Title

Functional Anabolic Network Analysis of Human-associated *Lactobacillus* Strains

Authors

Thomas J. Moutinho Jr.[a], Benjamin C. Neubert[a], Matthew L. Jenior[a], Maureen A. Carey[a,b], Gregory L. Medlock[a], Glynis L. Kolling[a], Jason A. Papin[a]#

[a] Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, USA

[b] Division of Infectious Disease and International Health, Department of Medicine, University of Virginia, Charlottesville, Virginia, USA

Running Head: Anabolic Network Analysis of Lactobacilli

#Address correspondence to Jason A. Papin, papin@virginia.edu

Abstract

Members of the *Lactobacillus* genus are frequently utilized in the probiotic industry with many species conferring demonstrated health benefits; however, these effects are largely strain-dependent. We designed a method called PROTEAN (Probabilistic Reconstruction Of constituent Anabolic Networks) to computationally analyze the genomic annotations and predicted metabolic production capabilities of 144 strains across 16 species of *Lactobacillus* isolated from human intestinal, oral, and vaginal body sites. Using PROTEAN we conducted a genome-scale metabolic network comparison between strains, revealing that metabolic capabilities differ by isolation site. Notably, PROTEAN does not require a well-curated genome-scale metabolic network reconstruction to provide biological insights. We found that predicted metabolic capabilities of lactobacilli isolated from the vaginal microbiota cluster separately from intestinal and oral isolates, and we also uncovered an overlap in the predicted metabolic production capabilities of intestinal and oral isolates. Using machine learning, we determined the most informative metabolic products driving the difference between predicted metabolic capabilities of intestinal, oral, and vaginal isolates. Notably, intestinal and oral isolates were predicted to have a higher likelihood of producing D-alanine, D/L-serine, and L-proline, while the vaginal isolates were distinguished by a higher predicted likelihood of producing L-arginine, citrulline, and D/L-lactate. We found the distinguishing products to be consistent with published experimental literature. This study showcases a systematic technique, PROTEAN, for comparing the predicted functional metabolic output of microbes using genome-scale metabolic network analysis and computational modeling and provides unique insight into human-associated *Lactobacillus* biology.

Importance

The *Lactobacillus* genus has been shown to be important for human health. Lactobacilli have been isolated from human intestinal, oral, and vaginal sites. Members of the genus contribute significantly to the maintenance of vaginal health by providing colonization resistance to invading pathogens. A wide variety of clinical studies have indicated that *Lactobacillus*-based probiotics confer health benefits for

several gut- and immune-associated diseases. Microbes interact with the human body in several ways, including the production of metabolites that influence physiology or other surrounding microbes. We have conducted a strain-level genome-scale metabolic network reconstruction analysis of human-associated *Lactobacillus* strains, revealing that predicted metabolic capabilities differ when comparing intestinal/oral isolate to vaginal isolates. The technique we present here allows for direct interpretation of discriminating features between the experimental groups.

Introduction

Lactobacillus is a diverse genus of bacteria with many member strains associated with the human body. Lactobacilli are Gram-positive, lactic acid-producing bacteria typically with a low GC content (1,2). They are known for their production of lactic acid, being facultative anaerobes, and are capable of being metabolically active in a large variety of conditions (3). There is evidence that human-associated lactobacilli colonize mucosal surfaces of the intestinal tract (4), vagina (5–12), and oral cavity (13,14). While strains of *Lactobacillus* have been isolated from all three of these body sites, it remains unknown which are permanent members of the resident microbiota (autochthonous) opposed to transient members (allochthonous). Transient intestinal lactobacilli are either resident members of the oral microbiota or have been ingested, most commonly from unpasteurized fermented foods (4,15).

Lactobacilli have been used for a broad range of applications primarily associated with human intestinal probiotics and industrial production of useful metabolites. *Lactobacillus*-based probiotics have been shown to confer health benefits in clinical studies for a variety of conditions including prevention of antibiotic associated diarrhea (16), *Clostridium difficile*-associated diarrhea (17), constipation (18), irritable bowel syndrome (19), and eczema/atopic dermatitis (20). Probiotics are controversial, likely due to claims made by currently marketed probiotics that lack FDA approval for the treatment of specific diseases (21,22). The primary benefits associated with lactobacilli-based probiotics may be a function of their presence in the gut, production of metabolites, and modulation of the immune system (23,24). Metabolism plays a key role in all three of these general mechanisms; therefore, a better understanding of their metabolic capabilities will help to elucidate the mechanisms contributing to probiotic effects (25).

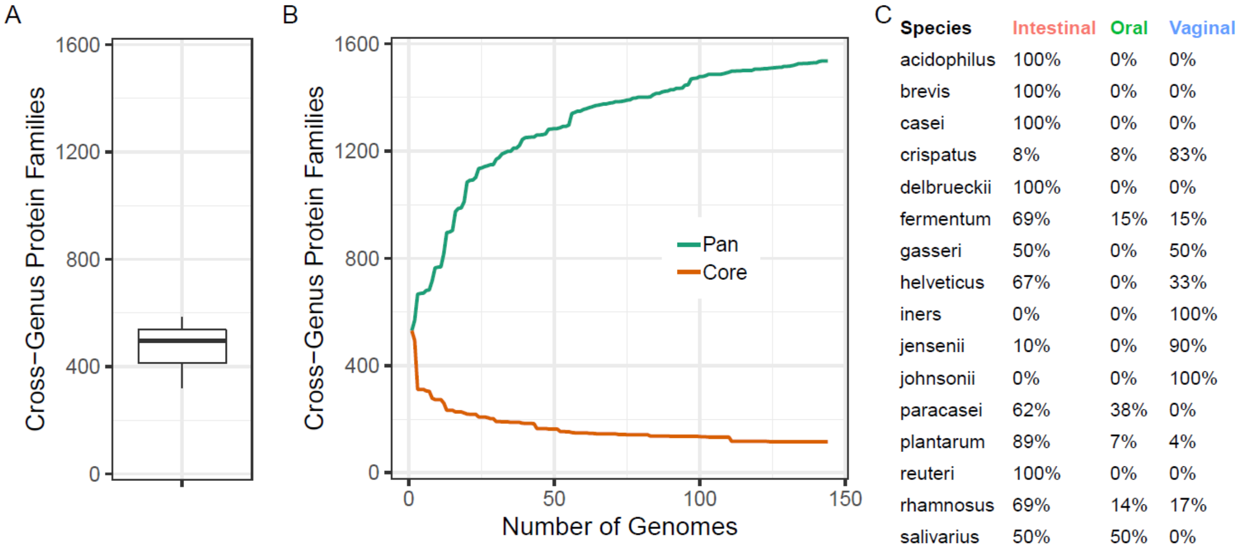
In recent years, there has been an explosion of genomic and metagenomic sequencing of human-associated microbiota, which provides a unique opportunity to apply genome-scale metabolic network reconstructions (GENREs) to enhance our current understanding of human-associated lactobacilli metabolism utilizing *in silico* techniques (25). Systems biology has the potential to advance design, selection, and delivery of *Lactobacillus*-based probiotics (26,27). GENREs are a powerful computational tool for mathematically modeling the metabolic processes within a cell at a systems-level, including all known metabolic reactions, metabolites, and metabolic genes in an organism (28). GENREs are created by referencing an annotated genome against biochemical databases, then integrating experimental data when available (29). There are several examples of *Lactobacillus*-specific comparative genomics studies (30–35); however, GENREs allow for a more functional perspective than genomics data alone because of the quantitative accounting for interactions between components in the network (25,36). Simulations with GENREs can accurately predict microbial growth yields and the metabolic pathways utilized for the production of metabolites during exponential growth of a microbe (37). A variety of analytical approaches can be applied to interrogate emergent properties of a GENRE. Flux Balance Analysis (FBA) and related methods have proven highly successful in the analysis of metabolic networks (38). FBA is a mathematical technique for analyzing the flow of metabolites through a GENRE; it can be used to identify a set of reaction fluxes that maximize growth in a specified media condition among other applications (28,39,40). Metabolic network reconstructions and FBA provide a mechanistic look into cellular metabolism and are increasingly used to study biochemical processes of single bacterial species as well as communities of organisms (41).

GENREs enable the computational prediction of metabolic capabilities of microbes, both catabolic and anabolic. Additionally, GENREs are capable of contextualizing large 'omic datasets (i.e. genomics, transcriptomics, and metabolomics) with known biochemistry and biological network architectures for improved understanding of the experimental data (42). An important recent finding demonstrated that metabolomics data alone can be used to differentiate between bacterial cultures at the strain level (43). We developed a computational method using GENREs to predict the metabolic products that a strain is likely able to produce. We used predicted production capabilities to then differentiate between different human-associated *Lactobacillus* strains. Just as metabolomics data can be used to differentiate bacterial strains, predicted production capabilities can be used for the same comparisons. We assessed the metabolic potential across a broad set of *Lactobacillus* species, consisting of 144 strains, which have all been isolated from three human-related body sites: intestinal, oral, and vaginal. We found that intestinal and oral isolates have a great deal of overlap in their metabolic functionality, while vaginal isolates have more unique metabolic production capabilities. These analyses can facilitate additional experimental interrogation of this important genus of bacteria.

Results and Discussion

Annotated metabolic genes associated with known metabolic functions are sufficiently represented among human-associated lactobacilli

In this study we predict the metabolic production capabilities of 144 lactobacilli strains. We utilized the PATRIC Cross-Genus Protein Families (PGfams) (4) for an initial genomic analysis. PGfams are comparable clusters of proteins that likely have similar functions. These clusters are intended to be used for cross-genus comparison due to their slightly relaxed clustering criteria. However, PGfams allow for the comparison of the large number of strains analyzed in this study. Lactobacilli consist of a broad range of species and thus using the PGfams was appropriate for an initial genomic comparison in this study. We first filtered the PGfams to only include metabolic gene families associated with known metabolic functions (see Methods). The distribution of total metabolic PGfams associated with each genome ranges from 340 to 580 and has a median value of 515 (Figure 1A). Across these 144 strains we found that they share 116 core metabolic PGfams, spanning a variety of cellular functions including, but not limited to, carbohydrate, nucleotide, and amino acid metabolism (Table S1). The pan set of metabolic PGfams, which represents the total set of unique PGfams, expanded to over 1500 after considering all strains utilized within this study (Figure 1B). The *Lactobacillus* strains we studied consisted of 16 species and were isolated from intestinal, oral, and vaginal human body sites (Figure 1C).



Species	Intestinal	Oral	Vaginal
acidophilus	100%	0%	0%
brevis	100%	0%	0%
casei	100%	0%	0%
crispatus	8%	8%	83%
delbrueckii	100%	0%	0%
fermentum	69%	15%	15%
gasseri	50%	0%	50%
helveticus	67%	0%	33%
iners	0%	0%	100%
jensenii	10%	0%	90%
johnsonii	0%	0%	100%
paracasei	62%	38%	0%
plantarum	89%	7%	4%
reuteri	100%	0%	0%
rhamnosus	69%	14%	17%
salivarius	50%	50%	0%

Figure 1: Known metabolic annotations are extensively sampled across the 16 *Lactobacillus* species included in this study. The genomic features used for this analysis are PATRIC Cross-Genera Protein families (PGfams), a standardized set of features across the PATRIC Database (4). (A) The number of metabolic PGfams for each genome are shown here, with the median value indicated by the middle line in the boxplot. (B) For the 144 strains from 16 species of *Lactobacillus*, we found that there are 116 protein families in the core set of metabolic PGfams, while the pan set of PGfams expands to over 1500 families. The nearly plateau shape of the curve for the pan set of PGfams curve indicates that this sampling represents a large portion of the genetic diversity among the 16 species included in the study. (C) This table shows the complete list of species used in this study and indicates the percentage of strains that were isolated from each human body site. Each strain in this study is a member from one of the 16 species and isolated from one of three human-associated body sites; intestinal, oral, or vaginal (Table S2).

Probabilistic Reconstruction Of constituent Anabolic Networks (PROTEAN)

We developed PROTEAN to predict the metabolic production capabilities of microbes based on genomic data alone. PROTEAN generates constituent metabolic production networks with maximum parsimony and probability to predict the production of a given metabolite with a defined set of input metabolites. PROTEAN is a combination of well-validated methods, including Parsimonious Enzyme Usage Flux Balance Analysis (pFBA) (37), likelihood-based gap filling (44), fastGapFill (45), and CarveMe (46). The algorithm uses the ModelSEED biochemical reaction database, a large set of known metabolic reactions, for constituent network generation (47). First, reaction likelihoods are calculated for each reaction in the ModelSEED database using Probannopy (48) (Figure 2). Reaction likelihoods correspond to the probability that a given reaction is catalyzed by an enzyme that is encoded for by the genome. We modified pFBA to utilize reaction likelihoods for weighted minimization of flux through each reaction, while still maintaining near-optimal flux through the objective function. Standard pFBA assumes that metabolism is optimized to minimize enzymatic turnover and thus the method is driven by a minimization of the total flux through the metabolic network (37). Weighted pFBA allows for the reconstruction of constituent anabolic networks while accounting for maximum genomic probability and resource parsimony (see Methods). The constituent anabolic networks output by PROTEAN consist of flux-carrying reactions required for the production of a certain metabolite with preferential flux through reactions that have higher reaction likelihoods. A constituent network represents a theoretically optimal biosynthetic network while accounting for the greatest genomic evidence for production of a given metabolite in a set media condition (Table S4). We represent the information from each constituent network using a single summary metric referred to as the Production Likelihood by calculating the average of all likelihoods of reactions that carry flux. The average of all reaction likelihoods in a metabolic pathway has been previously shown to be a valuable metric for making comparisons between networks (44).

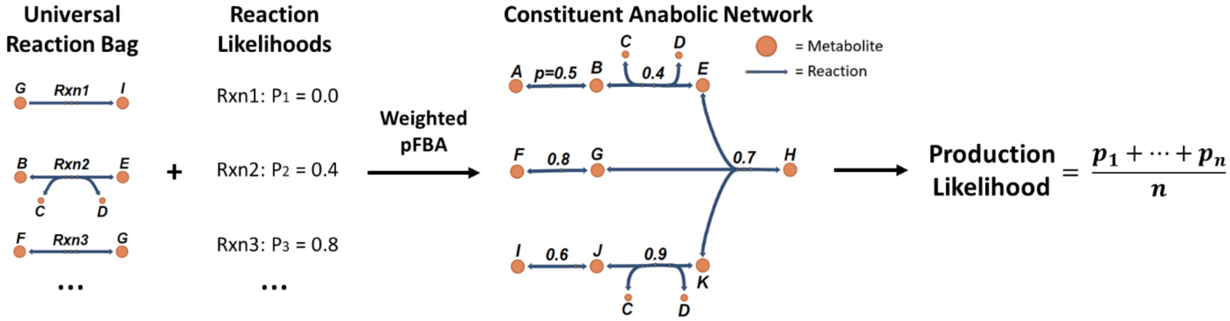


Figure 2: PROTEAN is an approach for quantifying the likelihood that a given metabolic network, derived exclusively from genomic evidence, is capable of synthesizing a particular metabolite. A modified version of Parsimonious Enzyme Usage FBA (weighted pFBA) was performed on a standardized set of reactions to generate constituent anabolic networks for each genome. Reaction likelihoods were used to weight the minimization of flux through each reaction in the network. Therefore, reactions with a greater likelihood were more likely to be included in the resulting constituent anabolic network. Each constituent network has a set of input metabolites representing the media condition (Table S4) and a demand reaction for a certain metabolic product. The resulting constituent network is the set of reactions that requires flux to produce the metabolic product in the given media condition. The production likelihood metric is an average of all the reaction likelihoods associated with the reactions included in the constituent network. This metric is used as a summary statistic that allows for the comparison of constituent networks across different metabolic products and strains, where a higher production likelihood corresponds with greater genetic evidence for that particular constituent anabolic network.

The Scaled Production Likelihood metric facilitates comparison of anabolic capabilities between species and strains

Predicted constituent anabolic networks were generated for a set of 50 biologically-relevant metabolic products for each of the 144 *Lactobacillus* strains. The 50 metabolites were selected based on known *Lactobacillus* biology (see Methods). For each metabolic product, we generated a constituent anabolic network (Table S3) across all strains. For each genome we scaled the Production Likelihoods metric by calculating the corresponding z-score. The standard deviation for the z-score calculation was across all metabolic products for each strain. This metric allows for a relative comparison of production capabilities across strains that does not rely on well-curated metabolic network reconstructions. The resulting Scaled Production Likelihood (SPL) is a metric indicating likelihood that a genome encodes for the cellular machinery required to produce a metabolite, given a specific media condition, relative to all of the other SPLs for the metabolic products per strain. For visualization, these data were grouped by species and summarized using the median of the SPLs across all of the strains within each species (Figure 3).

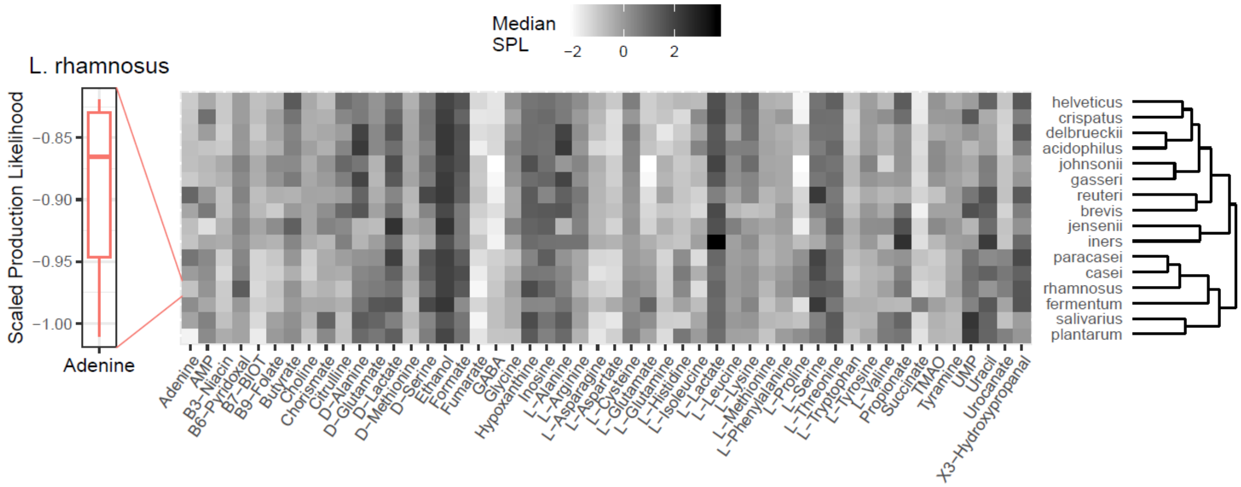


Figure 3: Predicted metabolic production capabilities with the Scaled Production Likelihood (SPL) metric align poorly with phylogeny. There is a single production likelihood for each genome associated with each metabolite. A median SPL can be calculated for a species that allows for more general comparisons across species, illustrated here by the distribution for one species (*L. rhamnosus*) and one metabolite (adenine). There are 50 metabolites used as features to allow for the comparison of predicted production capabilities across the lactobacilli analyzed.

The strains were grouped by species and clustered based on median SPLs. We found that across the 16 species, D- and L-lactate both have high median SPLs, as we would expect with lactobacilli. Additionally, fumarate and GABA have particularly low SPLs across all species. We were able to find several publications indicating GABA can be produced by select lactobacilli in specific environments (49,50). However, we were unable to find publications discussing the production of fumarate by lactobacilli. Additionally, we found that the dendrogram from clustering based on predicted metabolic production capabilities does not qualitatively align well with published phylogenetic trees generated using the 16S rRNA gene (34). The misalignment to established phylogenetic trees indicates that phylogeny is a poor indicator of metabolic production capabilities. It is likely that evolution of metabolic production capabilities is driven independently from classical genes used for phylogenetic comparisons, such as the 16S rRNA gene. Therefore, we need more precise computational tools to better understand the phenotypic differences between microbial species when interrogating metabolism. Perhaps phylogenetic analysis would be augmented with the consideration of metabolic genes in addition to the 16S rRNA gene.

Intestinal and oral *Lactobacillus* strains have different metabolic capabilities compared to vaginal strains

We performed principle coordinate analysis (PCoA) on the SPLs for each species and determined that the *Lactobacillus* strains cluster significantly by both species (Figure 4A) and isolation site (Figure 4B) (PERMANOVA; $P < 0.001$). The vaginal isolates differ from both the oral and gut cluster (Figure 4B). Substantial overlap was found between oral and gut isolates, specifically within *L. gasseri, L. rhamnosus,* and *L. salivarius*, likely due to the consistent transmission of orally colonized microbes to the intestines (15). It has been hypothesized that many of the lactobacilli isolated from the gut are actually transient strains that are colonized in the oral cavity (51). Our data supports this hypothesis by showing that oral isolates are metabolically similar to a portion of the intestinal isolates. However, there are lactobacilli, such as *L. reuteri*, which likely colonize the human intestines (52). Five of the 16 species in this study are only represented by strains isolated from the intestines; although this result is influenced by sampling bias in the PATRIC Database, it provides support that our data contains species that are only found in the intestines. The vaginal isolates cluster separately from the intestinal/oral isolates along the primary coordinate that accounts for 78% of the variation in these data. The vaginal microbiota is frequently dominated by several *Lactobacillus* species, such as *L. iners*, *L. crispatus*, and *L. jensenii* (53–55). This separation of vaginal isolates from intestinal/oral isolates indicates that these two main clusters have differences in their metabolic production capabilities. This result is to be expected because the intestinal/oral nutrient environment is drastically different from the vaginal environment and the dominant species appear to have metabolic capabilities that reflect this difference.

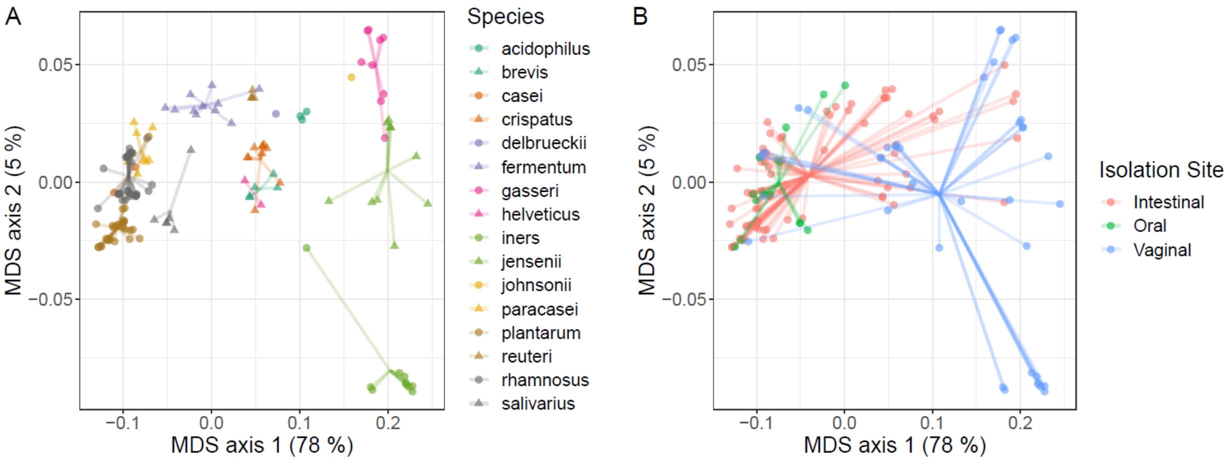


Figure 4: The Scaled Production Likelihood metric distinguishes metabolic functionality among species. (A) We found that *Lactobacillus* strains cluster significantly by species (PERMANOVA; P < 0.001). (B) Additionally, they cluster significantly by isolation site (PERMANOVA; P < 0.001). Both plots are PCoA using the Bray-Curtis distance metric of the SPLs for each isolate. Points in both panels are identical, but displayed with different color schemes.

In addition to distinguishing isolates by body site, the SPL metric is capable of defining collections of functional components that drive differences between groups. Using standard genomic analyses, differences between groups are typically defined by the differential gene content. Genes are intrinsically part of a larger network of metabolism where absence of specific functionality related to a gene may be compensated for within the system. Since our approach is based on Production Likelihoods of specific metabolites, it functions within a more complex metabolic framework compared to the analysis of genomic data without the network context. Using machine learning, we were able to identify the set of metabolites for which each group of strains is more likely to encode the cellular machinery required for production. We conducted a machine learning feature selection to determine the metabolites that are most likely to be produced by each group of strains, intestinal/oral strains and vaginal strains. We grouped the intestinal and oral strains together due to their inherent similarity (Figure 4B) and the observed transmission of oral strains to the intestines (15,51). We generated two separate area under the curve random forest (AUCRF) models to determine the metabolites that were more likely to be produced by each of the groups. Two models were necessary to enrich for the most discriminatory metabolites that were more likely to be produced in each of the groups, rather than simply identifying the metabolites that best classify the samples based on isolation site regardless of being more or less likely to be produced (See methods). The first model was generated to select the metabolites that are most likely to be produced by the intestinal and oral isolates compared to the vaginal isolates, while maximally discriminating the groups. The eight metabolites selected accurately classify greater than 90% of isolates to the correct group (Figure 5A). The second model was generated to select the metabolites that are most likely to be produced by the vaginal isolates compared to the intestinal and oral isolates, while maximally discriminating the groups. The seven metabolites selected accurately classify greater than 90% of the isolates to the correct group (Figure 5B).

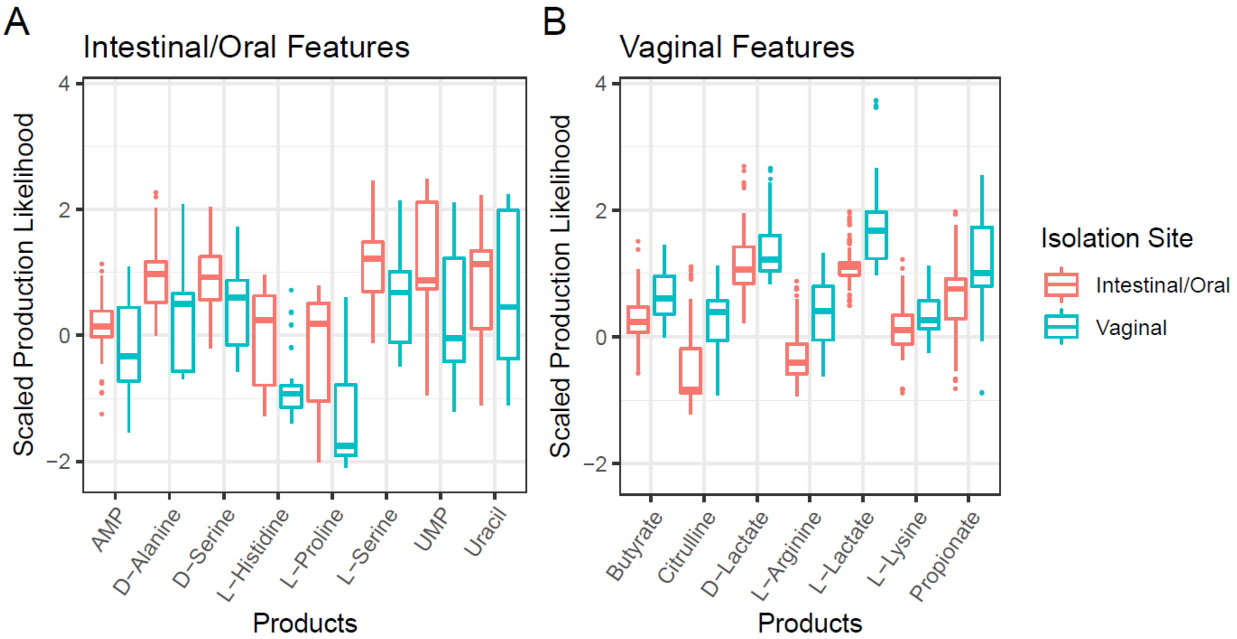


Figure 5: Machine learning of the SPL scores identifies metabolites that discriminate *Lactobacillus* strains. Machine learning feature selection identified the metabolites that are both most likely to be produced by each group and capable of classifying the strains into two groups, intestinal/oral and vaginal, with greater than 90% accuracy. (A) There are eight metabolites that are more likely to be produced by the intestinal/oral isolates compared to the vaginal isolates. (B) There are seven metabolites that are more likely to be produced by the vaginal isolates compared to intestinal/oral isolates. Both models are more than 90% accurate in predicting the membership to which the given isolate belongs using the SPLs of the metabolites listed.

Using SPLs as an input for AUCRF feature selection, we identified the metabolites that are most likely to be produced by the strains associated with the two isolate groups, intestinal/oral and vaginal. The selected metabolite products may contribute to how the strains interact with the mucosal tissues in each site. We hypothesize that these metabolites are related to key phenotypic differences between the two isolate groups. Four of the selected metabolites that are likely produced by intestinal/oral strains, D-alanine, D/L-serine, and L-proline (Figure 5A), have all been previously identified to have an impact on the human intestinal epithelium (23,24,56–58). Additionally, four of the selected metabolites that are likely produced by vaginal strains, L-arginine, citrulline, and D/L-lactate (Figure 5B), have been previously identified to have an impact on the human vaginal microbiome (59–62). The metabolites for which we have not found existing experimental evidence for are likely worth focusing on in future experimental studies.

For intestine-associated lactobacilli in this study, there is a connection between intestinal immune system regulation and D-alanine rich lipotechoic acid, a glycolipid expressed by some lactobacilli, such as *L. plantarum* (23,24). D-alanine rich lipotechoic acid, produced by lactobacilli*,* has been shown to down-regulate local colonic inflammation in a murine colitis model (23,24). With PROTEAN we identified that intestinal lactobacilli were more likely to produce D-alanine (Figure 5A). It is possible that a positive interaction with the intestinal host immune system would result in an evolutionary advantage by reducing local immune response. Additionally, serine rich serine-threonine peptides have been shown to have a similar regulatory effect on intestinal dendritic cells (56,57). These peptides expressed by *L. plantarum* are resistant to intestinal proteolysis and appear to be present in the colon of most healthy individuals (56,57). Similar to D-alanine, the production of D/L-serine would require a robust biosynthesis pathway present in those strains.

A final gut-related connection involves the biosynthesis of L-proline (Figure 5A). One of the primary stress responses in *L. acidophilus* to high osmotic pressure results in the accumulation of L-proline in the cell; there is little evidence that this response is a result of L-proline transport into the cell (58). These *Lactobacillus* strains are exposed to a large range of stressors in the gut, including suboptimal osmotic pressures. There is strong evidence that L-proline is used by *L. acidophilus* to tolerate suboptimal osmotic pressures and there is a lack of evidence for L-proline transporters. As such, the biosynthesis of L-proline may be advantageous for growth in the gut.

For the enriched metabolic products in vaginal isolates (Figure 5B), there is evidence for an arginine/ornithine antiporter and arginine deiminase in *L. fermentum* (59). These enzymes are part of the arginine deiminase pathway through which there is the production of citrulline which is exported from the cell and contributes to acid tolerance (59). It has also been demonstrated that treatment with probiotics containing arginine deiminase-positive lactobacilli can improve clinical symptoms of vaginosis in parallel with significant declines in polyamine (i.e. arginine, ornithine, and citrulline) levels in the vagina (60,61). The vaginal isolates in this study show enrichment for the cellular machinery required for the production of both citrulline and L-arginine (Figure 5B). The importance of lactate for the adequate maintenance of vaginal health in many individuals is known. The current hypothesis revolves around colonization resistance where vaginal lactobacilli establish an acidic environment by producing lactate (62). The acidic environment is generally inhospitable to invading pathogens as well as other microbes that are otherwise capable of residing in the vaginal environment (62). It has been shown that higher levels of D-lactate over L-lactate present in the vagina, produced by lactobacilli, further decrease the chance of infections in female patients (62). However, both isoforms of lactate remain important in maintaining vaginal health.

Conclusions

Microbial biosynthesis of metabolites has a broad range of applications, from bio-manufacturing to microbiome research (63). There is a wealth of well-curated and accessible knowledge stored in biochemical reaction databases such as ModelSEED (64). Genome-Scale Metabolic Network Reconstructions access this fundamental knowledge while accounting for systems-level interactions. This study represents one such application of GENREs that is a step toward predicting the metabolic production capabilities of understudied organisms. Experimental validation of the production capabilities predicted with PROTEAN will allow for conclusions to be made beyond the statement that a microbe is genetically likely to be able to produce a metabolite. Utilizing PROTEAN data, we found that human-associated lactobacilli strains cluster significantly by species and isolation site. Additionally, many of the metabolic products that drive the clustering of strains by the isolation sites have known physiological function and importance in the respective isolation sites.

Future applications of PROTEAN could include optimal strain selection for bio-manufacturing of a certain compound, generating predicted metabolomics data for an organism to generate a prioritized list of conditions that would be most worthwhile to validate experimentally, and predicting the metabolites that are most likely to be produced in a microbiota. Microbes can have a wide range of physiological impacts on human health; these impacts are, in part, a result of the metabolites that are or are not produced by members of a microbiota. One of the core limitations of this study includes the lack of reaction likelihoods for some reactions in the universal reaction bag we used from ModelSEED. The number of reactions we could generate likelihoods for was limited by the Probannopy reaction template. However, this template can be expanded to continue to improve the utility of PROTEAN. With the inclusion of validation data, additional analyses will be possible, such as determining metabolic production pathways lacking proper annotation. By determining the reactions that are most likely required for biosynthesis of a known product, it would be possible to generate additional hypotheses for enzyme annotation experiments. PROTEAN is an algorithm with potential for a wide range of applications in the study and use of microbial metabolic networks.

Methods

Constituent Anabolic Network Generation (PROTEAN)

Probabilistic pFBA-based constituent anabolic network generation was accomplished using three Python packages, Cobrapy (65), Mackinac (66), and Probannopy (48). The complete ModelSEED universal reaction bag was downloaded from the github repository and filtered based on the annotation quality score, including all reactions with an ‘OK’ quality status or better (64). For each reaction in the ModelSEED universal reaction bag, we used Probannopy to generate a reaction likelihood based on the FASTA file for each genome obtained from the PATRIC database (4). The Cobrapy implementation of Parsimonious Enzyme Usage Flux Balance Analysis (pFBA) was altered to allow for each reaction’s linear constraint to be set individually based on the reaction likelihood. The linear constraint for each reaction was set to one minus the reaction likelihood (a value between 0 and 1). There were reactions included in the universal reaction bag that were lacking from the Probannopy template model, therefore resulting in several gene-associated reactions lacking reaction likelihood scores. The reactions without likelihoods were left at a full minimization penalty (linear constraint value of 1). We chose to penalize the reactions without likelihoods to bias our results towards the construction of networks for which all reactions had evidence of presence. The linear constraints applied to each reaction based on likelihood acted as a weighting (inclusion penalty) for the minimization step in pFBA, resulting in the reactions with greater likelihood having a lower penalty for carrying flux; therefore, the reactions had a higher likelihood of being included in the constituent anabolic networks.

Using PROTEAN, we generated constituent anabolic networks by setting a certain input media condition (Table S4) and constraining flux through the single metabolite objective function (Table S3). We ran our likelihood-weighted pFBA flux minimization across the entire universal reaction bag and isolated the reactions that carried flux to get the desired product. The resulting networks consist of the direct reactions that would be part of a production pathway as might be shown in a typical biosynthesis pathway figure, while also accounting for all of the secondary and energy metabolites that are required for the production of the metabolite in consideration. Additionally, this algorithm is optimizing for three core characteristics in the constituent networks: 1) minimum flux through the network (loosely, the minimum number of reactions), 2) maximum average reaction likelihood across the constituent network, and 3) output flux within 90% of the optimal yield of the metabolic product. We chose to allow flux through any reaction in the universal reaction bag during the generation of the constituent anabolic production pathways rather than simply pulling from a GENRE that was first gapfilled to allow production of biomass. Using the universal reaction bag instead of a gapfilled model was important because the biomass function is difficult to define for understudied organisms and unnecessary for our applications.

Scaled Production Likelihood Metric

We represent the information from each constituent network using a single summary metric for ease of comparison, simply named the Production Likelihood. This metric is the average of the reaction likelihoods included in the constituent network. The average reaction likelihood for a metabolic pathway has been previously used for making comparisons between networks (44). The Production Likelihoods for all 50 metabolites are scaled for each given genome by calculating the z-score to create the Scaled Production Likelihoods used for the majority of the analysis in this study. The z-score is calculated for each individual strain using the median and standard deviation for the production likelihoods across the 50 metabolic products. The Scaled Production Likelihood allows for a ranked comparison of metabolic products across the genome set and corrects for annotation bias by essentially comparing the ranked z-score for each metabolic product.

Supporting data for pathway generation

The simulated media formulation was based on *in vitro* minimal media growth conditions for *L. plantarum* (Table S4) (67–69). The techniques used in this study do not assume that all species are capable of growth in the given media condition, therefore this media condition simply provides a standard reference for comparison. The product list was developed by identifying metabolites that have been shown to be produced by lactobacilli during *in vitro* growth experiments, in addition to other metabolites that have been shown to be related to human physiology (70–74).

Machine learning feature selection

Discriminating intestinal/oral and vaginal features were selected using area under the ROC curve random forest (AUCRF) using default parameters (75) (see Code). We generated two separate AUCRF models to determine the metabolites that were more likely to be produced by each of the groups, intestinal/oral and vaginal. Two models allowed us to enrich for likely products rather than simply selecting for the metabolites that provide the greatest discrimination between the groups but which may have poor likelihood scores. We conducted the enrichment for likely metabolic products for each model by reducing the feature set down to only metabolites that were more likely to be produced by the group of interest. Likely metabolic products were determined by comparing the median SPLs of each metabolite between the groups. Additionally, the feature sets were reduced to include only metabolites with a median value greater than zero for the group of interest. An AUCRF model was then generated to select the features that provided the greatest discrimination between the two groups.

Statistical modeling and figure generation

The principle coordinate analysis (PCoA) ordinations were created using the R vegan package (76), implemented with the Bray-Curtis dissimilarity metric. Statistical significance for comparing the PCoA clusters was determined using a PERMANOVA (R Adonis test). A variety of R packages were used for all figure generation (77–81).

Genome Quality and PATRIC Cross Genus Protein Family Data

Genomes used in the study were filtered for quality before being included in the analysis. Strains with greater than 0.2% unknown nucleotide calls in the genome were eliminated. Low quality genome assemblies with greater than 300 contigs were removed. Non-human associated *Lactobacillus* strains from the PATRIC database were used to determine the GC content range for each species (82,83), and significant outliers (plus or minus two percent) were removed to control for sequencing bias (84,85). Only isolates from the three human-associated sites (oral, intestinal, and vaginal) were included in the final dataset.

The inclusion of metabolic PATRIC cross genus protein families was conducted by filtering the PGfams for each genome based on the existence of an associated known reaction and Probannopy likelihood greater than 0. Pan and core metabolic PGfam sets were evaluated after the addition of all genomic features from each genome. The pan set of metabolic PGfams was defined as the total number of unique PGfams included in the data set after the above filtering steps. The core set of metabolic PGfams are those that existed within each genome included in this study.

Data and code availability

Genome FASTA files and metadata were downloaded from the PATRIC Database (4). Python and R code is available at: Github.com/Tjmoutinho/Lactobacillus

References

1. de Vos WM. Systems solutions by lactic acid bacteria: from paradigms to practice. Microb Cell Factories. 2011 Aug 30;10(1):S2.

2. de Vos WM, Hugenholtz J. Engineering metabolic highways in Lactococci and other lactic acid bacteria. Trends Biotechnol. 2004 Feb 1;22(2):72–9.

3. Ljungh Å, Wadström T. Lactobacillus Molecular Biology: From Genomics to Probiotics. Horizon Scientific Press; 2009. 217 p.

4. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014 Jan 1;42(D1):D581–91.

5. OHanlon DE. In vivo versus in vitro metabolomics profiling of vaginal lactobacilli for probiotic use. 2013 Jun 4 [cited 2018 Sep 24]; Available from: https://www.omicsonline.org/proceedings/in-vivo-versus-in-vitro-metabolomics-profiling-of-vaginal-lactobacilli-for-probiotic-use-785.html

6. O'Hanlon DE, Moench TR, Cone RA. Vaginal pH and Microbicidal Lactic Acid When Lactobacilli Dominate the Microbiota. PLoS ONE [Internet]. 2013 Nov 6 [cited 2018 Sep 24];8(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3819307/

7. Tachedjian G, Aldunate M, Bradshaw CS, Cone RA. The role of lactic acid production by probiotic Lactobacillus species in vaginal health. Res Microbiol. 2017 Nov 1;168(9):782–92.

8. Tachedjian G, O'Hanlon DE, Ravel J. The implausible "in vivo" role of hydrogen peroxide as an antimicrobial factor produced by vaginal microbiota. Microbiome [Internet]. 2018 Feb 6 [cited 2018 Sep 24];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5801833/

9. Parolin C, Foschi C, Laghi L, Zhu C, Banzola N, Gaspari V, et al. Insights Into Vaginal Bacterial Communities and Metabolic Profiles of Chlamydia trachomatis Infection: Positioning Between Eubiosis and Dysbiosis. Front Microbiol [Internet]. 2018 Mar 28 [cited 2018 Sep 24];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5883401/

10. Vitali B, Cruciani F, Picone G, Parolin C, Donders G, Laghi L. Vaginal microbiome and metabolome highlight specific signatures of bacterial vaginosis. Eur J Clin Microbiol Infect Dis. 2015 Dec 1;34(12):2367–76.

11. Gosmann C, Anahtar MN, Handley SA, Farcasanu M, Abu-Ali G, Bowman BA, et al. Lactobacillus-Deficient Cervicovaginal Bacterial Communities Are Associated with Increased HIV Acquisition in Young South African Women. Immunity. 2017 Jan 17;46(1):29–37.

12. Ratzke C, Gore J. Modifying and reacting to the environmental pH can drive bacterial interactions. PLOS Biol. 2018 Mar 14;16(3):e2004248.

13. Palmer RJ. Composition and development of oral bacterial communities. Periodontol 2000 [Internet]. 2014 Feb [cited 2018 Sep 24];64(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3876289/

14. Tannock GW. A Special Fondness for Lactobacilli. Appl Environ Microbiol. 2004 Jun;70(6):3189–94.

15. Schmidt TSB, Hayward MR, Coelho LP, Li SS, Costea PI, Voigt AY, et al. Extensive transmission of microbes along the gastrointestinal tract. Nieuwdorp M, editor. eLife. 2019 Feb 12;8:e42693.

16. Szajewska H, Ruszczyński M, Radzikowski A. Probiotics in the prevention of antibiotic-associated diarrhea in children: A meta-analysis of randomized controlled trials. J Pediatr. 2006 Sep 1;149(3):367-372.e1.

17. Hempel S, Newberry SJ, Maher AR, Wang Z, Miles JNV, Shanman R, et al. Probiotics for the Prevention and Treatment of Antibiotic-Associated Diarrhea: A Systematic Review and Meta-analysis. JAMA. 2012 May 9;307(18):1959–69.

18. Ford AC, Quigley EMM, Lacy BE, Lembo AJ, Saito YA, Schiller LR, et al. Efficacy of Prebiotics, Probiotics, and Synbiotics in Irritable Bowel Syndrome and Chronic Idiopathic Constipation: Systematic Review and Meta-analysis. Am J Gastroenterol. 2014 Oct;109(10):1547–61.

19. Nikfar S, Rahimi R, Rahimi F, Derakhshani S, Abdollahi M. Efficacy of Probiotics in Irritable Bowel Syndrome: A Meta-Analysis of Randomized, Controlled Trials. Dis Colon Rectum. 2008 Dec 1;51(12):1775–80.

20. Elazab N, Mendy A, Gasana J, Vieira ER, Quizon A, Forno E. Probiotic Administration in Early Life, Atopy, and Asthma: A Meta-analysis of Clinical Trials. Pediatrics. 2013 Sep 1;132(3):e666–76.

21. Berstad A, Raa J, Midtvedt T, Valeur J. Probiotic lactic acid bacteria – the fledgling cuckoos of the gut? Microb Ecol Health Dis [Internet]. 2016 May 26 [cited 2018 Sep 24];27. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4884264/

22. Suez J, Zmora N, Zilberman-Schapira G, Mor U, Dori-Bachash M, Bashiardes S, et al. Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT. Cell. 2018 Sep 6;174(6):1406-1423.e16.

23. de Vos WM. Lipotechoic acid in lactobacilli: D-alanine makes the difference. Proc Natl Acad Sci. 2005;102(31):10763–4.

24. Grangette C, Nutten S, Palumbo E, Morath S, Hermann C, Dewulf J, et al. Enhanced antiinflammatory capacity of a Lactobacillus plantarum mutant synthesizing modified teichoic acids. Proc Natl Acad Sci. 2005;102(29):10321–6.

25. Branco dos Santos F, de Vos WM, Teusink B. Towards metagenome-scale models for industrial applications—the case of Lactic Acid Bacteria. Curr Opin Biotechnol. 2013 Apr 1;24(2):200–6.

26. Le Barz M, Anhê FF, Varin TV, Desjardins Y, Levy E, Roy D, et al. Probiotics as Complementary Treatment for Metabolic Disorders. Diabetes Metab J. 2015 Aug;39(4):291–303.

27. Saulnier DM, Santos F, Roos S, Mistretta T-A, Spinler JK, Molenaar D, et al. Exploring Metabolic Pathway Reconstruction and Genome-Wide Expression Profiling in Lactobacillus reuteri to Define Functional Probiotic Features. PLOS ONE. 2011 Apr 29;6(4):e18783.

28. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol. 2012 Apr;10(4):291–305.

29. Haggart CR, Bartell JA, Saucerman JJ, Papin JA. Whole-genome metabolic network reconstruction and constraint-based modeling. Methods Enzymol. 2011;500:411–33.

30. Kant R, Blom J, Palva A, Siezen RJ, de Vos WM. Comparative genomics of Lactobacillus. Microb Biotechnol. 2011 May;4(3):323–32.

31. Drissi F, Merhej V, Angelakis E, El Kaoutari A, Carrière F, Henrissat B, et al. Comparative genomics analysis of Lactobacillus species associated with weight gain or weight protection. Nutr Diabetes. 2014 Feb;4(2):e109.

32. France MT, Mendes-Soares H, Forney LJ. Genomic Comparisons of Lactobacillus crispatus and Lactobacillus iners Reveal Potential Ecological Drivers of Community Composition in the Vagina. Appl Env Microbiol. 2016 Dec 15;82(24):7063–73.

33. Morita H, Toh H, Fukuda S, Horikawa H, Oshima K, Suzuki T, et al. Comparative Genome Analysis of Lactobacillus reuteri and Lactobacillus fermentum Reveal a Genomic Island for Reuterin and Cobalamin Production. DNA Res. 2008 Jun 1;15(3):151–61.

34. Zhang Z-G, Ye Z-Q, Yu L, Shi P. Phylogenomic reconstruction of lactic acid bacteria: an update. BMC Evol Biol. 2011 Jan 1;11:1.

35. Kleerebezem M, Vos WM de. Lactic acid bacteria: life after genomics. Microb Biotechnol. 2011 May 1;4(3):318–22.

36. Rau MH, Zeidan AA. Constraint-based modeling in microbial food biotechnology. Biochem Soc Trans. 2018 Mar 27;BST20170268.

37. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010 Jul 27;6:390.

38. Feist AM, Palsson BO. The biomass objective function. Curr Opin Microbiol. 2010 Jun;13(3):344–9.

39. Altafini C, Facchetti G. Metabolic Adaptation Processes That Converge to Optimal Biomass Flux Distributions. PLoS Comput Biol. 2015 Sep 4;11(9):e1004434.

40. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010 Mar;28(3):245–8.

41. Pinto F, Medina DA, Pérez-Correa JR, Garrido D. Modeling metabolic interactions in a consortium of the infant gut microbiome. Front Microbiol. 2017;8:2507.

42. Schmidt BJ, Ebrahim A, Metz TO, Adkins JN, Palsson BØ, Hyduke DR. GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics. 2013 Nov 15;29(22):2900–8.

43. Li H, Zhu J. Targeted metabolic profiling rapidly differentiates Escherichia coli and Staphylococcus aureus at species and strain level. Rapid Commun Mass Spectrom. 2017;31(19):1669–76.

44. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND. Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models. PLOS Comput Biol. 2014 Oct 16;10(10):e1003882.

45. Thiele I, Vlassis N, Fleming RMT. fastGapFill: efficient gap filling in metabolic networks. Bioinformatics. 2014 Sep 1;30(17):2529–31.

46. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 2018 Sep 6;46(15):7542–53.

47. Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AaronA, Henry C. Automated Genome Annotation and Metabolic Model Reconstruction in the SEED and Model SEED. In: Alper HS, editor. Systems Metabolic Engineering [Internet]. Humana Press; 2013 [cited 2017 Apr 6]. p. 17–45. (Methods in Molecular Biology). Available from: http://dx.doi.org/10.1007/978-1-62703-299-5_2

48. King B, Farrah T, Richards MA, Mundy M, Simeonidis E, Price ND. ProbAnnoWeb and ProbAnnoPy: probabilistic annotation and gap-filling of metabolic reconstructions. Bioinformatics. 2018 May 1;34(9):1594–6.

49. Komatsuzaki N, Shima J, Kawamoto S, Momose H, Kimura T. Production of γ-aminobutyric acid (GABA) by Lactobacillus paracasei isolated from traditional fermented foods. Food Microbiol. 2005;22(6):497–504.

50. Li H, Qiu T, Huang G, Cao Y. Production of gamma-aminobutyric acid by Lactobacillus brevis NCL912 using fed-batch fermentation. Microb Cell Factories. 2010;9(1):85.

51. Walter J. Ecological Role of Lactobacilli in the Gastrointestinal Tract: Implications for Fundamental and Biomedical Research. Appl Env Microbiol. 2008 Aug 15;74(16):4985–96.

52. Valeur N, Engel P, Carbajal N, Connolly E, Ladefoged K. Colonization and immunomodulation by Lactobacillus reuteri ATCC 55730 in the human gastrointestinal tract. Appl Environ Microbiol. 2004;70(2):1176–81.

53. Romero R, Hassan SS, Gajer P, Tarca AL, Fadrosh DW, Nikita L, et al. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. Microbiome. 2014 Feb 3;2(1):4.

54. Gajer P, Brotman RM, Bai G, Sakamoto J, Schütte UME, Zhong X, et al. Temporal Dynamics of the Human Vaginal Microbiota. Sci Transl Med. 2012 May 2;4(132):132ra52-132ra52.

55. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci. 2011 Mar 15;108(Supplement 1):4680–7.

56. Al-Hassi HO, Mann ER, Sanchez B, English NR, Peake STC, Landy J, et al. Altered human gut dendritic cell properties in ulcerative colitis are reversed by Lactobacillus plantarum extracellular encrypted peptide STp. Mol Nutr Food Res. 2014;58(5):1132–43.

57. Bernardo D, Sánchez B, Al-Hassi HO, Mann ER, Urdaci MC, Knight SC, et al. Microbiota/Host Crosstalk Biomarkers: Regulatory Response of Human Intestinal Dendritic Cells Exposed to Lactobacillus Extracellular Encrypted Peptide. PLOS ONE. 2012 May 14;7(5):e36262.

58. Jewell JB, Kashket ER. Osmotically regulated transport of proline by Lactobacillus acidophilus IFO 3532. Appl Env Microbiol. 1991 Oct 1;57(10):2829–33.

59. Vrancken G, Rimaux T, Weckx S, De Vuyst L, Leroy F. Environmental pH determines citrulline and ornithine release through the arginine deiminase pathway in Lactobacillus fermentum IMDO 130101. Int J Food Microbiol. 2009 Nov 15;135(3):216–22.

60. Famularo G, Pieluigi M, Coccia R, Mastroiacovo P, Simone CD. Microecology, bacterial vaginosis and probiotics: perspectives for bacteriotherapy. Med Hypotheses. 2001 Apr 1;56(4):421–30.

61. Rousseau V, Lepargneur JP, Roques C, Remaud-Simeon M, Paul F. Prebiotic effects of oligosaccharides on selected vaginal lactobacilli and pathogenic microorganisms. Anaerobe. 2005 Jun 1;11(3):145–53.

62. Witkin SS, Mendes-Soares H, Linhares IM, Jayaram A, Ledger WJ, Forney LJ. Influence of Vaginal Bacteria and d- and l-Lactic Acid Isomers on Vaginal Extracellular Matrix Metalloproteinase Inducer:

Implications for Protection against Upper Genital Tract Infections. mBio. 2013 Aug 30;4(4):e00460-13.

63. LeBlanc JG, Milani C, de Giori GS, Sesma F, van Sinderen D, Ventura M. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. Curr Opin Biotechnol. 2013 Apr 1;24(2):160–8.

64. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 2010 Sep;28(9):977–82.

65. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol. 2013 Aug 8;7(1):74.

66. Mundy M, Mendes-Soares H, Chia N. Mackinac: a bridge between ModelSEED and COBRApy to generate and analyze genome-scale metabolic models. Bioinformatics. 2017 Aug 1;33(15):2416–8.

67. Wegkamp A, Teusink B, De Vos W m., Smid E j. Development of a minimal growth medium for Lactobacillus plantarum. Lett Appl Microbiol. 2010 Jan 1;50(1):57–64.

68. Ricciardi A, Ianniello RG, Parente E, Zotta T. Modified chemically defined medium for enhanced respiratory growth of Lactobacillus casei and Lactobacillus plantarum groups. J Appl Microbiol. 2015 Sep 1;119(3):776–85.

69. Elli M, Zink R, Rytz A, Reniero R, Morelli L. Iron requirement of Lactobacillus spp. in completely chemically defined growth media. J Appl Microbiol. 2000 Apr 1;88(4):695–703.

70. Rowland I, Gibson G, Heinken A, Scott K, Swann J, Thiele I, et al. Gut microbiota functions: metabolism of nutrients and other food components. Eur J Nutr. 2018 Feb 1;57(1):1–24.

71. Neis EPJG, Dejong CHC, Rensen SS. The Role of Microbial Amino Acid Metabolism in Host Metabolism. Nutrients. 2015 Apr 16;7(4):2930–46.

72. Wu G. Intestinal Mucosal Amino Acid Catabolism. J Nutr. 1998 Aug 1;128(8):1249–52.

73. Rooj AK, Kimura Y, Buddington RK. Metabolites produced by probiotic Lactobacilli rapidly increase glucose uptake by Caco-2 cells. BMC Microbiol. 2010 Jan 20;10(1):16.

74. Belzer C, Chia LW, Aalvink S, Chamlagain B, Piironen V, Knol J, et al. Microbial Metabolic Networks at the Mucus Layer Lead to Diet-Independent Butyrate and Vitamin B12 Production by Intestinal Symbionts. mBio. 2017 Nov 8;8(5):e00770-17.

75. Urrea V, Calle M. AUCRF: variable selection with random forest and the area under the curve. R Package Version 11. 2012;

76. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'hara R, et al. vegan: Community ecology package. R Package Version. 2011;117–8.

77. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J Comput Graph Stat. 1996 Sep 1;5(3):299–314.

78. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer; 2016. 266 p.

79. Wickham H. tidyr: Easily Tidy Data with spread () and gather () Functions. Version 06 0. 2016;

80. Wickham H, Francois R, Henry L, Müller K. dplyr: A grammar of data manipulation. R Package Version 04. 2015;3.

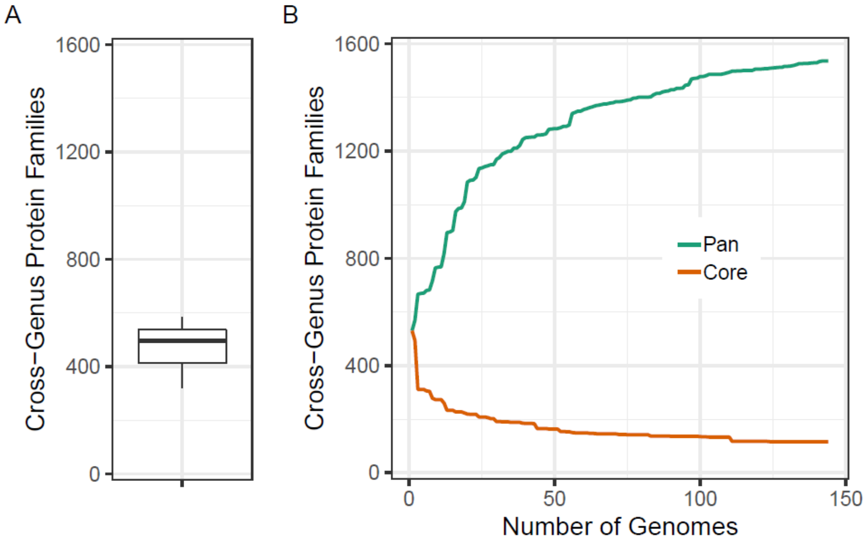
81. Neuwirth E, Brewer RC. ColorBrewer palettes. R Package Version. 2014;1–1.

82. Haywood-Farmer E, Otto SP. The Evolution of Genomic Base Composition in Bacteria. Evolution. 2003;57(8):1783–92.

83. Bentley SD, Parkhill J. Comparative Genomic Structure of Prokaryotes. Annu Rev Genet. 2004;38(1):771–91.

84. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012 May;40(10):e72.

85. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.



C

Species	Intestinal	Oral	Vaginal
acidophilus	100%	0%	0%
brevis	100%	0%	0%
casei	100%	0%	0%
crispatus	8%	8%	83%
delbrueckii	100%	0%	0%
fermentum	69%	15%	15%
gasseri	50%	0%	50%
helveticus	67%	0%	33%
iners	0%	0%	100%
jensenii	10%	0%	90%
johnsonii	0%	0%	100%
paracasei	62%	38%	0%
plantarum	89%	7%	4%
reuteri	100%	0%	0%
rhamnosus	69%	14%	17%
salivarius	50%	50%	0%

Figure 1: Known metabolic annotations are extensively sampled across the 16 *Lactobacillus* species included in this study. The genomic features used for this analysis are PATRIC Cross-Genera Protein families (PGfams), a standardized set of features across the PATRIC Database (4). (A) The number of metabolic PGfams for each genome are shown here, with the median value indicated by the middle line in the boxplot. (B) For the 144 strains from 16 species of *Lactobacillus*, we found that there are 116 protein families in the core set of metabolic PGfams, while the pan set of PGfams expands to over 1500 families. The nearly plateau shape of the curve for the pan set of PGfams curve indicates that this sampling represents a large portion of the genetic diversity among the 16 species included in the study. (C) This table shows the complete list of species used in this study and indicates the percentage of strains that were isolated from each human body site. Each strain in this study is a member from one of the 16 species and isolated from one of three human-associated body sites; intestinal, oral, or vaginal (Table S2).

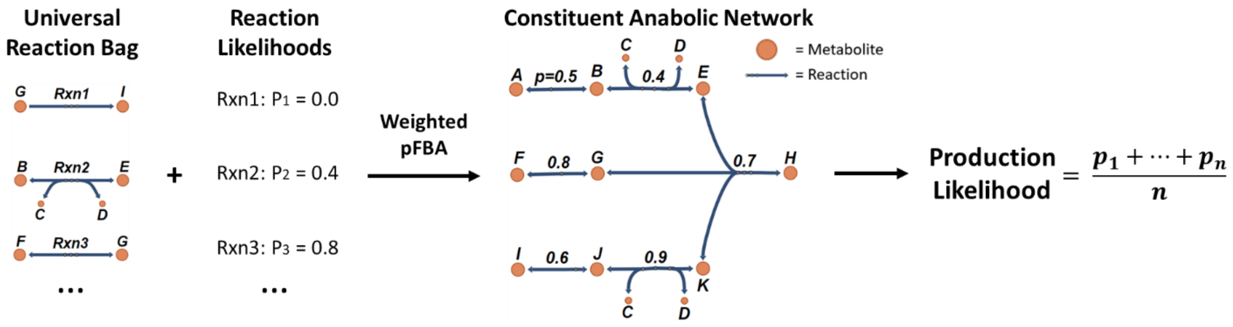


Figure 2: PROTEAN is an approach for quantifying the likelihood that a given metabolic network, derived exclusively from genomic evidence, is capable of synthesizing a particular metabolite. A modified version of Parsimonious Enzyme Usage FBA (weighted pFBA) was performed on a standardized set of reactions to generate constituent anabolic networks for each genome. Reaction likelihoods were used to weight the minimization of flux through each reaction in the network. Therefore, reactions with a greater likelihood were more likely to be included in the resulting constituent anabolic network. Each constituent network has a set of input metabolites representing the media condition (Table S4) and a demand reaction for a certain metabolic product. The resulting constituent network is the set of reactions that requires flux to produce the metabolic product in the given media condition. The production likelihood metric is an average of all the reaction likelihoods associated with the reactions included in the constituent network. This metric is used as a summary statistic that allows for the

comparison of constituent networks across different metabolic products and strains, where a higher production likelihood corresponds with greater genetic evidence for that particular constituent anabolic network.

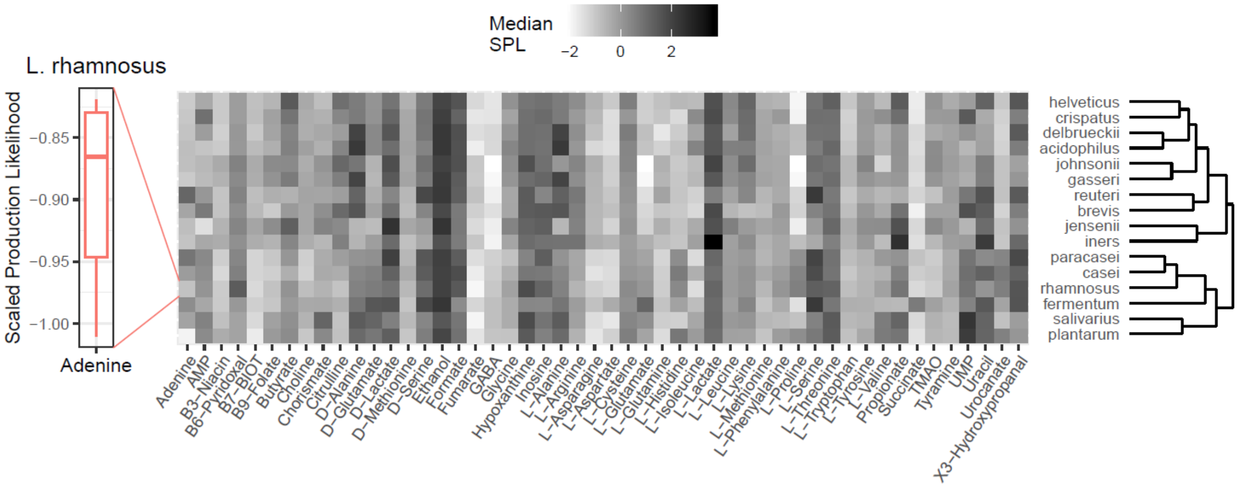


Figure 3: Predicted metabolic production capabilities with the Scaled Production Likelihood (SPL) metric align poorly with phylogeny. There is a single production likelihood for each genome associated with each metabolite. A median SPL can be calculated for a species that allows for more general comparisons across species, illustrated here by the distribution for one species (*L. rhamnosus*) and one metabolite (adenine). There are 50 metabolites used as features to allow for the comparison of predicted production capabilities across the lactobacilli analyzed.

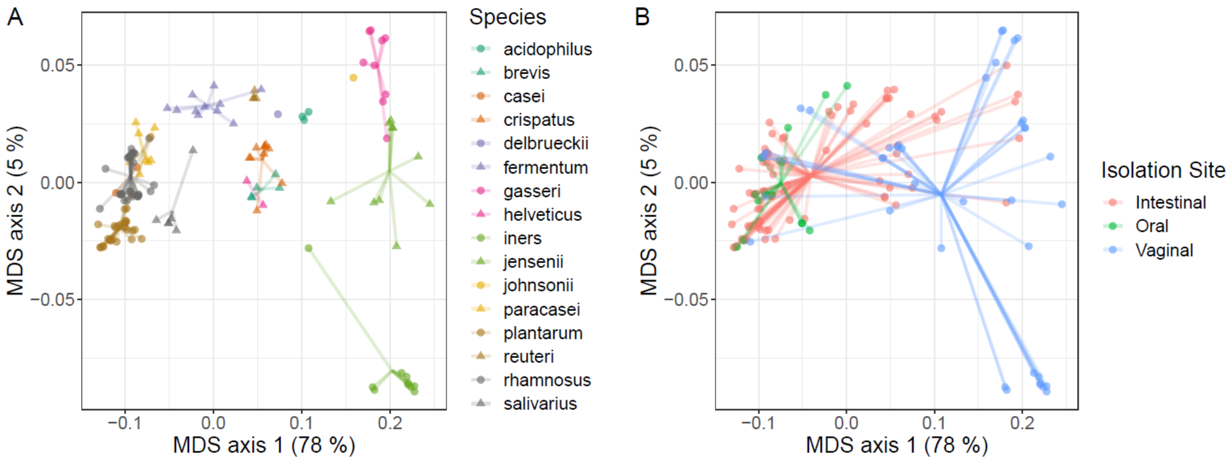


Figure 4: The Scaled Production Likelihood metric distinguishes metabolic functionality among species. (A) We found that *Lactobacillus* strains cluster significantly by species (PERMANOVA; P < 0.001). (B) Additionally, they cluster significantly by isolation site (PERMANOVA; P < 0.001). Both plots are PCoA using the Bray-Curtis distance metric of the SPLs for each isolate. Points in both panels are identical, but displayed with different color schemes.

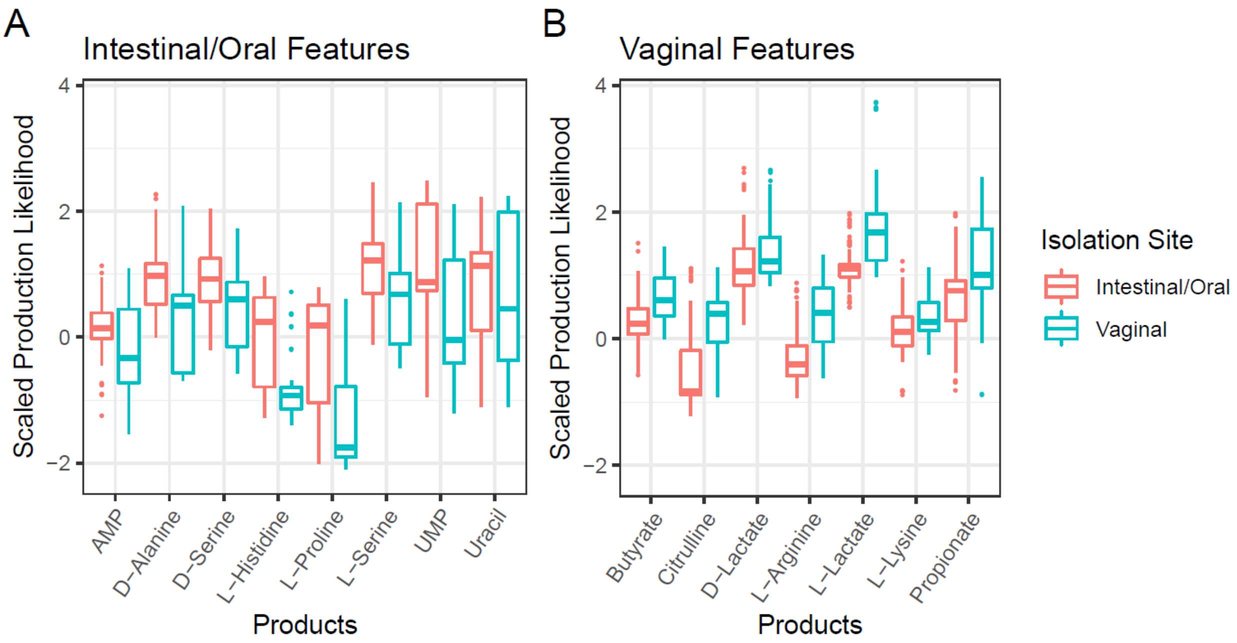


Figure 5: Machine learning of the SPL scores identifies metabolites that discriminate *Lactobacillus* strains. Machine learning feature selection identified the metabolites that are both most likely to be produced by each group and capable of classifying the strains into two groups, intestinal/oral and vaginal, with greater than 90% accuracy. (A) There are eight metabolites that are more likely to be produced by the intestinal/oral isolates compared to the vaginal isolates. (B) There are seven metabolites that are more likely to be produced by the vaginal isolates compared to intestinal/oral isolates. Both models are more than 90% accurate in predicting the membership to which the given isolate belongs using the SPLs of the metabolites listed.