Supplementary Information

June 3, 2020

Description of IBD metagenome study cohorts

Below we present a description of each of the six cohorts used in this metaanalysis. Each description is presented as was found in the original publication of each cohort.

iHMP (Lloyd-Price et al. 2019):

> Five medical centres participated in the IBDMDB: Cincinnati Children's Hospital, Emory University Hospital, Massachusetts General Hospital, Massachusetts General Hospital for Children, and Cedars-Sinai Medical Center. Patients were approached for potential recruitment upon presentation for routine age-related colorectal cancer screening, work up of other gastrointestinal (GI) symptoms, or suspected IBD, either with positive imaging (for example, colonic wall thickening or ileal inflammation) or symptoms of chronic diarrhoea or rectal bleeding. Participants could not have had a prior screening or diagnostic colonoscopy. Potential participants were excluded if they were unable to or did not consent to provide tissue, blood, or stool, were pregnant, had a known bleeding disorder or an acute gastrointestinal infection, were actively being treated for a malignancy with chemotherapy, were diagnosed with indeterminate colitis, or had undergone a prior, major gastrointestinal surgery such as an ileal/colonic diversion or j-pouch. Upon enrolment, an initial colonoscopy was performed to determine study strata. Subjects not diagnosed with IBD based on endoscopic and histopathologic findings were classified as 'non-IBD' controls, including the aforementioned healthy individuals presenting for routine screening, and those with more benign or non-specific symptoms. This creates a control group that, while not completely 'healthy', differs from the IBD cohorts specifically by clinical IBD status. Differences observed between these groups are therefore more likely to constitute differences specific to IBD, and not differences attributable to general GI distress.

PRJEB2054 (Qin et al. 2010):

> As part of the MetaHIT (Metagenomics of the Human Intestinal Tract) project, we collected faecal specimens from 124 healthy, overweight and obese individual human adults, as well as inflammatory bowel disease (IBD) patients, from Denmark and Spain.

SRP057027 (Lewis et al. 2015):

> Children and young adults less than 22 years of age were enrolled at the time of initiation of EN or anti-TNF therapy for treatment of active CD (defined as the Pediatric Crohn's Disease Activity Index [PCDAI] >10) at The Hospital for Sick Children in Toronto, ON, Canada; IWK Health Centre, Halifax, NS, Canada; and the Children's Hospital of Philadelphia, Pennsylvania. Participants in this observational cohort study were prescreened for eligibility and recruited from clinic or during inpatient hospitalization. Exclusion criteria included presence of an ostomy, treatment with probiotics within 2 weeks of initiating EN, treatment with anti-TNF therapy within 8 weeks of starting EN, or treatment with EN within 1 week of initiating anti-TNF therapy. The study protocol was approved by the institutional review boards at all participating institutions. Informed consent was obtained from all young adults and the parents/guardians of children less than 18 years of age.

PRJNA385949 (Hall et al. 2017):

> Samples from the PRISM study, collected at Massachusetts General Hospital: A subset of the PRISM cohort was selected for longitudinal analysis. A total of 15 IBD cases (nine CD, five UC, one indeterminate colitis) were enrolled in the longitudinal stool study (LSS). Three participants with gastrointestinal symptoms that tested negative for IBD were included as a control population. Enrollment in the study did not affect treatment. Stool samples were collected monthly, for up to 12 months. The first stool sample was taken after treatment had begun. Comprehensive clinical data for each of the participants was collected at each visit. At each collection, a subset of participants were interviewed to determine their disease activity index, the Harvey-Bradshaw index for CD participants and the simple clinical colitis activity index (SCCAI) for UC participants. Samples collected at Emory University: To increase the number of participants in our analysis, a subset of the pediatric cohort STiNKi was selected for whole metagenome sequencing including five individuals with UC and nine healthy controls. All selected UC cases were categorized as non-responders to treatment. Stool samples were collected approximately monthly for up to 10 months. The first sample from participants in the STiNKi cohort is before treatment started, and subsequent samples are after treatment started. Stool collection and DNA extraction methods are detailed in Shaw et al.

PRJNA400072 (Franzosa et al. 2019):

> PRISM cohort description and sample handling: PRISM is a referral centre-based, prospective cohort of IBD patients; 161 adult patients (>18 years old) enrolled in PRISM and diagnosed with CD, UC, and non-IBD (control) were selected for this study, with diagnoses based on standard endoscopic, radographical and histological criteria. The PRISM research protocols were reviewed and approved by the Partners Human Research Committee (re. 2004-P-001067), and all experiments adhered to the regulations of this review board. PRISM patient stool samples were collected at the MGH gastroenterolgy clinic and stored at -80C before DNA was extracted.
>
> Validation cohort description and sample handling: The validation cohort consisted of 65 patients enrolled in two distinct studies form the Netherlands; 22 controls were enrolled in the LifeLines DEEP general pipulation study and 43 patients with IBD were enrolled in a study at the Department of Gastroenterology and Hepatology at the University Medical Center Groningen. Patients enrolled in both studies collected stool using the same protocol: a single stool sample was collected at home and then frozen within 15 min in a conventional freezer. A research nurse visited all participants at home to collect home-frozen stool samples, which were then transported and stored at -80C. The stool samples were kept frozen before DNA was extracted.

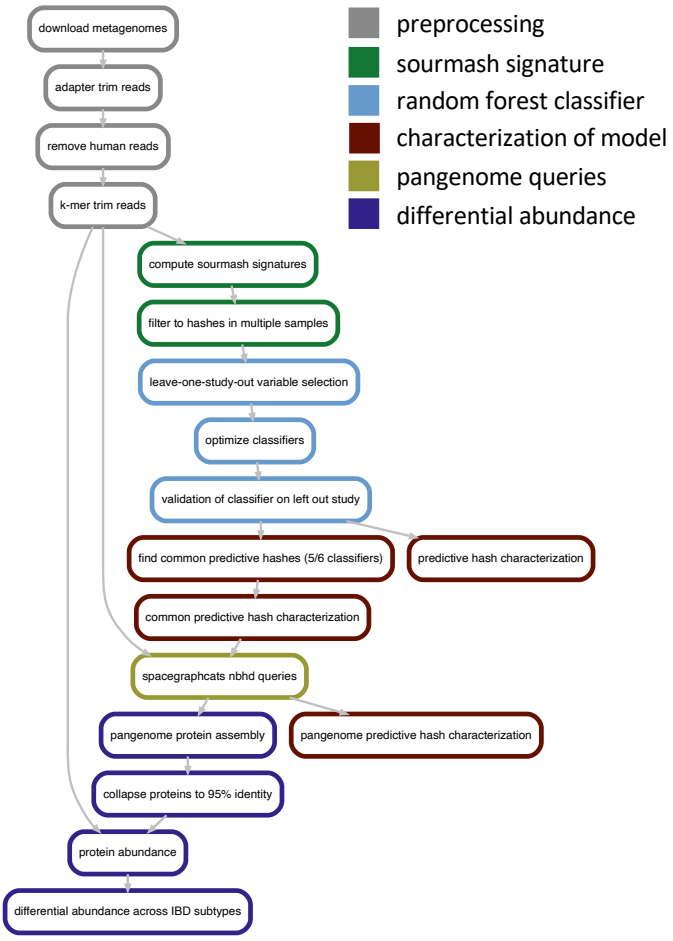
PRJNA237362 (Gevers et al. 2014):

> A total of 447 children and adolescents (<17 years) with newly diagnosed CD and a control population composed of 221 subjects with noninflammatory conditions of the gastrointestinal tract were enrolled to the RISK study in 28 participating pediatric gastroenterology centers in North America between November 2008 and January 2012.

Overview of pipeline

Construction of human microbiome metagenome assembled genome databases

While GenBank contains hundreds of thousands of isolate and metagenome-assembled genomes, we augmented the number of genomes by creating sourmash databases for all medium- and high-quality metagenome-assembled genomes from three recent human microbiome metagenome *de novo* assembly efforts (Pasolli et al. 2019; Nayfach et al. 2019; Almeida et al. 2019). The databases are available at in the OSF respository, "Comprehensive Human Microbiome Sourmash Databases" at the URL https://odf.io/hza89/. While we are aware that contamination in both GenBank and from these studies could introduce contamination into our



Simplified directed acyclic graph of the steps used in our pipeline, color coded by the section of the pipeline each step corresponds to. The steps in blue were performed six times, each time with a different validation study.

analysis, we reasoned that the increase we observed in identifiable hashes when we did not restrict ourselves to RefSeq was worth the trade.

To generate the databases, we downloaded the medium- and high-quality metagenome-assembled genomes and used sourmash `compute` with parameters `k 21,31,51`, `--track-abundance`, and `--scaled 2000`. We then used sourmash `index` to generate databases for `k = 31`. Below we detail the contents of each database.

- Pasolli et al. (2019): contains 70,178 high- and 84,545 medium-quality MAGs assembled from 9,428 human microbiome samples. Samples originate from stool (7,783), oral cavity (783), skin (503), vagina (88), and maternal milk (9). Original Data Download: http://segatalab.cibio.unitn.it/data/Pasolli_et_al.html
- Almeida et al. (2019): contains 40,029 high- and 65,671 medium-quality MAGs assembled from 11,850 human microbiome samples. All samples originate from stool. Original Data Download: ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/umgs_analyses/mags-gut_qs50.tar.gz
- Nayfach et al. (2019): contains 24,345 high- and 36,319 medium-quality MAGs assembled from 3,810 human gut microbiome samples. Original Data Download: https://github.com/snayfach/IGGdb

41 genome accessions and taxonomy

Available for download https://osf.io/ungza/

Supplementary References

Almeida, Alexandre, Alex L Mitchell, Miguel Boland, Samuel C Forster, Gregory B Gloor, Aleksandra Tarkowska, Trevor D Lawley, and Robert D Finn. 2019. "A New Genomic Blueprint of the Human Gut Microbiota." *Nature* 568 (7753): 499.

Franzosa, Eric A, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J Haiser, Stefan Reinker, Tommi Vatanen, et al. 2019. "Gut Microbiome Structure and Metabolic Activity in Inflammatory Bowel Disease." *Nature Microbiology* 4 (2): 293.

Gevers, Dirk, Subra Kugathasan, Lee A Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, et al. 2014. "The Treatment-Naive Microbiome in New-Onset Crohn's Disease." *Cell Host & Microbe* 15 (3): 382–92.

Hall, Andrew Brantley, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K Lagoudas, et al. 2017. "A Novel Ruminococcus Gnavus Clade Enriched in Inflammatory Bowel Disease Patients." *Genome Medicine* 9 (1): 103.

Lewis, James D, Eric Z Chen, Robert N Baldassano, Anthony R Otley, Anne M Griffiths, Dale Lee, Kyle Bittinger, et al. 2015. "Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease." *Cell Host & Microbe* 18 (4): 489–500.

Lloyd-Price, Jason, Cesar Arze, Ashwin N Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W Poon, Elizabeth Andrews, et al. 2019. "Multi-Omics of the Gut Microbial Ecosystem in Inflammatory Bowel Diseases." *Nature* 569 (7758): 655.

Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S Pollard, and Nikos C Kyrpides. 2019. "New Insights from Uncultivated Genomes of the Global Human Gut Microbiome." *Nature* 568 (7753): 505.

Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. "Extensive Unexplored Human Microbiome Diversity Revealed by over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle." *Cell* 176 (3): 649–62.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. "A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing." *Nature* 464 (7285): 59.