Immunology Letters 162 (2014) 62–68

Contents lists available at ScienceDirect

Immunology Letters

journal homepage: www.elsevier.com/locate/immlet

Review

Multi-omics analysis of inflammatory bowel disease

Hu Huang[a], Pajau Vangay[a], Christopher E. McKinlay[a,b], Dan Knights[b,c,*]

[a] Biomedical Informatics and Computational Biology, University of Minnesota, Minneapolis, MN 55455, United States
[b] Biotechnology Institute, University of Minnesota, Saint Paul, MN 55108, United States
[c] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, United States

ARTICLE INFO

Article history:
Available online 15 August 2014

Keywords:
Inflammatory bowel diseases
Microbiome
Multi-omics
Bioinformatics
Machine learning
Metagenomics

ABSTRACT

Crohn's disease and ulcerative colitis, known together as inflammatory bowel disease (IBD), are severe autoimmune disorders now causing gut inflammation and ulceration, among other symptoms, in up to 1 in 250 people worldwide. Incidence and prevalence of IBD have been increasing dramatically over the past several decades, although the causes for this increase are still unknown. IBD has both a complex genotype and a complex phenotype, and although it has received substantial attention from the medical research community over recent years, much of the etiology remains unexplained. Genome-wide association studies have identified a rich genetic signature of disease risk in patients with IBD, consisting of at least 163 genetic loci. Many of these loci contain genes directly involved in microbial handling, indicating that the genetic architecture of the disease has been driven by host–microbe interactions. In addition, systematic shifts in gut microbiome structure (enterotype) and function have been observed in patients with IBD. Furthermore, both the host genotype and enterotype are associated with aspects of the disease phenotype, including location of the disease. This provides strong evidence of interactions between host genotype and enterotype; however, there is a lack of published multi-omics data from IBD patients, and a lack of bioinformatics tools for modeling such systems. In this article we discuss, from a computational biologist's point of view, the potential benefits of and the challenges involved in designing and analyzing such multi-omics studies of IBD.

1. Introduction

Crohn's disease (CD) and ulcerative colitis (UC) are the two most common forms of inflammatory bowel diseases (IBD). CD involves intestinal inflammation mostly affecting the terminal ileum and colon, although it can also affect any other part of the gastrointestinal tract; UC is limited to the mucosa and submucosa of the colon. IBD is thus a complex chronic inflammatory disorder, likely caused by an series of interactions among multiple pathogenic factors, including genetics, environments and mucosal immune disorders [1–3]. The precise etiology of IBD in many cases is still unclear, and much ongoing research is based on defining causal mechanisms for known genetic susceptibility [1] and on understanding microbial risk factors [4].

The complexity of the IBD disease phenotype is driven by gene-environment interactions, possibly mediated by dysbiosis in the gut microbiota. Traditional bottom-up studies focus on individual candidate factors, such as candidate genes or microorganisms, and largely employ hypothesis-directed experimentation rather than exploratory research. Despite the large number of candidate factor studies, they do not readily provide holistic insight into the complexity of IBD etiology. IBD etiology is known to consist of intricate interacting networks formed by genetic and environmental factors, for instance diet, smoking status, age, medication history, and family history. Recent advances in high-throughput experimental "omics" data generation techniques along with necessary bioinformatics tools have resulted in new top-down research strategies to overcome these shortcomings in traditional approaches. Here "omics" refers to large-scale high-dimensional biological data generation aiming to comprehensively analyze one type of biochemical molecular species or interactions of molecules in a cellular system [5,6]. Experimental omics approaches, such as genomics, transcriptomics, proteomics and metabolomics, share several features when compared to traditional procedures [6,7]: (i) they are high-throughput, data-driven, holistic and top-down methodologies, in the sense that massive data are collected first with no prior hypothesis, and then the meaningful results are searched and explained within the obtained dataset; (ii) they consider the targets, such as cellular metabolism, as an "integrated system" by incorporating

* Corresponding author at: Keller Hall 4-192, 200 Union St SE, Minneapolis, MN 55455, United States. Tel.: +1 612 564 2762.
E-mail address: dknights@umn.edu (D. Knights).

http://dx.doi.org/10.1016/j.imlet.2014.07.014
0165-2478/

the relationships between different measurements; (iii) the analysis of these high-throughput data are statistically complicated and computationally expensive.

So, multi-omics: what's in it for IBD? Combining data collection across the host–microbiome interface allows the potential for modeling complex interactions between host immune response, host environmental factors including diet, and the gut microbiome. Large-scale genome-wide association studies (GWAS) [8–11] and extensive research in mouse models [12–14] have provided us with a substantial base of knowledge covering IBD and signaling pathways in the host. Similarly, we now have a substantial representation of gut microbiome biodiversity in publicly available functionally annotated bacterial reference genomes [8,15–18]. Yet there are limited systematic methods for connecting these two bodies of knowledge. Collecting and integrating multi-omics data has begun to receive IBD researchers' attention, but because of technical and financial reasons multi-omics research on IBD is still in its infancy [3]. IBD researchers wishing to collect multi-omics data to measure simultaneous behavior of the host immune system and the gut microbiome, for example, via dual RNA-Seq [19], will be forced to use ad-hoc bioinformatics methods or to develop novel bioinformatics tools.

Discovery of novel host–microbiome disease pathways in IBD will benefit from multi-omics study designs including paired host genomics or transcriptomics, bacterial metagenomics or metatranscriptomics, metaproteomics, immunomics, and other data types. However, this type of study design requires careful consideration of a number of analytical challenges, and substantial bioinformatics development is needed to enable rapid progress in the field.

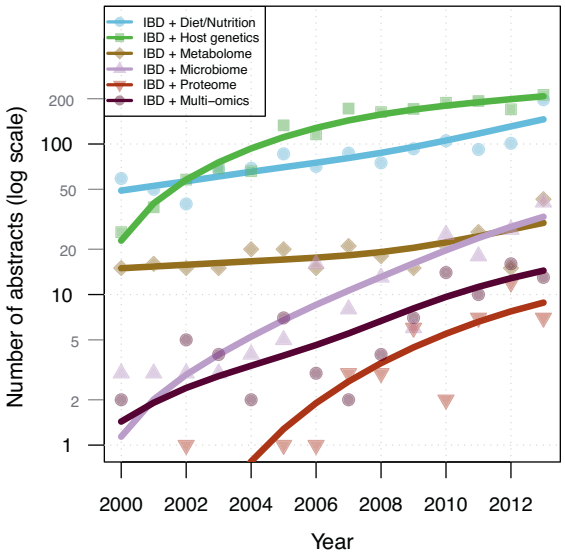

Fig. 1. Historical trends of 'omics terms in IBD-related research articles. The number of abstracts per year (log scale) for research articles mentioning both inflammatory bowel diseases and several kinds of high-throughput 'omics data since the year 2000. Trend data was mined using pattern matching within the results from a search of the PubMed database [85] for inflammatory bowel disease(s), Crohn's disease, or ulcerative colitis. We also mined abstracts for combinations of two or more of these "omes" (data types), but there were too few to show any pair individually on this plot. Instead we included a single series including IBD-related publications with any two or more different omics terms. This indicates that although high-throughput 'omics data are increasingly being collected for studying IBD, very little multi-omics IBD research has been published to date. This is likely due to a combination of the increased challenges associated with both multi-omics data collection and analysis.

2. Conventional IBD 'omes

Here we provide a brief overview of several important data types in IBD 'omics research. For each IBD "ome" we discuss important applications as well as potential pitfalls and challenges expected in modeling each data type. These data types have a history of established application in IBD research. Based on systematic mining of published abstracts from IBD research literature on PubMed, we found that these now common 'omics data types have been increasingly utilized to study IBD over the last ten years (Fig. 1). We found a noted lack, however, of published multi-omics IBD research. With a small number of exceptions (for example, Erickson et al. [20]), the following IBD 'omes have been studied independently of one another.

2.1. Genomics (host)

Host genetics has received the most attention of any of the "omics" data types we survey here. A single gene, nucleotide-binding oligomerization domain-containing protein 2 (NOD2), holds the largest component of disease risk for Crohn's disease (CD); NOD2 was the first known IBD-related gene [21–23]. Now thanks to very large-scale genotyping efforts the IBD genotype is known to contain at least 163 genetic loci [8], with some loci specific to CD and some to ulcerative colitis (UC). Only approximately 30% of the IBD loci are coding variants [24], and most of the loci have unknown function. Furthermore, most of the SNPs identified through GWAS are merely representative of a signal somewhere within a broader genetic locus. Fine-mapping studies of the exomes and regulatory landscape surrounding these putative IBD SNPs ("exomics") in specific cell types and tissues is expected to yield important insights into the functional roles of the loci in IBD pathogenesis [24]. Although knowledge of the genetics of IBD has grown exponentially for the last decade, the contribution of specific genes and the impact of their associated variation is largely still unclear, indicating that that genomics alone is not sufficient to uncover the causes of IBD.

2.2. Transcriptomics (host)

Transcriptomics refers to the study of the complete set of RNA molecules, or "transcripts", present in a population of cells. While genomics examines static DNA information, transcriptomics measures the dynamic expression of RNA molecules and their variation under different circumstances at the genome scale, hence reflecting the genes that are actively expressed at any given time [7], with the exception of mRNA degradation phenomena [25,26]. Techniques that allow assessment of RNA profiles on a genome-wide scale are microarrays [27], serial analysis of gene expression (SAGE) [28], and RNA-seq [29]. More recently, a "dual RNA-seq" technique has been proposed, which focuses on both host and microbiome transcriptome data [19].

Measuring the host transcriptomes in, for example, tissue from a mucosal biopsy may shed light on the phenotypic effects of the host genotype of immune system response under various exposures to bacterial products or other bioactive compounds. Transcriptomics via microarrays and RNA-Seq has been an essential component of IBD research for mapping immune regulatory and signaling networks and for understanding the etiology of IBD, often in controlled ex vivo experiments [29]. Investigation using high-density cDNA microarrays have identified several important regulatory molecules and genes associated with a disrupted immune response that may be involved in the pathogenesis of IBD [30]. Similar research on pediatric-onset IBD patients [31] confirmed the association of previously published IBD genes, but also expanded the number of immune-related genes differentially regulated in IBD. Moreover, transcriptomics has been applied to the study of the effect of anti-inflammatory diet interventions in Crohn's disease patients [32], and biomarker discovery using non-coding RNAs,