VTwins: Identifying Robust Associations from High-dimensional Data of Limited Samples

Qingren Meng

Beijing Institute of Genomics

Qian Zhou

Center for Synthetic Biology Engineering Research, Shenzhen Institutes of Advanced Technology

Qin Ma

The Ohio State University <https://orcid.org/0000-0002-3264-8392>

Jun Yu

Beijing Institute of Genomics, Chinese Academy of Sciences

Jun Chen

Department of Liver Diseases, The Third People's Hospital of Shenzhen, The Second Affiliated Hospital of Southern University of Science and Technology

Yu Kang (✉ kangy@big.ac.cn)

Beijing Institute of Genomics <https://orcid.org/0000-0001-5196-0376>

Article

Keywords: causality, high-dimension data, algorithm, metagenome, paired cohort, sampling, machine learning

Posted Date: February 6th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-116301/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Robust associations are strong indicators for causalities but challenging for identification from high-dimensional datasets. In examples of metagenomic research where microbiota is highly complex and variable, low concordance between studies in identifying disease-causative microbes has become the main obstacle in the field. Here, we develop a simple method—Virtual Twins (VTwins)—for inferring robust associations, imitating the twins in genetic research. From the original groups, paired samples of distinct phenotypes but matched taxonomical profiles are selected to reconstruct a “Twin” cohort, where statistical significance is often achieved. In direct comparison to current methods by revisiting the largest meta-analysis metagenomic dataset, VTwins can 10-fold reduce the sample-size for recalling disease-associated microbes robustly across-datasets and constructing machine-learning models of the same accuracy level as pooled samples in predicting disease status. In practice, VTwins is straightforward, powerful, and versatile in handling highly variable and high-dimensional datasets, suggesting potentials in mining causalities in the Big-data Era.

Introduction

Identification of robust associations in observation studies is essential for causal insight, the primary goal of science. Observational data collected nowadays in the Big-data Era are often high-dimensional, highly variable, and highly correlated, from which inferring causality is still a big challenge due to the lack of effective algorithms in dealing with the random perturbations of features. For instance, in metagenomic studies, a microbiome often comprises hundreds of microbial species, a typical kind of high-dimensional data where the abundance of constituents is highly variable and highly cross-related. Fast-growing evidence has highlighted the association of human microbiota to a broad range of disease conditions^{1,2}, which makes microbiota-targeted therapies one of those most promising ways to control them effectively^{3,4}. However, identifying the target microbes contributing to disease pathogenesis is still a difficult task. The difficulties are attributed mainly to the high interindividual-diversities, which often overwhelms the disease-associated differences in contemporary cross-sectional studies^{5,6}. Associations of false-positives caused by confounding cofactors and false-negatives due to insufficient sample size are commonly introduced, making low concordance of disease-associated microbial features identified among studies^{7,8}.

Several path-forward ideas have been proposed to limit confounding factors in metagenome studies, but they are either not applicable for human research⁹ or cross-sectional study¹⁰, or hard-to-use requiring a large control cohort and as more as possible information of confounding factors to be collected⁸. Here, by virtually generating paired samples of similar taxonomic profiles but in distinct phenotypes, we have developed an easily applicable tool—Virtual Twins (VTwins), to control microbial features caused by non-relevant cofactors and evaluate its effects and applicability to the cross-sectional metagenomic datasets.

Results

The principle of VTwins

The idea of VTwins has been inspired by the success of twin studies in identifying disease-associated genetic variations^{11, 12}. In terms of data characteristics, monozygotic twin samples, which are almost identical genetically, provide perfect control against genetic heterogeneity for the identification of disease-causal variation loci. Similarly, in the highly diverse metagenomic constituents, samples can be regarded as twins if bearing highly similar taxonomical profiles. Toward this end, based on the taxonomic profiles of samples from the original dataset, we attempt to construct a VTwins cohort composed of “twins” samples of distinct phenotypes. Phenotype-associated microbe features can then be identified by applying statistical tests for paired samples (such as the Wilcoxon signed-rank test). Specifically, we first construct a high-dimension space where the abundance of each taxon (here, we used the species level abundance) represents a dimension. All samples are then plotted into the high-dimension space according to each species’ abundance, and pairwise Euclidean distances for all samples are calculated. Finally, samples with the nearest distance but in the opposite phenotypic groups are selected to constitute a new VTwins cohort (see Methods and Fig. 1). Upon this VTwins cohort, microbe features consistently enriched in samples of one phenotype over the other and passing statistical tests are regarded as phenotype-associated or candidate causal microbes.

The accuracy of VTwins

Here, we revisited the up-to-date largest meta-analysis dataset of colorectal cancer (CRC) that includes eight study cohorts and a total of 771 samples (case: control = 387:384; the sample size of single cohort ranges 53 ~ 128, **Data Table S1**), for which two research groups have separately performed meta-analyses and published their works at the same time^{13, 14}. Using the pooled samples, a total of 13 species were both identified as CRC-associated microbes by the two studies, *i.e.*, within the top 20 species of largest effect size in one study¹³ and with $FDR < 10^{-5}$ in the Wilcoxon rank-sum test in the other¹⁴ (Fig. 2).

To test the power of VTwins, we first reconstructed the VTwins cohorts for each of the eight datasets and then compared the significance of each species from the VTwins cohorts (tested by the Wilcoxon signed-rank test for paired samples) with those from the original cohorts (tested by the Wilcoxon rank-sum test for group cohort). VTwins identifies 285 species on average that are significantly enriched either in the CRC group or the control group ($FDR < 0.05$, **Data Table S2-S9** for each of the eight cohorts), among which ~ 140 species are highly significant ($FDR < 0.001$). All the 13 CRC-associated species exhibit significant differences in all the eight VTwins cohorts (except *P. uenonis* in the ITA1 cohort) and achieve high significance of $FDR < 10^{-5}$ in most cases (Fig. 2).

In contrast, when using the original cohorts, although the 13 CRC-associated species exhibited high significance in the U test of pooled samples, every single cohort can only recall 4.9 (range 0 ~ 11) of the 13 species on average (Fig. 2). We further applied the Boruta algorithm, the best-performing variable selection methods based on ensemble machine learning random forest for high-dimensional

metagenomic datasets. Although the Boruta algorithm is much more sensitive than the U test by calculating each species' permutation importance and identified much more CRC-associated species, neither a single cohort could recall all the 13 species, nor a CRC-associated species was consistently recalled from all cohorts (Fig. 2). Therefore, VTwins saliently increases the statistical power and sensitivity for identifying disease-associated microbes compared to analysis with original cohorts even with machine learning.

We then ranked the 13 CRC-associated species according to their FDR values in the test of each VTwins cohort and compared them to their ranks of permutation importance inferred by Boruta algorithm from corresponding original cohorts. The results indicated the ranks of species are more stable across VTwins cohorts than their importance ranks across original cohorts, especially for the top species. A seven-species core panel that was proposed to represent the characteristic deviations in CRC microbiota in the previous meta-analysis¹³, including *P. stomatis*, *F. nucleatum*, *Parvimonas spp.*, *P. asaccharolytica*, *G. morbillorum*, *C. symbiosum*, and *P. micra*, were just the top seven species with the highest significance in VTwins analysis (**Figure S1**). The result suggests that species with higher rank significance in VTwins analysis are more probable disease-associated.

From the species that showed significant differences in VTwins analysis using the pooled sample, we select the top-ranking 30 species (FDR < 10^{-75} , **Data Table S10**) as our selection of CRC-associated microbes. Our list covers all the 13 CRC-associated species identified by the previous two studies, except *A. obesiensis* (FDR < 10^{-62}), and the other species in our selection include ten enriched in CRC and eight enriched in controls (**Figure S2**). Among the ten CRC-enriched species, *Clostridium hathewayi*¹⁵, *Bacteroides fragilis*¹⁶, and *Peptostreptococcus anaerobius*¹⁷ are well-known to produce tumorigenic toxins; *Clostridium bolteae*¹⁸, *Granulicatella adiacens*¹⁹, and *Ruminococcus gnavus*^{20, 21} are able to drive the mucosal inflammation that preludes cancer; *Eggerthella lenta* and *Flavonifractor plautii*²² are potential in degrading various anticarcinogenic compounds such as flavonoids. Meanwhile, all the eight species enriched in controls are butyrogenic species that produce the critical anti-inflammatory butyrate in colon²³, and species *Faecalibacterium prausnitzii* and *Eubacterium hallii* also produce beta-glucuronidase (B-GUS) and glycerol/diol dehydratase (GDH), respectively, to biotransform many tumorigenic dietary heterocyclic amine²⁴. Therefore, most of the species in our selection based on VTwins analysis are able to mechanically contribute to promote or protect against CRC, *i.e.*, causative microbes of the disease.

The robustness of VTwins-directed predictive model

Machine learning has been widely used in evaluating the importance of each microbiota feature in diagnosis. The two previous meta-analyses of CRC have used random forest¹³ and lasso¹⁴, respectively, to construct CRC-diagnosis classifiers with all the species abundance as input, and evaluated their performances with intra- and cross-cohort prediction validation. As random forest model exhibits a better accuracy between the two studies, we apply it to construct classifiers for direct comparison. Here, we use

the abundance of the top species selected by VTwins instead of a full count of the species and compared their diagnostic values and transferability across cohorts to see if these top species are more relevant to disease than to the specific cohorts.

In intra-cohort cross-validation, we gradually increase the number of features from 10 to 50 top-ranking species in constructing the classifiers and observe that the top 30 species deliver the best performance, *i.e.*, an average of 0.89 ± 0.09 in the area under the receiver operating characteristic curve (AUROC) for all the eight cohorts (Fig. 3A). The performance of our models appears better than that of constructed based on all species, which is only 0.81 on average (refer to Fig. 2a in Ref¹³). We then performed the cross-cohort (study-to-study) validation for the model of each cohort, where the classifier is trained by a given dataset and test against each of the rest seven cohorts. Here we also used the 30 top features selected by VTwins instead of all species features, and the accuracy is obviously improved from the AUCs of 0.69 ± 0.10 (refer to Fig. 2a in Ref¹³) to an average AUC of 0.78 ± 0.06 across all our study-to-study validations (Fig. 3B). The better performance in intra-cohort tests and transferability in cross-cohort tests of the VTwins-directed classifiers implies that the species features selected by VTwins are more relevant to the pathogenesis of CRC other than dataset-dependent features.

To further confirmed the predictive value of VTwins selected species, we perform the leave-one-dataset-out (LODO) validations where classifiers are trained on seven datasets combined and validated on the left-out dataset, for each dataset in turn²⁵. This approach, although a seven-fold increase in the training dataset, the accuracy was only marginally improved, producing AUC = 0.81 ± 0.06 for all eight datasets when 30 top-ranking species are used compared to AUCs of 0.77 ± 0.06 in cross-study validations (Fig. 3B). We also noticed that altering the number of features used had almost no effects on the AUCs in LODO tests (**Figure S3**), and nearly the same as the performance using all species features, which yields 0.81 ± 0.08 AUCs (refer to the LODO analysis in Fig. 2a in Ref¹³). We hypothesize that the classifier in the LODO tests reaches its supreme accuracy; and the probable existence of intermediate samples precludes a clear boundary between case and control samples no matter the number of features used in prediction. However, the VTwins-directed models in study-to-study transfer, which produces AUCs of 0.78 ± 0.06 , approaches the supreme accuracy level, *i.e.*, 0.81 AUC, with only an averagely one-seventh training sets, indicating the robust associations between VTwins-selected features in every single cohort and microbe-related CRC pathogenesis.

Additionally, 22 species are consistently recalled in each LODO validation. When sorted according to the rank of their importance in contributing to the model, the ranks of each species keep stable across validations (Fig. 3C). Consistent with the previous meta-analyses, very few control-associated species achieve top significance in VTwins analysis and rank high in the learning models (Fig. 3C), further highlighting that successful discrimination is achieved by CRC-associated rather than control-associated microbial features. These results show that, while great sample diversity embedded in the meta-cohorts, cross-validated AUCs can be high and stable for the prediction of CRC in both intra-dataset and cross-study, indicating the robustness and transportability of species features selected by VTwins in

association to CRC and the power of VTwins in diminishing batch effects that often confound metagenomic studies.

The sample size required for VTwins

As VTwins is much more sensitive in identifying disease-associated species than conventional analysis, we next tried to figure out the minimal sample size required for VTwins. First, we performed a random selection of a given number of samples (equal in the number of case and control samples) from the pooled cohorts and repeated ten times. For each selection, VTwins analysis was conducted using the selected samples, and the top 30 species with the highest significance were regarded as candidate disease-associated features. When we raised the sampling size gradually from 30 to 400, we observed that more of the seven species in the core panel were recalled and reached a plateau when the sample size was more than 120 (CRC : control = 60 : 60). If we counted the top 50 species, the required sample size was reduced to > 80 for the recall of the core panel (Fig. 4A). Even for the 13 CRC-associated species identified by the two meta-analyses, most could be recalled with randomly selected > 80 samples when the top 50 features were selected (**Figure S4**).

Moreover, for each sampling, we constructed random forest models using both 30 and 50 top species selected by VTwins and trained the model with the selected samples five times, then tested the model with all the rest samples. When the sampling size reaches 80, no matter the number of features used, an AUC $\approx$ 0.8 can be obtained in most validations, similar to the accuracy in our study-to-study transfer validation. Further increasing sampling size to 400, the accuracy is only marginally improved (Fig. 4B). As such, 80 samples (case : control = 40 : 40) is able to recall the core panel of CRC-associated microbes and to accurately identify disease-associated features by selecting those with top significance, thus met the minimal requirement in sample size for an effective VTwins analysis, *i.e.*, about 10-fold decrease than that required for original group cohort and conventional analysis.

VTwins in functional analysis

Finally, we tested whether VTwins was valid in identifying the CRC-associated functional deviations in the microbiome, which is characterized as increased amino acid putrefaction²⁶ and reduced carbohydrate degradation²³. To identify the pathogenesis-related metabolic functions instead of the hitcher pathways embedded in the genomes of CRC-associated microbes, we reconstructed another “twin” cohort based on the function-profile distance by means of pathway abundance and identified 214 differential pathways with high significance (Wilcoxon signed-rank test, FDR < 10^{-5} , **Data Table S11**). Thirteen of the top 30 pathways of highest significance (FDR < 10^{-22}) overlapped with the 20 pathways of the largest effect size (> 0.35) in conventional LDA analysis in the previous meta-analysis¹³, and the function of these 13 common pathways are all consistent with the previous understanding of the CRC-associated metabolic deviations (Fig. 5A). The 13 pathways often exhibit significant differences in every single cohort in VTwins analysis but show no differences in the corresponding original cohorts (**Figure S5**), confirming the high sensitivity and statistical power of VTwins in functional analysis as well.

The other pathways in our selection besides the 13 common ones keep the concordance with the CRC-associated metabolism deviations, which include another pathway of L-glutamate-degradation (V) and the urea-cycle enriched in CRC samples and sucrose-degradation (III) enriched in controls (Fig. 5A,5B). Additionally, the VTwins also identifies significantly reduced pathway abundance in the biosynthesis of IMP (I, II, & III) (Fig. 5C) and CDP-DAG (I & II) (Fig. 5D). IMP is the precursor of the de novo biosynthesis of all kinds of purines, which process utilizes amino acids instead of catabolizing them into tumorigenic metabolites in CRC conditions, and CDP-DAG is a liponucleotide that has been found of antiproliferative activity in vitro along with its analogs²⁷. The conspicuous deficiency of their biosynthesis in CRC samples provides further understanding of the deviation in CRC metabolism, which deserves further investigations in their roles in protecting against CRC; and the discovery of these pathways highlights the value of VTwins in exploring novel molecular mechanisms in the research of disease pathogenesis.

Conclusion

In conclusion, VTwins provides a straightforward approach to control the highly variable microbiota profiles in cross-sectional studies. By direct comparison to the conventional cohort-based analyses, we demonstrate the power and sensitivity of VTwins in identifying robust disease-associated microbes, where statistical power is saliently increased by transforming the original cohort into the paired VTwins cohort. By a similar twin strategy based on the functional profiles of samples, we have successfully identified the impairment of detoxification in autistic children in a former study²⁸. The success of VTwins may rely on the virtual twins it provides as mutual controls for the complexity of all the dimensions, as an adequate number of controls is the premise for causal-relationship mining and is hard to be fully satisfied in the case of hundreds of dimensions. Therefore, the more dimensions the data have, the more effective VTwins becomes. Moreover, the strategy of virtual twins cohort can be applied to other high-diversity and high-dimension datasets given that pairwise distances among samples can be calculated, which will facilitate its broader application in other fields than metagenomics, including genomics, transcriptomics, epigenomics, and even other scientific fields.

Methods

Study design and data availability

The preprocessed metagenomic data of six cross-sectional studies of CRC were directly downloaded from the curatedMetagenomicData²⁹ R package supplemental to the meta-analysis research by Thomas, A. M., et al¹³. The data includes information of metadata, taxonomic and functional composition, and relative abundance of each species and metabolic pathway, in which taxonomy and pathway profiling were annotated using MetaPhlAn2³⁰ and HUMAnN2³¹, respectively. Since the preprocessed metagenomic data for the other two validation studies used in the meta-analysis were not available, we downloaded their raw fastq files from the DNA Data Bank of Japan database (project No. DRA006684³²) and European Nucleotide Archive (project No. PRJEB27928¹⁴), respectively. To keep the consistency with the

other six cohorts, we performed raw data preprocessing strictly as described in the meta-analysis¹³, including quality and reads filtered, MetaPhlAn2³⁰ annotated taxonomic profiling, and HUMAnN2³¹ annotated pathway profiling.

For short, the eight studies were represented by the name of the country where the cohorts were recruited: ThomasAM_2018a— Italy (ITA 1)¹³, ThomasAM_2018b— Italy (ITA 2)¹³, FengQ_2015— Austria (AUS)³³, VogtmanE_2016— the United States (USA)³⁴, YuJ_2015— China (CHI)³⁵, and ZellerG_2014— France (FRA)³⁶, project No. DRA006684— Japan (JAP)³², and project No. PRJEB27928— Germany (GEM)¹⁴. In total, the pooled samples contain 387 CRC cases and 384 healthy controls.

Virtual Twins

The purpose of “Virtual Twins” is to construct a “twin” cohort from the original group cohort. To find the “twin” samples of opposite phenotypes but have a very similar taxonomic profile, we first constructed a high-dimension space, in which each species represented a dimension. Then, each sample was plotted into the space according to the relative abundance of the species within it. Pairwise Euclidean distances were used to indicate the similarity between the two samples, and we calculated the total distance of a sample to its K-Nearest Neighbors of the same group (intragroup KNN), of the opposite group (intergroup KNN), and of the total cohort (KNN), where K was the square root of the number of samples in the corresponding group.

We noticed that some samples are even closer to neighbors of opposite phenotypes than those of their own, implying that samples from the opposite groups are extensively entangled in the space. To maximally represent the ensemble population, three steps were taken in the selection of samples for the “twin” cohort: 1) outlier samples that were too far from other samples ($KNN > \text{mean KNN} + SD$) were excluded first to avoid including them in the “twin” cohort; 2) redundancies that were too close to neighbors of the same phenotypes ($KNN < \text{mean intragroup KNN} - SD$) were identified and removed to avoid the impact of redundant samples on statistical tests; 3) samples closer to neighbors in the opposite group than neighbors of own side group ($\text{intergroup KNN} < \text{intragroup KNN}$) were selected as representative samples that represent the most entangling part of the ensemble cohort, and the number of the representative samples of the two sides were often comparable.

Finally, the “Twin” cohort was reconstructed with all the representative samples and their neighbors of the opposite group: one representative sample and one of its K neighbors constituted a “twin”. Then redundant “twins” composed of the same two samples were eliminated. Although representative samples may be utilized for more than one time, *i.e.*, at most K times, the representation is equal for all representative samples and comparative for the opposite groups, which will not deviate the statistics result. Therefore, the “twin” cohort was reconstructed by an equal number of samples from both sides that maximally matched with each other.

Random forest classifier and permutation importance

All machine learning experiments used random forest, as this algorithm has been shown to outperform, on average, other learning tools for microbiome data. The species with top significance tested in the VTwins analysis were explored for the prediction value of their relative abundances.

The model was constructed based on the R package caret (<https://cran.r-project.org/web/packages/caret/index.html>) and e1071 (<https://cran.r-project.org/web/packages/e1071/index.html>), where the trainControl function was used to select the best model.

Receiver operating characteristic (ROC) curve was used to evaluate the performance of the classifier through R packages ROCR (<https://cran.r-project.org/web/packages/ROCR/index.html>) and indicated by the area under the curve (AUROC). The importance of the mean decrease in the impurity of each species which indicates its contributions to the classifier was directly output from the caret R package.

To conduct a direct comparison, the intra-dataset prediction capability was measured through 20-repeat tenfold cross-validation, resulting in 200 validation folds for each cohort. Here, the createFolds function in the caret R package was used to stratify each fold to contain a balanced proportion of controls and disease samples. In the cross-study validation, datasets were considered two by two: one is used for training the model, the other to validate. The Leave-one-dataset-out (LODO) approach consists of five-repeat training of the model on the pooled samples from all cohorts except the one used for model testing. We iterated along with all the cohorts, performing a LODO validation on each set of samples.

To calculate the reliable permutation feature importance, we applied the Boruta approach, the best-performing variable selection method based on ensemble machine learning random forest for high-dimensional metagenomic datasets. Boruta package in R was used to perform feature selection with 99 iterations and a confidence level cut-off of 0.01 for the p values.

Determination of the minimal required sample size of VTwins

A random sampling test was conducted to determine the minimal sample size requirement of VTwins. In the test, all the samples from the eight cohorts were pooled, and for each random sampling, an equal number of cases and controls were randomly selected to constitute a random set. Using sampling R package (<https://cran.r-project.org/web/packages/sampling/index.html>), we conducted a raising series of sampling size as 30, 40, 60, 80, 100, 120, 140, 160, 180, 200, and 400 samples in each set. For each sampling size, 10-repeat random sampling was conducted which includes a half-half number of cases and controls.

Then we treat the random set as a group cohort and perform VTwins analysis for it, *i.e.*, construct a corresponding “twin” cohort for a random set and select the top-ranking species features of highest significance. These top features were then used to construct a random forest model and tested with all the rest samples of the entire sample pool except the testing random set.

Statistics

The two-sided Wilcoxon signed-rank test was used to identify significant statistical differences in species abundance in VTwins analysis, and the two-sided Mann–Whitney U test (Wilcoxon rank-sum test) was applied to test the significance of each species in the original group cohort. All P values were corrected for multiple hypothesis testing by the Benjamini–Hochberg method.

Data and code availability

No datasets were generated during the current study, and the authors declare that all other data analyzed to support the conclusions of this study are available within the paper supplementary information tables. The Model Twins pipeline version 1.0 is archived at <https://github.com/mengqingren1992/VTwinss>.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Declarations

Acknowledgements

Funding: This study was supported by the National Natural Science Foundation of China (31970568), National Science and Technology Major Project (2018ZX10712001-018-002), The National Key Research and Development Program of China (2016YFC0903800), and Programs of the Chinese Academy of Sciences (QYZDY-SSW-SMC017).

Author contributions: Y.K. conceived and designed the project. Q.Meng. completed all the bioinformatic and statistical analyses of the microbiota. Q.Meng., and Y.K. conducted data acquiring and processing as well as figure predations. Y.K., Q.Ma, and J.Y. draft and revised manuscript. All the authors have revised and approved the manuscript submission.

Competing interests: The authors declare no competing interests.

References

1. Janney A, Powrie F, Mann EH. Host-microbiota maladaptation in colorectal cancer. *Nature* **585**, 509-517 (2020).
2. McCarville JL, Chen GY, Cuevas VD, Troha K, Ayres JS. Microbiota Metabolites in Health and Disease. *Annual review of immunology* **38**, 147-170 (2020).

3. Renz H, Skevaki C. Early life microbial exposures and allergy risks: opportunities for prevention. *Nature reviews Immunology*, (2020).
4. Finlay BB, Goldszmid R, Honda K, Trinchieri G, Wargo J, Zitvogel L. Can we harness the microbiota to enhance the efficacy of cancer immunotherapy? *Nature reviews Immunology* **20**, 522-528 (2020).
5. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).
6. Falony G, *et al.* Population-level analysis of gut microbiome variation. *Science (New York, NY)* **352**, 560-564 (2016).
7. Fischbach MA. Microbiome: Focus on Causation and Mechanism. *Cell* **174**, 785-790 (2018).
8. Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y. Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448-454 (2020).
9. Surana NK, Kasper DL. Moving beyond microbiome-wide associations to causal microbe identification. *Nature* **552**, 244-247 (2017).
10. Plantinga AM, Chen J, Jenq RR, Wu MC. pldist: ecological dissimilarities for paired and longitudinal microbiome association analysis. *Bioinformatics (Oxford, England)* **35**, 3567-3575 (2019).
11. Zhu X, *et al.* Identification of functional cooperative mutations of SETD2 in human acute leukemia. *Nature genetics* **46**, 287-293 (2014).
12. Polderman TJ, *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics* **47**, 702-709 (2015).

13. Thomas AM, *et al.* Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature medicine* **25**, 667-678 (2019).
14. Wirbel J, *et al.* Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature medicine* **25**, 679-689 (2019).
15. Xia X, *et al.* Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. *Microbiome* **8**, 108 (2020).
16. Lopez LR, Bleich RM, Arthur JC. Microbiota Effects on Carcinogenesis: Initiation, Promotion and Progression. *Annual review of medicine*, (2020).
17. Long X, *et al.* *Peptostreptococcus anaerobius* promotes colorectal carcinogenesis and modulates tumour immunity. *Nature microbiology* **4**, 2319-2330 (2019).
18. Campisciano G, *et al.* The Obesity-Related Gut Bacterial and Viral Dysbiosis Can Impact the Risk of Colon Cancer Development. *Microorganisms* **8**, (2020).
19. Gaiser RA, *et al.* Enrichment of oral microbiota in early cystic precursors to invasive pancreatic cancer. *Gut* **68**, 2186-2194 (2019).
20. Yu S, *et al.* Paneth Cell-Derived Lysozyme Defines the Composition of Mucolytic Microbiota and the Inflammatory Tone of the Intestine. *Immunity* **53**, 398-416.e398 (2020).
21. Henke MT, Kenny DJ, Cassilly CD, Vlamakis H, Xavier RJ, Clardy J. *Ruminococcus gnavus*, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory

- polysaccharide. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 12672-12677 (2019).
22. Gupta A, *et al.* Association of *Flavonifractor plautii*, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. *mSystems* **4**, (2019).
23. Topping DL, Clifton PM. Short-chain fatty acids and human colonic function: roles of resistant starch and nonstarch polysaccharides. *Physiological reviews* **81**, 1031-1064 (2001).
24. Zhang J, *et al.* Gut microbial beta-glucuronidase and glycerol/diol dehydratase activity contribute to dietary heterocyclic amine biotransformation. *BMC microbiology* **19**, 99 (2019).
25. Riester M, *et al.* Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *Journal of the National Cancer Institute* **106**, (2014).
26. Gerner EW, Meyskens FL, Jr. Polyamines and cancer: old molecules, new understanding. *Nature reviews Cancer* **4**, 781-792 (2004).
27. Brachwitz H, Thomas Y, Bergmann J, Langen P, Berdel WE. New nucleoside-5'-alkylphosphonophosphates and related compounds containing 2'-deoxycytidine, thymidine and adenosine as nucleoside component. Syntheses and their effects on tumor cell growth in vitro. *Chemistry and physics of lipids* **87**, 31-39 (1997).
28. Zhang M, *et al.* A quasi-paired cohort strategy reveals the impaired detoxifying function of microbes in the gut of autistic children. *Science advances* **6**, eaba3760 (2020).
29. Pasolli E, *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nature methods* **14**, 1023-1024 (2017).

30. Truong DT, *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**, 902-903 (2015).
31. Franzosa EA, *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nature methods* **15**, 962-968 (2018).
32. Yachida S, *et al.* Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nature medicine* **25**, 968-976 (2019).
33. Feng Q, *et al.* Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature communications* **6**, 6528 (2015).
34. Vogtmann E, *et al.* Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PloS one* **11**, e0155362 (2016).
35. Yu J, *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70-78 (2017).
36. Zeller G, *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology* **10**, 766 (2014).

Figures

Figure 1

The scheme and sensitivity of VTwins. (A) The input of Virtual Twins is metagenomic data of samples in two phenotype groups. (B) Features of species abundance are acquired through metagenome analysis tools such as MetaPhlAn2, Kraken, and constitute a matrix across samples. (C) A high dimension space for all samples is constructed with each feature as a dimension. Twin samples with close distance from each of the two phenotypes are searched. (D) Features are identified as candidates of causal features of a phenotype if they pass test (Wilcoxon signed rank test) and exhibits significant enrichment in samples of this phenotype.

Figure 2

Heat map of importance or FDR value of each species in tests for each original and corresponding twin cohort. Each row indicates the FDR values (top and middle panel) in Wilcoxon signed rank test or importance (bottom panel) calculated by Boruta algorithm. The last row of each panel represents values for pooled samples. Colors indicates the range of FDR or the decision in Boruta algorithm. NA, absent in all samples of the cohort according to the taxonomic annotation.

Figure 3

The robustness of VTwins. (A) Prediction performance of intra-cohort cross-validation based on random forest model and built with significance-ranked species selected by VTwins. The cohorts are color-coded and AUROC is displayed as a function of increasing features. (B) Predicting power of the VTwins-directed models (the top 30 species) in cross-cohort validation, as evaluated by AUROC on study-to-study transfer (model trained with a single study and validated with other cohorts separately; average and standard

deviation shown by bar height and error bars, respectively, n=7) and LODO validation (model trained with combined cohorts except the single validation cohort). (C) Importance- ranked species in LDOD evaluations. Only microbial species selected by at least 7 LODO validations are reported. AUROC, area under the receiver operating characteristic curve; LODO, leave-one-dataset-out validation. NA, species absent in all samples of the cohort.

Figure 4

The minimal sample size required for effective VTwins. Equal number of case and control samples are randomly selected and used for VTwins analysis. As sample size gradually increases, the minimal requirement is determined by (A) the number of species in core CRC panel recalled by VTwins in the top 30 (left) or 50 (right) species; and (B) predicting power of random forest model based on abundance of the top 30 (left) or 50 (right) species.

Figure 5

Application of VTwins in functional analysis. (A) Volcano plots of differential pathways as calculated by DEseq2 in original (left) and twin (right) cohort of pooled samples. Fold change of the pathway abundance and Benjamini-Hochberg-adjusted P value (FDR) are plotted for each pathway. Differential pathways beyond the P value cutoff (10^{-3} for original cohort, 10^{-22} , i.e., the top-30 for twin cohort) are colored according to the group where they are enriched. In the right panel, only pathways other than those identified in the original cohort are labeled. B-D. The relative abundance of pathways: urea cycle (B), IMP biosynthesis (C), and CDP-DAG biosynthesis (D) in control (CTRL) and CRC samples.

Supplementary Files

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

- [ExtendedDataTables.xlsx](#)
- [SupplementaryFigures.docx](#)