Reviewer's report

**Title:** SmileFinder: a resampling-based platform to evaluate signatures of selection from genome-wide sets of matching allele frequency data between populations

**Version:** 6  
**Date:** 28 January 2014

**Reviewer:** Emily Jane J McTavish

**Reviewer's report:**

This article describes a python script to identify regions of the genome displaying signatures of selection based on Fst metrics on the paired allele frequencies in two populations. This software complements a previously published ms demonstrating results of these analyses (Oleksyky et al. 2008). This method can be applied to allele frequency data even at low genomic coverage. This is a valuable alternative to methods that rely on very dense genomic coverage to locate potential regions of selective sweeps.

Major Compulsory Revisions:
I had a few general concerns with the manuscript, largely related to lack of clarity.

The text overstates the ability to locate selective sweeps, as opposed to locating regions with changes in Fst consistent with selective sweeps.

In addition, The introduction is not clear, and understanding the functioning of the software relies on a readers close reading of the related paper Oleksyky et al. 2008. This manuscript suffers from insufficiently clear description of what is meant by "variance of Fst". It is my understanding that they mean variance in Fst values among sliding windows, or SNPs within a sliding window, but a clear definition of what is meant here is essential to the manuscript's comprehensibility.

The manuscript states "When allele frequencies from a set of loci inside the region affected by the sweep are compared to exactly the same region in a related population with no history of selection, a rise in variance of allele frequencies between populations is observed." (lines 47-49)

While this may be true if only one population has undergone selection, Fixed differences between populations due to different fixed alleles should result in higher Fst's in those regions. However, this does not necessarily lead t higher variance in Fst in those regions. In fact if each population is fixed for alternative alleles the variance in Fst for that region would be 0. The authors do not address how gene flow would impact their metric.

Overall, I think too much weight is given to defining regions with high variance in Fst as "selective sweeps". If there is gene flow among populations any barrier to recombination (e.g. an inversion) would also result in regions with different fixed
alleles.

In addition, the methodology is not clearly described. More precise language about the calculations and "merging" of percentiles would make the program more comprehensible and easier to review.

Minor Essential Revisions

Manuscript:
line 22: between the ^frequencies of alleles in^ two or more populations?
line 31: "over representation of sweeps" is too specific. Over representation of variation in Fst would be more accurate.
line 63: By operative region do you mean selected SNP?
line 74-75: Diversity between two populations? wording unclear.
line 133: Baseline is re sampled from the data, does not necessarily represent "neutral scenario" if selection has been widespread
line 138: his->this
line 232: should not be bold
line 241: The manner in which the distributions are "merged" is not described in the methods.

Markdown file
line 33: count.py should not be written twice

Figure 1b. is wholly reproduced from Oleksyk et al. 2008.

Discretionary Revisions

1) Although this MS is a software note, it may be appropriate to address the many critiques of Fst as an estimator of selection, and other factors resulting likely to bias Fst
e.g.

2) Is the input format the input or output of any other software? Making it so would be valuable for interoperability.

Markdown file
line 33: suggests using "output.csv". but output is not a comma separated file. Tabs are sometimes used in .csv's but I think that leads to confusion.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare I have no competing interests.