Graph-based Analysis of Genetic Features Associated with Mobile Elements in Crohn's Disease and Healthy Gut Microbiomes

Julia Warnke-Sommer[1] and Hesham Ali[2]
[1]*Pathology and Microbiology, University of Nebraska Medical Center, Omaha, 68198 U.S.A.*
[2]*College of Information Science and Technology, University of Nebraska Omaha, 68106, U.S.A.*

Keywords: Next Generation Sequencing, Graph Theory, Metagenomics, Mobile Elements, Crohn's Disease.

Abstract: Horizontal gene transfer is a major driver of bacterial evolution and adaptation to niche environments. This holds true for the complex microbiome of the human gut. Crohn's disease is a debilitating condition characterized by inflammation and gut bacteria dysbiosis. In previous research, we analyzed transposase associated antibiotic resistance genes in Crohn's disease and healthy gut microbiome metagenomics data sets using a graph mining approach. Results demonstrated that there were significant differences in the type and bacterial distribution of transposase-associated antibiotic resistance genes in the Crohn's and healthy data sets. In this paper, we extend the previous research by considering all gene features associated with transposase sequences in the Crohn's disease and healthy data sets. Results demonstrate that some transposase-associated features are more prevalent in Crohn's disease data sets than healthy data sets. This study may provide insights into the adaptation of bacteria to gut conditions such as Crohn's disease.

1 INTRODUCTION

The human gut microbiome represents a wealth of biological organisms and their related functions. The composition and abundances of these microorganisms have been found to be associated with numerous important health characteristics. For example the gut microbiome has been found to help regulate immune function and development (Sommer, Bäckhed, 2013) influence dietary metabolism (Turnbaugh *et al*, 2006) and even modulate mood and behaviour (Foster, McVey Neufeld, 2013). The community of the gut microbiome is highly complex with a wide array of diverse microbial members that interact with each other as well as with the host. In a healthy state, this consortium of microorganisms is well adapted to perform multiple functions that are beneficial to the host as well as for maintaining the homeostasis of the gut environment (LeBlanc *et al*, 2013) (Kamada *et al*, 2013). These community members are specifically adapted to flourish in the gut environment and are flexible to adapt to changes to the environment such as diet alterations (David *et al*, 2014) (O'Sullivan *et al*, 2009).

One such mechanism that bacteria utilize to adapt to their environments is the transfer of small amounts genetic material called mobile genetic elements. Mobile genetic elements include plasmids, transposons, and bacteriophage related sequences. In particular, transposons are segments of DNA that can remove and insert themselves in different regions of the same or different genomic sequence. Also known as "jumping genes" these transposons many times utilize specialized genes called transposases to catalyse the excision and movement of the transposon sequence. In bacteria, transposons often carry accessory genes such as antibiotic resistance genes that can confer specialized functions to the organism in which it is integrated. The horizontal gene transfer of genes between bacteria is a major driver of bacterial evolution and allows for adaptation of a given bacterial species to its environmental niche (Ochman, Lawrence, Groisman, 2000).

Dysbiosis of the gut microbiome occurs in many disease states including inflammatory bowel disease (IBD) such as Crohn's disease and ulcerative colitis (Manichanh *et al*, 2012). According to research, this dysbiosis may facilitate increased horizontal gene transfer between bacterial members of the gut microbiome, promoting the spread of virulence

Warnke-Sommer J. and Ali H.
Graph-based Analysis of Genetic Features Associated with Mobile Elements in CrohnâĂŹs Disease and Healthy Gut Microbiomes.
DOI: 10.5220/0006254201040114
In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*, pages 104-114
ISBN: 978-989-758-214-1

factors and antibiotic resistance genes (Stecher, Maier, Hardt, 2013).

Next generation sequencing technologies can be applied to investigate the composition of the gut microbiome. These technologies are capable of producing millions of short DNA fragments called reads from a given input DNA sample. The reads are produced at such a high coverage of the original sample sequence that many overlap. Graph based tools called assemblers are used to assemble the reads into longer stretches of sequence called contigs that are used for downstream analysis (Nagarajan, Pop, 2013). We have developed an assembler tool called Focus (Warnke, Ali, 2014) that relies on a novel assembly graph called the hybrid graph. The hybrid graph models the overlap relationships between reads in a given data set at multiple levels of granularity.

In previous research (Warnke-Sommer, Ali, 2016), we demonstrated that transposase sequences are associated with graph structure due to their repetitive nature within a single genome or across multiple different genomes. The distribution of transposase sequences was studied in gut microbiomes of individuals with Crohn's disease and healthy individuals. Graph mining was used to explore genomic regions in proximity to the transposase sequences for antibiotic resistance genes. Regions around transposase sequences were enriched for different types of antibiotic resistances genes in the Crohn's disease versus the healthy gut microbiome. The distribution of these antibiotic resistance gene groups across genera is also different between Crohn's disease gut microbiomes and healthy gut microbiomes with most antibiotic resistance genes in the healthy gut samples coming from *Bacteroides*. The distribution of resistance genes in the Crohn's disease samples was more diverse across bacterial genera. This study was conducted to first demonstrate the capability of the hybrid graph to capture biological features in its graph structures. Finally this study also provided insights into the distribution of antibiotic resistance in the gut microbiome of Crohn's disease, a disease whose resulting complications are often treated with antibiotics regimens (Roy, Lichitiger, 2016).

This paper extends the analysis conducted in (Warnke-Sommer, Ali, 2016) to examine all gene features associated with transposase sequences in the Crohn's disease and healthy metagenomics data sets. Antibiotic resistance genes are commonly known to be in association with transposase sequences (Ghosh *et al*, 2013). However, it would be beneficial to examine additional genetic features associated with transposase sequences in Crohn's disease and healthy gut microbiome samples. The ability to determine transposase associated gene differences in Crohn's and healthy gut microbiomes provides valuable insights into potential niche adaptations of bacterial species in gut environments afflicted with Crohn's disease. Results demonstrate that transposase associated gene features are significantly different in Crohn's disease versus healthy gut microbiomes. Crohn's disease samples had significantly greater hits to several functional categories, including beta-glucoside metabolism, maltose and maltodextrin utilization, and heme, hemin uptake and utilization systems in gram positives.

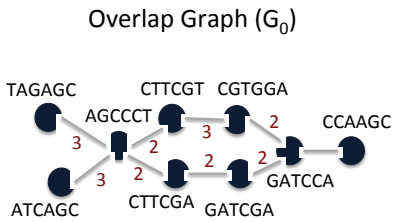

Figure 1: The overlap graph. Each read is mapped to a node in the overlap graph. Edges represent overlap relationships. Each edge is weighted according to the length of the overlap region shared between a pair of reads.

2 METHODS

In this section, the pipeline for analysing the transposase-associated gene features in the Crohn's disease and healthy gut microbiomes is discussed. First, the graph model applied for modelling the metagenomics reads and their overlap relationships is described. This graph model is part of the Focus assembler pipeline described in detail in (Warnke, Ali, 2014). Second, the database that is used to functionally annotate gene features is briefly discussed. Finally, the graph mining approach for extracting transposase related gene features from the hybrid graph is presented.

2.1 Graph Model

The overlap graph approach is commonly used to model reads and their corresponding read overlap relationships (Miller, Koren, Sutton, 2010). In this approach, each read is mapped to a node in the overlap graph and edges represent the overlap relationships between reads. The edges can be

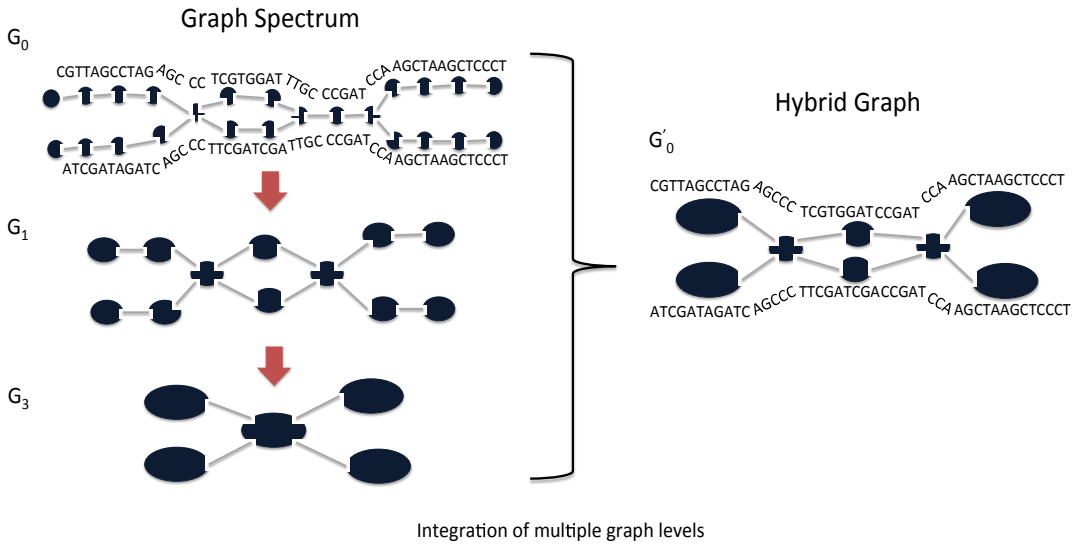

Figure 2: The multilevel graph set and hybrid graph. The initial overlap graph G0 is shown above. The DNA sequences beside G0 represent contigs that can be inferred from the overlap graph. Notice that the overlap graph captures the sequence variation between the two contigs. The multilevel graph set is created by reclusively matching and merging nodes to produces a succession of graphs G0 ... G3. The graph G3 is over reduced and does not capture the bubble structure in the original overlap graph. The hybrid graph integrates the three graph levels to produce a representation that is concise yet captures all sequence structure in the read set.

weighted to reflect overlap characteristics such as overlap length or identity. See Figure 1 for an example of the overlap graph.

The Focus assembly algorithm first performs seed-and-extend pairwise alignment to discover overlap relationships between reads (Warnke, Ali, 2014). Once the pairwise alignment module is completed, the overlap relationships and reads are loaded into an overlap graph representation, denoted as G_0. In this graph, edges are weighted by the lengths of the reads' shared overlaps.

The overlap graph G_0 is highly complex and can contain millions of node, making it difficult to recover any useful information from its structures. To address this issue, the overlap graph G_0 is recursively reduced with heavy edge matching and node merging to produce a series of reduced graphs G_1, G_2 … G_n, where $|G_1| \geq |G_2| \geq \ldots |G_n|$. Heavy edge matching is a maximal matching that is generated with preference for heavier edge weights (Karypis, Kumar, 1998). Focus attempts to find a maximal heavy edge matching that satisfies minimal edge weight thresholds set by the user. Once this matching is found, the endpoints of the edges in the matching are merged to form the new nodes in G_1. An edge between two merged nodes in G_1 is assigned the summed weight of the edges between the merged nodes' child nodes in G_0.

Figure 2 demonstrates graph reduction and multilevel graph set. Notice that G_3 does not capture a bubble in the overlap graph and is over reduced. Not all levels of the multilevel graph set will be appropriate for representing all regions of genomic sequence such as complex repeats or other sequence variation. To address this issue, nodes are selected from various levels of the multilevel graph set and integrated to create a novel graph model called the hybrid graph. The hybrid graph represents the input data set as concise as possible while still capturing input data set features. The nodes are selected from the multilevel graph set as follows. First the nodes of the coarsest graph G_n are evaluated. If a node is found to be a representative of a single contiguous sequence, then it will be added to the hybrid graph. If the reads that this node represents do not assemble into a contiguous sequence, then the algorithm does not add the node to the hybrid graph. Instead its child nodes in G_{n-1} are evaluated to examine whether their corresponding reads form a contiguous sequence. If they do, then they are added to the hybrid graph. If not, then their child nodes in G_{n-2} are evaluated. This process continues until G_0 is reached. The hybrid graph will contain all of the nodes in G_0, G_1, ... G_n that have been evaluated and found to form a contiguous region of sequence. Thus the hybrid graph represents the input data set as concisely as possible while capturing input sequence variation.

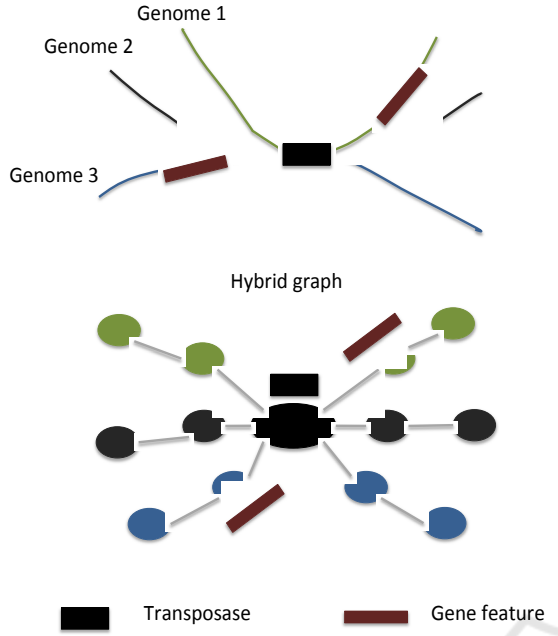

Figure 3: Graph mining for discovering biologically significant features. In (a) there are three genomes that share a common transposase sequence. In two of these genomes, a gene feature is associated with the transposase. In (b) the hybrid graph representing these genomes is shown. The common transposase region is reduced to a single node that has many branching paths representing the three unique genomes. Observe that the gene features are represented by nodes that are within a short path length from the node representing the shared transposase sequence. Any genetic features associated with this transposase can be determined by exploring the node neighbourhood centered on the node representing the repeated transposase sequence.

Please see (Warnke, Ali, 2014) for a formal description of the multilevel graph set and hybrid graph construction.

2.2 Seed Subsystems and ACLAME Database

The Seed Subsystems (Overbeek *et al*, 2014) is an organizational database for gene annotation. It is organized into several hierarchal levels of functional subsystems. This database has annotated protein sequences that are organized into families called Figfams based on sequence similarity and function (Meyer, Overbeek, Rodriguez, 2009). These Figfams are then annotated with the subsystem functional categories.

In this paper, the gene prediction software FragGeneScan (Rho, Tang, Ye, 2010) was used to predict genes within the metagenomics contigs produced by Focus. These predicted genes were then aligned against the seed subsystem using the DIAMOND software tool (Buchfink, Xie, Huson, 2015). The predicted genes were assigned annotations at subsystem level 2 according to the Figfam hit that they were most similar to. The minimum score for each protein hit was set to 40% minimum sequence identity with a minimum alignment length of 50.

The ACLAME database contains a comprehensive collection of prokaryotic mobile elements (Leplae *et al*, 2004). As before, the DIAMOND software was used to align the predicted genes against the ACLAME database with a minimum sequence identity score of 40% and minimum alignment length of 50. Any contigs with hits were annotated accordingly.

Finally, the nodes in the hybrid graph were annotated according to their corresponding contigs' annotations as determined by the SEED or ACLAME databases.

2.3 Graph Mining for Biological Features Associated with Mobile Elements

In (Warnke-Sommer, Ali, 2016) it was demonstrated that graph structure was associated with biologically significant features. In particular, sequence regions that are repetitive within the same genome or are present in multiple different genomes were found to be represented by nodes that have a large degree. As shown by Fig. 3, a region that is repeated within the same genome or is present in multiple different genomes will be reduced to a single node in the hybrid graph. In the hybrid graph, the number of paths containing a node representing a conserved or repeated sequence may indicate the number of distinct genomic regions that contain that repeated sequence. In (Warnke-Sommer, Ali 2016) the Shannon's diversity score was used to capture the number of paths entering and exiting from a node. The formula (Shannon, 2001) used is given by:

$$H = -\sum_{i=1}^{n} \frac{w_i}{w_{total}} \ln\left(\frac{w_i}{w_{total}}\right),$$

where n is the number of incident edges, w_i is the weight of the ith edge, and w_{total} is the total weight of all of the incident edges. For a given node, this formula captures both the number of incident edges as well as the evenness of edges' weights. Recall that the weight of a given edge represents the

summed length of the overlaps between the reads represented by the endpoints of that edge. Thus the Shannon's diversity score takes the read coverage of paths entering and exiting a node into consideration. The greater the number of edges incident to a given node and the more even the edges' weights are, the higher that the Shannon's score of that node will be. Any incident edges that are spurious such as a single read with poor end qualities will have minimal impact on the Shannon index score as a edge representing an overlap with this read will likely be low coverage in comparison to other paths entering and exiting the node. Thus the Shannon's index score is robust against false positive edges in the hybrid graph due to sequencing error and artefacts.

Observe that in Fig. 3, the nodes within a short path distance of the node annotated with the transposase sequence are annotated with additional genetic features. Analysing the graph neighbourhood centered on a given node annotated with a transposase sequence can discover any gene features related to that transposase sequence. Given a node that is annotated with a transposase sequence and meets requirements described in the next paragraph, all its neighbouring nodes within a path length of three are examined to determine if they are annotated with any additional gene features. Any gene features that are found are considered to be transposase associated.

In this paper, nodes annotated with transposase sequences and had a Shannon's score greater than 1 were chosen for neighbourhood analysis. The minimum Shannon's score of one was chosen because nodes that have a Shannon's degree greater than one are part of multiple paths, indicating that they may be part of multiple genomic sequences. Greater occurrence in multiple sequence regions may indicate that a particular transposase sequence has spread between multiple genomic regions and may have been or is currently a contributor of bacterial evolution and adaptation. The ability to analyse genetic features associated with transposase sequences that occur across multiple genomic regions, whether in the same genome or across different genomes, may provide insight into functions that have allowed bacterial species to adapt to their niche environmental conditions.

3 RESULTS

This section presents the results of examining the gene features associated with transposase sequences in the Crohn's disease and healthy gut microbiome data sets. First, gene features associated with transposase sequences are extracted as described in the methods. Each gene feature is assigned a level 2 subsystem classification. The counts for each subsystem are normalized per data set by dividing the subsystem count by the total number of nodes explored by neighbourhood analysis and then multiplying by a thousand to get number of subsystem hits per thousand nodes. Next the Mann-Whitney U test was used to determine whether there is a significant difference in subsystem category hits for the Crohn's disease versus the healthy data sets. For subsystems that have significantly different number of hits between the Crohn's disease and healthy data sets, the distribution of bacterial genera in which the subsystems occurs was determined. Finally, three of the significant features: Mannose binding, Beta galactoside, and Heme binding in gram positives are explored in more detail in the context of their possible roles in the gut microbiome of Crohn's disease individuals and previous literature.

3.1 Subsystem Analysis

This work extends the research on antibiotic resistance genes applied to the thirteen data sets in (Warnke-Sommer, Ali, 2016). The same data sets are used in this analysis. Please see (Warnke-Sommer, Ali, 2016) for a description of the data sets' characteristics and distribution of bacterial genera.

To begin, all transposase associated gene features were assigned a level 2 subsystem classify-cation. If a given subsystem category did not have at least five hits for at least two of the data sets it was not included in subsequent analysis. Fig. 4 displays the median subsystem category hits per thousand nodes for the significant subsystems. Many of these subsystems are related to metabolic functions. For example, lactate fermentation had many more hits in the the Crohn's disease data sets in comparison to the healthy data sets. In the next section, we see that this function is found in *Lactobacillus*. According to (Beiko, Harlow, Ragan, 2005) metabolic genes may be preferentially transferred horizontally between bacterial types, allowing for adaptation to novel energy sources.

Here we will give a brief description of each of the subsystems that were significantly different in the Crohn's and healthy gut microbiome data sets.

- Alginate metabolism: Alginate is a polysaccharide found in the cell walls of brown algae and capsules of soil bacteria (Draget, Smidsrød, Skjåk-Bræk, 2000). Alginates are

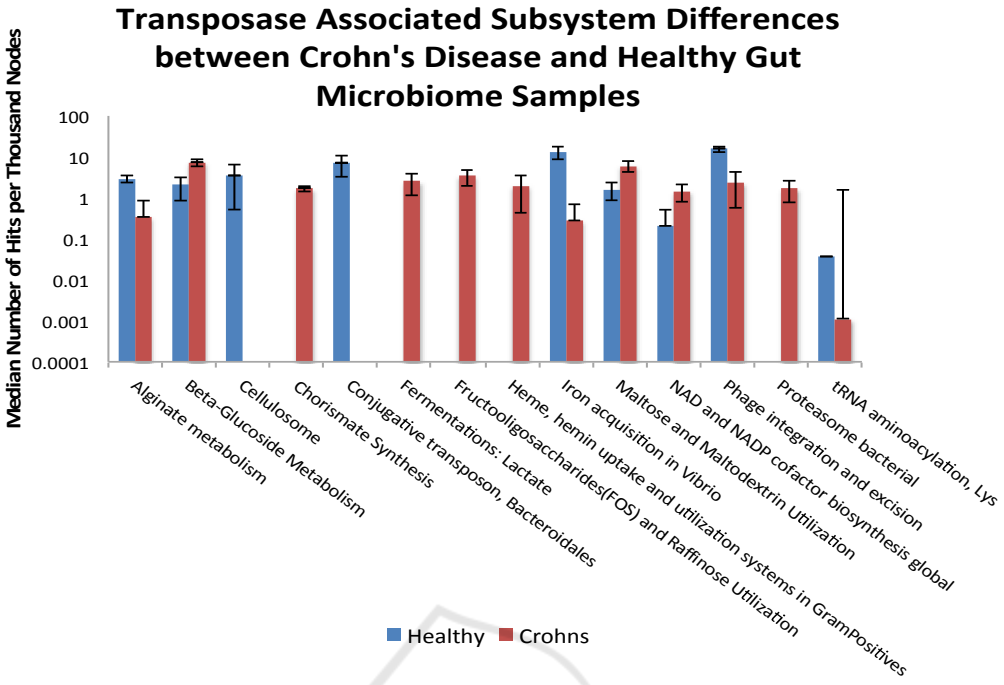

Figure 4: Transposase associated subsystem differences between Crohn's disease and healthy gut microbiome samples. Median gene hits to each significant subsystem are shown for Crohn's (red) and healthy (blue) samples.

commonly added to food as a thickener or stabilizer (Aliste, Vieira, Del Mastro, 2000).

- Beta-Glucoside metabolism: A glucoside is a glucose sugar molecule(s) that is attached to another non-sugar functional group. They are commonly found in plant material (Hollman et al, 1999).
- Cellulosome: Cellulosomes are large enzymes capable of digesting cellulose, a polysaccharide that is the major component of cell walls (Doi *et al*, 2003).
- Chorismate synthesis: Chorismate is a precursor in the synthesis of aromatic amino acids in most prokaryotes (Hopper, Rao, 2013).
- Conjugative transposon, Bacteroides: This is a DNA element that excises itself, forms a circular intermediate, and then reintegrates itself into the same genome or transfers between cells to a different genome (Salyers *et al*, 1995).
- Fermentations: Lactate: In bacteria, lactate is produced by the fermentation of carbohydrates Garvie, 1980).
- Fructooligosaccharides(FOS) and Raffinose Utilization: FOS and Raffinose are nondigestible oligosaccharides that can be found in plants (Barrangou *et al*, 2006).
- Heme hemin uptake and utilization systems in Gram Positives: Iron is often an enzyme cofactor in many prokaryotic biological processes. Many pathogens obtain iron through the uptake of heme (Anzaldi, Skaar, 2010).
- Iron acquisition in Vibrio: This includes Ton-B dependent transport of heme in Gram Negatives (Stojiljkovic, Hantke, 1992).
- Maltose and Maltodextrin Utilization: Maltose is a disaccharide that can be formed from the digestion of starch (Nichols *et al*, 2003). Maltodextrin is a modified starch that is commonly used as a food additive (Nickerson, McDonald, 2012).
- NAD and NADH cofactor biosynthesis global: The pyridine nucleotide redox pair NAD/NADH is an essential cofactor for all living organisms (Kurnasov *et al*, 2003).
- Phage integration and excision: Bacteriophages are capable of integrating their DNA into a host genome (Paatero *et al*, 2008).
- Proteasome bacterial: These perform protein degradation in bacteria to maintain homeostasis (Butler *et al*, 2006.
- tRNA aminoacylation: Aminoacyl-tRNA synthetases catalyse the addition of amino acid to a transfer RNA (Park, Schimmel, Kim, 2008).

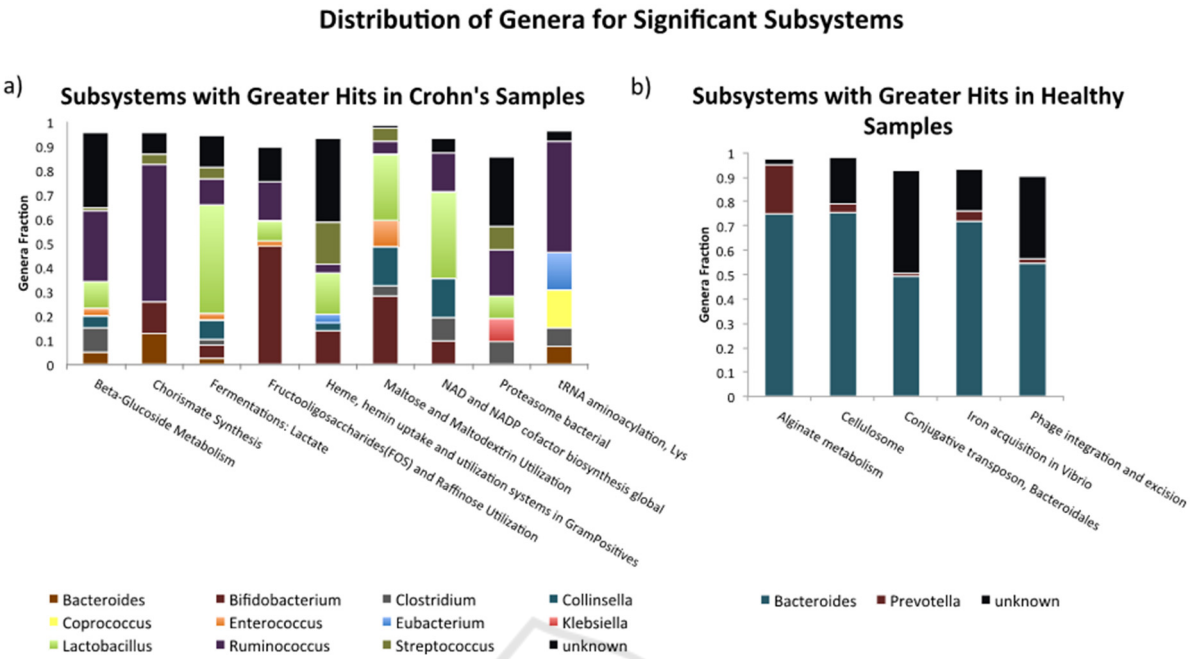

Figure 5: Distribution of Genera for Significant Subsystems. (A) shows the genera distribution for subsystems with a greater number of hits in Crohn's samples and (B) shows the genera distribution for subsystems with a greater number of hits in the healthy samples.

Fig. 5 displays the genera distribution for each of the significant subsystems. In the healthy samples, most of the transposase-associated subsystems were found in *Bacteroides.* The transposase-associated subsystems that were found in the Crohn's disease samples were more distributed across bacterial genera. For an example, the subsystem Fermentation:Lactate was found mostly in *Lactobacillus.* The subsystem Fructooligosaccharides (FOS) and Raffinose utilization was found commonly in *Bifidobacterium.* Species of *Bifidobacterium* are known to be capable of fermenting Fructooligosaccharides (Rossi *et al*, 2005).

Finally, Table 1 displays the subsystems that were not found to be significant between the Crohn's disease and healthy gut microbiome data sets. A few of these subsystems, such as At1g69340, appear to come from *Arabidopsis thaliana.* We would like to note that the SEED subsystem contains prokaryote orthologs of *Arabidopsis thaliana* (Gerdes *et al*, 2011).

3.2 in-Depth Analysis of Pathologically Relevant Subsystems

In this section, an in-depth summary of the Heme, Hemin Uptake and Utilization, Maltose and Maltodextrin Utilization, and Beta-Glucoside Metabolism subsystems is presented.

The subsystem Heme, Hemin, Uptake and Utilization are found most commonly in *Streptococcus and Lactobacillus* as well as in unknown genera. Iron is a cofactor of many enzymes in living systems. Its uptake is essential for pathogenic infection (Anzaldi, Skaar, 2010). In mammalian systems, the most abundant form of iron is bound in heme. Thus many bacterial systems have developed for the uptake of heme. The availability of host heme to bacterial pathogens can greatly increase the difficulty of clearing an infection (Contreras *et al*, 2014).

The subsystem Maltose and Maltodextrin is found in *Bifidobacterium, Lactobacillus, Collinsella,* and *Enterococcus*. Some studies have found that the food additive Maltodextrin can alter the bacterial homeostasis of the intestine (Nickerson, Chanin, McDonald, 2015). Maltodextrin has also been found to increase adhesion of Crohn's disease associated ecoli (Nickerson, McDonald, 2012) and promote *Salmonella* mucosal colonization and survival (Nickerson *et al*, 2014).

According to Fig. 5 beta-glucoside subsystem functions are found most commonly in *Ruminococcus, Clostridium,* and *Lactobacillus* genera. A large portion of the beta-glucoside functions occurs in unknown genera. Glucosidases are capable of producing metabolites that are implicated in colon cancer (Sobhani *et al*, 2013).

Table 1: Subsystems that were not significantly different between Crohn's disease and healthy gut microbiome samples. For each subsystem the median number of hits per thousand nodes is shown for the Crohns disease (C) and healthy (H) gut microbiome samples.

Subsystem	C.	H.	Subsystem	C.	H.
16S rRNA modification within P site of ribosome	1	1	Entner-Doudoroff Pathway	3	2
5-FCL-like protein	9	7	Folate Biosynthesis	5	2.5
ABC transporter oligopeptide (TC 3.A.1.5.1)	1	0	Galactosylceramide and Sulfatide metabolism	9	9
Alanine biosynthesis	7	0	Glycogen metabolism	6	2.5
Alkanesulfonate assimilation	0	4	Group II intron-associated genes	6	3.5
Ammonia assimilation	7	3.5	Heat shock dnaK gene cluster extended	3	6.5
Aromatic conversions and predicted Co2 transporter cluster	5	1	High affinity phosphate transporter and control of PHO regulon	10	1
At1g69340 At2g40600	2	3	Histidine Biosynthesis	4	0.5
At3g21300	8	0.5	L-fucose utilization	0	2.5
At5g63420	7	1	L-rhamnose utilization	2	3
ATP-dependent Nuclease	5	0	Lactose and Galactose Uptake and Utilization	9	2
Bacterial Cell Division	9	3.5	Mannose Metabolism	3	1.5
Bacterial Cytoskeleton	9	1	Methionine Biosynthesis	8	5
Beta-lactamase	1	3	Multidrug Resistance Efflux Pumps	10	5
Beta-lactamase cluster in Streptococcus	2	2	Oxidative stress	2	0.5
C jejuni colonization of chick caeca			Potassium homeostasis	8	3.5
Calvin-Benson cycle	6	2	Purine conversions	15	2.5
Cell Division Subsystem including YidCD	5	1	Restriction-Modification System	15	17.5
Cell division-ribosomal stress proteins cluster	3	1	Ribonucleotide reduction	4	0.5
Chitin and N-acetylglucosamine utilization	5	5.5	Ribosome LSU bacterial	10	1.5
D-Galacturonate and D-Glucuronate Utilization	4	7.5	Sialic Acid Metabolism	4	0
De Novo Pyrimidine Synthesis	4	1	Streptococcal group antigen operons	5	0
DMT transporter	2	1	Tetracycline resistance, ribosome protection type	3	0
DNA Repair Base Excision	7	2.5	Transcription factors bacterial	2	2.5
DNA repair, bacterial	13	4	tRNA modification Archaea	2	2
DNA repair, bacterial RecFOR pathway	0	4.5	Universal GTPases	5	2

Fig. 6 provides a more granular view of the level 3 subsystems classification counts for Heme, Hemin Uptake and Utilization, Maltose and Maltodextrin Utilization, and Beta-Glucoside Metabolism.

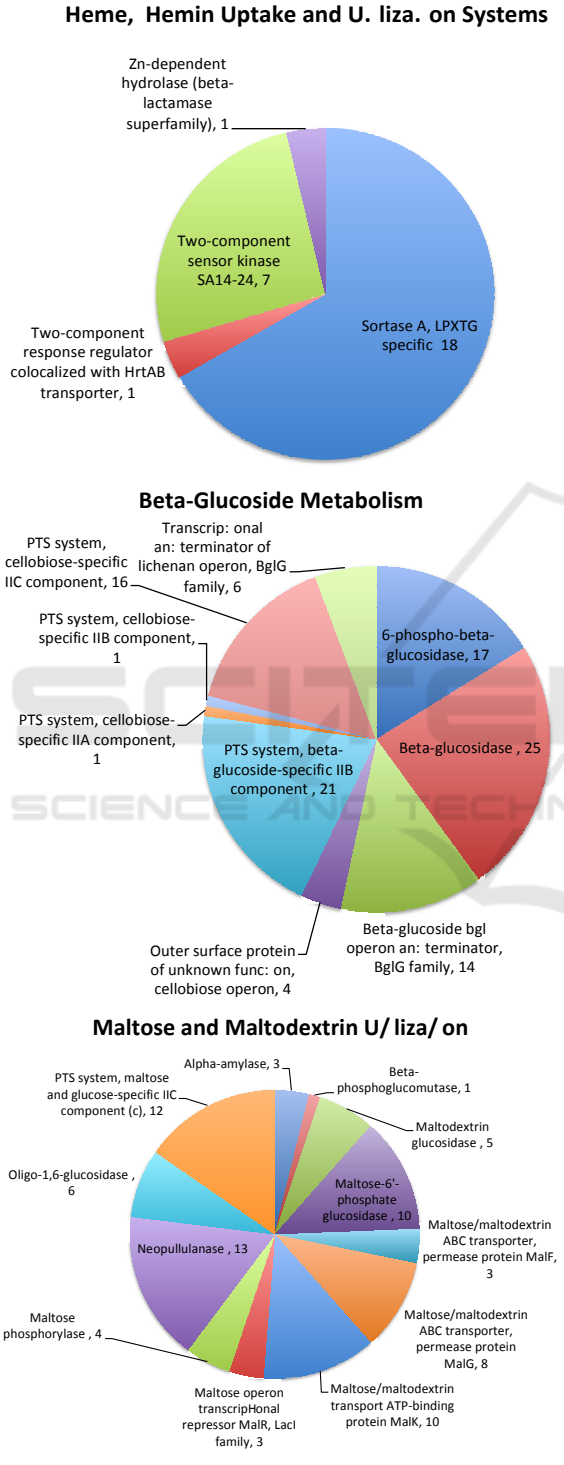

Figure 6: Level 3 subsystem counts for Heme, Hemin Uptake and Utilization Systems, Maltose and Maltodextrin Utilization., and Beta-Glucoside Metabolism.

4 CONCLUSIONS

In conclusion, this paper presents an extension of previous work using graph mining to explore transposase associated genetic elements in Crohn's disease and health individuals. Graph mining is shown to be a powerful tool that can be applied to extract biologically relevant features of the input dataset.

Flat fasta files of contigs loose information, as relationships between the contigs are lost in flat files. The hybrid graph maintains all of the global and local relationships within its structures. This is especially important for difficult to assemble regions such as repetitive regions, including transposes. Many assemblers produce fragmented results in these areas, as it is difficult to place repetitive regions. In contrast, the hybrid graph maintains all possible relationships allowing for extraction of information even from difficult to assemble areas.

Results demonstrate several important transposase associated genetic features that are more prevalent in Crohn's disease gut microbiome samples than healthy samples. Several of these functions have been implicated in previous research as biologically relevant to Crohn's disease and associated conditions. The results in this paper provide insights into gene features that may allow gut bacteria to adapt to their ecological niches.

ACKNOWLEDGEMENTS

Thank you to the UNO Bioinformatics lab for their valuable input and discussion on this paper.

REFERENCES

Aliste, A. J., Vieira, F. F., Del Mastro. N. L. Radiation effects on agar, alginates and carrageenan to be used as food additives. *Radiation Physics and Chemistry* 57.3 (2000): 305-308.

Anzaldi, L. L., Skaar, E. P. Overcoming the heme paradox: heme toxicity and tolerance in bacterial pathogens. *Infection and immunity* 78.12 (2010): 4977-4989.

Barrangou, R., et al. Global analysis of carbohydrate utilization by Lactobacillus acidophilus using cDNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America* 103.10 (2006): 3816-3821.

Beiko, R. G., Harlow T. J., Ragan, M. A. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of*

America 102.40 (2005): 14332-14337.

Buchfink, B., Xie, C., Huson. D. H., Fast and sensitive protein alignment using DIAMOND. *Nature methods* 12.1 (2015): 59-60.

Butler, S. M., et al. Self□compartmentalized bacterial proteases and pathogenesis. *Molecular microbiology* 60.3 (2006): 553-562.

Contreras, H., et al. Heme uptake in bacterial pathogens. *Current opinion in chemical biology* 19 (2014): 34-41.

David, L. A., et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505.7484 (2014): 559-563.

Doi, R. H., et al. Cellulosomes from mesophilic bacteria. *Journal of bacteriology* 185.20 (2003): 5907-5914.

Draget, Ingar K., Smidsrød, O., Skjåk□Bræk, G. Alginates from algae. *Biopolymers Online* (2005).

Foster, A., McVey Neufeld. K. A., Gut–brain axis: how the microbiome influences anxiety and depression. *Trends in neurosciences* 36.5 (2013): 305-312.

Garvie, E. I. Bacterial lactate dehydrogenases. *Microbiological reviews* 44.1 (1980): 106.

Gerdes, S., et al. Synergistic use of plant-prokaryote comparative genomics for functional annotations. *BMC Genomics* 12.S2. (2011): doi: 10.1186/1471-2164-12-S1-S2

Ghosh, T. S., et al. In silico analysis of antibiotic resistance genes in the gut microflora of individuals from diverse geographies and age-groups. *PLoS One* 8.12 (2013): e83823.

Hollman, P.C. H, et al. The sugar moiety is a major determinant of the absorption of dietary flavonoid glycosides in man. *Free radical research* 31.6 (1999): 569-573.

Hopper, P. P. W., Rao. R., Comprehensive database of Chorismate synthase enzyme from shikimate pathway in pathogenic bacteria. *BMC Pharmacology and Toxicology* 14.1 (2013): 1.

Kamada, N., et al. Control of pathogens and pathobionts by the gut microbiota. *Nature Immunology* 14.7 (2013): 685-690.

Karypis, G., Kumar, V., A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing* 20.1 (1998): 359-392.

Kurnasov, O., et al. NAD biosynthesis: identification of the tryptophan to quinolinate pathway in bacteria. *Chemistry & biology* 10.12 (2003): 1195-1204.

LeBlanc, Jean Guy, et al. "Bacteria as vitamin suppliers to their host: a gut microbiota perspective." *Current opinion in biotechnology* 24.2 (2013): 160-168.

Leplae, R., et al. ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic acids research* 32.suppl 1 (2004): D45-D49.

Manichanh, C., et al. The gut microbiota in IBD. *Nature Reviews Gastroenterology and Hepatology* 9.10 (2012): 599-608.

Meyer, F., Overbeek, R., Rodriguez., A., FIGfams: yet another set of protein families. *Nucleic acids research* 37.20 (2009): 6643-6654.

Miller, J. R., Koren, S., Sutton. G., Assembly algorithms for next-generation sequencing data. *Genomics* 95.6 (2010): 315-327.

Nagarajan, Niranjan, and Mihai Pop. "Sequence assembly demystified." *Nature Reviews Genetics* 14.3 (2013): 157-167.

Nichols, B. L., et al. The maltase-glucoamylase gene: common ancestry to sucrase-isomaltase with complementary starch digestion activities. *Proceedings of the National Academy of Sciences* 100.3 (2003): 1432-1437.

Nickerson, K. P., McDonald. C. Crohn's disease-associated adherent-invasive Escherichia coli adhesion is enhanced by exposure to the ubiquitous dietary polysaccharide maltodextrin. *PLoS One*7.12 (2012): e52132.

Nickerson, K. P., et al. The dietary polysaccharide maltodextrin promotes Salmonella survival and mucosal colonization in mice. *PloS one*9.7 (2014): e101789.

Nickerson, K.P., Chanin, R. McDonald., C., Deregulation of intestinal anti-microbial defense by the dietary additive, maltodextrin. *Gut microbe* 6.1 (2015): 78-83.

O'sullivan, O., et al. Comparative genomics of lactic acid bacteria reveals a niche-specific gene set. *BMC microbiology* 9.1 (2009): 1.

Ochman, H., Lawrence, J.G., and Groisman, E.A. Lateral gene transfer and the nature of bacterial innovation. Nature 405.6784 (2000): 299-304.

Overbeek, R., et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research* 42.D1 (2014): D206-D214.

Paatero, A. O., et al. Bacteriophage Mu integration in yeast and mammalian genomes. *Nucleic acids research* 36.22 (2008): e148-e148.

Park, S.G., Schimmel, P., Kim., S. Aminoacyl tRNA synthetases and their connections to disease. *Proceedings of the National Academy of Sciences* 105.32 (2008): 11043-11049.

Rho, M., Tang, H., Ye, Y., FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research* 38.20 (2010): e191-e191.

Rhodes, J. M., and B. J. Campbell. Inflammation and colorectal cancer: IBD-associated and sporadic cancer compared. *Trends in molecular medicine* 8.1 (2002): 10-16.

Rossi, M., et al. Fermentation of fructooligosaccharides and inulin by bifidobacteria: a comparative study of pure and fecal cultures. *Applied and environmental microbiology* 71.10 (2005): 6150-6158.

Roy, A., Lichtiger, S. Clostridium difficile Infection: A Rarity in Patients Receiving Chronic Antibiotic Treatment for Crohn's Disease. Inflammatory bowel diseases 22.3 (2016): 648.

Salyers, A. A., et al. Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. *Microbiological reviews* 59.4 (1995): 579-590.

Shannon, C. E., A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1 (2001): 3-55.

Sobhani, I., et al. Microbial dysbiosis and colon

carcinogenesis: could colon cancer be considered a bacteria-related disease?. *Therapeutic advances in gastroenterology* 6.3 (2013): 215-229.

Sommer, F., Bäckhed, F. The gut microbiota—masters of host development and physiology. *Nature Reviews Microbiology* 11.4 (2013): 227-238.

Stecher, B., Maier, L., Hardt. WD,. 'Blooming' in the gut: how dysbiosis might contribute to pathogen evolution. Nature Reviews Microbiology 11.4 (2013): 277-284.

Stojiljkovic, I., Hantke. K. Hemin uptake system of Yersinia enterocolitica: similarities with other TonB-dependent systems in gram-negative bacteria. *The EMBO journal* 11.12 (1992): 4359.

Turnbaugh, P. J., et al. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444.7122 (2006): 1027-131.

Warnke-Sommer, J., and Ali, H. "Graph mining for next generation sequencing: leveraging the assembly graph for biological insights." *BMC genomics* 17.1 (2016): 1.

Warnke, J., Ali H,. Focus: a new multilayer graph model for short read analysis and extraction of biologically relevant features. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014.