Early-stage colorectal cancer detection using artificial intelligence and whole-genome sequencing of cell-free DNA

Katherine Niehaus, Nathan Wan, David Weinberg, Brandon White, Ajay Kannan, Erik Gafni, Tzu-Yu Liu, Imran S. Haque, Girish Putcha
Freenome, Inc., South San Francisco, CA

BACKGROUND

• Despite population screening programs and availability of several stool-based, non-invasive screening methods, nearly 60% of colorectal cancer (CRC) cases are detected with regional or distant metastases (Siegel et al., 2018).
• Blood-based methods using cell-free DNA (cfDNA) are under development as an alternative to stool-based tests.
• Early-stage detection of cancer using only tumor-derived mutations in cfDNA (i.e., circulating tumor DNA, or ctDNA) is challenging for practical, technical, and biological reasons, such as the small proportion of ctDNA derived from tumor tissue (i.e., tumor fraction, or cfDNA/ctDNA) in early-stage disease (Haque et al., 2017).
• Using machine learning (ML) to discover signatures in cfDNA that may reflect both tumor and non-tumor (e.g., immune) contributions represents a promising direction for the early detection of cancer.
• Confounders, including variation in preanalytical and analytical processes, can affect the performance of ML models, especially in retrospective studies, and must be controlled to limit bias and improve generalizability.

OBJECTIVE

• As part of a program to develop a blood-based screening test for CRC, a machine learning approach for representing and learning associations between cfDNA profiles and cancer status was evaluated in a large cohort of non-cancer controls and early-stage CRC patients (predominantly stages I and II), with a focus on the importance of accounting for known confounding variables.

METHODS

• Sample collection: De-identified plasma samples were received from academic medical centers and commercial biobanks (Table 1).
• Whole-genome sequencing of cfDNA: cfDNA was isolated from 250 µL of plasma and converted into Illumina-compatible libraries, which were sequenced to a minimum of 400 million reads.
• Bioinformatics and feature generation: Reads aligning to annotated protein-coding genes were extracted, and read counts were normalized to Bioinformatics and feature generation: Reads aligning to annotated protein-coding genes were extracted, and read counts were normalized to N=456.
• Classifiers were trained using 85% of the samples for the training set and 15% of the samples for each stage (0.05% bootstrap confidence interval).
• Machine learning: ML models were trained using different cross-validation techniques including k-fold, k-batch, and balanced k-batch (Figures 2, 3).

RESULTS

Table 1. Clinical characteristics and demographics of patients with CRC and non-cancer controls

<table>
<thead>
<tr>
<th>Gender</th>
<th>CRC</th>
<th>Control</th>
<th>Total Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>377 (74%)</td>
<td>279 (61%)</td>
<td>656 (52%)</td>
</tr>
<tr>
<td>Male</td>
<td>41 (82%)</td>
<td>122 (27%)</td>
<td>533 (43%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>9 (18%)</td>
<td>55 (10%)</td>
<td>64 (5%)</td>
</tr>
</tbody>
</table>

Table 2. CRC performance by cross-validation methods in patients aged 50–84

<table>
<thead>
<tr>
<th>Method</th>
<th>Average Training Set Size (%)</th>
<th>Mean AUC (95% CI)</th>
<th>Mean Sensitivity at 85% Specifity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-fold</td>
<td>1128 (85%)</td>
<td>0.89 (0.87–0.91)</td>
<td>82% (75–85%)</td>
</tr>
<tr>
<td>k-batch</td>
<td>1128 (85%)</td>
<td>0.89 (0.87–0.91)</td>
<td>80% (75–85%)</td>
</tr>
<tr>
<td>balanced k-batch</td>
<td>502 (85%)</td>
<td>0.86 (0.83–0.89)</td>
<td>76% (68–81%)</td>
</tr>
</tbody>
</table>

• Early-stage colorectal cancer detection using artificial intelligence and whole-genome sequencing of cell-free DNA

Figure 1. Methods from sample processing to results

Figure 2. Model training and cross-validation (CV) procedures

Figure 3. Training schemas for k-fold, k-batch, and balanced k-batch

Figure 4. Sensitivity by CRC stage in patients aged 50–84

Figure 5. Sensitivity by tumor fraction in patients aged 50–84

Figure 6. Non-linear relationship between total number of samples in training set and AUC in test set

CONCLUSIONS

• A prototype blood-based CRC screening test using cfDNA and machine learning achieved high sensitivity and specificity in a predominantly early-stage CRC cohort (stages I and II).
• Classifier performance suggests contributions from both tumor and non-tumor (e.g., immune) derived signals.
• Assessing genome-wide cfDNA profiles at moderate depth of coverage enables the use of low-volume plasma samples.
• Cross-validation methods highlighted the importance of performing similar confounder analyses for retrospective (and prospective) studies.
• Prospective validation of a similar machine learning method using cfDNA is underway (NCT03289906), along with research evaluating the potential of a multi-cancer approach that integrates other cell-free, blood-based analytes (e.g., proteins) to improve performance.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Dr. Andrew Godwin and the Biospecimen Repository Core Facility staff (funded in part by the National Cancer Institute Cancer Center Support Grant [P30 CA040583] and NCI Research Support Grant SC1CA196464), Genentech Inc., Spectranet Inc., and Institute for this research by providing de-identified plasma samples. We also thank Signe Fransen for her extensive suggestions, feedback, and editorial support.

REFERENCES

Adalsteinsson et al. 2017; Nature Communications 8 (8): 3382
Haque et al. 2017; bioRxiv https://doi.org/10.1101/237758
Perlz et al. 2018; bioRxiv https://doi.org/10.1101/332025
Siegel et al. 2018; Cancer Journal for Clinicians 68 (7): 68–79

Presented at American College of Gastroenterology (ACG) 2018 Annual Scientific Meeting, October 8–10, 2018, Philadelphia, PA, USA