Manuscript Title

This manuscript (permalink) was automatically generated from taylorreiter/2021-paper-ibd@3b7bbf4 on November 23, 2021.

Authors

- **Taylor Elaine Reiter**
  0000-0002-7388-421X · taylorreiter · ReiterTaylor
  Department of Population Health and Reproduction, UC Davis · Funded by Grant XXXXXXXX

- **Luiz Irber**
  0000-0003-4371-9659 · luizirber · luizirber
  Graduate Group in Computer Science, UC Davis; Department of Population Health and Reproduction, UC Davis · Funded by Grant XXXXXXXX

- **Phillip T. Brooks**
  0000-0003-3987-244X · brooksph · brooksph
  Department of Population Health and Reproduction, UC Davis

- **Amy D. Willis**
  0000-0002-2802-4317 · ctb · AmyDWillis
  Department of Biostatistics, UW

- **C. Titus Brown**
  0000-0001-6001-2677 · ctb
  Department of Population Health and Reproduction, UC Davis

Abstract

This manuscript is a template (aka “rootstock”) for Manubot, a tool for writing scholarly manuscripts. Use this template as a starting point for your manuscript.

The rest of this document is a full list of formatting elements/features supported by Manubot. Compare the input (`.md` files in the `/content` directory) to the output you see below.

Basic formatting

Bold text

Semi-bold text

Centered text

Right-aligned text

Italic text

Combined *italics and **bold***

~~Strikethrough~~

1. Ordered list item
2. Ordered list item
    a. Sub-item
    b. Sub-item
        i. Sub-sub-item
3. Ordered list item
    a. Sub-item

- List item
- List item
- List item

subscript: H_2O is a liquid

superscript: 2^{10} is 1024.

unicode superscripts^{0123456789}

unicode subscripts$_{0123456789}$

A long paragraph of text. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Putting each sentence on its own line has numerous benefits with regard to editing and version control.

Line break without starting a new paragraph by putting
two spaces at end of line.

Document organization

Document section headings:

Heading 1

Heading 2

Heading 3

Heading 4

Heading 5

Heading 6

A heading centered on its own printed page

Horizontal rule:

`Heading 1`'s are recommended to be reserved for the title of the manuscript.

`Heading 2`'s are recommended for broad sections such as *Abstract*, *Methods*, *Conclusion*, etc.

`Heading 3`'s and `Heading 4`'s are recommended for sub-sections.

Links

Bare URL link: https://manubot.org

Long link with lots of words and stuff and junk and bleep and blah and stuff and other stuff and more stuff yeah

Link with text

Link with hover text

Link by reference

Citations

Citation by DOI [1].

Citation by PubMed Central ID [2].

Citation by PubMed ID [3].

Citation by Wikidata ID [4].

Citation by ISBN [5].

Citation by URL [6].

Citation by alias [7].

Multiple citations can be put inside the same set of brackets [1,5,7]. Manubot plugins provide easier, more convenient visualization of and navigation between citations [2,3,7,8].

Citation tags (i.e. aliases) can be defined in their own paragraphs using Markdown's reference link syntax:

Referencing figures, tables, equations

Figure 1

Figure 2

Figure 3

Figure 4

Table 1

Equation 1

Equation 2

Quotes and code

> Quoted text

> Quoted block of text
>
> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.

Code `in the middle` of normal text, aka `inline code`.

Code block with Python syntax highlighting:

```
from manubot.cite.doi import expand_short_doi

def test_expand_short_doi():
    doi = expand_short_doi("10/c3bp")
    # a string too long to fit within page:
    assert doi == "10.25313/2524-2695-2018-3-vliyanie-enhansera-copia-i-
        insulyatora-gypsy-na-sintez-ernk-modifikatsii-hromatina-i-
        svyazyvanie-insulyatornyh-belkov-vtransfetsirovannyh-geneticheskih-
        konstruktsiyah"
```

Code block with no syntax highlighting:

```
Exporting HTML manuscript
Exporting DOCX manuscript
Exporting PDF manuscript
```

Figures

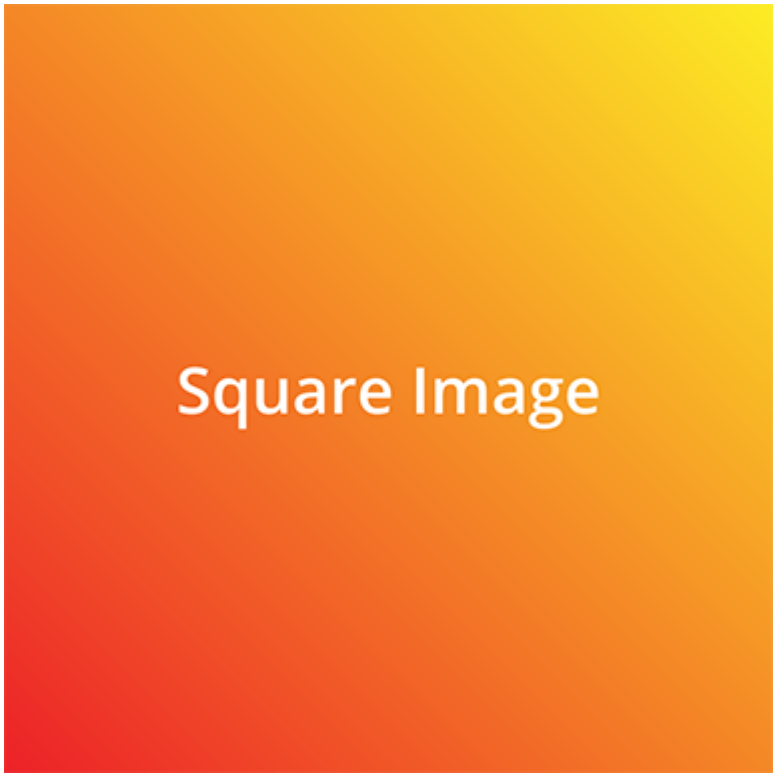

Figure 1: A square image at actual size and with a bottom caption. Loaded from the latest version of image on GitHub.

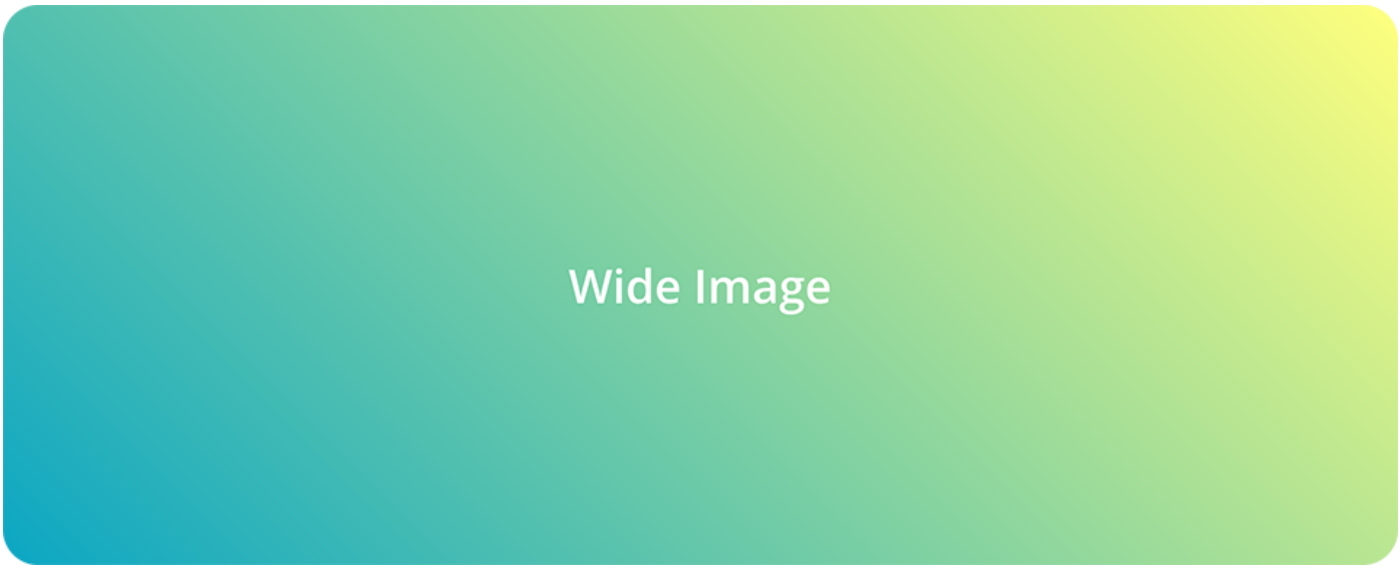

Figure 2: An image too wide to fit within page at full size. Loaded from a specific (hashed) version of the image on GitHub.

Figure 3: A tall image with a specified height. Loaded from a specific (hashed) version of the image on GitHub.

Figure 4: A vector `.svg` image loaded from GitHub. The parameter `sanitize=true` is necessary to properly load SVGs hosted via GitHub URLs. White background specified to serve as a backdrop for transparent sections of the image.

Tables

Table 1: A table with a top caption and specified relative column widths.

Bowling Scores	**Jane**	**John**	**Alice**	**Bob**
Game 1	150	187	210	105
Game 2	98	202	197	102
Game 3	123	180	238	134

Table 2: A table too wide to fit within page.

	Digits 1-33	**Digits 34-66**	**Digits 67-99**	**Ref.**
pi	3.14159265358979323 846264338327950	28841971693993751 0582097494459230	78164062862089986 2803482534211706	`piday.org`
e	2.71828182845904523 536028747135266	24977572470936999 5957496696762772	40766303535475945 7138217852516642	`nasa.gov`

Table 3: A table with merged cells using the `attributes` plugin.

	Colors	
Size	**Text Color**	**Background Color**
big	blue	orange
small	black	white

Equations

A LaTeX equation:

$$\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2} \tag{1}$$

An equation too long to fit within page:

$$x = a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 \tag{2}$$

Special

⚠ **WARNING** *The following features are only supported and intended for* `.html` *and* `.pdf` *exports. Journals are not likely to support them, and they may not display correctly when converted to other formats such as* `.docx`*.*

LINK STYLED AS A BUTTON

Adding arbitrary HTML attributes to an element using Pandoc's attribute syntax:

Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Adding arbitrary HTML attributes to an element with the Manubot `attributes` plugin (more flexible than Pandoc's method in terms of which elements you can add attributes to):

Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Available background colors for text, images, code, banners, etc:

`white` `lightgrey` `grey` `darkgrey` `black` `lightred` `lightyellow` `lightgreen` `lightblue` `lightpurple` `red` `orange` `yellow` `green` `blue` `purple`

Using the Font Awesome icon set:

✔ ? ★ 🔔 ⊗ •••

Light Grey Banner
useful for *general information* - manubot.org

Blue Banner
useful for *important information* - manubot.org

Light Red Banner
useful for *warnings* - manubot.org

Introduction

Sub-species groupings of microorganisms have functional differences that govern important genome-environment interactions across diverse ecosystems. For example, ecotypes of Escherichia coli have different gene complements that allow each group to thrive in diverse environments like the gut, soil, and freshwater [PNAS. 2011;108: 7200–7205]. Metagenomic sequencing data from communities of microorganisms contain information about specific strains present in a sample, but strain-resolved insights are lacking due to incomplete references or inability of current tools to retrieve such information (CITATIONS). Here we use *strain* to refer to within-species variation that generates taxonomic grouping below the species level.

Inflammatory bowel disease (IBD) is a spectrum of diseases characterized by chronic inflammation of the intestines that is likely caused by host-mediated inflammatory responses elicited in part by microorganisms [**kostic2014?**]. IBD is cyclical with periods of active disease and remission. IBD manifests in three subtypes depending on clinical presentation, including Crohn's disease (CD), which presents as discontinuous patches of inflammation throughout the gastrointestinal tract, ulcerative colitis (UC), which presents as continuous inflammation isolated to the colon, and undetermined, which cannot be distinguished as CD or UC. Diagnosis is often clinically difficult, with ramifications associated with over- or under-treatment that lead to decreased patient well-being. Detection of microbial signatures associated with IBD subtype may lead to improved diagnostic criteria and therapeutics that extend periods of remission. However, such a signature has thus far remained elusive [**kumar2019integrating?**].

The microbiome of CD and UC is heterogeneous, and studies that characterize the microbiome often produce conflicting results. This is likely in part driven by large inter- and intra-individual variation [**lloyd2019?**], but is also attributable to non-standardized laboratory, sequencing, and analysis techniques used to profile the gut microbiome [**kumar2019integrating?**]. Dysbiosis is frequently observed in IBD, particularly in CD [**kang2010dysbiosis?**,**machiels2014decrease?**,**lewis2015?**,**moustafa2018?**,**qin2010?**], however dysbiosis alone is not a signature of IBD [**lloyd2019?**]. Dysbiosis is defined as a decrease in gut microbial diversity that results in an imbalance between protective and harmful microorganisms, leading to intestinal inflammation [**weiss2017mechanisms?**].

Strain-level differences may account for some heterogeneity in IBD gut microbiome profiles. A recent investigation of time-series gut microbiome metagenomes found that one clade of *Ruminococcus gnavus* is enriched in CD [**hall2017?**]. Further, this clade produces an inflammatory polysaccharide [**henke2019ruminococcus?**]. While this clade is enriched in CD, its enrichment was previously masked from computational discovery by concomitant decreases in other *Ruminococcus* species in

IBD [hall2017?], highlighting the need for strain-resolved analysis of metagenomic sequencing in the exploration of IBD gut microbiomes.

Given these features of the IBD gut microbiome, strain-resolved analysis may improve insights into the XXX of these communities. The two biggest obstacles to strain-level analysis of short read data are data getting thrown away, either because it's not in reference databases or because it doesn't assemble or bin, and resolving genomes from communities with mixed populations of closely related but distinct genomes. While long reads have made strides toward resolving the latter issue (CITE: Bickhart), in habitats like the gut where communities are dominated by single strains of microbes (CITE: bork lab paper), the largest barrier to strain-level analysis is using all of the data. Here, we combine k-mer-based analysis with assembly graphs to not throw away the data.

K-mers, words of length *k* in nucleotide sequences, have previously been used for annotation-free characterization of sequencing data [sheppard2013genome?,dubinkina2016assessment?,standage2019kevlar?]. K-mers are suitable for strain-resolved metagenome analysis because they do not need to be present in reference databases to be included in analysis, they do not rely on marker genes which are largely conserved at the strain level, and they are suitable for species- and strain-level classification (CITE: metapallete, gather). Investigating all k-mers in metagenomes is more computationally intensive than reference-based approaches [benoit2016multiple?]., by data-reduction techniques like scaled MinHash sketching make k-mer-based analysis scalable to large-scale sequence comparisons [pierce2019?,rowe2019levee?]. MinHash sketching sacrifices the fine-scaled resolution of reference-based techniques but is representative of the full sequencing sample and complete databases, including strain-variable accessory elements that may be associated with diseases.

Assembly graphs complement sketch-based analysis [brown2020exploring?,jaillard2018fast?]. While both k-mers and assembly graphs can be used to represent all sequences contained within a metagenome, assembly graphs retain important sequencing context and known functional and taxonomic annotations, recovering critical information lost through the MinHash sketching approach. While assembly graphs have been leveraged in metagenome analyses [12, 13], their large size precluded analysis at scale. The *spacegraphcats* tool is designed to tackle this issue, encoding algorithms that scalably reduce the size of an assembly graph, enabling efficient querying and sequence retrieval (CITE SGC). These algorithms center around dominating sets, a subset of nodes that ensure that every node in the assembly graph is at most distance one from a node in the dominating set. Dominating sets partition the graph into *pieces* by assigning every node to exactly one of the closest nodes in the dominating set (CITE SGC). This simplified graph enables efficient queries: querying with a sequence that overlaps at least one k-mer in a compact de Bruijn graph (cDBG) node returns all k-mers (or all reads containing those k-mers) from the graph piece. We refer to sequences retrieved by a graph query as *neighborhoods* (CITE SGC). Genome queries often recover sequences not in reference databases or *de novo* assemblies, which disproportionately include sequences from low coverage regions or highly variable portions of the graph (CITE SGC). When a query has a Jaccard similarity between 10-2 and 10-3, 20-40% of a target genome sequence is recovered from a metagenome query (CITE SGC). This jumps to >80% when Jaccard similarity exceeds 10-1 (CITE SGC).

Here, we analyzed a meta-cohort of IBD gut microbiome metagenomes

Here we extend the functionality of spacegraphcats to enable differential abundance analysis on the simplified assembly graph. We perform this analysis on a meta-cohort of IBD gut microbiome metagenomes originating from six studies (260 CD, 132 UC and 213 healthy controls) (Table 1) [2,7,9,10,16,17]. We perform this analysis on sequences from species R. gnavus to determine which sections of the pangenome are more abundant in IBD (CD, UC) than non-IBD.

Here we use k-mer- and assembly graph-based techniques to perform a meta-cohort analysis of six studies of IBD gut metagenome cohorts comprising 260 CD, 132 UC and 213 healthy controls (see **Table @ref(tab:cohorts)**) [lloyd2019?,lewis2015?,hall2017?,franzosa2019?,gevers2014?,qin2010?]. SUMMARY OF TECHNIQUES WE BUILT? 1. LOOCV on scaled MinHash signatures from quality controlled metagenome data 2. Identification of consistent cross-study and cross-model signatures 3. Gather to anchor to genomes 4. SGC R1 nbhd queries to recover variable regions not in reference databases, and unite all the retrieved sequences under an organizing umbrella species 5. SGC R10 metapangenome graphs 6. Dominating set differential abundance analysis We demonstrate a weak but consistent signature of IBD subtype in fecal microbiome metagenomes. Only a small subset of all k-mers are predictive of UC and CD, and these k-mers originate from a core set of microbial genomes. We find that stochastic loss of diversity in this core set of microbial genomes is a hallmark of CD, and to a lesser extent, UC. While reduced diversity is responsible for the majority of disease signatures, we find signatures of strain enrichment in disease. Genes presumably associated with these strains occur more frequently in IBD metagenomes but are present in low abundance in nonIBD as well. Our findings highlight the need for strain-level analysis of metagenomic data sets, and provide future avenues for research into IBD therapeutics.

References

1. **Sci-Hub provides access to nearly all scholarly literature**
Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene
eLife (2018-03-01) https://doi.org/ckcj
DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

2. **Reproducibility of computational workflows is automated using continuous analysis**
Brett K Beaulieu-Jones, Casey S Greene
Nature biotechnology (2017-04) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/
DOI: 10.1038/nbt.3780 · PMID: 28288103 · PMCID: PMC6103790

3. **Bitcoin for the biological literature.**
Douglas Heaven
Nature (2019-02) https://www.ncbi.nlm.nih.gov/pubmed/30718888
DOI: 10.1038/d41586-019-00447-9 · PMID: 30718888

4. **Plan S: Accelerating the transition to full and immediate Open Access to scientific publications**
cOAlition S
(2018-09-04) https://www.wikidata.org/wiki/Q56458321

5. **Open access**
Peter Suber
MIT Press (2012)
ISBN: 9780262517638

6. **Open collaborative writing with Manubot**
Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
Manubot (2020-05-25) https://greenelab.github.io/meta-review/

7. **Opportunities and obstacles for deep learning in biology and medicine**
Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, … Casey S Greene
Journal of The Royal Society Interface (2018-04-04) https://doi.org/gddkhn
DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

8. **Open collaborative writing with Manubot**
Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653