Title: Cross-study analyses of gut microbiomes from healthy and obese individuals

Authors:

Maxime Déraspe*: maxime.deraspe.1@crchudequebec.ulaval.ca
Infectiology Research Centre RC-709
2705 Blvd. Laurier
Quebec City, G1V 4G2, QC, Canada

Charles Burdet: charles.burdet@inserm.fr
Département d'Epidémiologie Biostatistique et Recherche Clinique // Hôpital Bichat
AP-HP.Nord - Université de Paris
Inserm CIC-EC 1425 Epidémiologie Clinique
46 rue Henri Huchard, 75018 Paris, France

Juan Manuel Dominguez: juan-manuel.dominguez.1@ulaval.ca
Infectiology Research Centre RC-709
2705 Blvd. Laurier
Quebec City, G1V 4G2, QC, Canada

François Laviolette: Francois.Laviolette@ift.ulaval.ca
Computer Science and Software Engineering, Université Laval
1065, av. de la Médecine
Québec, Canada, G1V 0A6

Paul H Roy: Paul.Roy@crchudequebec.ulaval.ca
Infectiology Research Centre RC-709
2705 Blvd. Laurier
Quebec City, G1V 4G2, QC, Canada

Jacques Corbeil: jacques.corbeil@fmed.ulaval.ca
Infectiology Research Centre RC-709
Office 747, 2705 Blvd. Laurier
Quebec City, G1V 4G2, QC, Canada

* Corresponding author
Cross-study analyses of gut microbiomes from healthy and obese individuals

Maxime Déraspe1,2*, Charles Burdet1, Juan Manuel Dominguez1,2, François Laviolette2,6, Paul H Roy3,7 and Jacques Corbeil1,2,5,6

Abstract

Background With the advent of metagenomics, many large studies have been conducted with the quest of better understanding gut microbiota changes in relation to varying health conditions. Significant findings have been made for diseases such as cirrhosis, colorectal cancers, inflammatory bowel diseases and others, yet one that stands out is obesity for which conflicting results have been reported in the literature.

Methods Here, we built and analyzed a cross-study dataset of healthy and obese individuals looking for major changes in the taxonomic and functional composition of their metagenomes.

Results Our results suggest that the overweight and normal subjects have no strong dissimilarity in their metagenomes composition. Significant differences were observed when comparing the obese and the non-obese individuals in their functional and taxonomic profiles.

Conclusion In this study, we report the most significant changes that we observed and discuss their potential implication in the obesity condition.

Keywords: metagenomics; gut microbiota; obesity

Background

Obesity is a major epidemic with an economic cost to society exceeds 150 billion dollars in health resource allocation in the United State alone [1]. Since the 1980s in more than 70 countries, the prevalence of obesity has more than doubled, while increasing in most other countries [2]. At least three factors have a major impact on the body mass index (BMI) increases: first, the presence of trans fatty acids in food consumption; second, the high fructose levels in soda and fruit juices; and third, physical inactivity [1]. More importantly, increase in BMI is linked to several chronic diseases including cardiovascular diseases [3], diabetes mellitus (especially type 2) [4], chronic kidney disease [5], numerous cancers [6], and musculoskeletal disorders [7].

The gut microbiome is also, by all accounts, impacted by our lifestyle (diet, exercise, drug use) and is even considered, rightly so, as a human microbial organ [8]. Numerous studies relate dysbiosis of the gut microbiome with various pathological conditions such as obesity [9], type 2 diabetes [10, 11], inflammatory bowel disease [12, 13], colorectal cancer [14, 15, 16] or even mental disorders [17]. Pasolli and collaborators examined several diseases associated with microbiome datasets which they reanalyzed with machine learning aiming to provide an evaluation of the classification of ill and healthy individuals based on the gut metagenomics data [18]. In
all the assessed conditions (cirrhosis, colorectal, IBD, T2D and obesity [9]), obesity remains the worst predictive results with an AUC not surpassing 0.66. This same group also succeeded in improving their results by performing cross-study sample inclusion with additional control samples. However, this specific approach was not performed for the obesity cohort.

Several controlled studies have examined the relationship between obesity and the gut microbiota composition [19, 20, 21, 22, 23]. A general observation was the the ratio of the two taxonomic phyla Bacteroidetes and Firmicutes between obese and non-obese human or animal (mouse) [19, 20, 21, 24, 22, 23, 9]. The observation was not consistently reproduced as no significant relationship was found for the Bacteroidetes phylum in Duncan et al., and in Schwierz et al. whereas the Firmicutes to Bacteroidetes ratio was increased in favour of Bacteroidetes in obese subjects contrary to what was found in two studies by Ley et al. and the 2006 study of Turnbaugh et al. where the ratio was inverted.

In our study, we addressed and investigated the creation of a cross-study human gut microbiome dataset and the evaluation of their composition when analyzing individuals’ BMI. We used the MGnify protein database as the core resource for all our analyses [25]. We also sought to classify of obese and non-obese subjects based on each protein feature annotation of the gut metagenomes by evaluating different machine learning algorithms on these features. Note that for convenience, the words healthy and normal are used interchangeably in this manuscript and it refers to a BMI between 18.5 and 30 (healthy and overweight based on the CDC categories), as well as for the word obese that represent a BMI ≥ 30, without any distinction of the obesity subtypes if not mentioned otherwise.

**Methods**

**Cross-study samples selection**

The dataset was created by collating all the metadata available from the curatedMetagenomicData R package [26] and by selecting samples that respected the following selection criteria: subjects aged between 18 and 65 years old were considered healthy according to each study’s procedure (except for overweight), and there was no indication of drugs used to treat any underlying conditions. Their BMI was between 18.5 and 30, subjects were adults Similarly, the obese samples were chosen with these criteria except that their BMI exceeded 30. Subjects were chosen from different countries (Netherlands 378, Denmark 109, Madagascar 83, Spain 40, USA 19, Peru 11) to ensure a reasonable geographic distribution that would extend our analysis to different ethnic populations and therefore include different diet and genetics. Altogether, the gut metagenomes originate from four different studies with one unpublished (PasolliE.2018) [12, 27, 28]. The total number of samples summed up to 640 of which 221 were male subjects compared to 270 females and 149 of unknown gender. As for the age of the subjects, 58.9% were between 18 and 30, while the remainder were between 31 and 65.
Definition of overweight and obesity

The official CDC\cite{1} body fatness categories associated to BMI are < 18.5 for underweight, 18.5 to < 25 for normal, 25 to < 30 for overweight and 30 to < 35, 35 to < 40 and ≥ 40 for class 1, 2 and 3 obesity, respectively. In our reported results we have combined the obesity 2 and 3 levels due to the small sample size as only 3.3% of all samples have a BMI ≥ 35.

Metagenomics data preparation

The whole genome sequencing (WGS) raw sequence reads (FASTQ files) of the publicly available gut metagenomes from included subjects were downloaded from the NCBI\cite{2} sequence read archive (SRA). Quality filtering on the reads was performed using fastp\cite{29}, with a quality Phred score of 30. In order to normalize the depth of all the samples, the filtered reads were subsampled at 10 millions reads. We chose the Unified Human Gastrointestinal Protein (UHGP) catalogue clustered at 90% identity (UGGP-90) from MGnify databases \cite{25} for the metagenome annotation using kAAmer\cite{3}\cite{30} for protein identification. MGnify includes annotations from eggNOG \cite{31}, the enzyme commission database \cite{32}, and KEGG pathways \cite{33}, among others. Extraction of the significant hits was performed with the analysis scripts provided with kAAmer. We also assembled the subsampled reads into contigs using megahit \cite{34} and did a followup analysis with Ray Surveyor \cite{35} to compare the metagenomes DNA k-mer content.

Statistical analyses

We compared the subjects considered as normal for their body fatness condition (BMI < 30) with those classified as obeses (BMI ≥ 30). Statistical analyses were performed in Python using the scikit-bio, statsmodels and scipy libraries \cite{36, 37}. We compared the distribution of each protein feature (Taxa, Gene Ontology, Enzyme Commission, COG, KEGG pathways and modules) using the non parametric Wilcoxon - Mann & Whitney test, and p-values were corrected using the Benjamini-Hochberg procedure as implemented in the statsmodels Python package. Alpha diversity was computed using the Shannon index with the scikit-bio\cite{4} Python package and compared between groups using the non parametric Wilcoxon - Mann & Whitney test. Beta diversity computation was also achieved using scikit-bio with the Bray-Curtis dissimilarity and compared between cohorts with the analysis of similarities (ANOSIM) test.

Machine learning in the comparison of the microbiomes

We used the utility functions from the kAAmer analyses scripts to compute machine learning classifiers based on the protein feature annotations provided in the MGnify database (taxa, Gene Ontology, Enzyme Commission, KEGG pathways and modules, and Clusters of Orthologous Groups (COG)) \cite{25}. The machine learning

\footnote{Centers for Disease Control and Prevention: https://www.cdc.gov/obesity/adult/defining.html}
\footnote{National Center for Biotechnology Information: https://github.com/zorino/kaamer}
\footnote{http://scikit-bio.org/}
algorithms tested included gradient boosting methods (xgboost [38], lightgbm [39],
catboost [40]), ensemble method (random decision forests [41]) and the support
vector machine (SVM [42]). The number of subjects in both groups was unbalanced
with a ratio of approximately 6.8 in favour of the normal cohort - 558 normal indi-
viduals against 82 obese individuals - and the training was performed accordingly.
All hyperparameter searches were performed using the Optuna framework with a 10
fold cross-validation [43]. Similarly, the score metrics were evaluated with a 10 fold
stratified cross-validation. The cross-validation and score evaluations were achieved
using the scikit-learn package [44]. Feature importance from the models was also
extracted using the ELI5’s Python package[5].

Results
K-mer analyses of the metagenomes
Metagenomes comparison based on their complete genomics content is a complex
task due to the vast diversity of microbes found in individuals and the subsequent
very high number of genetic sequences obtained by NGS methods. Ray Surveyor
allows the direct comparison of metagenomes based on their $k$-mers content by
reconstructing a phenetic tree that provides clusters that can be associated with
known phenotypes [35]. Figure 1 illustrates the heatmap and hierarchical clustering
performed with Ray Surveyor on our dataset. The hierarchical clustering, shown on
the left and top axes of the heatmap, defines seven clusters with their colour ranging
from green to blue. On the same axes the associated BMI (top axis) and country (left
axis) of the individuals are identified by triangles. With regard to the BMI, the most
homogeneous clusters were found to be associated with the normal condition in the
colours: green (100% normal), red (96.94% normal), magenta (96.77% normal) and
yellow (90.54% normal). The two clusters with the most obese individuals are the
black and blue clusters with 38.6% and 35.23% respectively of obese individuals. The
cyano cluster was represented at 82.22% within the normal cohort. For the country of
the samples, the first three clusters (green, red and cyan) are ubiquitously composed
of people with Western (Europe and USA) origin with 353 individuals from the
Netherlands, 15 from Spain, 17 from Denmark and 12 from the United States (USA).
The magenta and yellow clusters are mostly composed of individuals with African
origin with 75 from Madagascar (71.4%) and Peruvian origin (9.5%) with the rest of
the samples originating from Denmark (17), the Netherlands (2) and the USA (1).
Finally, the black and blue clusters are mixed, although predominantly European,
with 77 individuals from Denmark, 24 from Spain, 28 from the Netherlands, 9 from
Madagascar, 6 from the USA and 1 from Peru.

Selecting the hierarchical clustering to the end of the tree ramification would
have yielded additional clusters and certainly improved the overall homogeneity of
the clustering. However, there is an interesting demarcation in the broadest clus-
ters between the lower-left group 1 (green, red and cyan colours) and the upper
right sections identified as group 2 (magenta, yellow, back and blue colours). In-
deed, the clustering suggests the existence of a relationship between the BMI and
the clustering of $k$-mers metagenomics content, even though the complexity that

[5] https://github.com/TeamHG-Memex/eli5
underpins obese conditions maybe too complex to be explained by the gut microbiota alone. Nonetheless, the metagenomics features that distinguish group 1 and group 2 overlap substantially with the one that discriminates normal from obese individuals as identified by the statistical and machine learning analyses of protein annotations. The next sections will present these characteristics, based on taxonomy, protein functions and pathway abundances, that contribute to the delineation of the metagenomes in the context of body fatness.

**Taxonomic analyses of the gut metagenomes**

Taxonomic analyses are one of the foundations of metagenome investigations using primarily 16S rRNA and WGS experiments. Investigating different taxonomic ranks often yields different results and is greatly influenced by the quality of the reference database, a common occurrence in most genomics studies [45]. We used the MGnify protein database because of its focus on the gut microbiome. It has been recently updated with the numerous protein annotations including Gene Ontology (GO), KEGG pathways, and Clusters of Orthologous Groups (COG) among others [25].

A multitude of microbes found in the gut have not been cultured and characterized [46], thus curation projects such as MGnify are needed to account for those uncharacterized bacterial species available through public WGS metagenomics studies. Understandably not every species is well characterized and therefore only a few proteins have a taxonomic identification down to the species or subspecies. In fact in MGnify, only 10% of the proteins are associated with a taxon down to the genus, and respectively 52%, 60%, 89% and 94% at the family, order, class and phylum taxonomic rank. The taxonomic classification can also be impacted by the protein sequences clustering at 90% identity that can regroup protein homologs found in different species. Nevertheless, the results of the annotations provide a representative portrait of the gut microbiome proteome which holds the essential information of its functional capabilities. Figure 2 shows the Bacteria domain cladogram down to the taxonomic family rank. It also indicates the most dominant phyla and families found in the gut metagenomes of 640 healthy and obese individuals. The table 1 along with figure 2 highlights the different bacterial families with their overall abundances and expression status in obese.

We noted an absence of significant differences when comparing the normal group as defined by the CDC (BMI between 18.5 and 25) with the overweight group (BMI between 25 and 30). Furthermore, including the overweight in the normal cohort yielded very similar p-values when comparing only the normal group with the obese 1, 2 and 3 groups (BMI ≤ 30) without including the overweight group. This observation remained valid for the other taxonomic ranks and protein annotations analyzed, thus justifying the BMI threshold at 30 to separate the groups in our experiments.

The first taxonomic analysis was at the phylum rank, as many but not all gut metagenomics studies have reported, the significance of their abundance and ratio in explaining obesity [19, 20, 21, 22, 23]. In figure 3, we reported the four most abundant phyla, ordered by the normal cohort individual mean, based upon their differential abundance between obese and normal individuals. Each boxplot represents a BMI group as defined by the CDC, except for obesity 3 and 4 that were
combined due to the limited number of samples in those two groups. Our results for the *Firmicutes* and *Bacteroidetes* ratio are in agreement with Schwertz *et al.* and contrast to observations reported in Ley *et al.* and Turnbaugh and collaborators. Indeed, our analyses suggest an unbalanced ratio in favour of the *Bacteroidetes* in obese individuals. Specifically, the mean relative abundance of *Bacteroidetes* was higher by 16% (41% to 25%, *P* < .001) in obese and in contrast *Firmicutes* were lower by 11% (52% to 63%, *P* < .001). On average, those two phyla represented 88.88% of the total microbes sequenced in the gut microbiomes. The third most abundant phylum, *Actinobacteria*, was also significantly (*P* < .001) reduced in obese individuals by 3.8% (from 5.5% to 1.7% on average). *Fusobacteria* were significantly more abundant in obese individuals, despite a low global abundance, going from 0.150% to 0.107% in healthy individuals (*P* < .001). Those four phyla abundance observations were also replicated when investigating the difference between group 1 (mostly normal) and group 2 (predisposition to obesity) from the hierarchical clustering. There were also additional abundant phyla with no significant differences for the BMI but that were capable of differentiating group 1 and group 2 clusters. Indeed, the phyla *Proteobacteria* (5.3x, *P* < .001), *Euryarchaeota* (1.5x, *P* < .001) and *Spirochaetes* (4.7x, *P* < .001) were all relatively enriched in group 2 compared to group 1, indicating potential dysbiosis at a greater propensity in group 2 as generally lower abundant taxa bloom significantly. To appreciate the overall abundances, the cladogram on figure 2 is coloured based on the phyla and the important families are highlighted when differently and significantly represented in the metagenomes of normal and obese. The feature extraction from the classification analyses yields an appreciation of which feature (herein the phyla) possesses the most weight in classifying the two cohorts. For the phylum classification, the most important feature is the *Actinobacteria* that outweighs the other phyla by at least twofold in the best performing algorithms. Another important feature that emerged from the classification was the *Deferribacteres* with an overexpression in obese individuals (*P* < .001), although it had a very low mean abundance of 0.0078% in all samples.

The investigation of the microbiome phylogenetic composition at the family taxonomic rank allowed us to further evaluate the important taxa related to body fatness. As in the phylum analysis, we selected the most abundant families that were significantly different in terms of abundance in obese and normal individuals. Figure 4 displays the top 9 families ordered by their abundance as found in the healthy group. The first two families *Lachnospiraceae* and *Ruminococcaceae* are from the *Firmicutes* phylum and constitute on average 25.4% and 22.6% of a normal microbiome composition, based on our criteria. Both were underrepresented in obese individuals by 6.7% and 8% respectively on average (*P* < .001). The next two most prevalent *Firmicutes* families were the *Clostridiaceae* and the *Eubacteriaceae* (both at ≈ 7%) and did not show significant differences between the obese and normal conditions yet had a significant difference between the group 1 and group 2 clusters of figure 1. The most abundant *Bacteroidetes* family was *Bacteroidaceae* with an overall abundance of 21.8% (second overall after the *Lachnospiraceae*) and a greater expression in obese gut metagenomes by more than 15% on average (19.9% → 35.2%, *P* < .001). The next most abundant families from the *Bacteroidetes*
phylum was Porphyromonadaceae with an overall abundance of 4.5% and overexpressed in obese (4.1% → 6.9%, \( P < .001 \)), Rikenellaceae at 1.7% with no significant difference, the Prevotellaceae at 0.31% and the Flavobacteriaceae at 0.25%. Interestingly, the Prevotellaceae is the only reported Bacteroidetes in the top 20 families that is underexpressed in obese metagenomes (0.32% → 0.27%, \( P < .001 \)). From the classification analyses, the Ruminococcaceae and Micrococcaceae are two families overrepresented in the predictions for obesity, and that result was obtained with multiple algorithms.

We also analyzed the most abundant genera, even though only 10% of the proteins in the MGnify database are annotated down to the genus rank, thus one needs to be cautious in overinterpreting the results. The six most abundant and differentially expressed genera, still ordered by their abundance in healthy individuals, are shown in supplementary figure 1. The three most abundant are Blautia, Dorea and Lachnoclostridium, which all come from the Lachnospiraceae family and constitute respectively on average 39.6%, 17.5% and 15.1% of the overall abundance of the gut metagenomes in our experimental group. The difference in expression of the Blautia and Dorea genera is in line with their family (Lachnospiraceae) tendency as they were underexpressed in obese individuals, respectively by 13.6% (41.4% → 27.8%, \( P < .001 \)) and by 3.6% (17.9% → 14.8%, \( P < .001 \)). In contrast, the Lachnoclostridium and Butyrivibrio genera from the same family were overexpressed in obese, respectively by a mean difference of 4.1% and 6.8% (\( P < .001 \)).

The next genus with the most identified proteins and being significantly different is Bacillus from the Bacillaceae family and similarly is overexpressed in obese by a mean margin of 3.4% (\( P < .001 \)). Finally, the last of the six genera of figure supp. 1 is Alloprevotella, an overexpressed Prevotellaceae even though the family was on average underexpressed in obese individuals (\( P < .001 \)). The top five genera identified in the machine learning analyses are conserved in the two best performing algorithms (LightGBM, Random Forest), and is constituted of Chryseobacterium, Blautia, Butyrivibrio, Flavobacterium and Oceanicola genera.

**Functional analyses**

The functional analyses is comprised of the abundances of protein annotation features, such as their Enzyme Commission (EC) numbers, their Gene Ontology (GO) characterization, their Clusters of Orthologous Group (COG) association and their associated pathway and module from the KEGG database.

The results for the KEGG pathways report a multitude of significant pathways (with \( P < .001 \)) that are differentially expressed in the obese and normal cohorts. Figure 5 shows the top twenty based on the lowest p-values as reported by the statistical tests. Interestingly, the most significant pathway is the butyrate metabolism which is decreased in obese individuals (\( P < .001 \)) and been shown previously to impact energy homeostasis in mice when comparing germ-free and normal mice [47]. Indeed, butanoate or butyrate is a short-chain fatty acid (SCFA) that has been considered as one of the main products of fermentation by bacteria in the colon [48]. Butyrate can regulate gene expression and could serve as a preferential energy source for the colonic epithelial cells (colonocytes) [49, 50, 51, 52, 53, 54]. SCFA are generally considered the result of resistant starch and dietary fiber fermentation from bacteria in the gut and are used in glucose and lipid biosynthesis [55, 56, 57].
The second most important pathway from the KEGG database analysis is the microbial metabolism in diverse environmental pathways (ko01120) which are also found to be decreased in the obese cohort. These pathways are related to several others that are involved in the metabolism or degradation of different metabolites, and which are also significantly underexpressed in obesity. To name a few, the glycolysis and gluconeogenesis pathway (ko00010), the glycine, serine and threonine metabolism pathway (ko00260), the carbon fixation pathway in prokaryotes (ko00720), and the nitrogen metabolism pathway, are all related to microbial metabolism and indeed underexpressed in the microbiome associated with obesity. This is a potential indication of metabolism efficiency reduction in obese individuals taking root in the gut microbes.

The importance of the glycolysis and gluconeogenesis pathway in obesity could be linked, one to food intake and satiety, and second as a key factor in insulin sensitivity [58]. Enzymes such as glucokinase (EC: 2.7.1.2) and pyruvate kinase (EC: 2.7.1.40), and their associated GO activities (GO:0004340, GO:0004743) were also found to be decreased in obese metagenomes. The glycine, serine and threonine metabolism pathway has also been studied with relationship to obesity. First, it has been suggested that the gut microbiota plays an important role in glycine availability for the host which would be impacted by diet and essential to multiple metabolic pathways of the host [59]. Second, the three amino acids were also significantly more available for the host in germ-free mice compared to wild type, suggesting the implication of the gut microbiota in their metabolism and availability [60]. Finally, threonine as well as cysteine has been shown to improve protection against colitis in rats by promoting mucin secretion and which is very probably linked to beneficial bacteria in the gut microbiota [61]. The carbon metabolism (ko01200) and fixation pathways (ko00720) are both underexpressed in obese individuals and the annotated proteins mainly come from the Ruminococcaceae family, a Firmicutes also less abundant in the obese. It has previously been demonstrated that not all bacterial species can utilise the same carbon source to produce fermented products beneficial to the host, such as SCFA [62]. Another noteworthy pathway that is less abundant in obese individuals is the one involved in the biosynthesis of amino acids (ko01230), which is also related to the metabolism of SCFA by providing the required amino acids for their synthesis [63, 64, 65].

On the other hand, the most important pathway for obesity delineation from the classification analyses was the selenocompound metabolism pathway (ko00450) as it appears as the top discriminant feature with different algorithms. It has been suggested by Hrdina and collaborators, when experimenting on mice, that bacteria compete with the host for selenium when availability becomes limiting [66]. A decreased expression in obese individuals could indicate a selenium-deficient diet as fewer bacteria involved in its metabolism were able to thrive. The gene ontology results also show evidence of selenocompound metabolism in relationship with obesity. Indeed, several activities involving selenocompounds were significantly less abundant in obese individuals such as the L-seryl-tRNASec selenium transferase activity (GO:0004125), the selenocysteinyl-tRNA(Sec) biosynthetic process (GO:0097056), the transferring selenium-containing groups transferase activity (GO:0016785), the
selenate reductase activity (GO:0033797), and the selenocysteine incorporation, insertion sequence binding, metabolic process and biosynthetic process (GO:0001514, GO:0035368, GO:0016259, GO:0016260) \((P < .001)\).

The vast majority of the top significative pathways were underexpressed in the obese cohort with the notable exceptions of the peroxisome pathway (ko04146), the NOD-like receptor signalling pathway (ko04621), the ferroptosis pathway (ko04216) and the biofilm formation in *Vibrio cholerae* pathway (ko05111). There is evidence that suggests that these four pathways could be related to immune response and pro-inflammatory reactions [67, 68, 69, 70]. Interestingly, the peroxisome pathway which is also involved in lipid metabolism has been suggested as an important factor to maintain gut epithelium homeostasis and renewal in *Drosophila* [71]. Indeed, dysfunctional peroxisomes in the host gut epithelial cells would trigger Tor kinase-dependent autophagy that would increase cell death and promote instability at the host-microbe interface in the gut. Similarly, the ferroptosis pathway would also be involved in intestinal epithelial cell death and could even lead to ulcerative colitis [72]. Even though those two pathways are usually considered as part of host cells, we found several bacterial proteins in the MGnify database that were classified as part of these pathways. The main enzyme that was predominantly part of both pathways was the long-chain-fatty-acid-CoA ligase (EC: 6.2.1.3), which also showed an increased abundance in obese individuals and could explain the pathway results. The NOD-like receptor signalling pathway genes characterized in the MGnify database are for the most part related to the thioredoxin system (such as *trxA*) and are predominantly found in the *Bacteroidaceae* and *Porphyromonadaceae* families, both overexpressed in obese gut metagenomes. Different studies have investigated the role of tetrathionate in the gut microbiota, including one suggesting that an upregulation of the thioredoxin reductase *trxA* may modulate the gut microbiome during inflammation by regulating the levels of tetrathionate [73], and another one from Winter et al. showing a competing growth advantage in inflamed gut for *Enterobacteriaceae*, such as *Salmonella* in presence of tetrathionate [74].

The COG results are also in agreement with the other annotation results, as the main significant categories are related to energy production and conversion and the amino acid and nucleotide transport and metabolism, and are all underrepresented in obese metagenomes \((P < .001)\). However, the COG annotations have a very broad categorization and do not accommodate for more specific functional analyses.

Finally, KEGG modules are another annotation resource that provides a functional annotation unit for the proteins, although they are related to KEGG pathways. The most significant module reported in the statistical test is associated with purine degradation (M00546) and is less abundant in obese individuals. The module is part of the purine metabolism KEGG pathway, which has not proved to be different in the pathway analyses. Interestingly, the degradation of purine is associated with gout, a condition that is characterized by the accumulation and reduced excretion of uric acid [75]. Indeed, Guo and collaborators found a disorder in purine degradation and butyric acid biosynthesis in the gut metagenomes. In agreement with our study, the butyric acid biosynthesis was less abundant in gout. However, the purine degradation pathway was enriched in gout and depleted in obese individuals. Nevertheless, body fatness (visceral fat) has been reported to strongly
correlate with gout [76]. In the context of obesity, the lack of bacterial species that metabolize purine into uric acid may contribute to the condition. Several other modules involved in the transport of metabolites such as cystine (M00234), glutamine (M00228), rhamnose (M00220), molybdate (M00189), glutamate (M00233) and others were also significantly depleted in obese individuals. The best performing classification algorithm also reported the chorismate reaction of the shikimate pathway as the most significant module that was enriched in normal individuals. The shikimate pathway is found exclusively in microorganisms and plants and is mainly dedicated to the production of aromatic amino acids (phenylalanine, tyrosine, and tryptophan) in bacteria [77]. Evidence has also shown that perturbation in the shikimate pathway is diet-related [78].

Obese metagenome classification

We also sought to classify the two groups: the obese, with a BMI of $\geq 30$, versus normal and overweight combined with a BMI $< 30$. The same protein annotation features that were used in the statistical analyses were used to train and evaluate different machine learning algorithms: xgboost, lightgbm, random forest, svm, decision tree, adaboost and the SCM [38, 39, 41, 79, 80, 42]. We hoped that the MGnify database, a very recently updated protein annotation database, which includes information from several other important databases, like KEGG, EC, and GO [33, 32, 81], would enable us to improve the annotation coverage of the metagenomic proteomes.

Similar work had been conducted in Pasolli and collaborators, for the meta-analysis of several diseases associated with the gut microbiota. Their results for the obesity condition were solely based on the taxonomic profiling obtained from Metaphlan2 [82] and the discrimination was made between lean (BMI $\leq 25$) and obese (BMI $> 30$) individuals without including the overweight condition. Unfortunately, obesity [9] was the least predictable condition as several others such as liver cirrhosis [83], colorectal cancer [14], inflammatory bowel diseases (IBD) [84], and type 2 diabetes [10, 11] reported superior classification scores. Nevertheless, based on the premise that cross-study sample inclusion was an improving factor, we built the obese and non-obese cohorts and evaluated classification accuracy. We did improve the overall results but not by a large margin for the cirrhosis, colorectal cancer and IBD conditions. However we obtained significant improvement for type 2 diabetes results and the obesity classification. The evaluation results for the machine learning algorithm are reported in table 2 (additional details are provided in supplementary tables). The tables are separated by protein annotation features, either functional or taxonomic, as statistically analyzed in the previous sections. The best performing features, based on the F1 score results, are the COG annotation (0.66) closely followed by the taxonomic genus (0.65), the gene ontology (0.64), the enzyme commission (0.63) and the KEGG module (0.62) annotations. From all the evaluated algorithms, lightgbm (gradient boosting) and random forest were the ones that yielded the best results overall. All individual features achieved better scores than the one reported in Pasolli et collaborators with the best F1 score being at 0.55 for the support vector machine. Our improved results could be attributed to the inclusion of samples from several studies, especially for the normal cohort, and
the usage of an updated metagenomics database for the protein annotations. The optimized results allowed the extraction of important features in the discrimination between obese and normal gut metagenomes as reported in the functional analyses section.

**Discussion**

In this study, we investigated the gut metagenome of 640 normal, overweight and obese individuals based on their BMI. We evaluated their metaproteomes seeking the difference in the functional and taxonomic annotation that would be additional indicators of the obesity condition. Overall, we found that the overweight group (BMI between 25 and 30) was not significantly different from the normal group (BMI between 18.5 and 25). Indeed, differences were readily noticed when comparing the obese group with the non-obese group, that is with a BMI cutoff at 30. It is important to take into account that BMI measurements alone have some limitations as a reflection of the percentage of body fat since its association varies between different ethnic groups, such as the Asian population when compared to Caucasian. [85]. Nonetheless, BMI is still used by the World Health Organisation (WHO) for the evaluation of obesity and excess weight. We used it for this study due to its wide availability and acceptability. The $k$-mer analysis showed two broad groups with a greater concentration of obese individuals in the second group. However, the clusters were also greatly representative of the ethnicity of the individuals with a distinction between Western and non-Western origins. It suggests that factors related to ethnicities such as diet, lifestyle and geography are contributing to shaping the gut microbiome, as previously suggested [86, 87]. No clusters were predominantly represented by obese metagenomes, with potential explanations being the unbalanced nature of our datasets, extrinsic factors such as diet, genetics and lifestyle and importantly that the BMI is not perfectly representative of body fatness, hence our effort to find better biomarkers [88].

The overall relative abundances of the major phyla, such as the *Firmicutes*, *Bacteroidetes* and *Actinobacteria*, were in agreement with previous studies. With regard to their relative abundances in normal versus obese individuals, contrasting results have been reported in the literature [19, 20, 21, 22, 23, 9]. Our results add support to an enriched proportion of *Bacteroidetes* in obese individuals when comparing the *Bacteroidetes* to *Firmicutes* ratio. Moreover, it is the *Actinobacteria* that best discriminated the obese and the normal groups, with a depleted abundance in the obesity group. *Actinobacteria* has been reported to be represented mainly by the *Bifidobacterium* genus in the gut metagenome [89] and their decrease is associated with several conditions in addition to obesity, such as types I and II diabetes, cystic fibrosis, hepatitis B and *Clostridium difficile* infections [90]. At the family taxonomic rank, the *Lachnospiraceae* (24.6%) and *Ruminococcaceae* (21.6%) were the major representatives of the *Firmicutes* phylum whereas the *Bacteroidaceae* family (21.8%) was the most abundant *Bacteroidetes*. Together, those three families represented an overall abundance of 68% of the metagenomes on average and their relative abundances were in line with their respective phyla when comparing the obese and non-obese groups.
The functional proteome analyses revealed the importance of short-chain fatty acid (SCFA) metabolism in the healthy condition in comparison with obesity. Indeed, butyrate metabolism was the most statistically significant pathway depleted in obese individuals. Butyrate along with propionate and acetate are SCFA produced by the fermentation of dietary fibers and starch by gut bacteria and play an important role in energy availability for the epithelial cells in the colon [91]. Other pathways, such as amino acid biosynthesis, were also decreased in obese metagenomes and are known to be related to SCFA production by providing the necessary building blocks of their synthesis. Another discriminant example is the bacterial shikimate pathway, which essentially produces aromatic amino acids, and from which one reaction (from KEGG modules) proved to be the most important feature in the machine learning analyses to separate the obese from the normal cohort. Furthermore, we also reported the importance of the selenocompound metabolism pathway, which is also related to diet and characterized by a selenium-deficient food intake from individuals with an obesity condition. On the other hand, the pathways that were more abundant in obese individuals were mostly involved in immune response and pro-inflammatory reactions.

**Conclusion**

Overall, by analyzing metagenomes collected from different studies, we were able to identify significant changes in obese versus non-obese individuals. By using protein annotations, we drew the taxonomic portrait and functional capabilities of the metagenomes that assist in interpreting the impact of the obesity condition on the gut microbiota. Additional metagenomics experiments on obesity with diverse cohorts of individuals, controlled diets and other measurements of body fatness would enhance our comprehension of the complex interactions between human obesity and the gut microbiome.

**Declarations**

**Ethics approval and consent to participate**

Not applicable

**Consent for publication**

Not applicable

**Availability of data and material**

All the data and software used in this study are publically available. See the supplementary tables for details on the metagenomics dataset.

**Competing interests**

The authors declare that they have no competing interests.

**Funding**

This study was financed by the Canada Research Chair in medical genomics (JC). MD was supported by the Fonds de recherche du Québec - Santé (#32279).
Authors’ contributions
Metagenomics sample selection was performed by MD and CB. MD and JMD analyzed the metagenomes. MD, CB and JC interpreted the analyses results. MD wrote the manuscript which has been critically reviewed by all the authors.

Acknowledgements
Not applicable

Author details
1Department of Molecular Medicine at Université Laval, Québec, Canada. 2Big Data Research Center at Université Laval, Québec, Canada. 3Infectious Disease Research Centre at Université Laval, Québec, Canada. 4Department of Computer Science at Université Laval, Québec, Canada. 5Center NUTRISS at Université Laval, Québec, Canada. 6Bioinformatics at Université Laval, Québec, Canada. 7Biochemistry, Microbiology, and Infectious Disease Research Centre at Université Laval, Québec, Canada. 8Institute Intelligence and Data at Université Laval, Québec, Canada.

References
1. Hurt, R.T., Kulisek, C., Buchanan, L.A., McClave, S.A.: The obesity epidemic: challenges, health initiatives, and implications for gastroenterologists. Gastroenterology & hepatology 6(12), 780 (2010)
2. Collaborators, T.G.O.: Health Effects of Overweight and Obesity in 195 Countries over 25 Years. New England Journal of Medicine 377(1), 13–27 (2017). doi:10.1056/NEJMoa1614362
3. The Emerging Risk Factors Collaboration: Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. The Lancet 377(9771), 1085–1095 (2011). doi:10.1016/S0140-6736(11)60105-0
4. Algoblan, A., Alaffi, M., Khan, M.: Mechanism linking diabetes mellitus and obesity. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 587 (2014). doi:10.2147/DMSO.S67400
5. Rhee, C.M., Ahmadi, S.-F., Kalantar-Zadeh, K.: The dual roles of obesity in chronic kidney disease. Current Opinion in Nephrology and Hypertension 25(3), 208–216 (2016). doi:10.1097/MNH.0000000000000212
6. Pischon, T., Nimpfuchs, K.: Obesity and Cancer. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature Reviews Gastroenterology and Hepatology 14(10), 573–584 (2017). doi:10.1038/nrgastro.2017.88
7. Onyemaechi, N., Anyanwu, G., Obikili, E., Onwuasoigwe, O., Nwankwo, O.: Impact of overweight and obesity on the musculoskeletal system using lumbosacral angles. Patient Preference and Adherence, 291 (2016). doi:10.2147/PPA.S90967
8. Baquero, F., Nombela, C.: The microbiome as a human organ. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 18 Suppl 4, 2–4 (2012). doi:10.1111/j.1469-0691.2012.03916.x
9. Le Chatelier, E., Nielsen, T., Qin, J., Pritfi, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., Leonard, P., Li, J., Burgdorf, K., Grarup, N., Jørgensen, T., Brandlund, I., Nielsen, H.B., Juncrer, A.S., Bertalan, M., Levenez, F., Pons, N., Rasmussen, S., Sunagawa, S., Tap, J., Tims, S., Zoetendal, E.G., Brunak, S., Clement, K., Doré, J., Kleerebezem, M., Kristiansen, K., Renaudt, P., Sicheritz-Ponten, T., de Vos, W.M., Zuckk, J.-D., Hansen, T., consortium, M., Bork, P., Wang, J., Ehrlich, S.D., Pedersen, O.: Richness of human gut microbiome correlates with metabolic markers. Nature 500(7464), 541–6 (2013). doi:10.1038/nature12506
10. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S.D., Nielsen, R., Pedersen, O., Kristiansen, K., Wang, J.: A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490(7418), 55–60 (2012). doi:10.1038/nature11450
11. Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., Bäckhed, F.: Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature 498(7452), 99–103 (2013). doi:10.1038/nature12198
12. Nielsen, H.B., Almeida, M., Juncrer, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J.-M., Quintanilha Dos Santos, M.B., Blom, N., Boruel, N., Burgdorf, K.S., Boumezbeur, F., Casellas, F., Doré, J., Dworzynski, P., Guarnier, F., Hansen, T., Hildebrand, F., Kaas, R.S., Kennedy, S., Kristiansen, K., Kultima, J.R., Leonard, P., Levenez, F., Lund, O., Moumen, B., Le Paslier, D., Pons, N., Pedersen, O., Pritfi, E., Qin, J., Raes, J., Serensen, S., Tap, J., Tims, S., Usery, D.W., Yamada, T., MetaHIT Consortium, Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., Ehrlich, S.D., MetaHIT Consortium: Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature biotechnology 32(8), 822–8 (2014). doi:10.1038/nbt.2939
13. Ni, J., Wu, G.D., Albenberg, L., Tomov, V.T.: Gut microbiota and IBD: Causation or correlation? Nature Reviews Gastroenterology and Hepatology 14(10), 573–584 (2017). doi:10.1038/nrgastro.2017.88
14. Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., Hercog, R., Koch, M., Luciani, A., Mende, D.R., Schneider, M.A., Schroth-King, P., Tournigand, C., Tran Van Nhung, J., Yamada, T., Zimmermann, J., Benes, V., Kloor, M., Ulrich, C.M., von Knebel Doeberitz, M., Sohann, I., Bork, P.: Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular systems biology 10, 766 (2014)
90. Arboleya, S., Watkins, C., Stanton, C., Ross, R.P.: Gut Bifidobacteria Populations in Human Health and Aging.

91. Litvak, Y., Byndloss, M.X., Büelmer, A.J.: Colonocyte metabolism shapes the gut microbiota. Science (New York, N.Y.) 362(6418) (2018). doi:10.1126/science.aat9076
Figure and table legend

Figure 1
Heatmap and hierarchical clustering of the metagenome based on their shared DNA content (k-mers of length 31). Red color indicate close similarity between metagenome as blue color indicate more dissimilarity. BMI groups are indicated by a right triangle in the upper portion of the figure and country of the individuals on the left side by upper triangles.

Figure 2
Bacteria cladogram of the gut metagenomics data found in 640 individuals down at the family taxonomic rank. Red bars indicate higher relative abundance of the taxa in obese individuals, while blue bars indicate higher representation in normal individuals. Stars indicate that the relative abundance was significantly different in both cohorts.

Figure 3
The top 4 most abundant phyla with significant changes in obese and non-obese gut microbiota.

Figure 4
The top 9 most abundant families with significant changes in obese and non-obese gut microbiota.

Figure 5
The top 20 KEGG pathways which show the most significant changes in abundances in obese and non-obese gut microbiota.

Table 1
Top 20 families based on their overall abundance in our dataset. The family codes serve as a reference for the families displayed in figure 2.

Table 2
Machine learning obesity classification results from metagenomics' protein annotations
**Figure 1** Heatmap and hierarchical clustering of the metagenome based on their shared DNA content ($k$-mers of length 31). Red color indicate close similarity between metagenome as blue color indicate more dissimilarity. BMI groups are indicated by a right triangle in the upper portion of the figure and country of the individuals on the left side by upper triangles.
Figure 2 Bacteria cladogram of the gut metagenomics data found in 640 individuals down at the family taxonomic rank. Red bars indicate higher relative abundance of the taxa in obese individuals, while blue bars indicate higher representation in normal individuals. Stars indicate that the relative abundance was significantly different in both cohorts.
Figure 3  The top 4 most abundant phyla with significant changes in obese and non-obese gut microbiota.

Figure 4  The top 9 most abundant families with significant changes in obese and non-obese gut microbiota.
Figure 5 The top 20 KEGG pathways which show the most significant changes in abundances in obese and non-obese gut microbiota.
### Table 1
Top 20 families based on their overall abundance in our dataset. The family codes serve as a reference for the families displayed in figure 2.

| Family Code | Family       | Phylum       | Overall Abundance (%) | Obese Expression (↑/↓) |
|-------------|--------------|--------------|-----------------------|------------------------|
| Lach        | Lachnospiraceae | Firmicutes   | 24.6                  | ↓                      |
| Bact        | Bacteroidaceae | Bacteroidetes| 21.8                  | ↑                      |
| R           | Ruminococcaceae | Firmicutes   | 21.6                  | ↓                      |
| C           | Clostridiaceae | Firmicutes   | 7.6                   |                        |
| Eub         | Eubacteriaceae | Firmicutes   | 7.1                   |                        |
| Pmo         | Porphyromonadaceae | Bacteroidetes | 4.5                  | ↑                      |
| O           | Oscillospiraceae | Firmicutes   | 2.9                   | ↓                      |
| Ent         | Enterobacteriaceae | Proteobacteria | 2.0                 |                        |
| Rik         | Rikenellaceae | Bacteroidetes| 1.7                   |                        |
| Baci        | Bacillaceae   | Firmicutes   | 0.85                  | ↑                      |
| Pst         | Peptostreptococcaceae | Firmicutes | 0.74                | ↓                      |
| Pbac        | Paenibacillaceae | Firmicutes  | 0.62                  |                        |
| S           | Sutterellaceae | Proteobacteria | 0.44            |                        |
| Lact        | Lactobacillaceae | Firmicutes | 0.40                  |                        |
| Pco         | Peptococcaceae | Firmicutes   | 0.36                  |                        |
| Eco         | Enterococcaceae | Firmicutes | 0.32                  |                        |
| Sco         | Streptococcaceae | Firmicutes | 0.31                  | ↓                      |
| Pre         | Prevotellaceae | Bacteroidetes| 0.31                  | ↓                      |
| F           | Flavobacteriaceae | Bacteroidetes | 0.25            | ↑                      |
| Sip         | Siphoviridae  | (Virus superkingdom) | 0.23            | ↓                      |
Table 2: Machine learning obesity classification results from metagenomics’ protein annotations.

| Taxa Phylum  | score/algo         | xgboost | lightgbm | random forest | support vector machine | decision tree | adaboost | set covering machine |
|--------------|--------------------|---------|----------|---------------|-------------------------|---------------|----------|----------------------|
| test_balanced_accuracy | 0.706              | 0.789   | 0.805    | 0.500         | 0.834                   | 0.712         | 0.630    |
| test_f1      | 0.498              | 0.561   | 0.544    | 0.049         | 0.581                   | 0.493         | 0.378    |
| test_roc_auc | 0.849              | 0.847   | 0.855    | 0.768         | 0.840                   | 0.754         | 0.630    |

| Taxa Family  | score/algo         | xgboost | lightgbm | random forest | support vector machine | decision tree | adaboost | set covering machine |
|--------------|--------------------|---------|----------|---------------|-------------------------|---------------|----------|----------------------|
| test_balanced_accuracy | 0.775              | 0.766   | 0.842    | 0.799         | 0.775                   | 0.696         | 0.705    |
| test_f1      | 0.569              | 0.602   | 0.591    | 0.511         | 0.498                   | 0.511         | 0.477    |
| test_roc_auc | 0.868              | 0.887   | 0.880    | 0.845         | 0.756                   | 0.840         | 0.705    |

| Taxa Genus   | score/algo         | xgboost | lightgbm | random forest | support vector machine | decision tree | adaboost | set covering machine |
|--------------|--------------------|---------|----------|---------------|-------------------------|---------------|----------|----------------------|
| test_balanced_accuracy | 0.754              | 0.805   | 0.726    | 0.500         | 0.723                   | 0.683         | 0.738    |
| test_f1      | 0.576              | 0.640   | 0.534    | 0.049         | 0.426                   | 0.464         | 0.519    |
| test_roc_auc | 0.841              | 0.837   | 0.871    | 0.815         | 0.766                   | 0.810         | 0.738    |

| Enzyme commission | score/algo         | xgboost | lightgbm | random forest | support vector machine | decision tree | adaboost | set covering machine |
|-------------------|--------------------|---------|----------|---------------|-------------------------|---------------|----------|----------------------|
| test_balanced_accuracy | 0.788              | 0.804   | 0.771    | 0.500         | 0.776                   | 0.759         | 0.648    |
| test_f1           | 0.630              | 0.632   | 0.606    | 0.177         | 0.489                   | 0.580         | 0.402    |
| test_roc_auc      | 0.881              | 0.871   | 0.889    | 0.870         | 0.775                   | 0.844         | 0.648    |

| Gene Ontology    | score/algo         | xgboost | lightgbm | random forest | support vector machine | decision tree | adaboost | set covering machine |
|------------------|--------------------|---------|----------|---------------|-------------------------|---------------|----------|----------------------|
| test_balanced_accuracy | 0.745              | 0.800   | 0.702    | 0.500         | 0.766                   | 0.692         | 0.609    |
| test_f1          | 0.582              | 0.652   | 0.518    | 0.049         | 0.523                   | 0.478         | 0.319    |
| test_roc_auc     | 0.882              | 0.873   | 0.894    | 0.851         | 0.710                   | 0.810         | 0.609    |

| COG              | score/algo         | xgboost | lightgbm | random forest | support vector machine | decision tree | adaboost | set covering machine |
|------------------|--------------------|---------|----------|---------------|-------------------------|---------------|----------|----------------------|
| test_balanced_accuracy | 0.824              | 0.813   | 0.830    | 0.669         | 0.813                   | 0.809         | 0.806    |
| test_f1          | 0.604              | 0.615   | 0.665    | 0.412         | 0.544                   | 0.610         | 0.627    |
| test_roc_auc     | 0.890              | 0.896   | 0.897    | 0.681         | 0.810                   | 0.813         | 0.806    |
| KEgg Pathway Score/Algo | xgboost | lightgbm | Random Forest | Support Vector Machine | Decision Tree | Adaboost | Set Covering Machine |
|-------------------------|---------|----------|---------------|------------------------|--------------|----------|---------------------|
| Test Balanced Accuracy  | 0.769   | 0.763    | 0.766         | 0.500                  | 0.785        | 0.730    | 0.708               |
| F1                      | 0.587   | 0.587    | **0.603**     | 0.049                  | 0.514        | 0.554    | 0.486               |
| ROC AUC                 | 0.860   | 0.869    | **0.889**     | 0.837                  | 0.791        | 0.855    | 0.708               |

| KEgg Module Score/Algo | xgboost | lightgbm | Random Forest | Support Vector Machine | Decision Tree | Adaboost | Set Covering Machine |
|------------------------|---------|----------|---------------|------------------------|--------------|----------|---------------------|
| Test Balanced Accuracy | 0.768   | **0.835**| 0.687         | 0.500                  | 0.790        | 0.737    | 0.718               |
| F1                     | 0.583   | **0.621**| 0.499         | 0.049                  | 0.528        | 0.553    | 0.505               |
| ROC AUC                | 0.860   | **0.895**| 0.893         | 0.872                  | 0.798        | 0.879    | 0.718               |