Caveats and Precautions

In this document we discuss and acknowledge the limitations of GWA analyses, and describe our approaches to address these to the extent possible. We wish to make the reader and user of CLIMtools aware of these limitations, so that correlation is not mistaken for causation: GWA studies demonstrate associations whose potential causative and adaptive relationships require validation. The caveats associated with GWAS have been extensively described in the literature[1-3]. Major caveats are summarized in the main text; here we provide an explication of these limitations in the context of our study and explain how we have addressed them.

1. Population structure

Perhaps the most studied limitation of genome wide-association studies is the significant presence of false positives resulting from the confounding effects of populations structure[4-6]. That is, natural variation may exhibit an elevated allele frequency due to selective pressure from either environmental factors and/or the genetic background[7]. Distinguishing the former from the latter presents a common challenge for GWA studies on any organism[7]. In recent years, several approaches have emerged to describe the population structure of an organism, detect the extent of inflation of p-values obtained from GWAS analysis due to population structure, and correct for it. Mixed models, as used in our study, address population structure, family structure and cryptic relatedness in GWAS analysis, reducing the rate of false positives while maintaining statistical power[8,9]. Mixed models include both fixed and random effects. Including random effects allows incorporation of information about relationships among individuals by using the kinship matrix (K).

Arabidopsis displays a well-studied population structure[10-15]. Mixed models outperform both principal component analysis and genomic control in correcting for population structure[16], and in Arabidopsis, mixed models are described as the most effective method to correct for population struture[9,17,18]. Population structure is particularly a cause for concern when studying associations with traits that vary geographically, e.g. the latitudinal cline in flowering time; and mixed methods have been successfully applied many times to correct for population structure in such traits[2,14]. Associations with environmental variables, as studied here, also fall into this category. We, therefore used a mixed model[19] to ameliorate the confounding effects of population structure.

Demonstrating the effectiveness of mixed models in accounting for population structure was not the aim of our study, as this has been demonstrated by many previous studies and has been well-reviewed[1,7,9,17,18,20-22]. Nevertheless, here we exemplify the effectiveness of mixed models in correcting for population structure in the context of our study with an analysis of the 19 BIO variables obtained from WorldClim 2[23]. We chose these environmental variables given their extensive previous use (previously as WorldClim 1[24]) in conjunction with mixed models[14,25,26]. We analyzed these 19 environmental variables using two different approaches that do not correct for population structure: a non-parametric test, Kruskal-Wallis, and a linear regression model[27]. We compared the distribution of the obtained *p-values*, with those obtained from the mixed model that we used (Fig. 1). We found that, as expected, when population structure is not considered in the association analysis, there is significant statistical inflation that results in the increase of false positives (Type I error). Based on these analyses, we conclude that the use of a mixed model successfully reduces statistical inflation derived from population structure, confirming the results from diverse previous studies.[1,8,9]

Given that we correct for population structure using a mixed model, we may also expect that there will be some number of false negative results arising from over-correction when addressing the confounding effects of population structure, especially for any environmental pattern that follows a discrete geographical pattern (Fig. 2). Because we were more concerned with reducing the chance for Type I error than Type II error (false negatives), we retained the mixed model approach of correction for population structure.

2. Sampling

Because populations from Sweden and the Iberian Peninsula, with particular interest from ecological and evolutionary perspectives, are overrepresented in the 1,001 Genomes sample[14], we can expect that the existing genetic variation is best described in those populations relative to under-represented populations. While the sampling is therefore not perfect, the 1,001 Genomes nevertheless represents the best collection of sequenced organisms with ecological relevance to date. The 1,001 Genomes represents a vast improvement on the sample relative to previous sequencing efforts in this species,[28] as it deliberately includes accessions from relevant locations based on prior knowledge of the population structure of the species[11-14,29,30]. As a result, the 1,001 Genomes provide a "*hierarchical collection of accessions with a range of geographic distances between nearest neighbors, and a few very densely sampled locales*[14]".

We include 879 genotypes (accessions) in our in silico analysis, far more than are feasible to include in typical wet bench GWAS on plant phenotypes. This provides the power to uncover associations with genetic variants that would have been missed in the case of standard experimental GWAS, conducted on subsets of Arabidopsis accessions. Given the large sample size and distribution, including very geographically distant accessions, our analysis has an advantage and a caveat. The massive amount of genetic and environmental variance included in the analysis, derived from a very large and widespread sample, increases its the power and confidence in the resulting candidates. The caveat has to do with an increased proportion of false negatives due to genetic heterogeneity. This is because in some cases, more than one co-correlated genetic variant, that differ in frequencies among different populations, underlie the same trait. Such would be the case of two different SNPs affecting the same codon, or two different SNPs causing the same effect on the same gene. This will result in a weakened correlation for each SNP that will yield an increased proportion of false negatives[1].

3. Sequencing data

Previous attempts to describe Arabidopsis genetic variation associated with the environment[28], used SNP arrays available at the time[26]. Despite the great value of these studies, it is challenging to argue that the genetic variance identified using SNP arrays reflects causality, for the simple reason that the vast majority of the genetic variance present in the sequenced organisms was not represented on the arrays. For this reason, SNP arrays over-identify non-causal SNPs that are in linkage disequilibrium with causative variants not present in the array. To exemplify this point, we conducted GWAS analysis on the 374 accessions used in our study of fully sequenced genomes[14],

that were also previously genotyped using the 250K SNP array[28]. For this purpose, we conducted a mixed model analysis as described before using both the full genomic sequences and the SNP array sequence datasets. We used the 19 BIO variables obtained from WorldClim 2[23] which were part of our study, given they were also used (as the previous version, WoldClim 1[24]) in analogous previous studies[14,25,26]. We found that the genetic variation present for these 374 accessions consisted of 205,936 SNPs when we used the 250K array; these SNPS constituted just 3.2% of the 6,432,557 SNPs presents in the fully sequenced genomes for the exact same accessions. After filtering for common variants (MAF>5%), 90.5% of the SNPs present in the fully sequenced genomes were missing in the 250K SNP array for these 374 accessions. Especially since there was no a priori information of causality used to identify the SNPs that were included on the SNP arrays, is difficult to argue that genetic variation within the 250K dataset and the associations obtained from it are likely to be meaningful causal variants. In Fig. S1 of the main text, we depict how the use of fully sequenced genotypes increases the numbers of significant association peaks that are missing from SNP array analysis and, conversely, reveals incomplete significant SNP "towers" that the 250K SNP array revealed as significant, but that were really in linkage disequilibrium with missing significant SNPs that were identified when using the fully sequenced dataset.

4. Synthetic associations

Synthetic associations are the association of non-causative markers that are in linkage disequilibrium with causative markers. In CLIMGeno, we provide the tools to explore the scores, q-values, allele frequencies and the predicted effect of any identified variant. Perhaps more importantly, to facilitate the identification of synthetic associations we provide the ability to select a genomic window of a customizable length around any variant of interest to visually explore the possible presence of synthetic associations.

5. Multiple comparisons

For every hypothesis tested, there is an inherent risk of accepting a hypothesis that is false. To reduce the probability of accepting a false positive, researchers usually define a level of significance for which they determine that the probability of obtaining a type I error is acceptably low. An issue with all GWA studies is that they do not test one single hypothesis, but rather a vast number of hypotheses, testing the association of hundreds of thousands of SNPs with the trait of interest. Accordingly, the presence of false positives in any GWAS is unavoidable. As described above, full-genome sequencing data facilitates the discovery of causal variants, but there is also a downside. This has to do to with the fact that the use of full-genome sequences vastly increases the number of SNPs tested, thus increasing type I error. Given this, it becomes important to estimate and control for the rate of false positives in such studies[31].

Accordingly, we followed a number of steps to reduce the number of variants considered, reduce type I error, and prioritize the most likely causal SNPs in the association lists that we provide: we filtered out SNPs in transposons, SNPs in intergenic regions, and SNPs in genic regions with a frequency < 5%. Furthermore, we include the information on the predicted effects of each SNP (e.g., synonymous vs. non-synonymous), locus annotations and description, as well as allele frequencies corresponding to all associated variants to facilitate a more informed prioritization of candidates.

Most importantly, to control type I error, we applied the qvalue[31] package in R using the Benjamini-Hochberg approach to calculate and provide q-values of the associated SNPs so that the user can impose any desired FDR threshold. We used each of the 19 BIO variables obtained

from WorldClim 2[23] to exemplify how the user can estimate the number of significant associations versus each q-value cut-off, and the number of expected false positives versus the number of significant tests (Fig. 3). We find that from a typically used FDR of 5%, we can expect a limited although always present number of estimated false positives, but also a very substantial number of significant outcomes. The user of CLIMtools may explore any of these parameters using FDRCLIM, which implements the qvalue R package[32-35] to calculate the FDR of all of the ExG associations described in our study.

Despite our measures to reduce, describe, and control type I error derived from multiple comparisons, users of our data should be aware that it is not possible to completely eliminate false positives in any GWA study.

6. Binary data

In the case of binary data, users should be aware that there is an inflation of p-values relative to continuous data[36] that should be considered when choosing a score threshold to select candidate associations, or to compare the strength of association of any variant of interest with other associated environmental variables[37]. In our data, 6 out of the 204 environmental variables provided reflect variables extracted from environmental categories, such as "Köppen-Geiger Csa hot summer Mediterranean climate" that were evaluated in terms of binary data. In this example accessions collected from hot, summer, Mediterranean climates were evaluated as 1, relative to all other accessions that were scored as 0. In these instances, the user should exercise caution given the exaggerated p-values obtained from GWAS analysis of binary data types.

7. Limitations addressed by removing rare variants

The power of GWAS depends on the phenotypic (and/or environmental) variance within the studied population as explained by the SNPs found in association. For this reason, rare variants, present in a limited number of individuals in the population, present problems for GWAS[38,39]. Other statistical caveats are relevant when considering rare variants as mixed models, which are effective for finding associations with common variants, have been found to be susceptible to spurious associations with variants with rare allele frequencies[21]. Additionally, rare variants are prone to be found in association with many other rare variants more often than with common variants[40]. For example, all variants present in just one Arabidopsis accession will necessarily be found in association with each other. Because here we are interested in likely adaptive variants, found in higher frequencies, we addressed this limitation by filtering out uncommon variants (MAF < 5%).

Despite great improvements in sequencing technologies, the error rate of Next Generation Sequencing is approximately 1%[41], to which is added the inherent human error rate derived from obtaining and manipulating sequencing data from diverse research facilities[42,43]. Our removal of variants with MAF < 5% also reduces the chances of inclusion of these errors in our datasets.

Literature Cited

1. Korte, A. & Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**, 29 (2013).
2. Atwell, S., Huang, Y.S., Vilhjalmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M. & Hu, T.T. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627-31 (2010).

3. Platt, A., Vilhjalmsson, B.J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045-52 (2010).
4. Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G. & Hirschhorn, J.N. Demonstrating stratification in a European American population. *Nature Genet* **37**, 868-72 (2005).
5. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037-48 (1994).
6. Hey, J. & Machado, C.A. The study of structured populations--new hope for a difficult and divided science. *Nat Rev Genet* **4**, 535-43 (2003).
7. Vilhjálmsson, B.J. & Nordborg, M. The nature of confounding in genome-wide association studies. *Nat Rev Genet* **14**, 1-2 (2012).
8. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. & Buckler, E.S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**, 203-8 (2006).
9. Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C. & Marjoram, P. An Arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**, e4 (2007).
10. Abbott, R.J. & Gomes, M.F. Population genetic structure and outcrossing rate of *Arabidopsis thaliana*. *Heredity* **62**, 411-18 (1989).
11. Beck, J.B., Schmuths, H. & Schaal, B.A. Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Mol Ecol* **17**, 902-15 (2008).
12. Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Ågren, J., Bossdorf, O., Byers, D. & Donohue, K. The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet* **6**, e1000843 (2010).
13. Schmid, K.J., Torjek, O., Meyer, R., Schmuths, H., Hoffmann, M.H. & Altmann, T. Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers. *Theor Appl Genet* **112**, 1104-14 (2006).
14. Consortium, G. 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481-91 (2016).
15. Hoffmann, M.H. Biogeography of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae). *J Biogeogr* **29**, 125-34 (2002).
16. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C. & Eskin, E. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-54 (2010).
17. Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M. & Holland, J.B. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**, 203-8 (2006).
18. Korte, A., Vilhjalmsson, B.J., Segura, V., Platt, A., Long, Q. & Nordborg, M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* **44**, 1066-71 (2012).

19. Seren, Ü., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K. & Korte, A. AraPheno: a public database for *Arabidopsis thaliana* phenotypes. *Nucleic Acids Res* **45**, D1054-D1059 (2017).
20. Hoffman, G.E. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PLoS ONE* **8**, e75707 (2013).
21. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459-63 (2010).
22. Sul, J.H. & Eskin, E. Mixed models can correct for population structure for genomic regions under selection. *Nat Rev Genet* **14**, 300 (2013).
23. Fick, S.E. & Hijmans, R.J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol* **37**, 4302–15.
24. Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965-78 (2005).
25. Lobréaux, S. & Melodelima, C. Detection of genomic loci associated with environmental variables using generalized linear mixed models. *Genomics* **105**, 69-75 (2015).
26. Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F. & Bergelson, J. Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* **334**, 83-6 (2011).
27. Seren, Ü., Vilhjalmsson, B.J., Horton, M.W., Meng, D., Forai, P., Huang, Y.S., Long, Q., Segura, V. & Nordborg, M. GWAPP: A web application for genome-wide association mapping in Arabidopsis. *Plant Cell* **24**, 4793-805 (2012).
28. Horton, M.W., Hancock, A.M., Huang, Y.S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N.W., Platt, A., Sperone, F.G. & Vilhjalmsson, B.J. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat Genet* **44**, 212-6 (2012).
29. Sharbel, T.F., Haubold, B. & Mitchell-Olds, T. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol Ecol* **9**, 2109-18 (2000).
30. Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N.A., Shah, C., Wall, J.D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M. & Bergelson, J. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* **3**, e196 (2005).
31. Johnson, R.C., Nelson, G.W., Troyer, J.L., Lautenberger, J.A., Kessing, B.D., Winkler, C.A. & O'Brien, S.J. Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* **11**, 724 (2010).
32. Storey, J.D. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol* **64**, 479-98 (2002).
33. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).
34. Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* **31**, 2013-35 (2003).

35. Storey, J.D., Taylor, J.E. & Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc Series B Stat Methodol* **66**, 187-205 (2004).
36. Mandt, S., Wenzel, F., Nakajima, S., Cunningham, J., Lippert, C. & Kloft, M. Sparse probit linear mixed model. *Mach. Learn* **106**, 1621-42 (2017).
37. Togninalli, M., Seren, Ü., Meng, D., Fitz, J., Nordborg, M., Weigel, D., Borgwardt, K., Korte, A. & Grimm, D.G. The AraGWAS Catalog: a curated and standardized *Arabidopsis thaliana* GWAS catalog. *Nucleic Acids Research* **46**, D1150-D1156 (2018).
38. Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annu Rev Genet* **44**, 293-308 (2010).
39. Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-45 (2011).
40. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol* **8**, e1000294 (2010).
41. Fox, E.J., Reid-Bayliss, K.S., Emond, M.J. & Loeb, L.A. Accuracy of next generation sequencing platforms. *Next Gener Seq Appl* **1**, e1000106 (2014).
42. Buell, C.R. Trust but verify: a lesson in technology limitations and error propagation. *Plant Cell* **30**, 515-6 (2018).
43. Sloan, D.B., Wu, Z. & Sharbrough, J. Correction of persistent errors in Arabidopsis reference mitochondrial genomes. *Plant Cell* **30**, 525-7 (2018).

Figures

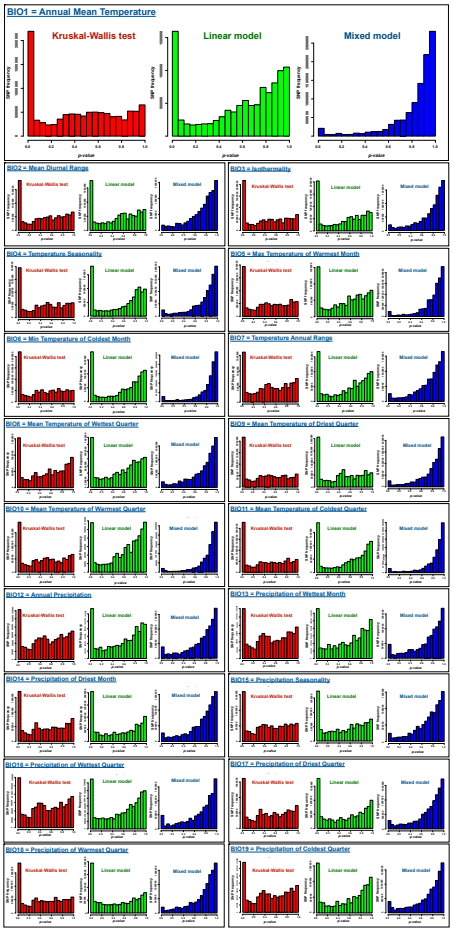

Fig. 1. The use of a mixed model reduces the inflation of p-values after correction for population structure. The excess of significant associations for the 19 WorldClim BIO variables, using two different association methods (Kruskal-Wallis and linear model) that do not correct for population structure, with a distribution of p-values skewed strongly towards 0, is expected due to the confounding effects of population structure which results in an inflation of significant yet spurious associations. As illustrated, the implementation of the mixed model successfully reduces this inflation and yields a distribution of p-values which show that, as expected, most of the variants are not significantly associated with the respective environmental variable, and a few significant variants remain.

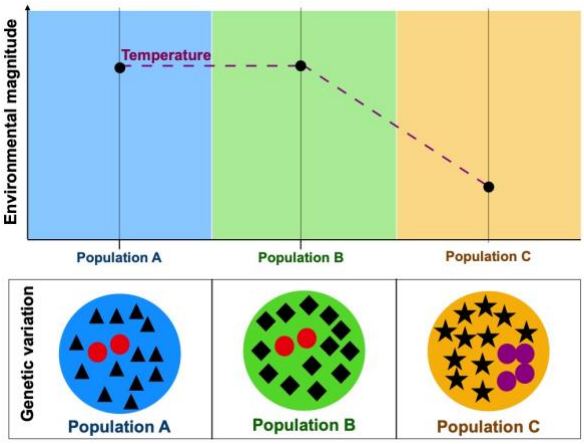

Fig. 2. As we correct for population structure using a mixed model, we may expect an increased number of false negatives resulting from over-correction. This is exemplified by a toy example in which we explore the genetic variants (SNPs) in a simplified scenario depicting genetic association with a single environmental variable (temperature) that occurs at different magnitude (y-axis) in three distinct and structured populations, with a "population C" adapted to lower temperatures relative to the other two ("population A", and "population B"). Triangles in "population A", rhombuses in "population B", and stars in "population C" represent the population structure distinctive of each population, in other words, the variants resulting from differences in genetic ancestry among the populations. Red circles denote variants adaptive to high temperature, and purple circles denote variants of the same gene adaptive to lower temperatures. When we study the genetic variation (SNPs) associated with temperature, we identify a number of these variants. In some cases, the associated variants will be unequivocally associated with high temperature (red circles). In other cases, we identify other variants that are only present in a single population (stars and purple circles in "population C"). The problem is that both of these associated variants exhibit an abrupt change in allele frequency in population C relative to Population A and Population B. It is therefore difficult to tell whether these changes in allele frequency are the result of genuine adaptation to low temperature (purple circles), or whether they just reflect a common genetic background of a structured population (stars). If we were not to correct for population structure, both the purple circles and the stars would be associated with low temperature; erroneously in the case of the stars. Thus, we would increase type I error in our results. Conversely, when we correct for population structure using a mixed model, we reduce this Type I error, as we reduce the inflation of significant associations and the number of false positives (stars). However, when we correct for population structure to reduce Type I error, we cannot avoid removing causative variants (purple circles) that result from genuine selection by the environmental variable which we are testing (temperature). We thus end up with increased type II (false negative) error. Because we were more concerned with reducing Type I error than reducing Type II error, in our analysis we correct for population structure and accept the increased possibility of type II error at the expense of reduced type I error. In reality, this illustration is an oversimplification given, among other reasons, that the structures of populations are not discrete and that the great majority of the environmental variables we studied are continuous. Accordingly, realistically, despite our best efforts, as in any GWA study we can expect both type I and II errors to be present to some extent in our datasets.

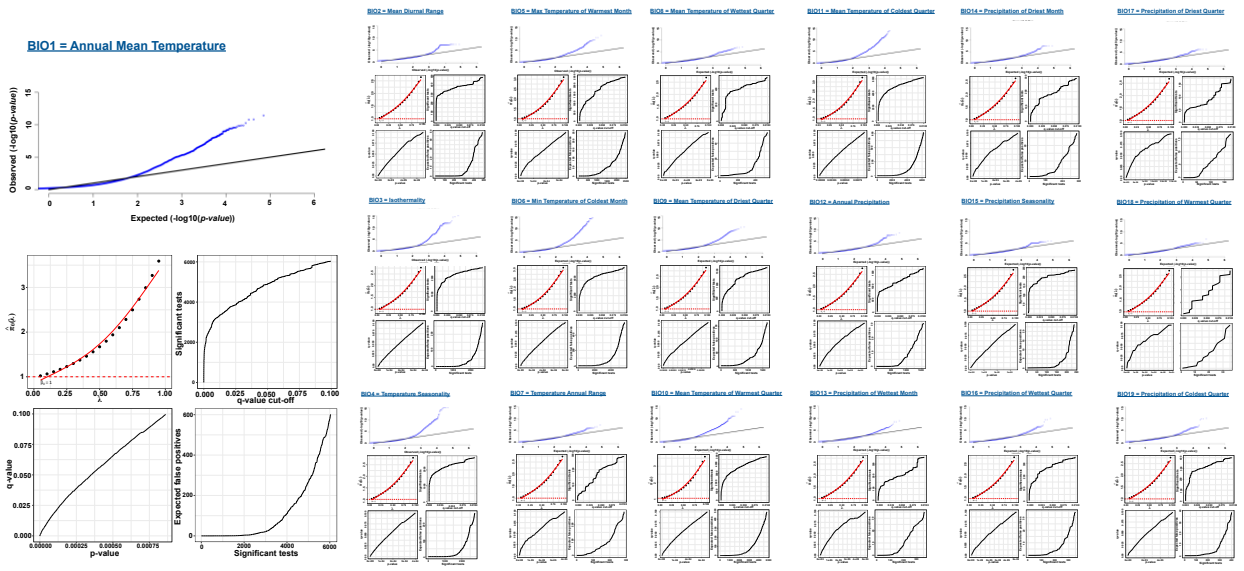

Fig. 3. We used the 19 BIO variables obtained from WorldClim 2[23] to exemplify the possible statistical inflation of *p-values* in our results and type I error. For each variable, Quantile-Quantile plots depict a distribution of *p-values* (expressed as scores, or -log(*p-value*)) following the expected distribution with a deviation from it as the observed scores increased in significance. A deviation from the expected scores over the whole range of theoretical quantiles would have suggested significant statistical inflation and a very significant presence of false positives. Also, for these variables, we describe the number of significant tests observed for each q-value cut-off, the relationship between p and q-values, as well as the relationship between expected false positives and number of significant tests. These results highlight the presence of true signals in our results, as well as the potential for type I error that will increase depending on the severity of the FDR threshold imposed.