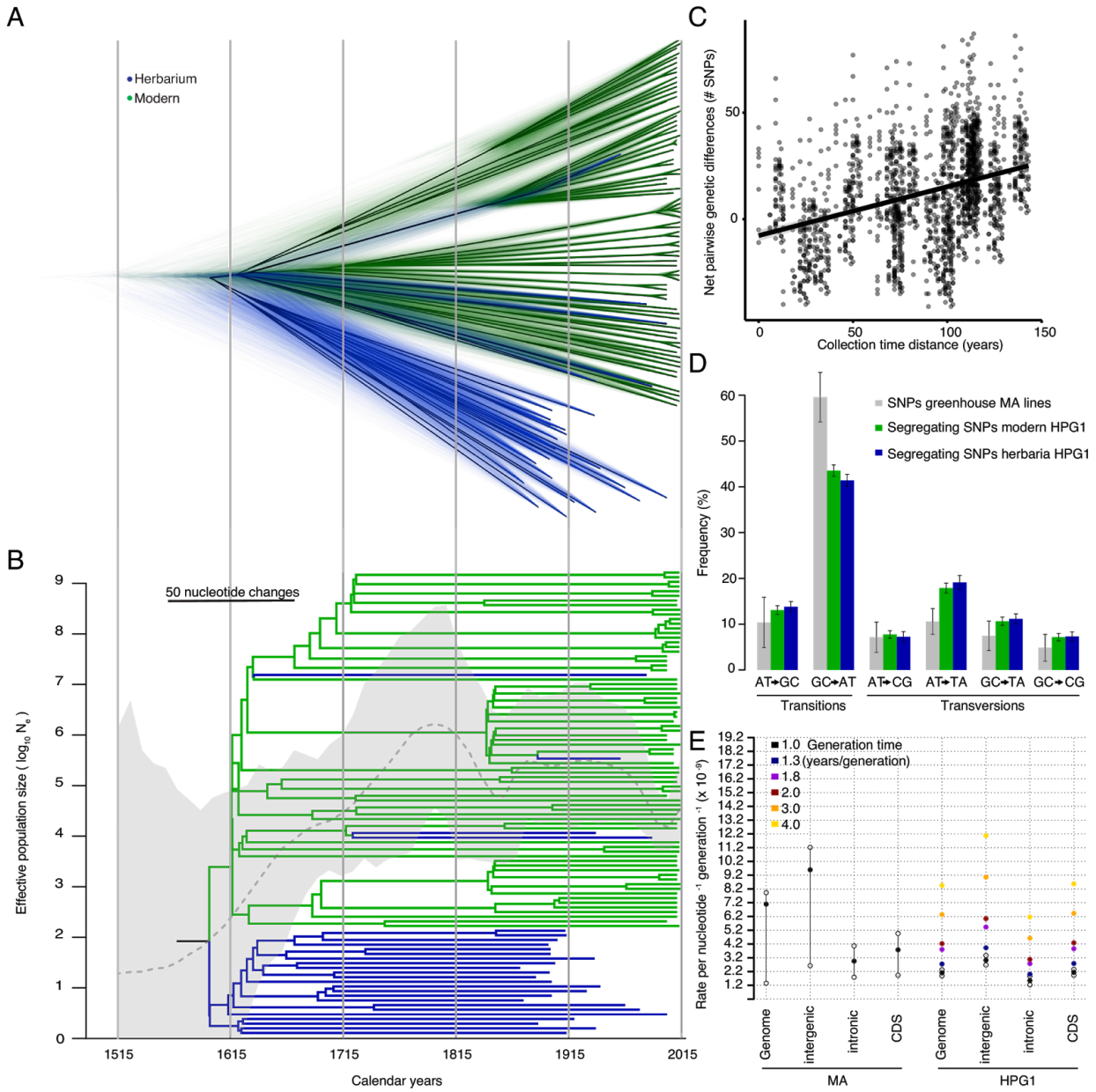

A
C
80 nt. changes
B
D
Herbarium
Modern
Dimension 2 (8.0%)
0.01
0.00
−0.00
−0.01
0.00
0.01
Dimension 1 (8.8%)

Figure 3


A
Herbarium
Modern
B
50 nucleotide changes
Effective population size ($\log_{10} N_e$)
Calendar years
1515
1615
1715
1815
1915
2015
C
Net pairwise genetic differences (# SNPs)
Collection time distance (years)
D
SNPs greenhouse MA lines
Segregating SNPs modern HPG1
Segregating SNPs herbaria HPG1
Frequency (%)
AT→GC
GC→AT
AT→CG
AT→TA
GC→TA
GC→CG
Transitions
Transversions
E
1.0 Generation time
1.3 (years/generation)
1.8
2.0
3.0
4.0
Rate per nucleotide $^{-1}$ generation $^{-1}$ (x 10^{-9})
Genome
intergenic
intronic
CDS
MA
HPG1

483

484

Table 1

485

Trait†	Location (chr-bp)	Gene	Anno–tation	Protein	aa change	LD	Bonf.
G	1-958,948	AT1G03810	nonsyn	Oligonucleotide/ oligosaccharide binding	A>P, 27	53	
D	1-13,994,958	AT1G36933	transposon	Copia		49	
S	1-20,324,050	AT1G54440	intronic	RRP6-LIKE 1		11	*
D	1-23,648,407	AT1G63740	nonsyn	TIR-NLR family	Y>S, 144	46	
G	2-358,395	AT2G01820	syn	RLK family		43	*
G	2-585,918	AT2G02220	syn	PSKR1		42	*
G	2-6,034,545	AT2G14247	syn	Expressed protein		38	*
G	2-7,047,529	AT2G16270	nonsyn	Unknown protein	P>A, 27	37	*
G	2-7,186,220	AT2G16580	intronic	SAUR8		36	*
G	2-10,495,275	AT2G24680	intronic	B3 family		34	*
G	2-12,415,084	AT2G28900	intronic	OEP16		32	
S	2-16,039,488	AT2G38290	3' UTR	AMT2		8	*
S	2-16,247,290	AT2G38910	nonsyn	CPK20	A>G, 60	7	*
G	2-16,333,662	AT2G39160	nonsyn	Unknown protein	A>G, 60	29	
G	3-2,500,258	AT3G07830	syn	PGA3		28	*
G	3-3,629,794	AT3G11530	intronic	VPS55		26	*
G	3-4,269,626	AT3G13229	5' UTR	DUF868 domain		25	*
D	3-11,873,293	AT3G30219	transposon	Gypsy		0	
G & D	4-4,228,138	AT4G07440	transposon	Oligonucleotide/ oligosaccharide binding		19	
G & D	4-9,046,942	AT4G15960	nonsyn	Alpha/beta-hydrolase superfamily	A>Q, 24	18	
G & D	4-15,646,341	AT4G32410	syn	ANY1		15	
G	4-15,845,001	AT4G32840	3' UTR	PFK6		14	
D	5-4,245,213	AT5G13260	syn	Unknown protein		12	
D	5-4,500,202	AT5G13950	nonsyn	Unknown protein	A>G, 60	11	
G	5-4,797,923	AT5G14830	transposon	Retrotransposon		10	
G	5-6,508,329	AT5G19330	nonsyn	ARIA	C>W, 215	0	
G	5-11,090,365	AT5G29037	transposon	Gypsy		4	
G	5-12,312,975	AT5G32630	pseudogene	–		3	
G	5-12,358,159	AT5G32825	transposon	CACTA		2	
S	5-16,024,197	AT5G40020	intronic	Thaumatin superfamily		2	*

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

Supplementary Information Guide for

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

SUPPLEMENTAL TEXT 3
1. Sample collection and preparation 3
2. Authenticity of aDNA 3
3. SNP calling thresholds 3
4. Resequencing of Col-0 Mutation Accumulation lines 3
5. Identification of bona fide HPG1 accessions and mutations 4
6. Extent of linkage disequilibrium and recombination 5
7. Substitution and mutation rate analyses 5
7.1 Greenhouse grown MA lines 5
7.2 Natural populations of HPG1 5
7.2.1 Net distances 5
7.2.2 Bayesian tip-calibration 6
7.2.3 Methylation status of mutated sites 6
8. Inference of genome-wide selection 7
9. Demography and migration of HPG1 8
9.1 Skygrid coalescent 8
9.2. Phylogeography 9
9.3. Isolation by distance 9
10. Phenotypic association analyses and dating of newly arisen mutations 9
10.1. Phenotyping 9
10.1.1 Root 9
10.1.2 Seed size 9
10.1.3 Flowering in the growth chamber 10
10.1.4 Fecundity in the field 10
10.2 Quantitative genetic analyses 11
10.2.1 Heritability 11
10.2.2 Linear Models 11
10.2.3 Evaluation of significance 12
10.2.4 Context of de novo mutations associated with phenotypes 12
10.2.5 Functional information 12
10.2.6 Proof of concept examples 13
REFERENCES 13
SUPPLEMENTAL FIGURES **17**
Figure S1. Ancient-DNA-like characteristics of unrepaired herbarium libraries. 17
Figure S2. Separation between HPG1 and other North American lineages. 18
Figure S3. Substitution spectrum and rates. 19
Figure S4. Relationship between methylation and substitutions. 21
Figure S5. Enrichment of low variants at putatively selected annotations. 23
Figure S6. Phylogeographic inference in HPG1. 24
Figure S7. Density of SNPs along all chromosomes and location of SNP hits. 26
Figure S8. Spatial and temporal emergence of root-associated mutations. 27
Figure S9. Linkage disequilibrium between SNPs with significant trait associations. 28
Figure S10. Correlations of SNP effects and p-values with frequency and age. 29
SUPPLEMENTAL TABLES 30

SUPPLEMENTAL TEXT

1. Sample collection and preparation

Seeds from modern accessions (Table S1) were bulked at the University of Chicago. Progeny for DNA extraction was grown at the Max Planck Institute for Developmental Biology. We used 2 to 8 mm^2 of dried tissue for destructive sampling from the herbarium specimens (Table S1).

2. Authenticity of aDNA

First, unrepaired sequencing herbarium libraries were screened for authenticity by sequencing at low coverage on Illumina HiSeq 2500 or MiSeq instruments. To verify the DNA retrieved from historical samples of *A. thaliana* was authentic, we checked the percentage of endogenous DNA of the sample (Fig. S1A) as well as typical postmortem DNA damages: high fragmentation of DNA (Fig. S1B), enrichment of substitution from C to T at the first base pair (Fig. S1C) as well as purine enrichment at breakpoints of DNA fragments (Fig. S1D) (for details see [1]). Sequencing to produce the final genomes (101 bp paired end) was carried out on an Illumina HiSeq 2000 instrument after DNA repair by uracil-DNA glycosylase[2–4].

3. SNP calling thresholds

To asses the effect of SNP calling thresholds on the mutation rate, we employed three different SHORE v0.9.0 quality thresholds following previous work (see Table S4 from ref. [5]): allowing at most one intermediate penalty in all strains (most stringent threshold; “32-32”); requesting that at least one strain had at most one intermediate penalty, while all others were allowed up to two high and one intermediate penalties (intermediate stringency, “32-15”); and finally allowing one high and one intermediate penalty for all strains (most lenient stringency, “24-24”). On top of that, we would either allow missing information per SNP in up to 50% of accessions, or request complete information (0% missing rate). Thus, the most rigorous case would be 32-32 quality and 0% missing rate, and the most relaxed 24-24 quality and 50% maximum missing rate. Substitution rate calculations (section 7.2) were done for datasets from all combinations of these quality parameters (Fig. S3), and we chose the regular 32_15 quality threshold and complete information for the final estimate (Fig 3 C, E).

4. Resequencing of Col-0 Mutation Accumulation lines

We also sequenced the genomes of twelve greenhouse-grown mutation accumulation (MA) lines, including ten that had been sequenced at lower coverage before[5,6] (Table S2). We called SNPs, indels and structural variants (SVs), following the workflow and parameters described[7], but without iterations. This procedure resulted in 2,203 polymorphisms shared by all lines, indicating errors in the

reference sequence (12% of variants replaced N's in the TAIR9 genome) or genetic differences in the founder plant of the MA population compared to the Col-0 reference genome. In addition, we identified 388 segregating variants across the twelve lines (Table S2), of which 350 were singletons. This analysis revealed on average 25.5 SNPs, 4.9 deletions and 3.2 insertions per MA line at the 31st generation (Table S2), compared to 19.6 SNPs, 2.4 deletions and 1.0 insertions previously detected in the 30th generation with shorter read length and lower read depth[8]. The genome length accessed in this sequencing effort, 115,954,227 bp, was used to scale the number of point mutations to a rate of 7.1 x 10^{-9} mutations site^{-1} generation^{-1} (Table S3, Fig. 3E).

5. Identification of *bona fide* HPG1 accessions and mutations

Before we could work with the colonizer group HPG1, we needed to carefully identify individuals that belong to other haplogroups or that have introgressions from them. We established the relationships among all samples at three levels of resolution: (i) the 149 nuclear SNPs used originally to define the HPG1 haplogroup in a global screening[9] (Fig. S2A), (ii) SNPs in the chloroplast genome (where we did not find any variants within HPG1), (iii) and all nuclear genome SNPs (Fig. S2B-C). At these three levels we performed a multidimensional scaling (MDS) analysis and built a neighbor-joining tree using the adegenet package in R (ref. [10]).

Having identified these *bona fide* HPG1 individuals, we wanted to confirm that the diversity has a legitimate origin from *de novo* mutations. For that we used the 1001 Genomes resource (1001genomes.org) to verify that the majority of HPG1-specific variants did not originate in the native Eurasian range. Subsetting the genomes from this resource to only European accessions, and limiting the SNP set to those with ≥1% frequency of alternative alleles, there were 338 variants out of all 5,181 HPG1 variants that were also found in Europe or Asia (6%). Only one of the reported SNPs associated with phenotypes (see section 10) was among these shared variants.

There are several scenarios that can explain these shared SNPs. One is that some HPG1 individuals were moved back to Europe by humans. Another one is that parallel mutations occurred in North America and Eurasia or that a reversion-mutation happened in some HPG1 individuals. Given that 10% of all sites in the genome are variable in the 1001 Genomes collection, this is not an implausible scenario for at least a fraction of shared SNPs. Two additional scenarios involve an origin from standing European variation: (1) the shared variants come from small introgression events that passed our filters above, or (2) the bottleneck was not complete, and while it left no diversity in the chloroplast genome, a few hundred SNPs were passed on in the nuclear genome (given the low number of variants, the colonizers could have been as many as two dozen seeds)

6. Extent of linkage disequilibrium and recombination

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. Linkage disequilibrium parameter *D*' did not decay with physical distance (intercept = 0.99, slope = 0.00, $p < 0.0001$) among all SNP pairs. Furthermore, only 0.02% of the nonsingleton SNP pairs were not in complete linkage disequilibrium (D'<1*)*, indicating extensive linkage between chromosomes. We also formally estimated recombination within HPG1. The estimate was much lower ($4N_er = \rho = 3.0 \times 10^{-6}$ cM bp^{-1}) than for a similar-sized collection of diverse *A. thaliana* individuals from the native range[11] ($\rho = 7.5 \times 10^{-2}$ cM bp^{-1}). The four-gamete test[12], which determines whether all four possible gametes (ab, aB, Ab, AB) are observed for two segregating loci, revealed that all configurations of SNPs could be explained with as few as 38 recombination events for the 100 genomes. We argue that this number of potential recombination events is sufficiently small to use phylogenetic methods with the 100 HPG1 genomes, even though such methods are normally not appropriate for genome-wide analyses. Indeed, other sources of failure of the four-gamete test and the violation of phylogenetic assumptions could be sequencing errors, or lineage sorting of segregating sites from the ancestral population. LD and recombination related statistics were determined using DnaSP v5 (ref. [13]) or plink v1.90b2n (ref.[14]).

7. Substitution and mutation rate analyses

7.1 Greenhouse grown MA lines

Mutation rates were estimated for each 31st generation greenhouse-grown MA line[5] as the number of mutations divided by the total bp length of the genome (or a given annotation) and by 31 generations (the two MA lines with only three generations were excluded from this analysis). Mean and confidence intervals across lines are reported (Table S3). 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).

7.2 Natural populations of HPG1

7.2.1 Net distances

For the "net genetic distances" method, we computed confidence intervals of the *b* regression slope coefficient ($D' = a+bT'$) using a bootstrap with replacement of 1,000 samples to avoid over-confident confidence intervals due to lack of independence of points[15]. We used either all SNPs or SNPs at specific annotations to calculate different substitution rates and scaled the slope into a per-base rate using all positions (of the given annotation) that passed alternative or reference call quality

thresholds rather than using a single value of genome length (Table S3). For all annotations we calculated substitution rates with three quality thresholds and either full information per SNP or allowing a maximum of 50% missing accessions per SNP (see section 3 and Fig. S3C). For some annotations substitution rates were not reliable. For instance, in 3' and 5' UTR regions, we did not have enough mutations (on average ~1 SNP difference between any pair), and thus do not report these regions' rates. Transposons showed unstable substitution rate estimates that we attribute to structural variation relative to the reference genome. This likely decreased our ability to map transposon reads correctly and subsequently call SNPs[16]. In contrast, in the MA dataset, transposon structural variation was probably fairly low since only 31 generation separate the Col-0 reference genome with each of the ten derived MA lines. This can be the reason that the number of transposon SNPs identified in the MA dataset is proportionally larger than in HPG1 (Table S2 and S3). Therefore, transposon substitution rates in HPG1 cannot be trusted.

7.2.2 Bayesian tip-calibration

For the second approach to estimate a substitution rate, the Bayesian phylogenetics tip-calibration approach, we performed systematic runs and chain convergence assessments of different demographic and molecular clock models. We found the Skygrid demographic model[17] and the lognormal relaxed molecular clock[18] [18] the most appropriate models. Under a relaxed molecular clock, the substitution rate is allowed to vary across branches with a lognormal distribution. The prior used for molecular clock was a Continuous-Time Markov Chain (CTMC)[17,19]. The analysis was carried out remotely at CIPRES PORTAL (v3.1 www.phylo.org) using uninformative priors. The run took about 1,344 CPU hours and performed 1,000 million steps in a Monte Carlo Markov Chain (MCMC), sampling every 100,000 steps. Burn-in was adjusted to 10% of the steps. To visualize the tree output we produced a Maximum Clade Credibility (MCC) tree with a minimum posterior probability threshold of 0.8 and a 10% burn-in using TreeAnnotator (part of BEAST package), and visualized the MCC tree using FigTree (tree.bio.ed.ac.uk/software/figtree/) (Fig. 3B). Additionally, we used DensiTree[20] to simultaneously draw the 10,000 BEAST trees with the highest posterior probability (Fig. 3A). Since all trees were drawn transparently, agreements in both topology and branch lengths appear as densely colored regions, while areas with little agreement appear lighter.

7.2.3 Methylation status of mutated sites

As in many other species, the spectrum of *de novo* mutations in the greenhouse-grown *A. thaliana* MA lines is biased towards G:C→A:T transitions[8], leading to an inflated transition-to-transversion ratio (Ts/Tv). This bias is less pronounced in recent mutations in a Eurasian collection of natural accessions[21] and in HPG1 accessions (Fig. 3D). A recent multigenerational salt stress experiment in

the greenhouse also showed a more balanced Ts/Tv (ref. [22]). These findings indicate that less benign conditions might promote a lower Ts/Tv, and one possible cause are methylation patterns, known to change under different environments[23].

We interrogated the potential evolutionary role of cytosine methylation in the mutability of cytosine bases in the HPG1 accessions. For reference DNA methylation data, we used previously generated bisulfite-sequencing data of HPG1 strains[7] and of Col-0 MA lines[5], respectively. For both datasets, methylation status was calculated as the fraction of reads with methylated cytosines by the total number of reads at a certain cytosine position in the genome. Our rationale was that if methylation affected mutability, the degree of methylation at positions were we find a new mutation should be higher. To be sure that a given site in HPG1 was a new mutation, we only considered positions for which we could determine that state by alignment to the *A. lyrata* genome[24]. The "tested sites" were positions in HPG1 that had a mutation both from *A. lyrata* and *A. thaliana* Col-0. These positions can be of two kinds, "fixed" if all HPG1 individuals carry the alternative, or "segregating" if both reference and alternative alleles exist in HPG1. As control, "control set", we used cytosine positions that did not vary across HPG1, *A. lyrata* and *A. thaliana*. To produce the methylation distribution of the control set we randomly chose 1,000 invariant cytosine positions. For the test sets, we averaged the methylation degree and compared it with the control distribution.

Ancestral cytosines with higher methylation in both *A. thaliana* Col-0 reference and HPG1 pseudo-reference methylome datasets were more likely to mutate to thymines in HPG1 (Fig. S4 A-D). Additionally, the methylation degree at substitutions inside genes was higher in the HPG1 methylome (Fig. S4 B,D). While some C→T changes could be explained by higher spontaneous deaminations known to happen more often at methylated cytosines, also C→A/G substitutions were more likely to have been methylated. If this process is common enough, the Ts/Tv ratio should decrease. We are far from understanding differences in Ts/Tv in natural and controlled conditions, but definitely methylation status seems to have a strong statistical connection with mutability.

8. Inference of genome-wide selection

Since we observed differences between the two mutation accumulation (MA) datasets (the laboratory Col-0 MA lines and the wild HPG1 lines), we tried to infer selection based on differences in polymorphisms and substitution rates. We compared the different substitution rates and the 95% bootstrap confidence intervals to assess how identical they were (Table S3). Genome-wide substitution rate in the HPG1 dataset was significantly lower than that in controlled greenhouse conditions, even after correcting by the mean generation time of 1.3 years[25] (Table S3). However, these differences could be due to differences in individual genomic annotations. For instance, coding regions and introns were virtually identical between the two datasets, but transposons and

222 intergenic were much lower in HPG1. Furthermore, if the true generation time is about 2-3 years (Fig.
223 3), differences in the mentioned annotations would also disappear. Therefore, we tried to investigate
224 the existence of purifying selection based on frequency equilibriums only within the HPG1
225 population. We did comparisons between pairs of annotation categories and between pairs of
226 frequency classes (i.e., low and common). In this way, genome-wide, genic and non-synonymous
227 polymorphisms were compared to matched putatively neutral annotations: intergenic and
228 synonymous sites (Fig. S5). We tested for an interaction using a Fisher's Exact test on 2x2 table
229 counts as:
230

Neutral annotation & low frequency	Neutral annotation & common frequency
Selected annotation & low frequency	Selected annotation & common frequency

235
236 Because any frequency cut-off is arbitrary, we computed the test with all cutoffs from 0 to
237 50% frequencies in 1% steps (Fig. S5). This test, which resembles in concept the MK or HKA tests,
238 evidences that there is a depletion of common frequency polymorphisms, which are more exposed
239 to selection, at the three putatively selected genomic levels compared to control (quasi-neutral)
240 regions: genome-wide, genes and nonsynonymous sites.
241
242 **9. Demography and migration of HPG1**
243 9.1 Skygrid coalescent
244 From the Bayesian phylogenetic analyses described previously (section 7.2.2), we studied the
245 demographic model estimated via Skygrid. We reconstructed a skyline plot that depicts changes in
246 effective population size, a measure of relative diversity, through time[26] (Fig. 3B). Sampling biases
247 could produce artefactual effects in the Skygrid plot. Nevertheless we expect these to be minor since
248 our dataset has a continuous sampling over a century (>2 samples per decade) instead of a
249 two-timepoints sampling, as it is common in ancient DNA studies where the rarity of the samples are
250 a limiting factor[27]. An additional BEAST run was performed only with modern samples to verify that
251 the corresponding part of the tree and population sizes matched (data not shown). Implementation
252 of non-phylogenetic methodologies for demographic inference exist, e.g. Multiple Sequentially
253 Markovian Coalescent (MSMC)[28], but after exploring them we concluded their resolution was
254 insufficient for analyses of the last several centuries. In order to compare with another method, we
255 got rough estimates of the diversity per decade as the average genetic differences between any two
256 samples per decade (Fig S6C).
257

258 9.2. Phylogeography

259 We performed another Bayesian phylogenetic analysis incorporating a geographic location trait[29,30].
260 For this, Brownian diffusion parameters are estimated by fitting a continuous gradient of geographic
261 locations along tree branches, starting from the leaves of the tree for which geographic locations are
262 known, i.e., the collection sites of our samples. We excluded three samples from the West coast of
263 the United States, separated over three thousand kilometers from the rest, since these do not fit a
264 gradual propagation by Brownian diffusion. We ran this analysis with the parameters described
265 previously (section 7.2.2) and sliced the resulting 3D (temporal and geographical) phylogeny at the
266 early 16th and late 18th century using SPREAD software[31] (Fig. S6B). Similar to before, we roughly
267 estimated "local diversity" for each sampling location by computing the average genetic differences
268 to the 10 closest neighbours (Fig. S6D).
269

270 9.3. Isolation by distance

271 We employed a heuristic search[32] using an isolation-by-distance pattern to find the origin of
272 diffusion of HPG1 in North America, and compared it to the phylogeography analyses. We performed
273 a regression between genetic distances on geographic distances for all pairs of samples (*genetic*
274 *distance ~ Euclidean geographic distances*). This pattern, known as isolation by distance pattern,
275 reflects that as individuals are more geographically apart, they differ more genetically. We evaluated
276 whether this relationship was still significant for each of our samples separately (i.e., from a focal
277 sample, does genetic distance increase as geographic distance increases?). Only the significant
278 samples were retained and plotted since those points are the expected origins of migrations (Fig. S6
279 E,F). Arrows can be plotted in the direction of the maximum slope to illustrate migration trajectories.
280 This was done separately for historic and modern samples.
281

282 **10. Phenotypic association analyses and dating of newly arisen mutations**

283 10.1. Phenotyping

284 *10.1.1 Root*

285 Fifteen root phenotypes were scored for ≥ 10 replicates per genotype over a time-series experiment
286 at the Gregor Mendel Institute in Vienna, using image analysis as described in detail elsewhere[33]. We
287 used the means per genotypes and per time series for association analyses.
288

289 *10.1.2 Seed size*

290 We spread the seeds of given genotypes on separate plastic square 12 x 12 cm Petri dishes. For faster
291 image acquisition we used a cluster of eight Epson V600 scanners. The scanner cluster was operated
292 by the BRAT Multiscan image acquisition tool

(www.gmi.oeaw.ac.at/research-groups/wolfgang-busch/resources/brat/). The resulting 1600 dpi images were analyzed in Fiji software. Scans were converted to 8-bit binary images, thresholded (parameters: setAutoThreshold("Default dark"); setThreshold(20, 255)) and particles analyzed (inclusion parameters: size=0.04-0.25 circularity=0.70-1.00). The 2D seed size was measured in square millimeters (parameters: distance=1600 known=25.4 pixel=1 unit=mm) for 2 plants per genotype, > 500 seeds per plant.

10.1.3 Flowering in the growth chamber

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

10.1.4 Fecundity in the field

To investigate variation in fecundity in natural conditions, we grew three replicates of each accession in a field experiment following a completely randomized block design. Seeds were sown from 09/20/2012 to 09/22/2012 in 66-well trays (well diameter = 4 cm) 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 (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° N., 86.63° W.), Michigan, USA on 10/13/2012 and 10/14/2012. Seedlings were watered-in and left to overwinter without further intervention. 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).

328

329 10.2 Quantitative genetic analyses

330 For 63 modern accessions, we measured time to bolting and flowering, seeds per plant, seed size,
331 and 15 root phenotypes in common chamber or common garden settings. For all 100 accessions,
332 climatic information from the bioclim database (http://www.worldclim.org/bioclim) was extracted
333 using their geographic coordinates. For historic samples, some locations were only known by county
334 name. In this case we assigned the geographic coordinate location of the centroid of the county.

335

336 *10.2.1 Heritability*

337 We performed association analyses using the R package GenABEL[34], with measured phenotypes (p =
338 25) and climatic variables (c = 18) as response variables and SNPs as explanatory variables. A
339 Minimum Allele Frequency (MAF) cutoff of 5% was used. The number of assessed SNPs was 391 in a
340 dataset of only modern samples but with imputed genotypes for missing data using Beagle v4.0 (ref.
341 [35]), and 456 SNPs with a dataset of modern and historic samples, without imputation. For all
342 associations, at least 63 individuals were genotyped for a specific SNP. We first investigated broad
343 sense heritability (H^2) of each trait using ANOVA partition of variance between and within lines using
344 replicates (Table S4). Significance was obtained by common F test in ANOVA. Secondly we used the
345 *polygenic_hglm* function to fit a genome wide kinship matrix to calculate a narrow sense heritability
346 estimate (h^2). Significance was calculated employing a likelihood ratio test comparing with a null
347 model. In principle, h^2 is a component of H^2, then its values should theoretically be $h^2 < H^2$; that is
348 not our case. Our result cannot be interpreted in this framework, since the calculation of both was
349 not done with the same samples: for the h^2 calculation we employed genotype means whereas for
350 the H^2 we used multiple replicated measurements per genotype. The averaging of replicates per
351 genotype in h^2 reduced environmental and developmental noise and thus we would expect $h^2>H^2$.
352 We did this so the climatic estimates of h^2, for which we only have one value per genotype, would be
353 comparable with the phenotypic h2 ones (Table S4).

354

355 *10.2.2 Linear Models*

356 For association analyses we first employed a linear mixed model that fitted the kinship matrix using
357 the *mmscore* function, and only three significant SNP hits were discovered using a 5% significance
358 threshold after False Discovery Rate correction (FDR). This was expected since we have few variants
359 and these would have originated in an approximated phylogeny structure. We concluded that fitting
360 the kinship matrix in our model was not appropriate since there would be no residual variation for
361 association with specific SNPs. With this rationale we employed a fixed effects linear model using the
362 *qtscore* function[36]. To reduce the false-positive rate we took a conservative permutation strategy by

363 carrying out association with over 1,000 randomized datasets (permuting phenotypes across
364 individuals) and used the resulting p-value distribution to correct p-values estimated with the original
365 dataset. SNPs with p-values below 5% in the empirical p-value distribution were considered
366 significant (Table S5). In climatic models, we included longitude and latitude as covariates to correct
367 for any spurious association between SNPs and climate gradients created by the migratory pattern of
368 isolation by distance.
369
370 *10.2.3 Evaluation of significance*
371 Significant SNPs were interspersed throughout the genome (Fig. S7) and their p-values and
372 phenotypic effects did not correlate with the minimum age of the SNPs nor with their allele
373 frequency (Fig. S10), something that could have indicated that the significance was merely driven by
374 the higher statistical power of intermediate frequency variants. Using QQ plots to assess inflation or
375 deflation of p-values, we observed generally that permutation corrected p-values were deflated.
376 Straight series of points in QQ plots indicate identical p-values for multiple SNPs, a pattern that we
377 attributed to long range LD, i.e. lack of independence (see Graphic Table S7 for trait distributions and
378 QQ plots from each association analysis). We also used a False Discovery Rate correction for the raw
379 p-values using *p.adjust* in R and, as a sanity check, we used a Bonferroni-corrected threshold, a
380 procedure considered over-stringent in association analyses (Table S5). This was calculated as: 5% /
381 (number of SNPs + number of traits) ~ 0.01%.
382
383 *10.2.4 Context of de novo mutations associated with phenotypes*
384 For each SNP in our dataset, we determined the ancestral and derived states, by identifying which
385 allele was found in the oldest herbarium samples. We compared the time of emergence and the
386 centroid of geographic distribution of the alternative alleles of SNP hits to random draws of SNPs
387 with the same MAF filtering (5%) (Fig. S8).
388
389 *10.2.5 Functional information*
390 On top of phenotypic and climatic associations of SNP hits, we also provide a likely functional effect
391 employing a commonly used amino acid matrix of biochemical effects[37]. Functional information of
392 gene name and ontology categorization of SNP hits was obtained from
393 www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/annotation_data.jsp and
394 www.arabidopsis.org/tools/bulk/go/ (Table 1 and Table S5).

395

396 *10.2.6 Proof of concept examples*

397 We argue that the power of an association approach relies on the fact that HPG1 lines resemble Near
398 Isogenic Lines (NILs) produced by experimental crosses[38] (Fig. S9A). Similar to genome-wide
399 association studies (GWA), power depends on many factors, namely the noise of phenotype under
400 study, architecture of phenotypic trait, quality of genotyping, population structure, sample diversity,
401 sample size, allele frequency, and recombination. On one hand, association analyses in NILs suffer
402 from large linkage blocks, but confident results can be achieved due to accurate measurement of
403 phenotypes, limited genetic differences between any two lines, and high quality genotypes. In
404 common GWA studies such as in humans, there are multiple confounding effects. Among the
405 confounders are (1) that any two samples differ in hundreds of thousands of SNPs, and (2) that
406 historical and geographic stratification produce non-random correlations among those SNP
407 differences. This considerably complicates the identification of phenotypic effects at specific genes,
408 and power relies greatly on large sample sizes to achieve the sufficient number of recombination
409 between markers.

410 To provide support for the non-synonymous SNP on chromosome 5, at position 6,508,329 in
411 AT5G19330, we looked for pairs of lines that carry the ancestral and the derived allele, but that differ
412 in few (or no other) SNPs in the genome. When considering all genic substitutions with a minimum
413 allele frequency of 5% (Fig. S9A), we identified 20 pairs of lines differing only in the AT5G19330 SNP
414 and another linked SNP (located on a different chromosome, association p-value > 0.4). The
415 phenotypic differences in mean gravitropic score of these almost-identical pairs were significantly
416 higher than phenotypic differences among all pairs of HPG1 lines, and genetically identical pairs
417 attending to substitutions inside genes (Fig. S9A). Furthermore, this SNP was not in complete linkage
418 with any other SNP hit ($r^2 < 0.5$) (Fig. S7D). The same approach was used to examine the SNPs in
419 AT1G54440 (Fig. S7E) and AT2G16580 (Fig. S7F), which represent an intermediate and a high LD
420 example.

421

REFERENCES

422
1. Weiß, C. L. *et al.* Temporal patterns of damage and decay kinetics of DNA retrieved from plant
423 herbarium specimens. *Royal Society Open Science* **3,** 160239 (2016).
424
2. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target
425 capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** db.prot5448 (2010).
426
3. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in
427 ancient DNA. *Nucleic Acids Res.* **38,** e87 (2010).

428 4. Kircher, M. in *Ancient DNA* (eds. Shapiro, B. & Hofreiter, M.) 197–228 (Humana Press, 2011).
429 5. Becker, C. *et al.* Spontaneous epigenetic variation in the Arabidopsis thaliana methylome.
430 *Nature* **480,** 245–249 (2011).
431 6. Shaw, R. G., Byers, D. L. & Darmo, E. Spontaneous mutational effects on reproductive traits of
432 arabidopsis thaliana. *Genetics* **155,** 369–378 (2000).
433 7. Hagmann, J. *et al.* Century-scale Methylome Stability in a Recently Diverged Arabidopsis
434 thaliana Lineage. *PLoS Genet.* **11,** e1004920–e1004920 (2015).
435 8. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in Arabidopsis
436 thaliana. *Science* **327,** 92–94 (2010).
437 9. Platt, A. *et al.* The scale of population structure in Arabidopsis thaliana. *PLoS Genet.* **6,**
438 e1000843 (2010).
439 10. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers.
440 *Bioinformatics* **24,** 1403–1405 (2008).
441 11. Choi, K. *et al.* Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene
442 promoters. *Nat. Genet.* **45,** 1327–1336 (2013).
443 12. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the
444 history of a sample of DNA sequences. *Genetics* **111,** 147–164 (1985).
445 13. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism
446 data. *Bioinformatics* **25,** 1451–1452 (2009).
447 14. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage
448 analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
449 15. Drummond, A., Pybus, O. G. & Rambaut, A. Inference of viral evolutionary rates from molecular
450 sequences. *Adv. Parasitol.* **54,** 331–358 (2003).
451 16. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational
452 challenges and solutions. *Nat. Rev. Genet.* **13,** 36–46 (2012).
453 17. Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model
454 for multiple loci. *Mol. Biol. Evol.* **30,** 713–724 (2012).
455 18. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating
456 with confidence. *PLoS Biol.* **4,** e88–e88 (2006).
457 19. Ferreira, M. a. R. & Suchard, M. a. Bayesian analysis of elapsed times in continuous-time
458 Markov chains. *Can. J. Stat.* **36,** 355–368 (2008).
459 20. Bouckaert, R. R. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26,**
460 1372–1373 (2010).
461 21. Cao, J. *et al.* Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat.*
462 *Genet.* **43,** 956–963 (2011).

22. Jiang, C. *et al.* Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. *Genome Res.* **24,** 1821–1829 (2014).
23. Wibowo, A. *et al.* Hyperosmotic stress memory in Arabidopsis is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. *eLife Sciences* **5,** e13546 (2016).
24. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43,** 476–481 (2011).
25. Falahati-Anbaran, M., Lundemo, S. & Stenøien, H. K. Seed dispersal in time can counteract the effect of gene flow between natural populations of Arabidopsis thaliana. *New Phytol.* **202,** 1043–1054 (2014).
26. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–1973 (2012).
27. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23,** 553–559 (2013).
28. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46,** 919–925 (2014).
29. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27,** 1877–1885 (2010).
30. Barton, N. H. & Wilson, I. Genealogies and geography. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **349,** 49–59 (1995).
31. Bielejec, F., Rambaut, A., Suchard, M. A. & Lemey, P. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics* **27,** 2910–2912 (2011).
32. Handley, L. J. L., Manica, A., Goudet, J. & Balloux, F. Going the distance: human population genetics in a clinal world. *Trends Genet.* **23,** 432–439 (2007).
33. Slovak, R. *et al.* A Scalable Open-Source Pipeline for Large-Scale Root Phenotyping of Arabidopsis. *Plant Cell* **26,** 2390–2403 (2014).
34. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23,** 1294–1296 (2007).
35. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* **98,** 116–126 (2016).
36. Aulchenko, Y. S., de Koning, D.-J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177,** 577–585 (2007).
37. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185,** 862–864 (1974).

38. Weigel, D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol.* **158,** 2–22 (2012).

SUPPLEMENTAL FIGURES

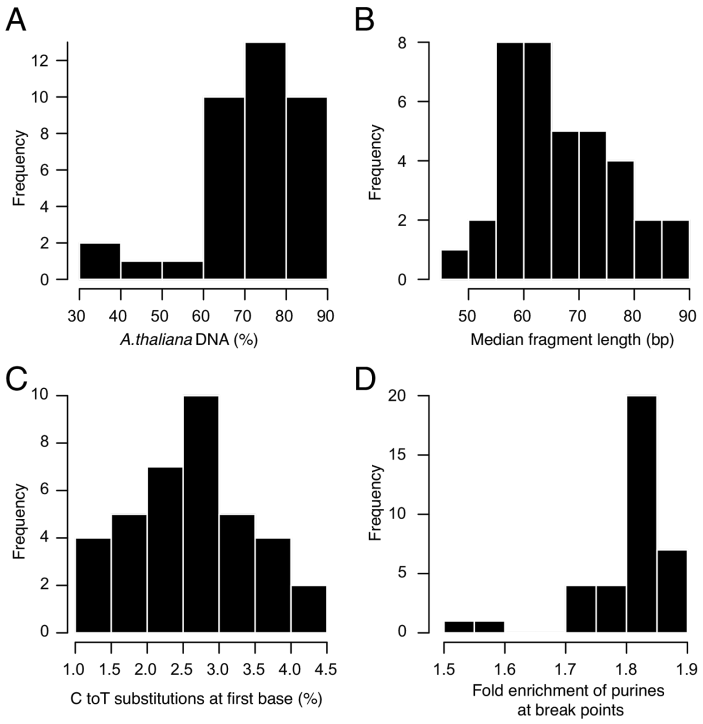


Figure S1. Ancient-DNA-like characteristics of unrepaired herbarium libraries.

(A) Fraction of *A. thaliana* DNA in sample. **(B)** Median length of merged reads. **(C)** Fraction of cytosine to thymine (C-to-T) substitutions at first base (5' end). **(D)** Relative enrichment of purines (adenine and guanine) at 5' end breaking points. Position -1 is compared with position -5 (negative numbers indicate genomic context before upstream reads' 5' end).

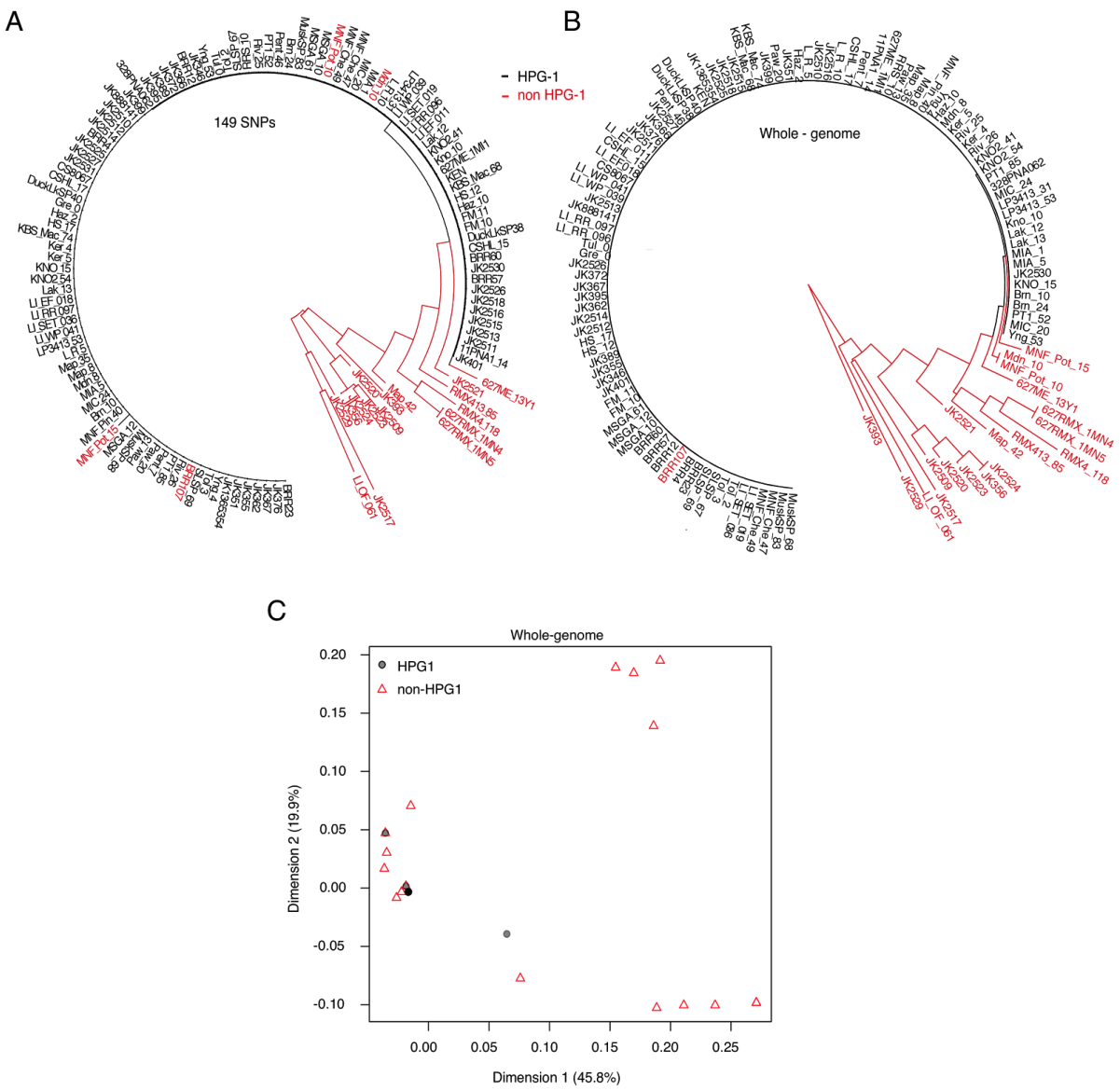


Figure S2. Separation between HPG1 and other North American lineages.

(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 (see explanation below for four HPG1-like accessions). **(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 (BRR7) within the HPG1 main branch was a recombinant removed from the analysis. **(C)** First two dimensions of a multidimensional scaling plot based on identity-by-state pairwise distances. Notice that black dots represent multiple transparent dots overlaid, a result of multiple almost-identical HPG1 genomes. Percentage of the variance explained by each dimension given in parentheses.

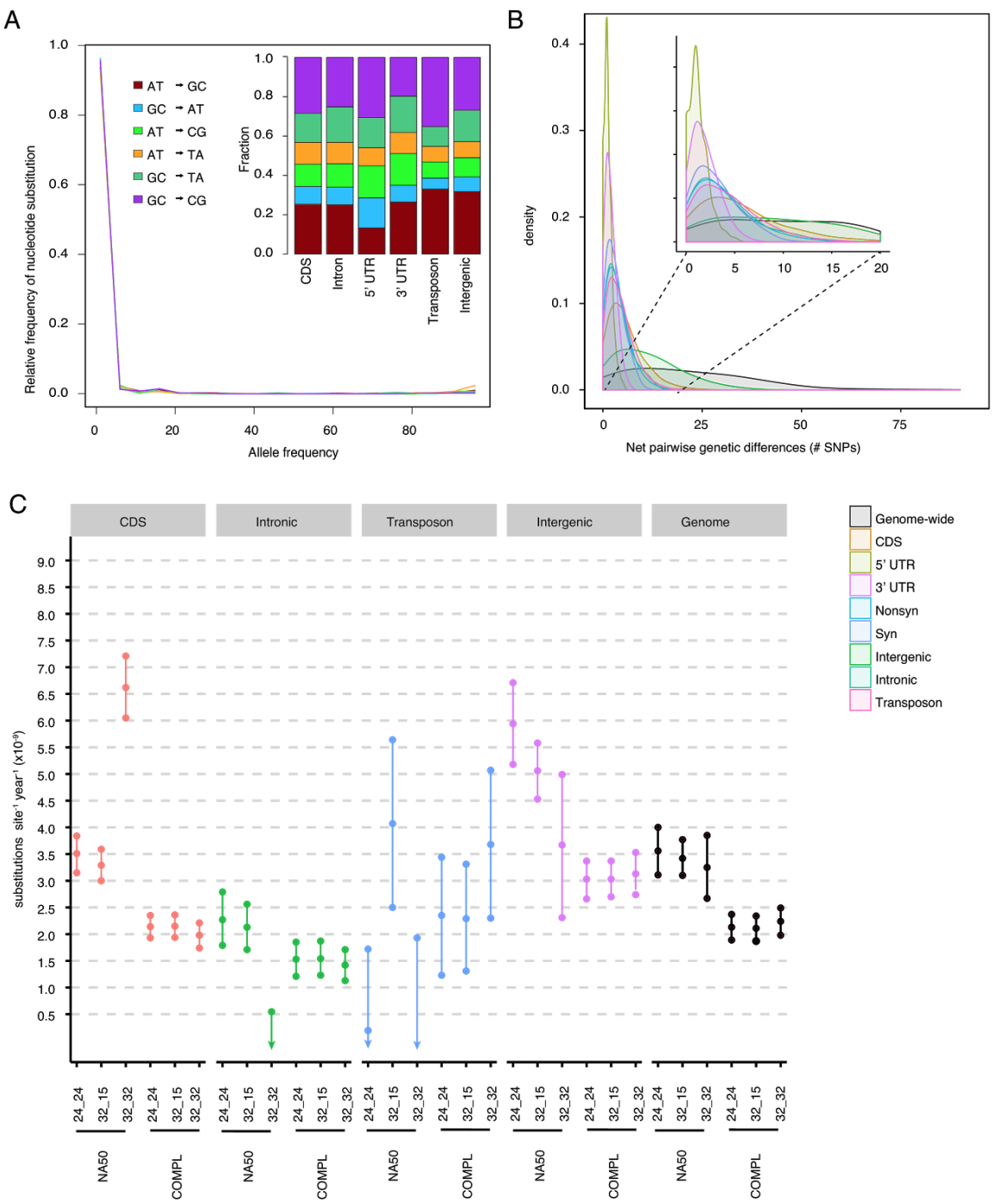


Figure S3. Substitution spectrum and rates.

(A) "Unfolded" site frequency spectrum using *Arabidopsis lyrata* as outgroup for all transitions and transversions. Bar plot shows proportions of different types of substitutions divided by genomic annotation. **(B)** Distributions of "net" pairwise genetic distances between historic and modern samples used to calculate mutation rates (from quality 32_15 and complete information per site). UTRs were excluded because of the small number of SNPs. **(C)** Mutation rates calculated for different genomic annotations and quality thresholds (32_32, 32_15, 24_24) and missing values (NA50:

maximum 50% missing data per SNP; COMPL: missing data 0%). Mean and 95% confidence intervals are shown.

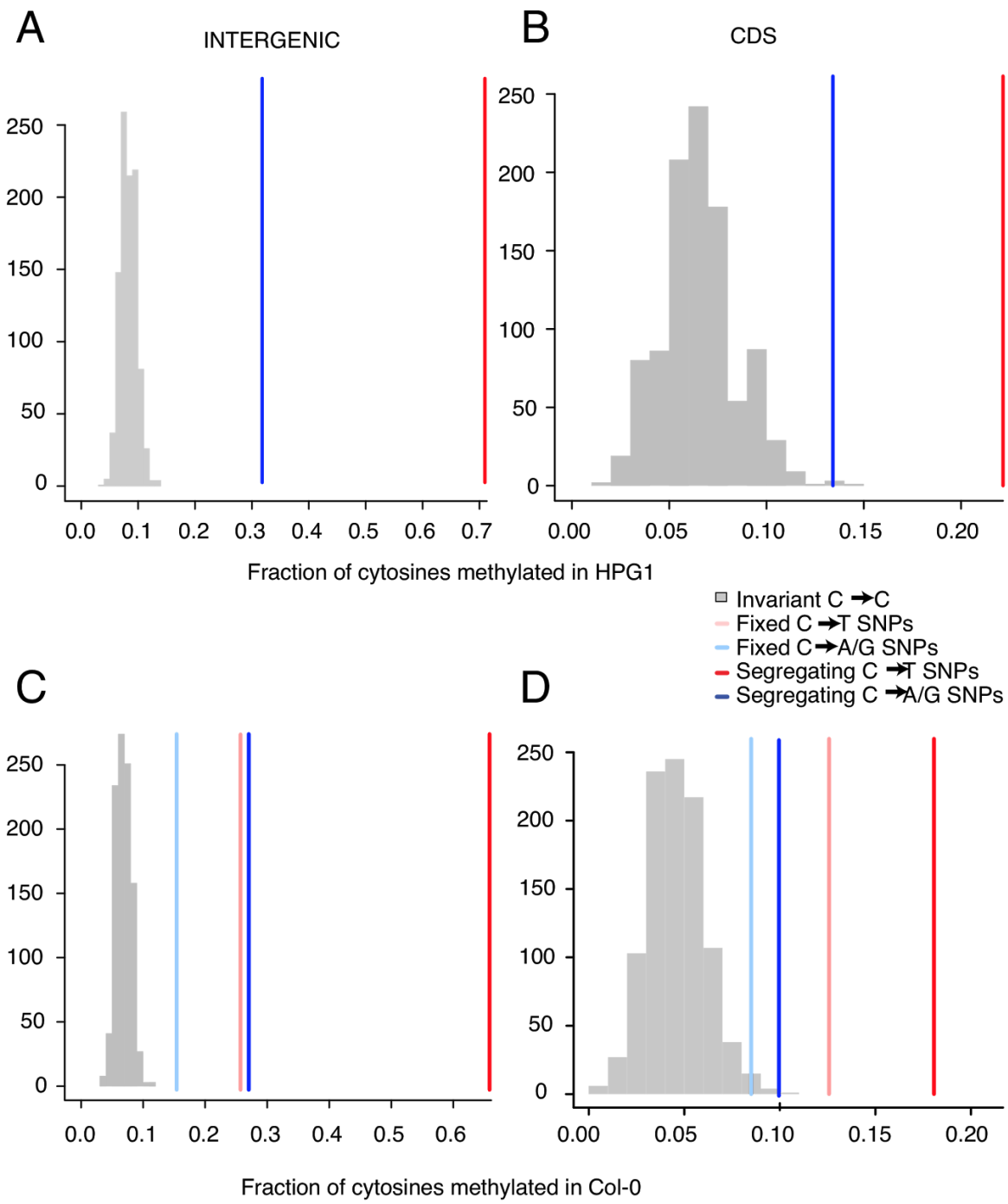


Figure S4. Relationship between methylation and substitutions.

(A, B) Fraction of methylation of cytosines in HPG1 pseudo-reference[7] at intergenic (A) or coding regions (B). **(C, D)** Fraction of methylation of cytosines in Col-0 reference genome[5] at intergenic (C) or coding regions (D). In each of the four comparisons, a grey histogram represents distribution of methylation of 1,000 random sets of invariant cytosines. Lines represent average methylation degree at those sites in HPG1 that changed from cytosine to thymine (red). We differentiate those substitutions that are shared - fixed - across all individuals (light red) or whose allele are present at

an intermediate - segregating - frequency (dark red). Likewise, average methylation is shown for sites that changed from cytosine to adenine (blue) that that are fixed (light blue) or segregating (dark blue). The fact that the average methylation is higher in new substitutions than in invariant positions supports a connection between methylation and mutability of sites.

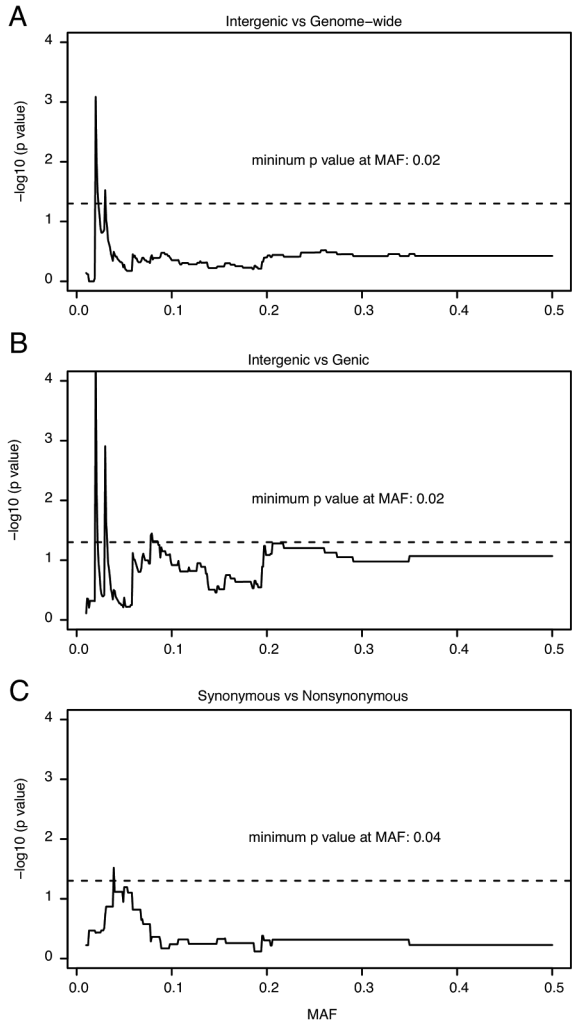


Figure S5. Enrichment of low variants at putatively selected annotations.

We tested for an interaction in a 2x2 table of counts of SNPs using Fisher's Exact Test. The tables were built with the number of SNPs falling into each of the two annotations: genome - intergenic **(A)**, genic - intergenic **(B)**, and synonymous - non-synonymous **(C)**; and two discrete allele groups assuming a minimum allele frequency (MAF) cutoff. We repeated the test by sliding the cut-off from 0 to 50% allele frequency (x axis) and we show the corresponding p-value (y axis). The dashed line indicates the 5% significance threshold.

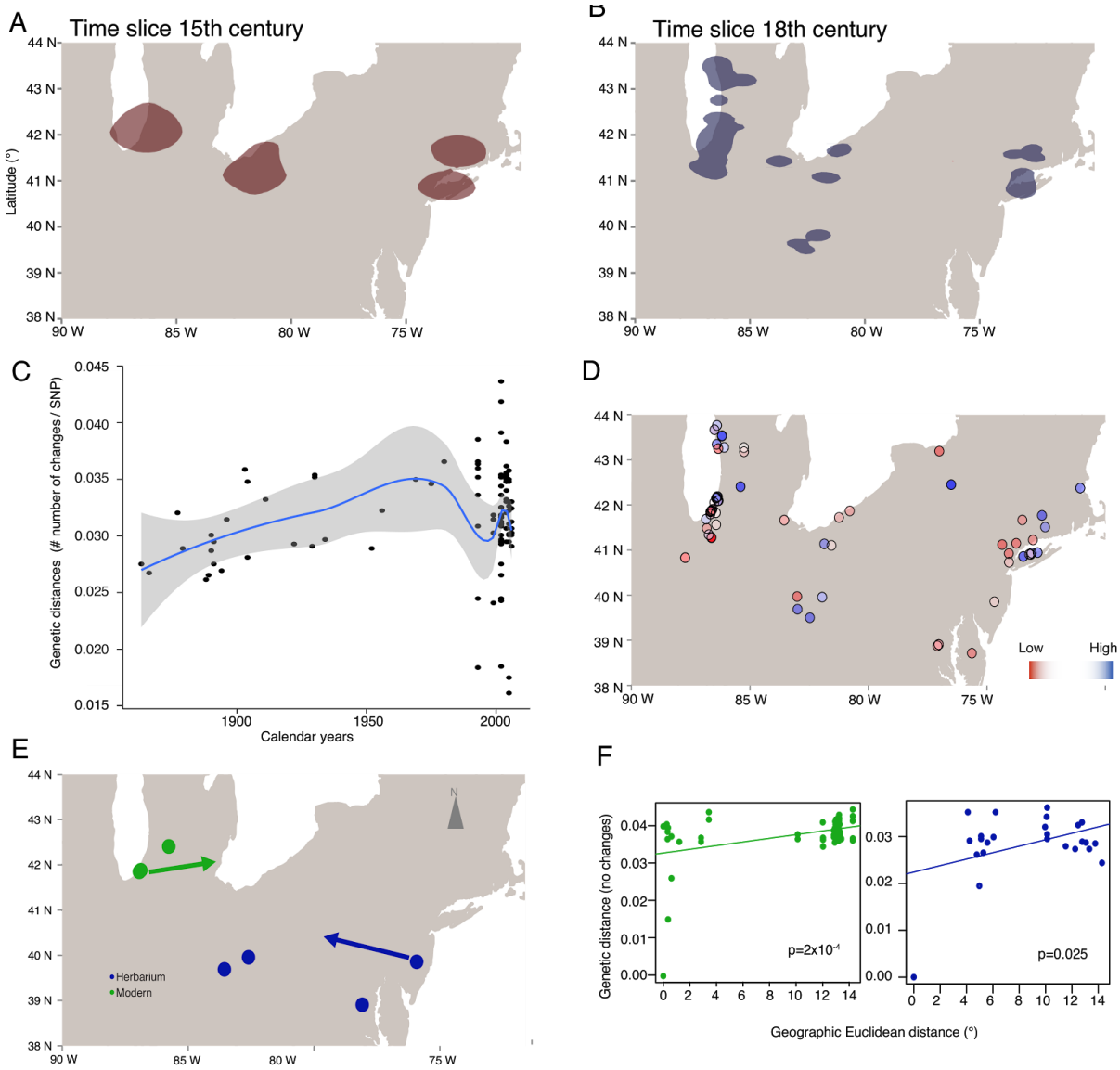


Figure S6. Phylogeographic inference in HPG1.

(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 (dark red/brown) summarizing the first ~100 years of the phylogenetic tree. Clouds represent the 95% interval of the Highest Posterior Probability Density of locations. **(B)** Current distribution map (dark blue) summarizing the last ~100 years. Clouds as in (A). **(C)** Diversity in time. Each point represents the average number of genetic changes between a sample and the other samples within a decade. The blue 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 number of genetic changes among the 10 geographically closest neighbors. Genetic distances are shown qualitatively from a red (low) to blue (high) gradient. **(E)** Origin of herbarium and modern geographic spread, determined using separate heuristic searches of

isolation-by-distance patterns. Three locations of modern samples and four locations of herbarium samples showed significant slopes ($p < 0.05$) in the isolation-by-distance pattern. That is, genetic distance increased when moving away from these geographic locations. For one sample of each subset, herbarium **(F)** and modern **(G),** a likely migration trajectory is depicted by an arrow and its isolation-by-distance pattern is shown.

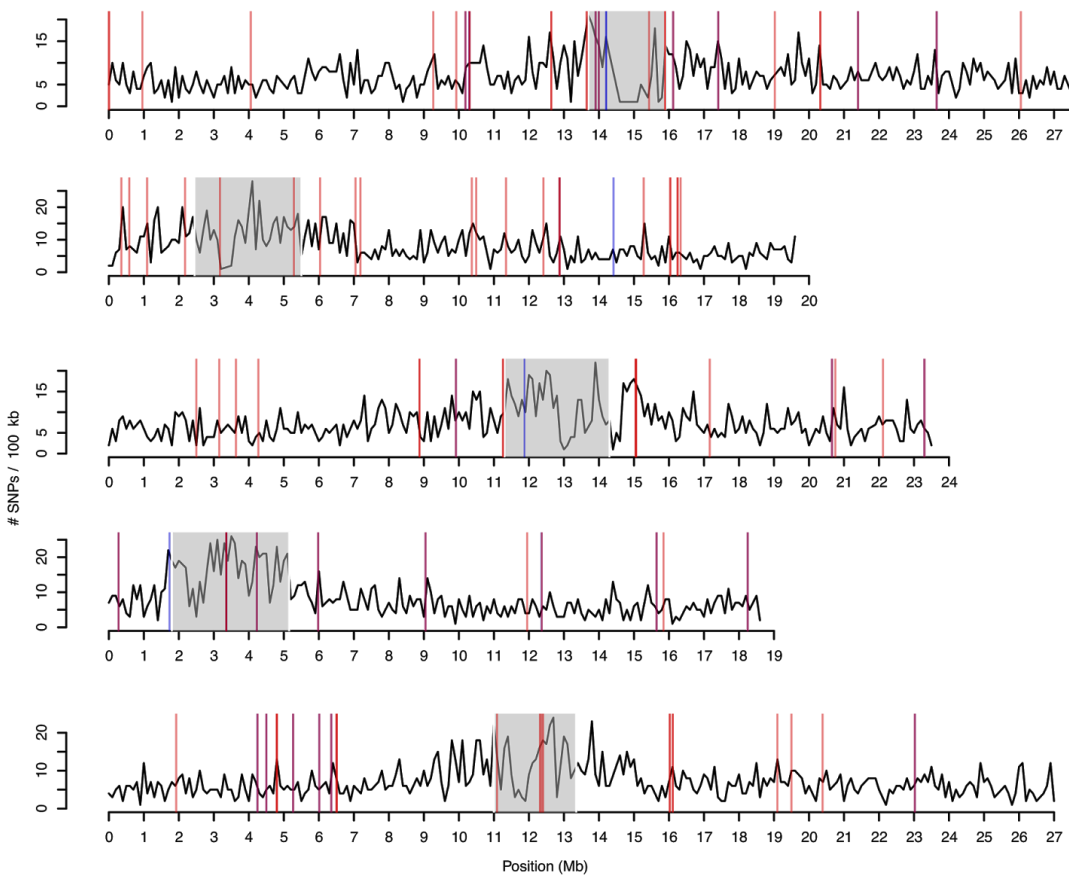


Figure S7. Density of SNPs along all chromosomes and location of SNP hits.

Black line shows number of SNPs per 100 kb window. Centromere locations are indicated by grey shading. Vertical lines indicate SNPs associated with root phenotypes (red) and climatic variables (blue) (see Table S5).

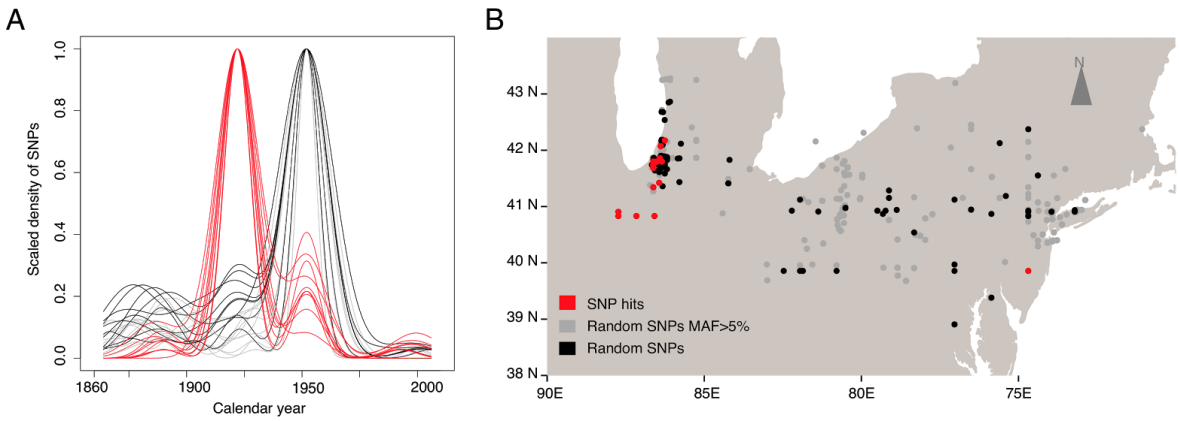


Figure S8. Spatial and temporal emergence of root-associated mutations.

(A) Age distribution of derived SNPs with a significant trait association (the herbarium sample in which they were first recorded) (red), compared with genome-wide SNPs with at least 5% minor allele frequency (grey), or without frequency cutoff (black). **(B)** Spatial centroid of all samples carrying a derived allele. Since it is an average location, centroids can be in a body of water. Ten random draws of 50 SNPs for each category were used to produce the density lines in (A) and points in (B).

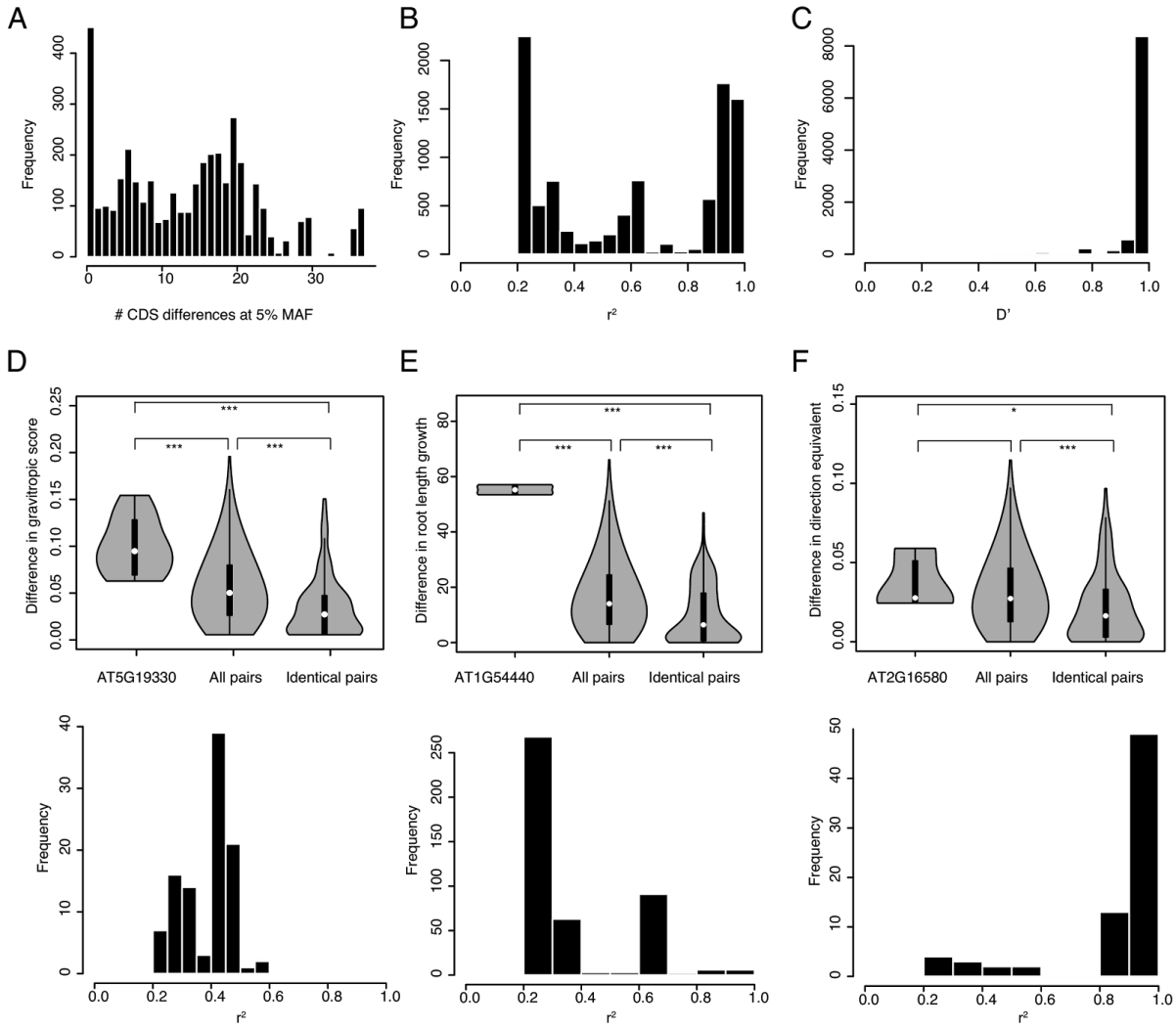


Figure S9. Linkage disequilibrium between SNPs with significant trait associations.

(A-F) Linkage disequilibrium between SNPs with significant trait associations. Histogram of genetic distances **(A)** between samples when evaluating only coding regions at 5% minimum allele frequency. Linkage disequilibrium between SNP hits measured as r^2 **(B)** and *D'* **(C)**. Three significant SNPs were further studied to exemplify the power of association analyses with HPG1. For each, phenotypic differences between accessions that differ in the focal SNP and that are otherwise virtually genetically identical are compared both with all pairs of accessions and with pairs of accessions completely identical for coding regions. Below each violin plot is the histogram of linkage disequilibrium of the focal SNP with all other SNP hits. The three focal SNPs evaluated are located in AT5G19330 **(D)**, AT1G54440 **(E)** and AT2G16580 **(F)**.

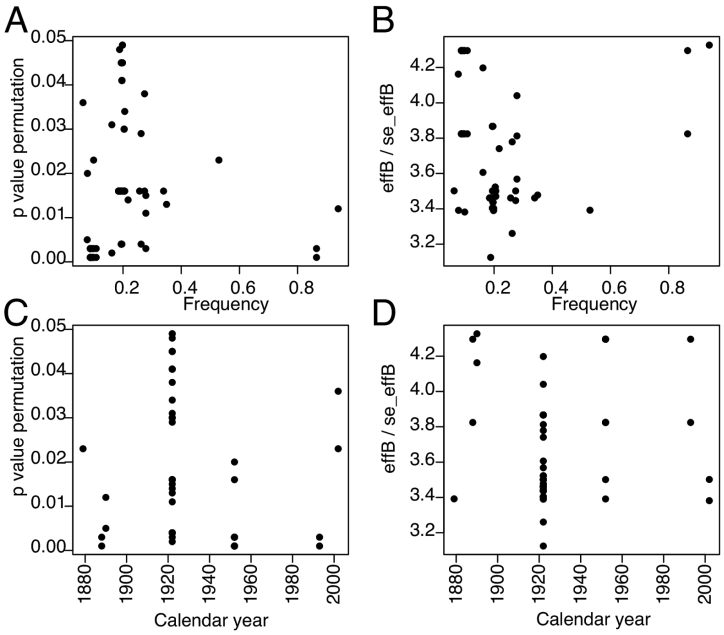


Figure S10. Correlations of SNP effects and p-values with frequency and age.

Correlation between SNP frequency and p-value **(A)**, frequency and effect **(B)**, age and p-value **(C)**, age and effect **(D)**. All cases were non-significant.

SUPPLEMENTAL TABLES

See appended .**pdf** file for **Tables S1-5**.

See appended .**pdf** file for **Graphic Table S7**: 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).

Table S1. Sample information.

(Abbreviation H* indicates herbarium samples that cluster with the mo
dern HPG1 clade rather than the historic HPG1 clade in Fig. 3., highlighted as a star in the map from Fig. 1. 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.)

Accession	Latitude (ºN)	Longitude (º E)	State	Date collected	Alternative name	Collector/ Herbarium	Average coverage (x)	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
JK399	38.7155	-75.635591	DE	1863	888124	NY	9	105,053,631	99,889,683	142	yes	H	101
JK366	43.1921	-77.0102	NY	1866	888144	NY	6.8	100,379,839	95,118,236	123	yes	H	94
JK395	38.9068	-77.036667	DC	1877	888134	NY	10.3	103,620,791	98,888,406	167	yes	H	100
JK888141	40.732007	-74.068455	NJ	1879	888141	NY	42	107,211,409	102,634,255	161	yes	H	103
JK389	38.9068	-77.036667	DC	1888	1365363	NY	9.9	106,042,465	100,826,958	151	yes	H	98
JK362	38.9068	-77.036667	DC	1889	1365364	NY	8.8	103,997,716	98,876,320	153	yes	H	93
JK367	40.9249	-74.0755	NJ	1890	1365344	NY	16.7	107,236,732	102,176,782	181	yes	H	95
JK372	41.1222	-74.3569	NJ	1890	1365332	NY	14.8	106,285,178	101,480,369	163	yes	H	96
JK1365354	38.8782	-77.09048	VA	1891	1365354	NY	36.4	106,718,326	102,458,166	169	yes	H	88
JK376	39.97	-83.01	NY	1891	1365337	NY	12.3	105,962,154	100,840,125	145	yes	H	97
JK351	41.15	-73.766667	NY	1894	1365333	NY	16.1	106,531,302	101,841,156	153	yes	H	90
JK355	35.99	-83.94	TE	1896	1365374	NY	14.3	106,391,637	101,455,311	192	yes	H	91
JK356	n/a	n/a	GA	1897	1365375	NY	5.3	90,426,010	89,296,191	n/a	no	H	92
JK393	n/a	n/a	NC	1897	1365370	NY	30.4	102,894,430	101,298,068	n/a	no	H	99
JK346	40.643136	-111.95177	UT	1903	102365	NY	29.1	107,223,283	102,450,446	222	yes	H	89
JK2525	41.224343	-73.06021	CT	1904	79391	UNCONN	12.5	105,025,845	n/a	138	yes	H	118
JK2529	n/a	n/a	OH	1904	176849	CFM	11.4	100,620,441	n/a	n/a	no	H	121

JK401	40.643136	-111.95177	UT	1904	102364	NY	10.4	99,572,736	94,661,828	216	yes	H	102
JK2513	41.102121	-81.560547	OH	1911	25	OSU	18.2	106,309,854	n/a	176	yes	H	108
JK2509	n/a	n/a	CT	1917	11	OSU	15.1	102,169,546	n/a	n/a	no	H	104
JK2530	41.482862	-86.822602	IN	1922	531679	CFM	22.2	107,043,540	n/a	161	yes	H	122
JK2526	41.666667	-73.508455	CT	1929	79409	UNCONN	16.3	107,026,827	n/a	161	yes	H	119
JK2515	41.137296	-81.863779	OH	1930	30	OSU	21.3	106,893,416	n/a	193	yes	H	110
JK2511	41.721618	-81.243317	OH	1934	14	OSU	5.6	95,822,372	n/a	109	yes	H	106
JK2523	n/a	n/a	OH	1940	25707	UNC	13.1	101,421,749	n/a	n/a	no	H	116
JK2520	n/a	n/a	OH	1945	54051	UNC	20.3	102,831,697	n/a	n/a	no	H	114
JK2524	39.856783	-74.686954	NJ	1952	63978	UNC	13.8	100,778,282	n/a	n/a	no	H	117
JK2512	39.95607	-81.953309	OH	1956	21	OSU	16.7	106,801,844	n/a	189	yes	H	107
JK2514	39.95607	-81.953309	OH	1969	27	OSU	28.4	107,044,415	n/a	219	yes	H	109
JK2517	n/a	n/a	OH	1981	34	OSU	21.7	102,643,436	n/a	n/a	no	H	112
JK2521	n/a	n/a	OH	1992	565960	UNC	2.9	62,673,938	n/a	n/a	no	H	115
JK2518	41.867643	-80.789021	OH	1993	40	OSU	14.8	106,578,197	n/a	177	yes	H	113
JK2531	39.856783	-74.686954	NJ	1952	1507461	CFM	15.1	106,158,181	n/a	177	yes	H*	123
JK2510	39.688861	-82.993218	OH	1930	13	OSU	21	106,305,970	n/a	178	yes	H*	105
JK2527	41.509059	-72.543694	CT	1975	79389	UNCONN	8.3	104,089,205	n/a	200	yes	H*	120
JK2516	39.500862	-82.472413	OH	1980	32	OSU	18.1	106,464,569	n/a	198	yes	H*	111
CSHL_15	40.8585	-73.4675	NY	1993	CSHL-15	ABRC	39.3	108,189,771	105,955,885	243	yes	M	16
CSHL_17	40.8585	-73.4675	NY	1993	CSHL-17	ABRC	41.5	108,194,960	105,982,511	240	yes	M	17
FM_10	42.4489	-76.5072	NY	1993	FM-10	ABRC	44.6	108,203,215	106,052,866	269	yes	M	20
FM_11	42.4489	-76.5072	NY	1993	FM-11	ABRC	44.4	108,214,008	106,040,276	288	yes	M	21
HS_12	42.373	-71.0627	MA	1993	HS-12	ABRC	48.8	108,230,030	106,124,249	251	yes	M	25
HS_17	42.373	-71.0627	MA	1993	HS-17	ABRC	55.3	108,242,062	106,155,362	254	yes	M	26
Kno_10	41.2816	-86.621	IN	1993	Kno-10	ABRC	39.4	108,198,601	105,985,288	226	yes	M	32
KNO_15	41.2816	-86.621	IN	1993	KNO-15	ABRC	43.6	108,219,683	106,069,077	231	yes	M	33
Gre_0	43.178	-85.2532	MI	1995	Gre-0	ABRC	44.6	108,209,345	106,032,827	207	yes	M	22
Tul_0	43.2708	-85.2563	MI	1995	CS6877	ABRC	31.2	108,140,393	105,806,418	221	yes	M	85
CS8067	41.3599	-122.755	CA	1996	Buckhorn Pas	ABRC	66.4	108,260,489	106,243,277	294	yes	M	15
Tol_2	41.6639	-83.5553	OH	1996	CS8022	ABRC	61	108,241,333	106,194,209	238	yes	M	83
Tol_3	41.6639	-83.5553	OH	1996	CS8023	ABRC	40.2	108,184,749	105,953,559	232	yes	M	84
MIA_1	41.7976	-86.6691	MI	1999	MIA-1	ABRC	73.1	108,279,881	106,291,612	234	yes	M	56

MIA_5	41.7976	-86.6691	MI	1999	MIA-5	ABRC	62.9	108,263,557	106,250,560	235	yes	M	57
MIC_20	41.8266	-86.4366	MI	1999	MIC-20	ABRC	39.9	108,200,416	106,010,135	237	yes	M	58
MIC_24	41.8266	-86.4366	MI	1999	MIC-24	ABRC	33.8	108,176,527	105,728,326	237	yes	M	59
Brn_10	41.9	-86.583	MI	2002	Brn-10	ABRC	33.3	108,177,381	105,905,097	243	yes	M	7
Brn_24	41.9	-86.583	MI	2002	Brn-24	ABRC	38.4	108,208,482	105,951,803	228	yes	M	8
Haz_10	41.879	-86.607	MI	2002	Haz-10	ABRC	33.8	108,154,100	105,903,700	230	yes	M	23
Haz_2	41.879	-86.607	MI	2002	Haz-2	ABRC	39.7	108,201,103	106,004,251	288	yes	M	24
Ker_4	42.184	-86.358	MI	2002	Ker-4	ABRC	32.1	108,132,127	105,806,486	261	yes	M	30
Ker_5	42.184	-86.358	MI	2002	Ker-5	ABRC	62.9	108,259,905	106,246,278	259	yes	M	31
L_R_10	41.847	-86.67	MI	2002	L-R-10	ABRC	22.4	108,062,944	105,496,224	186	yes	M	49
L_R_5	41.847	-86.67	MI	2002	L-R-5	ABRC	60.6	108,255,795	106,209,826	299	yes	M	50
Lak_12	41.8	-86.67	MI	2002	Lak-12	ABRC	37.8	108,176,901	105,775,999	237	yes	M	36
Lak_13	41.8	-86.67	MI	2002	Lak-13	ABRC	28.5	107,955,559	105,553,559	226	yes	M	37
Map_35	42.166	-86.412	MI	2002	Map-35	ABRC	64.7	108,265,863	106,224,216	290	yes	M	51
Map_42	42.166	-86.412	MI	2002	Map-42	ABRC	46	107,303,032	106,093,945	n/a	no	M	52
Map_8	42.166	-86.412	MI	2002	Map-8	ABRC	33.4	108,155,999	105,921,907	287	yes	M	53
Mdn_10	42.051	-86.509	MI	2002	Mdn-10	ABRC	34.9	108,106,772	105,906,924	n/a	no	M	54
Mdn_8	42.051	-86.509	MI	2002	Mdn-8	ABRC	37.4	108,199,679	105,940,666	266	yes	M	55
Paw_13	42.148	-86.431	MI	2002	Paw-13	ABRC	43	108,159,739	105,980,721	267	yes	M	70
Paw_20	42.148	-86.431	MI	2002	Paw-20	ABRC	41.3	108,218,762	106,059,867	241	yes	M	71
Riv_25	42.184	-86.382	MI	2002	Riv-25	ABRC	36.8	108,186,632	105,779,717	273	yes	M	76
Riv_26	42.184	-86.382	MI	2002	Riv-26	ABRC	35.7	108,194,281	105,958,738	260	yes	M	77
Yng_4	41.865	-86.646	MI	2002	Yng-4	ABRC	41.3	108,182,789	106,000,003	289	yes	M	86
Yng_53	41.865	-86.646	MI	2002	Yng-53	ABRC	46	108,230,553	106,125,861	191	yes	M	87
RRS_10	41.5609	-86.4251	IN	2003	RRS-10	ABRC	41.8	108,208,144	106,033,465	274	yes	M	80
DuckLkSP38	43.3431	-86.4045	MI	2004	DuckLkSP38	ABRC	37.1	108,171,751	105,932,415	253	yes	M	18
DuckLkSP40	43.3431	-86.4045	MI	2004	DuckLkSP40	ABRC	39.6	108,204,654	105,969,244	257	yes	M	19
KBS_Mac_68	42.405	-85.398	MI	2004	KBS-Mac-68	ABRC	41.3	108,181,390	105,870,424	259	yes	M	27
KBS_Mac_74	42.405	-85.398	MI	2004	KBS-Mac-74	ABRC	37.7	108,160,645	105,801,702	265	yes	M	28
MNF_Che_47	43.5251	-86.1843	MI	2004	MNF-Che-47	ABRC	27.6	108,093,393	105,596,885	281	yes	M	60
MNF_Che_49	43.5251	-86.1843	MI	2004	MNF-Che-49	ABRC	28.5	108,082,202	105,661,610	274	yes	M	61
MNF_Pin_40	43.5356	-86.1788	MI	2004	MNF-Pin-40	ABRC	47.9	108,238,775	106,099,919	287	yes	M	62
MNF_Pot_10	43.595	-86.2657	MI	2004	MNF-Pot-10	ABRC	61.4	108,189,553	106,228,588	n/a	no	M	63

MNF_Pot_15	43.595	-86.2657	MI	2004	MNF-Pot-15	ABRC	25.2	108,543,185	107,022,924	n/a	no	M	64
MSGA_10	43.2749	-86.0891	MI	2004	MSGA-10	ABRC	41.9	108,191,659	106,019,404	233	yes	M	65
MSGA_12	43.2749	-86.0891	MI	2004	MSGA-12	ABRC	42.8	108,227,214	106,032,928	240	yes	M	66
MSGA_61	43.2749	-86.0891	MI	2004	MSGA-61	ABRC	45.5	108,210,152	106,077,183	247	yes	M	67
MuskSP_68	43.2483	-86.3368	MI	2004	MuskSP-68	ABRC	25.8	108,063,297	105,588,467	215	yes	M	68
MuskSP_83	43.2483	-86.3368	MI	2004	MuskSP-83	ABRC	29.9	108,099,368	105,721,042	222	yes	M	69
Pent_46	43.7623	-86.3929	MI	2004	Pent-46	ABRC	48.3	108,227,763	106,099,890	238	yes	M	72
Pent_7	43.7623	-86.3929	MI	2004	Pent-7	ABRC	55.7	108,220,625	106,144,167	240	yes	M	73
SLSP_67	43.665	-86.496	MI	2004	SLSP-67	ABRC	53.5	108,238,880	106,143,530	245	yes	M	81
SLSP_69	43.665	-86.496	MI	2004	SLSP-69	ABRC	35.5	108,160,835	105,899,252	249	yes	M	82
KNO2_41	41.273	-86.625	IN	2005	KNO2.41	ABRC	44.7	108,209,694	106,063,235	219	yes	M	34
KNO2_54	41.273	-86.625	IN	2005	KNO2.54	ABRC	44	108,212,430	105,903,373	218	yes	M	35
LI_EF_011	40.9064	-73.1493	NY	2005	LI-EF-011	ABRC	68.6	108,267,109	106,250,331	259	yes	M	38
LI_EF_018	40.9064	-73.1493	NY	2005	LI-EF-018	ABRC	39	108,244,306	105,898,497	230	yes	M	39
LI_OF_061	40.7777	-72.9069	NY	2005	LI-OF-061	ABRC	58	104,897,841	105,729,196	n/a	no	M	40
LI_RR_096	40.9447	-72.8615	NY	2005	LI-RR-096	ABRC	63.5	108,264,679	106,251,487	261	yes	M	41
LI_RR_097	40.9447	-72.8615	NY	2005	LI-RR-097	ABRC	40.8	108,211,310	105,992,095	249	yes	M	42
LI_SET_019	40.9352	-73.114	NY	2005	LI-SET-019	ABRC	29.9	108,085,297	105,737,781	259	yes	M	43
LI_SET_036	40.9352	-73.114	NY	2005	LI-SET-036	ABRC	41.5	108,216,592	106,006,605	238	yes	M	44
LI_WP_039	40.9076	-73.2089	NY	2005	LI-WP-039	ABRC	104.8	108,301,282	106,273,259	239	yes	M	45
LI_WP_041	40.9076	-73.2089	NY	2005	LI-WP-041	ABRC	76.5	108,287,248	106,322,146	235	yes	M	46
PT1_52	41.3423	-86.7368	IN	2005	PT1.52	ABRC	50.6	108,240,431	106,154,252	219	yes	M	74
PT1_85	41.3423	-86.7368	IN	2005	PT1.85	ABRC	46.1	108,220,150	106,097,633	233	yes	M	75
RMX4_118	42.036	-86.511	MI	2005	RMX4.118	ABRC	41.8	106,178,554	105,685,651	n/a	no	M	78
11PNA1_14	42.0945	-86.3253	MI	2006	11PNA1.14	ABRC	47.5	108,227,783	106,133,372	276	yes	M	1
328PNA062	42.0945	-86.3253	MI	2006	328PNA062	ABRC	47.3	108,221,709	106,127,272	223	yes	M	2
627ME_13Y1	42.093	-86.359	MI	2006	n/a	ABRC	53.4	107,908,679	106,148,671	n/a	no	M	3
627ME_1MI1	42.093	-86.359	MI	2006	627ME-1MI1	ABRC	57.8	108,252,617	106,173,403	281	yes	M	4
627RMX_1MN4	42.0333	-86.5128	MI	2006	n/a	ABRC	43.6	106,799,549	105,789,469	n/a	no	M	5
627RMX_1MN5	42.0333	-86.5128	MI	2006	n/a	ABRC	50.6	106,885,430	105,897,441	n/a	no	M	6
BRR107	40.8313	-87.735	IL	2006	BRR107	ABRC	28.5	108,896,513	107,320,745	n/a	no	M	9
BRR12	40.8313	-87.735	IL	2006	BRR12	ABRC	43.9	108,190,572	106,031,493	232	yes	M	10
BRR23	40.8313	-87.735	IL	2006	BRR23	ABRC	30.7	108,095,072	105,726,913	236	yes	M	11

BRR4	40.8313	-87.735	IL	2006	BRR4	ABRC	44.7	108,180,840	106,033,507	219	yes	M	12
BRR57	40.8313	-87.735	IL	2006	BRR57	ABRC	28.4	108,093,033	105,630,963	225	yes	M	13
BRR60	40.8313	-87.735	IL	2006	BRR60	ABRC	42.9	108,281,285	106,199,572	229	yes	M	14
KEN	41.767	-72.677	CT	n/a	KEN	ABRC	55.2	108,233,232	106,158,223	249	yes	M	29
LP3413_31	41.6862	-86.8513	IN	n/a	LP3413.31	ABRC	55.9	108,244,332	106,190,596	227	yes	M	47
LP3413_53	41.6862	-86.8513	IN	n/a	LP3413.53	ABRC	51.2	108,157,453	105,994,665	245	yes	M	48
RMX413_85	42.036	-86.511	MI	n/a	RMX413.85	ABRC	38	106,816,221	105,483,632	n/a	no	M	79

Table S2. Sample information for Col-0 mutation accumulation lines.

Information about each Mutation Accumulation (MA) line and their number of SNPs at different annotations. Also the total number of SNPs, average number of mutations and total bp covered in the genome per annotation are reported.

MA line	Read depth	Generation	Total	SNPs	Deletions	insertions	CDS	Nonsyn	Syn	Intron	5' UTR	3' UTR	TE	Intergenic
0-4-26	57	3	7	6	1	0	0	0	0	0	0	0	1	5
0-8-87	49	3	7	5	0	2	1	1	0	1	0	0	0	3
30-109	45	31	31	23	7	1	3	3	0	3	0	0	2	15
30-119	45	31	33	26	2	5	1	1	0	1	2	0	4	18
30-29	51	31	39	26	10	3	2	1	1	3	0	1	5	15
30-39	48	31	28	18	7	3	1	1	0	1	0	1	4	11
30-49	50	31	30	23	3	4	4	4	0	0	0	0	6	13
30-59	40	31	46	31	8	7	5	2	3	2	0	0	6	18
30-69	50	31	26	21	3	2	4	3	1	1	1	1	6	8
30-79	50	31	31	25	3	3	6	4	2	2	0	0	8	9
30-89	39	31	35	27	5	3	4	3	1	1	1	0	2	19
30-99	44	31	37	35	1	1	6	5	1	2	0	2	8	17
Total SNPs				274			38	28	10	17	4	5	52	158
average (31st)			33.6	25.5	4.9	3.2	3.6	2.7	0.9	1.6	0.4	0.5	5.1	14.3
stdev (31st)			5.9	4.9	3.0	1.8	1.8	1.4	1.0	1.0	0.7	0.7	2.1	3.9
Total bp			115,954,227				30,753,966			17,446,837	4,289,789	2,508,199	9,267,413	48,090,487

Table S3. Mutation rate estimates for different annotations in HPG1 and mutation accumulation lines.

Mutation rates from MA lines are compared to HPG1 substitution rates from the dataset of 32_15 quality filter and complete information (see SOM) (Abbreviations: stat, descriptive statistic; bp, base pairs; lower and upper, lower and upper 95% CI; Nonsyn. and Syn., nonsynonymous and synonymous sites; UTR, untranslated region sites; HPG1 adj., substitution rate of HPG1 adjusted by a mean generation time of 1.3 years)

Dataset	stat	CDS	Syn.	Nonsyn.	Intronic	5' UTR	3' UTR	Transposon	Intergenic	Genome
MA	mean	3.776	n/a	n/a	2.958	3.008	6.431	17.752	9.592	7.094
MA	sem	1.928	n/a	n/a	1.786	5.258	9.094	7.420	2.628	1.352
MA	lower	2.581	n/a	n/a	1.851	-0.251	0.794	13.153	7.964	6.256
MA	upper	4.971	n/a	n/a	4.065	6.267	12.067	22.351	11.221	7.932
HPG1	mean	2.149	n/a	n/a	1.540	n/a	n/a	2.290	3.029	2.114
HPG1	sem	0.108	n/a	n/a	0.165	n/a	n/a	0.536	0.173	0.119
HPG1	lower	1.943	n/a	n/a	1.231	n/a	n/a	1.314	2.698	1.871
HPG1	upper	2.364	n/a	n/a	1.874	n/a	n/a	3.309	3.368	2.344
HPG1 adj.	mean	2.794	n/a	n/a	2.002	n/a	n/a	2.977	3.938	2.748
HPG1 adj.	sem	0.140	n/a	n/a	0.214	n/a	n/a	0.697	0.225	0.154
HPG1 adj.	lower	2.526	n/a	n/a	1.600	n/a	n/a	1.708	3.508	2.432
HPG1 adj.	upper	3.073	n/a	n/a	2.436	n/a	n/a	4.302	4.378	3.047
Distribution of pairwise SNP differences	min	0	0	0	0	0	0	0	0	0
	1st qu.	2	1	1	1	0	1	2	5	9
	median	5	3	3	3	1	2	4	10	18
	mean	5.6	3	3.1	3.8	1.2	1.9	4.3	11.3	21.1
	3rd qu.	8	5	4	5	2	3	6	16	31
	max.	27	17	11	15	5	7	22	43	87
Total number of SNPs		971	531	448	629	74	158	656	2498	5013
Total bp		32119233	n/a	n/a	18132262	2632130	4480510	6209512	43601507	108434034

Table S4. Description of phenotypic and climatic variables for association mapping analyses.

Mean and standard deviation (s.d.) across accessions for each phenotypic and climatic variables. Broad sense heritabilities (H2) were calculated from between line and within line (between replicate) variance in ANOVA. P-value corresponds to F test. Narrow sense heritabilities (h2) were calculated employing linear mixed models and kinship matrix from mean accession values. P-values correspond to Likelihood Ratio test.

Variable	Description	mean	s.d.	H2	p-value	h2	p-value
FT_V0	Time from germination until the first flower opens (days) under 0 days of vernalization	101	4.53	0.009	7.28E-03	0.017	1.97E-25
FT_V1	Time from germination until the first flower opens (days) under 14 days of vernalization	107	4.12	0.013	6.87E-04	0.395	1.83E-25
FT_V2	Time from germination until the first flower opens (days) under 28 days of vernalization	102	3.22	0.012	1.04E-03	0.429	3.37E-27
FT_V3	Time from germination until the first flower opens (days) under 63 days of vernalization	110	1.32	0.010	5.11E-03	0.226	9.52E-25
B_V0	Time from germination until the first developed bud (days) under 0 days of vernalization	88.8	4	0.013	8.99E-04	0.018	2.26E-25
B_V1	Time from germination until the first developed bud (days) under 14 days of vernalization	93.9	3.84	0.009	7.45E-03	0.340	3.98E-25
B_V2	Time from germination until the first developed bud (days) under 28 days of vernalization	89.2	2.13	0.005	6.92E-02	0.252	2.22E-25
B_V3	Time from germination until the first developed bud (days) under 63 days of vernalization	101	0.45	0.006	5.79E-02	0.177	1.99E-24
Fecundity	Pixel area of inflorescence (correlation with number of fruits, rho=0.84)	0.02	0.0042	0.001	3.56E-01	0.240	1.02E-22
seed_size	Average seed size (mm2)	0.134	0.0053	0.016	4.73E-03	0.149	3.58E-24
GR_rootLength	Average root growth rate	181	14.9	0.131	4.76E-77	0.640	3.13E-29
GR_shootArea	Average of shoot area growth rate	2279	253	0.053	2.33E-24	0.812	1.77E-31
rootLength	Average root length	467	35.8	0.048	2.01E-21	0.409	2.57E-28

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	0.544	1.14E-26
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	0.303	1.41E-25
meanRootWidth	Average root width	5.27	0.177	0.038	5.30E-16	0.359	1.52E-25
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	0.166	3.37E-25
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	0.291	1.76E-25
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	0.405	6.51E-26
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	0.381	5.47E-26
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	0.351	8.81E-26
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	0.210	4.68E-27
gravitropicScore	Average score for root angle intervals	0.1	0.0457	0.044	2.83E-19	0.642	7.56E-27
TotLen.EucLen	Average root tortuosity: Total root length divided by Euclidian length	1.1	0.0097	0.009	6.83E-03	0.422	2.53E-25
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	0.393	2.69E-24
BIO1	Annual Mean Temperature (ºC x 10)	98.1	12.8	n/a	n/a	0.066	3.22E-40
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	107	7.65	n/a	n/a	0.073	1.02E-40
BIO3	Isothermality (BIO2/BIO7) (x 100)	28.9	1.8	n/a	n/a	0.361	4.91E-39
BIO4	Temperature Seasonality (standard deviation x 100)	9169	483	n/a	n/a	0.383	4.68E-47
BIO5	Max Temperature of Warmest Month (ºC x 10)	283	10.1	n/a	n/a	0.152	3.78E-40
BIO6	Min Temperature of Coldest Month (ºC x 10)	-80.9	18	n/a	n/a	0.275	4.79E-42
BIO7	Temperature Annual Range (BIO5-BIO6) (ºC x 10)	364	17.5	n/a	n/a	0.239	6.31E-42
BIO8	Mean Temperature of Wettest Quarter (ºC x 10)	176	55.1	n/a	n/a	0.016	3.58E-43

BIO9	Mean Temperature of Driest Quarter (ºC x 10)	-7.11	48.7	n/a	n/a	0.000	3.58E-43
BIO10	Mean Temperature of Warmest Quarter (ºC x 10)	213	10.8	n/a	n/a	0.205	3.33E-40
BIO11	Mean Temperature of Coldest Quarter (ºC x 10)	-24.1	18.2	n/a	n/a	0.270	1.71E-41
BIO12	Annual Precipitation (mm)	990	109	n/a	n/a	0.219	3.94E-44
BIO13	Precipitation of Wettest Month (mm)	103	6.72	n/a	n/a	0.206	1.53E-40
BIO14	Precipitation of Driest Month (mm)	54.1	16.7	n/a	n/a	0.104	1.51E-40
BIO15	Precipitation Seasonality (Coefficient of Variation)	17.8	5.51	n/a	n/a	0.157	8.93E-40
BIO16	Precipitation of Wettest Quarter (mm)	291	19.7	n/a	n/a	0.269	1.55E-42
BIO17	Precipitation of Driest Quarter (mm)	191	44.8	n/a	n/a	0.084	3.67E-42
BIO18	Precipitation of Warmest Quarter (mm)	277	25.2	n/a	n/a	0.342	7.42E-44
BIO19	Precipitation of Coldest Quarter (mm)	197	47	n/a	n/a	0.022	2.68E-42

Table S5. SNP hits from association analyses and several descriptors.

SNP hits significant at the 5% level after permutation correction are shown. Additionally, if raw p-values pass a double Bonferroni threshold of 0.01% are marked with a "tick". (Abbreviations: nonsyn. and syn., nonsynonymous and synonymous changes; regular one-letter abbreviation was used for amino acid changes)

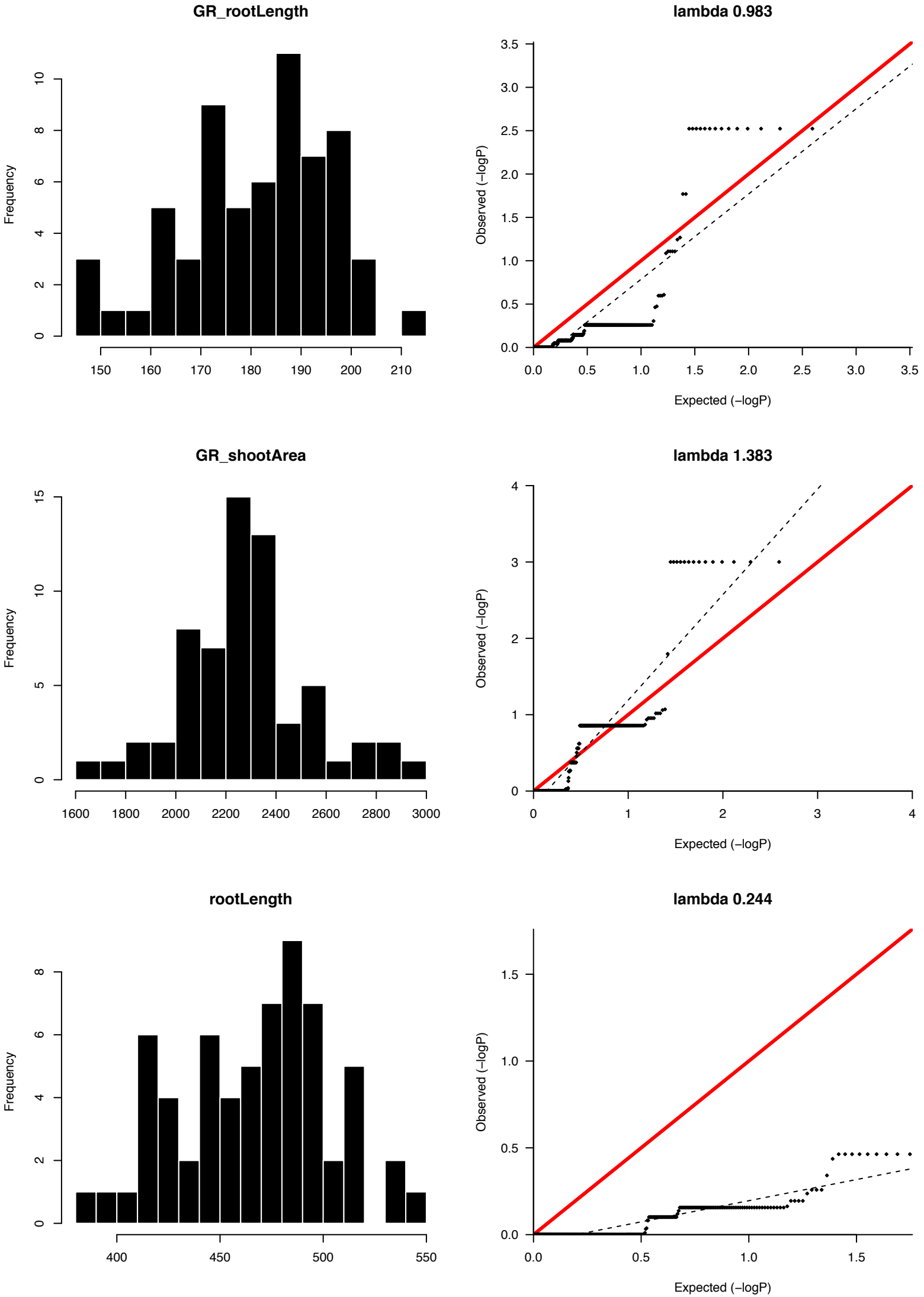
Trait	Chromosome	Position	Ancestral	Derived	Effect	Effect standard error	Sample size	p - value raw	p- value false discovery rate	p- value permutation corrected	Allele frequency	Allele frequency in modern set	Oldest herbarium individual	Longitude	Latitude	Substitution type	AA change	Gene	Biochemical effect (Grantham score)	Significant permutation	Significant double Bonferroni	LD
dirEquivalent	1	958948	G	T	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.186	0.227	1922	41.7	-85.3	nonsyn	A->P	AT1G03810	27	✓		53
gravitropicScore	1	9925177	C	T	0.033	0.010	63	7.10E-04	0.0651	0.016	0.078	0.092	1952	40.9	-82.3	interg.				✓		1
bio18	1	10187610	T	C	6.830	1.987	99	5.83E-04	0.0124	0.047	0.196	0.24	1922	41.7	-85.3	interg.				✓		52
GR_rootLength	1	12638692	C	T	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.087	0.105	1952	40.9	-81.3	interg.				✓	✓	13
GR_shootArea	1	12638692	C	T	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.087	0.105	1952	40.9	-81.3	interg.				✓	✓	13
GR_rootLength	1	13652509	C	A	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.093	0.107	1952	40.9	-82.9	interg.				✓	✓	12
GR_shootArea	1	13652509	C	A	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.093	0.107	1952	40.9	-82.9	interg.				✓	✓	12
bio18	1	13904611	C	T	6.570	1.756	90	1.83E-04	0.0124	0.016	0.217	0.237	1922	41.7	-85.3	interg.				✓		49
bio18	1	13994958	G	A	6.830	1.987	99	5.83E-04	0.0124	0.047	0.196	0.24	1922	41.7	-85.3	tranposon		AT1G36933		✓		49
bio18	1	17408807	C	T	6.830	1.987	99	5.83E-04	0.0124	0.047	0.196	0.24	1922	41.7	-85.3	interg.				✓		48
dirEquivalent	1	19024876	C	T	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.19	0.23	1922	41.7	-85.3	interg.				✓		47
GR_shootArea	1	20324050	G	A	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.087	0.105	1952	40.9	-82.9	interg.		AT1G54440		✓	✓	11
GR_rootLength	1	20324050	G	A	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.087	0.105	1952	40.9	-82.9	interg.		AT1G54440		✓	✓	11
bio18	1	23648407	A	C	6.830	1.987	99	5.83E-04	0.0124	0.047	0.196	0.24	1922	41.7	-85.3	nonsyn	Y->S	AT1G63740	144	✓		46
dirEquivalent	1	26052913	A	T	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.185	0.224	1922	41.7	-85.3	interg.				✓		45
GR_shootArea	1	29696198	G	A	-121.000	33.911	63	3.68E-04	0.0096	0.016	0.278	0.329	1922	41.5	-84.9	interg.				✓		42

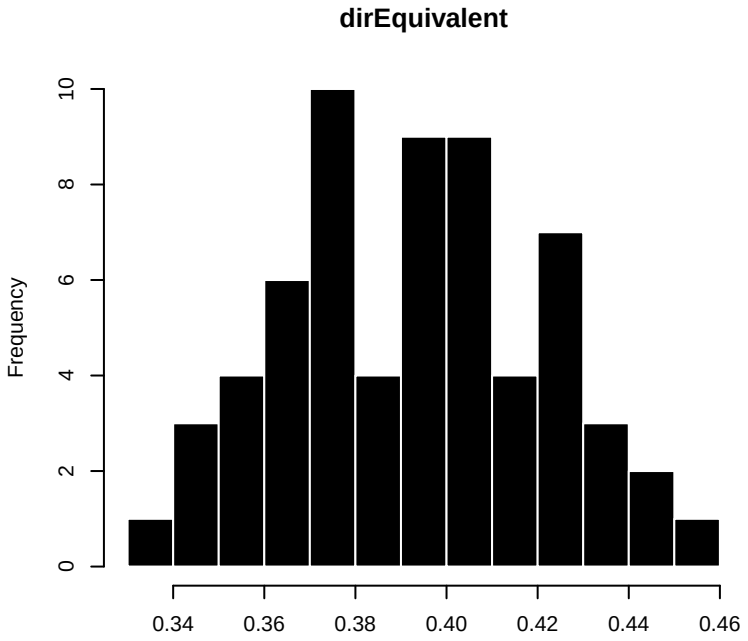
bio16	1	29696198	G	A	5.250	1.377	94	1.39E-04	0.0632	0.016	0.278	0.329	1922	41.5	-84.9	interg.				✓		42
bio18	1	29696198	G	A	6.340	1.569	94	5.36E-05	0.0124	0.004	0.278	0.329	1922	41.5	-84.9	interg.				✓		42
GR_rootLength	1	30015381	T	A	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.087	0.105	1952	40.9	-82.9	interg.				✓	✓	10
GR_shootArea	1	30015381	T	A	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.087	0.105	1952	40.9	-82.9	interg.				✓	✓	10
GR_rootLength	1	30143319	G	A	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.088	0.105	1952	40.9	-82.9	interg.				✓	✓	9
GR_shootArea	1	30143319	G	A	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.088	0.105	1952	40.9	-82.9	interg.				✓	✓	9
dirEquivalent	2	358395	C	T	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	syn.	V->V	AT2G01820		✓	✓	43
dirEquivalent	2	585918	C	T	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	syn.	G->G	AT2G02220		✓	✓	42
dirEquivalent	2	1093203	C	T	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	interg.				✓	✓	41
dirEquivalent	2	2176891	T	C	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	interg.				✓	✓	40
GR_rootLength	2	3174832	T	A	6.340	1.869	63	6.97E-04	0.017	0.017	0.529	0.566	1879	41.3	-84.3	interg.				✓		0
TotLen.EucLen	2	5285907	C	A	-0.006	0.002	63	3.05E-04	0.0241	0.037	0.162	0.194	1922	41.5	-85	interg.				✓	✓	39
dirEquivalent	2	5285907	C	A	-0.019	0.005	63	2.64E-05	0.0032	0.001	0.162	0.194	1922	41.5	-85	interg.				✓	✓	39
dirEquivalent	2	6034545	C	T	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	syn.	S->S	AT2G14247		✓	✓	38
dirEquivalent	2	7047529	G	T	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	nonsyn	P->A	AT2G16270	27	✓	✓	37
dirEquivalent	2	7186220	C	T	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	intron		AT2G16580		✓	✓	36
dirEquivalent	2	10369545	T	C	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	interg.				✓	✓	35
dirEquivalent	2	10495275	A	C	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.196	0.237	1922	41.7	-85.3	intron		AT2G24680		✓	✓	34
dirEquivalent	2	11346211	C	A	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.186	0.227	1922	41.7	-85.3	interg.				✓		33
dirEquivalent	2	12415084	T	A	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.186	0.227	1922	41.7	-85.3	intron		AT2G28900		✓		32
dirEquivalent	2	12876361	A	C	-0.015	0.004	63	1.56E-04	0.0041	0.006	0.262	0.29	1922	41.7	-84.6	interg.				✓		31
gravitropicScore	2	12876361	A	C	-0.021	0.006	63	1.08E-03	0.0651	0.027	0.262	0.29	1922	41.7	-84.6	interg.				✓		31
bio13	2	14417366	A	G	3.990	0.959	64	3.22E-05	0.0147	0.004	0.077	0	1890	39.5	-77.9	interg.				✓		1
dirEquivalent	2	15278350	A	G	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.186	0.227	1922	41.7	-85.3	interg.				✓		30
GR_shootArea	2	16039488	T	G	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.087	0.105	1952	40.9	-82.9	3' UTR		AT2G38290		✓	✓	8
GR_rootLength	2	16039488	T	G	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.087	0.105	1952	40.9	-82.9	3' UTR		AT2G38290		✓	✓	8
GR_rootLength	2	16247290	G	T	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.088	0.105	1952	40.9	-82.9	nonsyn	A->G	AT2G38910	60	✓	✓	7
GR_shootArea	2	16247290	G	T	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.088	0.105	1952	40.9	-82.9	nonsyn	A->G	AT2G38910	60	✓	✓	7
dirEquivalent	2	16333662	G	A	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.186	0.227	1922	41.7	-85.3	nonsyn	A->G	AT2G39160	60	✓		29
dirEquivalent	3	2500258	C	A	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	syn.	K->K	AT3G07830		✓	✓	28
dirEquivalent	3	3154804	C	T	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	interg.				✓	✓	27
dirEquivalent	3	3629794	C	T	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	intron		AT3G11530		✓	✓	26
dirEquivalent	3	4269626	G	T	-0.016	0.004	63	1.15E-04	0.0032	0.006	0.194	0.237	1922	41.7	-85.3	5' UTR		AT3G13229		✓	✓	25
GR_shootArea	3	8873116	C	T	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.097	0.118	1952	40.9	-81.9	interg.				✓	✓	6
GR_rootLength	3	8873116	C	T	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.097	0.118	1952	40.9	-81.9	interg.				✓	✓	6

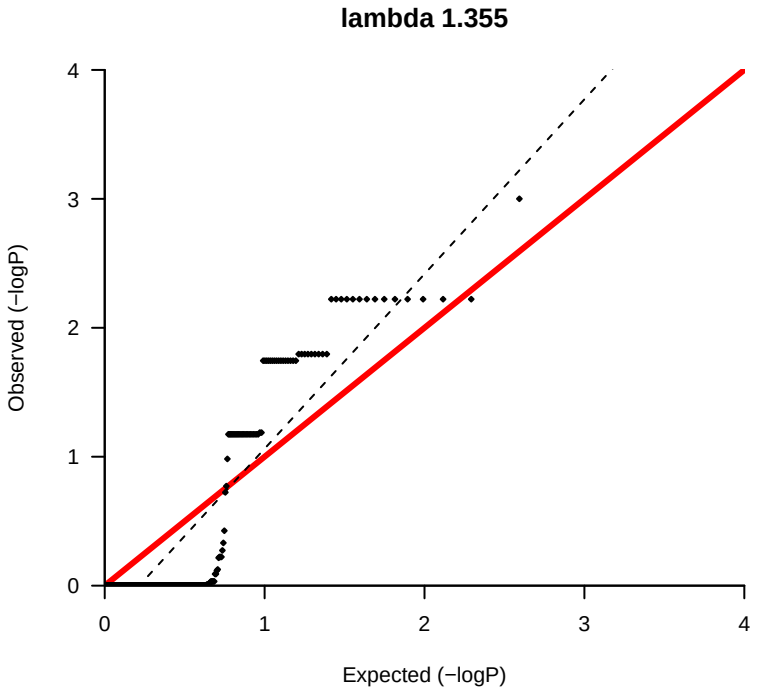
GR_rootLength	3	11259214	A	T	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.088	0.105	1952	40.9	-82.9	interg.					✓	✓	5
GR_shootArea	3	11259214	A	T	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.088	0.105	1952	40.9	-82.9	interg.					✓	✓	5
bio8	3	11873293	A	G	37.800	8.736	65	1.52E-05	0.0069	0.006	0.939	1	1890	41.8	-83.7	tranposon	AT3G30219				✓		0
GR_rootLength	3	15050751	G	A	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.108	0.105	1888	40.2	-82.5	interg.					✓	✓	4
GR_shootArea	3	15050751	G	A	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.108	0.105	1888	40.2	-82.5	interg.					✓	✓	4
dirEquivalent	3	17164638	C	A	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.19	0.227	1922	41.7	-85.3	interg.					✓		24
bio18	4	279210	T	G	6.830	1.987	99	5.83E-04	0.0124	0.047	0.196	0.24	1922	41.7	-85.3	interg.					✓		22
bio11	4	1732480	T	A	-5.550	1.564	79	3.89E-04	0.0195	0.045	0.063	0.068	2002	41	-87.5	interg.					✓		2
bio4	4	1732480	T	A	224.000	63.967	79	4.67E-04	0.0128	0.044	0.063	0.068	2002	41	-87.5	interg.					✓		2
dirEquivalent	4	3355152	C	G	-0.014	0.004	63	4.45E-04	0.0052	0.016	0.204	0.25	1922	41.7	-85.4	interg.					✓		21
bio18	4	3355152	C	G	6.850	1.944	##	4.25E-04	0.0124	0.035	0.204	0.25	1922	41.7	-85.4	interg.					✓		21
dirEquivalent	4	3355946	G	C	-0.014	0.004	63	4.45E-04	0.0052	0.016	0.204	0.25	1922	41.7	-85.4	interg.					✓		20
bio18	4	3355946	G	C	6.850	1.944	##	4.25E-04	0.0124	0.035	0.204	0.25	1922	41.7	-85.4	interg.					✓		20
dirEquivalent	4	4228138	A	G	-0.014	0.004	63	4.45E-04	0.0052	0.016	0.196	0.24	1922	41.7	-85.3	tranposon	AT4G07440				✓		19
bio18	4	4228138	A	G	6.830	1.987	99	5.83E-04	0.0124	0.047	0.196	0.24	1922	41.7	-85.3	tranposon	AT4G07440				✓		19
dirEquivalent	4	9046942	G	C	-0.014	0.004	63	4.45E-04	0.0052	0.016	0.204	0.25	1922	41.7	-85.4	nonsyn	H->Q	AT4G15960	24		✓		18
bio18	4	9046942	G	C	6.850	1.944	##	4.25E-04	0.0124	0.035	0.204	0.25	1922	41.7	-85.4	nonsyn	H->Q	AT4G15960	24		✓		18
dirEquivalent	4	11948961	T	A	-0.014	0.004	63	4.45E-04	0.0052	0.016	0.198	0.25	1952	41.7	-85.3	interg.					✓		17
dirEquivalent	4	12365323	C	T	-0.014	0.004	63	4.45E-04	0.0052	0.016	0.204	0.25	1922	41.7	-85.4	interg.					✓		16
bio18	4	12365323	C	T	6.850	1.944	##	4.25E-04	0.0124	0.035	0.204	0.25	1922	41.7	-85.4	interg.					✓		16
dirEquivalent	4	15646341	C	A	-0.014	0.004	63	4.45E-04	0.0052	0.016	0.206	0.25	1922	41.7	-85.4	syn.	E->E	AT4G32410			✓		15
bio18	4	15646341	C	A	6.720	1.936	99	5.14E-04	0.0124	0.042	0.206	0.25	1922	41.7	-85.4	syn.	E->E	AT4G32410			✓		15
dirEquivalent	4	15845001	A	T	-0.014	0.004	63	4.45E-04	0.0052	0.016	0.194	0.25	1922	41.8	-85.9	3' UTR		AT4G32840			✓		14
dirEquivalent	4	18249171	T	A	-0.014	0.004	63	4.45E-04	0.0052	0.016	0.274	0.328	1922	41.8	-85.9	interg.					✓		13
bio18	4	18249171	T	A	6.910	2.005	71	5.62E-04	0.0124	0.047	0.274	0.328	1922	41.8	-85.9	interg.					✓		13
bio18	5	4245213	A	T	6.830	1.987	99	5.83E-04	0.0124	0.047	0.196	0.24	1922	41.7	-85.3	syn.	I->I	AT5G13260			✓		12
bio18	5	4500202	G	A	6.830	1.987	99	5.83E-04	0.0124	0.047	0.196	0.24	1922	41.7	-85.3	nonsyn	A->G	AT5G13950	60		✓		11
dirEquivalent	5	4797923	A	T	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.188	0.227	1922	41.7	-85.3	tranposon		AT5G14830			✓		10
dirEquivalent	5	4797976	G	A	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.257	0.293	1922	41.7	-85.3	tranposon		AT5G14830			✓		10
dirEquivalent	5	4798526	A	G	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.339	0.362	1922	41.7	-85.3	interg.					✓		9
gravitropicScore	5	6508329	A	G	-0.020	0.006	63	5.20E-04	0.0651	0.008	0.35	0.447	1922	42	-85	nonsyn	C->W	AT5G19330	215		✓		0
dirEquivalent	5	11090365	T	A	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.186	0.224	1922	41.7	-85.3	TE		AT5G29037			✓		4
dirEquivalent	5	12312975	C	G	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.185	0.224	1922	41.7	-85.3	TE		AT5G32630			✓		3
dirEquivalent	5	12358159	C	T	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.186	0.224	1922	41.7	-85.3	tranposon		AT5G32825			✓		2
dirEquivalent	5	12409027	G	A	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.185	0.224	1922	41.7	-85.3	interg.					✓		1

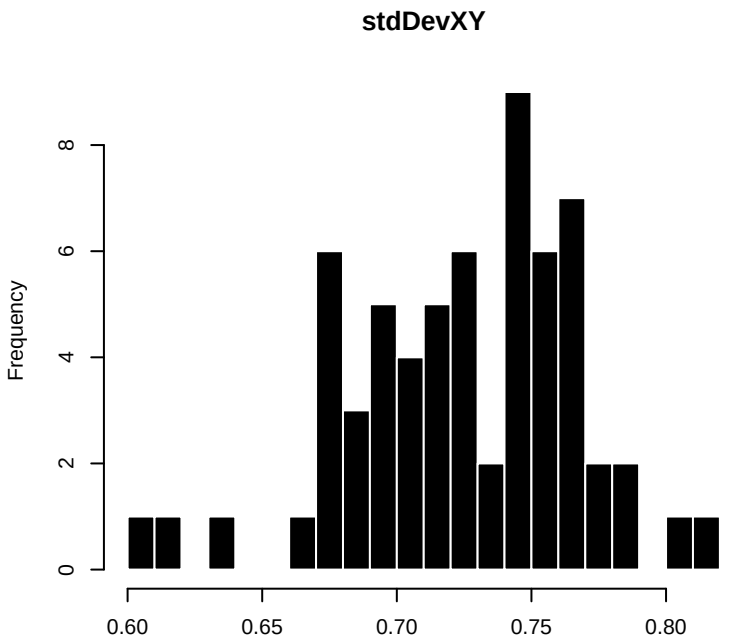
GR_rootLength	5	16024197	A	T	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.098	0.118	1952	40.9	-81.9	intron	AT5G40020	✓	✓	2
GR_shootArea	5	16024197	A	T	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.098	0.118	1952	40.9	-81.9	intron	AT5G40020	✓	✓	2
GR_shootArea	5	16109431	G	A	-231.000	53.774	63	1.75E-05	0.0005	0.001	0.865	0.877	1993	42.2	-84.4	interg.		✓	✓	1
GR_rootLength	5	16109431	G	A	-12.100	3.164	63	1.33E-04	0.0037	0.003	0.865	0.877	1993	42.2	-84.4	interg.		✓	✓	1
dirEquivalent	5	19099082	G	C	-0.014	0.004	63	5.30E-04	0.0052	0.018	0.186	0.227	1922	41.7	-85.3	interg.		✓		0
GR_rootLength	5	20388107	A	T	-10.700	3.164	63	6.94E-04	0.017	0.017	0.099	0.12	2002	41	-86.6	interg.		✓		0

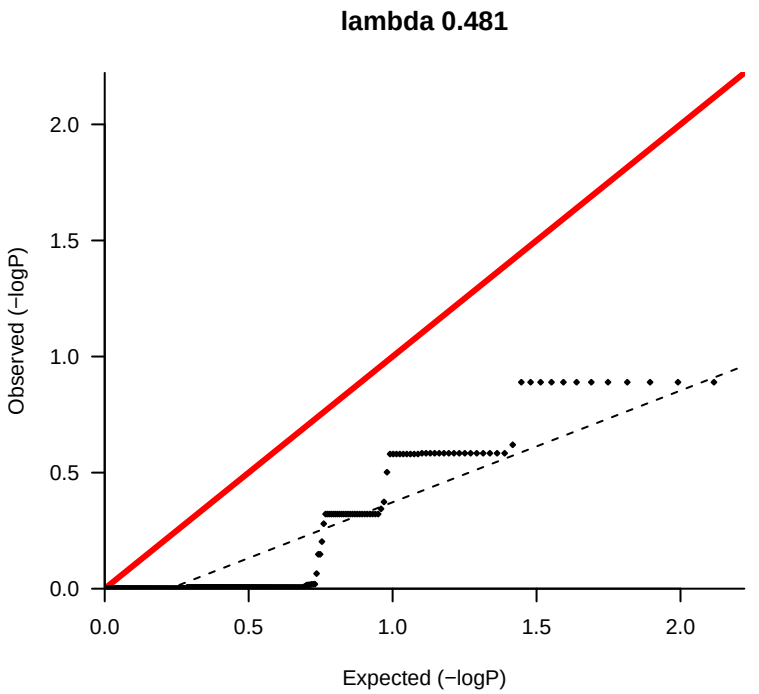
Graphic Table S7

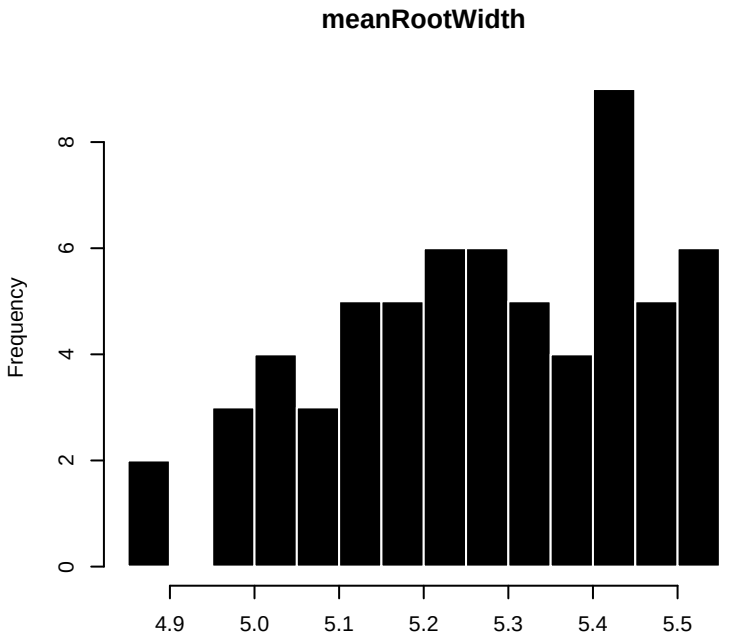


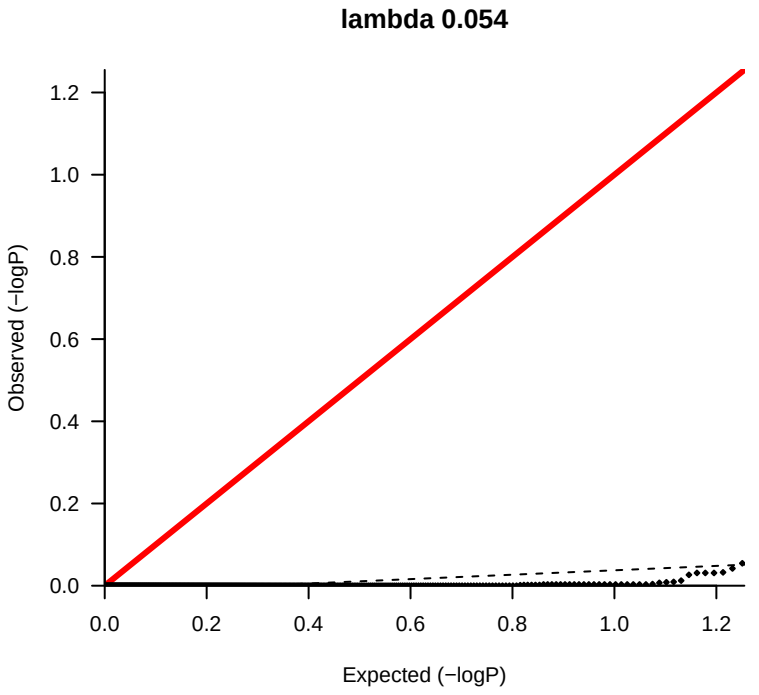

dirEquivalent
Frequency
0
2
4
6
8
10
0.34
0.36
0.38
0.40
0.42
0.44
0.46

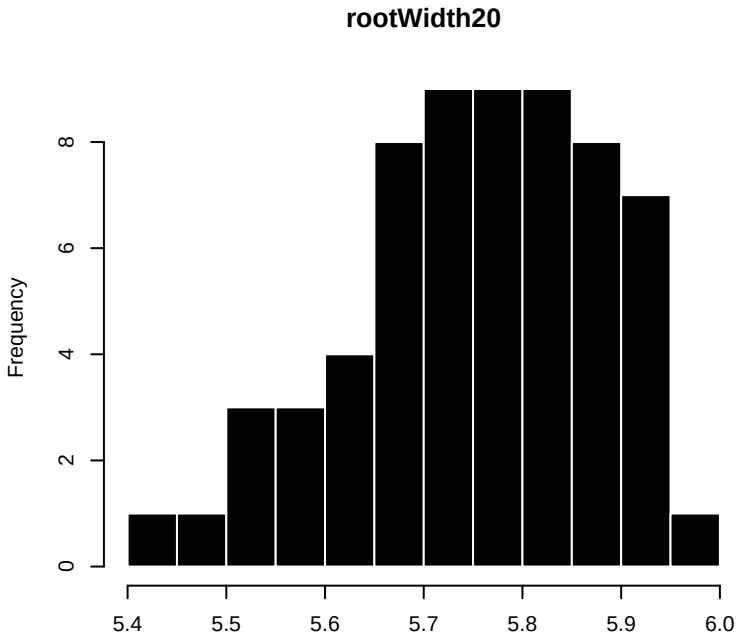

lambda 1.355
Observed (−logP)
Expected (−logP)
0
1
2
3
4

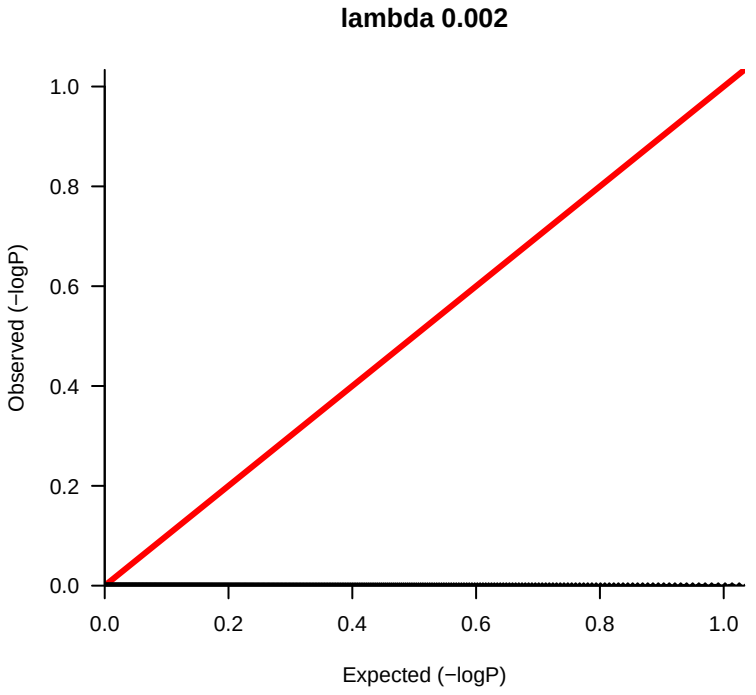

stdDevXY
Frequency
0
2
4
6
8
0.60
0.65
0.70
0.75
0.80

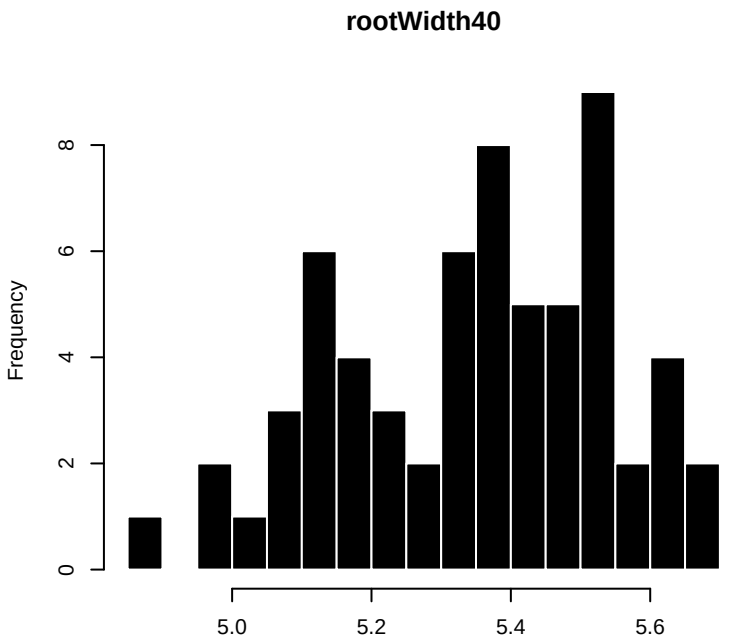

lambda 0.481
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

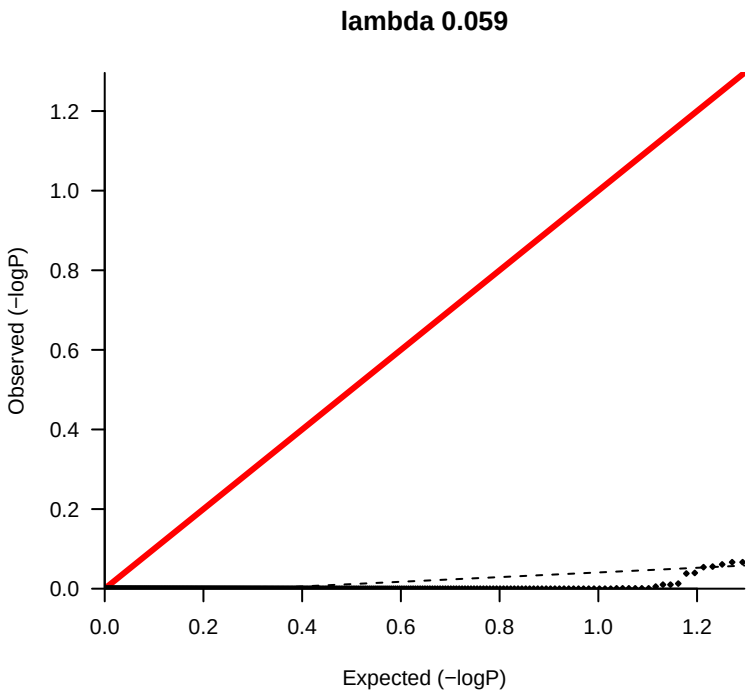

meanRootWidth
Frequency
0
2
4
6
8
4.9
5.0
5.1
5.2
5.3
5.4
5.5

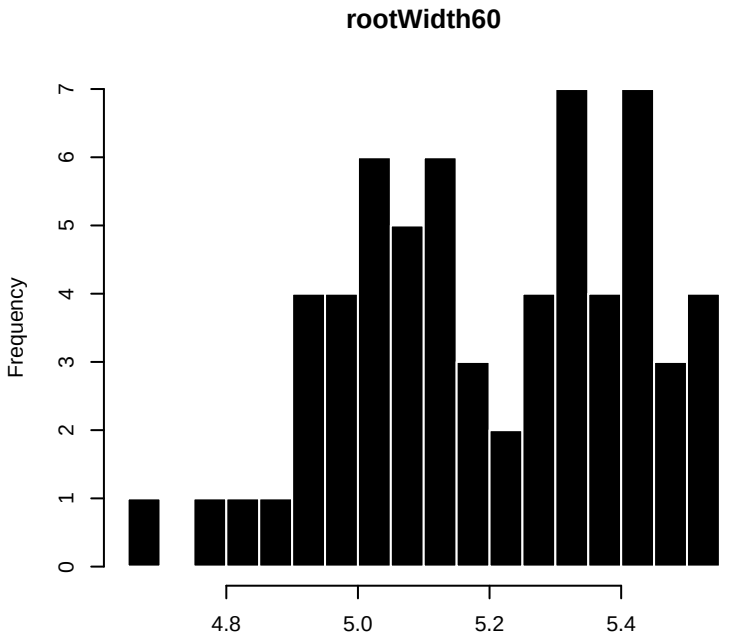

lambda 0.054
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2

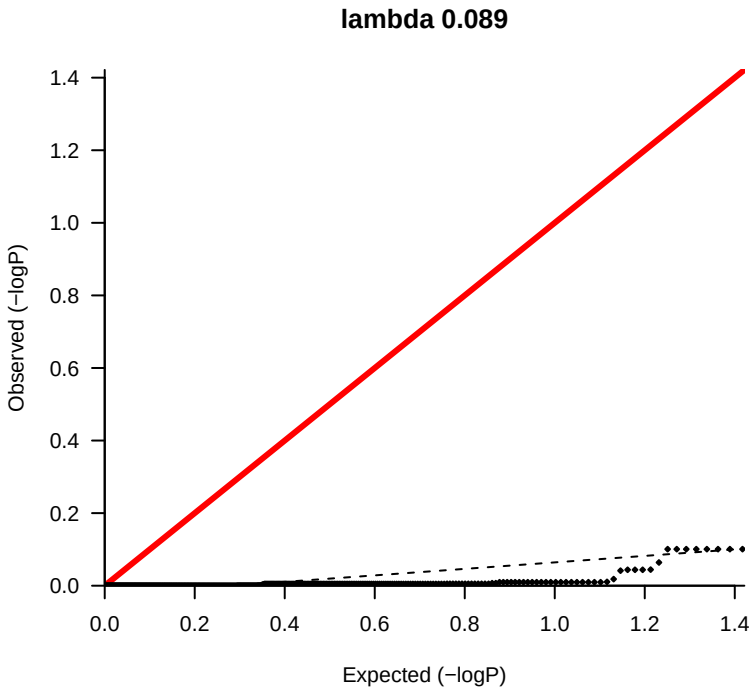

rootWidth20
Frequency
0
2
4
6
8
5.4
5.5
5.6
5.7
5.8
5.9
6.0

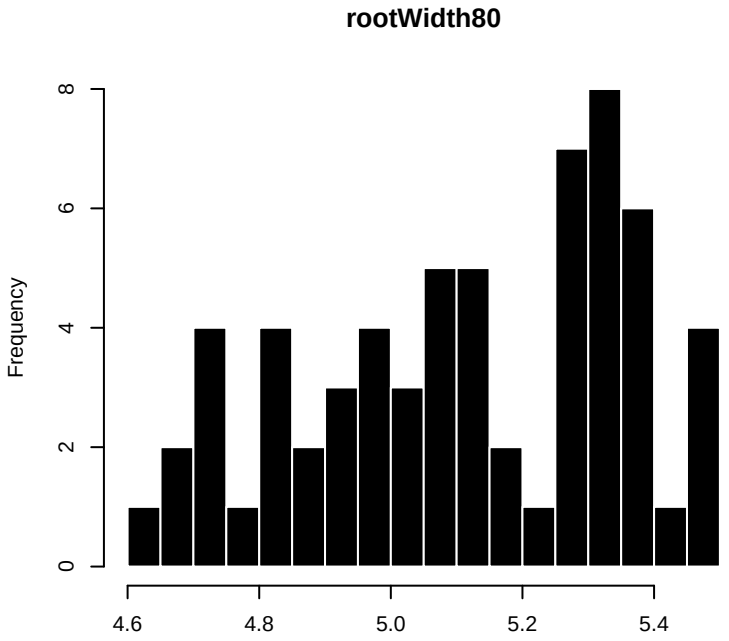

lambda 0.002
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0

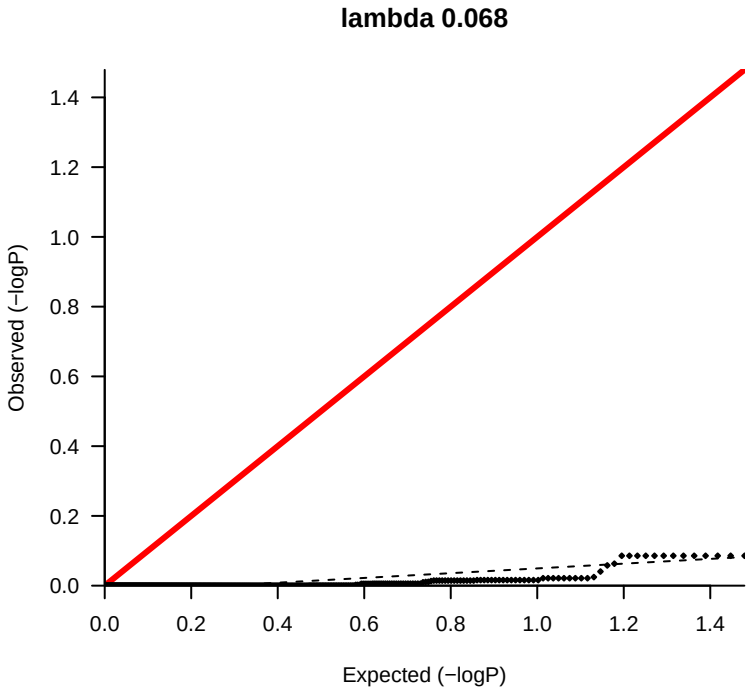

rootWidth40
Frequency
0
2
4
6
8
5.0
5.2
5.4
5.6

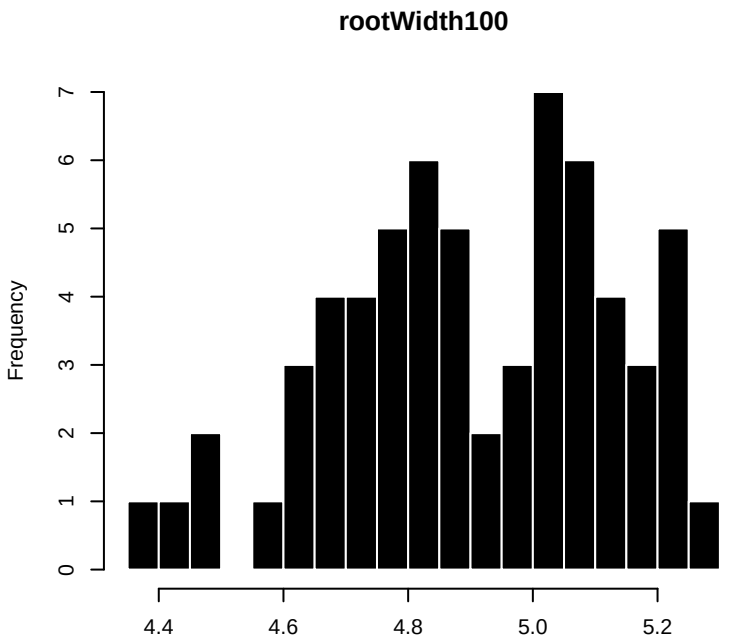

lambda 0.059
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2

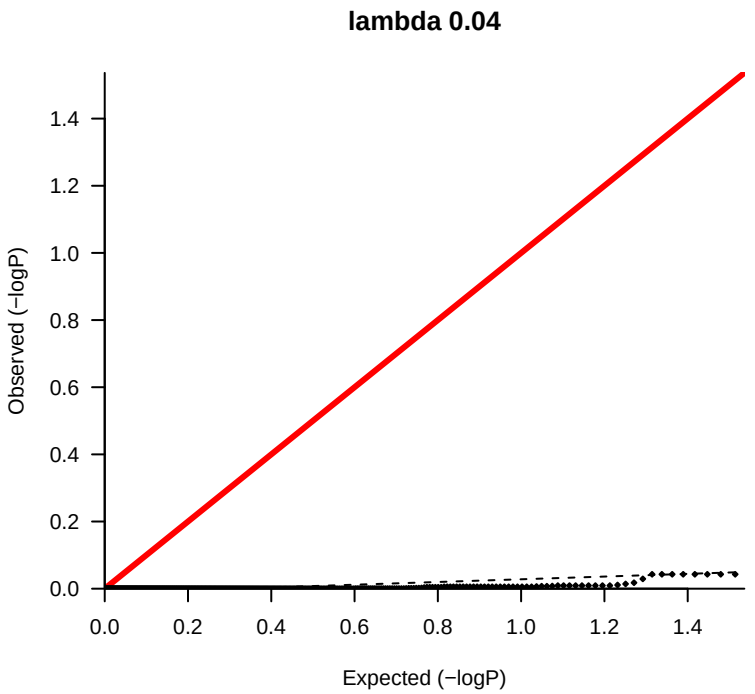

rootWidth60
Frequency
0
1
2
3
4
5
6
7
4.8
5.0
5.2
5.4

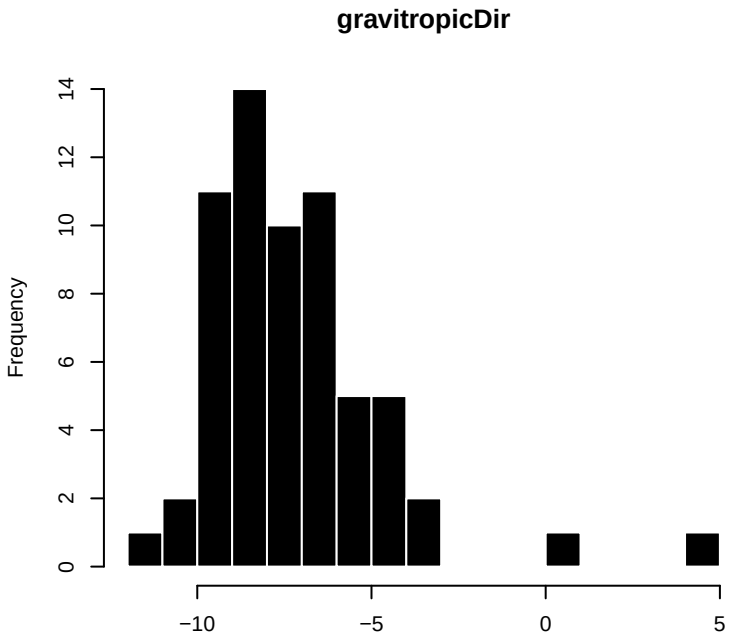

lambda 0.089
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

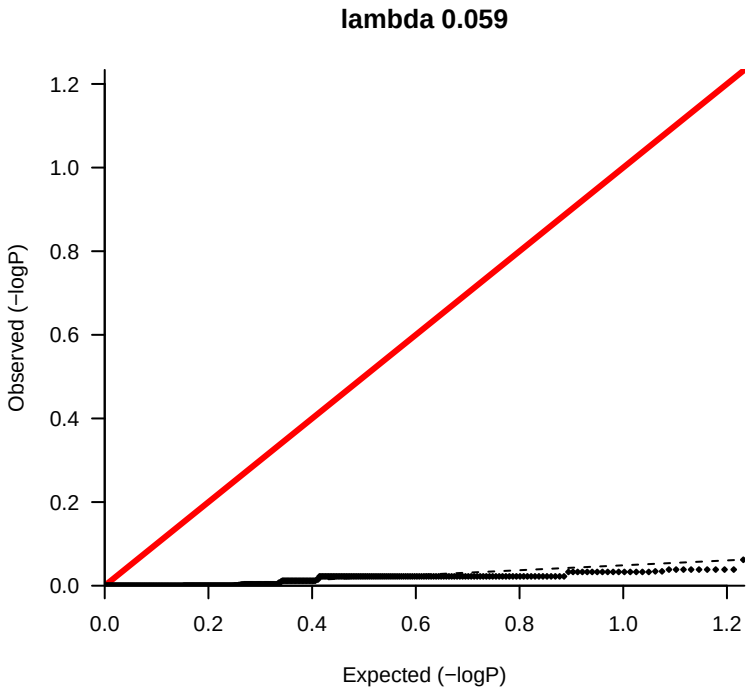

rootWidth80
Frequency
0
2
4
6
8
4.6
4.8
5.0
5.2
5.4

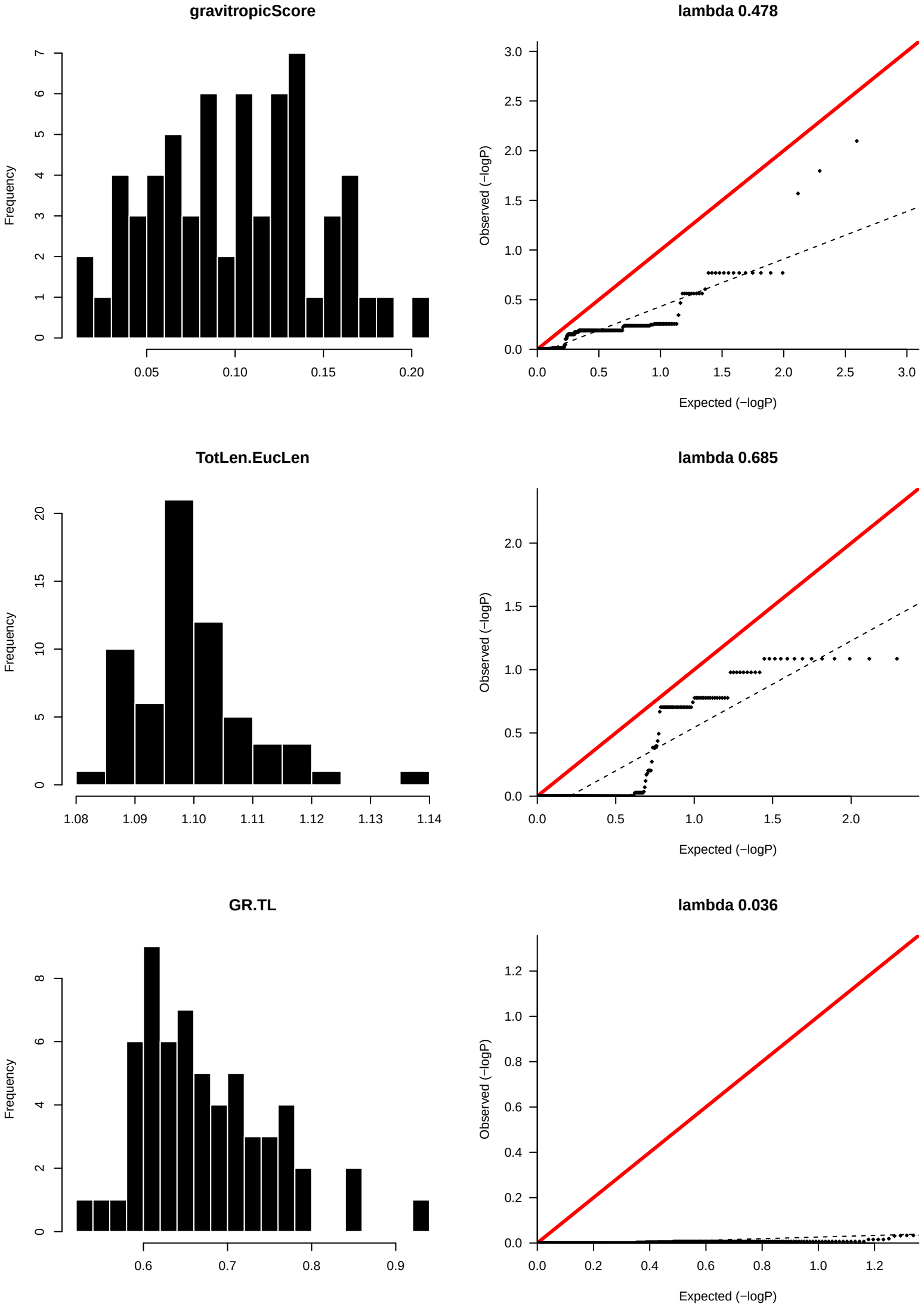

lambda 0.068
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

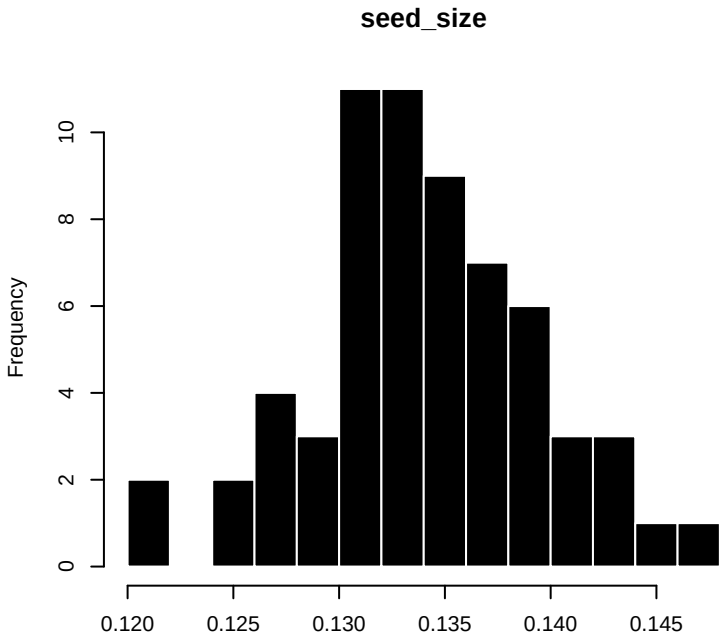

rootWidth100
Frequency
0
1
2
3
4
5
6
7
4.4
4.6
4.8
5.0
5.2

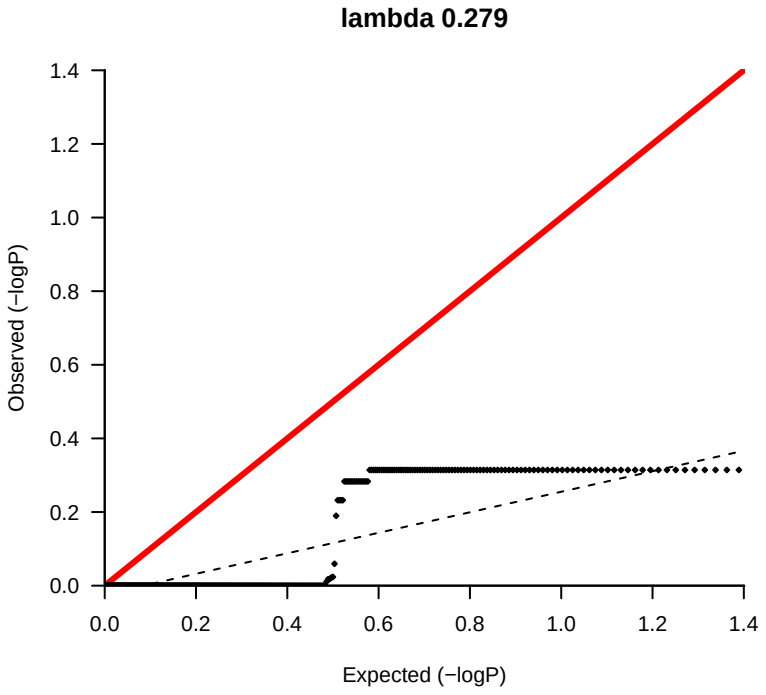

lambda 0.04
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

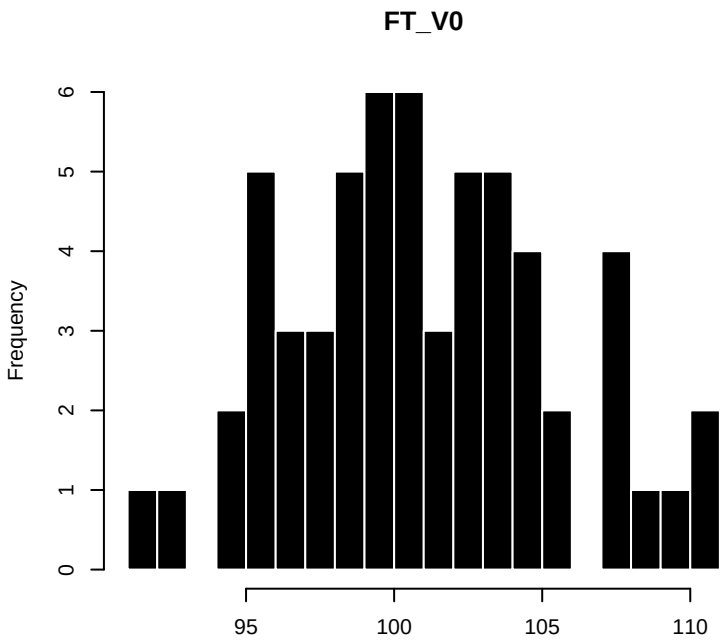

gravitropicDir
Frequency
0
2
4
6
8
10
12
14
−10
−5
0
5

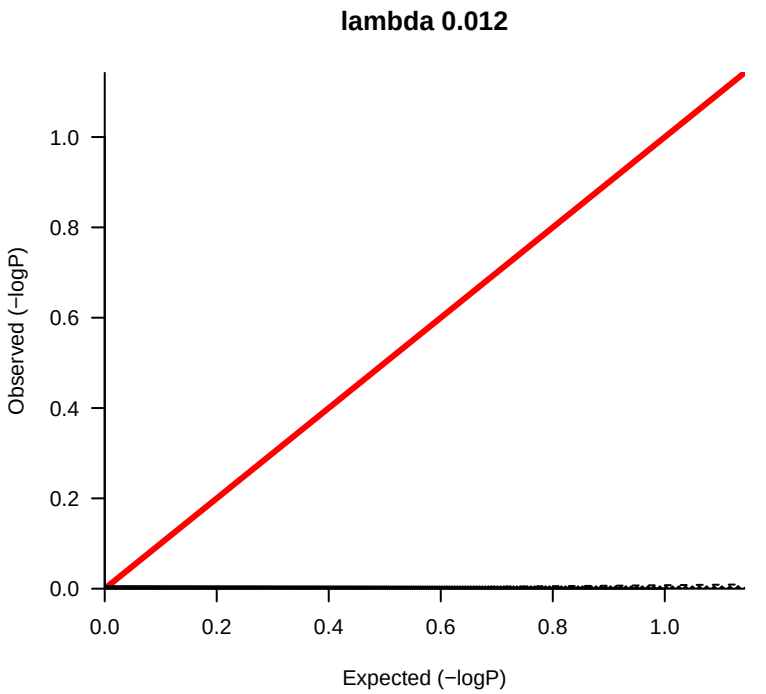

lambda 0.059
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2

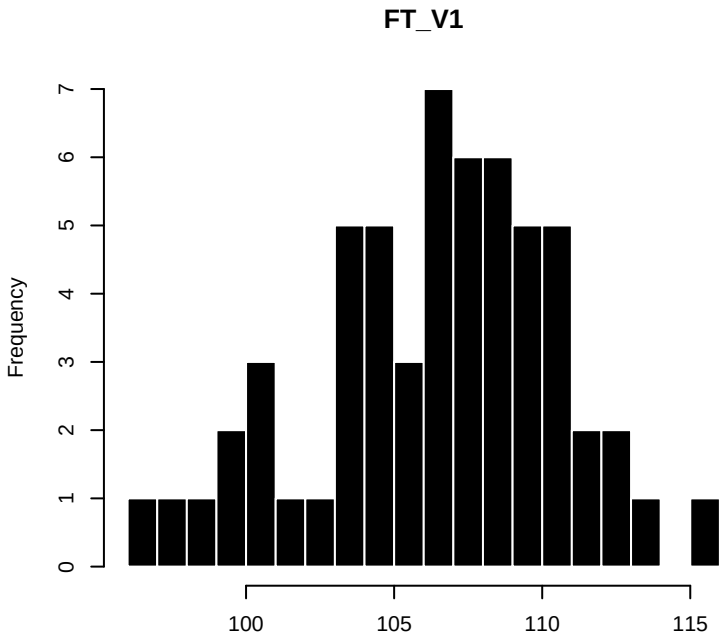

gravitropicScore
Frequency
0
1
2
3
4
5
6
7
0.05
0.10
0.15
0.20
lambda 0.478
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
TotLen.EucLen
Frequency
0
5
10
15
20
1.08
1.09
1.10
1.11
1.12
1.13
1.14
lambda 0.685
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
GR.TL
Frequency
0
2
4
6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

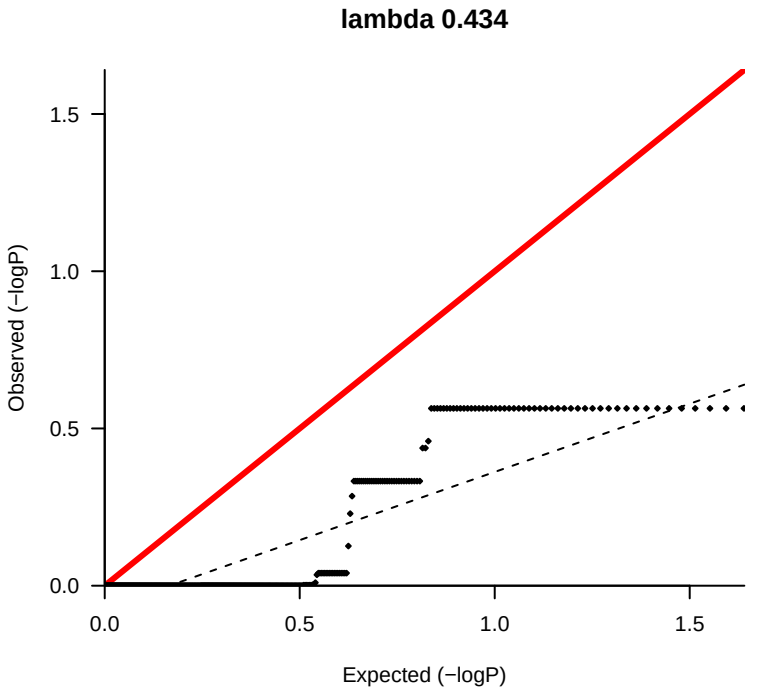

seed_size
Frequency
0
2
4
6
8
10
0.120
0.125
0.130
0.135
0.140
0.145

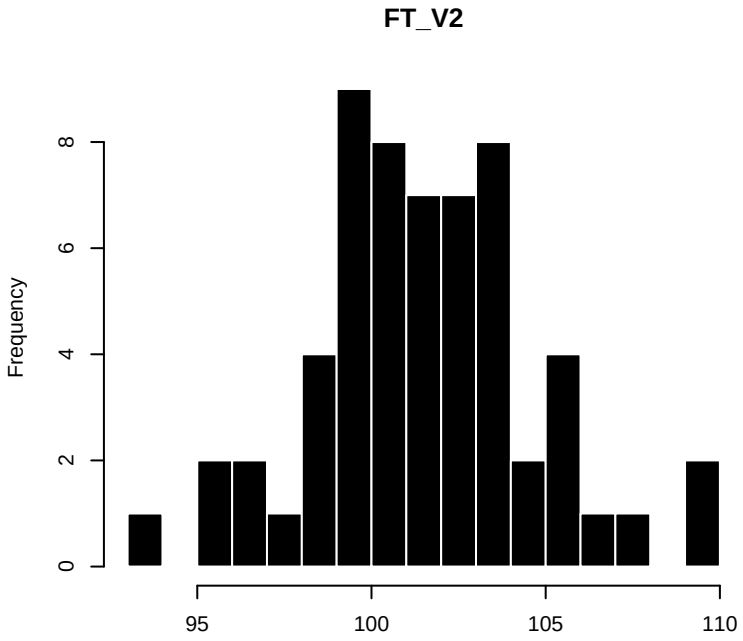

lambda 0.279
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

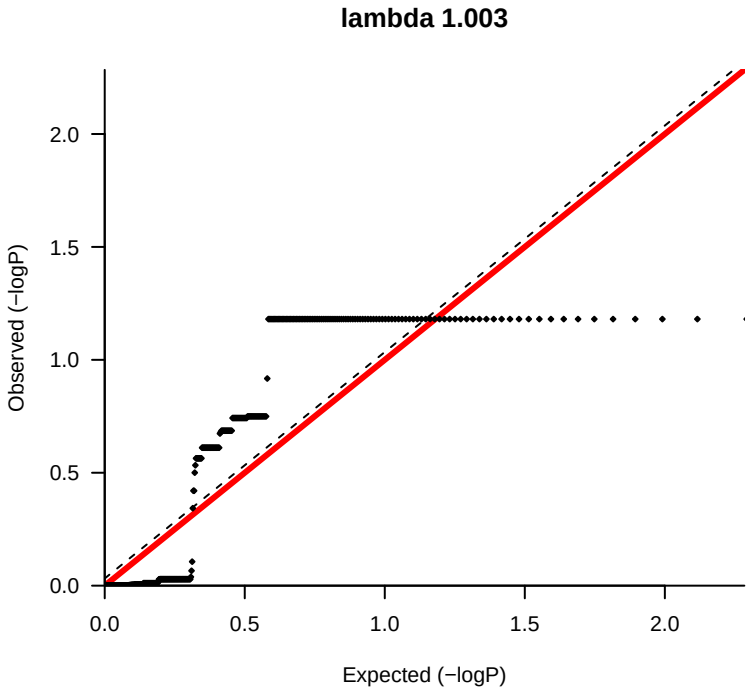

FT_V0
Frequency
0
1
2
3
4
5
6
95
100
105
110

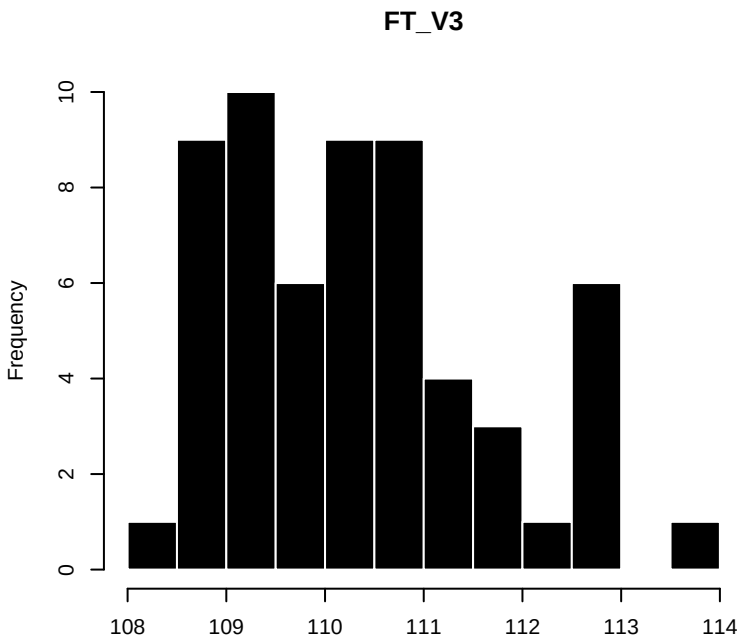

lambda 0.012
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0

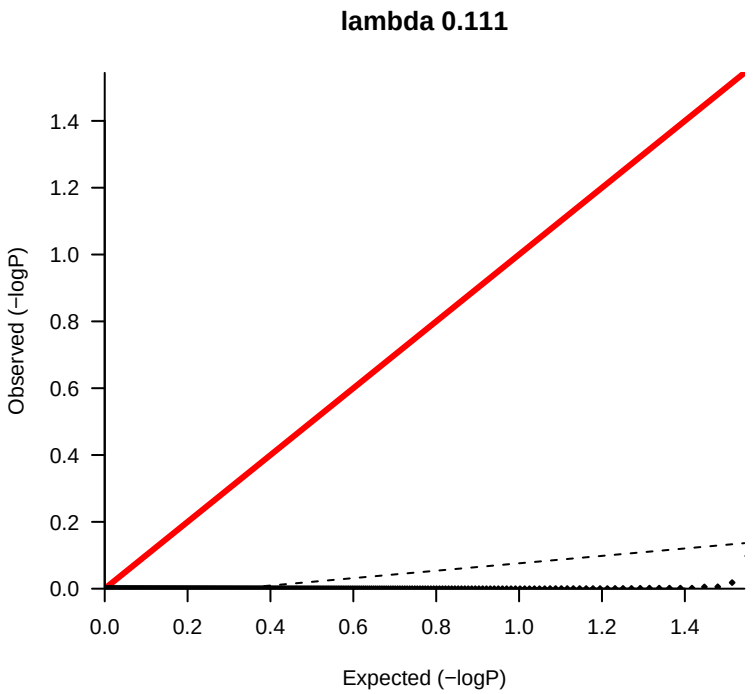

FT_V1
Frequency
0
1
2
3
4
5
6
7
100
105
110
115

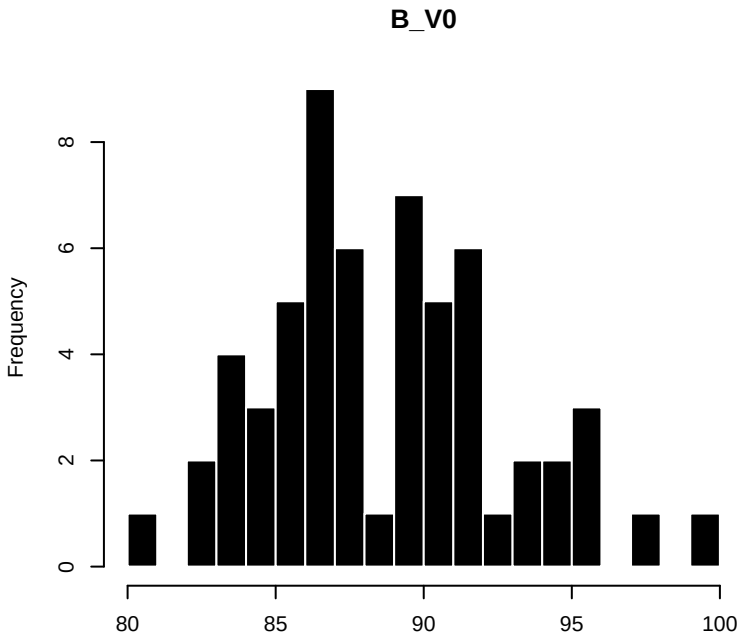

lambda 0.434
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

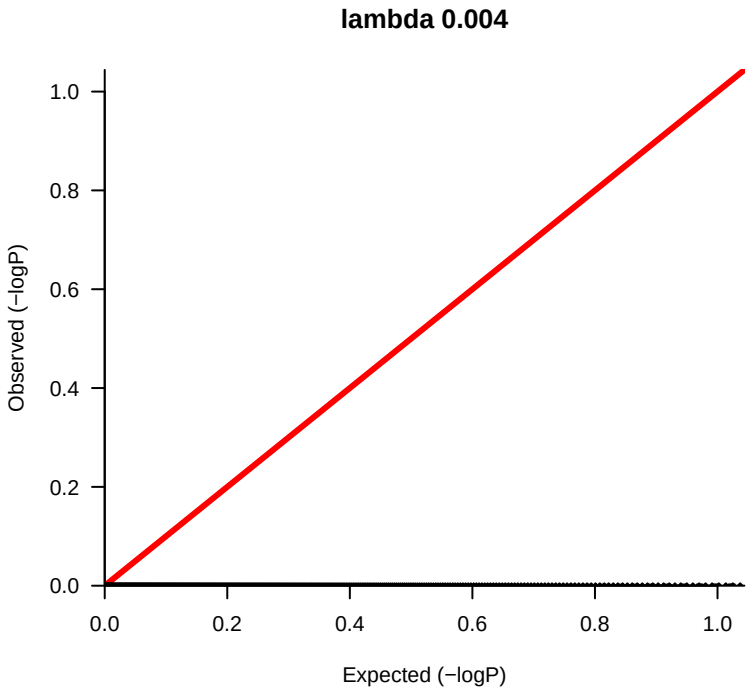

FT_V2
Frequency
0
2
4
6
8
95
100
105
110

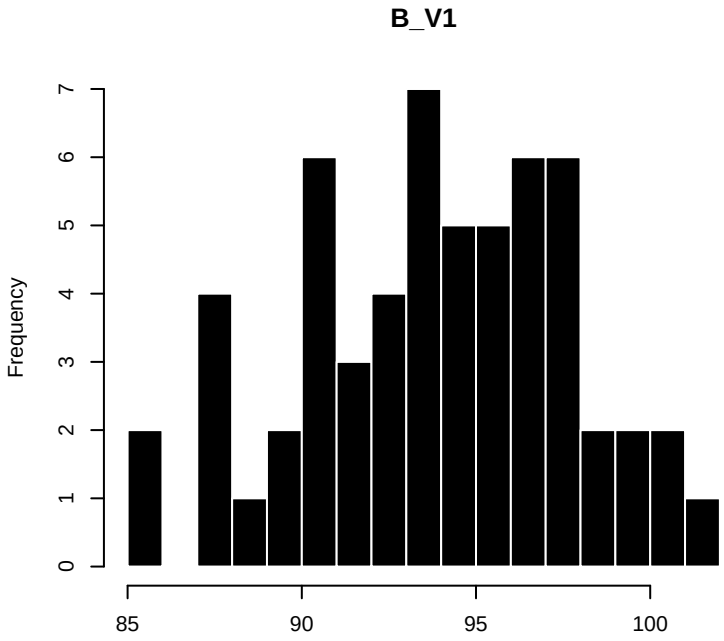

lambda 1.003
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

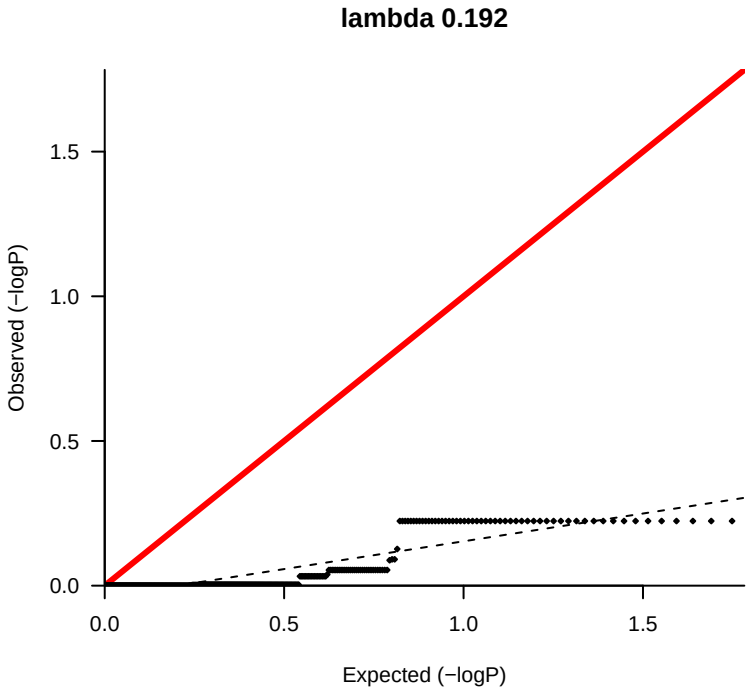

FT_V3
Frequency
0
2
4
6
8
10
108
109
110
111
112
113
114

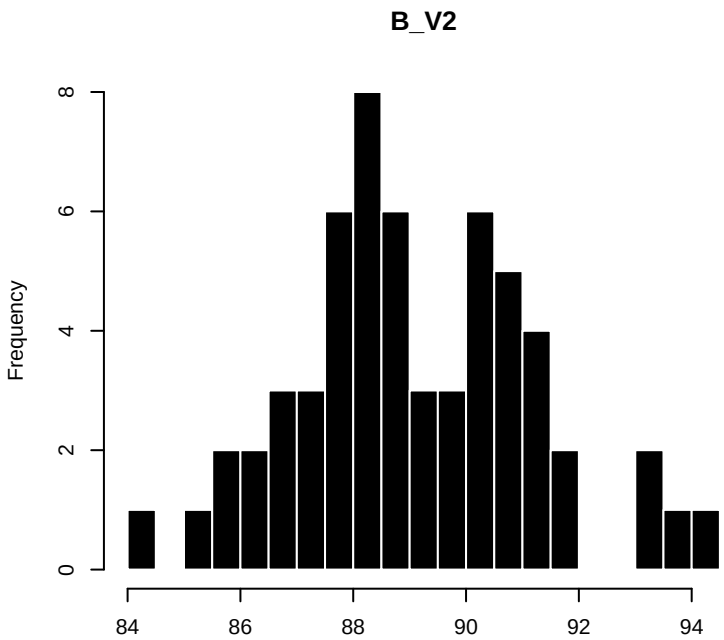

lambda 0.111
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

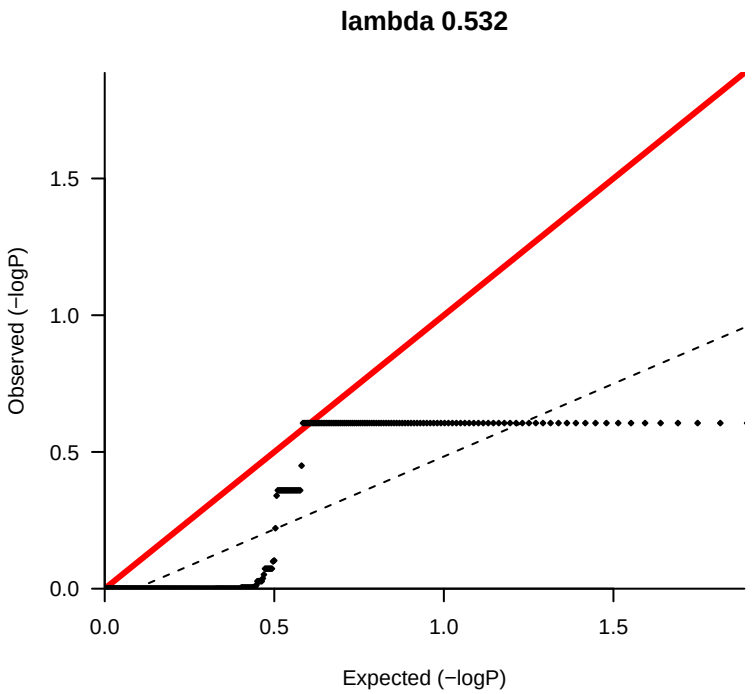

B_V0
Frequency
0
2
4
6
8
80
85
90
95
100

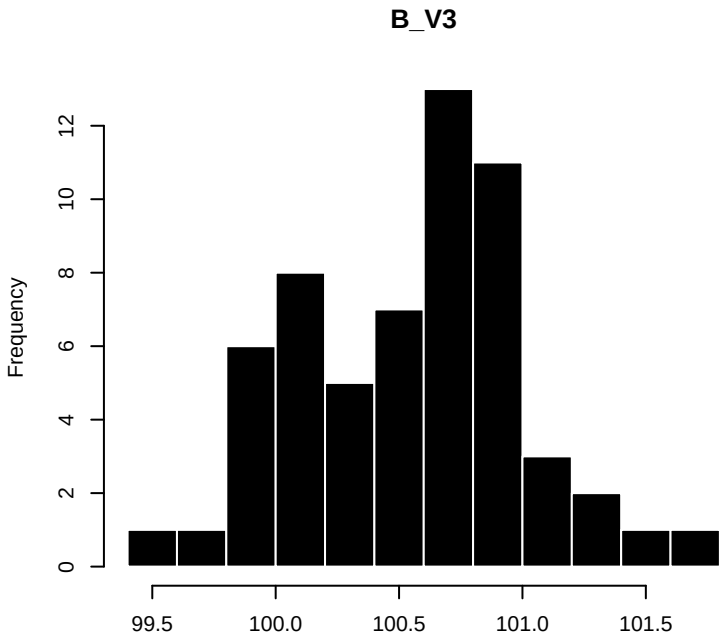

lambda 0.004
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0

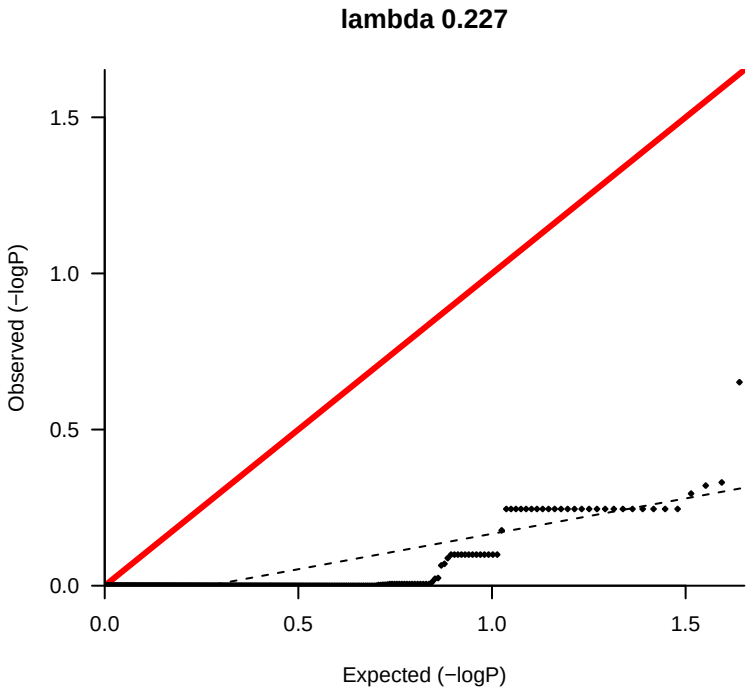

B_V1
Frequency
0
1
2
3
4
5
6
7
85
90
95
100

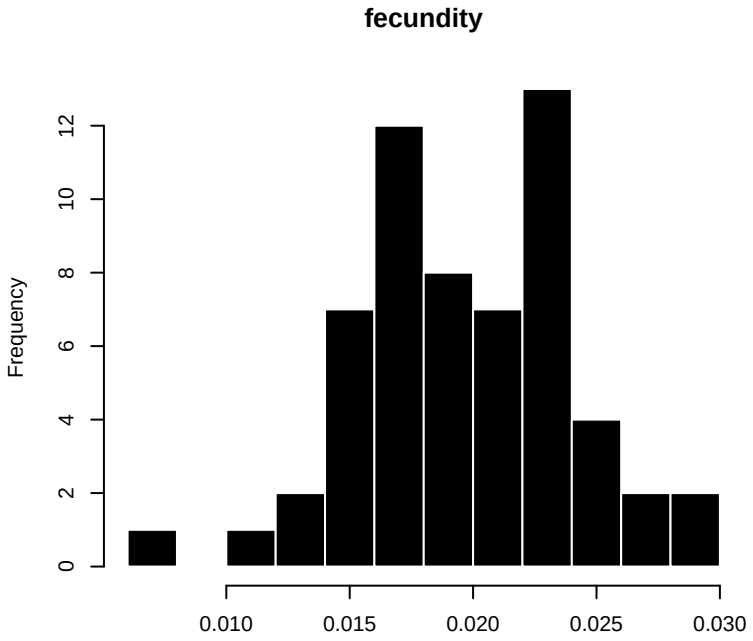

lambda 0.192
Observed (−logP)
0.0
0.5
1.0
1.5
Expected (−logP)

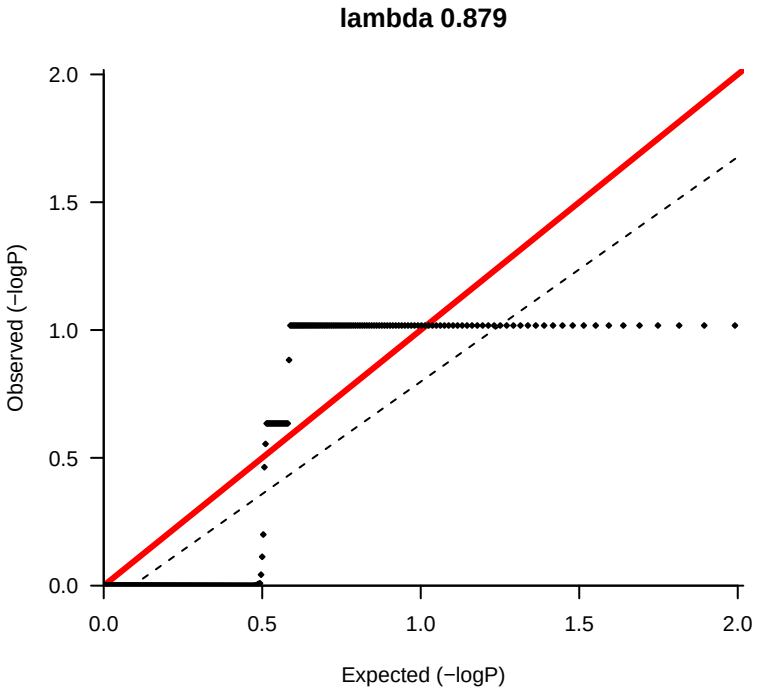

B_V2
Frequency
0
2
4
6
8
84
86
88
90
92
94

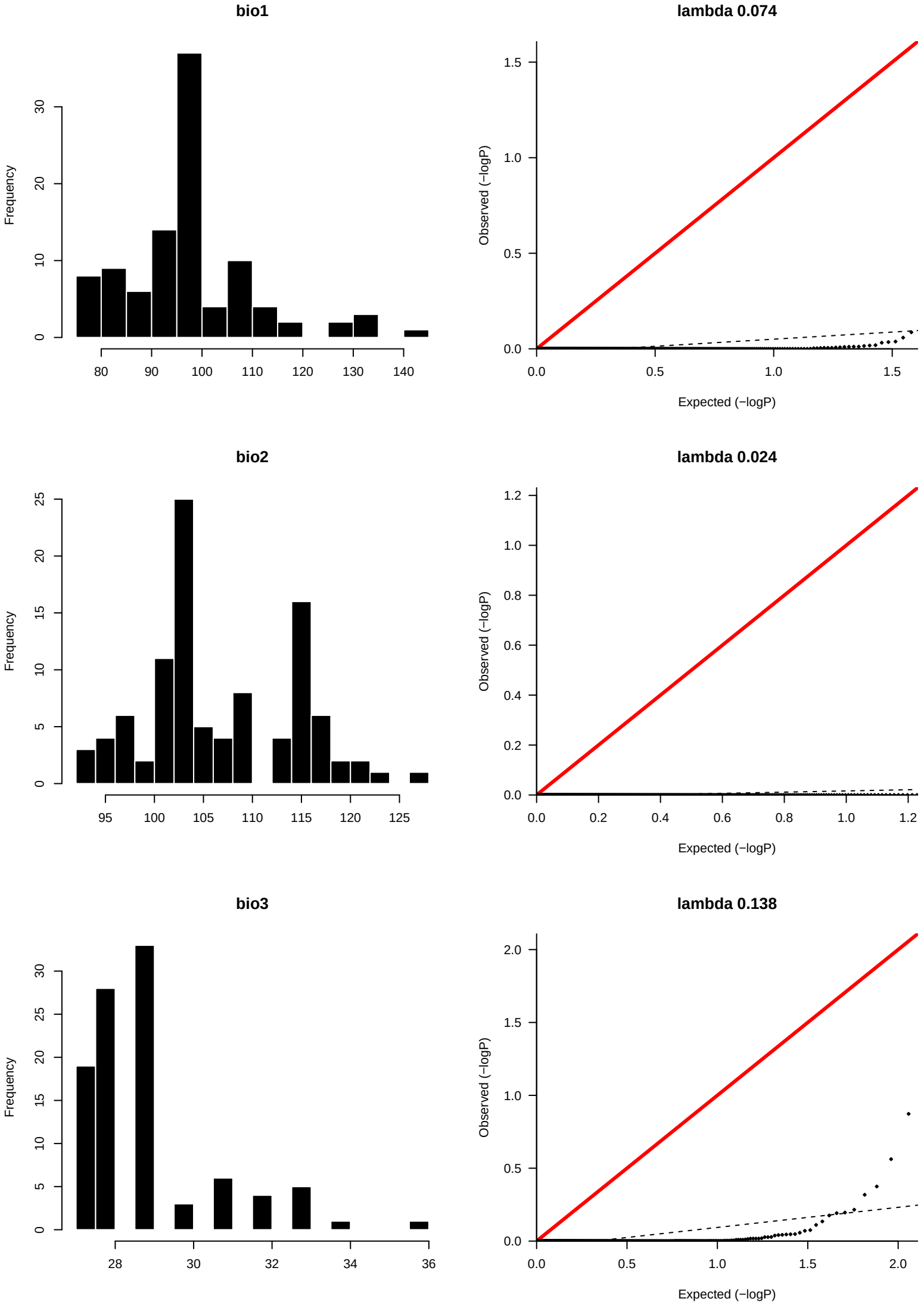

lambda 0.532
Observed (−logP)
0.0
0.5
1.0
1.5
Expected (−logP)

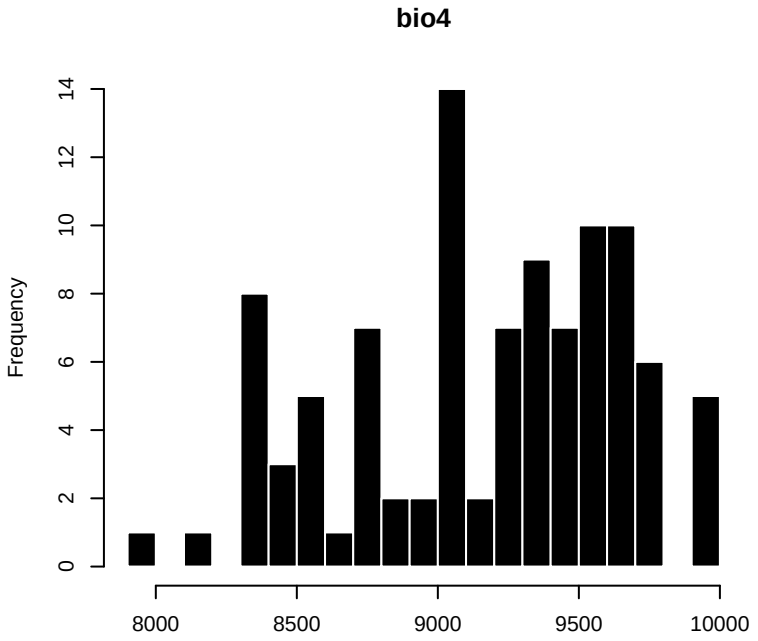

B_V3
Frequency
0
2
4
6
8
10
12
99.5
100.0
100.5
101.0
101.5

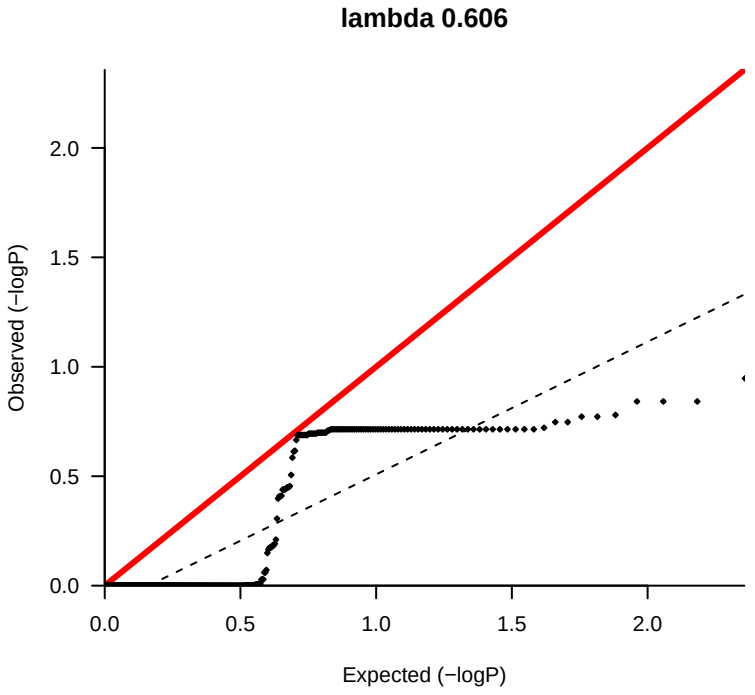

lambda 0.227
Observed (−logP)
0.0
0.5
1.0
1.5
Expected (−logP)

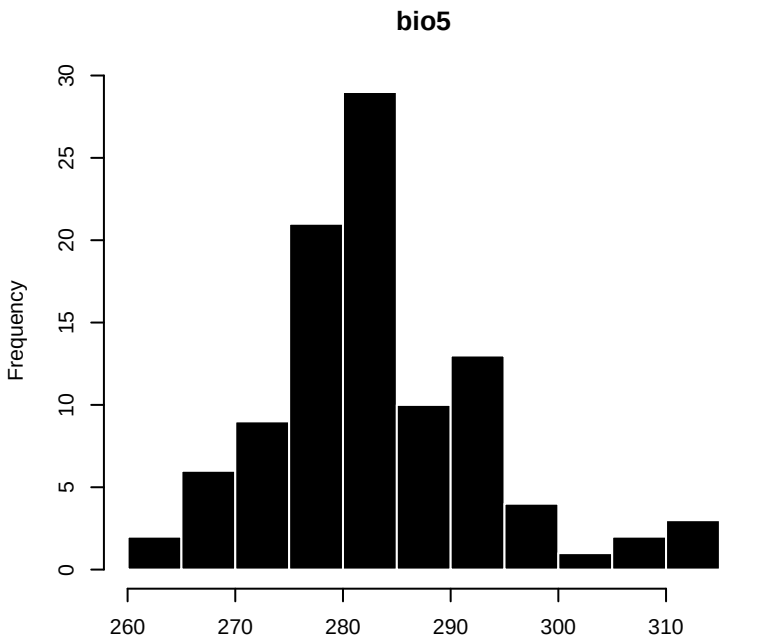

fecundity
Frequency
0
2
4
6
8
10
12
0.010
0.015
0.020
0.025
0.030

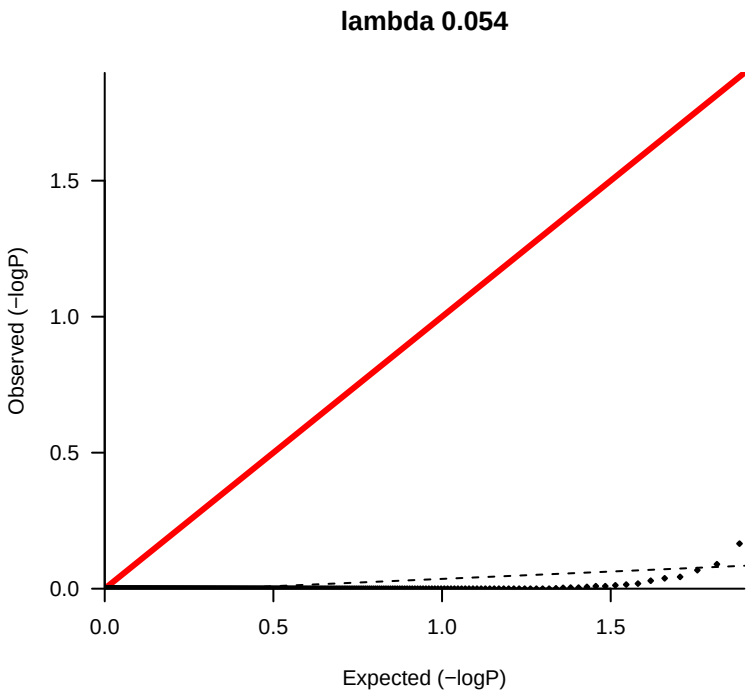

lambda 0.879
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

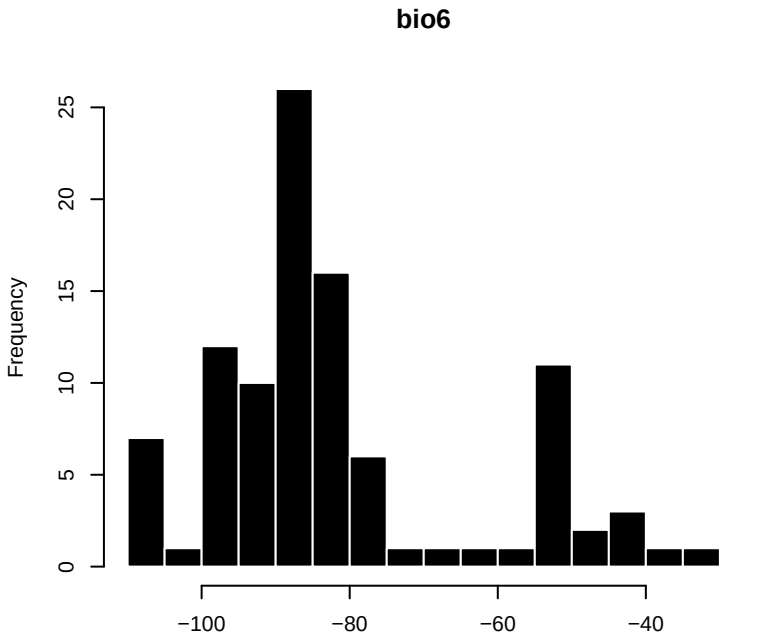

bio1
Frequency
0
10
20
30
80
90
100
110
120
130
140
lambda 0.074
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
bio2
Frequency
0
5
10
15
20
25
95
100
105
110
115
120
125
lambda 0.024
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
bio3
Frequency
0
5
10
15
20
25
30
28
30
32
34
36
lambda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

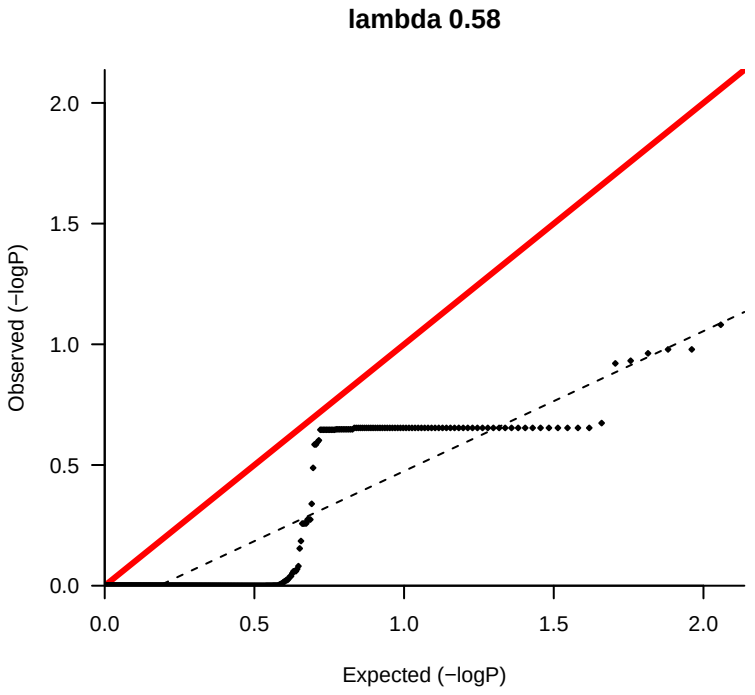

bio4
Frequency
0
2
4
6
8
10
12
14
8000
8500
9000
9500
10000

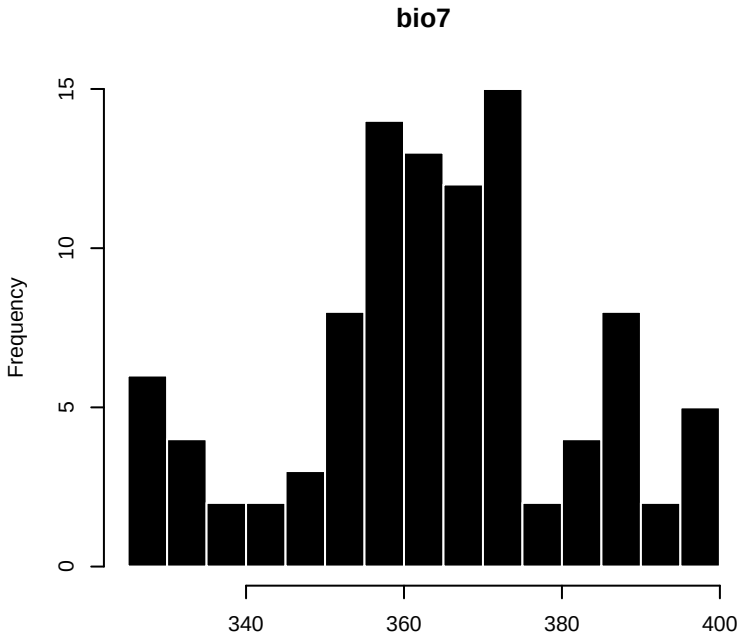

lambda 0.606
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

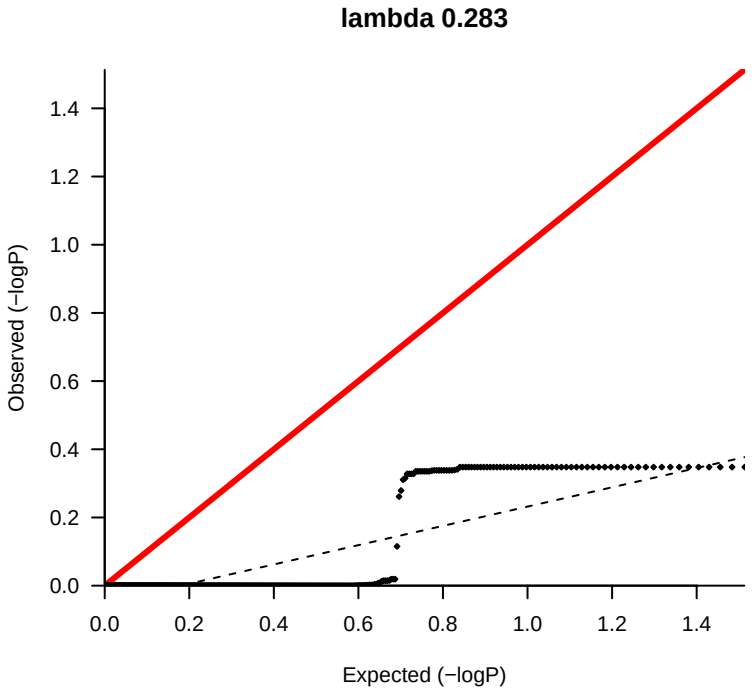

bio5
Frequency
0
5
10
15
20
25
30
260
270
280
290
300
310

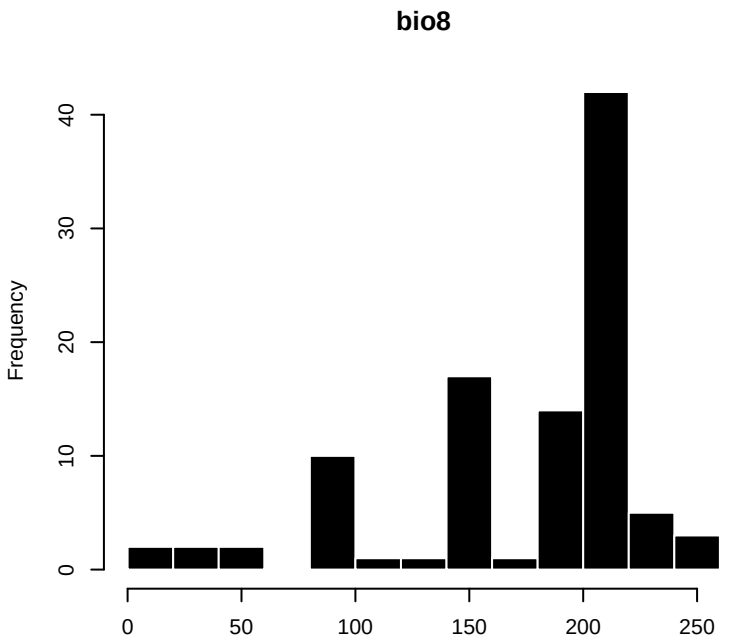

lambda 0.054
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

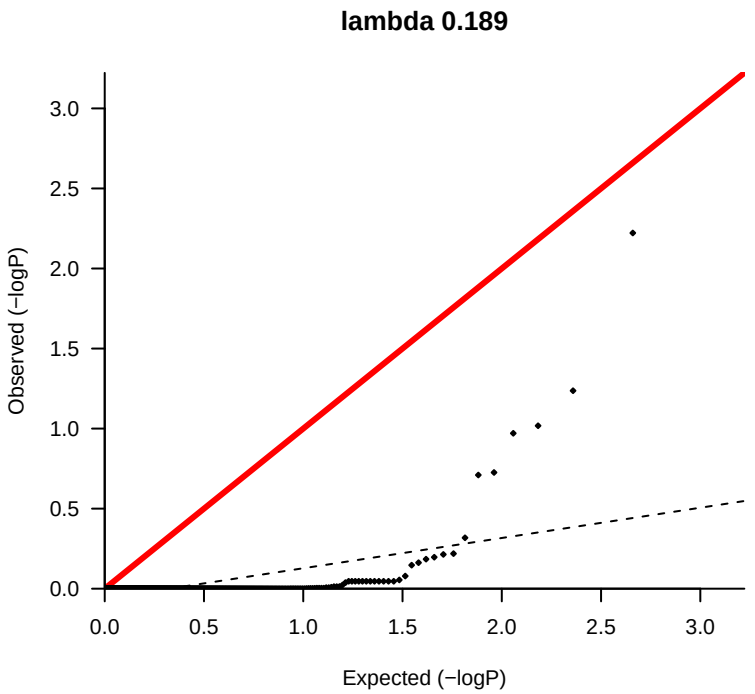

bio6
Frequency
0
5
10
15
20
25
−100
−80
−60
−40

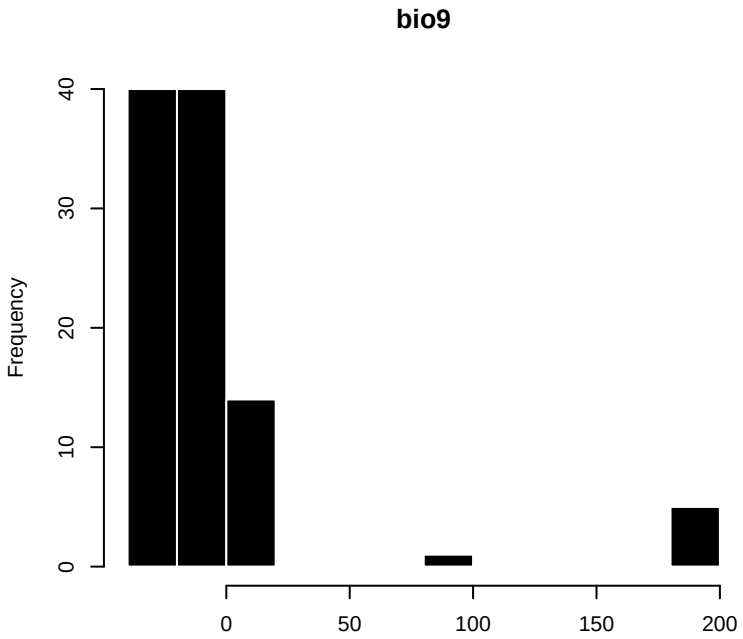

lambda 0.58
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0

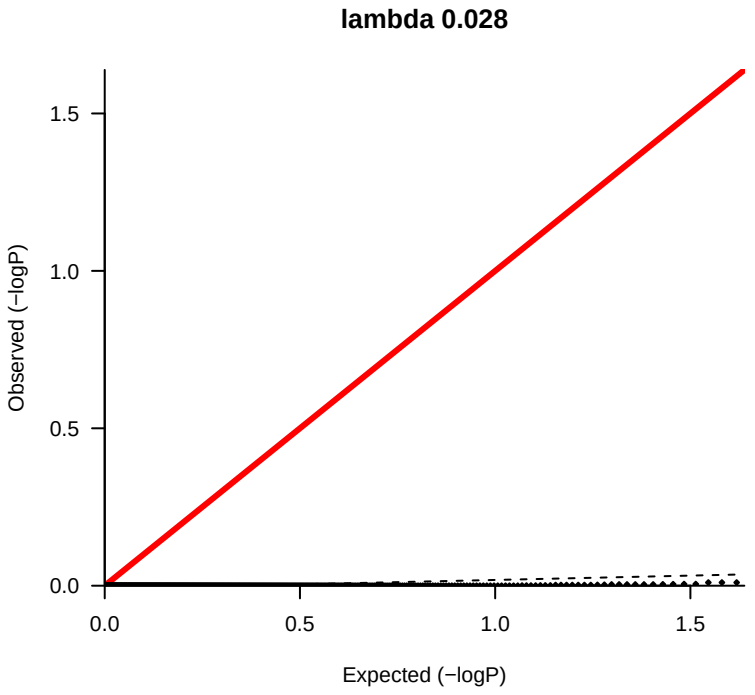

bio7
Frequency
0
5
10
15
340
360
380
400

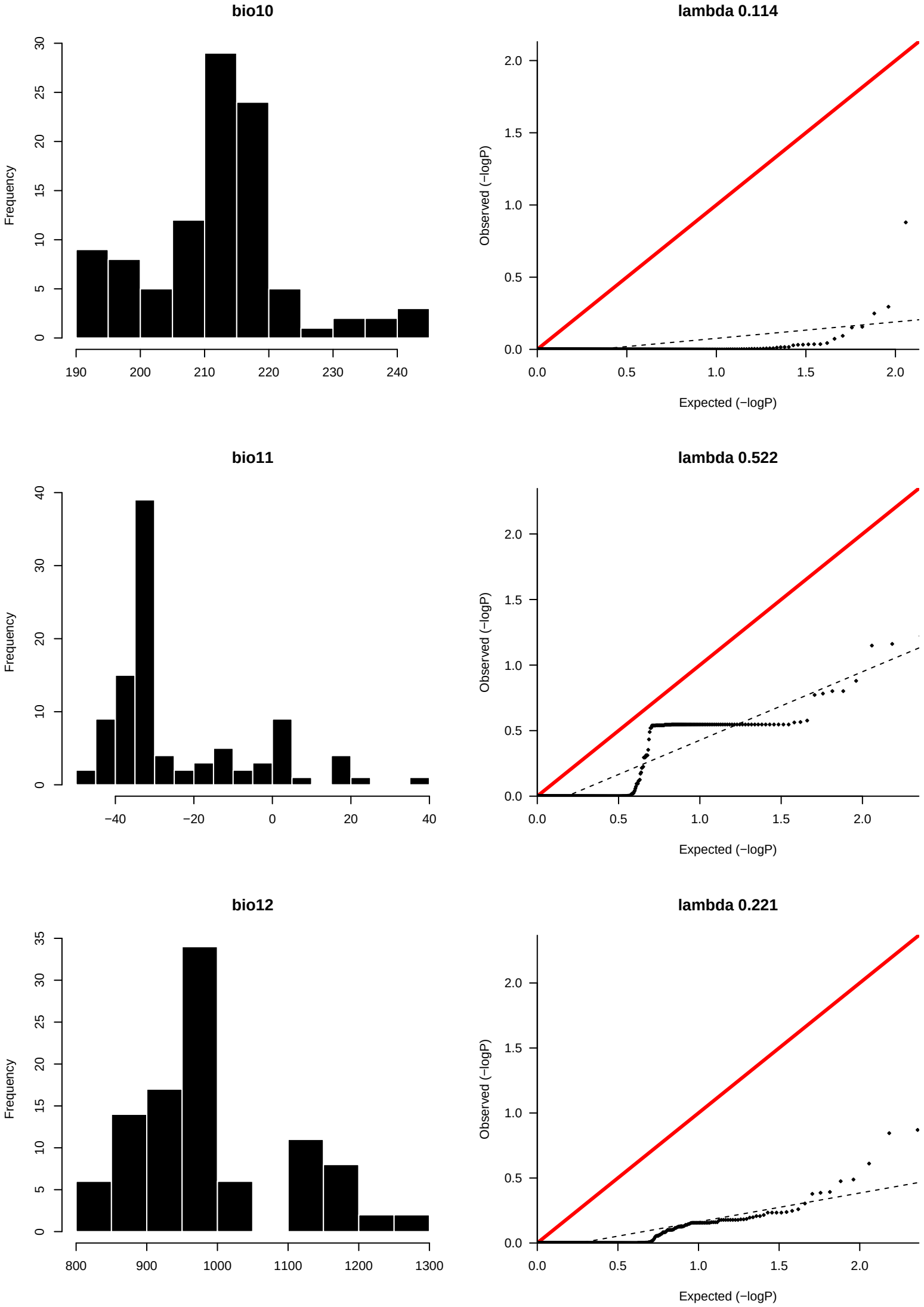

lambda 0.283
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

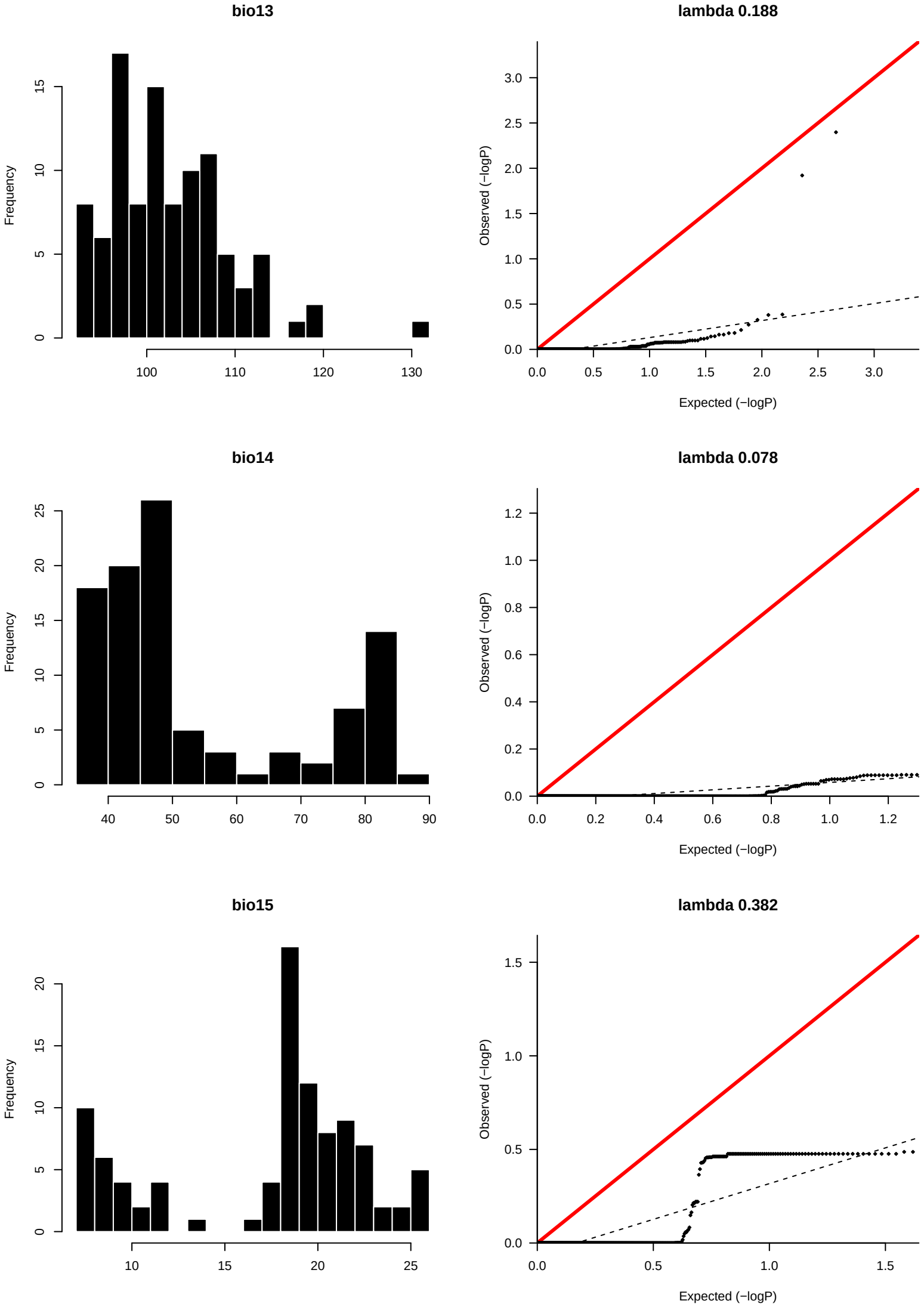

bio8
Frequency
0
10
20
30
40
0
50
100
150
200
250

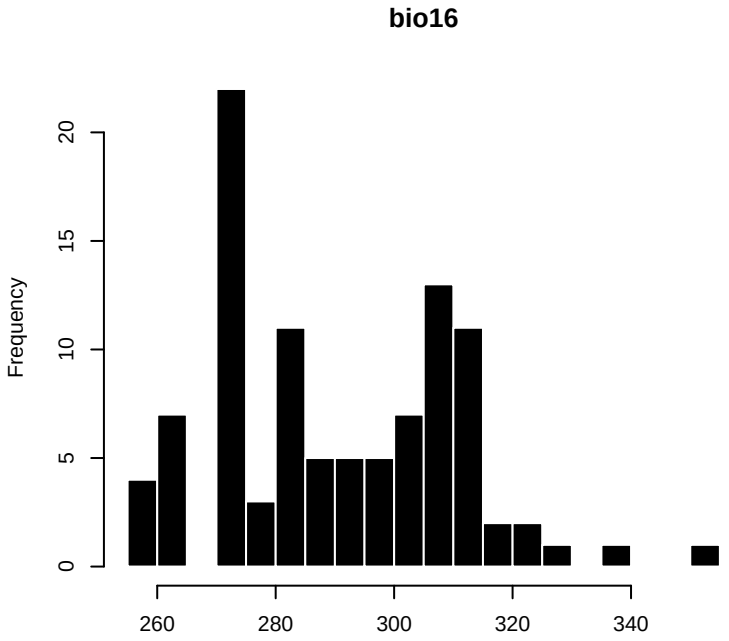

lambda 0.189
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

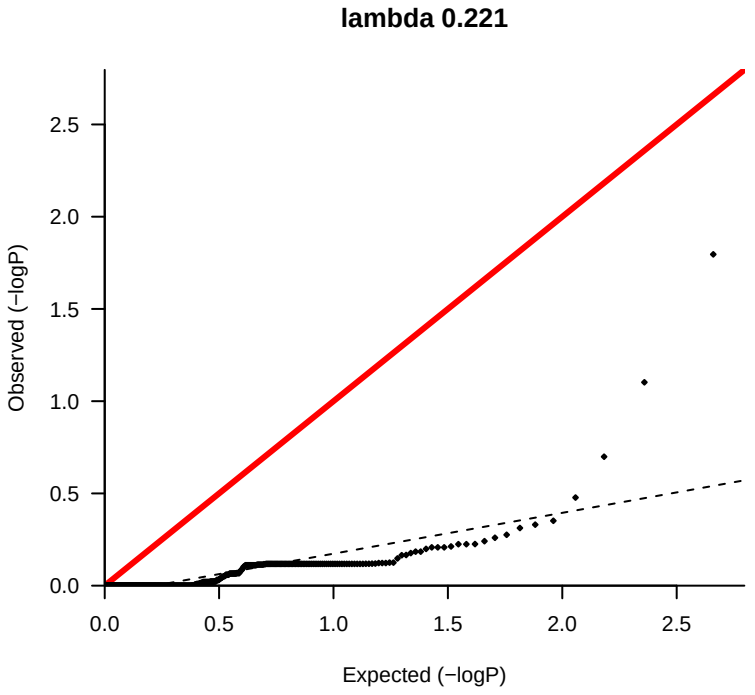

bio9
Frequency
0
10
20
30
40
0
50
100
150
200

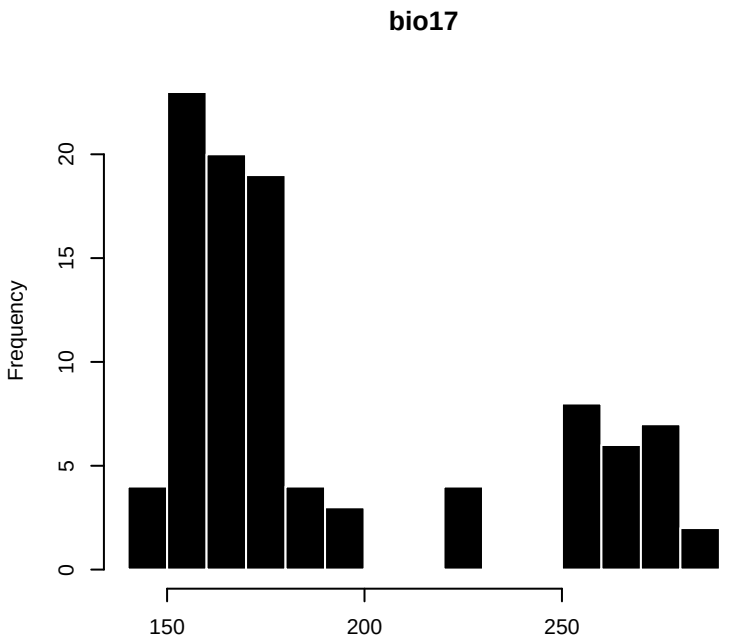

lambda 0.028
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5

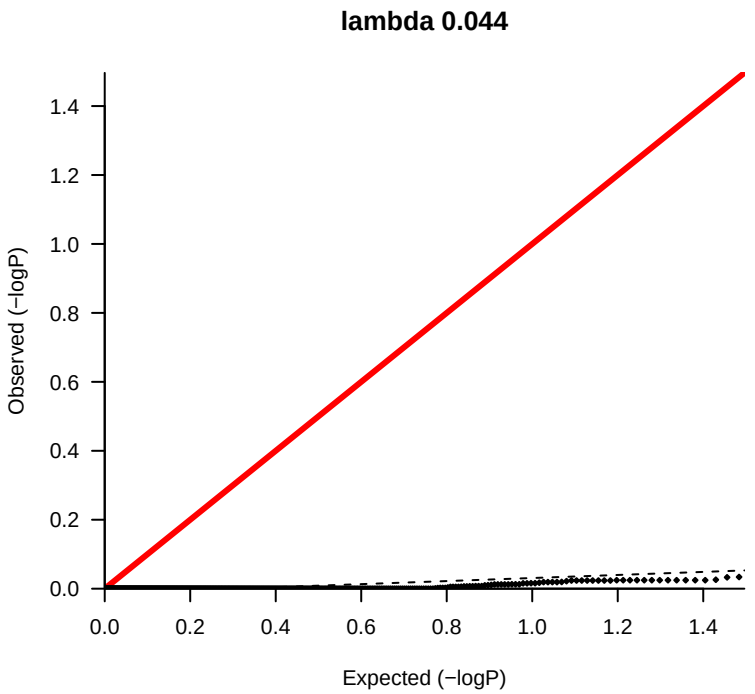

bio10
Frequency
0
5
10
15
20
25
30
190
200
210
220
230
240
lambda 0.114
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
bio11
Frequency
0
10
20
30
40
−40
−20
0
20
40
lambda 0.522
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
bio12
Frequency
0
5
10
15
20
25
30
35
800
900
1000
1100
1200
1300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

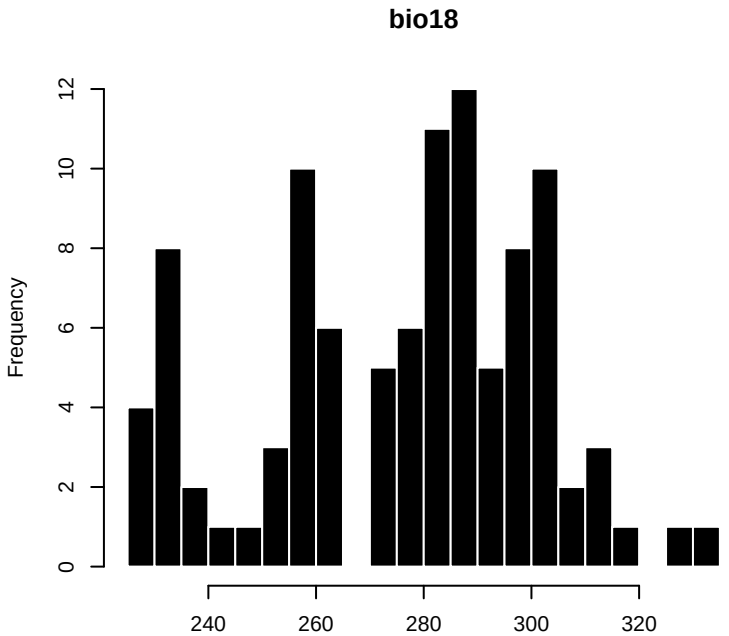

bio13
Frequency
0
5
10
15
100
110
120
130
lambda 0.188
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
bio14
Frequency
0
5
10
15
20
25
40
50
60
70
80
90
lambda 0.078
Observed (−logP)
Expected (−logP)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
bio15
Frequency
0
5
10
15
20
10
15
20
25
lambda 0.382
Observed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

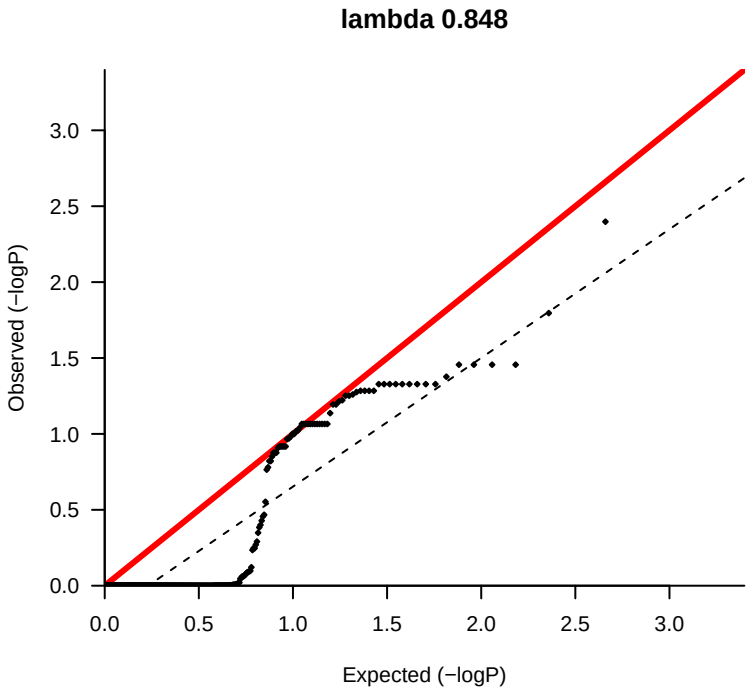

bio16
Frequency
0
5
10
15
20
260
280
300
320
340

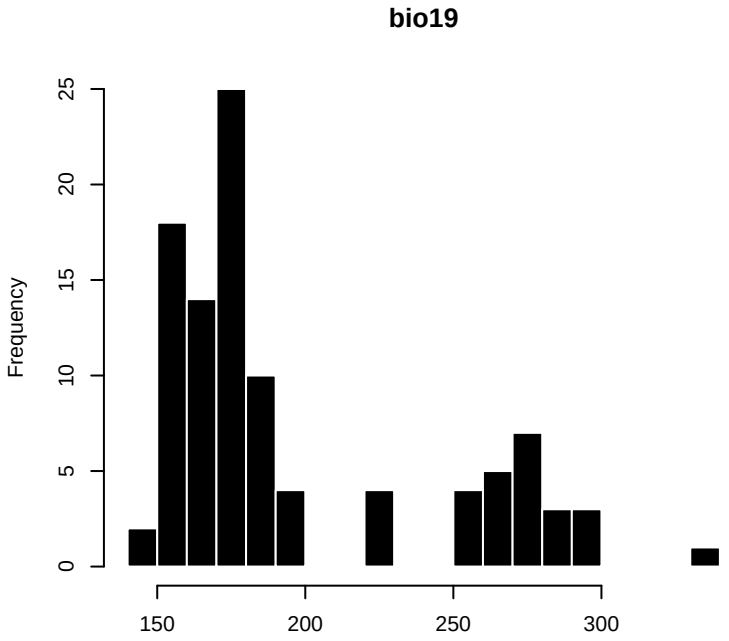

lambda 0.221
Observed (−logP)
Expected (−logP)
0.0
0.5
1.0
1.5
2.0
2.5

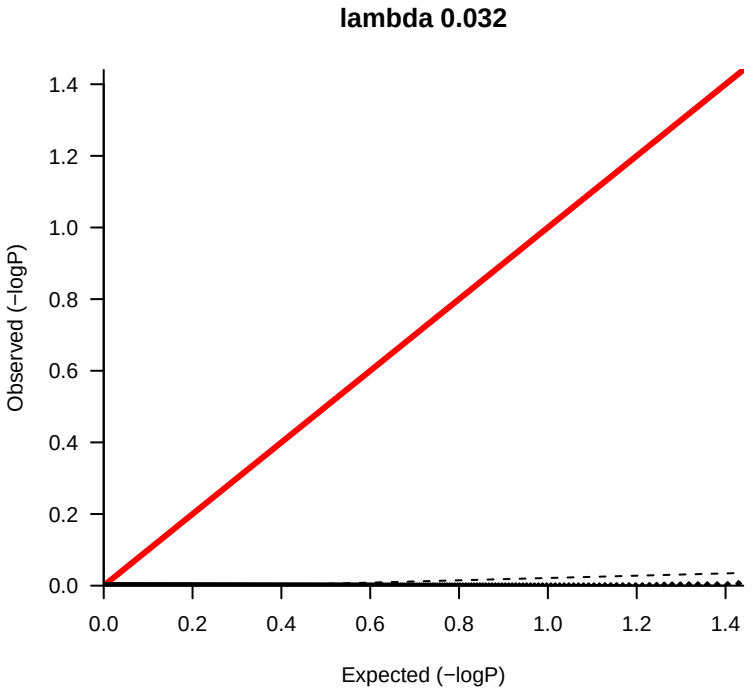

bio17
Frequency
0
5
10
15
20
150
200
250


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