Author's reply to Revision round #1

> I sincerely thank the reviewers for their thoughtful and constructive comments that have certainly improved the manuscript. As you will see, I addressed all of them carefully. In particular, I tested the influence of equilibrium vs. differentiation on the inference, implemented gene-flow among subpopulations, tested the influence of a hierarchical population structure and patched in an individual GWAS for comparison. This required rewriting of substantial parts of the code and re-running of the analyses, which took several months. As a result, the manuscript was substantially reshaped, in particular Results and Discussion, which made it not always straightforward to link the response to a particular comment with a particular change.

Decision for round #1 : *Revision needed*
Thank you for submitting your work to PCI Evolutionary Biology. The manuscript presents a novel mean-based GWAS approach whereby mean allele frequencies and mean trait values of many populations are utilized to understand the genetic basis of complex traits in natural populations. Whereas the two reviewers and myself find merits in the approach proposed, several points need to be addressed in a revised version of this manuscript. Both reviewers provide detailed and constructive criticisms on how to improve the manuscript. The most salient issues are a lack of realism in the simulations, failure to account for population structure in the analysis, and lack of clear, simulation-based comparisons with classic GWAA. Simulations were run for only a few generations, did not reach mutation-drift-selection balance and had no migration among sub-populations. Since population structure is a major confounder in GWAS, efforts to account for it in the analysis seems waranted. Likewise, claims of superiority of mean-GWAA over individual-based GWAA should be backed by clear comparisons and the advantages of mean-GWAA be clearly discussed. You will find more points that need to be addressed in the reviewers' comments.

I hope you will consider addressing the reviewers comments in a revised version of your manuscript. Please provide a point-by-point answer to the reviewers' comments.

*by **Frédéric Guillaume**, 20 Aug 2024 13:14*
Manuscript: **https://doi.org/10.1101/2024.06.12.598621**
version: 1

Review by Petri Kemppainen, 26 Jul 2024 15:23

The manuscript "On the potential for GWAS with phenotypic population means and allele-frequency data (popGWAS)" describes a simple method correlating population mean phenotypic values against population mean allele frequencies to identify putative causal nucleotides underlying traits of interest. The principle is the same as in GWAS but using populations, as opposed to individuals, as sampling units. The benefit of this (relative to regular GWAS), is the fact that sequencing (namely PoolSeq) and phenotyping (bulk phenotyping) is cheaper and easier for populations compared to individuals. In addition, by focusing on population, the variance around the phenotypic means is decreased potentially increasing the power of this method when heritability is low (high environmental variance). This approach is somewhat analogous to outlier analyses/genome scan approaches where associations between genotypes and phenotypes are tested indirectly by assuming different habitats have predictably different phenotypic means.

Here the focus is also on populations/habitats rather than the individuals, albeit typically these approaches are not readily applicable for PoolSeq data. It is for instance well known that studies of

parallel evolution (where similar phenotypic differences are predictably found across different habitats or environmental gradients across a species distribution range) are particularly powerful of disentangling the effects of natural selection from neutral processes such as genetic drift as a source of allele frequency differences between populations (e.g. Johannesson K. 2001. Trends Ecol Evol 16:148–153).

This the premise of this approach is tested in simple forward in time Wright-Fisher simulations, with initial allele frequencies for the ancestral population drawn from uniform distribution with range [0.1,0.9]. From this, sub-populations (500 individuals each) were colonized, each with phenotypic optima drawn from a normal distribution. The allelic effect sizes (for a varying number of loci) were drawn from different distributions. The subpopulations were then allowed to evolve for 2-50 generations, which, in the absence of migration between the subpopulations, resulted in different degree of population structuring between them (due to genetic drift).

Overall, the author reported high statistical power to detect phenotype x genotype associations in their simulations particularly when the number of causal loci was low and many populations were sampled, even when heritability was as low as ~30%. Promisingly, reasonable statistical power was also detected for moderately polygenic traits (up to~100).

Given the simulations, these results are not surprising for several reasons. We know that rapid adaptation in nature is highly dependent on standing genetic variation. However, in structured populations (the norm in natural populations) this variation is likely to differ between different geographic regions.

Thus, what set of alleles are available for adaption (e.g. when colonising new habitats or when the environment changes) can greatly differ from population to population. In the simulation presented in this manuscript, however, the initial allele frequency was the same for all populations (giving each the same probability of having the same set alleles available for adaptation).

> In principle, I agree with assessment of the reviewer here that it is difficult to identify causal loci, if the population differentiation is so large that the basis for adaptation to the same pressure is different. However, for modelling purposes, this simplifying assumption appeared justified because,
>
> i) Adaptation to a local optimum was used only as a vehicle to obtain populations with different population means; as stressed in the manuscript, any other mechanism creating phenotypic mean differences among populations should work as well.
> ii) At some point in the past, extant populations likely share a common ancestor population and if these ancestor population was already subjected to the same selection pressures, they should still share some of the underlying genetic variation, which can be detected. This might be different, if the same selection pressure arose only after population divergence (see below).
> iii) The limits of a reasonable population differentiation where the method still works were explicitly explored in the manuscript.
>
> Moreover, lack of common genetic variation questions also the value of using parallel selection (requiring a certain amount of divergence) as advocated by the reviewer above. As recently shown, phenotypic parallelism, even if derived from the same ancestral population, does not need to rely on the same genetic variants.
>
> Lai, W. Y., Nolte, V., Jakšić, A. M., & Schlötterer, C. (2024). Evolution of phenotypic variance provides insights into the genetic basis of adaptation. Genome Biology and Evolution, 16(4), evae077.

Pfenninger, M., Patel, S., Arias-Rodriguez, L., Feldmeyer, B., Riesch, R., & Plath, M. (2015). Unique evolutionary trajectories in repeated adaptation to hydrogen sulphide-toxic habitats of a neotropical fish (Poecilia mexicana). Molecular ecology, 24(21), 5446-5459.

Given the redundancy of the genomic basis of (most?) traits, I doubt whether *any* GWAS method is suitable for a scenario where different populations have only access to a different set of adaptive mutations, simply because there is likely no or only a very small common basis that could be detected. In other words, only the common genetic basis of a trait across the population tested can be identified by GWAS in the first place. This is therefore not a problem of the proposed method but a general feature of cross population GWAS analyses.

Furthermore, the initial allele frequencies were drawn from a uniform distribution, while allele frequencies in natural populations (i.e. populations reasonably close to mutation-drift equilibrium) are typically highly skewed towards low frequency variants (further reducing the chance that all populations have access to the same ancestral pool of adaptive alleles in the wild).

Thanks for pointing this out, this is a valid point that completely slipped my attention. Now I use a beta function with parameters alpha = beta = 0.5 to draw the initial allele frequencies. However, there are two points I'd like to raise on this issue:

i) a beta distribution of allele frequencies arises naturally for *neutral* alleles in a Wright-Fisher model. I'm not sure whether this applies also to functional sites under selection. Recent research indicates that functional standing variation is often maintained by balancing selection, which tends to lead to intermediate allele frequencies (which was also observed here).

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles in a finite population. Theoretical Population Biology, 3(1), 87-112.

Abdul-Rahman, F., Tranchina, D., & Gresham, D. (2021). Fluctuating environments maintain genetic diversity through neutral fitness effects and balancing selection. Molecular Biology and Evolution, 38(10), 4362-4375.

Rudman, S. M., Greenblum, S. I., Rajpurohit, S., Betancourt, N. J., Hanna, J., Tilk, S., ... & Schmidt, P. (2022). Direct observation of adaptive tracking on ecological time scales in Drosophila. Science, 375(6586), eabj7484.

Lynch, M., Wei, W., Ye, Z., & Pfrender, M. (2024). The genome-wide signature of short-term temporal selection. Proceedings of the National Academy of Sciences, 121(28), e2307107121.

Pfenninger, M., & Foucault, Q. (2022). Population genomic time series data of a natural population suggests adaptive tracking of fluctuating environmental changes. Integrative and Comparative Biology, 62(6), 1812-1826.

ii) and, more importantly, rare alleles, by definition, do not contribute much to the variance in phenotypic population mean, even when of large effect. This is of course different for individual GWAS (see a more detailed explanation below).

There was also a clear trend towards lower statistical power with increasing number of generations of adaptation (Fig. 4D), but the populations were only allowed to adapt for up to 50 generations.

The reviewer is right that the statistical power decreased with increasing number of generations. This is due to increasing independent evolution between populations as mirrored in increasing population structure (F_{ST}), not due to increasing adaptation, as the new simulation with gene-flow show (see below). Best possible adaptation (i.e. least achievable

mean deviation from the phenotypic optimum in drift-selection equilibrium) was already and regularly achieved with < 10 generations.

These circumstances, i.e. immediate colonization of populations form the same pool of ancestral genetic variation (with unnaturally high numbers of medium frequency alleles) that are allowed to adapt for only 50 generations is the type of scenario where the proposed approach is likely to perform well. Unfortunately, this is also a very unrealistic scenario in natural populations.

In the least, the simulations should include some form of a burn-in to allow allele frequencies to reach mutation-drift equilibrium, ideally with different levels of population structuring in the ancestral population and the populations should be allowed to evolve for much longer than 50 generations (and instead allowing different levels of gene flow between them), before any conclusions can be drawn from this study.

Inspired by this comment, I implemented several changes and performed additional simulations:

i) A series of initial simulations to evaluate the effect of the suggested "burn-in" generations on the drift/selection equilibrium. This revealed that best possible adaptation to the new optimum was in most cases reached after less than 10 generations and remained so. True to theoretical expectations, higher gene-flow among populations led to less perfect local adaptation for a given selective strength.

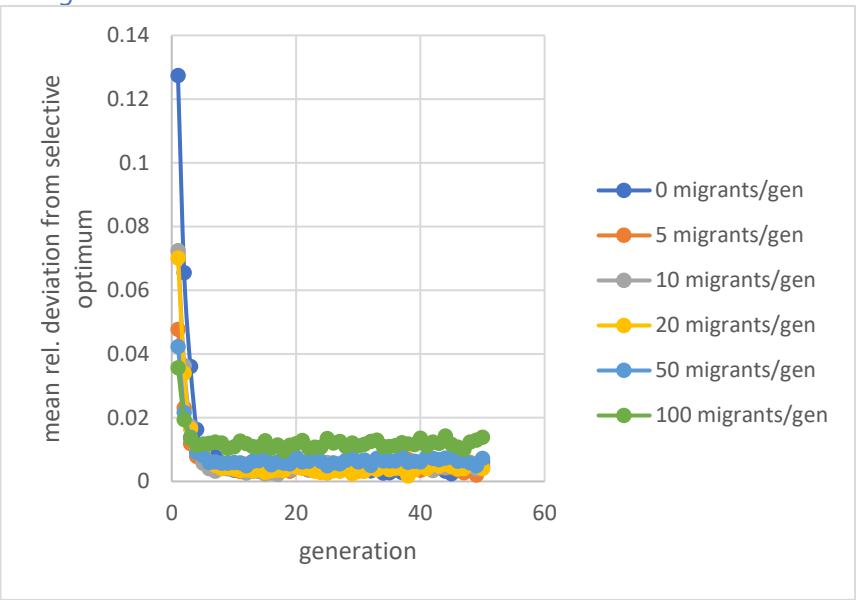
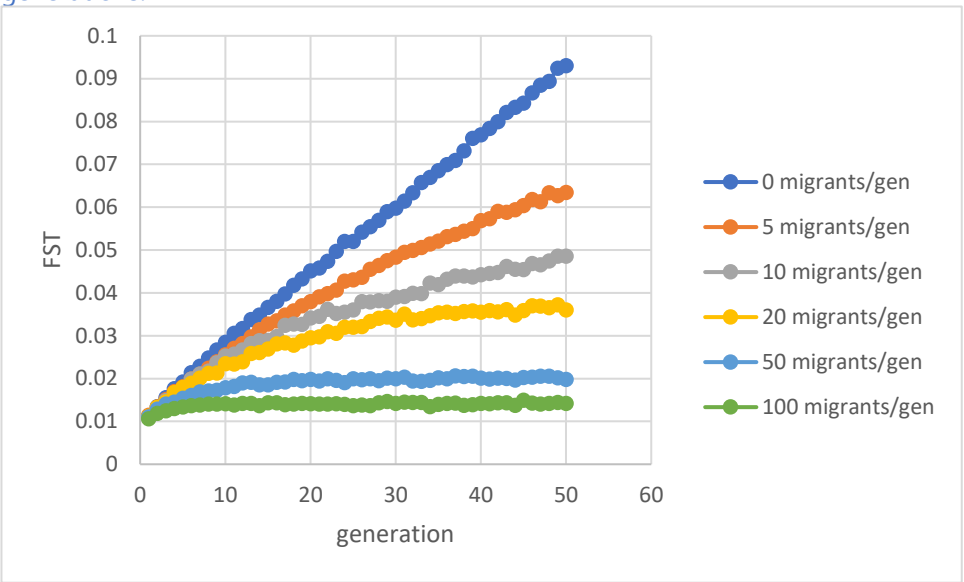


Figure: mean relative deviation of the populations from the selective optimum during the first 50 generations. For the given simulations, the equilibrium level is reached after less than 10 generations. The generally low (< ~2% deviation from local optimum) equilibrium level depended on the level of gene-flow among the subpopulation

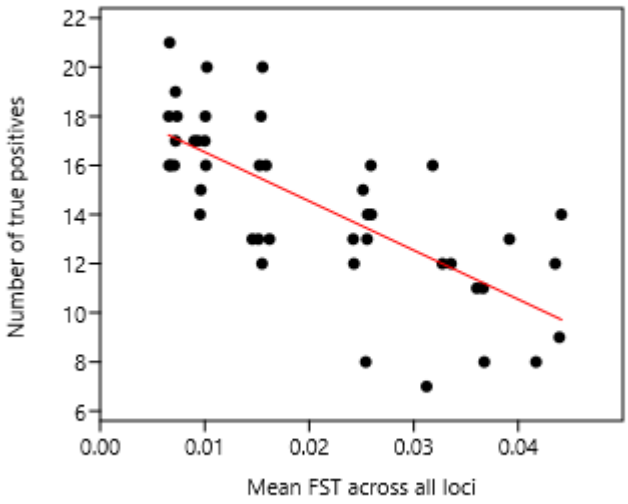
ii) For the main set of simulations, gene-flow among subpopulations, realised by recruiting a given number of individuals (0, 5, 10, 50, 100) randomly drawn from other subpopulations as parents per generation was implemented. This required substantial changes in the architecture of the script.

iii) Everything else being equal, the time to reach an equilibrium F_{ST} among subpopulations depended on the number of migrants per generation. It was quickly reached for high gene-flow and much slower for lower gene-flow. For zero migrants, the equilibrium F_{ST} would be 1, reached probably only after hundreds of

generations.



iv) However, predictive for GWAS performance was the F_{ST} among populations *at the time of analysis*, not its equilibrium value. This indicated that the method works also well for populations that are not in selection-migration-drift equilibrium. The use of a model with no gene-flow among subpopulations in the initial manuscript was therefore rather conservative.



v) The degree of adaptation was irrelevant for poolGWAS performance. More important was rather the extent of the phenotypic gradient – due to selection or not.

To keep a balance between attaining some equilibrium (at least for selection/drift equilibrium) and computation time, all new simulations were run for 30 generations.

This remark suggests that the reviewer assumes that the performance of the method depends on some equilibrium state of the respective populations. However, it is assumed that the individual phenotype depends on the (multilocus)genotype at the contributing loci. This relation is independent from the population the individual comes from. Likewise, both the

phenotypic population mean as well as the underlying allele frequencies at contributing loci are simple summary statistics over the individuals that make up the pool-sample – completely independent from the population genetic state of the populations of origin. In other words, the method should work as well for individuals pooled according to their phenotype rather than their population of origin. Initial simulations show that this is indeed the case, opening up a road to a similar GWAS approach suitable for experimental set-ups.

As mentioned before, perfect adaptation (see above) was regularly achieved within less than 10 generations. Moreover, natural populations are usually not in equilibrium:

> Müller, R., Kaj, I., & Mugal, C. (2022). A Nearly Neutral Model of Molecular Signatures of Natural Selection after Change in Population Size. Genome Biology and Evolution, 14. https://doi.org/10.1093/gbe/evac058.

Migration can lead to reduced fitness and persistent natural selection within recipient populations, affecting the migration-selection balance.

> Bolnick, D., & Nosil, P. (2007). NATURAL SELECTION IN POPULATIONS SUBJECT TO A MIGRATION LOAD. , 61, 2229 - 2243. https://doi.org/10.1111/j.1558-5646.2007.00179.x.

Rather, natural populations are constantly tracking variable selective optima. This is underlined by recent empirical investigations with population genomic time series.

> Pfenninger, Markus, and Quentin Foucault. "Population genomic time series data of a natural population suggests adaptive tracking of fluctuating environmental changes." Integrative and Comparative Biology 62.6 (2022): 1812-1826.
>
> Rudman, S. M., Greenblum, S. I., Rajpurohit, S., Betancourt, N. J., Hanna, J., Tilk, S., ... & Schmidt, P. (2022). Direct observation of adaptive tracking on ecological time scales in Drosophila. Science, 375(6586), eabj7484.

According to recent theoretical developments, such rapid adaptation leads to intermediate allele frequencies.

> Höllinger, I., Wölfl, B., & Hermisson, J. (2023). A theory of oligogenic adaptation of a quantitative trait. Genetics, 225(2), iyad139.

Mutation, however, was not considered because:

i) new or low frequency mutations will not substantially influence phenotypic population means and their AFs.
ii) mutations are rare (in our model species in the order of 10^{-9} per site and generation) and mutations affecting a particular functional trait even rarer.
iii) As mentioned before, adaptation of the subpopulations to some local optimum was introduced only to avoid setting arbitrary phenotypic differences among populations, not to study the adaptation process. The actual estimation takes place in a single generation on the genetic variation present in this generation. Mutation is therefore irrelevant, unless the time since divergence among subpopulation is so long, gene-flow so low and the differential mutation accumulation so strong that the genetic basis of the trait differs so much (i.e. F_{ST} is so high) that the application of the method (or any other GWAS) is not warranted anyways. But this limit was explicitly explored.
iv) This is common practice in quite some other studies on quantitative traits:
Matuszewski, S., Hermisson, J., & Kopp, M. (2015). Catch me if you can: adaptation from standing genetic variation to a moving phenotypic optimum. Genetics, 200(4), 1255-1274.
Hermisson, J., & Pennings, P. S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics, 169(4), 2335-2352.

v) From a more practical side: many (very) low frequency variants will probably not be called in a PoolSeq approach

Given these arguments, restricting the simulations to standing variation seemed reasonable.

A respective paragraph was inserted into M&M

A word on the realism of the simulations. The method was actually developed for application in forest trees with little to no population structure, due to high Ne (even larger than the demographic N), high gene-flow and long generation times. In addition, these species are managed, which means that the current canopy trees are derived from either natural regrowth after post-glacial recolonization, planting of seedlings from hatcheries or sowing of seeds from elsewhere. Accordingly, while they may or may not be locally adapted (most actively planted stands are likely not), they show high phenotypic variation both among local individuals as well as among stands and are certainly not in any sort of equilibrium. Yet, as the attached worked example shows, the method worked apparently very well.

The test is based on a simple linear regression performed independently for each SNP. From the GWAS literature, we know that population structuring lead to false associations (two populations that differ in phenotype may also differ at neutral loci due to genetic drift), such that great care is always taken to control for this.

The reviewer is perfectly right that confounding associations between two-allelic SNPs and a phenotype may easily arise in individual based GWAS (iGWAS hereafter), because phenotypically more similar individuals may be more similar because they share common ancestry as members of the same family/population/clade. Thus, the probability of wrongly associating a SNP that is shared among relatives with a phenotype is high and correction for it is warranted.

This is, because in iGWAS, a quantitative or qualitative phenotype (y) is related to a genotype (x) either by a linear or logistic model – with exactly three possible values of x (AA, Aa, aa). This leads inherently to low statistical power, requiring a very high number of individuals for a reliable statistical association. In popGWAS, the possible values of x are allele frequencies in different populations (0-1) and thus limited only by sampling effort, which gives the approach a strong inherent statistical advantage – in particular by identifying neutrally drifting loci that are not expected to be associated with the phenotype.

"Confounding" in the case of linear associations between mean population phenotypes and population AFs as in popGWAS means that actually neutral loci would show a range of AFs that covary with the differences in phenotype thus leading to false positives. As the simulations to statistical power show, this is by pure chance certainly the case to some extent, provided enough independent loci are tested.

How can population structure nevertheless interfere with popGWAS? This could be the case, if the focal trait is effectively neutral and thus the constituting loci follow the overall drift pattern. In this case, the matrix of mean phenotypic population differences (or Q_{ST}) should show a positive correlation with the overall F_{ST} matrix among populations. Alternatively, one could look for correlations between significant principal components on the AF matrix and the trait in question. A trait showing this easily testable pattern may therefore *a priori* not be suitable for analysis with the method. However, in any case, where the Q_{ST} is larger than the F_{ST} and uncorrelated to it, this should not be a problem.

Another way to break potential chains of population structure would be (in cases where there was no bulk phenotyping) to arrange the individuals in phenotypic classes and prepare the pools accordingly for these classes (see above).

The above aspects are now discussed.

There are two major ways to achieve this. One is to either include relatedness as a random effect or add population (or PC coordinates) as a co-variate. The other is to perform genomic control to account any residual p-value inflation. Notably, residual p-value inflation may exist even if relatedness is otherwise accounted for (it may not have succeeded to account for everything), and should always be performed. Thus, in all association studies/outlier analyses/genome scans I expect to see some quantile-quantile plots of expected (uniform distribution with range [0,1]) vs. observed -log10 p-values. P-value inflation exist when the slope of a linear regression of these data points is >>1 and genomic control is simply dividing the observed -log10 values by this slope. Without seeing these plot, it is not clear whether p-value inflation exists in the data from the simulations. In more realistic simulations (as suggested above) certainly some level of p-value inflation is expected and would need to be accounted for.

> I thank the reviewer for raising the issue, because thinking about the response led to a deeper understanding of the statistical properties of the proposed and other methods. According to the suggestion of the reviewer, qq-plots were integrated into the analysis. I report some summary statistics, because showing the plots for all x-thousand simulation runs is neither possible nor informative. Generally, the mean slope of the qq-plots was 1.28 and did not exceed 2, which I deem not much larger than 1. Please see below some qq-plot examples.
>
> Generally, a qq plot compares distribution of samples: if the slope is ~1, the two samples probably come from the same distribution. Using a random uniform distribution as reference for a qq-plot in GWAS implicitly assumes that any deviation of an individual data point from the null-hypothesis is random and therefore meaningless.
>
> In the present case, neutral (or better: non-contributing) loci are not expected to show a linear relation with the phenotype; the neutral expectation on these loci is therefore adequately described by a random uniform distribution. Any observed respective association is therefore a coincidence; the slope for these loci compared to a uniform random distribution should be ~1. For the contributing loci, however, the expectation is that their AFs are linearly associated to the phenotypic means, i.e. they are expected to deviated from the tested null-hypothesis. The tested sample is therefore a composite from two different processes and thus different distributions: the non-contributing loci that should conform to the random uniform expectation and the contributing loci for which the strength of association in each case depends on many factors, but whose p-values are expected to be on average much smaller than those of the non-contributing loci. It is therefore inevitable that the p-value distribution of such a mixed sample in *any* association study must deviate to some degree upwards from a slope of one, in particular in the upper part – if the method has any power to distinguish the two classes of loci (see Figure below). The degree of deviation should therefore depend on i) the power of the method to detect deviations from null-hypothesis and ii) the ratio between contributing and non-contributing loci.
>
> In the below example, the proportion of contributing loci is relatively large (50:1000), thus raising the slope probably stronger than in real life data sets with relatively much more neutral loci. Still, the slope is not much larger than 1.
>
> Please note that the underlying data for the three following examples of analysis methods is identical, coming from the very same simulation run. The results are therefore absolutely comparable.

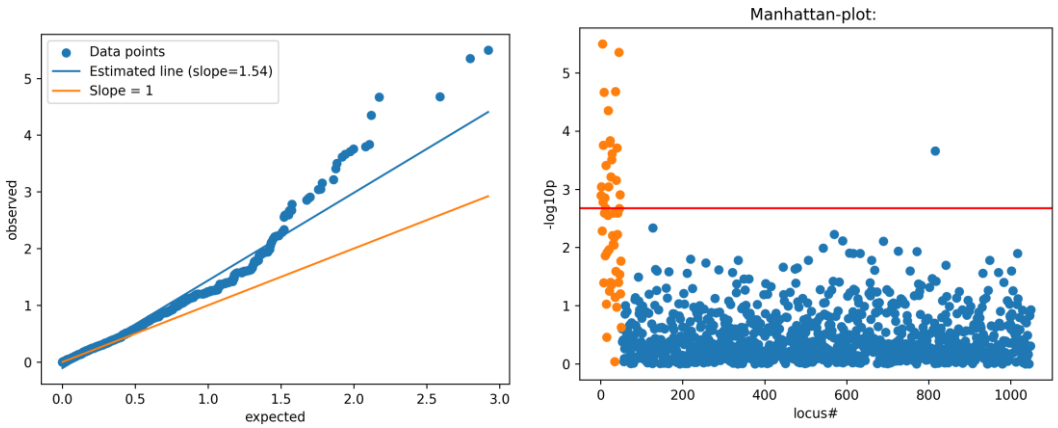


Figure left: popGWAS qq-plot for a simulation run with 24 populations, 50 contributing and 1000 neutral loci, 20 generations, flat allelic contribution and a per generation gene-flow of 20 individuals. The orange line indicates a slope of 1, the blue line is the estimated slope for the observed data.

Figure right: The respective Manhattan-plot. The orange dots represent the a priori known contributing loci, the blue ones known non-contributing loci. The red horizontal line indicates the upper 2% quantile used for further analysis. It is clearly visible that the two classes of loci belong to different p-value distributions, which made the identification of many true positive loci possible.

I implemented a simple iGWAS as well; see reviewer 2. This was done by testing for a linear relation between the genotypes and the phenotypes of a sample of individuals. The number of individuals was chosen such that sequencing them with a coverage of 15X for proper genotyping would match the sequencing effort necessary for the respective popGWAS approach. Individuals were randomly chosen from all subpopulations. It was possible to study the statistical behaviour of this approach as well. With iGWAS, the slope of the qq-plot did not deviate much from 1 for again the exactly same data set. Accordingly, the power to detect true outlier loci was rather limited such that only four true positives and multiple false negatives were detected with the given thresholds.

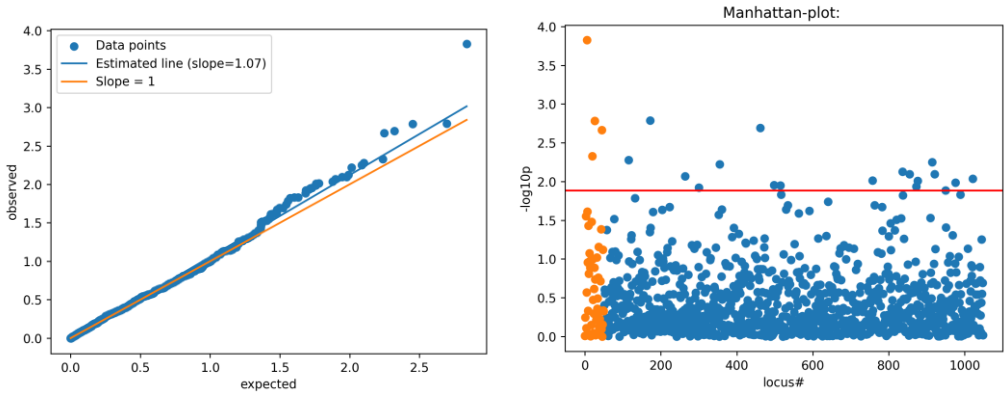


Figure above: iGWAS qq-plot for the same simulation run as above. The orange line indicates a slope of 1, the blue line is the estimated slope for the observed data.

Figure below: The respective Manhattan-plot. The orange dots represent contributing loci, the blue ones non-contributing loci. The red horizontal line indicates the upper 2% quantile used for further analysis.

In a 2017 paper of the reviewer, differences in allele-frequencies before and after selection were used to identify functional candidate loci and there, substantial p-value inflation was observed. To see whether the current simulation used can reproduce this pattern, I implemented a test on AF differences between the phenotypically most extreme simulated populations (upper and lower 10%), which approximates the approach in the reviewer's paper. I used a two-sided Fisher's Exact test (FET) to test for deviation from a uniform expectation. Here, strong p-value inflation (slope = 130) over all loci was observed in all simulations. In addition, the distributions of the two classes of loci were not obviously different and therefore the statistical power to identify contributing loci negligible; except for one, all loci above the 2% outlier threshold were false positives.

> Kemppainen, P., Rønning, B., Kvalnes, T., Hagen, I. J., Ringsby, T. H., Billing, A. M., ... & Jensen, H. (2017). Controlling for P-value inflation in allele frequency change in experimental evolution and artificial selection experiments. Molecular ecology resources, 17(4), 770-782.

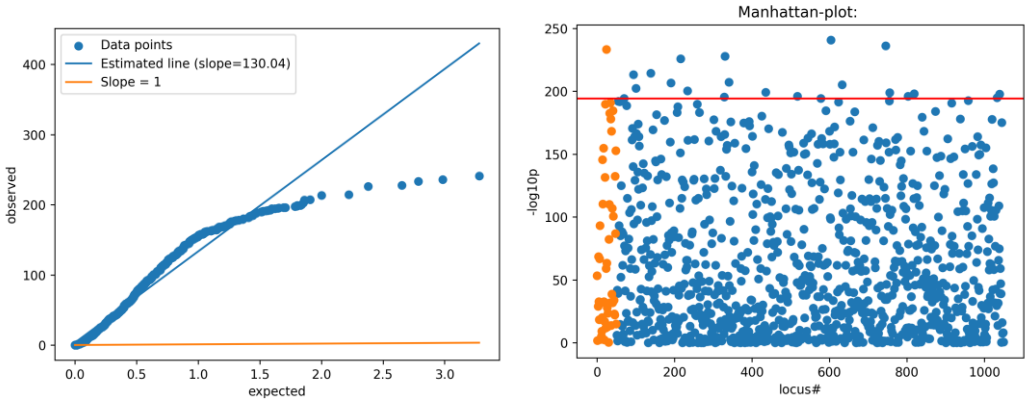


Figure left: extreme value GWAS qq-plot for the same simulation run as above. The orange line indicates a slope of 1, the blue line is the estimated slope for the observed data.

Figure right: The respective Manhattan-plot. The orange dots represent contributing loci, the blue ones non-contributing loci. The red horizontal line indicates the upper 2% quantile used for further analysis.

On the need to account for p-value inflation (or other measures to account for multiple testing): Given the results above, I doubt whether this is generally necessary, if one does not use all significant loci above a given, predefined threshold significance threshold, but a certain quantile of all loci regardless of their p-value, i.e. an *a priori* fixed, rather conservatively small number of outliers. The -log10p value is such a case used rather as a goodness-of-fit measure to a linear model to rank the loci (taking into account data quality e.g. missing data) than for the statistical rejection of a certain null-hypothesis. Correcting for GC by dividing by the slope is a linear procedure that would change nothing on the ranking (similar to corrections for multiple testing).

The novelty of this approach is that there currently does not exist any method that can utilise PoolSeq data to test for genotype x phenotype associations and as such I would certainly want to see this method tested with more realistic simulations (and ideally also complemented with empirical data as hinted by the author).

> Thank you for these positive words. I now extended the simulations according to the suggestions. Moreover, I now additionally compare the popGWAS approach to individual

GWAS and added additional summary stats. Please find included a very early, confidential draft manuscript implementing the new method to identify the genomic basis of leaf-out and leaf-shedding day differences among nation-wide stands of European beech.

I expect it to perform similar to other outlier methods out there with the benefit that this is specifically tailored for PoolSeq data (I do not expect it to do well on polygenic traits, but this is perfectly fine).

See remarks about the definition of polygenicity to reviewer 2

However, seeing as this method is so related to standard GWAS, the knowledge from this field should be utilised better, notably by showing to what extent the approach is susceptible for p-value inflation and then, in the least perform GC (which will reduce false positive rates but also power), but ideally directly control for relatedness e.g. by including relatedness as a random effect, which is expected to reduce false positive rates but not statistical power (at least not to the same extent as GC). As an example, in Fang et al (2021,Mol Biol Evol 38:msab144) such an approach (EMMAX) was successfully used to teste for association between genotype and habitat treated as a binary trait.

As shown above and in the results, p-value inflation is not an issue with the proposed method and correction for relatedness among individuals is likewise not applicable for summary stats among populations.

Some minor comments (in no particular order):
In the abstract there is no hint on what the test is based on, only how it performs. The title gives some idea, but not enough to be of any help when reading the abstract.

Changed according to the reviewer's suggestion.

I believe that the "phenotypic plasticity parameter" used in this study is simply environmental variance. Phenotypic plasticity implies a genotype x environment interaction which was, to my understanding, not tested here.

Changed according to the reviewer's suggestion throughout the manuscript.

L318 "The number of effectively independently evolving loci in a population depends on genome size, effective population size (including all factors that affect it locally and globally) and LD structure". It is not clear what LD structure means. I my view, all the cited processes effect LD structure so it should not be mentioned as a separate process here.

Changed to recombination rate

The proposed method only performed well when genome size was small (the number of independently segregating loci in the simulations) and/or the number of sampled populations was large. In large(ish) outbreeding natural populations LD typically declines relatively quickly such that likely >>100.000 (the largest genome size simulated here) SNPs are typically required for the sampled loci to be in sufficiently high LD with the causal loci of interest for them to be useful, especially for polygenic traits.

Please see Table in the Supplementary info: 100.000 independently evolving loci cover large parts of biodiversity, depending on LD (genome sizes up to 3 Gb, if mean LD extends over 50 kb as e.g. in humans). Typical PoolSeq datasets yield > 1 Million scorable loci, many of which are of course in LD. The mean genome size in the Animal Genome Size Database is ~500 Mb; the Kew Royal Botanical Garden genome size database lists ~4000 species with genome sizes below 1.5 Gb.

Thus, even if similar performance as presented here can be replicated in more realistic simulations (see above), it seems the proposed method still relies on a large sample size to detect anything but SNPs with large or medium effect sizes (on par with most outlier methods/genome scans approaches that I'm aware of which is not a problem, see above).

Given the parallels with methods used to detect parallel/convergent evolution, I'm surprised that this is not discussed more. In particular, the simulations used in this literature are typically much more realistic (e.g. Fang et al. 2020. Nat Ecol Evol 4:1105–1115, Kemppainen et al. 2021. Mol Ecol).

> See remarks above.

L233 "For the assessment of the effect of population structure, the subpopulations could evolve in complete isolation from each other for predetermined number of generations…" I think it is a bit misleading to consider the number of generations of adaptation in isolation as a proxy for "population structure". Population structure implies some level of migration-mutation-drift balance in a meta-population setting and the level of population structure should then be modelled by different levels of migration/dispersal.

> As the reviewer rightly says, AFs (and this is the only thing that matters here) are influenced by migration and drift (as detailed above, mutation appears not to be relevant here). It is unclear, why one cannot be replaced by the other (moreover since e.g. in the 2016 paper, the reviewer and his/her co-authors manipulated Ne instead of m to simulate dispersal). In any case I now included migration explicitly into the simulations.

I am not aware of the F1-score, and I would be grateful for some more information about this in the main text (what does it tells us?).

> The F1 score is the harmonic mean of the precision and recall. It thus symmetrically represents both precision and recall in one metric. Information added.

Although I found the manuscript well and clearly written, it was not completely clear how the different sections in the materials and methods were connected. It seems there were several sections describing different parts of the same simulation. It would benefit with some text giving a better overview of the methods/simulations that were used. Also, whenever range of parameter values is mentioned, it would be great to have a reference to Table 1. Or make it clear early that all parameter values are described in detail in Table 1 so the reader is not left hanging.

> Reading the manuscript after a while again, I perfectly see the point of the reviewer. I tried to accommodate it by expanding the introductory paragraph and changed the hierarchy of the subheadings in the hope it becomes clearer now. The reference to the parameter tables was inserted throughout.

I did not see any information about how many sub-populations were simulated? I assume, since there was no gene flow between subpopulations, there was no need to simulate more than were used for the analyses. This should be clear from the text.

> I now added gene flow to the simulations and mention this in the methods.

Review by anonymous reviewer 1, 09 Aug 2024 13:43

The manuscript by Markus proposes a GWAS approach to identify the genetic basis of complex traits in natural populations. The author uses extensive simulations to show that a moderate number of true positive QTL can be identified using allele frequency data from PoolSeq and phenotypic means instead of individual-level genotypes and phenotypes. He also shows that given a large number of independent populations scored, a reasonably good prediction score can be achieved in new populations.

Although the simulation results are convincing regarding the effect of different parameters on the power of QTL detection and prediction, it is not clear whether this approach has any real and practical advantage compared to an individual-based GWAS, and/or whether it overcomes the well-known limitations of GWAS. A proper comparison is never made, and I think that besides a discussion

explicitly addressing this point, a proper simulation-based comparison is needed to support the claims made by the author regarding the advantages of this approach.

> Please note that no claim of superiority of the popGWAS method over iGWAS in performance was ever made in the first manuscript version; only the effectiveness in terms of necessary effort was pointed out (see below).
>
> However, according to the reviewer's demand, I've included a simple iGWAS that was performed for each simulation run on exactly the same simulated data and the same procedures for prediction to ensure comparability (see details below). In summary, popGWAS performed better in every single tested instance.

Main points that should be addressed:

1. What is the actual feasibility of using the proposed experimental design in wild populations?

Based on the results and discussion (line 530), to have a moderate chance to detect some QTLs >60 populations need to be sampled (min 50 individuals for pool DNAseq and phenotypes). How many species could actually be sampled in such way?

> We did this sort of sampling with a forest tree species in Central Germany with three persons within a week (please find an early draft for the respective worked example study in the material for review), it can be easily done with many other plant species. Colleagues of mine work with tiny soil species, that occur by the millions on every square meter, but can be individually genotyped only with extreme effort, also freshwater and marine plankton and -benthos that can be (automatically) sorted, many insect species (e.g. from malaise traps) would come immediately to mind, series of museum collections, hair trappings e.g. from wildcats, differentially evolved populations from experimental evolution and probably many more instances that I am currently not aware of. I concede that it might be difficult for many other species, but it is up to the according researcher to decide which method to use. My new simulation results give a good idea on the robustness of results depending on the number of populations sampled and using poolGWAS.

Is this actually an experimental design that can be implemented by researchers? Which are the phenotyping approaches that the author imagine will make mean phenotyping feasible? Etc.

> Satellite or any other remote sensing approach, flow cytometry, bulk measurements of metabolic rates, metabolomes, transcriptomes, HPLC or GC bulk measurements, automated video-based phenotyping, mass CT scanning of museum series… as cited, automated phenotyping is currently a very intense field of research that will likely open up many new possibilities for many traits.
>
> A respective paragraph was introduced.

2. What are the actual sequencing costs of this approach?

It is stated in the intro (line 80-84) and in the discussion (line 627) that this approach requires 'marginal sequencing efforts compared to individual based' approaches. However, the author doesn't show any calculations that actually support this point. The author should give precise estimates so the reader is convinced that there is actually a cost advantage. For example, if for a moderate-powered pool-based GWAS we need 60 populations DNA pools sequenced at 50x (minimum, to get 1 read per individual, based on the 50 individuals used in the simulations), we will need a total of 3000x coverage for all pools. Now, for an individual-based GWAS, if we sample 600 individuals and sequence at 5x (which will

give me enough confidence to call individual genotypes), that results in the same sequencing effort of 3000x coverage.

> I agree that normalising on the amount sequenced is a fair comparison and that's what I implemented in the simulations. However, I strongly disagree that 5x coverage is sufficient for reliable genotype calling. ~20% of heterozygous SNPs will be inevitably wrongly called, which would be detrimental in species with high Ne, high recombination rates and thus low LD (see comments of reviewer 1). We are using a minimum 15X mean coverage for reliable genotyping in individual resequencing and that is what I took to calculate the number of individuals in iGWAS.
>
> However, most importantly, the effective sequencing costs for many species depend not so much on the amount of actual sequencing, but rather on the number of sequencing libraries that need to be prepared (for the organisms that we are working with (beech ~500 Mb genome size, Chironomus ~200 Mb), the library preparation costs regularly exceed the costs for high (20X) coverage individual resequencing. In the numerical example of the referee above, the WGS approach would cost almost 5x times more for *Chironomus* given our current real prices (moderate 7,500 € for popGWAS vs. 37,200 € for iGWAS).
>
> However, I refrained from inserting an actual calculation of sequencing costs into the manuscript, because i) sequencing and library preparation costs vary considerably among countries, labs and commercial services, ii) they vary (fall) over time and would be thus soon outdated and iii) would give away the conditions we currently get, which would be against the contract with the service company (and therefore please keep the numbers given above private).
>
> I inserted a more generic sentence on the differential requirements in terms of necessary library preparations.

3. How does this approach overcomes GWAS limitations when addressing actual complex traits?

a. It is well known that rare alleles have larger effects when it comes to complex traits, and therefore large sample sizes are needed to identify such rare alleles. How is this method better at doing this than individual-based GWAS?

> Rare alleles seem to play an important role in human diseases; it is far from clear, whether this is true for traits in natural populations under strong purifying selection; theory suggests that this is rather not the case.
>
> > Sella, G., & Barton, N. H. (2019). Thinking about the evolution of complex traits in the era of genome-wide association studies. Annual Review of Genomics and Human Genetics, 20, 461–493.
>
> But more importantly, rare alleles play by definition no relevant role for the phenotypic mean of a population. This is of course different for individuals (see comments to reviewer 1).
>
> And finally, the simulations have shown that popGWAS performs in any single case better than the respective iGWAS. However, please note that I do not claim that popGWAS is the golden approach with 100% accuracy. This approach, as well as any other GWAS approach will have its shortcoming when it comes to systems with strong population structure and highly polygenic traits (in the latter case some loci will most likely be detected but by far not all). I show that this is a promising approach which leads to fairly robust results with manageable effort.

b. Population structure and relatedness between individuals is a huge issue in GWAS, and therefore GLM that include GRMs must be used to assure that the results are not dominated by false positives. How is the pool-based GWAS addressing this point? Specially given that it requires that tens of independent populations be genotypes and phenotyped. In the current proposed model there is nothing addressing structure. Is the pool based data able to control for structure across populations and within populations?

> The issue of population structure is now extensively addressed in the simulations (see comments above).

c. related to the above, estimates of false positives for the QTL mapping results should be presented and discussed. I didn't see them, maybe they are in the supplement?

> The false positive rate is explicitly or implicitly addressed throughout the manuscript (by showing PPV which is 1- false positive rate)

The author should provide an actual (simulation based) comparison between both approaches to prove that pool-based GWAS is better at the known GWAS limitations that he describes in the introduction.

> Again, it was never claimed in the initial paper that popGWAS performs better than iGWAS; I explicitly tried to avoid any comparison and referred to the lower effort necessary. Now, with the suggested simulations backing this up, such a claim can be rightfully made.

4. Why is the prediction power of this approach so much better than widely-used (but mostly underperforming) polygenic scores in humans?

The proposed approach sounds very similar to polygenic risk scores and we know that when used in populations different from the one where scores where estimated they terribly underperform. The author should discuss why his approach is so successful compared to PRS. If the traits the approach is targeting are complex traits, and are confounded by structure and environmental effects, shouldn't the performance be similar to what we know so far in real populations (a.k.a. poor).

This is indeed a very interesting issue that initially surprised me as well. I think that this is mainly because the target of prediction is different. Here, I predict phenotypic population trait means, while PRS usually tries to predict a quantitative trait of individuals.
For several reasons, it is much easier to predict the former:

i) The difference in means among subpopulations is governed by only a couple of loci, because the range of phenotypic population means extends over only a relatively small part of the theoretically possible range –both in the simulations but most likely also in nature it is improbable to encounter populations that are (nearly) fixed for the respective alternative alleles over all or even most contributing loci. Consequently, already the allele frequency differences of a couple of loci actually need to follow a linear model to explain most of the variance among population means – the rest is free to vary and behave (almost) like a neutral locus (see Manhattan plots above). As long as these important loci are the same over the majority of subpopulations (either due to common descent or gene-flow), accurate predictions of population trait means from a handful of loci is therefore possible.

ii) Such loci likely have an intermediate frequency, because alleles of very low or very high frequency by definition cannot have a large impact on the phenotypic population mean, even when of large effect (see discussion above).

iii) Conversely, as the reviewers at several points rightly point out, the individual phenotype can be determined by rare alleles of large effects. Moreover, a locus whose AF is important for the difference in means among populations might have rather low predictive power for the individual, because as a locus of likely intermediate frequency, all genotypes are necessarily realised with high probability within a population. The accurate prediction of individual phenotypes therefore likely requires much more loci than the phenotypic population mean. Or put differently, popGWAS might preferentially detect contributing loci that are likely not easily discovered in iGWAS.

iv) Errors in phenotyping (either by measurement or environmental influence) play a more important role in individuals than the use of robust population means over many individuals.

v) Polygenic risk scores are mainly used in humans with their complex population history, low Ne, weird LD structure and for diseases where low frequency alleles with high phenotypic effect are the main players. I presume that polygenic scores would perform reasonably well in humans if applied to "normal", i.e. purifyingly selected traits.

A respective paragraph was included in the discussion.

5. Is this approach actually good for complex traits? From the results it seems that it does moderately well for traits with a few QTL.

This depends on the definition of polygenic; the use in literature is not uniform. I tend to stick to the following grading:

< 10: oligogenic

>10 < 50 mildly polygenic

> 50 < 100 moderately polygenic

> 100 highly polygenic

This reflects the arguing in Sella & Barton (2019), where they state that most quantitative traits are directly influenced by several dozen genes, while many if not all additionally expressed genes have some background influence. But there is no generally agreed scale for polygenicity.

Moreover, the usefulness of any method largely depends on the goal:
- if it is the goal to infer *all* loci in a genome contributing to a quantitative trait and to quantify their relative influence, probably no existing (single) GWAS method and no experimental design will meet this goal, even only because not all functionally contributing loci also need to contribute to observed variance among scrutinised entities.

- if it is the goal to accurately predict *individual* quantitative phenotypes with the inferred genomic basis, e.g. for medical or breeding purposes, the suggested method may not lead very far, as detailed above, but can add a substantial number of candidate loci in addition to other methods.

- if it is, however, the goal to infer the reasons of the quantitative differences in ecologically relevant traits observed among natural populations and to accurately predict the mean phenotypes of unmeasured populations or their evolutionary potential under changed environmental conditions, popGWAS could be your method of choice.

As respective paragraph was added to the Discussion.

6. The discussion section should address the comparison between mean-based and individual-based GWAS approaches given that it is on this front that the manuscript promises to do better. The current discussion barely addresses any of the points I mentioned above, and therefore it is not clear whether there is actually any practical advantage of using one or the other.

My major claim was that GWAS is currently not widely used in biodiversity research and that this is probably in part due to the lack of data and resource efficient methods. The major point is whether it is i) good enough to yield reliable associations as starting point for further investigation and ii) feasible for the average Molecular Ecology lab. In the new version of the manuscript, given the additional simulation results and modified discussion, I think this should be more clearer now.

Minor comments:

- It is not clear from the abstract what is the innovation of the proposed approach. I will suggest that the problem is clearly described first and then how the proposed method solves such problem.

Changed according to suggestion

- line 63. "Few empirical studies are currently available" – this needs citations of the successful studies doing GWAS in wild pops, e.g. Johnston et al 2011 Mol Ecol, Pallares et al 2014 Mol Ecol, etc

cited

- I couldn't access the supplementary material, the link provided doesn't work.

Sorry for that, the link is now fixed and should work, but I hope you had access to the Supplement via the editorial manager

- fig 2, given that the main factor is number of populations scored, it will be good to make this plot separating each x-axis factor into # of population scored. The current way of presenting this data looks like PPV is really high overall, but given that the one and only thing a researcher can control is the number of pops she can sample, it will be actually very useful to show, for each simulation parameter, what's the PPV given # pops scored.

I don't agree here, because this would require taking study results *a priori* for granted.

- line 402-402 reads funny, re-phrase.

Rephrased.

- Line 423, first time FP is mentioned.

Explanation added.

- Fig 4, what does it mean that r are negative? And for some panels (e.g. panel C, 500; panel E, 12), basically 50% of the simulations show opposite signs. Please explain how shall we understand this, predictions cannot be trusted at all?

Yes, exactly. There you find the limits of the method, which should be explored in every method evaluation. However, the median predictive accuracy over all simulations was 0.86 (Suppl. Figure 4B) and only a very small proportion of simulations with (now known) unsuitable parameter combinations was actually actively misleading.

- line 494, yes, the trend is that the correlation is positive but it approximates zero pretty quickly. It will be important if the author explicitly tests for significance of such correlations and states what's the actual limit in which he thinks the method is actually useful (e.g. less than 200 loci? Less than 20?).

> Proper choices made for one system may be inappropriate for another, depending on population structure, genome size and other factors. There is likely no "one-rule-fits-all". Therefore, giving exact recommendations is likely not helpful. I prefer to relegate this responsibility to my fellow researchers applying the method. However, Supplemental Figure 7 gives a recommendation for a statistically optimal choice of outlier numbers.

It is clear that the researcher doesn't know a priori the genetic architecture of the trait, but this would make clear whether the proposed approach is truly useful for polygenic traits where hundreds or thousands of loci are expected to be involved.

> I guess that this comment is referring to Figure 1. This depends on the definition of polygenic (see above), but recommendations are clearly given and yes, for some extremely polygenic traits with miniscule contribution of each of hundreds of loci, this method may fail completely - as would likely any other method.
>
> I now distinguish in the Introduction and Discussion much more clearly between the two goals of GWAS: understand the function by identifying the underlying loci on the one hand and accurately predicting the genetic component of the phenotype on the other. As Shmueli (2010) correctly and elegantly works out, these goal are by no means identical. popGWAS certainly achieves more on the prediction side than on the complete identification of underlying loci, but is provenly much better in many situations than comparable methods.

- line 500. From Fig 1, while the trend again is consistent across distribution of effect sizes, it is clear than under the most realistic model (strong exponential), working with traits with more than 20 QTLs is not really a good idea under this model.

> 1) Whether a strongly exponential allele contribution distribution is really the most realistic, is an open question. (Sella and Barton 2019) tend not to think so and I tend to agree. However, even with a strongly exponential allelic contribution distribution, predictions are extraordinarily good.
>
> 2) As detailed above, whether the reliable identification of ~20 loci underlying ecologically important trait differences among populations and accurate prediction of population trait means is sufficient strongly depends on the goals of the study.
>
> After the extension of the simulations I hope that the results will help to get an idea on the effect of various factors affecting the robustness of the results guide researchers to plan their studies and interpret their results.

- the discussion uses a lot of imprecise statements that don't make clear the actual limitations and potential for the method. For example, line 609 " provided a sufficiently high number of populations is screened" how many?. Line 560 "increasing number of samples led to diminishing returns in statistical power beyond a certain threshold" which is such threshold? The discussion should be precise about the interpretation of the results and their implications for the implementation of this proposed approach in the field of complex traits in wild populations.

> A) I extended the simulations, results and discussion.
> B) See above; I don't think that proposing fixed values is appropriate since these may be very strongly system dependent.