Comparative assessment of biobank datasets for the study of inflammatory bowel disease

Bri Nett[1], Nicholas Beckloff[1], Jack Wimberley[2], Chris Hargarten[1], Javad Zahiri[2], Matt Peterson[2], Ashley Faber[1], Srikant Sarangi[2], and Barry Wark[1]

[1] Ovation.io, 2 Union Street, Portland ME 04101 [2] Paradigm4, 281 Winter Street, Suite 360, Waltham MA 02451 & Affiliations

P129

INTRODUCTION

Lack of available multiomics datasets has slowed progress in understanding the relationship between genomic variants, gene expression, and disease progress in inflammatory bowel disease (IBD). Biobanks provide an opportunity to produce population-scale datasets to accelerate understanding of these diseases. Ovation's IBD Omics Data is a multiomics dataset including whole genome sequencing (WGS) and RNA sequencing (RNA-seq) from inflamed and non-inflamed tissues, derived from clinical biopsies of patients with IBD. Omics data is supplemented with patient phenotypic data including diagnoses, therapies, procedures, and routine laboratory results. We report initial results of analysis of the first 212 patients from this cohort to identify differentially expressed genes (DEGs) in patients with IBD, and to explore variants in these genes in both the Ovation IBD Omics Data and UK Biobank whole exome sequencing (WES) data.

METHODS

Cohort construction
Ovation IBD Omics Data is a commercially licensable dataset diagnosed with the following ICD-10 codes: K50 - Crohn's disease, K51 - Ulcerative colitis, K50 & K51 - CD & UC, or K52 - Other and unspecified noninfective gastroenteritis and colitis. Cases were selected from Ovation's biobank based on the availability of inflamed and non-inflamed tissue from gastrointestinal biopsies, and linked phenotype data including diagnoses, therapy exposures, procedures, and relevant routine laboratory results.

Omics
DNA and RNA were extracted from fixed formalin paraffin embedded (FFPE) tissue samples of 212 IBD patients by isotachophoresis. Inflamed and non-inflamed tissues were identified by pathologist review of tissue blocks. For WGS, 275–325 million 150bp paired-end reads of genomic DNA from non-inflamed tissue were acquired resulting in depths of coverage 15–30X. For RNA-seq, 30 million 150bp paired-end reads of ribose-free RNA were acquired from inflamed and non-inflamed tissue for each patient.

Comparison Datasets
UK Biobank whole exome sequencing capture experiment data was provided by Paradigm4. DEG lists were collected from previously published studies: Paired inflamed vs non-inflamed snap-frozen mucosa from 171 IBD patients from Hu et al., 2021; Inflamed ulcerative colitis vs noninflammatory bowel disease controls based on meta-analysis of eight datasets from biopsy tissue Linggi et al., 2021; and IBD (treatment included "infliximab") vs healthy control from Kaddoura et al., 2023 based on a meta-analysis of two microarray datasets from colonic mucosal biopsies (GSE14580 & GSE73661).

Variant Analysis
DNA FASTQ files were partitioned and aligned to the GRCh38.p14 reference with bwa-mem2 (Vasimuddin, 2019). Variants were called with GATK (McKenna, 2010), then these were matched to known variants within the dbSNP database with assigned RefSNP ID numbers (rsIDs), and finally all variants were stored in Ovation's array-based variant store in Variant Call Format (VCF). VCFs were ingested into the Paradigm4 REVEAL platform along with phenotypic data variables. Paradigm4 filtered Ovation's variants based on the same target regions used for UK Biobank WES (~39 Mbp; Resource 3803) so the two variant sets could be accurately compared. Minor alleles (i.e. the 2nd most common allele in a measured population) were identified and their frequencies (MAFs) were calculated within the Ovation cohort and the UK Biobank cohort.
Annotations were generated using the Ensembl Variant Effect Predictor.
Differential Gene Expression Analysis
Ribose-free RNA FASTQs were aligned with the STAR aligner and counts were computed using FeatureCounts. Normalization followed by differential expression analysis was performed using PyDESeq2 (Muzellec, 2022).

Acknowledgements
Thank you to the operations, product, and engineering teams at Ovation.io and Paradigm4, and our partners in the Ovation Research Network for making this analysis possible.

Correspondence
Barry Wark, PhD barry@ovation.io

SUMMARY

- Ovation's dataset provides incremental diversity to the existing public UK Biobank datasets in IBD
- Biobank datasets such as the UK Biobank and Ovation IBD Omics Data can provide novel insights into disease biology for complex diseases like IBD, especially when combining genome, expression, and rich phenotype data
- PFKFB3 is up-regulated in inflamed gastrointestinal tissue from IBD patients as compared to their neighboring non-inflamed tissue. It is also up-regulated within inflamed tissues in severe vs mild cases.
- A deeper understanding of the genetic variants that exist in PFKFB3 and their influence on gene expression could be informative in the development of new therapeutic strategies
- Further study is indicated to elucidate the relationship between unique genomic variants identified, expression profiles, and associated metadata to identify candidates of potential clinical relevance

DEMOGRAPHICS BY BIOBANK

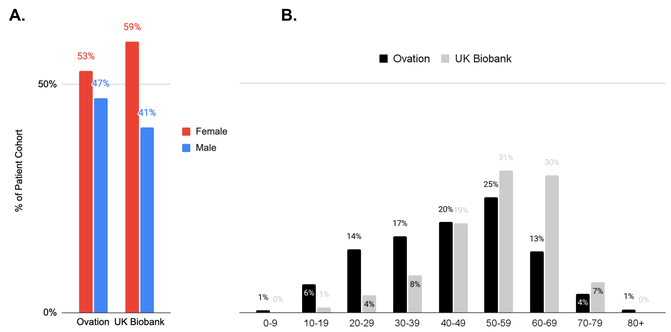

Figure 1. IBD Patient demographics by Biobank | Demographics for patients diagnosed with an ICD-10 code of K50, K51, and/or K52 in the two biobanks are represented for Ovation (N = 3,642) and for the UK Biobank (N = 8,338). **A.** Proportion of patients is broken down by sex. **B.** Compares the age group at which the first IBD related ICD-10 code was assigned for each patient

DIFFERENTIAL EXPRESSION: INFLAMMATION

Analysis of the global gene expression levels from Ovation's IBD dataset (N = 212) showed 824 DEGs inflamed and non-inflamed tissue which passed our threshold ($p_{adj} < 0.001$ & $\log_2$ fold-change > |0.6|; **Figure 2**). These results were compared to two other DEG lists to look for concordance between our dataset and existing datasets. The first was a list from Hu et al. with a similar sample size and experimental design (paired inflamed vs non-inflamed in IBD patients). From their list of 922 DEGs, 310 DEGs were overlapping: 260 were up-regulated in both studies, 46 were down-regulated in both studies and four genes (CPS1, MME, TM4SF20, & MALRD1) were up-regulated in Ovation's DEG analysis and down-regulated in the Hu et al. analysis. The second comparison we made was to the 174 DEG consensus list from two compiled studies comparing IBD patients treated with infliximab vs healthy control samples (Kaddoura, 2023). Out of this consensus list, 84 genes overlapped with Ovation's DEG list. Ovation's RNA-seq data and analysis is in high agreement with existing datasets, but also adds value in conjunction with existing datasets by contributing new and sometimes conflicting results that give insight into the biological processes that underlie IBD. Together these results highlight the complementary value of Ovation's IBD RNA-seq data with existing resources.

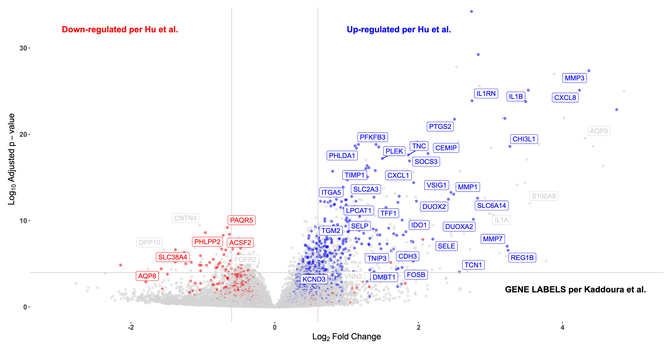

Figure 2. DEG Analysis of Ovation cohort shows concordance with previous studies | Log_2 fold-changes (FC) plotted by $-log_{10}$ adjusted p-values by gene. Grey lines indicate our thresholds (log_2 FC > |0.6| & $p_{adj} < 0.0001$). Coloring indicates genes which were reported to be significantly up- (blue) or down-regulated (red) by Hu et al. 2021. Labels highlight a subset of the 84 significant DEGs and were noted as DEGs in IBD patients vs healthy controls (Kaddoura, 2023).

BIOBANK COMPARISON: GENETIC VARIATION

To understand the genetic variation in our IBD cohort and its overlap with the UK Biobank's IBD cohort, we extracted the corresponding "exomic" variants from our WGS data so it could be compared directly with the UK Biobank's WES. From the top 1000 IBD DEGs (inflamed vs non-inflamed) described above, we counted the number of unique minor alleles per gene and contrasted this list to the top DEGs from Hu et al. because of its similar sample size and scope. Because our patient cohort (sampled from the Ovation Research Network) was expected to be more genetically diverse than the UK Biobank's primarily European cohort, we hypothesized that our cohort would capture more common minor alleles (with MAFs > 5%) for our DEGs compared to those within the UK Biobank cohort. Our data revealed that the total number of common minor alleles were similar between the Ovation and UK Biobank cohorts. However, common minor alleles occurred in twice as many genes in Ovation's cohort (**Table 1** & **Figure 3**). Therefore, the variation highlighted by these common minor alleles was more evenly distributed across genes in the Ovation cohort.

Minor alleles in top 1000 DEGs
IBD:Inflamed vs Non-inflamed

IBD Cohort	Cohort Size	# of minor alleles	# of genes (≥ 1 minor allele)
Ovation	212	2644	**648**
UK Biobank	9842	2666	298

Table 1. Number of common minor alleles by cohort for top DEGs | Minor alleles with MAFs > 5% were counted as "common minor alleles" for all genes in the top DEG list. The number of genes containing common minor alleles are listed along with the total number identified per IBD cohort (Ovation and UK Biobank).

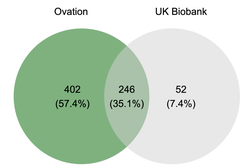

Figure 3. Genes with common minor alleles by biobank cohort | Genes with at least one common minor allele (MAFs > 5%) were counted by biobank for the top 1000 DEGs (IBD: inflamed vs non-inflamed) and plotted to illustrate overlap between the two cohorts (% are out of genes with at least one common minor allele).

DIFFERENTIAL EXPRESSION: SEVERITY

To identify IBD-related genes of interest, we conducted a subsequent differential expression analysis contrasting severe vs mild cases of IBD within inflamed samples. Results were filtered based on genes that had Ovation specific minor alleles (**Figure 4C**, colored points). This list was compared to the 174 DEGs (IBD vs healthy control) compiled in Kaddoura et al.

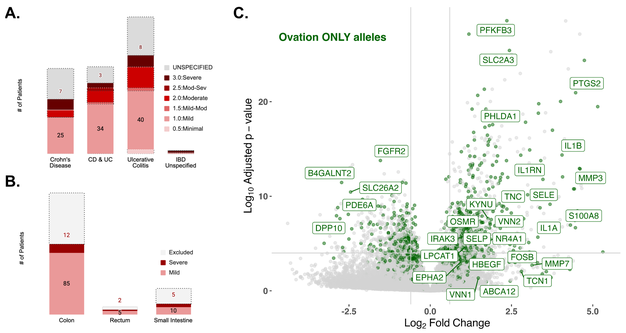

Figure 4. Comparison of Severe vs Mild inflamed IBD samples across anatomic site | A. Patients (N = 212) characterized by disease severity and diagnosis, with counts listed for the severe vs mild cohort. **B.** Samples collected from the three anatomic sites included in the severe vs mild cohort. **C.** Results for DEG analysis of severe vs mild inflamed IBD samples. Colored points indicate genes with minor alleles (MAFs > 5%) that were found in the Ovation cohort, but not the UK Biobank cohort. DEGs in IBD vs healthy controls (Kaddoura, 2023) are labeled. Thresholds marked at log_2 FC > |0.6| & $p_{adj} < 0.0001$.

EXAMPLE TARGET: PFKFB3

Significant up-regulation of PFKFB3 was observed in association with increased disease severity in inflamed IBD tissues (**Figure 5**). This up-regulation was consistently noted across all differential expression comparisons that we made (see Methods for descriptions of external DEG lists we compared to). Furthermore, direct links between PFKFB3 and IBD have been established, with studies indicating that reduced expression of this gene can lead to improvements in experimental colitis in mouse models (e.g. Duan, 2023; Zhou, 2022). Therefore, a deeper understanding of the genetic variants that exist in PFKFB3 and their influence on gene expression could be informative in the development of new therapeutic strategies.

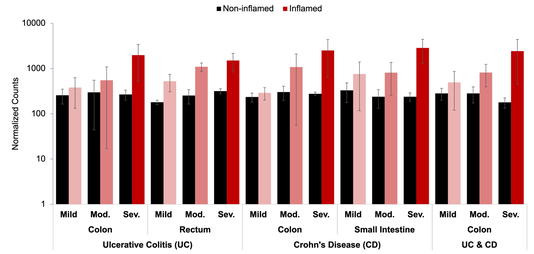

Figure 5. PFKFB3 is up-regulated in severely inflamed tissue | Normalized transcript counts were plotted by severity, anatomic site of collection and disease diagnoses. Means of normalized counts ± standard deviation are plotted (N = 186).

To identify unexplored sequence variants that might be associated with PFKFB3 expression, we considered genomic variants identified in WGS data. More than 100,000 variants were identified in PFKFB3 in the Ovation IBD cohort (N = 212). Filtering removed major allele loci, minor allele loci with a high likelihood of being sequencing errors (only observed in one sample), described variants (i.e. annotated with rsIDs), and unannotated variants that overlapped the assigned rsIDs from our variant list, resulting in a final target set of 187 unannotated minor allele loci (**Figure 6**). Notably, one minor allele with a MAF > 5% and 13 with MAFs > 1% were identified in Ovation's derived WES dataset, while zero minor alleles were found in the UK Biobank WES. PFKFB3 underscores the potential of the Ovation dataset to target genes of interest and gives us the ability to further interrogate the role of less-explored genetic variants in patients afflicted with IBD.

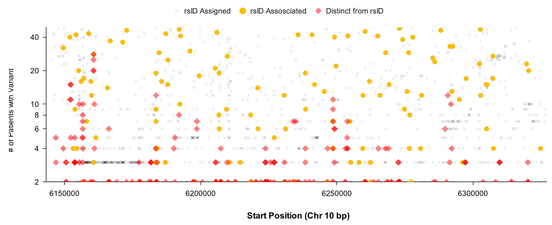

Figure 6. PFKFB3 minor alleles by position (minor allele loci) | Minor allele loci (frequency > 1%) were organized by genomic start position and divided into three groups: grey points mark variants with an assigned rsID (N = 1427), yellow points mark variants without a rsID that share a start position with one identified in this group (N = 124), and red points mark the remaining variants (N = 187).

References

- Duan et al., Macrophage LMO7 deficiency facilitates inflammatory injury via metabolic-epigenetic reprogramming. Acta Pharmaceutica Sinica B, Volume 13, Issue 12, 2023. doi:10.1016/j.apsb.2023.09.012.
- Hu et al., Inflammation status modulates the effect of host genetic variation on intestinal gene expression in inflammatory bowel disease. Nat Commun 12, 1122 , 2021. doi:10.1038/s41467-021-21458-z
- Kaddoura et al., Identification of Specific Biomarkers and Pathways in the Treatment Response of Infliximab for Inflammatory Bowel Disease: In-Silico Analysis. Life (Basel). 2023 Mar 2;13(3):680. doi: 10.3390/life13030680.
- Linggi et al., Meta-analysis of gene expression disease signatures in colonic biopsy tissue from patients with ulcerative colitis. Sci Rep. 2021 Sep 14;11(1):18243. doi:10.1038/s41598-021-97366-5.
- McKenna A., Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20:1297-303. DOI: 10.1101/gr.107524.110.
- Muzellec et al., PyDESeq2: a python package for bulk RNA-seq differential expression analysis, bioRxiv 2022.12.14.520412; doi:10.1101/2022.12.14.520412.
- Valatas et al., Editorial: Stromal and immune cell interactions in intestinal inflammation and fibrosis. Frontiers in Immunology, 14, 2023. doi:10.3389/fimmu.2023.1152140.
- M. Vasimuddin, S. Misra, H. Li and S. Aluru, "Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems," 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), Rio de Janeiro, Brazil, 2019, pp. 314-324, doi: 10.1109/IPDPS.2019.00041.
- Zhou et al., Increased stromal PFKFB3-mediated glycolysis in inflammatory bowel disease contributes to intestinal inflammation. Front Immunol. 2022 Nov 2;13:966067. doi:10.3389/fimmu.2022.966067.
- Zhou et al., P028 Inhibition of stromal glycolysis by targeting PFKFB3 decreases experimental colitis, Journal of Crohn's and Colitis, Volume 16, Issue Supplement_1, January 2022, Page i150, doi:10.1093/ecco-jcc/jab232.157.