Bioinformatics, 35(19), 2019, 3812–3814
doi: 10.1093/bioinformatics/btz123
Advance Access Publication Date: 2 March 2019
Applications Note

OXFORD

Genome analysis

MetaMarker: a pipeline for *de novo* discovery of novel metagenomic biomarkers

Mohamad Koohi-Moghadam[1,2,3], Mitesh J. Borad[4], Nhan L. Tran[5], Kristin R. Swanson[6], Lisa A. Boardman[7], Hongzhe Sun[2,*] and Junwen Wang[1,8,*]

[1]Department of Health Sciences Research and Center for Individualized Medicine, Mayo Clinic, Scottsdale, AZ 85259, USA, [2]Department of Chemistry, The University of Hong Kong, Hong Kong SAR, China, [3]Center for Genomic Sciences, The University of Hong Kong, Hong Kong SAR, China, [4]Department of Hematology, [5]Department of Cancer Biology, [6]Department of Neurologic Surgery, Mayo Clinic, Scottsdale, AZ 85259, USA, [7]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA and [8]College of Health Solutions, Arizona State University, Scottsdale, AZ 85259, USA

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on November 7, 2018; revised on January 24, 2019; editorial decision on February 7, 2019; accepted on February 27, 2019

Abstract

Summary: We present MetaMarker, a pipeline for discovering metagenomic biomarkers from whole-metagenome sequencing samples. Different from existing methods, MetaMarker is based on a *de novo* approach that does not require mapping raw reads to a reference database. We applied MetaMarker on whole-metagenome sequencing of colorectal cancer (CRC) stool samples from France to discover CRC specific metagenomic biomarkers. We showed robustness of the discovered biomarkers by validating in independent samples from Hong Kong, Austria, Germany and Denmark. We further demonstrated these biomarkers could be used to build a machine learning classifier for CRC prediction.

Availability and implementation: MetaMarker is freely available at https://bitbucket.org/mkoohim/metamarker under GPLv3 license.

Contact: wang.junwen@mayo.edu or hsun@hku.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Whole-metagenome sequencing (WMS) is a novel approach to study microbial communities. Using WMS, scientists have found that the human microbiome has a tight relationship with different diseases, such as colon cancer, bacterial vaginosis, diabetes and Crohn's disease (Cho and Blaser, 2012). Differences in the bacterial patterns (composition and abundance) between healthy and diseased individuals can be considered as novel biomarkers for disease initiation and prognosis. In searching for metagenomic biomarkers, researchers have focused on finding a family of strains, genes, metabolites or pathways that can robustly distinguish two or more microbial communities. For example, *Campylobacter jejuni* in human intestine were shown to be indicative of immunoproliferative small intestinal disease (Lecuit *et al.*, 2004), moreover, the level of intestinal bacterial lipopolysaccharide was shown to be positively associated with colorectal cancer (CRC) (Schuerer-Maly *et al.*, 1994).

Although WMS has been used to discover new metagenomic biomarkers, current computational methods discard unmapped reads, and consequently bias the results. These methods first map raw reads to a reference database to generate an operational taxonomic unit (OTU) table as a common pre-processing step. This step excludes unknown bacterial sequences that might not yet be in reference databases, but could be potentially important for disease. To address this issue, we developed MetaMarker, a *de novo* approach for discovering metagenomic biomarkers from WMS without

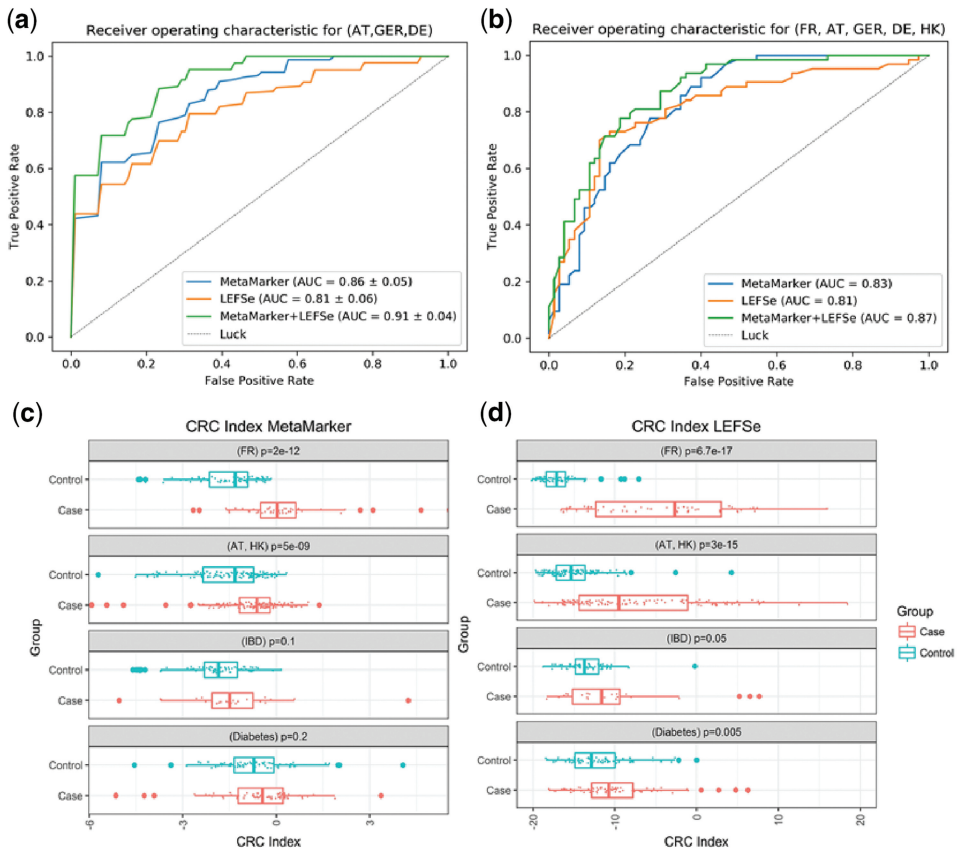

Fig. 1. (**a**) The 10-fold cross validation result of random forest model using an independent population (Austria, Germany and Denmark), (**b**) a test-train model using random forest model on a multiracial population (70% of samples to train and 30% for test), (**c**) using Mann–Whitney test to compare CRC index of MetaMarker in four different populations. (**d**) Using Mann–Whitney test to compare CRC index of LEFSe in four different populations

needing to use an OTU table. Biomarkers found by MetaMarker are conserved sequence fragments that are shared by a subgroup of bacteria. These fragments usually have specific biological functions and play important roles in microbial communities, but they might not yet be collected in any reference database (Supplementary Fig. S1).

2 Methods and implementation

In MetaMarker we used a two-step approach to overcome memory limitation of processing large number of reads in WMS samples. In the first step, we discovered short-markers by profiling normalized abundance of *k*-mers in case and control groups. Here, the *k*-mers which are statistically different in case and control have been assembled to generate the short-markers. In the second step, we used the short-markers to generate long-markers. We finally cleaned and ranked the long-markers to generate the list of biomarkers (Supplementary Fig. S2). MetaMarker takes FASTA format of WMS samples as input files and generates two files as final biomarkers (case-enriched and control-enriched). MetaMarker extracts biomarkers from the population from France using almost constant memory (10 GB) in a reasonable time (Supplementary Table S2). The detailed pipeline description is available in Supplementary Methods.

3 Results

We applied MetaMarker on WMS CRC stool samples from France (Zeller *et al.*, 2014) to extract CRC specific biomarkers. We selected the top 70 biomarkers of MetaMarker for further analysis. These biomarkers ranged from 279 to 3175 bp (median 1101 bp) (see Supplementary Results). We used Hong Kong and Austria cohorts as the independent population to evaluate the discovered biomarkers. Of the top 70 biomarkers for cohort from France, 29 were also significant (P-value <0.05) in both the Austria and Hong Kong populations (Supplementary Figs S7 and S9), and 53 were replicated in at least one of these two independent cohorts (Supplementary Figs S8 and S10).

We then built a classifier using the discovered biomarkers on the independent population from Austria, Germany and Denmark (84 cases and 134 controls). The 10-fold cross validation result showed that these biomarkers can distinguish CRC samples from healthy individuals with AUC = 0.86 (Fig. 1a). We compared the result of our pipeline with LEFSe (Segata *et al.*, 2011) which works based on OTU table. We generated the top 70 biomarkers of France population using LEFSe. Biomarkers from LEFSe ranged from 150 to 2115 bp (median 801). The 10-fold cross validation result showed, LEFSe has AUC = 0.81 on this population (Fig. 1a). Moreover, as top 70 biomarkers of MetaMarker and LEFSe do not have any global overlap, we merged them to build a new dataset. The result showed AUC = 0.91 using these 140 biomarkers (Fig. 1a, see Supplementary Results).

We built a model based on test-train sets and included samples from Hong Kong (Yu *et al.*, 2017) (74 cases and 54 controls). We then performed a stratified sampling and selected 30% of the samples as unseen testing data (138 samples). We used the remaining

70% of the samples (322 samples) for training. The result showed AUC = 0.83 for MetaMarker and AUC = 0.81 for LEFSe. The merged set of the biomarker showed AUC = 0.87 on this multi ethnic dataset (Fig. 1b, see Supplementary Results).

We finally used a CRC index score (Yu *et al.*, 2017) to find the specificity of the biomarker (see Supplementary Results). Here, we used two other populations to compare with the CRC cohorts. One cohort from China (Qin *et al.*, 2012) included 71 WMS samples from patients who had diabetes and 74 control samples from healthy individuals. The second cohort from Spain (Qin *et al.*, 2010) had 25 WMS samples from patients with inflammatory bowel disease (IBD). We used Wilcoxon–Mann–Whitney test to compare CRC index. The results showed the top 70 biomarkers from MetaMarker could specifically differentiate CRC in samples from the French (*P*-value $<2e-12$), Austria–Hong Kong cohorts samples (*P*-value $<5e-9$), but failed to distinguish the IBD and diabetes samples (*P*-value >0.1 and *P*-value >0.2, respectively; Fig. 1c). However, LEFSe was better able to detect CRC cases from the French and Austria–Hong Kong cohorts (*P*-value $<6.7e-17$ and *P*-value $<6.7e-17$, respectively) than MetaMarker, but LEFSe misclassified samples from IBD and diabetes patients as CRC (*P*-value <0.05 and *P*-value <0.005, respectively; Fig. 1d). Thus, MetaMarker has a higher specificity for classifying CRC than LEFSe.

4 Conclusions

To the best of our knowledge, MetaMarker is the first *de novo* approach to discover biomarkers from WMS samples with large number of reads. Discovering these biomarkers may provide opportunities to develop new approaches to combat disease-causing bacteria or to improve human health by promoting disease-protective bacteria. Our results showed that MetaMarker has better performance than LEFSe in distinguishing CRC samples in a multiracial population. This may be rooted in the nature of biomarkers which are discovered by MetaMarker. These biomarkers are conserved sequence fragments which can be shared by different families of bacteria while LEFSe defines biomarkers for each family of bacteria separately. Upon pooling biomarkers from MetaMarker and LEFSe, our machine learning model improved classification performance beyond that of either model. This indicates the two pipelines complement each other and it is recommended to use both.

Funding

The work was supported by the Research Grants Council, Hong Kong SAR [17121414 M, 17307017P] and start-up funds from the Mayo Clinic (Mayo Clinic Arizona and Center for Individualized Medicine); and The National Institutes of Health [R01CA170357, U01CA220378, 1U54CA210180, P30CA015083].

Conflict of Interest: none declared.

References

Cho,I. and Blaser,M.J. (2012) The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.*, **13**, 260–270.

Lecuit,M. *et al.* (2004) Immunoproliferative small intestinal disease associated with *Campylobacter jejuni*. *N. Engl. J. Med.*, **350**, 239–248.

Qin,J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

Qin,J. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.

Schuerer-Maly,C. *et al.* (1994) Colonic epithelial cell lines as a source of interleukin-8: stimulation by inflammatory cytokines and bacterial lipopolysaccharide. *Immunology*, **81**, 85–91.

Segata,N. *et al.* (2011) Metagenomic biomarker discovery and explanation. *Genome Biol.*, **12**, R60.

Yu,J. *et al.* (2017) Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*, **66**, 70–78.

Zeller,G. *et al.* (2014) Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.*, **10**, 766.