Author’s response to reviews

Title: Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA

Authors:

Nathan Wan (nathan.wan@freenome.com)
David Weinberg (david.weinberg@freenome.com)
Tzu-yu Liu (joyce.liu@freenome.com)
Katherine Niehaus (kniehaus8188@gmail.com)
Eric Ariazi (eric.ariazi@freenome.com)
Daniel Delubac (Ddelubac@gmail.com)
Ajay Kannan (ajay.Kannan@freenome.com)
Brandon White (brandon.white@freenome.com)
Mitch Bailey (mitch.bailey@freenome.com)
Marvin Bertin (marvin.bertin@freenome.com)
Nathan Boley (nathan@ravel.bio)
Derek Bowen (derek.bowen@freenome.com)
James Cregg (james.cregg@freenome.com)
Adam Drake (adam.drake@freenome.com)
Riley Ennis (riley.ennis@freenome.com)
SIgne Fransen (signe.fransen@freenome.com)
Erik Gafni (egafni@gmail.com)
Loren Hansen (loren.hansen@freenome.com)
Yaping Liu (yaping.liu@freenome.com)
Gabriel Otte (gabe@freenome.com)
Dear Dr. Madabhushi,

Thank you for sending the referees’ reports and for your interest in publishing a revised version of our manuscript in BMC Cancer. Please find attached our revised manuscript, “Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA”, as well as point-by-point responses to all referee comments.

Our revisions address all of the concerns of the referees. In addition, in the revised manuscript we improved our analysis of the relationship between tumor fraction and classifier performance. While the manuscript was under review, we were able to further optimize our IchorCNA-based tumor fraction estimate from cfDNA WGS data, reducing the limit of detection from 3% down to ~0.8%. This strengthened our conclusion that TF alone does not explain our classification results and led to minor changes to Figure 3D, Supplementary Figure 2, and Supplementary Figure 3.

We thank you and the reviewers again for your critical review of our manuscript, which is much improved as a result. We hope and trust you will find it suitable for publication in BMC Cancer.

Sincerely,
Response to Reviewers comments on manuscript, “Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA” by Wan et al.

Reviewer reports:

Hiba Abdul-Adheem, Msc/Asis.Prof. (Reviewer 1):

1. I think it is best to use the term ‘readings’ instead of ‘reads’.

In this particular context and for consistency with cited work, we believe that ‘reads’ is a more appropriate term.

2. Some abbreviations have been mentioned and have not been mentioned before, like the following:
   a. Line 183, TF.
   b. Line 299, IU.
   c. Line 254, SVM.
   d. Line 271, ROC.

We have thoroughly reviewed our use of abbreviations in the manuscript and revised accordingly. In particular, we double-checked that TF, IU, and SVM are correctly defined (on lines 106, 161, and 196, respectively) and we added the definition of ROC (“Receiver Operating Characteristic”) on first use near the bottom of page 10.

3. Samples collected from Europe and United States, but the residences did not take in consideration.

In the revised Methods section, we now describe that “samples were collected from institutions and biobanks located in the United States, Germany, and Scotland.

4. Line 217, (operating characteristic curve (AUC)), Wrong abbreviation.

We have clarified in the text that AUC stands for “area under the curve”, where “curve” refers to the receiver operating characteristic curve.
5. Line 234, the term "mappability" is best to be re-checked.

We have confirmed that “mappability” is the appropriate term to use in this context.

6. Line 265, there are no data of standard deviation in table 1.

On line 265, the text refers to Supplementary Figure 1 in which the standard deviations are listed in columns 4 and 6.

7. There is an error page number (1) on the table 3 page.

We thank the reviewer for identifying this error, and we have removed this erroneous page number.

8. In methods, which genes or gene primers was used in PCR?

We have revised the Supplemental Methods section to indicate that we used Illumina-compatible unique dual index primers for the PCR step.

9. Why the authors used plasma samples particularly? Why they did not use serum of whole blood in DNA extraction?

We chose to use plasma rather than serum samples for our study because plasma is less likely to be contaminated with DNA from blood cells (http://mcr.aacrjournals.org/content/14/10/898) and is therefore more suitable for analysis of circulating cell-free DNA.

Shailesh Advani, MD, PhD (Reviewer 2):

Abstract:

Can authors summarize how the stratification impacted AUCs, particularly in reference to age or institution?

While we were unable to add this level of detail to the abstract due to length limitations, we include those details in the main text of the manuscript. Specifically, Table 3 describes AUC when stratified by age (binned-age) and institution (balanced k-batch). While a decrease in AUC was observed when stratifying by institution, a mean AUC of 0.83 was still achieved despite a significantly smaller data set (average of 263.6 samples per fold in training versus 653.6 samples per fold with k-fold or k-batch, as shown in Supplementary Table 1).

Introduction:

a) Can authors provide some information on background of CRC epidemiology?
We appreciate the reviewer’s suggestion and have revised the text accordingly. In the first paragraph of the Introduction, we now describe that “Although the burden of CRC has been decreasing, CRC remains the third leading cause of cancer-related deaths in men and women in the United States.”

b) How are previous studies on ctDNA or cfDNA in colorectal cancer different from your paper? Or how has this method studied previously in CRC, early and late stage?

We address differences in our approach versus that of others (e.g., ctDNA-based mutation detection vs cfDNA WGS) in the third paragraph of the Background section and second paragraph of the Discussion section of the revised manuscript.

Methods:

a) None

Results:

a) Well written, no comments. May be dividing results into separate sections will be helpful

We appreciate the reviewer’s suggestion and have added 3 sub-sections to our Results. We now have the following 2 sub-headings: Performance of confounder covariates (line 233), Performance by cross-validation procedures (line 246), Detailed comparison of performance by clinical parameters (line 270).

Discussion:

a) Can other confounders affect your results? Synchronous vs metachronous, tumor mutation profiles?

While other confounders may exist, we focused our methods on the known or suspected confounders for which we have the relevant metadata; in this case, we focused on age, sequencing batch, and institution. Given that synchronous vs. metachronous labels and tumor mutation profiles are only relevant to cancer patients but not the non-cancer controls, we believe this should not confound the cancer vs. non-cancer classification results reported in our manuscript.