Reviewer 2

Major comments:

The authors spent most of the Results section on testing correlation between *B* and various summary statistics, and interpreted a lack of correlation between *B* and a statistic as evidence for BGS having a weak effect. There are several problems with this approach.

First, *B* is for predicting the diversity-reducing effects of BGS in a panmictic population of constant size. But the authors were using it when the population was subdivided. It is perhaps unsurprising that most of the correlation reported were weak, because *B* is not meant to be used in this context. Hence, it is unclear what the strength of the correlation means, and equating the strength of this correlation to the effects of BGS on a statistic of interest seems unfounded, if not dangerous. If this type of correlation is to be calculated, using the equations derived by Zeng and Corcoran (2015) makes more sense, because at least we know population structure is included in the equations. But even then, equating the strength of the correlation to the strength of BGS is probably unwarranted because it ignores the fact that different statistics have different levels of variability, and the strength of correlation is expected to be weaker for statistics such as F_{ST}, which tends to be noisy because it involves the ratio between other statistics.

Second, *B* was used as a measure of the strength of BGS. However, this may not be true. *B* only tells us the expected reduction in diversity due to BGS (in the ancestral population under the current model). However, the level of reduction does not necessarily predict how strong an effect BGS may have on the shape of the site frequency spectrum (SFS), which probably have an effect on some of the statistics considered. It's possible that distortion in the SFS is maximal for intermediate values of *B*, and this may be affected by, e.g., the divergence time. Whether this is the case requires careful examination.

Third, average *B* values are probably uninformative about whether BGS will cause a problem with F_{ST} outlier tests. Take the "no recombination" case as an example. All sites in the region will have the same *B*. In this case, we would not expect BGS to have any effect because there is no variation in N_e across the region. In contrast, BGS may have

an effect when there is significant variation in *B*. Thus, measuring how variable *B* is across the region may be more informative. But as argued below, the effect is likely weak under the models considered here.

We completely agree with all of the three points. We have re-structured the paper so that our first result is that the treatments with and without selection do not differ in the mean (and general distribution of) F_{ST}. We have highlighted this in the new figure 1.We include the correlations with *B* as additional evidence. Having said that, the equations of Zeng and Cocoran (2015) heavily rely on *B* in their derivation of the F_{ST} with BGS, so this seems like the best summary of the strength of background selection available. An argument for making an important point about these correlations is that they are more directly compared to works that look for such correlation between *B* and H_S (but not between *B* and F_{ST}) such as McVicker et al. (2009) and Elashiv et al. (2016).

The observation that BGS does not cause a big problem in F_{ST} outlier tests is perhaps not surprising given the models used. This can be seen from the fact that the 10kb focal region "contained on average 0.44 genes for the human genome and 3.15 genes for the stickleback genome". In other words, most of the simulated region were consisted of neutral sites. In addition, within the selected region, where BGS is strongest, the mutations under purifying selection are not expected to differentiation between populations. Since about 75% of the sites in coding regions are nonsynonymous, only the remaining 25% of sites in the selected region are expected to be affected by BGS (it's unclear whether there were neutral sites in the selected region; see below). Thus, relatively few sites are affected by BGS in regions where its effect is strongest. Given that F_{ST} tends to noisy, it may not be unexpected to see BGS having little effect on false positive rates. Although showing the BGS probably has a weak effect on this scale is useful, the authors should acknowledge the limitations of their experimental set-up.

The gene density that we use is the gene density typically found in eukaryotic species and is based on empirically determined recombination maps. It is therefore seems to us to be biologically relevant. If the goal is to formulate expectations about the effect of BGS in real populations, then considering this gene density is appropriate. It is much more appropriate than considering selection on every site as it has been done in previous studies. This reviewer's intuition is correct that BGS has little effect on genome

scans, but this is not the majority opinion in the literature. We believe that this is a biologically relevant and very useful negative result.

From the above, the authors should interpret their results much more cautiously. At the very least, the results presented should not be used to interpret genome-scale results, but the authors attempted to do this towards the end of the Discussion. Caution is needed because the simulations did not consider cases in which there were large regions with reduced recombination as well as regions with normal recombination. Examples include comparing X to autosomes or large genomic islands with other regions of the genome. These seem to be a focus of discussion in the literature (e.g., Wolf and Ellegren 2016 Nat Rev Genet; Burri 2017 Evol Lett). But the models considered here do not deal with these cases. Thus repeatedly emphasising that BGS has little effect may be misleading. The authors' suggestions could well be right. But additional simulation results will need to be presented.

Again, we draw recombination maps directly from two well-studied species. Different results are of course possible with another recombination map, although we show that even with no recombination at the scale of 10 cM the F_{ST} is also not greatly affected.

Note also that we are explicit in many places that we are not attempting to model the effects of BGS at a genome-wide scale on F_{ST}. Our main goal is to investigate the role of BGS as a source of variation in F_{ST} among loci, and therefore all of our simulations mimic only the effects of linked selection within 10cM of the focal regions. We have added more qualifiers in this regard throughout the discussion.

Minor Comments:

L176-177: what's the average percentage of sites under selection in this 2 genomes?

The info appears under the section "selection" at lines 217 and 218

L220-222: the effects of BGS is strongest under intermediate level of selection. What's the fraction of sites with $1 < 4N_e s < 10$?

Info added at lines 225-227

L246: it's somewhat strange that the "human" treatment used the genetic map and annotation from the human genome, but retained N = 1, 000 from the "default" treatment. Given that the authors had emphasised their intention to make the simulations realistic, why not use N = 10, 000 in the "human" treatment? In the same vein, could the author consider a case where one of the subpopulations has a smaller N?

The label treatment that we used in the first version for this treatment was unintentionally misleading. This treatment was not intended to mimic all aspects of human biology, but merely indicated that we had used the genetic map from humans. We modified "human" to "human genetic map" to clarify this issue.

What the symbol B refers to seems to be inconsistent. In the Methods (p. 15), B is the reduction in N_e per site. However, in the results, B seems to refer to average B across the focal region.

In the methods, lines 289-290, it is stated: *We computed B for all sites in the focal region and report the average B for the region.*

F_{ST} outlier tests: please clarify whether all sites (i.e., focal and flanking regions) or just sites in the focal region were used.

Just the focal region was used. We have modified this section to clarify this.

Why the SNPs were put into groups of 500? Does this choice have any effect on the result?

The fdist approach requires as set of loci as input, as though results were coming from a sampling of a whole genome. We grouped loci into sets of 500 to mimic a (small) genome of data.

Shouldn't there be 12 false positive rates per treatment, given that 6 α values were listed on p. 16?

We have modified this section. We used a single value of alpha for the results that we present, but also confirmed that this choice of alpha was representative of other choices.

Were all sites in genic regions under selection?

The distribution of selection coefficients for those regions is explained in the methods in the section *Selection.* As per the gamma distribution considered, many sites have an extremely low selection coefficient so that they are essentially neutral. The overall distribution (over genic and non-genic sequences) of selection coefficient is presented in figure S1. As another reviewer argued that our selection pressures were relatively strong, we have now added a treatment in which selection pressures have been reduced.

Did you differentiation synonymous and nonsynonymous changes?

As per the Gamma distribution considered, some mutations are neutral (or quasi neutral) and could be interpreted as synonymous mutations. But we have not explicitly identified nucleotides as synonymous or nonsynonymous.