Opportunistic detection of *Fusobacterium nucleatum* as a marker for the early gut microbial dysbiosis

Ji-Won Huh
Pohang University of Science and Technology
Tae-Young Roh (tyroh@postech.edu)
Pohang University of Science and Technology https://orcid.org/0000-0001-5833-0844

Research

Keywords: Fusobacterium nucleatum , inflammatory bowel diseases (IBD), colorectal cancer (CRC), opportunistic detection, microbial experience, Integrative Human Microbiome Project (iHMP)

Posted Date: December 18th, 2019

DOI: https://doi.org/10.21203/rs.2.19123/v1

Version of Record: A version of this preprint was published at BMC Microbiology on July 13th, 2020. See the published version at https://doi.org/10.1186/s12866-020-01887-4.

Abstract

Background Gastrointestinal microbiome has a great impact on human health and diseases. Fusobacterium nucleatum , an oral resident gram-negative microorganism, has been reported to differentially prevail in colorectal cancer (CRC). However, the effect of its enrichment was not clearly explored in terms of microbial homeostasis and stability at the early stage of disease development.

Result Our analysis on longitudinal metagenomic data generated by the Integrative Human Microbiome Project (iHMP) showed that F. nucleatum significantly proliferated in IBD subjects with decrease in microbial diversity. Using non-parametric LDA effect size (LEfSe) algorithm, 12 IBD- and 14 non-IBD-specific marker species were identified. IBD-specific marker species were found more in the F. nucleatum -experienced subjects than in not-experienced ones. In addition, F. nucleatum experience severely abrogated intra-personal stability of microbiome in IBD patients and induced highly variable gut microenvironment. From the longitudinal comparison between prior and posterior distributions to F. nucleatum detection, 41 species could be regarded as indicative “classifiers” of dysbiotic state. Multiple logistic regression models were built to predict the probability experiencing F. nucleatum and showed that the high probability at the points following F. nucleatum observation was significantly associated with microbial dysbiosis. Finally, clustering all detected microbes based on the longitudinal distribution revealed that marker species for IBD and non-IBD conditions as well as pre-documented signature biomarkers of CRC were well distinguishable and could be utilized for explaining gut symbiosis and dysbiosis.

Conclusion F. nucleatum opportunistically appeared under early gut dysbiosis, and a group of discriminative marker species associated with F. nucleatum , screened from the longitudinal analysis of microbial changes, was successfully applied to predict early microbiota alterations in both IBD and non-IBD conditions. Our prediction model and development of microbial biomarkers for estimating gut dysbiosis should provide a novel aspect of microbial homeostasis/dynamics and useful information on non-invasive biomarker screening.

Background

The microbial communities in the gastrointestinal tract play pivotal roles in maintaining many biological functions such as food digestion, metabolism, and immunological regulations as well as developing diseases like ulcers, bowel perforation, inflammatory bowel diseases, irritable bowel syndrome, other inflammatory conditions, metabolic syndromes, and even colorectal cancer.

F. nucleatum was initially identified as a non-motile obligate anaerobe commonly residing on the tooth surface of healthy individuals and bridging bacterial species to form dental plaque biofilm [1, 2]. Many researches have indicated that *F. nucleatum* is also found in other organs and associated with several disorders such as oral inflammation, adverse pregnancy outcomes, IBD, Lemierre’s syndrome, cardiovascular diseases, atherosclerosis, Alzheimer’s disease, and cancer [3–8]. IBD refers to as chronic

conditions describing a group of inflammatory disorders in intestines. Patients with IBD tend to show a high level of *F. nucleatum* in the colon and are at significantly high risk of CRC. It has been demonstrated that *F. nucleatum* is related with and promotes the growth of CRC [9–17].

CRC is the fourth most incident cancer in the world. The rates of CRC incidence and mortality are still rising in developing countries and in relatively young people in the United State. [18, 19]. Chronic inflammation at large intestine is a significant risk factor of CRC [30, 31]. CRC development is increased in IBD people by six times compared with control group and 1 out of 7 IBD patients died by CRC [32]. Furthermore, the occurrence of CRC after negative findings from a recent colonoscopy is three times higher in IBD patients than in non-IBD controls, indicating rapid carcinogenesis with chronic gut inflammation [33]. For early detection of CRC, endoscopic surveillance is required and accompanied by discomfort due to invasiveness. Thus, there is a realistic need for development of non-invasive biomarkers for the diagnosis [20]. Although *F. nucleatum* is detected significantly more in CRC tissues, the efficacy of fecal *F. nucleatum* as a potential non-invasive biomarker was not consistent because it is rarely detected [21–28]. However, corroborating evidences that *F. nucleatum* is a promising biomarker for CRC, have been accumulated and more effort have been made to understand its role in the carcinogenesis.

The Integrative Human Microbiome Project (iHMP) released extensive longitudinal datasets of disease-specific cohorts to understand the interaction between the microbiome and host using multi-omics technologies. Among them, there are whole-metagenome shotgun sequencing data from more than one hundred fecal samples of IBD and non-IBD participants over one-year collection [29]. These data seem to provide a comprehensive profiling on overall microbiome with IBD but the establishment of cancer-associated microbiome in IBD has not been investigated [34].

Here we explored whether *F. nucleatum* and its associated pathobionts might become prevalent under dysbiotic environment by analyzing the longitudinal metagenomic data and predicted if the occurrence of *F. nucleatum* could play a function as the indicator reflecting disease condition.

Results And Discussion

Metagenomic profiling of IBD or non-IBD participants

As shown in Fig. 1, the overall analysis includes filtering, profiling, longitudinal dissection, biomarker screening, modeling, and microbial dynamics. The data used in this study was the fecal metagenome dataset downloaded from the Inflammatory Bowel Disease Multi-'Omics Database (IBDMD) of iHMP, which were longitudinally generated from 130 participants. The data quality was tested and only significant samples were considered for the analysis (Additional file 9).

Microbial taxonomy was assessed at the species level using MetaPhlAn2, and the quality of the compositional data were controlled according to the specific conditions mentioned in the Method (Additional file 10) [35]. The number of filtered data was 1526 samples from 106 participants, and the

basic information of the participants such as sex, age, and collection days were comparable between IBD patients and non-IBD subjects (Additional file 11: Table S1). The overall distribution did not show any specific tendency to sex, IBD-activity, subject, and data generation sites (Additional file 1: Figure S1).

Consistent with the previous reports, two major phyla in human gut, *Firmicutes* and *Bacteroidetes* displayed a reciprocal proportion in principal coordinate analysis (PCoA) (Fig. 2a) [36]. The metagenome samples from IBD and non-IBD subjects were clearly distinguished. Samples from non-IBD subjects mainly localized in a left-lower quadrant and samples from IBD patients were more widely distributed along PC1 axis (probability value of IBD vs. non-IBD, $P_{\text{IBD-Non (PC1)}} < 2.2e{-}16$, Fig. 2b). The representative subtypes of IBD, ulcerative colitis (UC) and Crohn's diseases (CD), were not significantly discriminated by PC1 and PC2 axes (probability values of UC vs. CD, $P_{\text{UC-CD (PC1)}} = 0.1726$, $P_{\text{UC-CD (PC2)}} = 0.0988$), implying that the two idiopathic inflammatory disorders shares similar microbial community (Fig. 2b). Overall microbiome seemed to be largely distinct among subjects and stable over time (Fig. 2c). As grouped by K-means clustering, most of non-IBD samples belonged to cluster C3, suggesting that microbiome from non-IBD subjects should be relatively convergent relative to those from UC or CD (Odd Ratio, $OR_{\text{nonIBD-C3}} = 4.42$, $OR_{\text{UC-C3}} = 2.30$, $OR_{\text{CD-C2}} = 2.15$).

Gut microbiota homeostasis is maintained under normal condition but stress conditions induce a decrease of microbial diversity, leading to dysbiosis [37]. Multiple measurements of alpha diversity indices like Shannon diversity, Pielou's evenness, and richness (the number of detected species per sample) were lower in samples with IBD than in those without IBD. There were no significant differences in alpha diversity indices between CD and UC. Interestingly, the active state of IBD had lower Shannon diversity and richness (Fig. 2d, additional file 2: Figure S2a and b).

Additionally, the fraction of human sequence reads mixed with gut metagenomic data was higher in IBD and much higher in active state of IBD, which means that a leakage of human genome into gut lumen might reflect the severity of disorders (Fig. 2e). The fraction of human sequence reads was also positively correlated with diseases severity scores such as simple clinical colitis activity index (SCCAI) for UC and Harvey-Bradshaw index (HBI) for CD (Additional file 2: Figure S2c, d).

Detection of *F. nucleatum* and its longitudinal dissection

F. nucleatum is not frequently found in gut microbiome but its detection is important as an indicator of gut disorders. *F. nucleatum* occurred 41 times with marginal preference to chronic inflammation (OR = 1.79, Fisher's one-sided $P_{\text{detect}} = 0.1062$) (Fig. 3a). Considering relative abundance upon detection, *F. nucleatum* was favorably detected in IBD samples (Wilcoxon test $P_{\text{detect}} = 0.02891$) (Fig. 3b).

The low detection frequency of *F. nucleatum* in healthy state is not a favorable property for early diagnosis. Thus we tested whether the detection frequency and abundance are related with detection consistency using 44 replicated samples. Highly abundant species were mostly captured in both duplicated sample pair but less abundant ones were not. About one-fourth of total species appeared only

in one sample of duplicate pairs. *F. nucleatum* was a relatively rare microbe observed only 4 times in three replicates of total 44 test samples. The abundance and the number of detection showed a positive correlation, as expected. The recovery rate, which is the ratio of the number of pairs having a certain microbe in both duplicated samples vs. total number of sample pairs with that microbe, was higher for abundant species. Recovery rate of *F. nucleatum* was only 33.3%, supporting inconsistent fecal detection (Additional file 3: Figure S3).

To overcome the limitation of snapshot-based approach, the samples collected from each subject over one year were organized according to their chronological point relative to *F. nucleatum* (Fig. 3c). At first, subjects were classified into *F. nucleatum*-experienced or –innocent (non-experienced) groups, and the samples from *F. nucleatum*-experienced subjects were sub-divided into prior or posterior distribution to the detection point of *F. nucleatum* as well as proximal or distal distribution toward *F. nucleatum* as illustrated. The samples from *F. nucleatum*-experienced subjects were displayed in PCoA plot and showed highly dispersed distribution (Fig. 3d, g). Experiencing *F. nucleatum* seemed to lead to reduced Shannon diversity and Pielou's evenness. Particularly, the samples either proximal or posterior to *F. nucleatum* detection exhibited reduced alpha diversity and increased human read fraction (Fig. 3e, h, i, additional file 2: Figure S2e, f). Longitudinal tracking of *F. nucleatum*-experienced subjects revealed that the microbial diversity was reduced in non-IBD subjects as well (Fig. 3f). These results implies that *F. nucleatum* might appear under the perturbed gut microenvironment with low microbial diversity.

Identification of biomarkers in IBD/non-IBD and their correlation with *F. nucleatum.*

In order to clarify whether *F. nucleatum* was truly associated with inflammatory environment, we tried to screen biomarkers of IBD and non-IBD conditions. Using a non-parametric Linear discriminant analysis Effect Size (LEfSe) algorithm which emphasizes not only the differential abundance of features among the classes but also the biological consistency within the same class, 12 IBD and 14 non-IBD biomarkers were selected at species level. As expected, these markers were specifically detected in either IBD or non-IBD samples (Fig. 4a, b, additional file 12). In *F. nucleatum*-experienced subjects, the ratio of IBD markers to non-IBD markers was significantly increased, and non-IBD subjects who have experienced *F. nucleatum* held increased number of IBD biomarkers as well (Fig. 4c). The prevalence of IBD markers over non-IBD markers was also distinct in *F. nucleatum*-posterior samples. The subjects who have experienced *F. nucleatum* at least once over one year showed more IBD biomarkers in non-IBD conditions (the median of IBD marker $\#_{F.\ n.\text{-unexp}} = 6.54$, the median of IBD marker $\#_{F.\ n.\text{–exp}} = 9.23$, $P < 2.2e{-}16$) (Fig. 4d, e). These results suggested that experience of *F. nucleatum* should be tightly connected with IBD. The association of biomarker species with *F. nucleatum* was assessed by calculating Spearman's correlation coefficients. All 14 non-IBD biomarkers were negatively correlated with *F. nucleatum,* having very significant enrichment p-values, and IBD biomarkers showed mostly positive correlation with some exceptions (Fig. 4f). Collectively, the absolute correlation coefficient of a certain microbe with *F. nucleatum* had strong relationship with its enrichment p-values in either IBD or non-IBD conditions (□□ = 0.33, $P = 3.8e{-}15$), (Fig.

4f). When the longitudinal abundance of the biomarker species was examined, two representative non-IBD marker species, *Alistipes shahii* and *Alistipes putridinis,* showed the decreasing patterns of abundance along the temporal axis centered at *F. nucleatum*-detection. In contrast, IBD markers like *Clostridium symbiosum* and *Clostridium bolteae* had opposite abundance pattern, low at prior and high at posterior to *F. nucleatum*-detection along the temporal axis (Fig. 4g). The abundance of these four biomarkers was changed significantly before and after *F. nucleatum* experience only in IBD condition, which means that the abundance of the cardinal microbes should be perturbed by inflammatory circumstances. In addition to these biomarkers, two IBD markers, *Flavonifractor plautii* and a unclassified species in *Oscillibater* genus, and three non-IBD markers, *Alistipes finegoldii, Roseburia hominis, Roseburia inulinivorans,* exhibited similar patterns of abundance over time (Additional file 4: Figure S4).

Microbial destabilization after *F. nucleatum* detection

Human gut microbiota is a sort of indicators of human health and understanding of their behavior is important for diagnosis and prevention of various diseases. The microbial imbalance, called dysbiosis, is believed to cause or be associated with several metabolic and inflammatory diseases [38, 39]. Thus, we examined intra- and inter-individual perturbation of microbiome according to experience of *F. nucleatum* and temporal distribution relative to its detection.

Intra-individual dissimilarity of microbiome was measured by calculating pairwise Bray-Curtis distance after random sampling in a given participant (Fig. 5a). Consistent with the previous literature, IBD subjects regardless of *F. nucleatum* experience, showed much longer microbial distance than non-IBD subjects at any given time intervals, supporting that IBD is related with microbial destabilization [34]. More importantly, IBD patients who have experienced *F. nucleatum* had more unstable microbiomes than *F. nucleatum*-unexperienced group (Fig. 5b). For temporal analysis of microbial stability, the time when *F. nucleatum* was detected was set as the initial point of the pairwise distance (Fig. 5c). *F. nucleatum*-experienced subjects showed significantly higher dissimilarity in samples ahead of *F. nucleatum*-detection than in those behind the detection, which was not shown in *F. nucleatum*-innocent subjects with random initial points ($P_{|x|<20w}$ = 3.5e−05 in *F. nucleatum*-experienced; $P_{|x|<20w}$ = 0.1905 in *F. nucleatum*-innocent; Fig. 5d). When examining 20 *F. nucleatum*-experienced subjects, all samples from non-IBD subjects were localized in a confined area of PCoA plot but certain IBD subjects (for example, C3009, H4015, M2034, and P6009) showed dramatic shift of microbiome prior to *F. nucleatum* detection (Additional file 5: Figure 56).

Two different participants in the same class were randomly selected to compare inter-individual microbial distance (Fig. 5e). The dissimilarity between IBD patients was higher than non-IBD subjects (d_{IBD} = 0.5696, $d_{non\text{-}IBD}$ = 0.5000), and that of *F. nucleatum*-experienced subjects also higher than non-experienced ones (d_{exp} = 0.5927, $d_{non\text{-}exp}$ = 0.5401). The microbial distance on the temporal distribution was higher at posterior than prior to *F. nucleatum*-detection ($d_{posterior}$ = 0.5816, d_{prior} = 0.5372) (Fig. 5f-h).When *F. nucleatum*-detected samples were compared with another sample from other *F. nucleatum*-

experienced subjects, the inter-personal microbial distance was gradually elevated until 20 weeks after *F. nucleatum* detection (Fig. 5i, j).

Collectively, these results suggested that highly variable microbiome might be pre-established in *F. nucleatum*-colonizing environment, and potentiate dysbiosis with chronic inflammation. On the other hand, a convergent microbiome before *F. nucleatum* detection is transformed into unstable and divergent one with *F. nucleatum* detection, possibly leading to the formation of pathogenic microbiome.

Identification of classifier microbes prior and posterior to *F. nucleatum* detection

To figure out representative microbes in samples before and after *F. nucleatum*-detection, 317 samples from *F. nucleatum*-experienced subjects were partitioned by iterating createDataPartition function in caret R package 1,000 times. A total of 258 microbes, identified at least 5 times across *F. nucleatum*-experienced subjects, were tested for their discriminative ability for samples prior or posterior to *F. nucleatum*-detection. The values of area under the Receiver Operating Character (ROC) curve (AUC) was calculated, and 41 significant species with average AUC value above 0.5 in multiple logistic regression models (FDR < 0.001) were regarded as classifiers (Fig. 6a). Among the classifier microbes, 26 species were enriched in *F. nucleatum*-posterior samples and 15 species in *F. nucleatum*-prior samples (Fig. 6b, additional file 13: Table S2). The posterior-enriched classifiers, including 3 IBD marker species, were found more in IBD samples, and the prior-enriched classifier with 4 non-IBD marker species were preferentially observed in non-IBD samples (Fig. 6b, c).

A recent fecal metagenome analysis suggested 29 core signature bacteria enriched in CRC metagenomes including three *F. nucleatum* strains [40]. Among them, 18 signature species were detected in our data, and most of them (13 out of 17 signatures except *F. nucleatum)* were positively correlated with *F. nucleatum* (additional file 14: Table S3). Of note, five CRC signature species were ranked as potent classifiers enriched in *F. nucleatum*-posterior samples (P_{CRC} = 0.0164) (Fig. 6b). The CRC-enriched classifiers included three *Clostridium* species, *C. symbiosum, C. bolteae, C. clostridioforme, F. nucleatum,* and *Peptostreptococcus stomatis (*$AUC_{C.\ sym.}$ = 0.6574, $AUC_{C.\ bolt.}$ = 0.6427, $AUC_{C.\ clostri.}$ = 0.6102, $AUC_{F.\ nuc.}$ = 0.6043, $AUC_{P.\ sto.}$ = 0.5406). Especially, *C. symbiosum,* the most significant *F. nucleatum*-posterior classifier in our study, was proposed as a potent fecal biomarker of CRC even superior to *F. nucleatum* [41].

When considering discriminative significance of all microbes found in *F. nucleatum*-experienced subjects, all 11 CRC biomarkers, detected more than 5 times in *F. nucleatum*-experienced subjects, could discriminate *F. nucleatum*-posterior from *F. nucleatum*-prior samples (P_{CRC} = 0.00091). Additionally, a majority of IBD and non-IBD markers showed a discriminative power (P_{IBD} = 0.02285, $P_{non\text{-}IBD}$ = 0.0732) (Fig. 6d).

Among prior-enriched classifier species, *Dorea longicatena* with the highest AUC (AUC = 0.7224) was recently proposed as potential probiotic of metabolic disorder and also reported to be over-represented in remissive CD patients after ileocolonic resection when compared to recurrent cases [42, 43]. *Coprococcus comes* (AUC = 0.7143)was reported to show down-regulation in CRC patients, and three *Roseburia* speciesincluding *R. hominis, R. inulinivorans,* and *R. intestinalis (*$AUC_{R.hom.}$ = 0.6594, $AUC_{R.inul.}$ = 0.6576, $AUC_{R.intest.}$ = 0.6140)*,* were well-documented to shape beneficial gut microflora by fermenting dietary polysaccharides [44–47]. Most biomarkers found in this analysis exhibited meaningful p-values and were differentially enriched in either *F. nucleatum*-posterior or -prior samples, supporting that *F. nucleatum*-oriented approach is advantageous to the identification of effective biomarkers (Fig. 6e).

Estimation of *F. nucleatum* experience and dysbiosis level in *F. nucleatum*-innocent subjects

Among 41 classifiers, top 13 classifiers except *F. nucleatum,* satisfying average AUC > 0.6 and FDR < 1e–07, were used to construct a prediction model for the estimation of the probability experiencing *F. nucleatum* as illustrated (Fig. 7a). Multiple general linear modeling (GLM) was tested with 100 randomly partitioned training datasets. The 10^{th} model was chosen as the best by considering AUC, Akaike information criterion (AIC), accuracy, sensitivity, precision, and specificity (Fig. 7b, additional file 6: Figure S6). The 10 species used for building the 10^{th} model were *Dorea longicatena, Coprococcus comes, Lachnospiraceae* bacterium 3_1_46FAA*, Clostridium symbiosum, Roseburia hominis, Roseburia inulinivorans, Alistipes shahii, Bacteroides stercoris, Clostridium bolteae,* and *Veillonella parvula* in descending order of mean AUC (Additional file 13: Table S2). When applying this model to *F. nucleatum*-experienced subjects for validation, the probability experiencing *F. nucleatum,* so called “posterior probability”, was gradually increased and reached decision threshold to 0.5 before actual *F. nucleatum*-detected points, which means that this model can predict the exact point of *F. nucleatum* detection (Fig. 7c). Interestingly, samples with posterior probability above 0.5 from *F. nucleatum*-innocent subjects were assigned as putative *F. nucleatum*-posterior group, where we observed clear manifestations of dysbiosis: decreased alpha-diversity, increased biomarkers of IBD and CRC, and decreased non-IBD biomarkers (Fig. 7d). The posterior probability was significantly correlated with both Shannon diversity and the ratio of IBD to non-IBD markers (Spearman correlation, $\rho_{shannon} = -0.29$, $\rho_{ratio} = 0.53$; Fig. 7e). As shown in PCoA plots of top 12 “dynamic” subjects selected according to the standard deviation of posterior probability, several IBD patients including E5009, H4015, H4032, H4044, P6009, P6010, and P6025, displayed dramatic microbial shift with increasing posterior probability (Additional file 7: Figure S7). In addition to top 12 dynamic subjects, top 32 subjects corresponding to 70^{th} percentile, also confirmed the negative correlation between posterior probability and Shannon diversity with intra-personal perspective along timeline (Fig. 7f, additional file 8: Figure S8). The samples with low posterior probability were converged in the lower left side but the samples with high probability were scattered, indicating that our prediction model explained microbial variance well (Fig. 7g).

Application of potential biomarkers to the evaluation of fecal microbiome

To see the distribution of individual species, we formulated 5 features associated with *F. nucleatum* and inflammation as follows; 1) microbial abundance correlation with *F. nucleatum,* 2) enrichment in IBD condition, 3) enrichment in *F. nucleatum*-experienced subjects, 4) enrichment in samples posterior to *F. nucleatum* detection, 5) discriminative significance for samples prior or posterior to *F. nucleatum.* The biomarkers of IBD and non-IBD conditions were distinguishable in PCA plot, and the pre–documented CRC marker species were largely distributed at the right side of plot with moderate dispersion (Fig. 8a, b).

To verify the validity of our IBD/non-IBD biomarkers as well as CRC markers in the longitudinal analysis, all microbes detected were grouped by k-mean clustering. Among 9 clusters, cluster 1 harbored most non-IBD biomarkers and cluster 6 had many of both CRC and IBD biomarkers. Cluster 4 had both CRC and non-IBD markers. Cluster 8 and 9 contained IBD markers (Fig. 8c). Notably, the cluster 1 and 6 were positioned in distal area of PCA plot, and the cluster 8 and 9 were localized near cluster 6, which indicated that the microbes belonging to the disease-marker containing cluster share similar correlation and dynamic characteristics (Fig. 8d).

The number of detected microbes along temporal proximity to *F. nucleatum* was decreased in clusters 1, 4, and 7 where non-IBD biomarkers are engaged (Fig. 8e). Clusters 6 and 8, which have both CRC and IBD biomarkers, increased in IBD condition. Interestingly, although the cluster 2 and 3 showed significant decrease in detected microbe number regardless of inflammatory conditions, they did not contain any biomarkers. In accordance with Fig. 4e, dysbiosis-associated biomarkers changed only in IBD condition. Clusters 1, 4, and 7 were negatively associated with the posterior probability, but the clusters 6 and 8 were positively related (Fig. 8f). Furthermore, the clusters 1 and 6 were reciprocally distributed each other in terms of microbial abundance and detection frequency (Fig. 8g).

Together, our work illuminated previously unrecognized knowledge to understand microbial dynamics as dysbiosis is underway by focusing on the opportunistic colonization of *F. nucleatum.* Although further experiments were needed to prove detailed microbial interactions, we anticipated that analyzing chronological alteration of microbiome would greatly improve biomarker screening and diagnosis of microbiota-associated diseases.

Methods

Data curation and taxonomy assignment

A total of 1,638 fecal metagenomic dataset (1,338 HMP data and 300 HMP pilot data), longitudinally collected from 130 participants was downloaded from IBDMD (https://ibdmdb.org/) [29]. (Additional file 15). To keep the validity of longitudinal sampling, the data from participants who provided fecal samples more than 5 times was considered. Technical replicates were not used in this study. After filtering,

metagenomic analysis of 1,560 fecal sample data (243 HMP pilot and 1,317 HMP) from 106 participants was performed at species-level resolution by MetaPhlAn2 software [35]. To improve the resolution of metagenomic data and to reduce outlier-driven statistical distortion, the following three conditions for quality control were examined and finally 1,526 samples were selected for the further analysis.

- Species level explains more than 90% of total microbiome.
- Total bacterial abundance accounts for 70% of whole metagenome.
- Minimum number of bacterial species is greater than 17.

Sample categorization based on diseases activity indices

Simple complex colitis activity index (SCCAI) and Harvey-Bradshaw index (HBI) were available in 413 UC-derived samples and 650 CD-derived samples, respectively (Supplementary data 1). HBI is a simpler version of the Crohn's disease activity index (CDAI), which enables patients to self-diagnose diseases severity. To match two different activity indices, we generated a combined criteria like below to categorize them, considering common guidelines (https://www.igibdscores.it/en/info-hbi.html) [59, 60]. 1) Remission: SCCAI ≤2, and HBI ≤3; 2) Border: 3≤ SCCAI ≤5, and 4≤ HBI ≤7; 3) Active: SCCAI ≥6, and HBI ≥8.

Principal Coordinate Analysis

Microbial abundance data was log10-transformed after adding 1e−05 pseudo-abundance, where 1e−05 is the half of minimum abundance detected across whole samples. Then, integer 5 was added to remove negative values and use Bray-Curtis dissimilarity. Principal coordinates analysis were conducted using vegdist and cmdscale function in R.

To examine whether samples are distinguishable in 2-dimensional PCoA plot by participant categories, we performed analysis of variance (ANOVA) for the particular classes against principal coordinates. Then, samples from IBD/non-IBD subjects, UC/CD patients, or *F. nucleatum*-innocent/experienced subjects were examined whether they were differentially distributed according to their classes.

To conduct principal component analysis (PCA) between microbes, five microbial features in distributional dynamics were used as described below. These features were used for k-mean clustering of microbes as well.

1) P-value for Spearman correlation coefficients with *F. nucleatum*

2) P-value for the differential enrichment in IBD condition

3) P-value for the differential enrichment in samples from *F. nucleatum*-experienced subjects

4) P-value for the differential enrichment in samples after *F. nucleatum* detection

5) P-value for discriminating samples posterior to *F. nucleatum* detection from those prior to *F. nucleatum* detection in 100 random sub-samples.

Because the significances for *F. nucleatum*-posterior enrichment and classifying samples were measured for microbes detected in *F. nucleatum*-experienced subjects at least 5 times, 258 out of 533 total species were plotted in PCA plot.

K-mean clustering

To test whether three origins of samples, non-IBD, UC or CD subjects, were distinguishable by their microbial composition, we conducted k-mean clustering using kmean function in R after adding 1e−05 pseudo-abundance and log10 transformation. Then, we tested whether each condition is over-represented in a particular cluster compared to others using Fisher's exact test, and odd ratios for each cluster were compared.

To classify the microbes based on their distributional features, we clustered 258 species, detected at least 5 times in *F. nucleatum*-experienced subjects, using k-mean clustering. The best number of cluster was determined using NbClust package in R, and cluster number was chosen by vote. Features used for clustering are same as annotated in PCA plot section.

Screening microbial marker species for IBD or non-IBD condition

To identify microbial biomarkers that were differentially enriched in IBD or non-IBD conditions, we applied a web-based linear discriminant analysis effect size (LEfSe) algorithm (http://huttenhower.sph.harvard.edu/galaxy/), which emphasizes not only the differential abundance of features among the classes but also the biological consistency within a same class [61]. Here, by setting IBD subtypes (UC and CD) as a sub-class of IBD, we could obtain common inflammatory biomarkers to capture shared intestinal circumstances by chronic inflammation rather than the specific alteration in UC or CD. Significance thresholds were 0.05 for both Krustal-Wallis test among classes and pairwise Wilcoxon test between sub-classes. LDA score threshold of discriminative microbes was 2.5. Detailed results was included in an additional file 12.

Microbial dissimilarity analysis

Microbial distance was calculated by pairwise Bray-Curtis dissimilarity. For intra-individual dissimilarity of microbiome, one subject was randomly selected for 10,000 times and two samples were chosen from the same subject. For inter-individual dissimilarity test, the subjects were divided into three groups based

on their classification categories such as inflammatory condition, *F. nucleatum* experience, or longitudinal distribution toward *F. nucleatum* observation, and selected each sample from two random subjects. As a control of inter-individual distance, two samples were randomly selected regardless of categories. To examine microbial distance by temporal proximity toward *F. nucleatum*-detected point, the *F. nucleatum*-detected samples was set as the initial point and another random sample was selected from the same subject. Two randomly picked samples from a *F. nucleatum*-innocent subject were served as control. Detailed codes for the analysis are included in a supplementary material, and in GitHub (https://github.com/JW-Huh/F.nucleatum-project).

Logistic regression model

To construct a prediction model for *F. nucleatum* experience, we first screened "classifier" microbes that discriminate between *F. nucleatum*-prior and *F. nucleatum*-posterior samples. After partitioning 317 samples from *F. nucleatum*-experienced subjects 1,000 times using a createDataPartition function in caret R package, 41 classifiers, having average AUC value significantly higher than 0.5 (FDR < 0.001), were selected among 258 microbes which were detected at least 5 times across *F. nucleatum*-experienced subjects. Here, to improve the number of samples, 41 *F. nucleatum*-detected samples were considered as *F. nucleatum*-posterior group. Detailed information for classifier species were included in an additional file 13.

Among 41 classifiers, top–13 potent classifiers except *F. nucleatum* (average AUC > 0.6 & classifying FDR < 1e–07) and inflammatory condition of subjects were selected as model features. To find out the best set of classifiers, we added up the microbes from the top to the 13^{th} in a decreasing order of average AUC, resulting in 13 different feature sets. In a similar way of classifier screening, samples from *F. nucleatum*-experienced subjects were divided into training and test set 100 times using caret R package, and multiple logistic regression models were generated (total 1,300 models; one model/training set with 100 training sets and 13 feature combinations). The best performer was selected by averaging performance ranks of cross-validation AUC with training set, AUC with test set, Akaike information criterion (AIC), and four prediction statistics with decision threshold at 0.5 (accuracy, sensitivity, specificity, precision). The selected model 10 was used for subsequent analysis. Detailed codes for the classifier screening and modeling were presented in GitHub (https://github.com/JW-Huh/F.nucleatum-project).

Conclusions

This study revealed that sporadic observation of *F. nucleatum* in fecal metagenome reflected dysbiotic environment in the gut. Distribution of 12 IBD and 14 non-IBD biomarkers was significantly altered by *F. nucleatum* experience. Among the longitudinal metagenomes, samples posterior to the *F. nucleatum* detection showed high intra- and inter-individual dissimilarity, indicating that occurrence of *F. nucleatum* might serve as a trigger for perturbed stability and increased divergence of microbiome. The 41 classifier species related with prediction of *F. nucleatum* occurrence were identified and their effectiveness was

validated in *F. nucleatum*-innocent subjects. They also included the previously suggested CRC core signature species. A prediction model constructed with the classifier species successfully estimated microbial dysbiotic state and colonization of diseases-associated microbes. Thepotentialprobability experiencing *F. nucleatum* was significantly associated with the distribution of our markers, microbial diversity and variance. Based on the distribution characteristics, all microbes were classified to suggest potential biomarkers for symbiosis and dysbiosis. Our results provide a novel layer of information to develop conditional biomarkers focused on a specific microbe and to understand the microbial dynamics during perturbation of metagenomic stability.

Declarations

Acknowledgements

The authors thank Dr. Seokjin Ham and Byung Hee Kang for their kind advices in designing the project.

Funding

This work is supported by the grant from the National Research Foundation of Korea (NRF–2014M3C9A3064548, NRF–2017M3C9A6047625), BK21 Plus funded by the Ministry of Education, Republic of Korea (10Z20130012243), and the Technology Development Program of MSS (S2632274)

Authors contributions

J.-W. H. and T.-Y. R. conceived and designed this project. J.-W. H. collected the data and performed the analysis. J.-W. H. and T.-Y. R. wrote and edited manuscript.

All authors read and approved the final manuscript.

Ethnic approval and consent to participate

Not applicable

Competing interest

The authors declare that they have no competing interests.

Availability of data and materials

All publically available datasets were describe in Methods. Supplementary figures are provided separately.

Reference

1.Bradshaw DJ, Marsh PD, Watson GK, Allison C: *Role of Fusobacterium nucleatum and Coaggregation in Anaerobe Survival in Planktonic and Biofilm Oral Microbial Communities during Aeration. Infect Immun* 1998, *66(*10):4729–4732.

2.Saygun I, Nizam N, Keskiner I, Bal V, Kubar A, Acikel C, Serdar M, Slots J: *Salivary infectious agents and periodontal disease status. J Periodontal Res* 2011, *46(*2):235–239.

3.Williams MD, Kerber CA, Tergin HF: *Unusual presentation of Lemierre's syndrome due to Fusobacterium nucleatum. J Clin Microbiol* 2003, *41(7*):3445–3448.

4.Barak S, Oettinger-Barak O, Machtei EE, Sprecher H, Ohel G: *Evidence of periopathogenic microorganisms in placentas of women with preeclampsia. J Periodontol* 2007, *78(*4):670–676.

5.Han YW, Shen T, Chung P, Buhimschi IA, Buhimschi CS: *Uncultivated bacteria as etiologic agents of intra-amniotic inflammation leading to preterm birth. J Clin Microbiol* 2009, *47(*1):38–47.

6.Figuero E, Sanchez-Beltran M, Cuesta-Frechoso S, Tejerina JM, del Castro JA, Gutierrez JM, Herrera D, Sanz M: *Detection of periodontal bacteria in atheromatous plaque by nested polymerase chain reaction. J Periodontol* 2011, *82(*10):1469–1477.

7.Coppenhagen-Glazer S, Sol A, Abed J, Naor R, Zhang X, Han YW, Bachrach G: *Fap2 of Fusobacterium nucleatum is a galactose-inhibitable adhesin involved in coaggregation, cell adhesion, and preterm birth. Infect Immun* 2015, *83(*3):1104–1113.

8.Strauss J, Kaplan G, Beck P, Rioux K, Allen-Vercoe; RPRDTLE: *Invasive potential of gut mucosa-derived fusobacterium nucleatum positively correlates with IBD status of the host.* 2011.

9.Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, Clancy TE, Chung DC, Lochhead P, Hold GL *et al: Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. Cell Host Microbe* 2013, *14(*2):207–215.

10.Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW: *Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin. Cell Host Microbe* 2013, *14(*2):195–206.

11.Fukugaiti MH, Ignacio A, Fernandes MR, Ribeiro Junior U, Nakano V, Avila-Campos MJ: *High occurrence of Fusobacterium nucleatum and Clostridium difficile in the intestinal microbiota of colorectal carcinoma patients. Braz J Microbiol* 2015, *46(*4):1135–1140.

12.Yamamura K, Baba Y, Nakagawa S, Mima K, Miyake K, Nakamura K, Sawayama H, Kinoshita K, Ishimoto T, Iwatsuki M *et al: Human Microbiome Fusobacterium Nucleatum in Esophageal Cancer Tissue Is Associated with Prognosis. Clin Cancer Res* 2016, *22(*22):5574–5581.

13.Rubinstein MR, Baik JE, Lagana SM, Han RP, Raab WJ, Sahoo D, Dalerba P, Wang TC, Han YW: *Fusobacterium nucleatum promotes colorectal cancer by inducing Wnt/beta-catenin modulator Annexin A1. EMBO Rep* 2019, *20(*4).

14.Zhang S, Yang Y, Weng W, Guo B, Cai G, Ma Y, Cai S: *Fusobacterium nucleatum promotes chemoresistance to 5-fluorouracil by upregulation of BIRC3 expression in colorectal cancer. J Exp Clin Cancer Res* 2019, *38(*1):14.

15.Chen Y, Peng Y, Yu J, Chen T, Wu Y, Shi L, Li Q, Wu J, Fu X: *Invasive Fusobacterium nucleatum activates beta-catenin signaling in colorectal cancer via a TLR4/P-PAK1 cascade. Oncotarget* 2017, *8(*19):31802–31814.

16.Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA *et al: Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res* 2012, *22(*2):299–306.

17.Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J *et al: Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res* 2012, *22(*2):292–298.

18.Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin* 2018, *68(*6):394–424.

19.Siegel RL, Miller KD, Jemal A: *Colorectal Cancer Mortality Rates in Adults Aged 20 to 54 Years in the United States, 1970–2014. JAMA* 2017, *318:*572–574.

20.Ramos M, Llagostera M, Esteva M, Cabeza E, Cantero X, Segarra M, Martín-Rabadán M, Artigues G, Torrent M, Taltavull JM *et al: Knowledge and attitudes of primary healthcare patients regarding population-based screening for colorectal cancer. BMC Cancer* 2011, *11(*408).

21.Flanagan L, Schmid J, Ebert M, Soucek P, Kunicka T, Liska V, Bruha J, Neary P, Dezeeuw N, Tommasino M *et al: Fusobacterium nucleatum associates with stages of colorectal neoplasia development, colorectal cancer and disease outcome. Eur J Clin Microbiol Infect Dis* 2014, *33(*8):1381–1390.

22.Eklof V, Lofgren-Burstrom A, Zingmark C, Edin S, Larsson P, Karling P, Alexeyev O, Rutegard J, Wikberg ML, Palmqvist R: *Cancer-associated fecal microbial markers in colorectal cancer detection. Int J Cancer* 2017, *141(*12):2528–2536.

23.Flemer B, Lynch DB, Brown JM, Jeffery IB, Ryan FJ, Claesson MJ, O'Riordain M, Shanahan F, O'Toole PW: *Tumour-associated and non-tumour-associated microbiota in colorectal cancer. Gut* 2017, *66(*4):633–643.

24.Liang Q, Chiu J, Chen Y, Huang Y, Higashimori A, Fang J, Brim H, Ashktorab H, Ng SC, Ng SSM *et al: Fecal Bacteria Act as Novel Biomarkers for Noninvasive Diagnosis of Colorectal Cancer. Clin Cancer Res* 2017, *23(*8):2061–2070.

25.Suehiro Y, Sakai K, Nishioka M, Hashimoto S, Takami T, Higaki S, Shindo Y, Hazama S, Oka M, Nagano H *et al: Highly sensitive stool DNA testing of Fusobacterium nucleatum as a marker for detection of colorectal tumours in a Japanese population. Ann Clin Biochem* 2017, *54(*1):86–91.

26.Wong SH, Kwong TNY, Chow TC, Luk AKC, Dai RZW, Nakatsu G, Lam TYT, Zhang L, Wu JCY, Chan FKL *et al: Quantitation of faecal Fusobacterium improves faecal immunochemical test in detecting advanced colorectal neoplasia. Gut* 2017, *66(*8):1441–1448.

27.Shah MS, DeSantis T, Yamal JM, Weir T, Ryan EP, Cope JL, Hollister EB: *Re-purposing 16S rRNA gene sequence data from within case paired tumor biopsy and tumor-adjacent biopsy or fecal samples to identify microbial markers for colorectal cancer. PLoS One* 2018, *13(*11):e0207002.

28.Zhang X, Zhu X, Cao Y, Fang JY, Hong J, Chen H: *Fecal Fusobacterium nucleatum for the diagnosis of colorectal tumor: A systematic review and meta-analysis. Cancer Med* 2019, *8(*2):480–491.

29.Integrative HMPRNC: *The Integrative Human Microbiome Project. Nature* 2019, *569(*7758):641–648.

30.Hanahan D, Weinberg RA: *Hallmarks of cancer: the next generation. Cell* 2011, *144(*5):646–674.

31.Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F: *Global patterns and trends in colorectal cancer incidence and mortality. Gut* 2017, *66(*4):683–691.

32.Mattar MC, Lough D, Pishvaian MJ, Charabaty A: *Current Management of Inflammatory Bowel Disease and Colorectal Cancer. Gastrointest Cancer Res* 2011, *4(*2):53–61.

33.Soetikno R, Sanduleanu S, Kaltenbach T: *An atlas of the nonpolypoid colorectal neoplasms in inflammatory bowel disease. Gastrointest Endosc Clin N Am* 2014, *24(*3):483–520.

34.Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ *et al: Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature* 2019, *569(*7758):655–662.

35.Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C: *Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods* 2012, *9(*8):811–814.

36.Ley RE, Turnbaugh P, Klein S, Gordon JI: *Human gut microbes associated with obesity. Nature* 2006, *444:*1022–1023.

37.Kriss M, Hazleton KZ, Nusbacher NM, Martin CG, Lozupone CA: *Low diversity gut microbiota dysbiosis: drivers, functional implications and recovery. Curr Opin Microbiol* 2018, *44:*34–40.

38.Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL *et al: The long-term stability of the human gut microbiota. Science* 2013, *341(*6141):1237439.

39.Carding S, Verbeke K, Vipond DT, Corfe BM, Owen LJ: *Dysbiosis of the gut microbiota in disease. Microb Ecol Health Dis* 2015, *26:*26191.

40.Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R *et al: Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med* 2019, *25(*4):679–689.

41.Xie YH, Gao QY, Cai GX, Sun XM, Sun XM, Zou TH, Chen HM, Yu SY, Qiu YW, Gu WQ *et al: Fecal Clostridium symbiosum for Noninvasive Detection of Early and Advanced Colorectal Cancer: Test and Validation Studies. EBioMedicine* 2017, *25:*32–40.

42.Mondot S, Lepage P, Seksik P, Allez M, Treton X, Bouhnik Y, Colombel JF, Leclerc M, Pochart P, Dore J *et al: Structural robustness of the gut mucosal microbiota is associated with Crohn's disease remission after surgery. Gut* 2016, *65(*6):954–962.

43.Brahe LK, Le Chatelier E, Prifti E, Pons N, Kennedy S, Hansen T, Pedersen O, Astrup A, Ehrlich SD, Larsen LH: *Specific gut microbiota features and metabolic markers in postmenopausal women with obesity. Nutr Diabetes* 2015, *5:*e159.

44.Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M *et al: The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe* 2014, *15(*3):382–392.

45.Roberfroid M, Gibson GR, Hoyles L, McCartney AL, Rastall R, Rowland I, Wolvers D, Watzl B, Szajewska H, Stahl B *et al: Prebiotic effects: metabolic and health benefits. Br J Nutr* 2010, *104 Suppl 2:*S1–63.

46.Patterson AM, Mulder IE, Travis AJ, Lan A, Cerf-Bensussan N, Gaboriau-Routhiau V, Garden K, Logan E, Delday MI, Coutts AGP *et al: Human Gut Symbiont Roseburia hominis Promotes and Regulates Innate Immunity. Front Immunol* 2017, *8:*1166.

47.Riviere A, Selak M, Lantin D, Leroy F, De Vuyst L: *Bifidobacteria and Butyrate-Producing Colon Bacteria: Importance and Strategies for Their Stimulation in the Human Gut. Front Microbiol* 2016, *7:*979.

48.Hsieh YY, Tung SY, Pan HY, Yen CW, Xu HW, Lin YJ, Deng YF, Hsu WT, Wu CS, Li C: *Increased Abundance of Clostridium and Fusobacterium in Gastric Microbiota of Patients with Gastric Cancer in Taiwan. Sci Rep* 2018, *8(*1):158.

49.Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, Hurley E, O'Riordain M, Shanahan F, O'Toole PW: *The oral microbiota in colorectal cancer is distinctive and predictive. Gut* 2018, *67(*8):1454–1463.

50.Coker OO, Dai Z, Nie Y, Zhao G, Cao L, Nakatsu G, Wu WK, Wong SH, Chen Z, Sung JJY *et al: Mucosal microbiome dysbiosis in gastric carcinogenesis. Gut* 2018, *67(*6):1024–1032.

51.Pushalkar S, Ji X, Li Y, Estilo C, Yegnanarayana R, Singh B, Li X, Saxena D: *Comparison of oral microbiota in tumor and non-tumor tissues of patients with oral squamous cell carcinoma. BMC Microbiol* 2012, *12(*144):1–15.

52.Wang K, Lu W, Tu Q, Ge Y, He J, Zhou Y, Gou Y, Van Nostrand JD, Qin Y, Li J *et al: Preliminary analysis of salivary microbiome and their potential roles in oral lichen planus. Sci Rep* 2016, *6:*22943.

53.Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner AC, Yu WH, Lakshmanan A, Wade WG: *The human oral microbiome. J Bacteriol* 2010, *192(*19):5002–5017.

54.Moore WEC, Moore LH: *Intestinal Floras of Populations That Have a High Risk of Colon Cancer. J Appl Environ Microbiol* 1995, *61(*9):3202–3207.

55.Saitoh S, Noda S, Aiba Y, Takagi A, Sakamoto M, Benno Y, Koga Y: *Bacteroides ovatus as the predominant commensal intestinal microbe causing a systemic antibody response in inflammatory bowel disease. Clin Diagn Lab Immunol* 2002, *9(*1):54–59.

56.Lucke K, Miehlke S, Jacobs E, Schuppler M: *Prevalence of Bacteroides and Prevotella spp. in ulcerative colitis. J Med Microbiol* 2006, *55(*Pt 5):617–624.

57.Deng X, Li Z, Li G, Li B, Jin X, Lyu G: *Comparison of Microbiota in Patients Treated by Surgery or Chemotherapy by 16S rRNA Sequencing Reveals Potential Biomarkers for Colorectal Cancer Therapy. Front Microbiol* 2018, *9:*1607.

58.Kasai C, Sugimoto K, Moritani I, Tanaka J, Oya Y, Inoue H, Tameda M, Shiraki K, Ito M, Takei Y *et al: Comparison of human gut microbiota in control subjects and patients with colorectal carcinoma in adenoma: Terminal restriction fragment length polymorphism and next-generation sequencing analyses. Oncol Rep* 2016, *35(*1):325–333.

59.Harvey RF, Bradshaw MJ: *Measuring Crohn's diseases activity. Lancet* 1980, *315(*8178):1134–1135.

60.Walsh AJ, Ghosh A, Brain AO, Buchel O, Burger D, Thomas S, White L, Collins GS, Keshav S, Travis SP: *Comparing disease activity indices in ulcerative colitis. J Crohns Colitis* 2014, *8(*4):318–325.

61.Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C: *Metagenomic biomarker discovery and explanation. Genome Biol* 2011, *12(*R60):1–18.

Additional Files

Additional file 1: Figure S1. Microbial variation by sample categories. **(a)** Sex. **(b)** Disease severity. The severity was classified based on their diseases scores. **(c)** Participant. **(d)** Institutes. Five different

institutes have collected fecal samples of IBD and non-IBD participants.

Additional file 2: Figure S2. Microbial diversity and human read fraction. (a) Pielou's evenness, (b) Richness, (c) simple clinical colitis activity index (SCCAI) for UC, (d) Harvey-Bradshaw index (HBI) for CD, (e) Pielou's evenness for *F. nucleatum*-experience, (f) Richness for *F. nucleatum*-experience.

Additional file 3: Figure S3. Low detection probability of opportunistic microbes. (a) Microbial abundance and its detection frequency in 44 duplicated samples, (b) Proportion of half-recovered species among total detected species, (c) Correlation between microbial abundance and detection number. Dot color indicates recovery rate of a certain microbe in pairs.

Additional file 4: Figure S4. Abundance changes for microbial biomarkers. (a) non-IBD markers, (b) IBD markers. Line color indicates sample conditions (red line for IBD, blue line for non-IBD). * indicates p-value < 0.05, ** p < 0.011, *** p < 0.001, **** p < 0.0001

Additional file 5: Figure S5. PCoA plot of 20 *F. nucleatum*-experienced subjects. Line color indicates temporal proximity to *F. nucleatum*.

Additional file 6: Figure S6. Model performance comparison.

(a) AUC, (b) AIC, (c) accuracy, (d) sensitivity, (e) precision, and (f) specificity

Additional file 7: Figure S7. Individual alteration of microbiome in 12 dynamic subjects by inflammatory conditions and posterior probability. Line color indicates posterior probability.

Additional file 8: Figure S8. Intra-individual change of posterior probability and Shannon diversity in 70th percentile dynamic subjects. Pearson correlation coefficients were shown at the bottom of each participant panel.

Additional file 9: Filtering step: removing replicated samples or participants with insufficient number of collections.

Additional file 10: Quality control: removing samples with poor taxonomic assignment.

Additional file 11: Table S1. Basic information of participants.

Additional file 12: LEfSe biomarker screening results.

Additional file 13: Table S2. Classifier species enriched prior or posterior to the detection point of *F. nucleatum*.

Additional file 14: Table S3. Correlation coefficients with *F. nucleatum* and multiple enrichment tests for global biomarker species of colorectal cancer (CRC).

Additional file 15: Metadata.

Additional file 16: Summary of microbial correlation with *F. nucleatum* and enrichment tests.

Figures

Figure 2

Schematic diagram of metagenome analysis. Longitudinal metagenome data from IBDMD were filtered by indicated criteria, and basic characteristics of microbiome were profiled. Based on longitudinal experience of F. nucleatum and temporal distribution toward F. nucleatum-detected samples, microbial characteristics was compared. Using LEfSE algorithms, microbial biomarkers of non-IBD or common IBD condition were screened, and correlation of the marker species with F. nucleatum was assessed. After identifying classifier microbes, which significantly differentiate F. nucleatum-observed point, probability of experiencing F. nucleatum was estimated in F. nucleatum-innocent subjects using multiple logistic regression models. At last, microbes were classified into 9 clusters according to five longitudinal features associated with inflammatory conditions and F. nucleatum experience. Particular clusters contained a significant number of disease-associated marker species or well-known probiotics.

Figure 4

Characteristics of IBD and non-IBD microbiome data. (a) Reciprocal patterns of Firmicutes and Bacteroidetes on principal coordinate analysis (PCoA) plot. (b) Distribution of IBD and non-IBD (ulcerative colitis (UC), and Crohn's disease (CD) samples on PCoA plot. P-values indicate a significance in pairwise comparison between two groups against principal coordinates. (c) Logarithmic abundance heatmap of 1526 samples. Pseudo-abundance (1e-05) was added to avoid infinite value. Samples were ordered by participant and visit number information. (d) Shannon diversity of samples. According to disease severity score, UC and CD samples were categorized into three stages (remission, border, and active). (e) Logarithmic human read fraction of samples.

Figure 6

Transient colonization of F. nucleatum is a sign of intestinal disturbance. (a) IBD and non-IBD frequency by F. nucleatum observation. (b) Logarithmic abundance of F. nucleatum upon observation. Wilcoxon rank-sum test was conducted. (c) Sample classification by F. nucleatum experience, temporal proximity, and directionality. (d) Distribution of samples collected from F. nucleatum-experienced subjects. (e) Shannon diversity by F. nucleatum-oriented classification. (f) Shannon diversity of samples from F. nucleatum-experienced subjects based on temporal proximity to F. nucleatum-detected point. (g) Distribution of samples collected before or after the F. nucleatum-detected samples. (h) Logarithmic

human read fraction of samples by F. nucleatum-oriented classification. (i) Logarithmic human read fraction of samples from F. nucleatum-experienced subjects based on temporal proximity to F. nucleatum-detected point.

Figure 8

Microbial biomarkers for inflammatory conditions highly correlated with F. nucleatum. (a) Screening non-IBD or IBD marker species by LEfSE algorithm. Y-axis indicated logarithmic linear discriminant analysis (LDA) score. (b) Number of detected marker species per sample by inflammatory condition. (c) Logarithmic detection ratio of IBD/non-IBD marker species detected depending on the experience of F. nucleatum. Pseudo-count 1 was added to denominator and numerator to avoid infinite value. (d) Logarithmic detection ratio of IBD/non-IBD marker species by temporal distribution toward F. nucleatum detection. (e) Distribution of IBD and non-IBD marker species along temporal proximity to F. nucleatum observation in F. nucleatum-experienced subjects. Dotted lines indicate the median number of detected marker species in F. nucleatum-innocent subjects. (f) Relationship between Spearman correlation coefficients of biomarker species with F. nucleatum and differential enrichment p-value of the microbes in IBD or non-IBD condition. Circle size denotes the number of detection (NOD) of the microbe across whole samples. (g) Logarithmic abundance of four representative IBD and non-IBD marker species along the temporal axis centered at F. nucleatum-detection. Blue line indicates non-IBD and red line, IBD. Font color for microbes indicates marker classes (blue for non-IBD; darkred for IBD). * indicates p-value < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001.

Figure 10

F. nucleatum experience is associated with microbial destabilization. (a) Analytic scheme for calculating intra-individual stability of microbiome. (b) Intra-individual dissimilarity of microbiome with different time intervals facetted by F. nucleatum experience and inflammatory conditions. (c) Analytic scheme for calculating intra-individual stability of microbiome with fixed initial point. (d) Intra-individual dissimilarity of microbiome with fixed initial time point. (e) Analytic scheme for calculating inter-individual dissimilarity of microbiome. (f-h) Inter-individual dissimilarity of microbiome by inflammatory conditions, F. nucleatum experience, and temporal distribution toward F. nucleatum observation, respectively. (i) Analytic scheme for calculating inter-individual dissimilarity of microbiome with fixed initial point. For F. nucleatum-innocent control, initial points were randomly selected. (j) Inter-individual dissimilarity of microbiome by temporal proximity to F. nucleatum.

Figure 12

F. nucleatum-oriented dynamics is informative of capturing biomarkers for IBD or CRC. (a) Schematic illustration of screening classifier species in F. nucleatum-experienced subjects. (b) List of classifier microbes enriched in F. nucleatum-posterior or prior samples. Fisher's exact test was performed. Triangles are CRC signature species. (c) Number of detected posterior- or prior-enriched classifiers in IBD or non-IBD samples. (Wilcoxon test. **** < 0.0001). (d) Classifying significance of microbes in F. nucleatum-experienced subjects. Red circle indicated CRC signature species detected in F. nucleatum-experienced subject at least 5 times. Fisher's exact test was performed. Gray dotted line indicated p-value = 0.05. (e) Average AUC of microbes and their logarithmic p-value for differential enrichment in F. nucleatum-posterior (upper right) or -prior samples (lower right).

Figure 14

Estimation of F. nucleatum-experience and dysbiosis level in F. nucleatum-innocent subjects. (a) Schematic illustration of constructing multiple generalized linear regression model (b) Average rank of model performance. (c) Model validation using F. nucleatum-experienced subjects. Line color indicates inflammatory condition of subjects. (d) Characterization of predicted F. nucleatum-posterior or -prior groups in F. nucleatum-innocent subjects. Wilcoxon test. ns indicates non-significant (p-value>0.5), ** p < 0.01, *** p <0.001, **** p <0.0001 (e) Correlation between posterior probability and IBD/non-IBD marker ratio or Shannon diversity. Dot color indicates Shannon diversity. Spearman correlation between two parameters and its significance was described at the top of scatter plot. (f) Intra-individual change of posterior probability of Shannon diversity in the top 12 dynamic individuals. Pearson correlation coefficients between posterior probability and Shannon diversity were shown. (g) Posterior probability of whole samples in PCoA plot.

Figure 16

Clustering all detected microbes based on longitudinal distribution. (a) Distribution of IBD/non-IBD marker species on PCA plot. Euclidean distances between species were measured. (b) Distribution of CRC marker species. (c) K-mean clustering of microbes and biomarkers. Blue star marks for cluster 1 and red star for cluster 5 (d) Microbial distribution by clusters. Clusters 1 and 6 were encircled. (e) Detected number of cluster component per sample along temporal proximity to F. nucleatum observation. Line color indicates sample condition. Spearman correlation and its significance were calculated. (f) Correlation between posterior probability and the number of detected microbes by clusters. Spearman correlation and its significance were calculated. (g) Distribution of clusters 1 and 6 in PCoA plot of samples. Logarithmic abundance and the number of detected species were displayed.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile10qualitycontrol.xlsx
- Additionalfile8figureS8.tif
- Additionalfile12LEfSEbiomarkerscreening.xlsx
- Additionalfile7figureS7.tif
- Additionalfile11TableS1Participantsinformation.docx
- Additionalfile4figureS4.tif
- Additionalfile6figureS6.tif
- Additionalfile3figureS3.tif
- Additionalfile9filtering.xlsx
- Additionalfile9filtering.xlsx
- Additionalfile12LEfSEbiomarkerscreening.xlsx
- Additionalfile14TableS3CRCmarkerspecies.docx
- Additionalfile1figureS1.tif
- Additionalfile1figureS1.tif
- Additionalfile13TableS2Classifier.docx
- Additionalfile13TableS2Classifier.docx
- Additionalfile3figureS3.tif
- Additionalfile11TableS1Participantsinformation.docx
- Additionalfile4figureS4.tif
- Additionalfile14TableS3CRCmarkerspecies.docx
- Additionalfile16Microbetestsummary.xlsx
- Additionalfile2figureS2.tif
- Additionalfile5figureS5.tif
- Additionalfile6figureS6.tif
- Additionalfile8figureS8.tif
- Additionalfile16Microbetestsummary.xlsx
- Additionalfile2figureS2.tif
- Additionalfile5figureS5.tif
- Additionalfile10qualitycontrol.xlsx
- Additionalfile15metadata.xlsx
- Additionalfile7figureS7.tif
- Additionalfile15metadata.xlsx