Evolution at two-time frames shape structural variants and population structure of European plaice (*Pleuronectes platessa*)

This paper explores population structure and variation in two structural variants (SVs) in European plaice. Previous work identified these two SVs on chromosome 19 and 21 and this study further explores the variation by incorporating additional sampling locations and attempting to date the SVs by comparison with other species. The paper also investigates whether the SVs could have been introduced through introgression from another species that often hybridizes with plaice. Further, the paper also uses the new data to re-assess population structure in the species and examine demographic histories to investigate the strong divergence that has been found between Icelandic populations and other European populations.

The results of the population structure and demographic analyses reveal that the divergence of Icelandic population may be explained by the possibility of a different glacial refugium (different from other European populations). This is supported by their genetic divergence, demographic modelling, and by no clear reduction in genetic diversity in the Icelandic population. In addition, the study finds support for isolation by distance particularly when excluding SVs from the analysis. The authors claim this is the strongest IBD documented for a marine fish. Although population structure is generally weak. In addition, the study finds that the two SVs are old (>220,000 years) and shows that the frequency of the derived haplotype increases at the range edge along with a reduction in genetic diversity (genome-wide). The SVs were diverged from other species and the study finds no support of introgression of SVs.

Overall, I think the paper is interesting. Given the importance of SVs to population structure, I think this paper would be of interest to many readers. While the study investigates an important question, I have several concerns that the authors should address. These concerns are all detailed below; however, a few major points include:

1- Flow and organization of the manuscript: The introduction provides extensive information on structural variation, but this is not the only thing that the authors are interested in. The authors also investigate overall population structure and diversity, as well as investigate demographic histories. This is not well established in the Introduction, and (as detailed below) it is not clear from the Introduction why the authors investigate demographic history with the Icelandic population. This needs better context and requires some reorganization of the Introduction and Methods to improve flow.
2- Methods – I think there are a few cases where parts of the methods need to be clarified. Why were loci pruned for LD if the primary point of the paper is to investigate SVs, which will have many loci in high LD? And by removing loci that are out of HWE, could this remove some loci that are of interest for complex structural variation? This is important for neutral population structure, but authors don't attempt to remove other outliers before investigating population structure. I would also suggest that authors investigate overall population structure with neutral loci (not all loci). This also relates to the methods that were chosen here to

investigate population structure (PCA). Structuring can be explained by a few important loci. It is not clear which loci contribute to the population structure overall, as there is no information on the loadings or significance of loci on PC axes. Was it many loci or just a few loci that are under selection that separate the populations?

3- In addition, I am not sure about the dating of the SVs. Selection and reduced recombination on SVs can quickly lead to divergent haplotypes, I am not sure if the branch length of trees will be informative for dating the inversion. My thought would be that the earliest date that can be given to the formation of these SVs is at the timing of the split between Iceland and other populations (what is this timing, it is not clearly stated in the text?). However, if there were two refugia it is possible that the derive haplotype could have evolved only in the European refugium (southern) and then have been introduced into Iceland after the last glacial maximum (through secondary contact), so it may be even more recent than the split. I would like the authors to explain how these analyses with the SVs do not violate the assumptions. If they do, then I am not sure how informative they are. Further, the trees are compared to a collinear region of chromosome 19 to represent genome-wide average. But using this chromosome may be problematic given that the SV exists on this chromosome (only 5MBP away). I would suggest using a chromosome that does not contain SVs.
4- There is a very short section on functional annotation in the SV regions. But not much detail is provided and I could not find access to Supplemental File 1 (List of genes). I would like to see a bit more discussion on the genes in the SVs. Have they been found in other SVs in other species (particularly marine fish)? Did the authors perform an enrichment analysis? With over 1800 genes in the SVs, all that the authors say is that many are associated with ion transport and other functions like sexual recognition. There are >1800 genes, so how many are associated with ion transport? This section needs more details and these details may provide some hints as to the function of the SVs. And as mentioned below, it may be most interesting to examine which genes are found within the SNPs where FST and Ho are 1.

More detailed comments are provided here below. Authors should also be careful to correct grammatical errors throughout the text. In addition, there are many areas throughout the text where authors should be more specific, instead of saying “many” or “some” or “several”, please indicate how many there actually are (see comments below for examples).

Abstract – The abstract requires more details as many parts of the study are not presented in the Abstract. The findings of the Icelandic population representing a potential different glacial refugium seem important in the text, but I don’t see it mentioned in the Abstract? Additionally, I understand that the range of dates for the SVs is 550-220 kya, but the Abstract only indicates that they evolved around 220 kya. Also, there is no mention of the tests for introgression.

Line 13-15 – Indicate that this is known from previous work.

Line 14 – and “shows” strong genetic differentiation…

Introduction – There is a lot of information on SVs here, which provides a nice summary of a lot of the literature. However, the Introduction misses an important part of the study. The introduction focuses solely on SVs, which is not the only thing that this paper investigates. There is no mention of the Icelandic population here and analyses of demographic history. The inclusion of analyses on population structure and demographic history need to be better outlined in the Introduction. It is not clear from the Introduction why the dadi analysis (demographic history) was done with Northern Europe and Icealand. This needs more background and better context in the Introduction. In the study, the authors suggest a previously unknown refugium may have existed in Iceland, but it is not clear why this question was even investigated. There should be more discussion about other aspects such as population structure and not just a focus on SVs.

Line 34-35 – Is this always true? If an inversion is introduced through secondary contact/hybridization, couldn’t both homozygotes be present initially? Maybe indicate “*de novo*” or “initially”, or reword for clarity.

Line 40 – “SVs are likely to evolve incompatible alleles”. Why? Explain.

Line 43-44 – I’m not sure if I understand what is meant by “become trapped in environmental gradients”. Also, physical barriers to gene flow would imply that populations are NOT fully connected? Please clarify this section.

Line 49 – Is “evolving” necessary here? Consider rewording this part.

Line 54 – Consider changing to “maladapted” instead of “unadapted”

Line 56 – Fix wording here.

Lines 61-63 – Not sure why allele frequency clines are relevant in this sentence? Why does having mutations in SVs make clines more likely? Not clear.

Line 70-71 – But this can only occur if genes with functional relevance are found within that region of the chromosome/SV?

Line 106-108 –Provide references for these biological characteristics.

Line 108 – What type of markers/how many? Microsatellites?

Line 115 – in European plaice specifically? Or in all four species? This was not clear. I think more information on this study in European plaice is needed here (Lines 112-118). For the “larger geographical scale (line 118)” indicate where these new samples are from to provide better context.

Line 119 – There is no mention of the Icelandic population here and the examination of demographic histories.

Methods

Line 138 – Samples were collected during spawning. What do we know about the distribution of these fish outside of the spawning season? I am wondering if they remain close to these spawning grounds or are they highly migratory?

Line 182- What does present mean here? Does it mean that it was genotyped in >80% individuals (missing in <20%)? Or something else?

Line 185 – I wonder if removing SNPs that are out of HWE could remove potentially informative SNPs associated with complex structural variation. Can the authors comment on this?

Line 196 – Population structure should also be explored for neutral markers. I suggest performing analysis with all loci, neutral loci, and SV dataset(s). Or perhaps a neutral, outlier, and SV dataset. Although there are many ways to identify outliers, it might be easiest to use a method such as PCAdapt and remove loci that are outliers on the first couple of axes that explain a lot of the variation. It is difficult to know whether the patterns observed in Figure 1b are due to potential adaptive differences among populations or due to neutral structure. In a PCA differences can be driven by a few important loci. If such is the case, it would also be nice to see which part of the genome are responsible for these differences observed here (Fig. 1b). Other types of analyses for population structure may also be helpful, such as STRUCTURE (or ADMIXTURE).

Line 199 – If you are looking for structural variants, shouldn't you not prune for LD? This may remove loci that are linked due to SV?

Line 211-222 – It was not clear from the Introduction why this analysis is being conducted. It needs to be clear how this is related to SVs (the focus of the paper).

Line 238 – How many loci had H_O and FST of 1 and for which inversions? These loci may be especially important for understanding the functional differences between genotypes. The genes that these SNPs are located within may be especially informative.

Line 253-255- It is not clear to me what was done here to calculate average/smoothed FST?

Line 280 – Why did you choose a collinear region of chromosome 19 to represent genome-wide average? Why not choose a different chromosome that does not contain any SVs, as the dynamics of collinear regions on chromosome 19 could still be influenced by the presence of the SV (which is only 5Mbp away). This does not seem like the best approach.

Line 295 – These inversions are likely not neutral (or at least that is what this study suggests). I have often wondered how to appropriately date inversions. I don't see how these analyses can be applied to inversions without violating many assumptions. Please explain.

Results

Line 317 – Typo (need space between 'the' and 'Transition')

Line 333 – Provide information about timing in the text.

Line 334-342 – It seems confusing to go back to Figure 1 here after this population structure has already been discussed above (Lines 302-307). This is an area where better organization could improve the manuscript. As mentioned, the demography history (above; Lines 327-333) seems out of place in the middle of this. This is where providing better context and describing all objectives in the Introduction would help with the flow of the manuscript.

Line 336-337 – Manhattan plot or data to support this statement?

Figure 3 (D1/2) – What does R2 indicate in this figure? The mean R2 for the locus with all other loci on the chromosome?

Line 360 – How much lower?

Line 360 – Be specific in the text. "Several SNPs"-> How many?

Line 362-363 – Not clear where the reduction in FST is in the figure? Do you mean the huge gap in SV21 (where there is low LD), which likely due to differences in position along the chromosome compared to the reference genome of flounder? Or another region? I also don't see a specific region of low FST for SV19? Moreover, I don't see a concordant increase in π. This needs to be better labeled in the figures.

Line 376-377 – There are over 1800 genes here, how did you determine that ion transport was important? Looked for over enrichment? Also I could not find access to Supplementary File 1?

Line 379-390 – Again, I am not sure that using chromosome 19 is the best option for the genome-wide estimates.

Lines 391-400 – How many SNPs were used here? It seems clear from the tree that this is not a case of ancient introgression (Line 398-400)?

Discussion

Line 411 – derived "form"

Line 412 – edge “of” the plaice distribution…

Line 417 – How much greater is this IBD relative to other marine fish? What is the magnitude of difference?

Line 417-420 – Was the geographic distribution of previous studies similar to this one? It seems surprising that if this is the strongest IBD detected for a marine fish that all types of markers should be able to resolve this pattern.

Line 433- Are there any mitochondrial data from this species that would support a different glacial refugium? Or is such a scenario observed in other marine species? This would be useful to know. In addition, a scenario of secondary contact was most likely. How does secondary contact fit into this history? I don’t see it mentioned in the Discussion here.

Line 449 – References for these time frames?

Line 421-460 – This section seems a bit out of place. The entire Introduction focuses primarily on structural variation, so the population structure doesn’t flow well with the general goals of the paper (as discussed above).

Line 462 – I’m not sure that it was shown that the large SVs were responsible for the ‘main population differences’. They explain individual differences, but without their inclusion you see more IBD? It seems based on FST they are important for differentiating the North Sea and Baltic, but other comparisons don’t seem to show as high FST for both SVs? Based on the PCA, it is not clear what genomic regions are driving population differences. The PCA should be done without the SVs at the very least (and also with outliers removed).

Line 469-471 – There are no environmental data in this paper. I understand that the environmental break in the Baltic is well known but you need environmental data to back up this conclusion here. It would be possible to incorporate data from online databases. MARSPEC and Bio-ORACLE have salinity and temperature data that could be used here to better explore these associations. These can be accessed online and also using the ‘sdmpredictors’ package in R. This may be useful to the authors or even just to show on a map to those who are unfamiliar with this region. Further, since the fish were collected during spawning, do the environmental conditions at their capture site reflect the environmental conditions they would usually experience? Or are they migratory species?

Line 471-472 – Why not examine selection in your study? Test of selection could be performed on the genotype groups?

Line 498 – 499- Please provide examples of this phenomenon in other species/inversions? I was not aware of this pattern (which is not easy to see in the figures). In another recent study with a large inversion in a salmonid, they actually find

an increase in FST in the middle of the inversion and a drop in nucleotide diversity (https://www.biorxiv.org/content/10.1101/504621v1.abstract).

Line 506 - wording

Line 512-514 – If this is true, would there not be drift occurring in these SVs? Should Iceland look different compared to the rest of the populations given the long divergence time?

Line 515 – "do not include effects from selection or recombination" – does this mean it is likely shorter or longer? I am not sure that you can accurately date an inversion by branch length given that inversion are much different than other parts of the genome. Selection and reduced recombination could quickly lead to very divergent haplotypes. Assuming that Hap 2 is the ancestral allele, then given that the SV19 inversion is polymorphic in Iceland and that Iceland contains Hap1 (derived) for SV21, the best date you may be able to give both inversions is only as early as the time of the split using other loci. What is the split date between Iceland and other populations (this was obtained from demographic modeling?). Demographic history also suggested secondary contact after isolation, so it could be possible that variation in the inversion was introduced into Iceland after the split and was not present before.
I'd be happy to learn how these methods are appropriate and also better than just using other loci to date the split.

Line 522-529 – This is a really cool idea. Are there any other candidate species that could represent a possible source of introgression of this SV? Could you provide examples of some candidate for future studies?

Line 55-547 – Why not test for selection in this region? Perhaps the composite likelihood ratio (CLR) method in SweeD would be appropriate here (it is more robust to variation in recombination, ascertainment, and demography than Tajima's D). You could test it within each haplotype group (homozygotes for Hap1 and Hap2).

Line 550 – I would also be interested to know if there are some life history traits that may be associated with this inversion. It seems many inversions underlie complex phenotypes (e.g., mating and migration strategies). Are there any traits that could be of interest, such as migration behaviour? Or age-at-maturity? Colour morphs? I am not sure. But it would be interesting to know about any possibilities.

Line 578 - Fix wording here.

Line 596 – There is not really much discussion on the genes in the inversions. I think this would be interesting to explore a bit further. Were any processes over-enriched? Are any of the genes in the inversions found within inversions in other species?

Figure S6 – What does "mean FST" represent? What is the comparison that the points represent? Needs more information in caption.

Figure S7 – Include something to indicate the position – or at least which end is the start of the chromosome.