Gut Pathogens

RESEARCH

Open Access

Using whole-genome sequencing (WGS) to plot colorectal cancer-related gut microbiota in a population with varied geography

Han Shuwen[1,3], Wu Yinhang[3], Zhao Xingming[4], Zhuang Jing[3], Liu Jinxin[4], Wu Wei[3] and Ding Kefeng[1,2*]

Abstract

Background: Colorectal cancer (CRC) is a multifactorial disease with genetic and environmental factors. Regional differences in risk factors are an important reason for the different incidences of CRC in different regions.

Objective: The goal was to clarify the intestinal microbial composition and structure of CRC patients in different regions and construct CRC risk prediction models based on regional differences.

Methods: A metagenomic dataset of 601 samples from 6 countries in the GMrepo and NCBI databases was collected. All whole-genome sequencing (WGS) data were annotated for species by MetaPhlAn2. We obtained the relative abundance of species composition at the species level and genus level. The MicrobiotaProcess package was used to visualize species composition and PCA. LEfSe analysis was used to analyze the differences in the datasets in each region. Spearman correlation analysis was performed for CRC differential species. Finally, the CRC risk prediction model was constructed and verified in each regional dataset.

Results: The composition of the intestinal bacterial community varied in different regions. Differential intestinal bacteria of CRC in different regions are inconsistent. There was a common diversity of bacteria in all six countries, such as *Peptostreptococcus stomatis* and *Fusobacterium nucleatum* at the species level. *Peptostreptococcus stomatis* (species level) and *Peptostreptococcus* (genus level) are important CRC-related bacteria that are related to other bacteria in different regions. Region has little influence on the accuracy of the CRC risk prediction model. *Peptostreptococcus stomatis* is an important variable in CRC risk prediction models in all regions.

Conclusion: *Peptostreptococcus stomatis* is a common high-risk pathogen of CRC worldwide, and it is an important variable in CRC risk prediction models in all regions. However, regional differences in intestinal bacteria had no significant impact on the accuracy of the CRC risk prediction model.

Keywords: Colorectal cancer, Intestinal bacteria, Regional difference

Introduction

Cancer incidence and death rates are increasing worldwide. The GLOBOCAN 2020 database showed that lung cancer incidence is high in East Asia, Eastern Europe and southern Europe, nonmelanoma skin cancer incidence is highest in the United States, Australia and Canada, cervical cancer incidence is high in Central Africa, and breast cancer incidence is the highest in most other countries [1]. Cancer is a multifactorial disease, and current

*Correspondence: dingkefeng@zju.edu.cn

[1] Department of Colorectal Surgery and Oncology, Key Laboratory of Cancer Prevention and Intervention, Ministry of Education, The Second Affiliated Hospital, Zhejiang University School of Medicine, 88 Jiefang Road, Building 6 Room 2018, Hangzhou 310009, Zhejiang, China
Full list of author information is available at the end of the article

research is focused on determining the risk factors for cancer. Cancer risk factors can be classified as genetic and environmental [2]. Different distributions of cancer risk factors, especially environmental factors and geographical differences, are the main factors that cause the difference in the highest incidence of cancer in different countries.

Globally, CRC ranks third in incidence and second in mortality among all malignancies [3]. The incidence of CRC increases with the level of economic development [4]. According to statistics, the incidence of colon cancer varies approximately 9 times in different regions of the world. Northern Europe has the highest incidence (25.3% for men and 16.4% for women). The incidence of rectal cancer has a similar regional distribution, and the lowest incidence was found in East Asia (2.8% for men and 1.9% for women). Overall, the incidence of CRC is low in most parts of Africa and Central and South Asia (all less than 9.0%) [1]. In addition, the incidence and mortality rates of CRC vary by race, with black individuals having the highest rates (45.7%) and Asians having the lowest (30.0%) [5].

CRC is a cancer caused by a combination of genetic and environmental factors. For example, RAS [6] and BRCA [7] gene mutations and mismatch repair genes MLH1, MSH2, MSH6, and PMS2 mutations [8] are associated with CRC risk. Most CRCs are sporadic cancers caused by environmental factors, among which the main risk factors include obesity, physical inactivity, poor diet, alcohol consumption and smoking [9]. Regional and population differences in CRCs lead to regional differences in cancer-related risk factors, including socioeconomic factors, diet, intestinal microbial changes, immune microenvironment changes, and genetic mutations [10]. In addition, the coverage ratio of cancer screening and targeted intervention measures are important factors in influencing the regional difference in CRC incidence [11]. Genome-wide sequencing association studies identified risk-associated loci for CRC, but these loci did not differ significantly between regions [12, 13]. Existing studies are still unable to account for regional variations in CRC risk.

There is considerable evidence that microbial dysregulation in the human gut is an important risk factor for CRC. Decrease in propbiotic and increase in pathogenic bacteria was identified in CRC incidence. Probiotics (such as *Bifidobacterium*, *Lactobacillus* and *Bacteroidetes*) decreased, while pathogenic bacteria (such as *enterotoxigenic Bacteroidetes*, *Escherichia coli* and *Clostridium difficile*) increased [14]. The main causes of gut microbiota dysbiosis include diet, drugs, environmental pollutants and gut immune dysfunction and so on [14–16]. For example, Chen et al. reported that most patients with advanced colorectal adenoma had a low fiber diet. Meanwhile, the abundance of *Clostridium*, *Roseburia*, and *Eubacterium spp.* in the advanced colorectal adenoma group with a low fiber diet increased, while the abundance of *Enterococcus*, *Streptococcus spp.* and butyric-producing bacteria decreased [17]. Increased toxins secreted by the bacteria lead to intestinal mucosal damage and chronic inflammation, which ultimately induces CRC [18]. Intestinal microbes are directly affected by the genetic and environmental CRC risk factors mentioned above [19].

In environmental risk factors, for example, obesity may contribute to CRC through LPS-mediated systemic inflammation and a decrease in short-chain fatty acid (SCFA)-producing bacteria [20]. High-fat diets induce intestinal microecological disorders, leading to an increase in pathogenic bacteria (*Alistipes sp. Marseille-P5997*, *Alistipes sp. 5CPEGH6*) and a decrease in beneficial bacteria (*Parabacteroides distasonis*, *Parabacteroides sp. CT06*). Finally, it damages intestinal barrier dysfunction and promotes the occurrence of CRC [21]. High-fiber diets inhibit CRC by increasing bacterial metabolites that produce SCFA [22]. Long-term alcohol consumption or smoking can reduce the abundance of *Bifidobacteria*, *Bacteroides* and *Firmicute*s and increase the abundance of *Proteus* and *Actinomycetes* [23, 24]. Genetic mutations also play a role in differences in gut microbial composition and abundance. Mutation of the KRAS gene can change the abundance of *Roseburia*, *Parabacteroides*, *Metascardovi*a, *Staphylococcus*, and *Bacillale* and affect the composition of the intestinal bacterial community [25]. Mutation of APC is closely related to changes in intestinal microbiota (such as *Faecalibacterium prausnitzii* and *Fusobacterium mortiferum*) and serum metabolites [26]. Some intestinal microorganisms can also inhibit the inhibitory effect of P53 on the WNT pathway and promote the occurrence of CRC [27]. Importantly, gene mutations of CRC are also affected by region or ethnicity [28]. Yong et al. found that BRAF and KRAS mutation rates in CRC in Asian patients were lower than those in North American CRC patients [29]. Guda et al. found 15 novel mutated genes in African-American patients with CRC, which are rarely mutated in Caucasians with CRC [30]. Therefore, differences in CRC risk factors (both genetic and environmental) caused by regional or geographical factors can directly lead to differences in intestinal microbiota among populations of different regions, and these differences lead to differences in CRC risk among different regions.

This study included a large number of samples worldwide, compared the different intestinal microflora of healthy people and CRC patients in different regions, and identified regional differences in the intestinal microflora of CRC patients. This study provides new ideas for the study of the incidence and etiology of regional differences in CRC.

Methods

Data sources and acquisition

CRC-related metagenomic data were collected from GMrepo [31] and NCBI databases (https://www.ncbi.nlm.nih.gov/sra) [32]. The samples with detailed country and age information were screened out, and the samples with missing information such as BMI and age were removed. All WGS sequencing data was uniformly annotated using MetaPhlAn2. The relative abundance table of species level and genus level was obtained for downstream analysis. Samples with less than 2% of the species were removed. Finally, we recruited 601 samples, including 279 CRC samples and 322 healthy controls, as shown in Table 1 and Additional file 1: Figure S1. The quality control steps refer to the article of Wu et al. [31]. The design and workflow of this study are shown in Additional file 2: Figure S2.

Descriptive research (composition ratio)

The MicrobiotaProcess package was used to visualize species composition and PCA according to CRC/healthy grouping. The sample composition mainly shows the top 25 species and top 20 genera (in order of species prevalence and average abundance), and the remaining species were classified as others. Second, the logarithmic PCA dimension reduction diagram of the relative abundance of species was calculated. To further display the differences, the differential species of the second part were proposed, and the species composition and PCA diagram were drawn according to the same rules.

Difference analysis

For the separate difference analysis of each dataset, the species with a prevalence < 0.01 and the maximum relative abundance < 0.001 in all samples were filtered out. LEfSe analysis (http://huttenhower.sph.harvard.edu/galaxy) [33] was performed for the CRC/healthy group for all relative abundance of species composition differences between sample inspection. The threshold value of effect size LDA score (log10) > 2 was used to screen out the differential species, visualize the bar chart of effect size of the differential species (positive and negative only represent the direction, and the magnitude of effect size is the absolute value, and the larger the absolute value is, the greater the difference is), and draw the phylogenetic branching diagram (the maximum level is only labeled to genus).

Table 1 Overview of data information

NCBI project.id	Country	CRC samples	Healthy controls	Total
PRJDB4176	Japan	40	40	80
PRJEB10878	China	72	54	126
PRJEB12449	USA	49	51	100
PRJEB27928	Germany	21	59	80
PRJEB6070	France	51	59	110
PRJEB7774	Austria	46	59	105
		279	322	601

Correlation analysis

Spearman correlation coefficients were calculated separately for groups of different disease states, and FDR multiple test correction was performed by BH. Only the correlations with an absolute value of correlation coefficients > 0.3 and P value < 0.05 were retained, and the rest were assigned 0. Plot correlation heatmaps were made using the Corrplot package.

Construction of the CRC risk prediction model

According to the risk grouping label of the sample phenotype, a binary classifier was established by using the SIAMCAT package [34] three times fivefold nested cross-validation LASSO algorithm. For the training set, species with a maximum relative abundance < 0.001 and mean relative abundance < 0.0001 in all samples were filtered out, and logarithmic standardization was adopted to obtain the internal cross-validation AUC of self-training. In addition, the external validation AUC can be obtained by using the model as external validation with datasets from other regions, and finally, the AUC result graph can be drawn.

The heatmap describes the difference in bacteria in patients with CRC in different regions

To further highlight the differences in the bacteria of CRC patients in different regions, we selected the species with specific differences in the above regions (LDA Score (LOG10) absolute value > 3 + log2FC > 2 or < -2 species) and the proportion of common bacteria in all samples as pie charts. The MAPS package and the GGploT2 package were used to display the map.

Results

The composition of the intestinal bacterial community varies in different regions

We analyzed the species composition of intestinal bacteria in healthy groups and CRC groups from Japan, China, the USA, Germany, France and Austria at the species level and genus level, respectively. The overall composition of gut bacteria was found to differ between the six countries. For example, at the genus level, *Phocaeicola*, *Prevotella* and *Subdoligranulum* are more prevalent in Japan. *Faecalibacterium*, *Mediterraneibacter* and *Roseburia* are more common in China. *Escherichia*, *Adlercreutzia* and *Lachnoclostridium* are more prevalent in the USA. *Roseburia*, *Mediterraneibacter* and *Collinsella* are more common in Germany. *Bacteroides*, *Ruminococcus* and *Bifidobacterium* are more common in France. There are many *Roseburia*, *Mediterraneibacter* and *Collinsella* in the population of Austria (Fig. 1 and Additional file 3: Figure S3).

At the species level, *Subdoligranulum sp., [Ruminococcus] torques* and *Collinsella aerofaciens* are more prevalent in Japan. *Faecalibacterium prausnitzii, Lachnospira eligens* and *Bacteroides caccae.* are more common in China. *Bacteroides ovatus*, *Bacteroides fragili* and *Roseburia hominis* are more prevalent in the USA. *Collinsella aerofaciens, Roseburia intestinalis* and *Roseburia inulinivorans* are more common in Germany. *Ruminococcus bromii, Lachnospira eligens* and *Bifidobacterium longum* are more common in France. *Ruminococcus sp. 5_1_39BFAA*, *Escherichia coli*, and *Methanobrevibacter smithii* are more prevalent in Austria (Additional file 4: Figure S4).

There are differences in intestinal bacteria between CRC patients and healthy people in different regions

LEfSe analysis was used to test the relative abundance composition of all species among CRC/healthy group samples for the datasets of each country, and it was found that the CRC diversity bacteria were different in different regions (Fig. 2).

In Japan, 36 species CRC differential bacteria at species level, including *Subdoligranulum sp.* and *[Ruminococcus] torque* were found. CRC differential bacteria at genus level, including *Prevotella* and *Subdoligranulum*, of 21 species were identified.

In China, there were 61 types of CRC differential bacteria (*Bacteroides fragilis, Bacteroides caccae*, etc.) at species level and 30 types of CRC differential bacteria (*Mediterraneibacter*, *Eikenella*, etc.) at genus level.

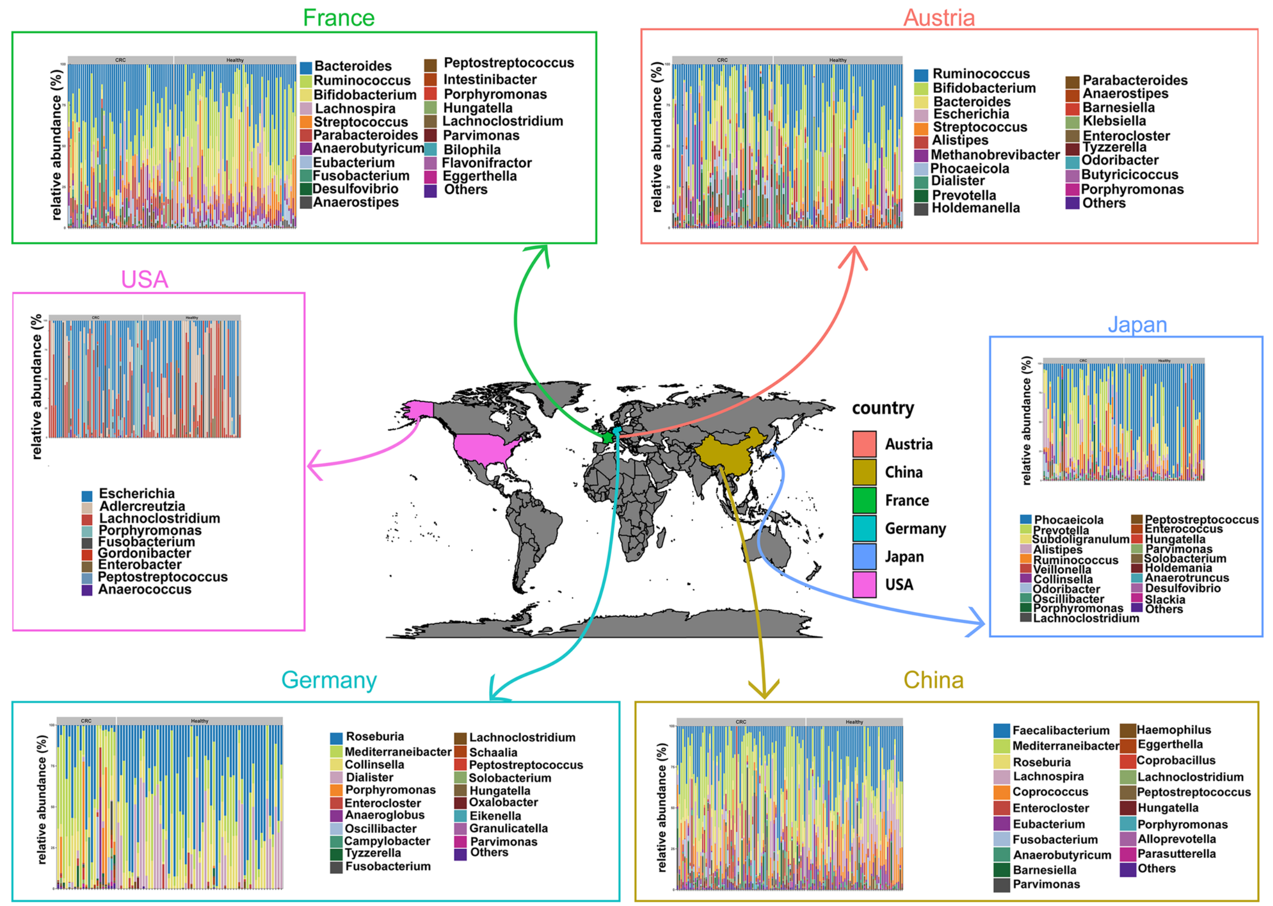

Fig. 1 Composition of the intestinal microbiome of populations in different regions. The figure shows the species composition diagram of CRC and healthy people from Japan, China, the USA, Germany, France and Austria

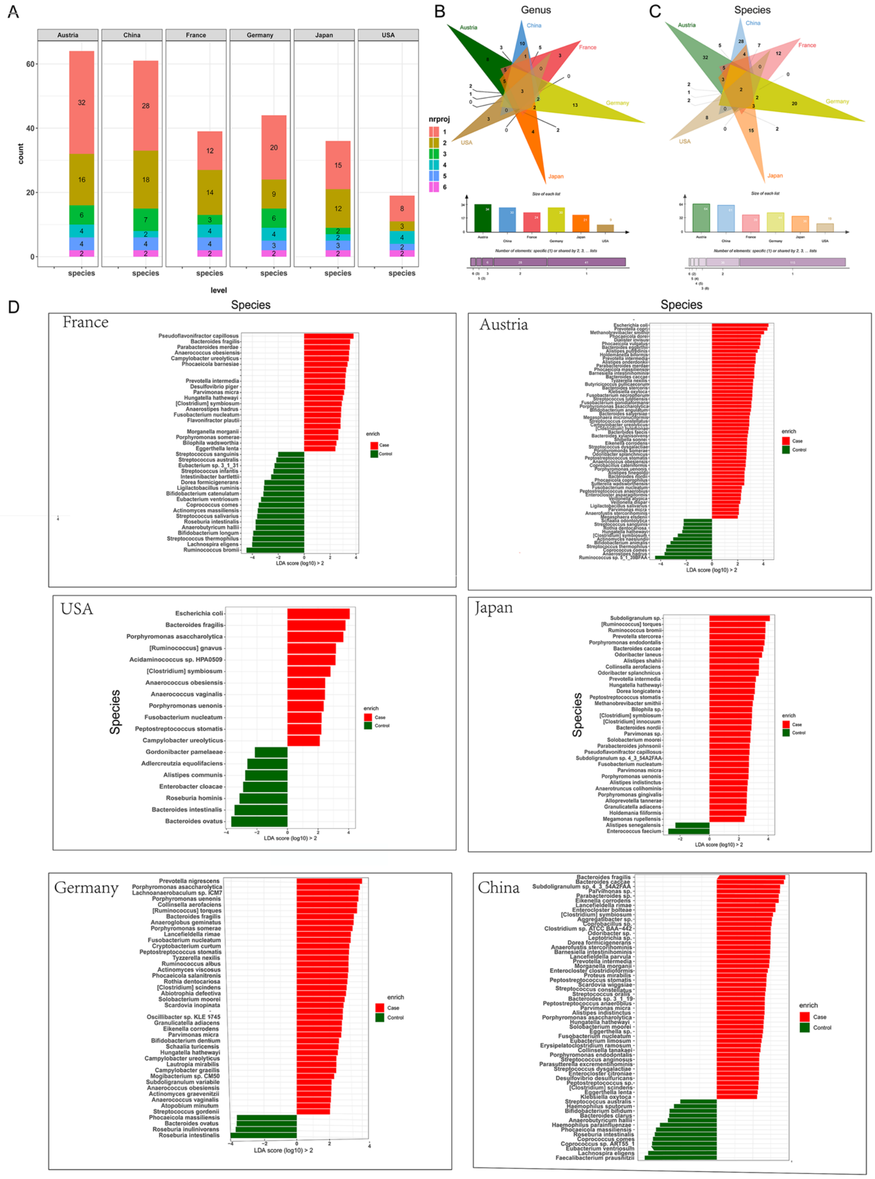

Fig. 2 Differences in CRC intestinal bacteria in different regions. Figure **A** CRC differential species frequency statistics of Japan, China, the USA, Germany, France, and Austria at the genus and species levels. Figures **B** and **C** Venn plots of CRC differential intestinal bacteria from six countries at the genus and species levels, respectively. Figure **D** Differential CRC bacteria at the genus and species levels in six countries are described in the map. The LDA discriminant histogram is in the box. The greater the LDA score is, the greater the impact of species abundance on the differential effect

In USA, a total of 19 kinds of CRC different bacteria at species level, including *Escherichia coli* and *Bacteroides fragilis*, and 9 kinds of CRC different bacteria, including *Escherichia* and *Porphyromonas* at genus level were identified.

In Germany, a total of 44 kinds of CRC different bacteria at species level, including *Prevotella nigrescens* and *Porphyromonas asaccharolytica*, and 30 kinds of CRC different bacteria at genus level, including *Mediterraneibacter* and *Porphyromonas* were found.

In France, there were 39 kinds of different bacteria at species level, including *Pseudoflavonifractor capillosus* and *Bacteroides fragili*, and 24 kinds of different bacteria at genus level, including *Bacteroides* and *Fusobacterium*.

In Austria, there were 64 kinds of different bacteria at species level, including *Escherichia coli* and *Prevotella*

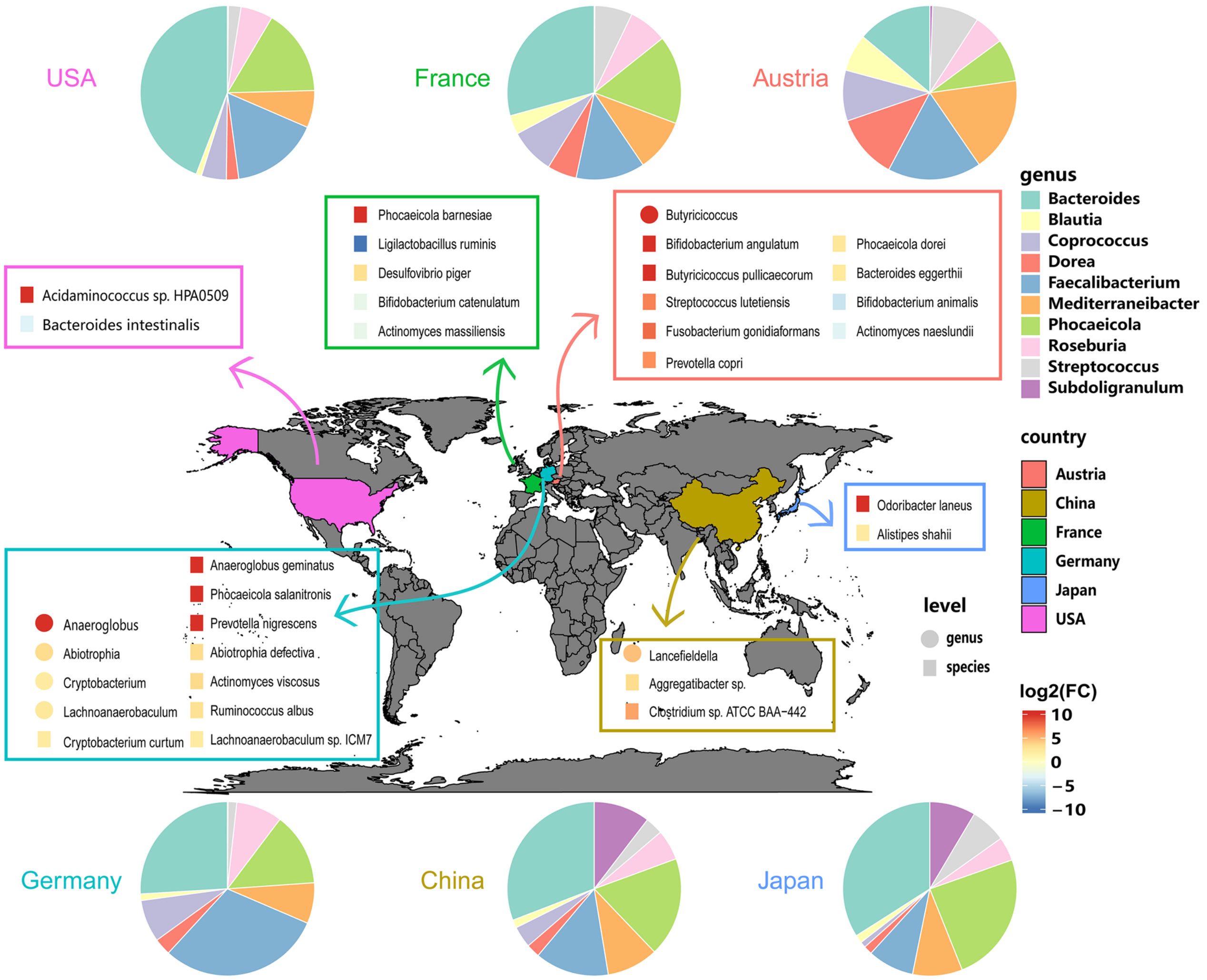

Fig. 3 Differences in disease bacteria among different regions. Heatmap describing bacterial differences in CRC patients in different regions. The pie chart shows the percentages of 10 common bacteria at the genus level in six countries

copri, and 34 kinds of different bacteria at genus level, including *Escherichia* and *Prevotella*.

These results indicate that regional differences affect intestinal microflora.

In addition, we found common bacteria in six countries, including *Peptostreptococcus*, *Porphyromonas* and *Lachnoclostridium* at the genus level (Fig. 2B) and *Peptostreptococcus stomatis* and *Fusobacterium nucleatum* at the species level (Fig. 2C). The USA has the fewest bacteria in common with other regions. The overlap between China and France was highest (Genus: *Eubacterium*, *Lachnospira*, *Anaerobutyricum*. *Morganella*. *Eggerthella*. Species: *Morganella morganii*, *Lachnospira eligens*, *Dorea formicigenerans*, *Eggerthella lenta*, *Streptococcus australis*, *Eubacterium ventriosum*, *Anaerobutyricum hallii*).

Finally, we screened 10 common bacteria in all the samples, *Bacteroides*, *Blautia*, *Coprococcus*, *Dorea*, *Faecalibacterium*, *Mediterraneibacter*, *Phocaeicola*, *Roseburia*, *Streptococcus* and *Subdoligranulum*, and calculated the proportion of these 10 bacteria in 6 countries. The results showed that except for in Austria, the proportion of *Bacteroides* was large. *Faecalibacterium* is more prevalent in Germany but less prevalent in other countries. *Subdoligranulum* is only high among people in China and Japan. In addition, we drew a heatmap to visually describe the microflora differences of CRC patients in different regions (Fig. 3). As shown in the figure, CRC-associated bacteria are different in all six countries.

Interaction of CRC-related intestinal bacteria in different regions

In addition, we analyzed the correlation of CRC-differential bacteria in six regions (Figure 4).

At the species level, *Peptostreptococcus stomatis* is an important CRC-related bacterium that is related to other bacteria in different regions. In China, Germany and France, *Peptostreptococcus stomatis* was positively correlated with *Parvimonas sp.* (Spearman=0.78 in China,

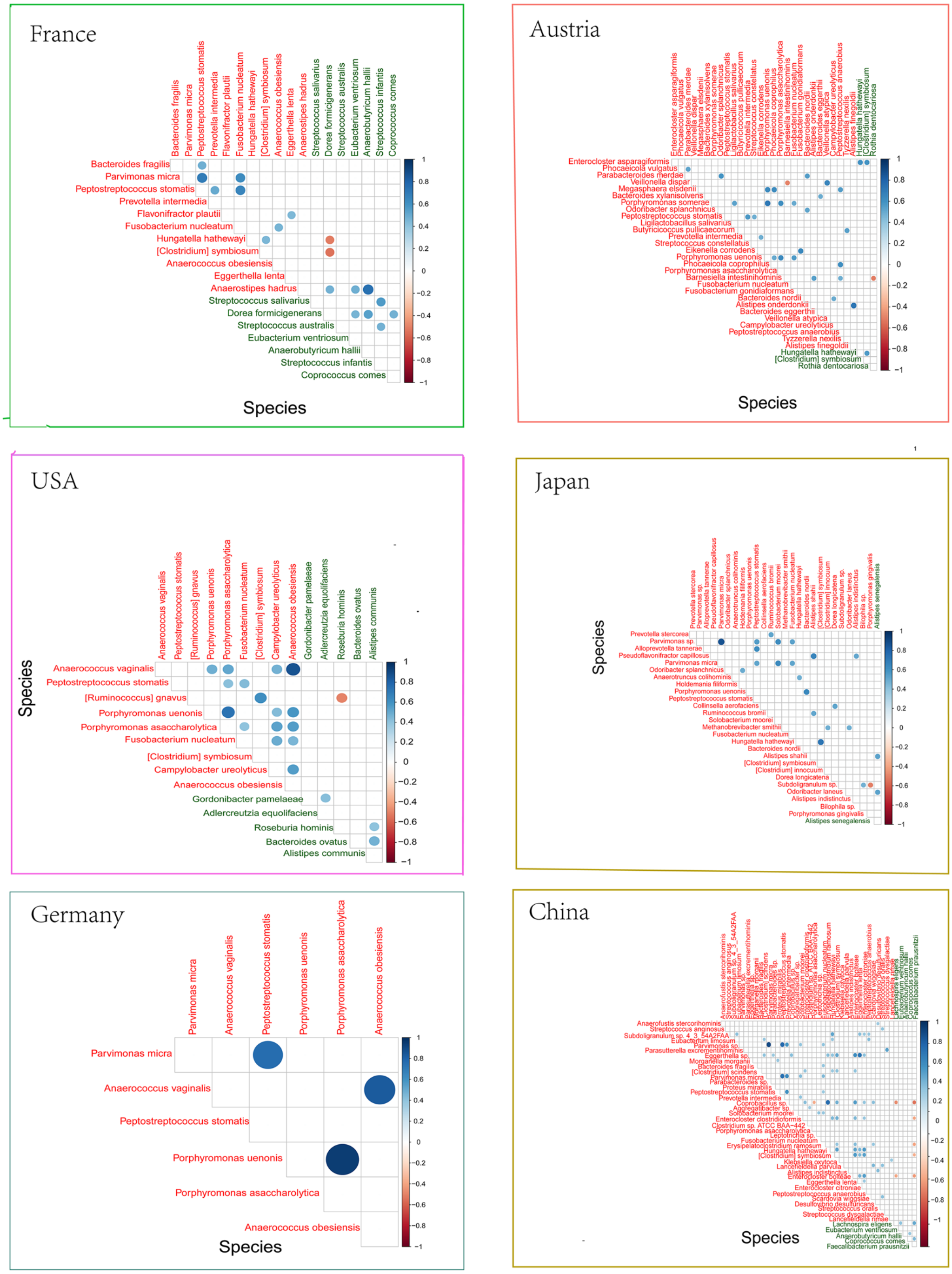

Fig. 4 Correlation of different disease-related bacteria. Interactions between differentially abundant bacteria in CRC and healthy groups at the genus and species levels in the map. The dots represent correlations, blue represents positive correlations, and red represents negative correlations. The darker the color and the larger the dots are, the greater the correlation coefficient

Spearman=0.77 in Germany, Spearman=0.70 in France). In France, *Peptostreptococcus stomatis* and *Fusobacterium nucleatum* were also positively correlated (Spearman=0.65). Moreover, in Japan and China, *Parvimonas micra* and *Parvimonas sp.* were the most relevant (Spearman=0.86 in Japan, Spearman=0.90 in China).

Similarly, at the genus level, *Peptostreptococcus* was also related to other bacteria. The correlation between *Peptostreptococcus* and *Parvimonas* was high (Spearman=0.84 in Japan, Spearman=0.73 in China, Spearman=0.70 in Germany, Spearman=0.69 in France). In the USA, *Peptostreptococcus* was also associated with *Anaerococcus* (Spearman=0.48).

Regions do not affect the accuracy of CRC risk prediction models

We constructed CRC risk prediction models based on intestinal bacteria in each region, conducted internal cross-validation and external validation of data from other regions, and ranked the importance of intestinal bacteria in each region model. The final findings are given below.

The CRC risk prediction model was constructed based on all regional samples, and the AUCs at the genus and species levels were 0.783 and 0.84, respectively. The intestinal bacteria with the highest importance of CRC risk prediction model variables were *Peptostreptococcus stomatis* at the species level and *Peptostreptococcus* at the genus level (Fig. 5A and Additional file 5: Figure S5B).

Based on the single-region CRC risk prediction model, although its accuracy decreased when applied to other regions, the overall accuracy was within the acceptable range (Fig. 5B and Additional file 5: Figure S5A). Moreover, the accuracy of the CRC risk prediction model based on all regional samples was not lower than that of the single-region model. *Peptostreptococcus* at the genus level and *Peptostreptococcus stomatis* at the species level are still characteristic CRC intestinal bacteria in all regions (Fig. 5C, D).

Discussion

Based on metagenomic data, this study screened the bacterial characteristics of CRC in different regions and established a CRC risk prediction model. It was found that the composition of the intestinal bacterial community at the species and genus levels was different in the populations of the six countries. CRC-differentiated bacteria were also different in different regions. However, there were also differences in bacteria shared by the six countries, including *Peptostreptococcus stomatis* and *Fusobacterium nucleatum*, at the species level. There are few overlapping intestinal bacteria in the USA with other regions. *Peptostreptococcus stomatis* (species level) and *Peptostreptococcus* (genus level) are important CRC-related bacteria that are related to other bacteria in different regions. There was no significant difference in the accuracy of CRC risk prediction models based on a single region and all regions. The important intestinal bacteria in the CRC risk prediction model are *Peptostreptococcus stomatis* at the species level and *Peptostreptococcus* at the genus level.

Peptostreptococcus has been shown to increase significantly in the intestines of CRC patients [35]. Studies have confirmed that the relative species abundance of the oral microbiota in the intestinal tract of CRC patients is significantly increased, including *Peptostreptococcus stomatis*, *Fusobacterium nucleatum* and *Parvimonas micra* [36]. A study on 16S rRNA gene sequencing of intestinal microorganisms in CRC patients showed that there were consistent CRC-related intestinal bacteria in developed and developing countries, and *Parvimonas*, *Peptostreptococcus* and *Fusobacterium* were important in distinguishing CRC from healthy people [37]. Long et al. found that *Peptostreptococcus anaerobius* adsorbed in putative cell wall binding repeat 2 (PCWBR2), targeting the α2/β1-PI3K-Akt-NF-κB signaling axis, which drives CRC [38]. Moreover, Yu et al. confirmed known associations of *Peptostreptococcus stomatis*, *Parvimonas micra*, and *Fusobacterium nucleatum* with CRC [39]. In addition, there was a significant association between *Peptostreptococcus stomatis* and other species, and microbial marker compositions could improve the accuracy of the early diagnosis of CRC [40]. This study analyzed common intestinal bacteria with high CRC correlation in different regions, indicating that *Peptostreptococcus stomatis* is highly carcinogenic to the intestinal tract. However, the carcinogenic mechanism of *Peptostreptococcus stomatis* is not yet known.

In addition, differences were found in CRC-associated bacteria in different regions, and there was little overlap

(See figure on next page.)

Fig. 5 CRC risk prediction model and importance of variables. Figure **A** Overall CRC risk prediction model based on the genus and species levels in all regions. Figure **A**-a1: ROC curve at the genus and species levels. Figure **A**-a2: Top 20 characteristic model interpretation diagrams. The left side represents the relative weights of the corresponding features in the 15 cross-validation submodels in the training model, the middle is the normalized abundance values of each species among the grouped samples, and the right side is the boxplot of the ratio of the 20 features with nonzero coefficients among the 15 submodels. Figure **B** Cross-validation of disease risk prediction models between species datasets in different regions. Figure **C** Venn diagram of the top 20 characteristic bacteria between datasets at the species level. Figure **D** Top 20 characteristic model interpretation diagrams of the species-level CRC risk prediction model in each region in the map

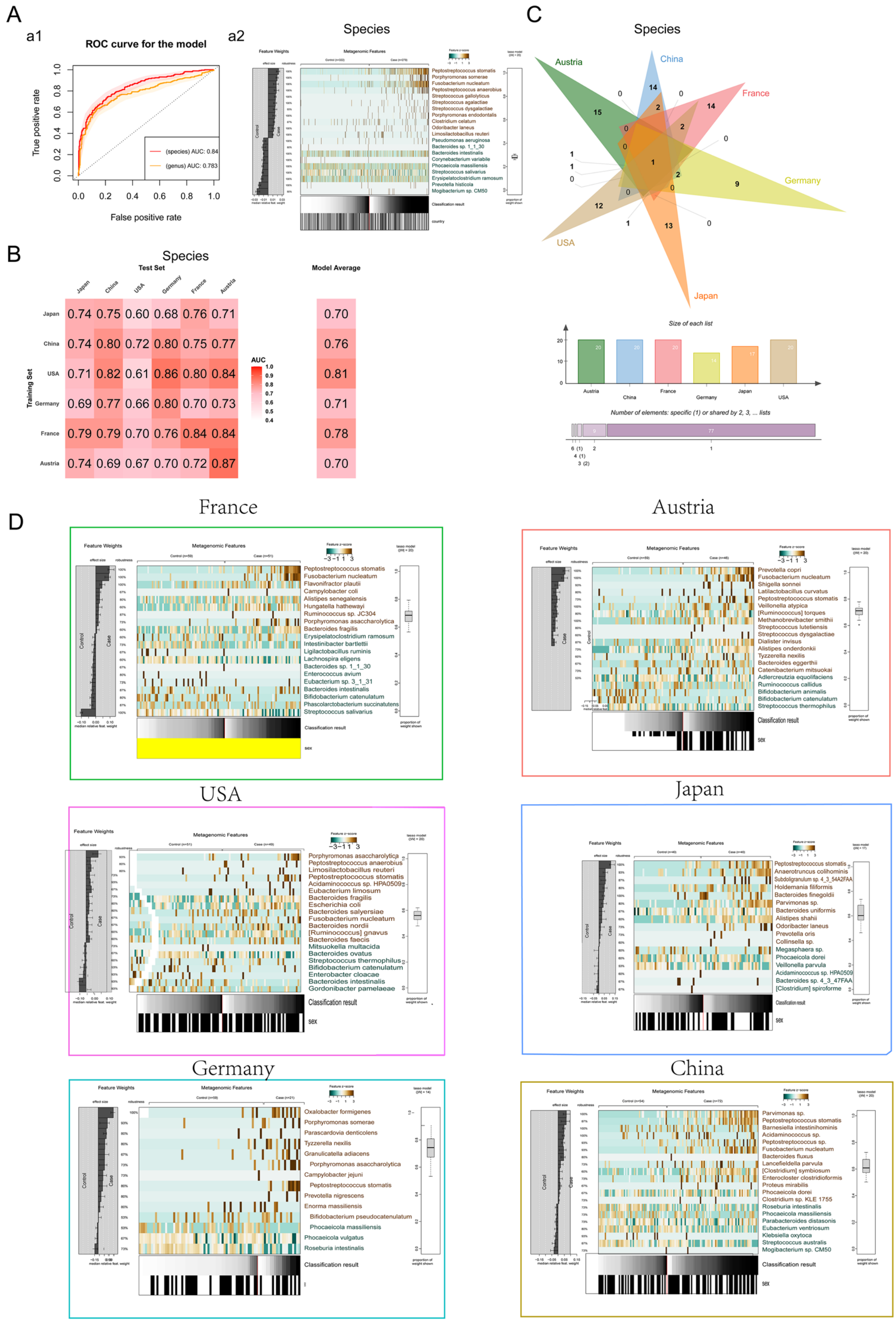

Fig. 5 (See legend on previous page.)

between the USA and the other five countries in CRC differential intestinal bacteria. There are differences between CRC-associated bacteria in China and Japan. The taxonomy and functional composition of the gut microbiome across populations in terms of health or disease are critical to unearthing its role in maintaining human health. Large-scale, global microbiome projects have revealed changes in gut microbiome composition in healthy individuals due to geographic location, host genetics, delivery mode, age, nutrition, diet, and lifestyle [41]. Genetics has been thought to play an important role in determining differences in the microbiome between people. Genes determine the environment the microbiome occupies, and each particular environment allows certain strains of bacteria to grow. Moreover, the diversity and composition of gut microbes vary among different ethnic groups. Deschasaux et al. analyzed the intestinal microbiota of six ethnic groups (439 Dutch, 367 Ghanaians, 280 Moroccans, 197 Turks, 443 African Surinamese, and 358 South Asian Surinamese) living in the Netherlands. Race was found to be an important marker of differences in the composition of human fecal microbiota [42]. However, Rothschild et al. reported that host genetic factors play a minor role in determining the composition of the microbiome, with 98% of the variation in the human gut microbiome determined by environmental factors [43]. A previous study also used metagenomic analysis of the CRC fecal microbiome to obtain microorganisms from CRC cohorts of different races [39]. In addition, different geographical locations within the same country or region can also lead to differences in human gut microbes. For example, Bramble et al. used shotgun metagenomic sequencing to perform a large-scale comparison of gut microbiome profiles in 180 children (from the urbanized capital Kinshasa to extreme rural areas in the southwest of the Democratic Republic of the Congo, including children affected by Konzo disease from prone villages). It was found that the intestinal microbiome structure varied greatly in different regions, but there was no significant difference in intestinal microbiome or functional enrichment between konzo-prone regions [44]. Yan et al. characterized the intestinal microbiota of the population from 14 regions in one province and found that the influence of region on intestinal microbiota was much greater than other factors, and regional differences affected the cross-regional application of disease models [45]. Therefore, the USA is not geographically adjacent to other countries, which may account for the obvious differences in CRC-associated microorganisms between the USA and other regions. Although China and Japan are both Asian countries, there are still differences in CRC-related bacteria. This may be due to the large gap in dietary habits between the two countries, and the main diet structure of Japan includes raw food.

In this study, an overall CRC risk prediction model and a single-region CRC risk prediction model were constructed based on intestinal microorganisms. Although the accuracy of the single-region model fluctuates when applied in other regions, it is within the acceptable range. The accuracies of the multi-region and single-region models are basically the same. In addition, the bacteria common to the six countries identified in this study also play an important role in the classification of the CRC risk prediction model. Based on a CRC metagenomic dataset of 1,368 samples from different geographic cohorts, Liu et al. identified 16 markers in multiple regions (China, Italy, and the United States), including 11 bacteria, 4 fungi, and 1 archaea. The CRC diagnostic model based on microbial characteristics performed well in different geographic cohorts (AUROC = 0.83) [46]. Thus, the prediction of CRC risk using gut microbes does not appear to be affected by the differential bacteria, even though there are differences in CRC-associated bacteria within different regions.

Compared with 16S variable region sequencing of bacteria, metagenomic sequencing can more accurately locate information at the bacterial species level in microbiome studies. Big data provides the basis for this research, indicating that promoting data sharing and making full use of the advantages of big data can promote new discoveries and new meaningful phenomena. Several studies have investigated links between the gut microbiome and CRC through metagenomic data. Thomas et al. performed a meta-analysis of fecal metagenomic datasets from reported cohorts involving five countries and two new cohorts from Italy. The composition and functional characteristics of CRC-associated gut microbiota were identified. CRC prediction model for 16 species was constructed and validated [47]. Wirbel et al. conducted a similar meta-analysis of eight geographically and technically diverse fecal shotgun metagenomic studies of CRC [48]. These two studies confirmed that there was heterogeneity in CRC microbiota characteristics in different populations, but comprehensive analysis of flora markers obtained from multiple cohorts could improve diagnostic accuracy. The difference was that present study highlighted the role of regional differences in CRC gut microbiota, even though the accuracy of our single-region model and multi-region model did not differ significantly.

We collected 601 valid metagenomic data samples of intestinal bacteria from six national datasets (Austria, China, Japan, USA, France and Germany) on three continents (Asia, Europe, and the Americas), deeply explored the differences in intestinal flora in different geographical locations, and further analyzed the relationship between regional differences in intestinal flora and CRC risk. While studies have found that long-term diet, food diversity, and overall nutrition may be important factors in these differences, these ideas have not been proven. This study needs to expand the sample size and include more factors, such as lifestyle, diet and disease, to further analyze the factors related to regional differences in CRC intestinal bacteria.

Conclusion

In the present study, WGS sequencing data of intestinal bacteria from 601 samples from 6 countries were included and analyzed at the species and genus levels in Japan, China, the USA, Germany, and France. To be confirmed, it was found that the intestinal bacterial community composition and CRC differential bacteria in different regions were different, and regional differences in CRC were identified. In addition, *Peptostreptococcus stomatis* is a CRC-associated bacterium common in all regions. Regional differences in intestinal bacteria had no significant impact on the accuracy of the CRC risk prediction model. In conclusion, region is an important factor leading to differences in CRC-related intestinal bacteria. The study provides new ideas for the study of CRC etiology from the perspective of regional differences.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13099-022-00524-x.

Additional file 1: Fig. S1. Basic information and characteristics of each region.

Additional file 2: Fig. S2. The flow chart.

Additional file 3: Fig. S3. PCA diagram of CRC and healthy people from different regions. PCA diagram of CRC and healthy people from Japan, China, the USA, Germany, France and Austria. In the PCA dimensionality reduction diagram, the horizontal and vertical coordinates were the first and second principal components (explanatory variances in parentheses), and the top 5 features with the largest contribution to the first and second principal components are shown in the figure.

Additional file 4: Fig. S4. Composition of the intestinal bacterial community at the species level in each region

Additional file 5: Fig. S5. CRC risk prediction model and importance of variables at the genus level.

Acknowledgements

Not Applicable.

Author contributions

DK conceived and designed the study. HS and WY wrote the manuscript. ZX and LJ carried out the data acquisition, date analysis and statistical analysis. ZJ and WW designed and draw figures. All authors read and approved the final manuscript.

Funding

This work was supported by the Key Research and Development Project of Zhejiang Province (No.2022C03026), National Natural Science Foundation of China (No. 82072624) and the Fundamental Research Funds for the Central Universities (No. 226–2022-00009).

Availability of data and materials

The datasets generated during the current study are not publicly available but obtained from corresponding authors on reasonable request.

Declarations

Ethics approval and consent to participate

NA.

Consent for publication

Not applicable.

Competing interests

The authors declare that no potential conflicts of interest exist.

Author details

[1] Department of Colorectal Surgery and Oncology, Key Laboratory of Cancer Prevention and Intervention, Ministry of Education, The Second Affiliated Hospital, Zhejiang University School of Medicine, 88 Jiefang Road, Building 6 Room 2018, Hangzhou 310009, Zhejiang, China. [2] Cancer Center Zhejiang University, Hangzhou, Zhejiang, China. [3] Huzhou Central Hospital, Huzhou, Zhejiang, China. [4] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China.

Received: 30 August 2022 Accepted: 19 December 2022
Published online: 28 December 2022

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Wu S, Zhu W, Thompson P, Hannun YA. Evaluating intrinsic and non-intrinsic cancer risk factors. Nat Commun. 2018;9(1):3490.
3. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin. 2021;71(1):7–33.
4. Fidler MM, Soerjomataram I, Bray F. A global view on cancer incidence and national levels of the human development index. Int J Cancer. 2016;139(11):2436–46.
5. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2020. CA Cancer J Clin. 2020;70(3):145–64.
6. Osumi H, Shinozaki E, Suenaga M, Matsusaka S, Konishi T, Akiyoshi T, Fujimoto Y, Nagayama S, Fukunaga Y, Ueno M, et al. RAS mutation is a prognostic biomarker in colorectal cancer patients with metastasectomy. Int J Cancer. 2016;139(4):803–11.
7. Oh M, McBride A, Yun S, Bhattacharjee S, Slack M, Martin JR, Jeter J, Abraham I. BRCA1 and BRCA2 gene mutations and colorectal cancer risk: systematic review and meta-analysis. J Natl Cancer Inst. 2018;110(11):1178–89.
8. Hampel H, Frankel WL, Martin E, Arnold M, Khanduja K, Kuebler P, Nakagawa H, Sotamaa K, Prior TW, Westman J, et al. Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). N Engl J Med. 2005;352(18):1851–60.
9. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol. 2019;16(12):713–32.
10. Carethers JM. Racial and ethnic disparities in colorectal cancer incidence and mortality. Adv Cancer Res. 2021;151:197–229.
11. Ashktorab H, Kupfer SS, Brim H, Carethers JM. Racial disparity in gastrointestinal cancer risk. Gastroenterology. 2017;153(4):910–23.

12. Galvan A, Ioannidis JP, Dragani TA. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. Trends Genet. 2010;26(3):132–41.
13. Lu Y, Kweon SS, Tanikawa C, Jia WH, Xiang YB, Cai Q, Zeng C, Schmit SL, Shin A, Matsuo K, et al. Large-scale genome-wide association study of East Asians identifies loci associated with risk for colorectal cancer. Gastroenterology. 2019;156(5):1455–66.
14. Song M, Chan AT, Sun J. Influence of the gut microbiome, diet, and environment on risk of colorectal cancer. Gastroenterology. 2020;158(2):322–40.
15. Yang J, Yu J. The association of diet, gut microbiota and colorectal cancer: what we eat may imply what we get. Protein Cell. 2018;9(5):474–87.
16. Jin Y, Wu S, Zeng Z, Fu Z. Effects of environmental pollutants on gut microbiota. Environ Pollut. 2017;222:1–9.
17. Chen HM, Yu YN, Wang JL, Lin YW, Kong X, Yang CQ, Yang L, Liu ZJ, Yuan YZ, Liu F, et al. Decreased dietary fiber intake and structural alteration of gut microbiota in patients with advanced colorectal adenoma. Am J Clin Nutr. 2013;97(5):1044–52.
18. Janney A, Powrie F, Mann EH. Host-microbiota maladaptation in colorectal cancer. Nature. 2020;585(7826):509–17.
19. Tsilimigras MC, Fodor A, Jobin C. Carcinogenesis and therapeutics: the microbiota perspective. Nat Microbiol. 2017;2:17008.
20. Song M, Chan AT. Environmental factors, gut microbiota, and colorectal cancer prevention. Clin Gastroenterol Hepatol. 2019;17(2):275–89.
21. Yang J, Wei H, Zhou Y, Szeto CH, Li C, Lin Y, Coker OO, Lau HCH, Chan AWH, Sung JJY, et al. High-fat diet promotes colorectal tumorigenesis through modulating gut microbiota and metabolites. Gastroenterology. 2022;162(1):135-149.e132.
22. Bishehsari F, Engen PA, Preite NZ, Tuncil YE, Naqib A, Shaikh M, Rossi M, Wilber S, Green SJ, Hamaker BR, et al. Dietary fiber treatment corrects the composition of gut microbiota, promotes scfa production, and suppresses colon carcinogenesis. Genes. 2018;9(2):102.
23. Tsuruya A, Kuwahara A, Saito Y, Yamaguchi H, Tsubo T, Suga S, Inai M, Aoki Y, Takahashi S, Tsutsumi E, et al. Ecophysiological consequences of alcoholism on human gut microbiota: implications for ethanol-related pathogenesis of colon cancer. Sci Rep. 2016;6:27923.
24. Biedermann L, Brülisauer K, Zeitz J, Frei P, Scharl M, Vavricka SR, Fried M, Loessner MJ, Rogler G, Schuppler M. Smoking cessation alters intestinal microbiota: insights from quantitative investigations on human fecal samples using FISH. Inflamm Bowel Dis. 2014;20(9):1496–501.
25. Sui X, Chen Y, Liu B, Li L, Huang X, Wang M, Wang G, Gao X, Zhang L, Bao X, et al. The relationship between KRAS gene mutation and intestinal flora in tumor tissues of colorectal cancer patients. Ann Transl Med. 2020;8(17):1085.
26. Liang S, Mao Y, Liao M, Xu Y, Chen Y, Huang X, Wei C, Wu C, Wang Q, Pan X, et al. Gut microbiome associated with APC gene mutation in patients with intestinal adenomatous polyps. Int J Biol Sci. 2020;16(1):135–46.
27. Kadosh E, Snir-Alkalay I, Venkatachalam A, May S, Lasry A, Elyada E, Zinger A, Shaham M, Vaalani G, Mernberger M, et al. The gut microbiome switches mutant p53 from tumour-suppressive to oncogenic. Nature. 2020;586(7827):133–8.
28. Global, regional, and national burden of colorectal cancer and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Gastroenterol Hepatol 2022, 7(7):627–647
29. Yoon HH, Shi Q, Alberts SR, Goldberg RM, Thibodeau SN, Sargent DJ, Sinicrope FA. Racial differences in BRAF/KRAS mutation rates and survival in stage iii colon cancer patients. J Natl Cancer Inst. 2015;107(10):djv186.
30. Guda K, Veigl ML, Varadan V, Nosrati A, Ravi L, Lutterbaugh J, Beard L, Willson JK, Sedwick WD, Wang ZJ, et al. Novel recurrently mutated genes in African American colon cancers. Proc Natl Acad Sci U S A. 2015;112(4):1149–54.
31. Wu S, Sun C, Li Y, Wang T, Jia L, Lai S, Yang Y, Luo P, Dai D, Yang YQ, et al. GMrepo: a database of curated and consistently annotated human gut metagenomes. Nucleic Acids Res. 2020;48(D1):D545-d553.
32. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The sequence read archive: a decade more of explosive growth. Nucleic Acids Res. 2022;50(D1):D387-d390.
33. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):R60.
34. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, Bork P, Sunagawa S, Zeller G. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol. 2021;22(1):93.
35. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, Jia W, Cai S, Zhao L. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. Isme j. 2012;6(2):320–9.
36. 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):1454–63.
37. Young C, Wood HM, Seshadri RA, Van Nang P, Vaccaro C, Melendez LC, Bose M, Van Doi M, Piñero TA, Valladares CT, et al. The colorectal cancer-associated faecal microbiome of developing countries resembles that of developed countries. Genome Med. 2021;13(1):27.
38. Long X, Wong CC, Tong L, Chu ESH, Ho Szeto C, Go MYY, Coker OO, Chan AWH, Chan FKL, Sung JJY, et al. Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity. Nat Microbiol. 2019;4(12):2319–30.
39. Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, Tang L, Zhao H, Stenvang J, Li Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut. 2017;66(1):70–8.
40. Baxter NT. Ruffin MTt, Rogers MA, Schloss PD: Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med. 2016;8(1):37.
41. Dhakan DB, Maji A, Sharma AK, Saxena R, Pulikkan J, Grace T, Gomez A, Scaria J, Amato KR, Sharma VK. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. Gigascience. 2019. https://doi.org/10.1093/gigascience/giz004.
42. Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, Tremaroli V, Bakker GJ, Attaye I, Pinto-Sietsma SJ, et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. Nat Med. 2018;24(10):1526–31.
43. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, Costea PI, Godneva A, Kalka IN, Bar N, et al. Environment dominates over host genetics in shaping human gut microbiota. Nature. 2018;555(7695):210–5.
44. Bramble MS, Vashist N, Ko A, Priya S, Musasa C, Mathieu A, Spencer A, Lupamba Kasendue M, Mamona Dilufwasayo P, Karume K, et al. The gut microbiome in konzo. Nat Commun. 2021;12(1):5371.
45. He Y, Wu W, Zheng HM, Li P, McDonald D, Sheng HF, Chen MX, Chen ZH, Ji GY, Zheng ZD, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med. 2018;24(10):1532–5.
46. Liu NN, Jiao N, Tan JC, Wang Z, Wu D, Wang AJ, Chen J, Tao L, Zhou C, Fang W, et al. Multi-kingdom microbiota analyses identify bacterial-fungal interactions and biomarkers of colorectal cancer across cohorts. Nat Microbiol. 2022;7(2):238–50.
47. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med. 2019;25(4):667–78.
48. 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):679–89.

Publisher's Note