



HHS Public Access

Author manuscript

ACM BCB. Author manuscript; available in PMC 2019 October 04.

Published in final edited form as:

ACM BCB. 2019 September ; 2019: 619–625. doi:10.1145/3307339.3343476.

An Integrated Approach for Efficient Multi-Omics Joint Analysis

Massimiliano S. Tagliamonte,

Dept. of Path., Imm., and Lab. Med, College of Medicine, UF, Gainesville, FL, USA

Sheldon G. Waugh,

Army Public Health Center, Aberdeen Proving Ground, Aberdeen, MD, USA

Mattia Prosperi,

Department of Epidemiology University of Florida Gainesville, FL, USA

Volker Mai

Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, Emerging Pathogens Institute, University of Florida

Abstract

The challenges associated with multi-omics analysis, e.g. DNA-seq, RNA-seq, metabolomics, methylomics and microbiomics domains, include: (1) increased high-dimensionality, as all -omics domains include ten thousands to hundreds of thousands of variables each; (2) increased complexity in analyzing domain-domain interactions, quadratic for pairwise correlation, and exponential for higher-order interactions; (3) variable heterogeneity, with highly skewed distributions in different units and scales for methylation and microbiome. Here, we developed an efficient strategy for joint-domain analysis, applying it to an analysis of correlations between colon epithelium methylomics and fecal microbiomics data with colorectal cancer risk as estimated by colorectal polyp prevalence. First, we applied domain-specific standard pipelines for quality assessment, cleaning, batch-effect removal, et cetera. Second, we performed variable homogenization for both the methylation and microbiome data sets, using domain-specific normalization and dimension reduction, obtaining scale-free variables that could be compared across the two domains. Finally, we implemented a joint-domain network analysis to identify relevant microbial-methylation island patterns. The network analysis considered all possible species-island pairs, thus being quadratic in its complexity. However, we were able to pre-select the unpaired variables by performing a preliminary association analysis on the outcome polyp prevalence. All results from association and interaction analyses were adjusted for multiple comparisons. Although the limited sample size did not provide good power (80% to detect medium to large effect sizes with 5% alpha error), a number of potentially significant association (dozens in the uncorrected analysis, reducing to just a few in the corrected one) were identified. As a last step, we linked the network patterns identified by our approach to the KEGG functional

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

mstagliamonte@ufl.edu.

ontology, showing that the method can generate new mechanistic hypotheses for the biological causes of polyp development.

Keywords

Methylation; microbiome; bioinformatics; principal component analysis; dimension reduction; network; correlation; joint analysis

1 Introduction

Colorectal cancer (CRC) is among the leading causes of cancer and cancer mortality in the United States [1]. While the human gut microbiota is thought to contribute to maintaining health, detailed mechanisms often remain elusive. In terms of CRC risk, it has been hypothesized that microbiota activities can contribute to changes in the gut epithelium that might drive carcinogenesis. While several species have been associated with CRC [2], for a few there are proposed mechanisms of induction. Several bacterial species, such as *Acidovorax spp.*, *Bacteroides fragilis*, *E. coli*, and *Fusobacterium nucleatum* might play a role in CRC development by inducing inflammation and host DNA damage [3–7]. Other authors proposed more generic mechanisms related to the influence of gut microbiota on folate levels [8–10], and its potential effect on cells DNA methylation and repair mechanisms [11].

A more comprehensive approach is needed to clarify how the overall microbiome might affect host gene expression and cells transformation; various high throughput analysis platforms now allow for an in depth analysis of correlations between multiple domains of complex data variables with health outcomes of interest. However, an integrated analysis approach for these highly complex data domains to derive new mechanistic insights into disease mechanisms remains a challenge. Adenomatous colorectal polyps represent an important risk factor preceding CRC and as such they form the basis for current CRC screening approaches. While gut epithelial RNA expression adapts quickly to changes in the local gut environment and thus is inherently unstable, gut epithelial methylation patterns in these tissues appear more stable. Thus, we aimed to develop an analytical approach to explore microbiota associations with gut epithelium methylation pattern to gain insight into potential CRC mechanisms and to develop the foundation for improving on current CRC screening modalities.

The challenges associated with joint-domain analysis such as multi-omics, e.g. methylation and microbiome data, include: [1] increased high-dimensionality, as both -omics domains include ten thousands to hundreds of thousands of variables each; [2] increased complexity in analyzing domain-domain interactions, quadratic for pairwise correlation, and exponential for higher-order interactions; [3] variable heterogeneity, with highly skewed distributions in different units and scales for methylation and microbiome. Here, we present an efficient strategy that tackles all three issues, by analyzing first datasets within their domain, reducing dimensionality, and then looking at relevant inter-domain associations using pairwise correlation networks. We anticipate that, although a number of higher-order interactions are

neglected at the inter-domain step (yet not at the intra-domain), the method is still able to explain a large portion of joint variance and association with outcomes.

2 Methods

2.1 Sample Collection and case definition

Stool samples were stored at -80°C pending analysis. Biopsy samples were placed in RNAlater stabilization solution (ThermoFisher) and fixed for 24 hours at 4°C . Samples were then transferred to -20°C pending analysis. For the current analysis aimed at linking microbiota with methylation pattern, cases ($N=13$) were defined as having at least one polyp identified by biopsy.

Case/high-risk classification was based on the presence of at least a total of 5mm of polyp size, either by one large or multiple smaller polyps [12]. Classified samples were matched on BMI and age, alleviate differential effects based on these variables. All samples and data were obtained from a previously completed colorectal cancer screening study performed at UF.

2.2 Methylation Analysis

After rinsing the biopsy surfaces and further cleaning by vortexing with 3mm glass beads to remove any mucus and attached bacteria, human DNA were extracted from biopsy samples using Qiagen DNA Blood and Tissue kit (Qiagen). Methylation status was determined using the Infinium MethylationEPIC BeadChip. Quality controlled methylation datasets, provided by HudsonAlpha, were filtered and normalized utilizing the R package Watermelon. Quality control (QC) [13], preprocessing, and normalization of Illumina HumanMethylationEPIC 850K BeadArray methylation data was performed using the minfi bioconductor package [14]. To ensure high-quality methylation data for downstream analyses, quantile normalization was applied to normalize the between array technical variation based on internal control probes. Additionally, batch and chip effect adjustments were implemented using an empirical Bayes batch-correction method (ComBat) [15] as implemented in the Bioconductor sva package [16]. Probes with a median detection p-value >0.01 were identified and excluded from subsequent statistical analyses, as well as SNP-associated, i.e. reported SNPs at the queried CpG site and at the single-base extension (SBE) site, and cross-reactive CpGs. Finally to reduce invariant sites, following probe exclusions, a total of 794,271 CpG sites were retained and used in downstream statistical analyses.

2.3 High Throughput Microbiota Sequencing and Analysis

DNA was isolated from fecal samples using the Qiagen stool DNA extraction protocol, modified to include includes a bead-beating step. Illumina MiSeq (2×250 bp) sequencing was performed using barcoded primers with 50 % PhiX spike. The 16S rRNA gene (V1–V2 Region) was amplified using validated primers MiSeq 27F 5' - AATGATACGGCGACCACCGAGATCTACAC TATGGTAATT CC AGMGTTYGATYMTGGCTCAG-3' containing 5' Illumina adapter; primer pad; primer linker; and forward primer MiSeq-338R (reverse primer) PCR primer sequence (each sequence contains different barcode) was used with the sequence 5' -

CAAGCAGAAGACGGCATAACGAGAT TCCCTTGTCTCC AGTCAGTCAG AA
GCTGCCTCCCGTAGGAGT-3' containing reverse complement of 3' Illumina adapter;
Golay barcode; primer pad; primer linker; and reverse primer. The PCR conditions were the following: initial melting step at 95 °C for 2 min, followed by 20 cycles of 95 °C for 30 s, 50 °C for 30 s, and 72 °C for 1 min 30 s. Two 50 µl PCR reactions for each sample were combined together and the PCR products were purified with Agencourt AMPure XP system (Beckman Coulter). Cleaned PCR products were pooled in equimolar amounts and submitted for sequencing. Microbiome analyses were performed using the Quantitative Insights into Microbial Ecology (QIIME) tool [17] and the Brazilian Microbiome Project (BMP) [18], binning sequences into Operational Taxonomic Units (OTUs) using a 98 percent similarity level.

2.4 Feature Reduction, Correlational and Classification Methods

Our methodology focused on three separate methods for reducing dimensionality and identifying associations between methylation and microbiome data. In detail, Figure 1 illustrates the flowchart for intra- and joint-domain analysis including quality control, dimension reduction, correlation filtering, and network build-up.

We utilized Sparse Independent Principal Component Analysis (sIPCA) on the methylation dataset to correct for high dimensionality and noisy characteristics. sIPCA uses a feature selection tool that chooses a user specified number of variables per Independent Principal Component (IPC). After initial quality control (removal of CpGs sites with beta values less than the average variance of the methylation dataset), quantile normalization and the removal of CpG sites known to interact with SNPs, the sIPCA analysis selected one percent of the total CpG sites to represent each IPC.

To identify relevant joint hotspots between methylation and gut microbiota patterns, we first identified strong (and statistically supported) intra-domain independent associations with the high-risk outcome, followed by an inter-domain correlation analysis. The strength and significance of associations were determined by Spearman's correlation analysis, which is more robust than Pearson's with respect to rank (not necessarily linear). Correlational analyses were conducted between CPGs and OTUs, OTUs and IPCs, IPCs and Genera, and Genera and CPGs.

Visualization of the correlational networks was performed using Cytoscape 3.6.0 [19], displaying the strength, significance and directionality of each pairwise correlation.

For each created principal component, we utilized Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis [20–22] to identify the genes and protein pathways associated with the selected CpG sites and OTUs. For the methylation components, the R package missMethyl [23] was used to reference the selected CpG sites into its respective KEGG functional group.

Random forest models were generated using the AUC-RF algorithm [24] for feature reduction and maximizing model performance. The most predictive OTUs were determined based on mean decrease in accuracy when removed from the model. The area under the

curve (AUC) of receiver operator characteristic (ROC) curves were compared using the pROC R package [25].

3 Results

A total of 27 subjects' biopsy samples were analyzed for methylation patterns; of these samples, 25 of the biopsy were also analyzed for microbiome; in addition, 23 stool samples from the same subjects were also utilized for microbiome analysis.

3.1 Microbiome Analysis

For the analysis, we retained a total of 7,597,922 high-quality sequences with a mean of 48,088 sequences/sample and an average length of 327 nucleotides. After removal of OTUs containing less than 10 sequences we retained 10,564 OTUs at the 98 percent similarity level. After removal of unclassified OTUs and retaining of the ones constituting 70% of the abundance (core 70%), 234 OTUs spanning 70 genera remained

3.2 Methylation Analysis

After removing poor quality probes, and performing quantile and batch-effect normalization, 794,421 CpG sites were retained for analysis. An initial principal component analysis demonstrated some separation between the high and low risk samples. After the secondary quality control with the removal of CpG sites with a variance less than the mean (512,211 CpG sites) and CpG sites with known SNP expression (60,515 CpG sites), 166,331 total CpG sites remain for further analysis.

3.3 SIPCA Analysis

The SIPCA analyses yielded eight components accounting for approximately 75% of all the variance in the methylation dataset. Eight components were eventually selected due to the sudden decrease in kurtosis measurement between the eighth and ninth component.

3.4 Correlation Analysis with SIPCA

The correlation analyses between the OTUs and CPG sites yielded 9,442 pairwise comparisons with a p-value smaller than 0.00005 (Figure 2). Figure 3A visualizes the pairwise comparisons utilizing a force-undirected graph setup. Correlation analyses between the created IPCs and genera yielded 21 pairwise connections with a p-value below 0.05. Within these connections, we observed IPC5, IPC4 and IPC2 having multiple positive and negative correlations with the genera *Sutterella*, *Megamonas*, *Pseudomonas*, *Enhydrobacter*, *Eubacterium*, *Prevotella*, *Citrobacter* and *Catenibacterium* (Figure 3B).

Correlation analyses between individual OTUs and IPCs yielded 52 pairwise connections with a p value below 0.05, with 115 OTUs and all eight IPCs.

To potentially decipher the hidden gene functions behind the combinations of CPG sites, among the eight IPCs, we conducted a KEGG analysis on the selected CPG sites, in each IPC to determine the primary function of each component.

The IPCs accounting for the most variance were associated with pathways involved various kinds of cancer, such as cellular senescence, p53 signaling, and cytoskeleton and junctions regulation. A more in-depth analysis of these correlations is needed; a larger sample size might lead to more specific associations based on the severity of the lesions (i.e. cases vs high-risk CRC).

3.5 Classification Analysis with SIPCA

Observing the random forest prediction models, we compared the predictive ability of using the IPCs, OTU and a combination of both to determine risk classification. The combination model performed better than the IPC or genera bases model with an AUC of 0.897. The genera-only based model performed the next best with an AUC of 0.846, with the IPC-only model performing last with an AUC of 0.692. Importance plots for the three models show consistency in the important contributors to correctly predicting risk classification. The top four contributors to the combination model (Genera: *Provetella*, *Disllaister*, IPC2 and IPC8), consistently score high in the combo model and their independent group models, respectively, with the highest Mean Decrease of Accuracy scores. While limited by the small sample size, we have shown here that our approach facilitated the integration of multiple biological data domains to explore novel mechanisms associated with CRC risk. From the observation of large degrees of variation in this multi-omics dataset we gained insight into the sample size requirements of sufficiently powered future studies. Such studies can further explore mechanisms derived from multi-omics data that have might have potential utility in developing better CRC screening approaches.

4 Conclusions

It is widely accepted that gut microbiome composition has an effect on human health, although the mechanisms are not always clear. Bacteria can affect phosphorylation and acetylation status of specific promoters, and different species might counteract each other effects [26–28]. While specific bacteria have been associated with CRC at different stages [29–32], possible mechanisms of causation remain yet to be cleared Future analyses on a larger sample size might lead to an insight of pathways involved in specific methylation changes.

From a methodological point of view, we demonstrated that the modular intra- and inter-domain analysis is relatively scalable and the reduce dimensionality yet explains a good portion of the variance, and retains fair classification accuracy The network approach is also useful to generate new mechanistic hypotheses, especially in conjunction with the KEGG pathways.

Further investigations, with a larger sample size, are needed to have a clearer insight on the microbiota impact on CLC risk, although the use of larger datasets may effect scalability

Data and code used in the present study can be find at the following address: <https://osf.io/v4hns/?viewonly=e0f7c9c1ea054c768226aa2bbb5b5f21>

ACKNOWLEDGMENTS

This work was funded in part by NIH NCI R21CA195251-01A1 and NIH NIAID 1R01AI141810-01.

REFERENCES

- [1]. Prevention CfDCA. United States Cancer Statistics: Data Visualizations [Accessed on 06/23/2019]. Available from: <https://gis.cdc.gov/Cancer/USCS/DataViz.html>.
- [2]. Peters BA, Dominianni C, Shapiro JA, Church TR, Wu J, Miller G, et al. The gut microbiota in conventional and serrated precursors of colorectal cancer. *Microbiome*. 2016;4(1):69. [PubMed: 28038683]
- [3]. Tanaka N, Che FS, Watanabe N, Fujiwara S, Takayama S, Isogai A. Flagellin from an incompatible strain of *Acidovorax avenae* mediates H₂O₂ generation accompanying hypersensitive cell death and expression of PAL, Cht-1, and PBZ1, but not of Lox in rice. *Mol Plant Microbe Interact*. 2003; 16(5):422–8. [PubMed: 12744513]
- [4]. Tosolini M, Kirilovsky A, Mlecnik B, Fredriksen T, Mauger S, Bindea G, et al. Clinical impact of different classes of infiltrating T cytotoxic and helper cells (Th1, th2, treg, th17) in patients with colorectal cancer. *Cancer Res*. 2011;71(4): 1263–71. [PubMed: 21303976]
- [5]. Arthur JC, Perez-Chanona E, Muhlbauer M, Tomkovich S, Uronis JM, Fan TJ, et al. Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science*. 2012;338(6103): 120–3. [PubMed: 22903521]
- [6]. Buc E, Dubois D, Sauvanet P, Raisch J, Delmas J, Darfeuille-Michaud A, et al. High prevalence of mucosa-associated *E. coli* producing cyclomodulin and genotoxin in colon cancer. *PLoS One*. 2013;8(2):e56964. [PubMed: 23457644]
- [7]. Rubinstein MR, Wang X, Liu W, Hao Y, Cai G, Han YW. *Fusobacterium nucleatum* promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin. *Cell Host Microbe*. 2013;14(2):195–206. [PubMed: 23954158]
- [8]. D’Aimmo MR, Mattarelli P, Biavati B, Carlsson NG, Andlid T. The potential of bifidobacteria as a source of natural folate. *J Appl Microbiol*. 2012;112(5):975–84. [PubMed: 22335359]
- [9]. Pompei A, Cordisco L, Amaretti A, Zanoni S, Matteuzzi D, Rossi M. Folate production by bifidobacteria as a potential probiotic property. *Appl Environ Microbiol*. 2007;73(1): 179–85. [PubMed: 17071792]
- [10]. Sugahara H, Odamaki T, Hashikura N, Abe F, Xiao JZ. Differences in folate production by bifidobacteria of different origins. *Biosci Microbiota Food Health*. 2015;34(4):87–93. [PubMed: 26594608]
- [11]. O’Reilly SL, McGlynn AP, McNulty H, Reynolds J, Wasson GR, Molloy AM, et al. Folic Acid Supplementation in Postpolypectomy Patients in a Randomized Controlled Trial Increases Tissue Folate Concentrations and Reduces Aberrant DNA Biomarkers in Colonic Tissues Adjacent to the Former Polyp Site. *J Nutr*. 2016;146(5):933–9. [PubMed: 27075913]
- [12]. Kessler WR, Imperiale TF, Klein RW, Wielage RC, Rex DK. A quantitative assessment of the risks and cost savings of forgoing histologic examination of diminutive polyps. *Endoscopy*. 2011;43(8):683–91. [PubMed: 21623556]
- [13]. Pidsley R, YW CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14:293. [PubMed: 23631413]
- [14]. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10): 1363–9. [PubMed: 24478339]
- [15]. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27. [PubMed: 16632515]
- [16]. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882–3. [PubMed: 22257669]

- [17]. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335–6. [PubMed: 20383131]
- [18]. Pyro VS, Roesch LF, Ortega JM, do Amaral AM, Totola MR, Hirsch PR, et al. Brazilian Microbiome Project: revealing the unexplored microbial diversity--challenges and prospects. *Microb Ecol*. 2014;67(2):237–41. [PubMed: 24173537]
- [19]. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003; 13(11):2498–504. [PubMed: 14597658]
- [20]. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27–30. [PubMed: 10592173]
- [21]. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45(D1):D353–d61. [PubMed: 27899662]
- [22]. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 2019;47(D1):D590–d5. [PubMed: 30321428]
- [23]. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*. 2016;32(2):286–8. [PubMed: 26424855]
- [24]. Calle ML, Urrea V, Boulesteix AL, Malats N. AUC-RF: a new strategy for genomic profiling with random forest. *Hum Hered*. 2011;72(2): 121–32. [PubMed: 21996641]
- [25]. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011; 12:77. [PubMed: 21414208]
- [26]. Opitz B, Puschel A, Beermann W, Hocke AC, Forster S, Schmeck B, et al. Listeria monocytogenes activated p38 MAPK and induced IL-8 secretion in a nucleotide-binding oligomerization domain 1-dependent manner in endothelial cells. *J Immunol*. 2006; 176(1):484–90. [PubMed: 16365441]
- [27]. Schmeck B, Beermann W, van Laak V, Zahlten J, Opitz B, Witzernath M, et al. Intracellular bacteria differentially regulated endothelial cytokine release by MAPK-dependent histone modification. *J Immunol*. 2005; 175(5):2843–50. [PubMed: 16116170]
- [28]. Haller D, Holt L, Kim SC, Schwabe RF, Sartor RB, Jobin C. Transforming growth factor-beta 1 inhibits non-pathogenic Gram negative bacteria-induced NF-kappa B recruitment to the interleukin-6 gene promoter in intestinal epithelial cells through modulation of histone acetylation. *J Biol Chem*. 2003;278(26):23851–60. [PubMed: 12672795]
- [29]. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *Isme j*. 2012;6(2):320–9. [PubMed: 21850056]
- [30]. Drewes JL, White JR, Dejea CM, Fathi P, Iyadorai T, Vadivelu J, et al. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *NPJ Biofilms Microbiomes*. 2017;3:34. [PubMed: 29214046]
- [31]. Nakatsu G, Li X, Zhou H, Sheng J, Wong SH, Wu WK, et al. Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat Commun*. 2015;6:8727. [PubMed: 26515465]
- [32]. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun*. 2015;6:6528. [PubMed: 25758642]

CCS CONCEPTS

- Applied computing~Biological networks
- Applied computing~Bioinformatics
- Applied computing~Molecular sequence analysis

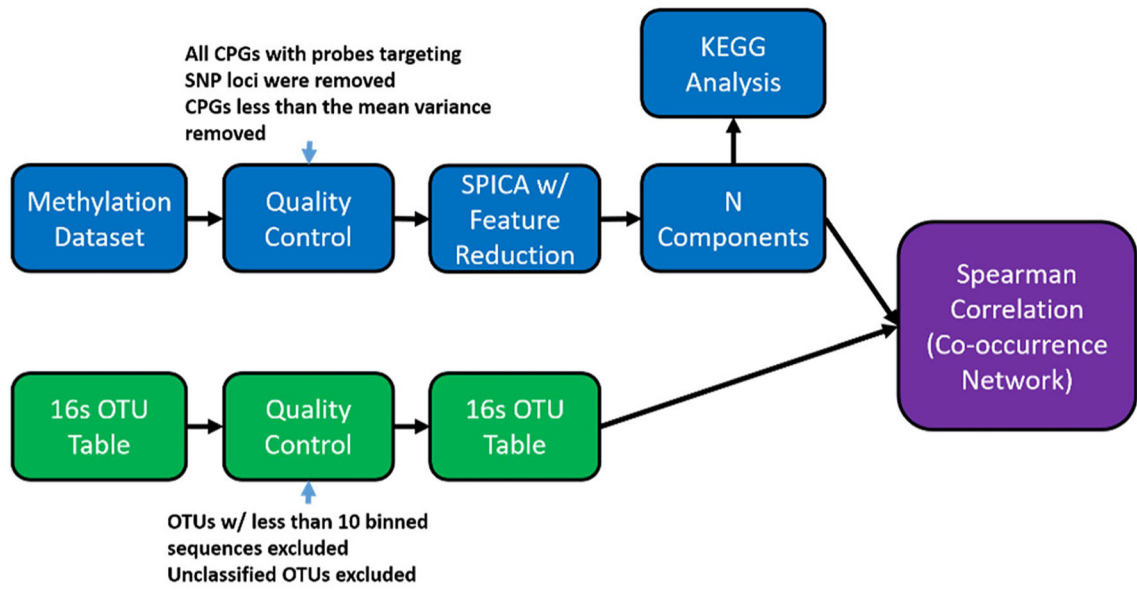
Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Correlation Analysis with SIPCA



Classification with SIPCA and AUCRF

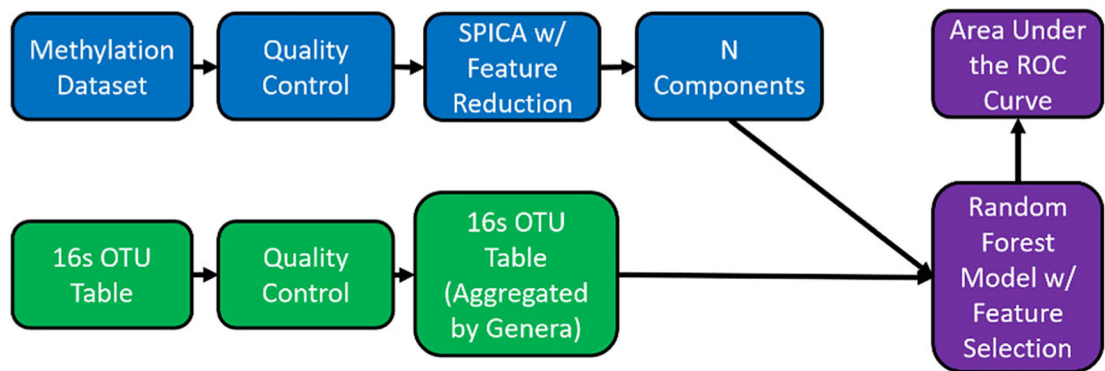


Figure 1. Flowchart for intra- and joint-domain analysis including quality control, dimension reduction, correlation filtering, and network build-up.

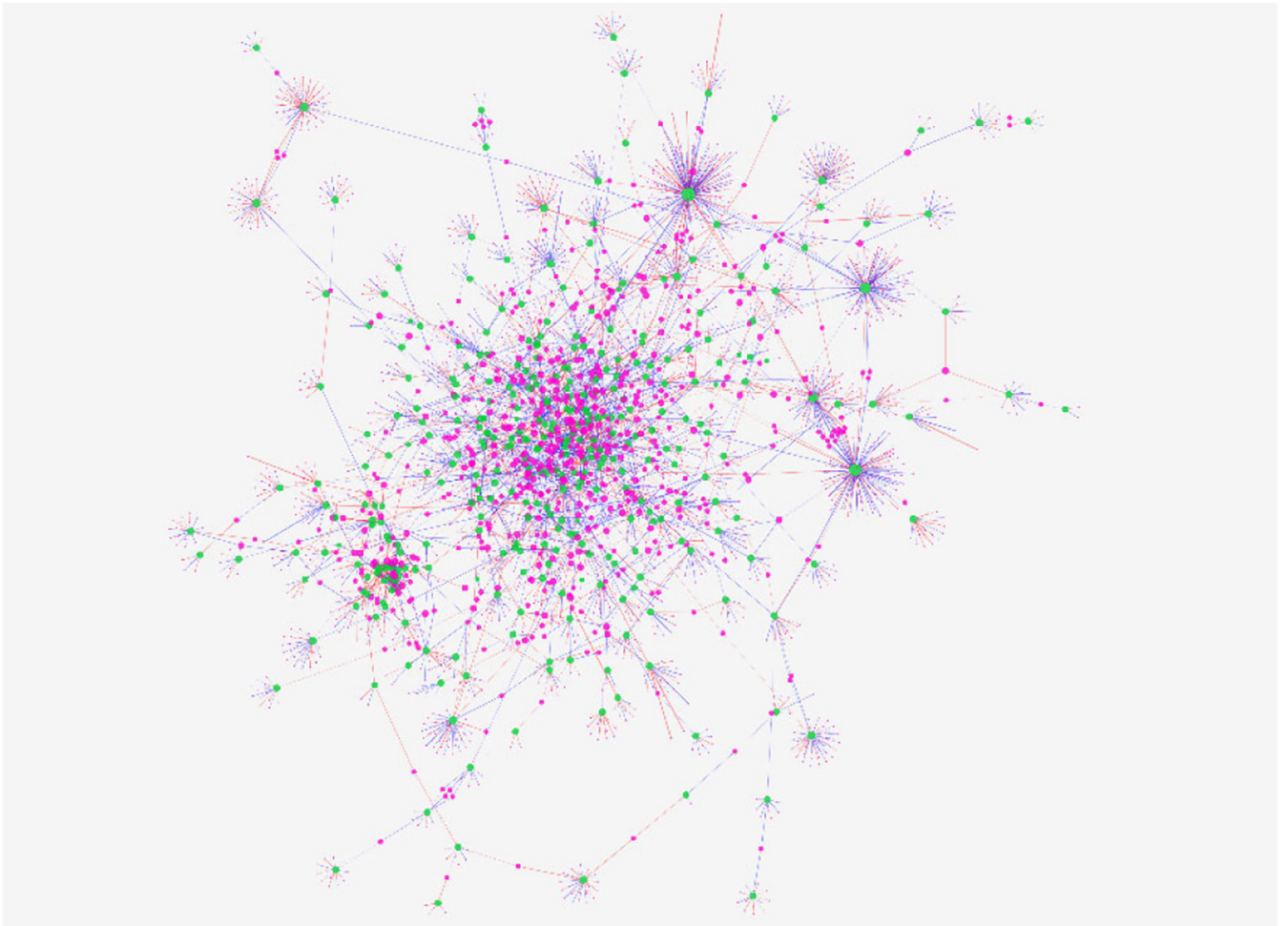


Figure 2. Correlation-focused analysis of Operational Taxonomic Units and methylation CpG sites positively or negatively associated each other.

The size of nodes (OTUs) is scaled by the number of network degree, where larger diameter circles indicate a greater abundance of network connections. The color of the nodes is stratified by belonging to the OTU (green) or CpG (pink) group. The correlation connections are depicted by either a positive or negative correlation (Blue: Positive, Red: Negative) and the overall strength (thickness of the network link).

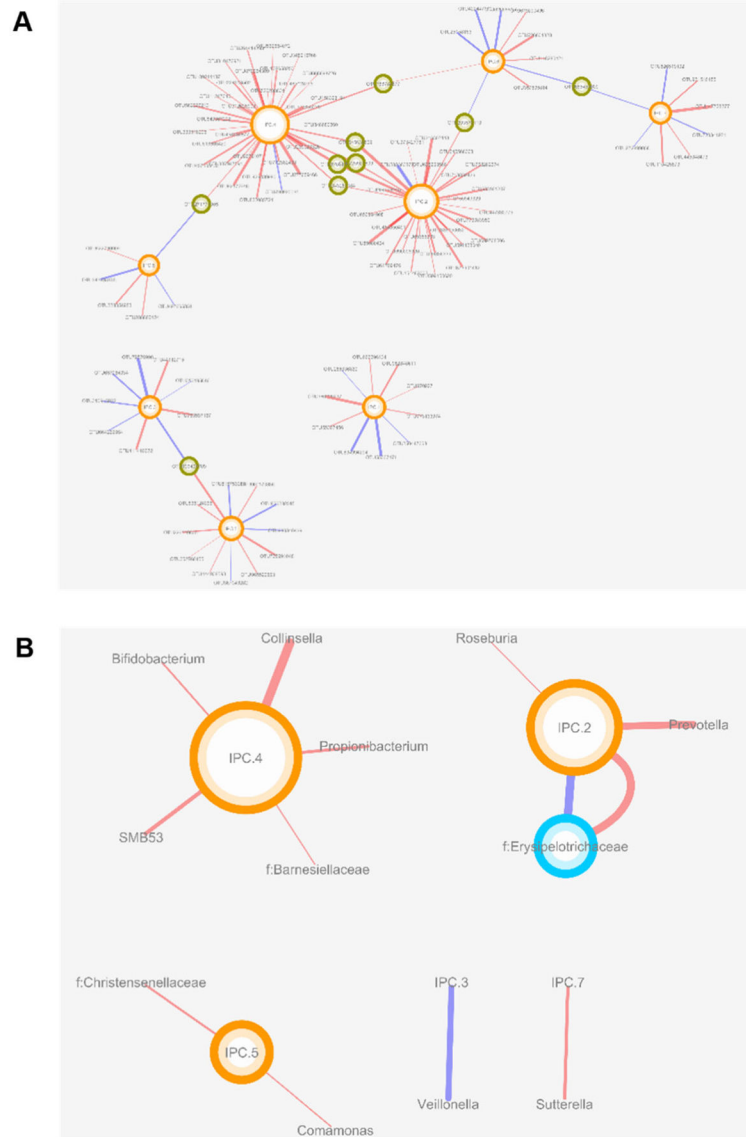


Figure 3. Correlation-focused analysis of Operational Taxonomic Units and methylation CpG sites.

The size of nodes is scaled by the number of network degree, where larger diameter circles indicate a greater abundance of network connections. The correlation connections are depicted by either a positive or negative correlation (Blue: Positive, Red: Negative) and the overall strength (thickness of the network link). **A:** Correlation-focused analysis of Operational Taxonomic Units and methylation CpG sites positively or negatively associated each other. The color of the nodes is stratified by belonging to the OTU or CpG group. **B:** Correlation-focused analysis of Genera and Independent Components positively or negatively associated each other. The color of the nodes is stratified by belonging to the Genera or IPC group.