Dear recommender,

Below are our responses to first round of reviews by PCI Evol Bio. Authors responses are in blue underlined text.

Please note that the formatting of the manuscript has changed, as it is being submitted for review to *Genome Biology*'s Special Issue in "Evolutionary Genomics."

Thank you,

Katie Lotterhos (lead author)

Table of Contents:

Round #1

Decision

by Sebastian Ernesto Ramos-Onsins, 2018-01-03 16:00

Manuscript version: https://doi.org/10.1101/202481

This preprint merits a revision

This manuscript addresses an important subject regarding the differential effect of genes in diverse environments, their role in local adaptation and the interference between genes that are physical linked but have divergent behaviour in different environments. The authors aim to contrast the hypothesis that "evolution in complex environments should select for modular genetic architectures with limited pleiotropy among modules". The manuscript uses innovative approaches in this field to analyse the interaction between the environmental information of each individual, the genetic linkage of genes and the genetic polymorphisms. The use of co-association networks and galaxy biplots are especially informative and show the main results of this work. Also, it is of crucial importance the simulation analysis performed in this manuscript to understand the possible expected patterns under different scenarios; this analysis includes three different demographic histories (isolation by distance from (i) equilibrium, expansion from (ii) single-refugia or from (iii) two-refugia) and explore the allele frequencies of loci under neutral and under positive selection, contrasted with information on a number of generated environmental variables.

The authors used as a model species for such analysis the species Pinus contorta (lodgepole pine). Using co-association networks identified several non-overlapping modules of genes

associated with environmental factors (Aridity, Freezing, Geography and Multi, this last not clearly associated to a single environmental factor). Surprisingly, physical linkage was observed between genes associated with different climate modules but seem not to affect importantly the different modularity of genes to different environments. The results obtained here using novel approaches in this field may help to understand the effects of complex and heterogeneous environments over the genetic variability and over the gene interactions in the genome.

Nevertheless, the manuscript needs to be more clearly explained. Although there are many concepts extensively described in the introduction, it is often difficult to understand the meaning of sentences including words such as modularity, architecture and pleiotropy, which can have different meanings in different context. The two reviewers also coincide in the difficulty to understand some sections of the manuscript.

We admit that these terms are difficult to write about because, as noted, they are used in many different ways in the literature. We have taken the valuable advice of the reviewers and (i) rewritten the introduction, (ii) added a table to help with terminology, and (iii) reframed and expanded on major areas of the discussion.

I am especially curious about the simulation methods. I believe the methodology used for simulation, as well as the specific parameters used, should be included with more detail in order to facilitate the replication of the analysis. I am wondering about the possibility of including the effect of recombination in such simulations, which may also help to the interpretation of the data.

Since these simulations have already been published and code used to reproduce all analyses will be included with the manuscript, certain details were omitted in the first version. We have improved the manuscript with more details about this section.

Finally, the two reviewers give an essential number of comments that the authors must follow before having the recommendation of PCI in Evolutionary Biology. Please answer all the comments of the reviewers separately in a separated text and detail all the modifications included in the manuscript. I encourage the authors to revise this manuscript following comments from this round of review and resubmit to PCI in Evol. Biol.

We hope you find our revised manuscript worthy of recommendation.

Dr. Sebastian E. Ramos-Onsins

Reviewer #1

Reviewed by Tanja Pyhäjärvi, 2017-12-05 14:45

Review of Modular environmental pleiotropy of genes involved in local adaptation to climate despite physical linkage

Katie E. Lotterhos, Kathryn Hodgins, Sam Yeaman, Jon Degner, Sally Aitken

https://doi.org/10.1101/202481

The paper presents (to my knowledge) a novel approach to analyse environmental adaptations using genetic polymorphism and environmental data. The main idea is to inspect the modularity

of genetic architecture by joint co-association network analysis of environment and genetic polymorphism. The authors use experimental data from Pinus contorta exome sequencing with climate data. In addition, they simulate multiple demographic, selective and neutral scenarios to test the behaviour of the method. The concept and evolutionary implications of multivariate nature of climatic variation were nicely explained. Also, discussion on the caveats of using the principal components (or equal approach) to summarize environmental variation's effect on biological organisms was important to bring up. The new approach is attractive and applicable to wide set of systems. Clarifying some aspects of the analysis would improve getting the main message through.

Major comments:

The preprint presents what I presume to be a reanalysis of exome resequencing genetic polymorphism data across 281 Pinus contorta populations from Yeaman et al. 2016 (Science). However, the authors should state clearly whether this is the same data or something that has not been published yet.

This is now more clearly stated in the 2nd to last paragraph of the "Background" section and in the "Methods".

Because the modularity is taken into account for both gene groups as well as for environment, it was a bit confusing at some points which modularity was referred to. First I assumed that the paper only presents modularity in environment (as suggested by the title). However, the abstract presents modularity mostly for the genetic architecture. The basic idea of modularity of genetic architecture could be presented more clearly from the beginning (or in the title?) via an example or maybe even a figure presenting the both to make it easier to grasp from the beginning. Please explain clearly what is meant and is there a difference between: "evolutionary modules of loci", "selectional modularity of architecture", "modularity if selectional pleiotropy", "developmental/functional modules" and "environmental module" or are they essentially the same thing.

Thank you for these useful suggestions to improve terminology and accessibility of the manuscript - we agree that these concepts are wrought with different meanings in the literature, which makes them difficult to write about clearly. To be clear, the co-association network is based on modularity of loci in their associations with the environment (the genetic architecture), and not the modularity of the environment itself. Since this is a new presentation of pleiotropy not previously addressed in the literature, we coined the term "environmental pleiotropy" to capture what the co-association networks are measuring. To clarify terminology, we have added a table (Table 1) and a new conceptual figure.

Why were SNPs that were already top candidates based on preliminary analysis only used in the co-association analysis? If the univariate model is likely to miss some signals of adaptation, why limit only to SNPs based on correlation with single environmental variable?

To answer the first question, the simulations show that neutral loci can form interesting co-association networks, and so it is important to exclude them and only use SNPs with high confidence for adaptation. The co-association network approach represents a framework for interpretation once top candidates are identified, not another genome scan method to search for outliers. This is now more clearly stated in the paragraph stated on lines 117.

In response to the second question, we do not state in the manuscript that univariate models are likely to miss some signals of adaptation. We do state that associations with principal components may miss some signals of adaptation that are detected with univariate associations. Univariate methods are probably going to miss adaptation loci, but that question was not what we were interested in addressing with this analysis. The goal isn't to identify new loci, but rather to use clustering to better interpret how selection has shaped the variation.

The authors state that the results are not sensitive to the distance threshold used in the clustering of the networks. What other thresholds were tested and why 0.1 was chosen as a threshold?

We chose 0.1 as a threshold based on the simulations. The simulated data showed that this threshold enriched for selected loci to be strongly connected to each other in a co-association network, and for a lack of connections among neutral loci with each other or among neutral and selected loci. This is now shown more clearly in a new figure (Supplementary Figure 13).

While we recognize that it may be advantageous for all studies to use the same threshold, at the same time we do not want to advocate that this threshold be used for all studies because there may be some nuances for genomes from different species.

Further, as the co-association analysis is in the core of the paper, it would be interesting to know, why the authors ended up using the division into 4 clusters. Are the conclusions dependent on the number of clusters?

To be clear, this division into four main clusters is for the purposes of presentation only, and was necessary (i) for clear data visualization and (ii) computational limits. This is now more clearly stated in the "Environmental response modules" section of "Results". We believe that each individual module - as visualized with the co-association network - are more meaningful representation of groups of loci that respond similarly to the environment.

To emphasize this better in the revised manuscript, we re-colored and re-ordered loci in Figure 1 (and subsequent figures) according to their co-association module (rather than by the number of outlier SNPs in the first version). We hope you will find this new visualization a better representation of the data, as we have.

For example, the distance between "Freezing" and "Geography" is not very large and the clusters are actually quite similar based on visual inspection.

The "Geography" group differs from the "Freezing" group in two important ways that are shown in the clustering/network plot and galaxy plots. Most Geography SNPs have associations with latitude but not frost variables, while most Freezing SNPs have associations with frost variables but not latitude.

Also, will the clustering be identical when structure corrected associations are used?

Interestingly, the clustering is not perfectly identical with the structure-corrected associations were used, although it results in broadly similar associations as shown in the Supp Mat. We argue that raw associations would be the correct associations to use, because sometimes

correction for structure changes the sign of the association because of the orientation of the SNP frequency relative to the neutral structure in multivariate space (as can be visualized by comparing the raw and corrected patterns the galaxy plots).

Minor comments:

Simulations were a great addition. Please give more details about the demographic simulations. For example, what was the strength and duration of selection in relation to demographic history in the simulations?

We added these details to the methods.

Exactly how many individuals were sampled per population? Is each of the 281 populations represented by single individual? If only one per population was sampled, how were the allele frequencies for Bayenv2 obtained? Do seedlots refer to a set of seeds from multiple trees, e.g. can it be assumed to be a random sample from the population?

We added these details to the methods and to the last two paragraphs of the "Background" section. Basically, a "seedlot" is a sampling location.

How does your approach relate to the idea of analysing gene networks jointly when identifying the genetic basis of local adaptation that was presented e.g. by Daub et al. 2013?

We assumed the reviewer was referring to the paper. "Evidence for Polygenic Adaptation to Pathogens in the Human Genome." In this paper they use a gene set enrichment approach, in which *a priori* sets of specific biological pathways are tested for enrichments of associations with phenotypes.

Our approach is not a gene set enrichment approach, and we would not be able to identify specific biological pathways in our data because many of our top candidates are not annotated. In addition, co-association networks are not an enrichment approach, but a framework for understanding multivariate patterns of associations across candidate genes. We did test for enrichment of GO terms in our candidate set, and these are presented in the results and supplementary tables.

Our approach is more similar to the use of co-expression networks used to identify modules in RNAseq data. Our approach is distinct from the application of bipartite networks previously applied to the study of pleiotropy, and this is now discussed in the section "Benefits and caveats of co-association networks" in the Discussion.

Add citation to Hill and Robertson 1966

This paper is cited in the "Background" section.

Structure correction is mentioned briefly, but what is the overall structure pattern in P. contorta?

The overall structure pattern is one of isolation by distance, which is now indicated in the 2nd to last paragraph of "Background".

"Across" repeated in Linkage disequilibrium part.

Fixed.

What do you mean by "SNPs from most genes associated with only a single climate module"? Based on figure 1, each SNP can only be associated with one network.

This sentence was rewritten for clarify.

Could you present what kind of results would reject the Hypothesis of Modular Pleiotropy?

We now use Figure 1 to compare the patterns expected under modular vs. universal pleiotropy.

Replace “Thus, it..”, with “Thus, it is…” in Discussion.

This sentence has been revised.

How sensitive the method is to the choice of environmental variables or SNPs? Are the networks and modules highly dependent on single variables? Also, how many nuisance variables (environmental variation that are not selectively relevant) the analysis tolerates?

From the simulations we performed, we can at least say that co-association networks correctly cluster selected loci when many variables are measured, but the true causal environment is excluded (see “Results: Simulations”). Thus, the approach appears to be able to tolerate nuisance variables (see also “Discussion: Benefits and caveats of co-Association networks”), although we did not quantify “how many”.

A fuller evaluation is beyond the scope of this manuscript, as they require a set of more complex simulations that the lead author has been working on for some time. So, a more comprehensive evaluation on more realistic simulations is forthcoming.

Figure 1a, text in the gray background is tiny and hard to read.

We removed this text from the figure and added a table (Table 2) to explain the abbreviations.

Figure 3, the shading of the quadrant does not reproduce in the printouts.

Thank you for letting us know. The shading of the quadrants did reproduce on our printers, but we darkened the shading slightly for the revised manuscript.

Figure 2, The among-group LD patterns are almost invisible in the screen and completely invisible in the print-out.

Note that the LD data had to be transformed to be plotted with the recombination rates, and this transformation affected some of the visualization. Generally, this made patterns of LD less than 0.05 less visible, but these are negligible. We also provide a supplementary figure of LD patterns.

Tanja Pyhäjärvi, University of Oulu, Finland

Thank you so much for your useful comments!

Reviewer #2

SUMMARY of Modular environmental pleiotropy of genes involved in local adaptation to climate despite physical linkage

The authors present an important advance in how we think about and measure pleiotropy using co-association networks and galaxy plots. As a geneticist studying pleiotropy and someone who is deeply interested in the genotype to phenotype map, I am very intrigued by these results and novel methods. The figures are beautiful and well thought out. The section in the introduction about the debate surrounding how to define traits (cold, dry, cold-dry) is among the clearer parts of the paper. A major impact of the paper is the 'co-association network' approach, which solves the aforementioned debate by finding clusters of genes that define relevant traits (or in this case, relevant environmental responses).

A major problem with the paper is that it is often unclear and important sections are difficult to follow or seemingly left out. Because the authors are studying pleiotropy, they necessarily have to deal with multiple complicated topics at once, including (in their words), "the modularity of the architecture (number of distinct climate factors), overlap among modules, and physical linkage among loci". There are many times when I am not sure which of these topics is being addressed. Some of this confusion comes from terminology, for example, genetic architecture usually refers to genome structure or linkage, but occasionally the authors also use architecture to refer to the number of distinct climate factors. The word modularity can also be very tricky because it can refer to the number of distinct climate modules or the number of distinct genetic modules that respond to these climates. The terminology in this manuscript should be used more carefully.

We agree that it is difficult to write about these terms which are used in many different ways in the literature. In this new version of the manuscript we have carefully edited for terminology and semantics, including a new introduction, table for terminology, and conceptual figure.

The first sentence of the abstract is another good example of the unclearness: "Physical proximity among alleles shaped by different sources of selection is a fundamental aspect of genetic architectures critical for predicting their evolution". I did not understand this sentence at all. In addition to terminology concerns (e.g. genetic architecture is not yet defined), there are major grammatical errors that prevent a reader from discerning the object and subject of the sentence (e.g. are the allele frequencies shaped by selection, or is their proximity shaped by selection? which of these are you trying to predict?).

This sentence refers to physical linkage among genes and has been revised for clarity. Both allele frequencies and the proximity of alleles may be shaped by selection, and in this manuscript we are not trying to predict either but rather describe them in a way for understanding pleiotropy and modularity.

Further, this sentence summarizes many different, difficult concepts simultaneously and it is unclear which is the main focus. It would be better to start with one central concept, and then step through the next concepts one at a time. i.e. "Are there modules of genes that respond to particular environmental changes? This is important to understand for reasons X, Y, and Z. We might be able to answer this question by looking at [describe data]. But physical linkage among loci can obscure this signal..."

The entire introduction has been revised and reorganized to emphasize the main focus of the study. We have also added a conceptual figure to explain the framework and guide the reader.

Given this paper has the potential to demonstrate high impact findings, it is worth the time to restructure the manuscript, stepping through each topic more slowly and making clear to the reader (perhaps with numbered lists in text) what are the difficulties in measuring the extent of pleiotropy and how this method deals with those difficulties. Sometimes, instead of relying on words like "pleiotropy," perhaps it would improve clarity to talk about "genetic changes that affect multiple environments", and the genotype- phenotype map. Instead of staying so close to the terminology, and to the literature regarding modular pleiotropy, I wonder if the authors might use less technical terms to explain the motivation for this work, and its significance to a broader audience.

We appreciate that you highlight the potential for high impact of this research. We want ~~it~~ this study to reach a broad audience, while at the same time guiding the reader through the different ways these terms have been used in the literature and how terms relate to our study. In this new version we added a Table for terminology, and taken care to use less technical terms in the introduction.

In sum, this manuscript presents a novel approach to studying the genotype-phenotype map, and also a novel way of thinking about and defining phenotype. I recommend the authors re-structure the paper, stepping through its important contributions one at a time to improve clarity.

INTRODUCTION

This sentence in the introduction indicates an omission of the literature: "Although there is emerging agreement that organisms have modular organization of genes in their effects on phenotypes." There does not seem to be agreement at all and I think the authors should acknowledge this. The cited paper from Boyle et al 2017 did not acknowledge this and authors apologized over Twitter. There are many papers by Hill (1,2) and by Rockman (3) that do not support the modular pleiotropy model. Further, these are many "reply to" papers that highlight just how controversial this topic remains (1,4,5).

The introduction is a little bit unfocused. The authors review a great deal of literature and terminology, but the main message of the paper is a little lost. I think this is a difficult task because so many issues are tackled in this introduction. A single sentence early on explaining a single, simple important contribution might help a reader focus.

The introduction has been completely revised, with the intent to provide a roadmap for the reader.

There seems to have been another study that generated some important data used in the current study. I don't understand, even after re-reading the paragraph many times, what these data look like. This seems problematic, as these data are key to the current study. What are "candidates"? Are these "candidate loci" or "candidate trees"? How many are there?

In this revision, we have taken care throughout to specify whether we are talking about "candidate SNPs" or "candidate genes."

Why is convergent adaptation relevant here? This paragraph should start off with a comment about pleiotropy, and how these data will allow authors to study pleiotropy. Not all details from this dataset need to be described, but the details relevant to the insights about modularity and pleiotropy should be described here. How do these data differ from those used in previous studies of pleiotropy? Does this feature of the data give you more power?

Because this is a re-analysis of the same dataset, it is important to describe how this approach is yielding insight not achieved in the previous publication where convergent adaptation was the focus. We added some of the details requested by the reviewer to show the uniqueness of our dataset in the last two paragraphs of the "Background".

Authors sometimes omit the word "frequencies" from the phrase "allele frequencies". Authors should be clearer when they are talking about an allele itself versus its measured frequency across environments.

We found two instances where the word "frequency" was omitted for this phrase and these were revised.

How does a co-association network analysis "characterize pleiotropy and linkage of the architecture?" What features does it use? Is this a new approach? What are its benefits? What novel insights might it provide? I understand that this is the introduction, so you don't want to give away the results. But you must state something about this novel approach and why it is a good idea.

We addressed these points through a new conceptual figure, now Figure 1.

METHODS:

This is very difficult for anyone to read who does not study plants.

What is a seedlot?

How do you have needle tissue? I thought you only have seeds ("seed"lot)

Did you grow these trees in different environments? Or did you sample trees from different environments and calculate how allele frequency changes across environments? This part of the methods is really unclear.

A seedlot can be thought of as a sampling location. Seeds were collected from different sampling locations, grown in a common garden, and needle tissues were collected and sequenced. This is now clearer in the "Methods" section.

Why did you obtain samples and extract sequence information if Yeaman et al 2016 already did this. Do you have different samples? What is the difference between your work and the Yeaman study?

The same dataset was used, but the Yeaman 2016 focused only on comparative genomics and convergent adaptation with another species. This is now clearer in the last two paragraphs of "Background". We added additional clauses to the methods to highlight which analyses were the same as those already reported.

The co-expressed genes section is also missing relevant details. What 7 climates? How many plants per cluster? How did you choose the plants in each cluster?

The co-expression data was previously published, so some relevant details were glossed over in the first draft. We have now included more details and clarified our analysis with respect to the gene expression patterns. In the gene expression analysis, the individual plants were not clustered, but rather genes were clustered based on their similarity in expression profile across the seven climate treatments tested in the earlier paper.

RESULTS:

In the first paragraph, again it is not very clear what is relevant about these top candidates in terms of pleiotropy or modularity. Why are you studying them? You provide information about how many there are before really explaining why they are in the paper.

The study of pleiotropy and modularity is relevant only to loci under selection and not neutral loci, which are putatively the top candidate genes and the top candidate SNPs within the genes. We added text in "Background" and a sentence at the beginning of the "Results" to clarify this.

In paragraph 2, again the motivation is not clear. The co-association method is described nicely in the methods. But I am left to infer what these SNP-environment associations mean and why you might want to cluster them. At the end of the co-association network section, perhaps add a final sentence to summarize the observation. Which of your hypotheses was supported and what does this mean about selection and about pleiotropy?

We hope that you find the new conceptual figure helpful in providing the motivation for this section and summarizing the hypotheses. Much of the "Discussion" has been expanded to address these questions in further detail.

I think the section on PCA is appropriate to include as it finally explains some of the benefits of co-association networks over other approaches.

The section on benefits and caveats of co-association networks also seems beneficial to include. However, it focuses on yet another impact of this study, identifying high confidence SNPs. It does not focus much on pleiotropy or modularity, these words are not even used. It would be very beneficial to pick a single contribution of this paper and continually come back to it in the text. It is not necessarily confusing to include multiple contributions and insights within a single manuscript, but one in particular should be highlighted as the main focus. Otherwise the paper reads as unfocused and the motivation/impact is unclear.

The entire "Discussion" has been revised and expanded to illustrate the three main insights gained by using the approach (sections of "Discussion" starting with "Genetic architecture of adaptation..."). We hope this clarifies the motivation and impact of the study.

The last two sections of the "Discussion" focus on the benefits and caveats of the approach, which have been revised to relate more directly to the interpretation of the main insights.

In sum, the co-association analysis is a novel and important idea. I hope with some re-structuring and decisions about how to focus the manuscript this idea will reach a broader audience.

Thank you for the useful suggestions!

1. Hill WG, Zhang X-S. Assessing pleiotropy and its evolutionary consequences: pleiotropy is not necessarily limited, nor need it hinder the evolution of complexity. Nat Rev Genet. Nature Publishing Group; 2012 Apr 1;13(4):296–6.

2. Hill WG, Zhang X-S. On the Pleiotropic Structure of the Genotype–Phenotype Map and the Evolvability of Complex Organisms. Genetics. Genetics; 2012 Mar 1;190(3):1131–7.

1. Rockman MV. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution. 2012 Jan;66(1):1–17.

2. Paaby AB, Rockman MV. Pleiotropy: what do you mean? Reply to Zhang and Wagner. Trends Genet. Elsevier; 2013 Jul;29(7):384.
3. Wagner GP, Zhang J. Universal pleiotropy is not a valid null hypothesis: reply to Hill and Zhang. Nat Rev Genet. Nature Publishing Group; 2012 Apr 1;13(4):296–6.