Inflammatory Bowel Disease Transcriptome and Metatranscriptome Meta-Analysis (IBD TaMMA) framework

Luca Massimino
Humanitas University https://orcid.org/0000-0003-3975-9148
Luigi Lamparelli
Humanitas University
Yashar Houshyar
Humanitas Clinical and Research Center
Silvia D'Alessio
PhoenixLAB s.r.l.s.
Laurent Peyrin-Biroulet
Inserm Ngere and Nancy University Hospital
Stefania Vetrano
Humanitas University
Silvio Danese
Humanitas University
Federica Ungaro (federica.ungaro@humanitasresearch.it)
Humanitas University https://orcid.org/0000-0001-5395-7795

Brief Communication

Keywords: inflammatory bowel disease (IBD), gut disorders, aetiogenesis

Posted Date: May 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-478844/v1

Version of Record: A version of this preprint was published at Nature Computational Science on August 20th, 2021. See the published version at https://doi.org/10.1038/s43588-021-00114-y.

Abstract

Inflammatory bowel disease (IBD) is a class of chronic inflammatory gut disorders whose aetiogenesis is still unknown. Despite the high number of omics studies, the RNA sequencing (RNA-Seq) data produced for a better IBD pathogenesis understanding cannot be compared because of the experimental variability and different data analysis approaches. To overcome this hurdle, we here introduce the open-source IBD Transcriptome and Metatranscriptome Meta-Analysis (TaMMA) framework, a comprehensive survey of publicly available IBD RNA-Seq datasets. IBD TaMMA will expedite the profiling of the IBD-associated transcriptome and metatranscriptome, holding out the strong promise of being of high impact for the IBD community.

Main

Inflammatory bowel disease (IBD), including Ulcerative Colitis (UC) and Crohn's Disease (CD), is a class of multifaceted chronic inflammatory gut disorders characterized by an uncontrolled, resolution-failing inflammation that leads to bowel demages[1]. Many efforts have been made to better comprehend IBD etiogenesis, so that, in the last decades, advances in computational biology have boosted preclinical and clinical research studies[2], generating a high number of RNA sequencing (RNA-Seq) data. Nevertheless, because of the experimental variability and different data analysis approaches, no comparison among the studies has been possible so far. To exploit the efforts made over the years by IBD experts in the field of Next Generation Sequencing (NGS), we here introduce the first meta-analysis web app, the IBD Transcriptome and Metatranscriptome Meta-Analysis (TaMMA) platform (https://ibd-meta-analysis.herokuapp.com/; link is temporarily kept private for the revision process; username: danese; password: steam). IBD TaMMA is a comprehensive survey of publicly available RNA-Seq datasets from IBD-derived and control samples across different tissues, all analyzed with the same pipeline and batch corrected for data harmonization and simultaneous comparison among the different studies. By exploiting this tool, the scientific community will benefit from a user-friendly, open-source platform where the profiling of the IBD-associated transcriptome and metatranscriptome will become quicker and statistically more powerful, also holding out the strong promise of being of high impact for the IBD community.

Results

Different meta-analyses of IBD patient gene expression profiles from microarray datasets have already identified dysregulation of expression of genes encoding for several inflammatory factors and RNA-binding proteins[3–5]. Nevertheless, these studies focused on a limited number of genes, lacking not only the whole-transcriptome but also metatranscriptome profiling, the latter newly emerged as successful to uncover novel gut-populating microbial entities[6].

Therefore, to provide a wider picture of the whole-transcriptome and metatranscriptome at different tissue/cell levels in both UC and CD patients, publicly available RNA-Seq datasets were collected and

analyzed. Being 26 independent studies, we predicted an experiment-dependent bias which was counteracted through a well-established batch correction algorithm, in accordance with source and tissue of origin. Of note, we also tried to batch correct the different library construction strategies utilized, but their variance was already fully explained by the source study. The meta-analysis performed was used as the core to design the IBD TaMMA web app (**Supplementary Fig. 1a**), which allows quick access to differential gene expression and gene ontology functional enrichment results, among the different conditions (https://ibd-meta-analysis.herokuapp.com/; full guide available at https://ibd-tamma.readthedocs.io/). Sample dispersion within the UMAP, easily attainable through the IBD TaMMA platform, shows clustering in accordance with the tissue of origin but not with the source study (Fig. 1a and Fig. 1b), indicating successful data harmonization. Consistently, housekeeping gene expression levels were found to be comparable across the different tissues and conditions (**Supplementary Fig. 1b**).

IBD TaMMA pinpoints a strong differential gene expression among UC, CD, and healthy (control) groups in the ileum, colon, and rectum, as shown in Fig. 1c. Of note, IBD-specific proinflammatory signatures were confirmed. Indeed, dysregulations between IBD-derived intestinal and the healthy tissues in the expression levels of Tumor Necrosis Factor-alpha (*TNF*), Interferon-gamma (*IFNG*), Interleukin 12B (*IL12B*), Integrin alpha 4 (*ITGA4*), Integrin beta 7 (*ITGB7*), already known as drivers of chronic inflammation and thus exploited as therapeutic targets for IBD patients[7], were confirmed (Fig. 1d and **Supplementary Fig. 1c**). Likewise, *S100A8* and *A9* transcripts encoding for the two subunits of the fecal biomarker calprotectin[8] and the recently emerged *S100A12*[8] were increased in intestinal samples from CD and UC as compared to the healthy (Fig. 1e and Fig. 1f). These results are well in line with most of the studies reporting these molecules as biomarkers of inflammation in patients with IBD[8].

Metatranscriptomics performed on IBD and healthy stools paralleled previous metagenomics analysis, confirming the *Bacteroidetes* and *Firmicutes phyla*, followed by the *Actinobacteria* and *Proteobacteria*, as the main colonizers of the fecal microbiota[9] (Fig. 1g, upper bars). IBD TaMMA also highlighted IBD and healthy intestinal samples to be colonized by the same *phyla*, although with different proportions (Fig. 1g, lower bars). Moreover, the decreased intestinal microbiota diversity, a well-known feature of IBD pathogenesis[1], was confirmed in IBD stools as compared to the healthy (Fig. 1h), paralleled by the decreased diversity also in colon and ileum from UC and in the colon from CD (Fig. 1i and Fig. 1j). Interestingly, the CD ileum showed increased microbiota diversity compared to the other groups (Fig. 1j), providing a novel insight into the disease location-dependent microbiota composition in CD patients. Of note, the IBD TaMMA also confirmed virome dysbiosis with the expansion of *Caudovirales* in both pediatric IBD and UC samples[10,11] (Fig. 1k) as well as the increased levels of *Herpesviridae* family in IBD-derived samples and of the *Hepadnaviridae* family in UC ileum as compared to the healthy, as previously reported[12,13] (**Supplementary Fig. 1d** and **Supplementary Fig. 1e**).

It is noteworthy that during the analysis most of the human unmapped reads failed to be classified by metatranscriptomics profiling and therefore were considered as *NGS dark matter* (**Methods**). Although its analysis goes beyond the scope of this work, a dedicated submission was done as we think these data

can also contribute to the understanding of gene and microbial entities not yet known but that may be the aim of future investigations (i.e., discoveries of new microbial entities).

Conclusively, altogether these pieces of evidence establish the IBD TaMMA as a reliable platform confirming well-known features of IBD pathogenesis and resulting in a useful open-source tool for developing further insights into IBD pathogenesis.

Discussion

Currently, NGS and multi-omics approaches are accelerating IBD investigations, allowing researchers to simultaneously explore disease intricacy from many points of view[2]. Nevertheless, although numerous whole-transcriptome analyses of IBD samples have been performed so far, no integrative study has been finalized, thus dispersing both insights and evidence often because of the weak statistical power.

IBD TaMMA introduces the first integrative analysis of all IBD-related publicly available RNA-Seq datasets, computationally uniforming their batch differences due to obvious experimental variability. IBD TaMMA validates specific pro-inflammatory and microbiota signatures confirming what is already well-established in the field and therefore proving to be a reliable platform of IBD transcriptomic and metatranscriptomic investigations. IBD TaMMA is easily exploitable thanks to a guide that can drive its users, finally resulting in an open-access resource that will expedite future studies, also generating new hypotheses and novel insights. Conclusively, the whole IBD community will benefit from TaMMA in the immediate future as a key web app for data analysis.

Methods

The full methodology is available in the Supplementary material.

Meta-analysis

FASTQ reads from 3,853 RNA-Seq data from different tissues, namely ileum, colon, rectum, mesenteric adipose tissue, peripheral blood, and stools, were mined from NCBI GEO/SRA and passed the initial quality filter. FASTQ files that passed the initial quality filter (see metadata table in TaMMA, https://ibd-meta-analysis.herokuapp.com/; link is temporarily kept private for the revision process; username: **danese**; password: **steam**) were mapped to the human reference genome and initial gene quantification was performed. Since these data came from 26 different studies made in different laboratories, we counteract the presumptive bias through a batch correction in accordance with source and tissue of origin. Once the gene counts were adjusted, samples were divided into groups in accordance with the tissue of origin and patient condition prior to differential expression analysis and gene ontology functional enrichment. Finally, the reads failing to map to the human genome were subjected to metatranscriptomics profiling by taxonomic classification using exact *k*-mer matching either archaeal, bacterial, eukaryotic, or viral genes.

Web app design

The web app was developed with Dash and Plotly, stored in GitHub, and running in Heroku. The complete guide on how to use the TaMMA web app can be found at https://ibd-tamma.readthedocs.io/.

Declarations

Data availability

Relevant datasets mentioned in this article are available in the *summary* and *metadata* tabs within the IBD TaMMA web app (https://ibd-meta-analysis.herokuapp.com/; the link is temporarily kept private for the revision process; username: danese; password: steam).

The *results described in the manuscript* are available in the *literature tab* within the TaMMA web app.

The web app *underlying code* and *data* are available at https://github.com/Humanitas-Danese-s-omics/ibd-meta-analysis and https://github.com/Humanitas-Danese-s-omics/ibd-meta-analysis-data, respectively.

The *complete guide* on how to use the TaMMA web app can be found at https://ibd-tamma.readthedocs.io/.

The IBD TaMMA NGS dark matter can be found at https://dataverse.harvard.edu/dataverse/tamma-dark-matter.

Ethical consideration

Please, refer to the original articles for the ethical approval of the human studies mentioned in this manuscript.

Acknowledgments

The authors would like to acknowledge the “Fondazione Cariplo per la ricerca Biomedica” under the grant agreement #2018-0112 to FU; “Fondazione AMICI ONLUS” for the research prize 2020 to FU; INNOVATIVE MEDICINES INITIATIVE (IMI) 2020-2024 (NUMBER 853995 — ImmUniverse) to SD and SV.

Author contributions

LM and FU: conceptualization and writing original draft; LM, LAL, YH: formal analysis, investigation; SDA, LPB, SV, SD, and FU review and editing; SV, SD, and FU supervision, resources, and funding acquisition.

Competing interest statement

SD has served as a speaker, consultant, and advisory board member for Schering-Plough, Abbott (AbbVie) Laboratories, Merck and Co, UCB Pharma, Ferring, Cellerix, Millenium Takeda, Nycomed,

Pharmacosmos, Actelion, Alfa Wasserman, Genentech, Grunenthal, Pfizer, AstraZeneca, Novo Nordisk, Vifor, and Johnson and Johnson. The other authors declare no conflicts of interest.

References

1. Aldars-García, L., Marin, A. C., Chaparro, M. & Gisbert, J. P. The Interplay between Immune System and Microbiota in Inflammatory Bowel Disease: A Narrative Review. *Int. J. Mol. Sci.* **22**, (2021).
2. Seyed Tabib, N. S. *et al.* Big data in IBD: big progress for clinical practice. *Gut* **69**, 1520–1532 (2020).
3. Li, X. *et al.* Meta-Analysis of Expression Profiling Data Indicates Need for Combinatorial Biomarkers in Pediatric Ulcerative Colitis. *J. Immunol. Res.* **2020**, 8279619 (2020).
4. Naz, S. *et al.* Transcriptome meta-analysis identifies immune signature comprising of RNA binding proteins in ulcerative colitis patients. *Cell. Immunol.* **334**, 42–48 (2018).
5. Vennou, K. E., Piovani, D., Kontou, P. I., Bonovas, S. & Bagos, P. G. Multiple outcome meta-analysis of gene-expression data in inflammatory bowel disease. *Genomics* **112**, 1761–1767 (2020).
6. Ungaro, F., Massimino, L., D'Alessio, S. & Danese, S. The gut virome in inflammatory bowel disease pathogenesis: From metagenomics to novel therapeutic approaches. *United European Gastroenterol. J.* **7**, 999–1007 (2019).
7. Argollo, M., Kotze, P. G., Kakkadasam, P. & D'Haens, G. Optimizing biologic therapy in IBD: how essential is therapeutic drug monitoring? *Nat. Rev. Gastroenterol. Hepatol.* **17**, 702–710 (2020).
8. Liu, F., Lee, S. A., Riordan, S. M., Zhang, L. & Zhu, L. Global studies of using fecal biomarkers in predicting relapse in inflammatory bowel disease. *Front Med (Lausanne)* **7**, 580803 (2020).
9. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
10. Fernandes, M. A. *et al.* Enteric virome and bacterial microbiota in children with ulcerative colitis and crohn disease. *J. Pediatr. Gastroenterol. Nutr.* **68**, 30–36 (2019).
11. Zuo, T. *et al.* Gut mucosal virome alterations in ulcerative colitis. *Gut* **68**, 1169–1179 (2019).
12. Wang, W. *et al.* Metagenomic analysis of microbiome in colon tissue from subjects with inflammatory bowel diseases reveals interplay of viruses and bacteria. *Inflamm. Bowel Dis.* **21**, 1419–1427 (2015).
13. Ungaro, F. *et al.* Metagenomic analysis of intestinal mucosa revealed a specific eukaryotic gut virome signature in early-diagnosed inflammatory bowel disease. *Gut Microbes* **10**, 149–158 (2019).

Figures

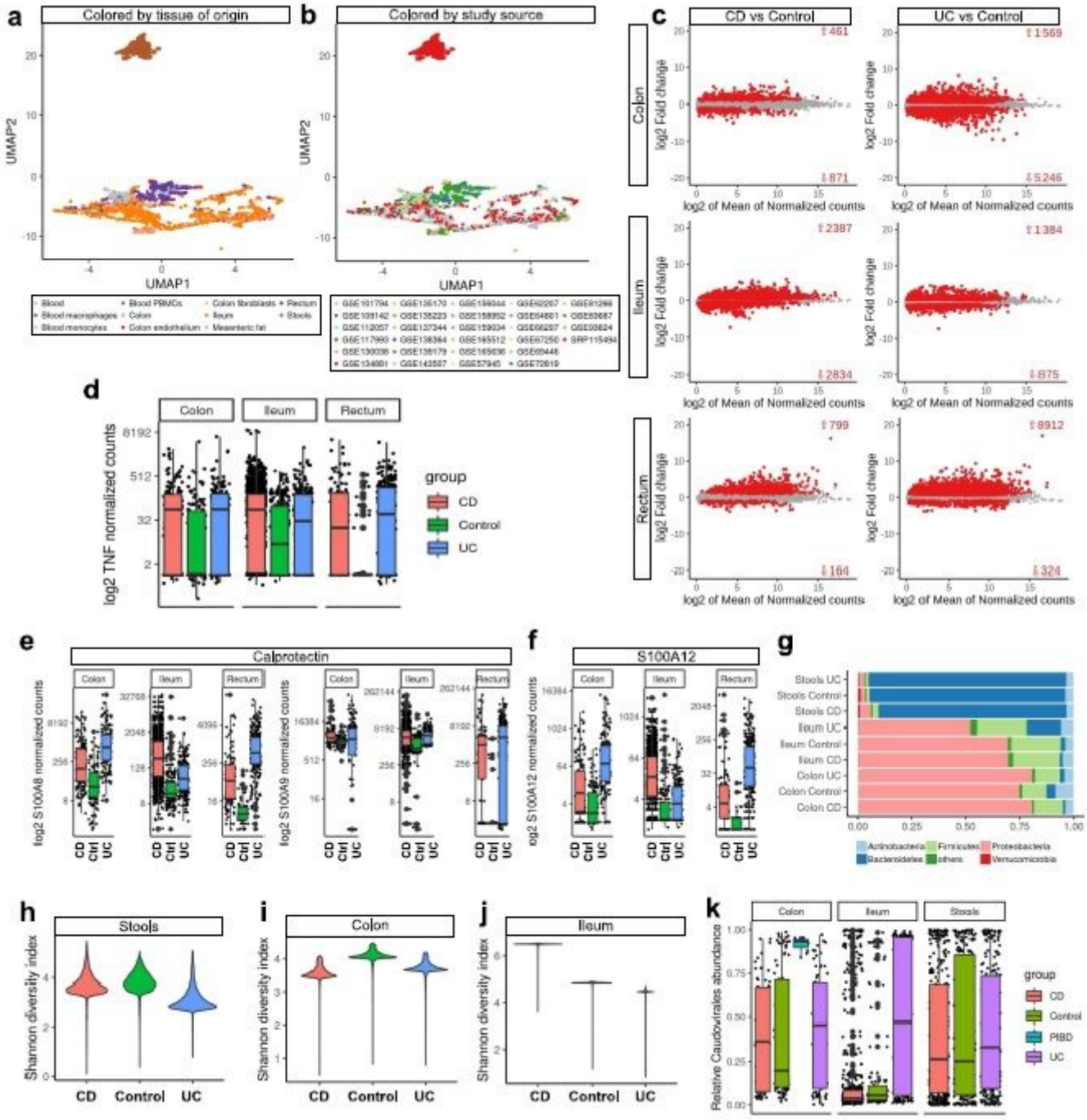

Figure 1

IBD TaMMA confirms IBD-specific signatures. (a,b) Multidimensional scaling of human whole-transcriptome by UMAP from UC, CD patients, and healthy (control) subjects, showing clustering in accordance with the tissue (a), but not with the data source of origin (b). (c) MA plots showing the differential gene expression results in the indicated comparisons. Red dots represent genes being differentially expressed with high statistical significance (FDR<1e-10). (d) Box plots showing differential TNF normalized expression among UC, CD, and healthy ileum, colon, and rectum. (e,f) Box plots showing

differential calprotectin (S100A8 and S100A9) (e) and S100A12 (f) encoding gene normalized expression among UC, CD, and healthy (ctrl) ileum, colon, and rectum. (g) Bar plots showing the relative abundance of the indicated bacterial phyla in stools, ileum, and colon from UC and CD patients, and healthy subjects (control). (h-j) Violin plots showing Shannon diversity indices among CD, healthy (control), and UC in stools (h), colon (i), and ileum (j). (k) Box plots showing relative Caudovirales order abundance in colon, ileum, and stools from healthy subjects (control) and CD, pediatric (P) IBD, and UC patients. All box plots represent sample distribution with the median, min, max, first, and third quartiles. An interquartile range of 1.5 has been used to define outliers. For the statistics, please refer to Supplementary Fig. 2.

Supplementary Files

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

- SupplementaryMaterial.docx
- SupplementaryInformation.pdf
- flatUngarospc.pdf
- flatUngaroepc.pdf