WAGENINGEN
UNIVERSITY & RESEARCH

In collaboration with:

University Medical Center Groningen

MSc thesis report

SPECIALISED METABOLITE PATHWAY ASSOCIATION STUDY ON INFLAMMATORY BOWEL DISEASES

Wouter Lokhorst, BSc

BIOINFORMATICS GROUP, WAGENINGEN UR, THE NETHERLANDS

Specialised metabolite pathway association study on inflammatory bowel diseases

Wouter Lokhorst[1], Dick de Ridder[2] and Marnix Medema[2]

1. MSc student, Department of Bioinformatics, Wageningen UR, Wageningen, The Netherlands
2. Supervisor, Department of Bioinformatics, Wageningen UR, Wageningen, The Netherlands

Keywords: specialised metabolism, microbiome, bacteria, metagenome, inflammatory bowel disease, biosynthetic gene cluster, gene cluster family, classification

Abstract

A major factor of influence for inflammatory bowel diseases are the bacteria residing in the gut. This is caused by changes in the metabolism of the microbiome, such as a lowered short-chain fatty acid availability to the host. It is shown here that pathways of these specialised metabolites can be detected from metagenomics data through automated analyses and that healthy and diseased individuals can be accurately classified based on biosynthetic gene cluster abundances. Nonetheless, improvements on the pipelines could be beneficial to understanding the mechanisms of the diseases.

Introduction

The inflammatory bowel diseases (IBDs) are a group of chronic inflammatory autoimmune disorders that involve environmental, host genetic, and microbial factors. They can be differentiated by the region of affected tissue of the gastrointestinal tract and the severity of their symptoms [1]. The two most typical forms are Crohn's Disease (CD) and Ulcerative Colitis (UC). Correct diagnosis can be difficult given their high similarity and no cures have been found so far [2]. A major factor of influence for IBDs are the bacteria residing in the gut [3][4]. For example, a lack of bacteria producing short-chain fatty acids (SCFAs), can negatively influence SCFA pools available to the host [5]. This causes gut mucus permeability to increase. Hence, SCFAs are associated with the pathogenesis of IBD [6]. In turn, the host genetics influence the microbial species composition [7]. Dominant phyla present in the gut of healthy Dutch individuals are, amongst others, *Bacteroidetes* (8.1% and 7.4% based on metagenomics sequencing (MGS) and 16S rRNA, respectively) and *Actinobacteria* (22.3% and 12.3%, MGS and 16S rRNA) [8]. In IBD patients however, colonic samples are depleted of *Bacteroidetes* and enriched in *Actinobacteria* [9]. This directly relates to the beneficial effects of the immunostimulatory and immunoregulatory activities of *Bacteroides fragilis*, i.a. through production of polysaccharide A [10]. This and many other saccharides, along with the SCFAs, belong to the specialised metabolism of bacteria. Under disease conditions as described here, microbial function can be affected much more than species composition [11]. Oppositely, human gene expression can also be influenced by the gut microbiome [12] and changes in the specialised metabolism of the microbiome could affect the onset and course of a disease.

Up until now, research on the functional changes of specialized metabolism in the gut microbiome has not been plentiful [13]. This is partly caused by a lack of tools for large-scale analysis of genomic data regarding metabolism. With the development of antiSMASH [14] in 2011 (currently at version 4.0 [15]), the first step towards these analyses was made. It is used to discover Biosynthetic Gene Clusters (BGCs), which indicate potential production of specialised metabolites and has already led to the discovery of a family of antibiotics in the human microbiome [16]. A subsequent analysis on the Human Microbiome Project (HMP) data has revealed tremendous numbers of BGCs (which included ClusterFinder for prediction of BGCs in genomes [17], nowadays integrated into antiSMASH). Among the results was an oligosaccharide ligand for nucleotide-binding oligomerization domain–containing protein 2 (NOD2) [18], which is one of the genes strongly associated with CD [19].

More steps towards a fully automated analysis were taken with the release of BBSplit [20] (read

binning tool for metagenomes that uses the short-read aligner BBMap [21]) to facilitate multi-reference input for mapping and finally, the release of BiG-SCAPE [22] (generates similarity networks for BGCs), which makes it possible to create Gene Cluster Families (GCFs) from BGCs. These GCFs are highly related groups of BGCs that produce metabolites with the same function. As such, BiG-SCAPE allows for homology-based comparisons between BGCs. This shows that tools have become available to facilitate metabolite analyses through metagenomics.

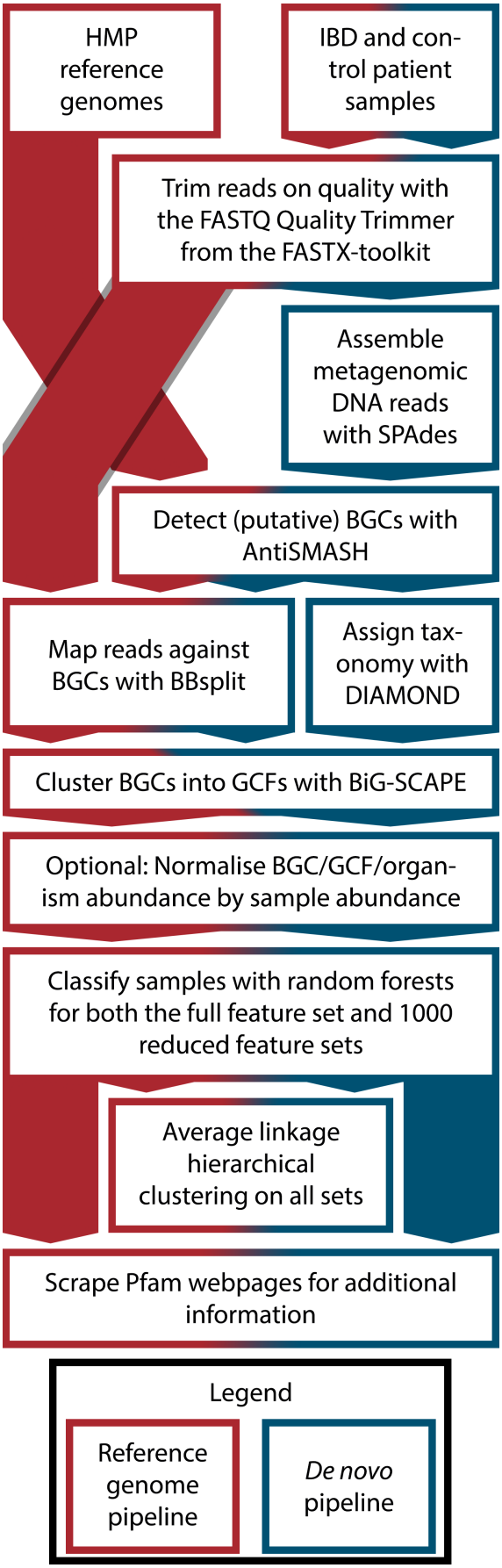
Here, those tools are used to elucidate changes in the genomic abundances of specialized metabolite pathways of the human gut microbiome between healthy and diseased individuals through an automated pipeline. Two methods are attempted. One uses reference genomes as a means of guided discovery, by which well-annotated BGCs are thought to be found. The other uses a *de novo* approach, meaning that any discovered BGCs come from the metagenomic data of the individuals themselves. This can lead to more detailed discovery of the specialized metabolites present in the gut.

Methods

Two analysis pipelines were constructed for detecting changes in specialised metabolite production with metagenomics data (Figure 1). Tool version numbers, used commands and scripts, etc. can be found in Supplementary file 1.

The first pipeline detects BGCs in reference genomes. It will therefore be called the reference genome pipeline from now on. In this study, the complete set of 457 genomes of the HMP gastro-intestinal tract reference dataset [23] have been selected for analysis. Although the HMP does contain microbial genomes extracted from other regions of the human body, those have not been included here because of time and resource constraints. Inclusion of the other genomes would result in a higher amount of detected BGCs. Around 44,000 BGCs have been detected in the entire HMP dataset [16], but this will probably include many similar and redundant BGCs, or even irrelevant to this study (BGCs of species that do not occur in the gastro-intestinal tract).

The other pipeline detects BGCs in the contigs of assembled reads, therefore dubbed the *de novo* pipeline. The reference genome pipeline yields BGCs with a better annotation than the *de novo* pipeline, yet less certainty regarding the presence



of a gene in the real genome. To elaborate, species in the sample may have lost one or more genes of

a BGC and gained others. The *de novo* pipeline therefore yields more accurate results, but far more often unannotated.

The reference genome pipeline

The University Medical Center Groningen (UMCG) has provided a set of MGS samples (paired-end reads) taken from the human gut, consisting of 65 individuals (20 healthy and 45 diseased). The reads were trimmed to a minimum quality of 28 with the FASTQ Quality Trimmer from the FASTX-toolkit [24]. Afterwards, they were mapped to the BGCs from the HMP. Multi-mapped reads were also included for determining the BGC abundance. Uniquely mapped reads were counted once, but multi-mapped read counts were divided by the number of BGCs they map to. This was done to prevent distorted organism abundances. Non-normalised read counts were used here (as well as normalised), because there is plausible evidence that microbial load has predictive potential for IBD [25]. It is suggested that read counts of organisms normalised by the total read count per sample give a false indication that certain organisms have been enriched while, in fact, they are not. Rather, read counts normalised by cell counts should be used, though such data was not available here. Therefore, it is impossible to account for possible technical sequencing artefacts and no definitive choice can be made between normalised and non-normalised read counts, so both are included in the analysis. To reduce over-fitting to sample-specific BGCs, the maximum read count of a BGC must at least be 10 for a single sample. However, BGCs below the threshold were still included in GCF clustering, because they might hold useful information about related BGCs and influence which BGCs end up in the same GCF. BiG-SCAPE generates similarity networks of these GCFs with several distance cut-offs. The network that shows the smallest amount of multi-product GCFs while keeping the number of GCFs low, was chosen for further analysis. Here, a cut-off of 0.55 was used. After this clustering, reads were binned to the GCFs. Reads were also binned to the species of origin (the genome from which a BGC was predicted). Classification was then performed with each of these binning methods and on the BGCs without binning.

A random forest algorithm with bootstrapping [26] was used to classify samples as either diseased or healthy based on the BGC abundances. The random forest was created 1000 times, with

Figure 1. Schematic overview of the pipelines.

randomly divided training and test sets (50 and 15 samples, respectively) each time. Each random forest consists of 1000 decision trees and these were trained on either BGCs, GCFs, or organisms. The average percentage correctly classified samples of all random forests of one binning method combined is the final result for that method. Reduced feature sets were used as well. These were randomly sampled 1000 times from the entire feature set (BGCs, GCFs or organisms) and a random forest was created 100 times to determine which reduced set of 25 features performed best. Again, the average result of a method is used as the final result. Since future additional samples may differ from the samples used here, average linkage hierarchical clustering was used to see if the current diagnostic classes were resembled without the class labels. This is indicative for the likelihood of new samples being correctly classified. Furthermore, this clustering method does not assign weights to the features. Thus, it is not likely to divide control and IBD samples correctly with the full feature set. However, it does add another measure for robustness of the features of the reduced set, because it can be considered to be a simulation of shifting weights (e.g. when adding more samples). In that regard, having a smaller feature set gives an advantage, because it contains less noise and, when chosen correctly, is probable to be relevant to IBD.

More information was needed to check the relevance to IBD for the highest-scoring reduced feature sets. Pfam [27] webpages (and the integrated InterPro [28] webpages) of the domains contained in their BGCs (in case of GCFs and organisms, their BGCs were obtained first) were searched for words containing one or more of the following terms: sacchar, fatty, inflamma, immune and antib. These give further clues about the function of the BGC. These results, along with the percentage of correctly classified samples, are indicators for the suitability of the classification model as a means of identifying specialised metabolites involved with IBD.

The de novo pipeline

Only additional stages to the reference genome pipeline are explained here, see Figure 1 and the previous chapter for an explanation on the other stages.

After trimming, the reads were assembled with SPAdes [29]. metaSPAdes [30] was not chosen here due to a mistake in determining the type of reads (single-end instead of paired-end). There was not enough time to re-run the assembly. BGCs were detected in the contigs and taxonomy was assigned to the BGCs with DIAMOND [31]. The taxonomy is needed for binning to organisms. At the same time, the reads were mapped to the BGCs, equal to the reference genome pipeline. The remaining stages were also equal to that pipeline.

Results and discussion of the reference genome pipeline

The reference genome pipeline started with running antiSMASH against each reference genome separately, to ultimately result in 10455 BGCs. Reads were trimmed and mapped against these BGCs and a total of 7341 BGCs were present in the samples. The BGCs were binned into 3074 GCFs by BiG-SCAPE and 2991 GCFs had up to 10 BGCs in them. The largest three GCFs contained 243, 193 and 124 BGCs (Figure 3) and their most often predicted product types were putative, saccharide and putative, respectively. The large number of predicted putative BGCs underlines the need for enhanced annotation of BGCs, which will be discussed later. The BGCs were binned to organisms as well and 446 genomes of the original 457 had reads mapped against them. The three genomes with the most BGCs were Streptomyces sp. HGB0020, Streptomyces sp. HPH0547 and Pseudomonas sp. 2_1_26, with 119, 97 and 58 BGCs, respectively (Figure 2).

BGC abundances decrease in IBD patients

The average BGC abundance across control samples is 310.74 reads, whereas it is 182.40 reads for IBD samples, with standard deviations of 137.12 and 67.66 reads, respectively. Briefly, a student t-test was run. The resulting p-value of less than 0.0001 suggests that technical variation cannot cause this alone. Another, yet unlikely, cause is that DNA could be easier to extract from the control samples. Therefore, it is probable that read count itself counts as a biomarker for IBD. Still, the decrease might not be evenly distributed over the BGCs and classification can help uncover this.

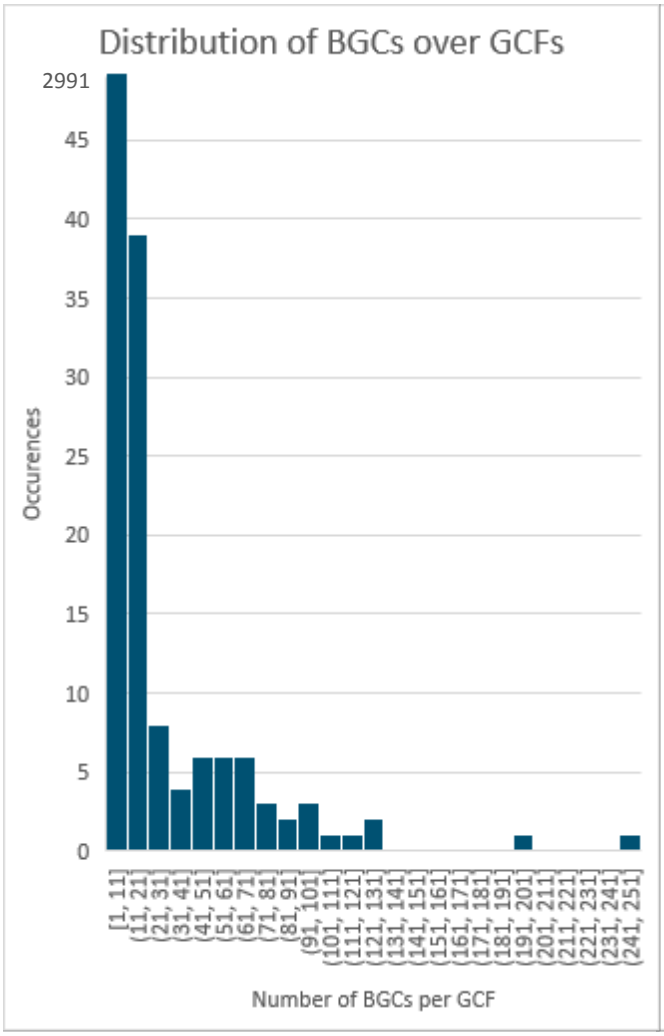


Figure 3. Distribution of BGCs over GCFs.

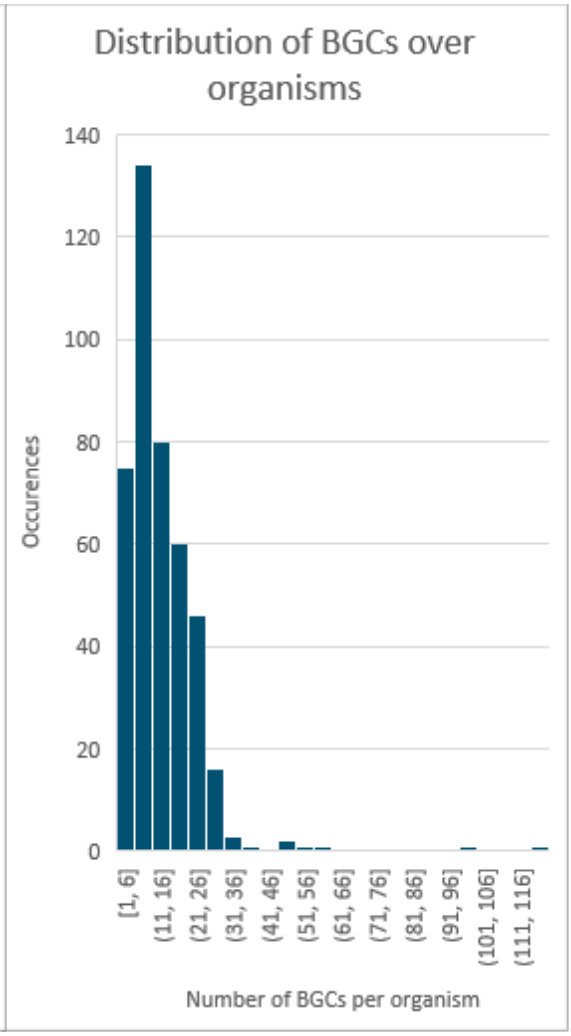


Figure 2. Distribution of BGCs over organisms.

Accurate classification of IBD is possible

The classification was done with several different binning methods, on full and reduced feature sets, with non-normalised and normalised read counts (Table 1), from here on referred to as "settings". The highest number of correct classifications was achieved with the reduced feature set, of which the non-normalised read counts without binning scored best. Near-equal percentages were achieved with the non-normalised read counts of binning by GCFs
and by organisms. The increased overall score of the reduced feature set settings might be due to over-fitting to some of those features (Supplementary file 2). For example, the feature with the second highest weight for the normalised reduced feature set without binning has average read counts of 0.0002% for IBD and 0.0006% for control samples (Figure 4). Such features are not likely to retain their classification weight when the number of samples is increased. On the other hand, the non-normalised reduced feature set with binning by organisms has *Roseburia intestinalis* L1-82 as one of the features. This is interesting, because a decrease in *Roseburia spp* leads to a higher IBD genetic risk score [7] and a decrease is measured here as well (an average read count of 5541.97 for IBD samples versus 7919.98 for control samples). This shows that even though the binning by organisms was not as successful as the other methods, valuable information is still gained from it. Binning BGCs as organisms can skew the abundance of those organisms, though it does make it easier to extract the most important BGCs per organism. Follow-up research could investigate if mapping the reads directly to the genomes yields different results. Another reason the classification on the organisms performed worse, was because

Table 1. Percentage correctly classified samples per setting.

BGCs binned as		Full feature set		Reduced feature set	
		Non-normalised read counts	Normalised read counts	Non-normalised read counts	Normalised read counts
No binning	**Training set**	100%	100%	100%	100%
	Test set	90.05%	84.54%	95.47%	90.53%
GCFs	**Training set**	100%	100%	100%	100%
	Test set	89.76%	84.82%	94.47%	91.33%
Organisms	**Training set**	100%	100%	100%	100%
	Test set	88.76%	82.13%	93.73%	90.00%

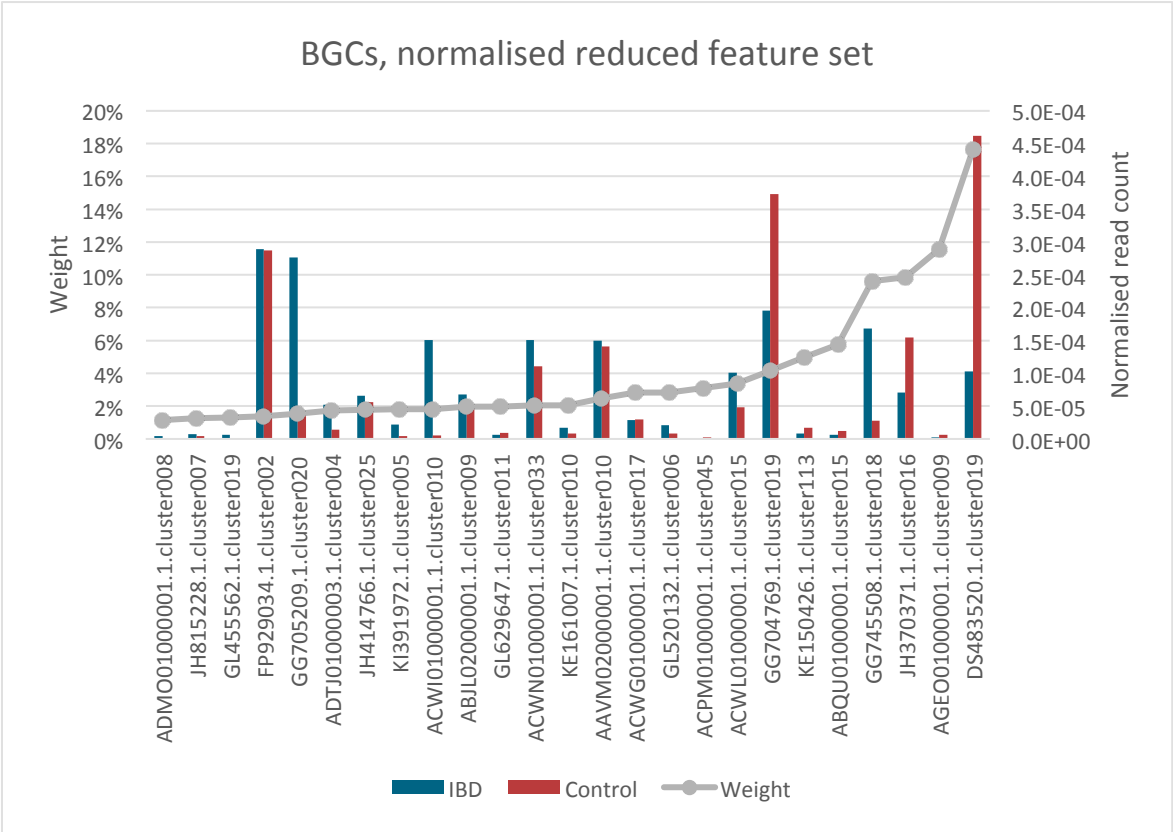


Figure 4. Feature weights and normalised read counts for the reduced feature set without binning.

the genomes used in the reference genome pipeline were not always the actual organism present in the sample. A BGC can be found in the reference organism, but the actual organism might have lost it, while retaining the other genes of the BGC. Possibly, another organism present in the sample might have gained it. Such genomic rearrangements have their effect on the total organism counts and subsequently on its classification performance. The de novo pipeline is likely to perform better in this regard.

Forecasts of the classification score

Future research may include more or different samples and it is therefore necessary to know how well the classification will perform during follow-up research. To investigate this, clustering was performed with the same settings as classification, using both the full feature set as well as the reduced feature set from classification. The resulting dendrograms can be found in Supplementary file 3. None of the settings result in a clear distinction between control and IBD samples, meaning that false classifications are bound to happen if future samples do not resemble the currently included samples.

Another measure for the classification performance for the reduced feature sets, was the biological background of the features. Their functions could affirm or refute their eligibility as a biomarker. Protein domains were retrieved for each gene in each BGC for each of the aforementioned settings (for the GCFs and organisms, BGCs were first extracted). This yielded 702 unique domains. Next, Pfam database webpages of these domains were searched for the terms sacchar, fatty, inflamma, immune and antib. The full list of these domains with links to their Pfam webpages and the weights of the BGCs they derive from, for each of the six settings, can be found in Supplementary file 4. This search has provided a lot of extra information about the BGCs and helps in determining their products. Take GG730105.1.cluster032 for example. This putative BGC has the term sacchar in 11 out of 13 of its Pfam webpages and is part of the GCF cf_saccharide_235. This combination of web scraping and BiG-SCAPE gives great confidence that this BGC produces a saccharide metabolite. False positives occur as well, such as the putative BGC AGEO01000001.1.cluster009, which has the term antib on 6 of its 21 domain webpages, but seems to be involved in transcription, rather than antibiotic or antibacterial activities. Furthermore, this BGC is the second-most important feature for the normalised reduced feature set without binning. It is safe to say that this BGC is not likely to keep its classification weight when the sample size is increased, as it is easily interchangeable with other BGCs that indicate an overall loss of microbes in the gut. It also means that the current web scraping method is far from perfect. Development of a tool for accurately scraping both Pfam and InterPro databases is advised.

Currently, the non-normalised reduced feature set with binning by GCFs seems to be the best method. The features and their weights are least likely to change out of the six investigated settings. This robustness is caused by adequate numbers of reads for the GCFs (most importantly, the ones with a higher weight) and because they consist of multiple BGCs. Even when one BGC of a GCF becomes absent, another similar member may arise and take its place, keeping the overall GCF read count equal. Accurate GCFs are crucial here. As such, future research may incorporate the MIBiG database [32] (which holds well-annotated BGCs) in the GCF construction stage, or rather, including the database should become available as an option in BiG-SCAPE.

Results and discussion of the de novo pipeline

The *de novo* pipeline seems promising in theory, but in practice it required too much time and resources to be fully run during this research study. The results obtained thus far are reported here.

After trimming, the reads were assembled per sample. This resulted in an average N50 of 3797 for IBD samples and 1463.58 for control samples (**Error! Reference source not found.**). Simultaneously, IBD samples have a larger amount of contigs than the control samples. In all of the samples combined, 91 BGCs were detected. This is a stunningly low number compared to the 10455 with the reference genome pipeline. As mentioned in the Methods chapter, this is due to the use of the wrong assembly tool. This also means that nothing can be said here about differences between IBD and control samples with this pipeline, nor about differences between the two pipelines until it has been re-run with the correct tool.

	IBD	Control
# contigs (>= 0 bp)	**135637.42**	**82606.98**
# contigs (>= 1000 bp)	**16869.49**	**11445.31**
# contigs (>= 10000 bp)	**856.02**	**548.07**
N50	**3797.00**	**1463.58**

Table 2. Assembly statistics.

Wisdom lies in the sands of time

Running times and server loads were major constraints during this study and made it impossible to run this pipeline within reasonable time. Briefly, the code of BiG-SCAPE was improved, which resulted in a running time of roughly 1200 seconds with this dataset. This is an improvement of around 25% compared to the old code. Further improvements are still possible, but may require part of the code to be written in C++ rather than Python. Such improvements may also be possible for antiSMASH. For example, by using the best programming language for a certain process (e.g. Perl for parsing, Python or C++ for calculations and Bash scripts when executing other programs). Nonetheless, it is a great tool for detecting BGCs and such problems only arise when dealing with large datasets in combination with running many of the prediction modules of the tool.

If future studies on this subject apply the suggested improvements and recommendations given here, it will lead to fast comparative analyses of metagenomic samples. This will include a well-performing, robust classification model for distinguishing between diseased and healthy individuals and highly-accurate detection of involved specialised metabolite pathways. In turn, these pathways can provide a role in development of medicines, in deciding which and what dosage of medicine to use on a patient or in dietary advice for both diseased and healthy people.

References

[1] Phenotype of Inflammatory Bowel Disease at Diagnosis in the Netherlands: A Population-based Inception Cohort Study (the Delta Cohort). Nuij *et al.* Inflammatory Bowel Diseases, Volume 19, Issue 10, pages 2215–2222 (2013). http://journals.lww.com/ibdjournal/Abstract/2013/09000/Phenotype_of_Inflammatory_Bowel_Disease_at.22.aspx

[2] Management of Inflammatory Bowel Disease Using Stem Cell Therapy. Irhimeh *et al.* Current Stem Cell Research & Therapy, Volume 11, Issue 1, pages 72-77 (2016). http://www.eurekaselect.com/133547/article

[3] Gut flora in health and disease. Guarner *et al.* The Lancet, Volume 361, Issue 9356, pages 512-519 (2003). http://www.sciencedirect.com/science/article/pii/S0140673603124890

[4] The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead. Kostic *et al.* Gastroenterology, Volume 146, Issue 6, Pages 1489-1499 (2014). http://www.sciencedirect.com/science/article/pii/S0016508514002200

[5] Diet, the Gut Microbiome, and Epigenetics. Hullar *et al.* The Cancer Journal, Volume 20, Issue 3, pages 170–175 (2014). http://journals.lww.com/journalppo/Abstract/2014/05000/Diet,_the_Gut_Microbiome,_and_Epigenetics.2.aspx

[6] Dietary metabolites and the gut microbiota: an alternative approach to control inflammatory and autoimmune diseases. Richards *et al.* Clinical & Translation Immunology, Volume 5, Issue 5, e82 (2016). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4910123/

[7] Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. Imhann *et al.* BMJ Gut, Published online (2016). http://gut.bmj.com/content/early/2016/11/03/gutjnl-2016-312135

[8] Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Zhernakova *et al.* Science, Vol. 352, Issue 6285, pp. 565-569 (2016). http://science.sciencemag.org/content/352/6285/565.full

[9] Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Frank *et al.* PNAS, Volume 104, Issue 34, Pages 13780-13785 (2007). http://www.sciencedirect.com/science/article/pii/S1074761314001939

[10] Relevance of Commensal Microbiota in the Treatment and Prevention of Inflammatory Bowel Disease. Dasgupta *et al.* Inflammatory Bowel Diseases, Volume 19, Issue 11, pages 2478–2489 (2013). http://journals.lww.com/ibdjournal/fulltext/2013/10000/Relevance_of_Commensal_Microbiota_in_the_Treatment.25.aspx

[11] Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Morgan *et al.* Genome Biology, 13, page R79 (2012). https://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-r79

[12] The effect of host genetics on the gut microbiome. Bonder *et al.* Nature Genetics 48, pages 1407–1412 (2016). http://www.nature.com/ng/journal/v48/n11/full/ng.3663.html

[13] Challenges of metabolomics in human gut microbiota research. Smirnov *et al.* International Journal of Medical Microbiology, Volume 306, Issue 5, Pages 266-279 (2016). http://www.sciencedirect.com/science/article/pii/S1438422116300212

[14] antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Medema *et al.* Nucleic Acids Research, 39, W339-W346 (2011). https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr466

[15] antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Blin *et al.* Nucleic Acids Research, Published online ahead of print (2017). https://academic.oup.com/nar/article/3778252/antiSMASH-4-0-improvements-in-chemistry-prediction?searchresult=1

[16] A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics. Donia *et al.* Cell, Volume 158, Issue 6, p1402–1414 (2014). http://www.cell.com/cell/fulltext/S0092-8674(14)01102-7
[17] Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. Cimermancic *et al.* Cell, Volume 158, Issue 2, p412–421 (2014). http://www.cell.com/cell/fulltext/S0092-8674(14)00826-5
[18] Small molecules from the human microbiota. Donia *et al.* Science, Vol. 349, Issue 6246, 1254766 (2015). http://science.sciencemag.org/content/349/6246/1254766.full
[19] Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Jostins *et al.* Nature, 491, 119–124 (2012). http://www.nature.com/nature/journal/v491/n7422/full/nature11582.html
[20] Introducing BBSplit. Bushnell. http://seqanswers.com/forums/showthread.php?t=41288
[21] Long Read RNA-seq Mapper. Marić. http://bib.irb.hr/datoteka/773708.Josip_Maric_diplomski.pdf
[22] BiG-SCAPE: exploring biosynthetic diversity through gene cluster similarity networks. Yeong *et al.* http://edepot.wur.nl/381865
[23] Human Microbiome Project Database. https://www.hmpdacc.org/hmp/HMRGD/
[24] FASTX-toolkit. http://hannonlab.cshl.edu/fastx_toolkit/
[25] Quantitative microbiome profiling links gut community variation to microbial load. Vandeputte *et al*. Nature, 551, 507-511 (2017). https://www.nature.com/articles/nature24460
[26] Scikit-learn Random Forest Classifier. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[27] Pfam database. http://pfam.xfam.org/
[28] InterPro database. https://www.ebi.ac.uk/interpro/
[29] SPAdes. http://cab.spbu.ru/software/spades/
[30] metaSPAdes: a new versatile metagenomic assembler. Nurk *et al*. Genome Research, 27, 824-834 (2017). http://genome.cshlp.org/content/27/5/824.long
[31] Fast and sensitive protein alignment using DIAMOND. Buchfink *et al*. Nature Methods, 12, 59-60 (2014). https://www.nature.com/articles/nmeth.3176
[32] MIBiG database. https://mibig.secondarymetabolites.org/