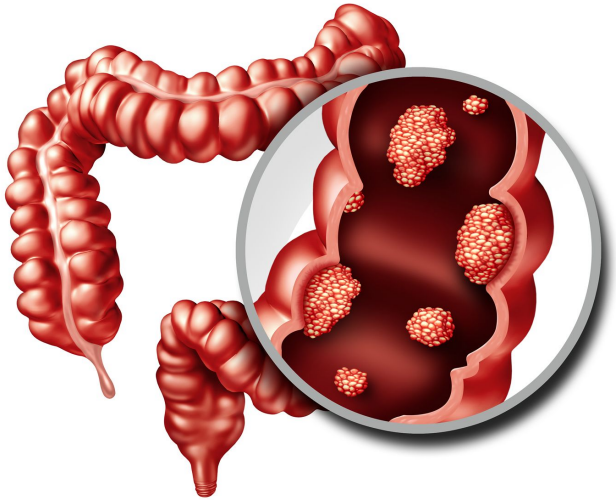
Functional Activity of the Human Gut Microbiome to Classify Colorectal Cancer

Kelly Sovacool

Aug. 2020

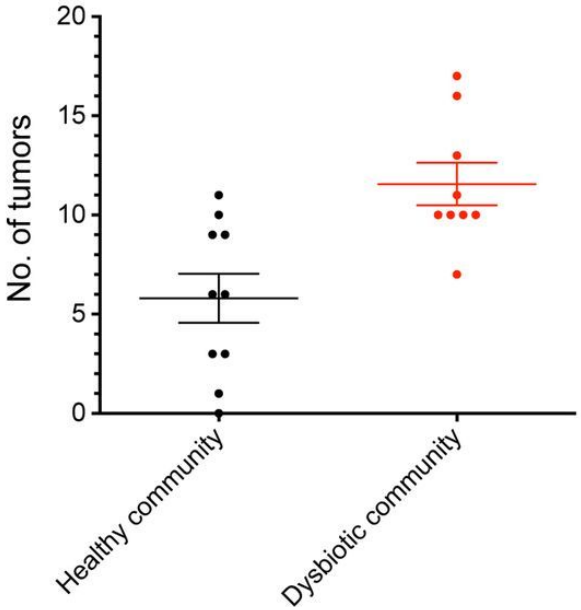
Colorectal cancer

- Colorectal cancer (CRC) is responsible for the second-most cancer deaths after lung cancer.
- CRC can be caught early with colonoscopy, but patient compliance is low due to invasiveness and cost.
- Fecal immunochemical test (FIT) is less invasive, but also less sensitive than colonoscopy.
- There is a need for a sensitive and non-invasive diagnostic test.



The gut microbiome changes in CRC

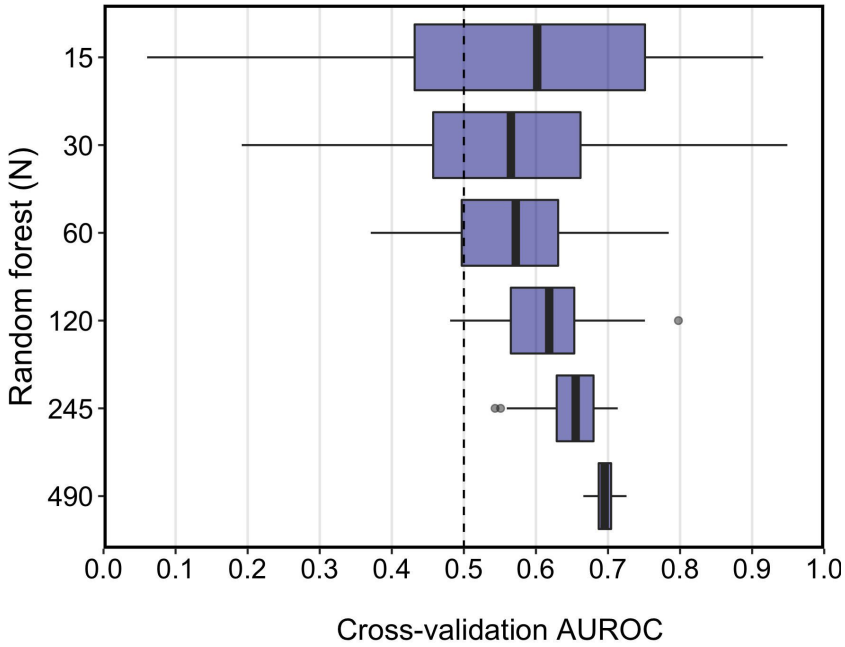
- Fecal matter transplants from CRC mice increase tumor formation in germ-free mice.
- Changes in the taxonomic composition of gut microbiomes have been observed in CRC.
  - e.g. increased *Fusobacterium* in some CRC datasets
- However, changes are not consistent across all CRC samples or datasets.



Zackular et al 2013 mBio

Taxonomic composition for CRC classification

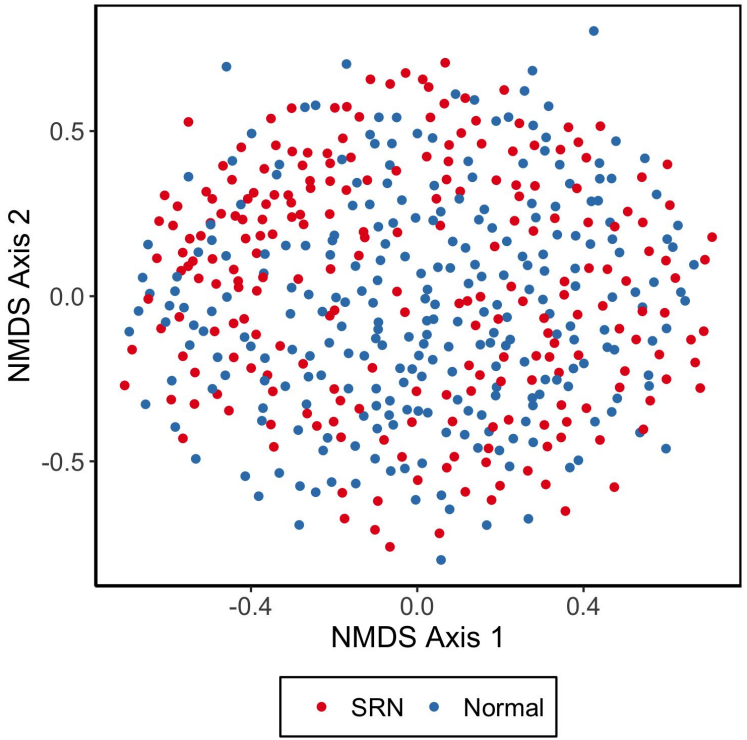
- Taxonomic composition is often characterized by clustering 16S rRNA gene sequences into Operational Taxonomic Units (OTUs).
- OTU-based machine learning models have modest performance on classifying stool samples as healthy, adenomatous, or cancerous.



Topçuoğlu et al. 2020 mBio

Taxonomic changes are inconsistent

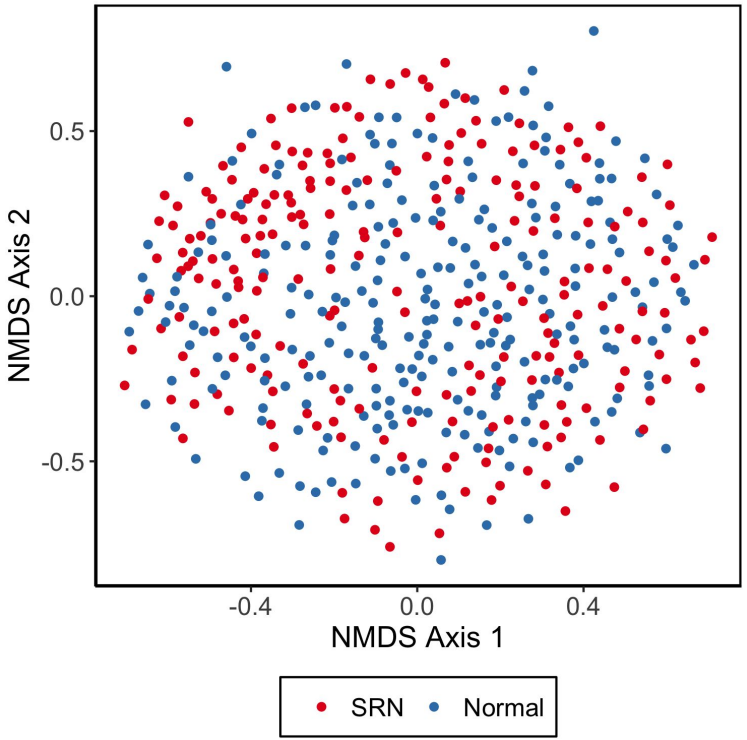
- Microbiome changes in disease are inconsistent because there is high interpersonal variability in microbiome composition.



Topçuoğlu et al. 2020 mBio

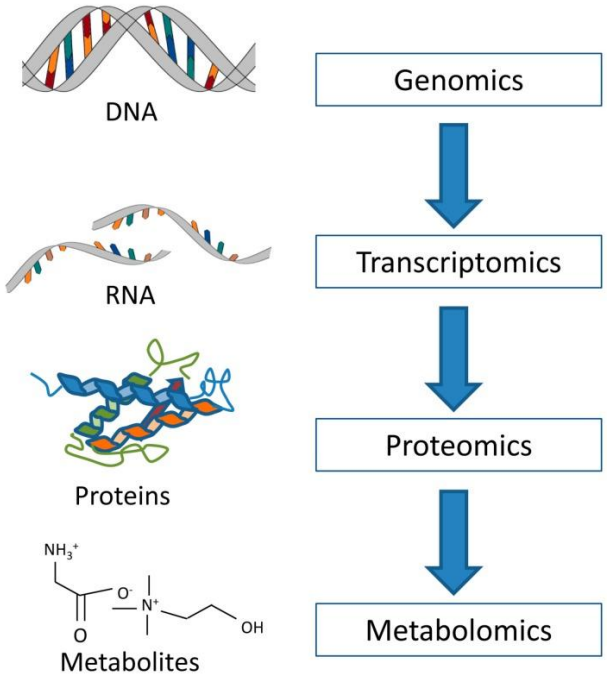
Taxonomic changes are inconsistent

- Microbiome changes in disease are inconsistent because there is high interpersonal variability in microbiome composition.
- Possible explanation: **functional redundancy**, where different species can perform the same function.
- Thus, communities with **different taxonomic composition** can have the **same functional composition**.



Topçuoğlu et al. 2020 mBio

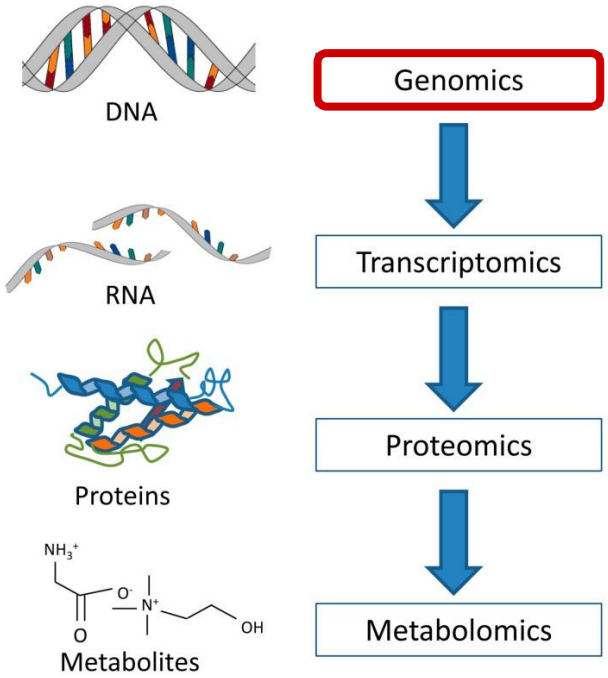
Profiling microbiome function



Profiles of **functional potential** can be built by annotating microbial genomes with known pathways.

Haukaas et al. 2017 Metabolites

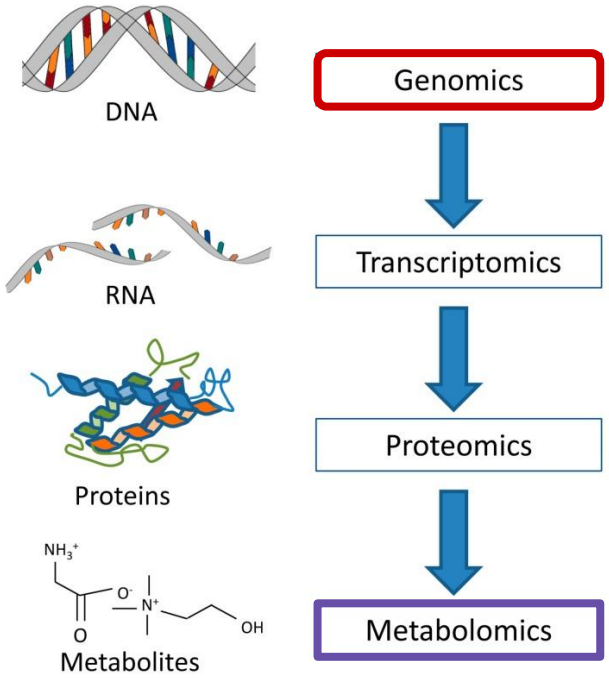
Profiling microbiome function



Profiles of **functional potential** can be built by annotating microbial genomes with known pathways.

Haukaas et al. 2017 Metabolites

Profiling microbiome function



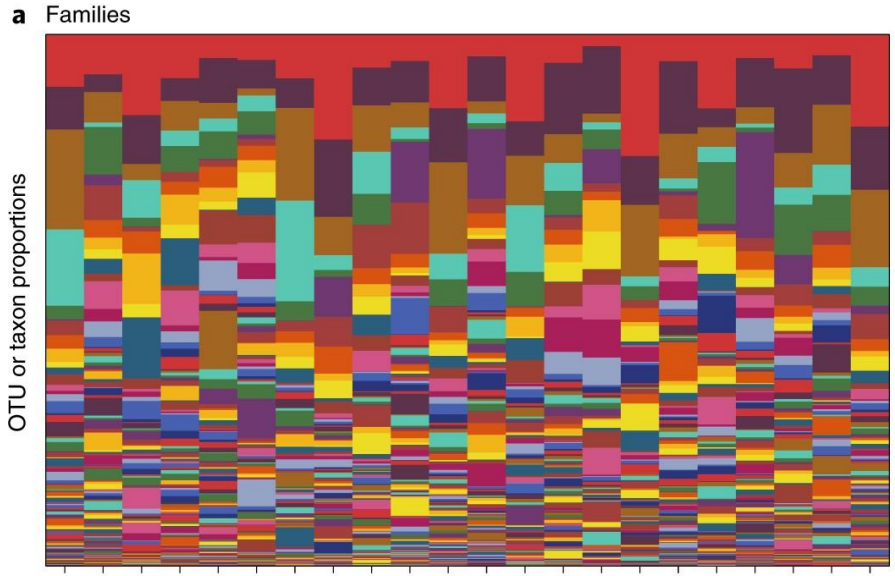
Profiles of **functional potential** can be built by annotating microbial genomes with known pathways.

Profiles of **active function** can be built by annotating metabolites with the pathways they are products of.

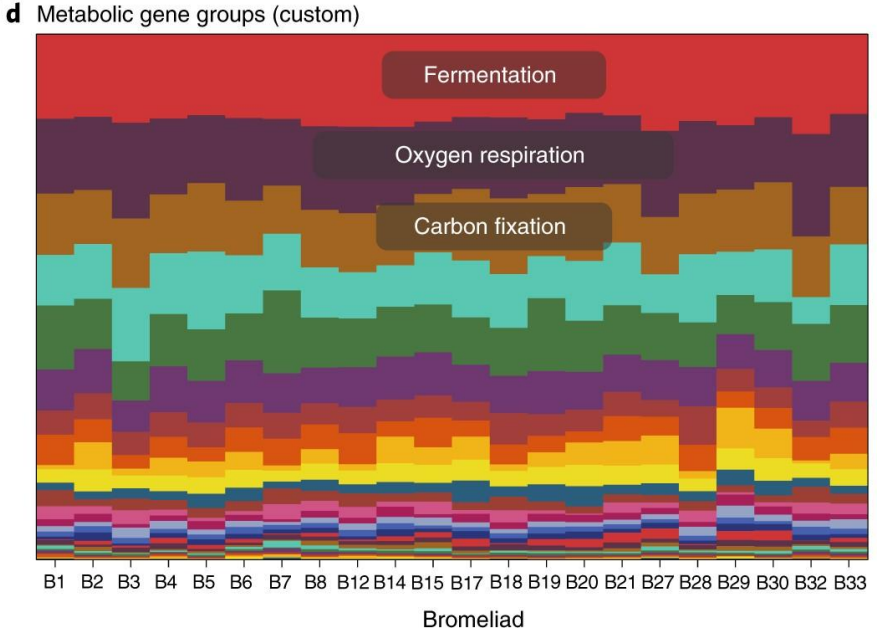
Haukaas et al. 2017 Metabolites

Taxonomic variability and functional stability

Taxonomic composition



Functional (potential) composition



Louca et al. 2018 Nature Ecol & Evol

Aim 1: Impact of functional redundancy of the gut microbiome on CRC classification.

Aim 2: Impact of integrating active metabolites with functional potential on CRC classification.

GLNE 007 Dataset

- 211 stool samples from patients with CRC.
- 223 stool samples from patients confirmed non-cancerous.
- Exclude adenomas, IBD, other active cancers.
- 16S rRNA gene sequencing already performed; stool left over for additional analyses.

Aim 1. Impact of functional redundancy of the gut microbiome on CRC classification.

Hypothesis: Using functional profiles instead of only taxonomic profiles improves classification modeling of stool samples as CRC or non-cancerous because of functional redundancy in the gut microbiome.

1. Build taxonomic & functional potential profiles.
2. Compare taxonomic & functional potential within & between disease states.
3. Build ML models with taxonomic profiles, functional profiles, or both and compare performance.

Aim 1A: taxonomic and functional potential profiles

- Build taxonomic profiles with 16S rRNA gene amplicon sequences; process and cluster into OTUs with mothur.
  - Output: table of OTU relative abundances for each sample.
- Build functional potential profiles with whole metagenome shotgun sequences; process with HUMAnN2.
  - Output: table of metabolic pathway relative abundances for each sample.

HUMAnN2 functional potential profiles

- Metagenomic reads are mapped to reference genomes to assign gene families.
- Gene families are mapped to the metabolic pathways they encode with the MetaCyc database.
- To avoid overestimating pathways, MinPath algorithm determines the minimum set of pathways that explain the genes present.
- HUMAnN2 output: table of metabolic pathways and samples

Aim 1B: functional redundancy in CRC

- No consensus on how to define or quantify functional redundancy with omics data.
- A practical way to define functional redundancy:
  - differences in **taxonomic** composition within and between disease states are not distinguishable, while:
  - differences in **functional** composition are greater between disease states than within.

Bray-Curtis dissimilarity index

- Calculate Bray-Curtis dissimilarity on OTU abundances of pairwise samples:

$$b_{ii'} = \frac{1}{2} \sum_{j}^{J} |r_{ij} - r_{i'j}|$$

Bray-Curtis dissimilarity index

- Calculate Bray-Curtis dissimilarity on OTU abundances of pairwise samples:

Total number of OTUs

Bray-Curtis index between samples i and i′

$$b_{ii'} = \frac{1}{2} \sum_{j}^{J} |r_{ij} - r_{i'j}|$$

Relative abundance of OTU j in sample i

Bray-Curtis dissimilarity index

- Calculate Bray-Curtis dissimilarity on OTU abundances of pairwise samples:

Total number of OTUs

Bray-Curtis index between samples i and i′

$$b_{ii'} = \frac{1}{2}\sum_{j}^{J} |r_{ij} - r_{i'j}|$$

Relative abundance of OTU j in sample i

- Range of $b_{ii'}$
  - 0 - all OTUs are shared at same abundances between samples.
  - 1 - no OTUs are shared between samples.
- Result: matrix of dissimilarities between all pairs of samples.

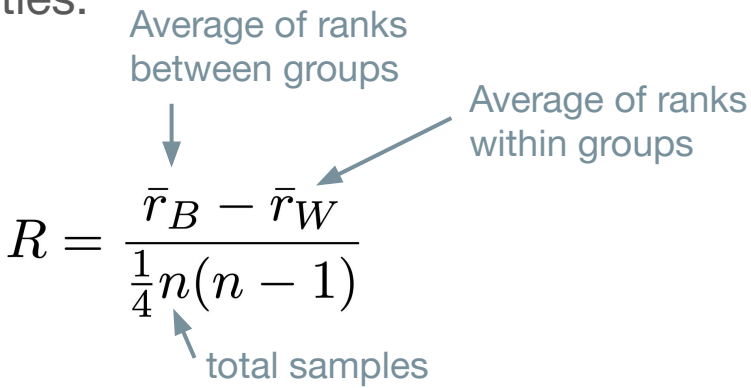
Analysis of Similarities (ANOSIM)

- Rank Bray-Curtis dissimilarities.
- Calculate the test statistic:

$$R = \frac{\bar{r}_B - \bar{r}_W}{\frac{1}{4}n(n-1)}$$

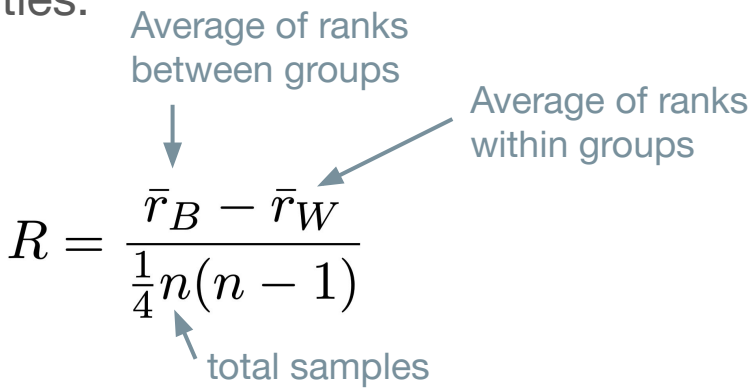
Analysis of Similarities (ANOSIM)

- Rank Bray-Curtis dissimilarities.
- Calculate the test statistic:



Analysis of Similarities (ANOSIM)

- Rank Bray-Curtis dissimilarities.
- Calculate the test statistic:



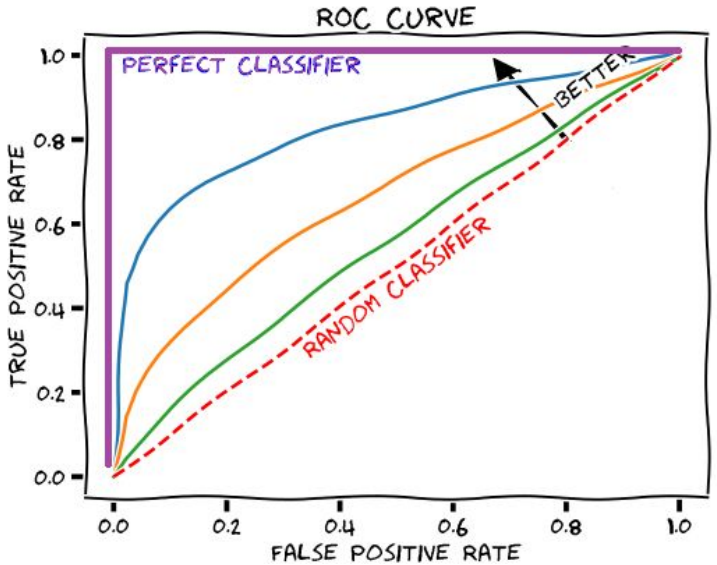
- Range of R
  - 1 - between-group dissimilarities are greater than within-group
  - 0 - no difference
  - -1 - within-group dissimilarities are greater than between-group
- Determine P value with a permutation test.

Aim 1B: functional redundancy in CRC

- Calculate Bray-Curtis dissimilarity on OTU abundances of pairwise samples.
- Calculate Bray-Curtis dissimilarity on potential pathway abundances of pairwise samples.
- Evaluate statistical significance with Analysis of Similarity (ANOSIM).
- Visualize dissimilarities with Nonmetric Multidimensional Scaling (NMDS).
- If there is functional redundancy:
    - differences in **taxonomic** composition within and between disease states are not distinguishable, while:
    - differences in **functional** composition are greater between disease states than within.

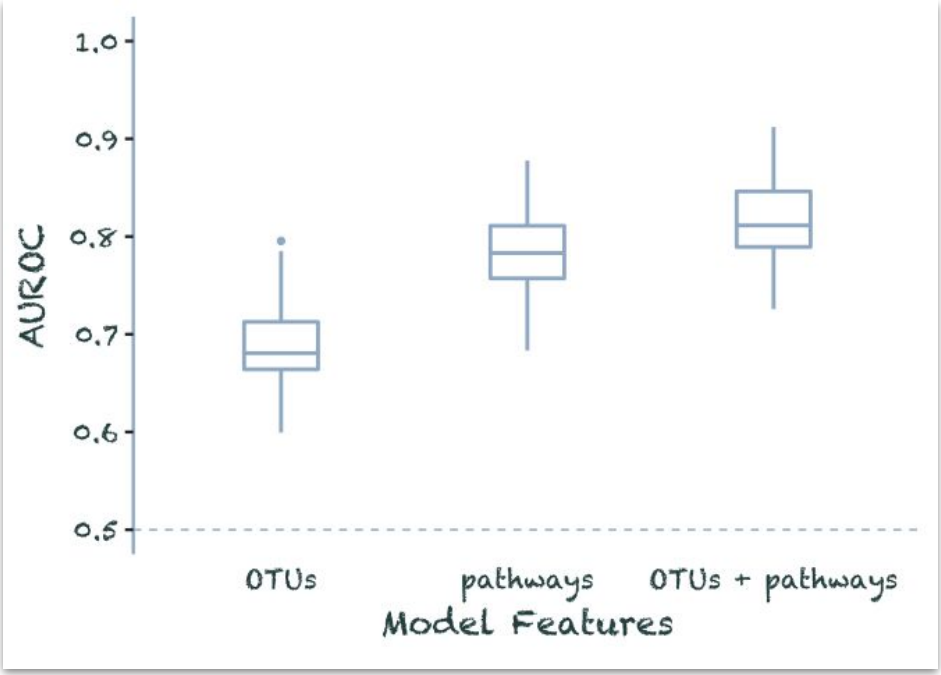
Aim 1C: CRC classification with taxonomic and functional profiles

- Build random forest models with OTUs, pathways, or both as model features.
  - Train on random data split with 80% training and 20% testing.
  - Calculate AUROC on held-out test data.
  - Repeat for 100 iterations.
- Wilcoxon test for significant differences of distributions of AUROCs between models:
  - Null hypothesis: AUROCs have the same distribution.



Aim 1 outcomes

[AUROC pathways > AUROC OTUs]



If models with functional potential perform better than taxonomic models, it suggests the importance of functional redundancy in CRC.

Aim 1 outcomes

If models with functional potential perform no better or worse than taxonomic models:

$$[\ \text{AUROC pathways} \leq \text{AUROC OTUs}\ ]$$

- There may be microbial genes of unknown function, which are entirely missed by this analysis, that are important in CRC.
- Functional redundancy may not be sufficient to discriminate disease states.
- Functional potential may not be a close enough approximation to true function to discriminate disease states.

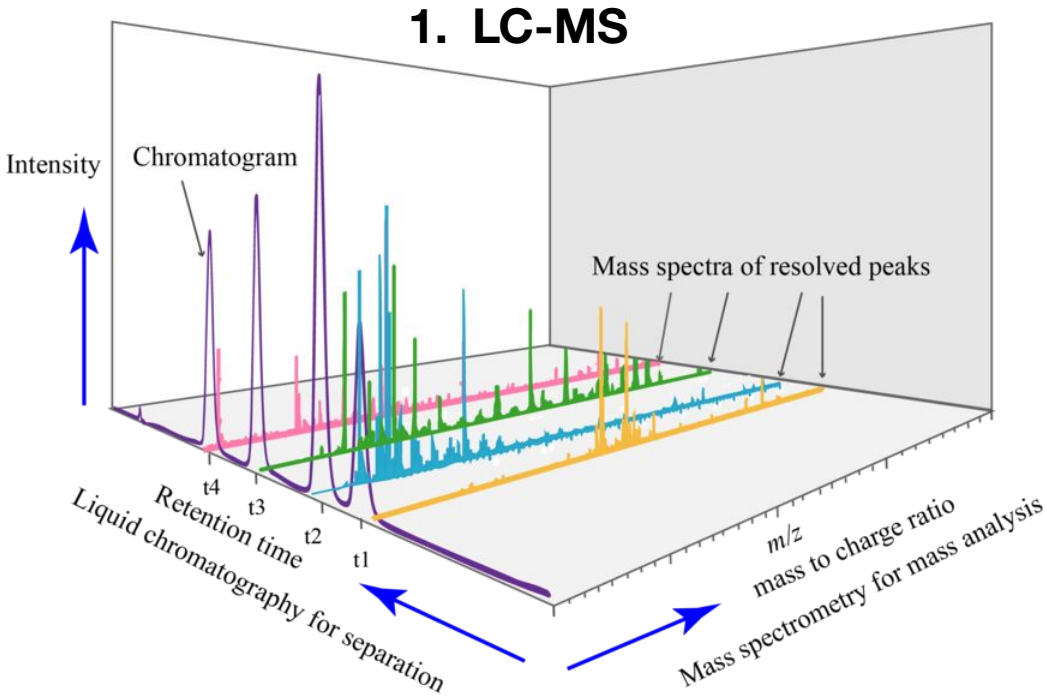
Aim 2. Impact of integrating active metabolites with functional potential on CRC classification.

Hypothesis: Using active metabolic pathways confirmed with mass spectrometry instead of all potential metabolic pathways from metagenomes improves the classification modeling of stool samples as CRC or non-cancerous.

1. Do untargeted metabolomics and annotate known metabolites.
2. Identify metabolites that could be produced by microbiota.
3. Build ML models with active metabolic pathways or all potential pathways and compare performance.

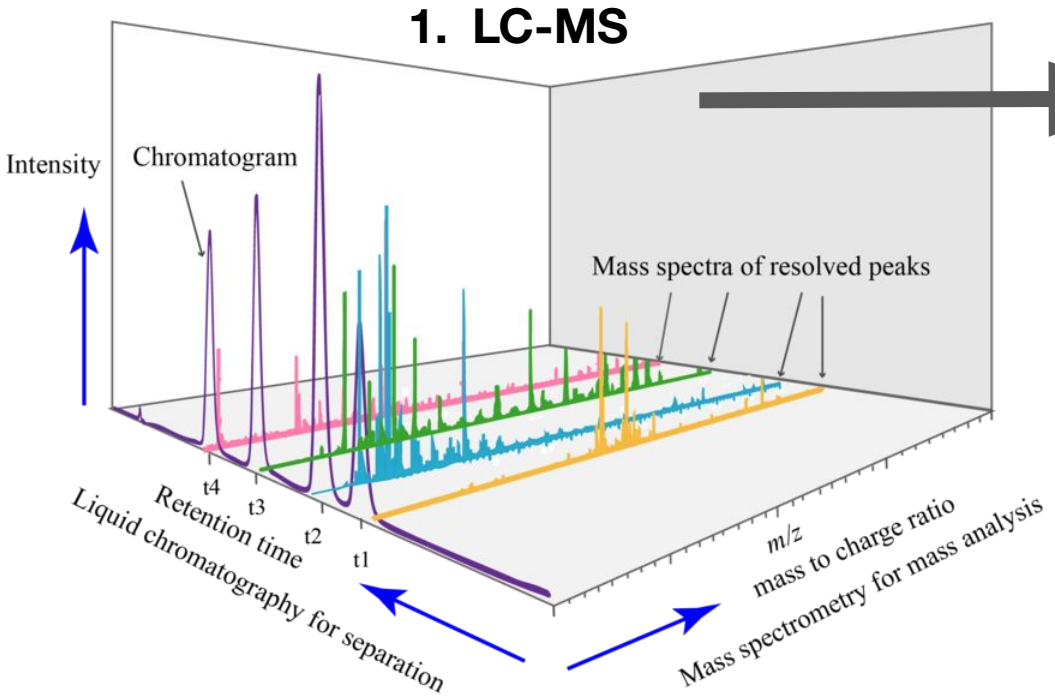
Aim 2A: untargeted metabolomics

Liquid chromatography tandem mass spectrometry (LC-MS/MS)



Aim 2A: untargeted metabolomics

Liquid chromatography tandem mass spectrometry (LC-MS/MS)



2. Fragment precursor ions

3. Second round of MS on fragments

Aim 2A: untargeted metabolomics

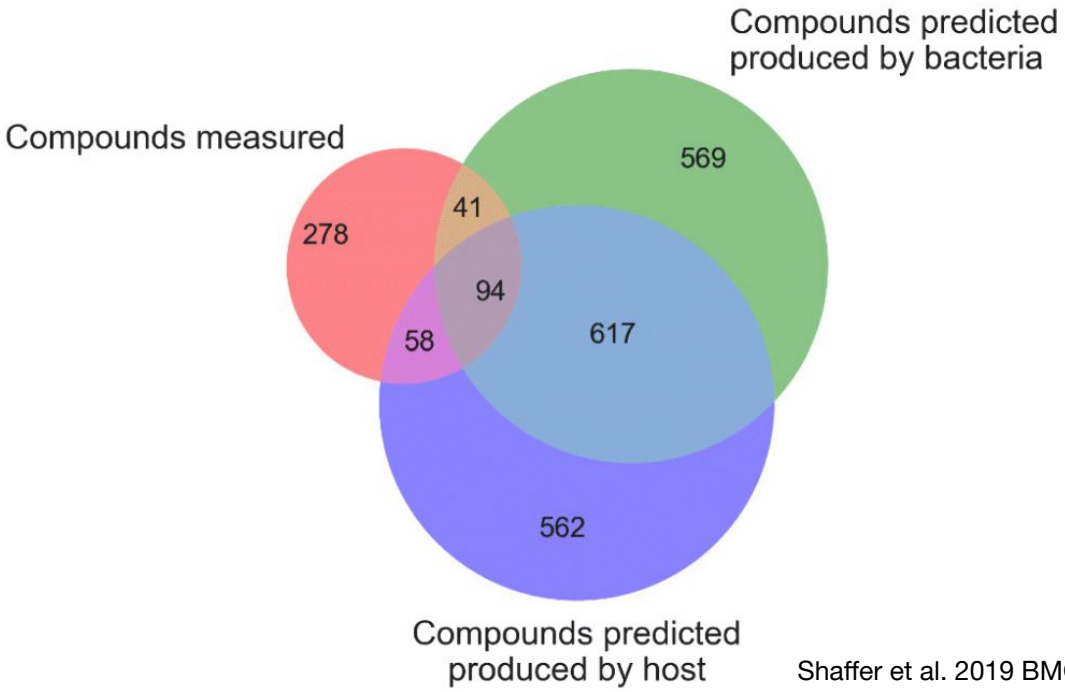
- Analyze LC-MS/MS data with GNPS.
- GNPS queries spectra against all reference spectral libraries to find near-exact matches and annotate matched compounds.
- As of 2016, GNPS had 18,163 known compounds in its database.
- Trained users can contribute new spectral libraries to GNPS, so it is constantly growing.

Aim 2B: known active bacterial metabolites

- Already have potential metabolic pathways from metagenomes analyzed by HUMAnN2 with the MetaCyc database.
- For all potential pathways, query MetaCyc to generate set of potential metabolic products.
- Intersect the set of potential metabolites from MetaCyc with annotated metabolites from LC-MS/MS to get metabolites that are:
  - Known to be products of bacterial metabolism in general.
  - Capable of being produced by these particular microbial communities.
- Output: set of pathways that produce known active bacterial metabolites

Aim 2B: known active bacterial metabolites

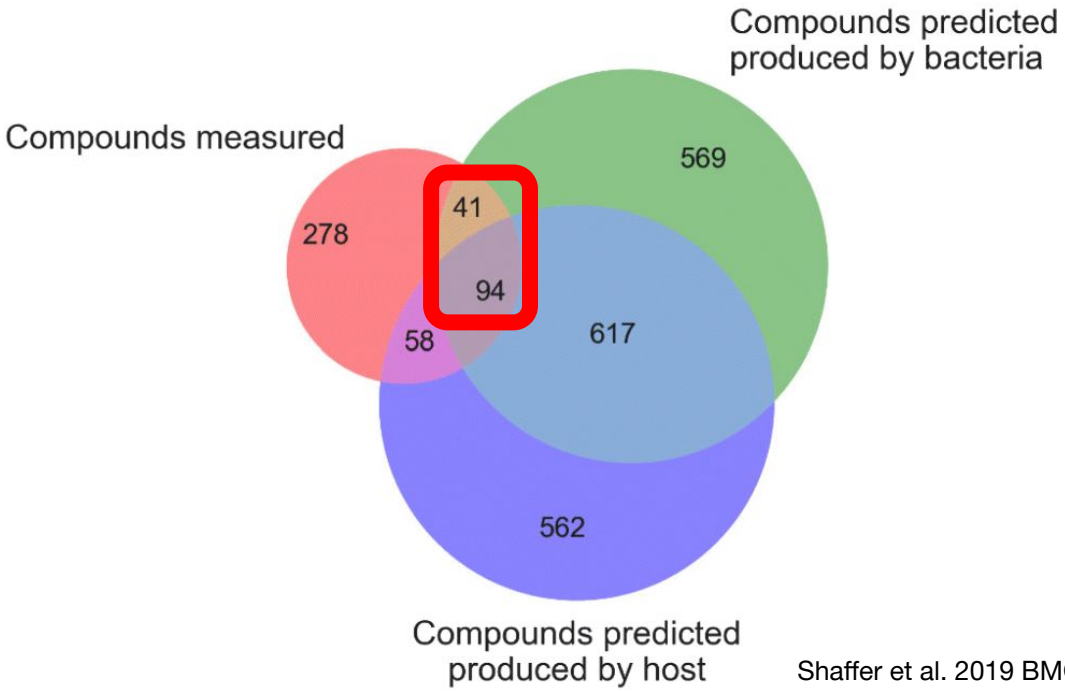
Set intersection of potential metabolites with active metabolites



Shaffer et al. 2019 BMC Bioinf

Aim 2B: known active bacterial metabolites

Set intersection of potential metabolites with active metabolites



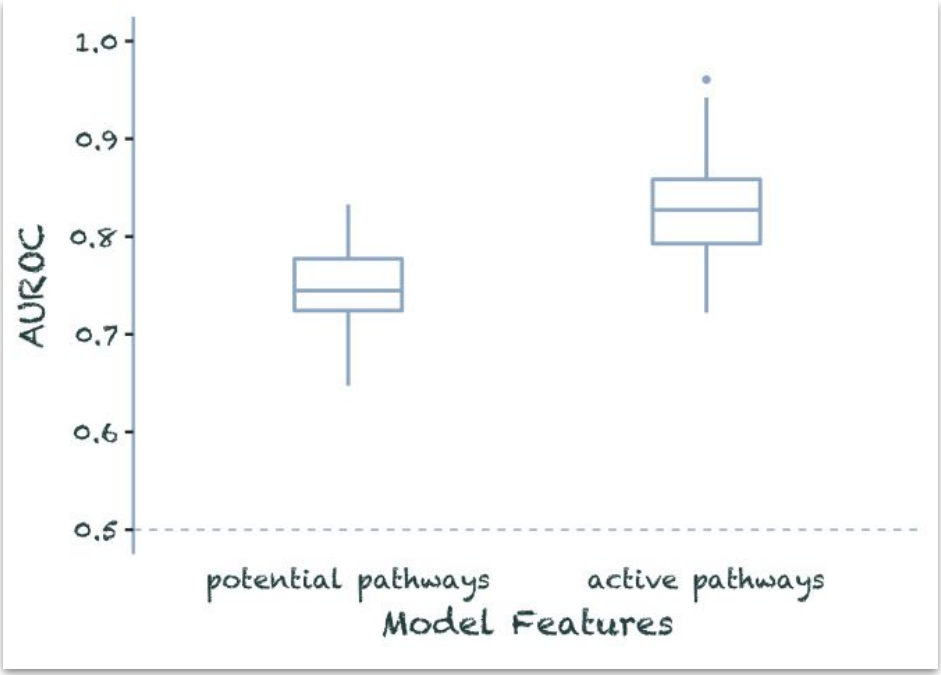
Shaffer et al. 2019 BMC Bioinf

Aim 2C: CRC classification modeling with confirmed active pathways

- Build random forest models with either only confirmed active pathways, or with all potential pathways.
  - Train on random data split with 80% training and 20% testing.
  - Calculate AUROC on held-out test data.
  - Repeat for 100 iterations.
- Wilcoxon test for significant difference in distributions of AUROCs between models:
  - Null hypothesis: AUROCs have the same distribution.

Aim 2 outcomes

[AUROC active > AUROC potential]



If models with active pathways outperform models with all potential pathways, it suggests functional potential from metagenomes is not a close enough approximation to real function.

Aim 2 outcomes

- If models with all potential pathways outperform models with active pathways, metagenomics data may compensate for unknown metabolites or low abundance metabolites missed by LC-MS/MS.

  [AUROC active < AUROC potential]

- If both models perform poorly, there may be microbial genes of unknown function that are important in CRC classification.

  [AUROC active & AUROC potential ≈ 0.5]

Additional limitations

- Stool samples are proxies for the actual gut environment.
- These data are not longitudinal.
- These analyses only consider microbial genes, pathways, and metabolites. Ignoring host genetics and risk factors completely.
- Metagenomics and metabolomics are expensive.

Congratulations!

you've unlocked the backup slides

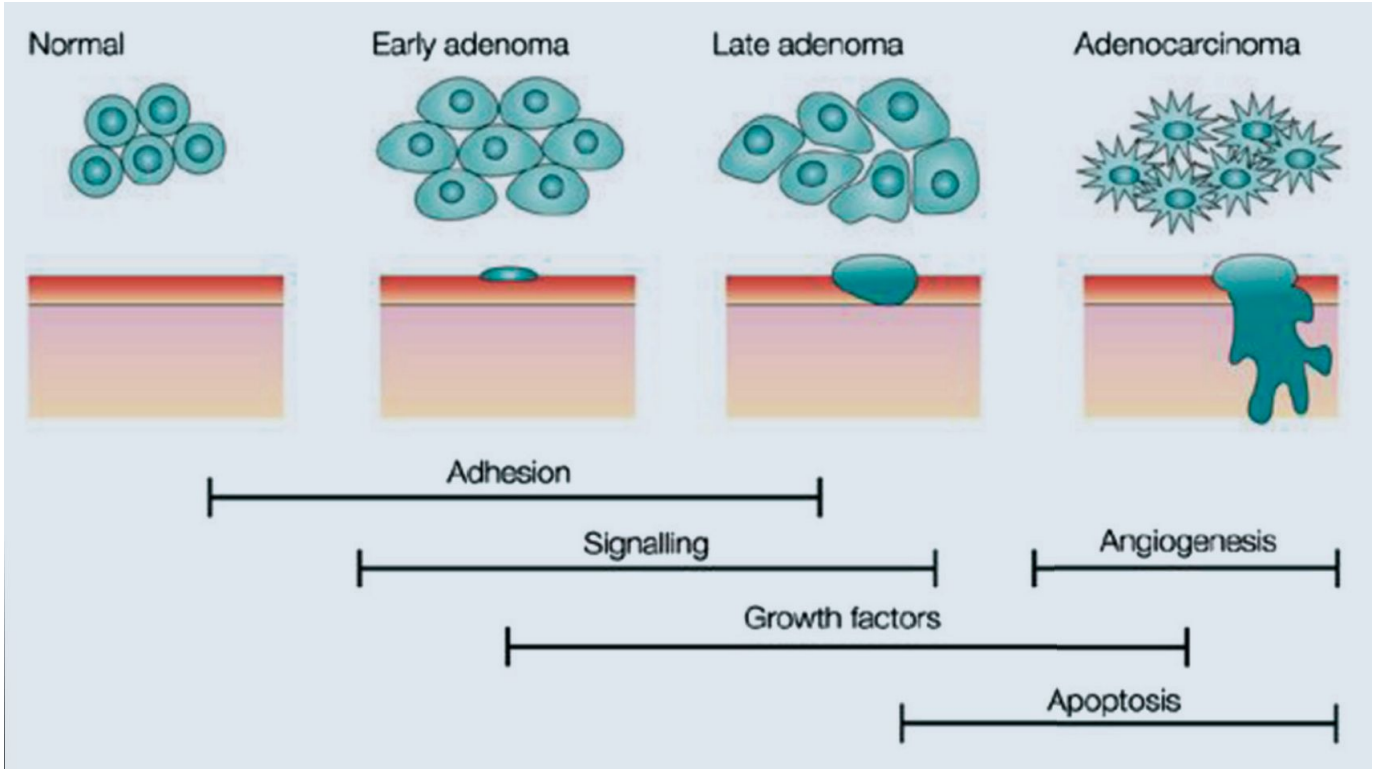
Microbiome changes in CRC

- *Fusobacterium nucleatum* - adhesion protein
- *Bacteroides fragilis* - enterotoxin
- *Pks+ Escherichia coli* - colibactin, induces DNA double-strand breaks
- *Clostridium* species - conversion of primary to secondary bile acids, associated with liver cancer

All vary broadly in abundance, significance, and enrichment across studies

Wirbel et al. 2019 Nature Med.

Adenoma-Carcinoma Sequence



Armaghany et al. 2012 GCR

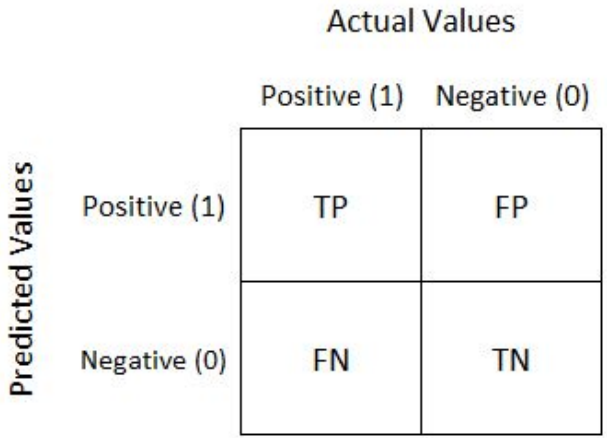
OTU clustering

credit: M Brodie Mumphrey

Mothur clustering algorithm: OptiClust

- *De novo* clustering: no reference database.
- Sequence pairs are considered similar if > 97% sequence similarity.
- Algorithm iteratively assigns samples to OTUs by maximizing the MCC.
- Matthews Correlation Coefficient:

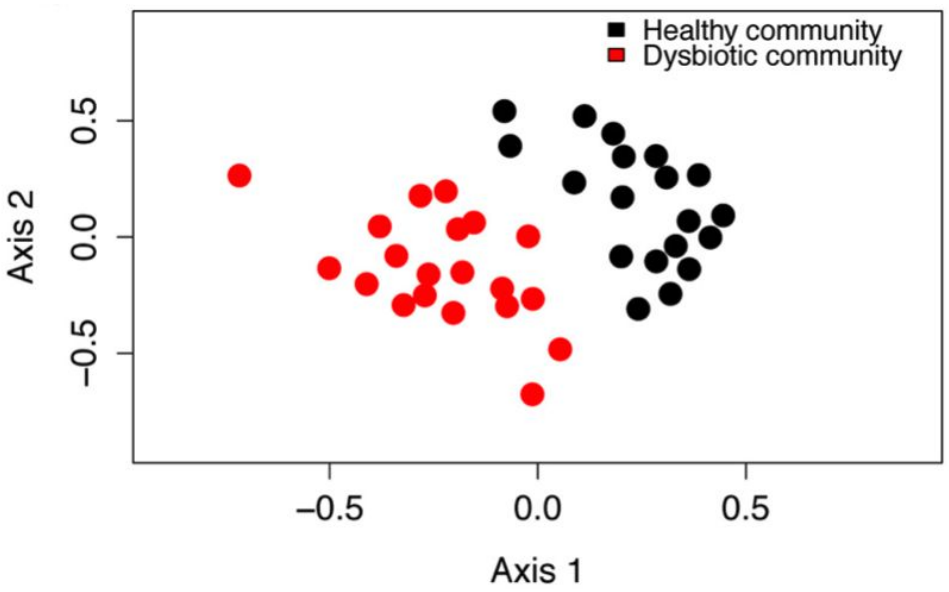
$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$



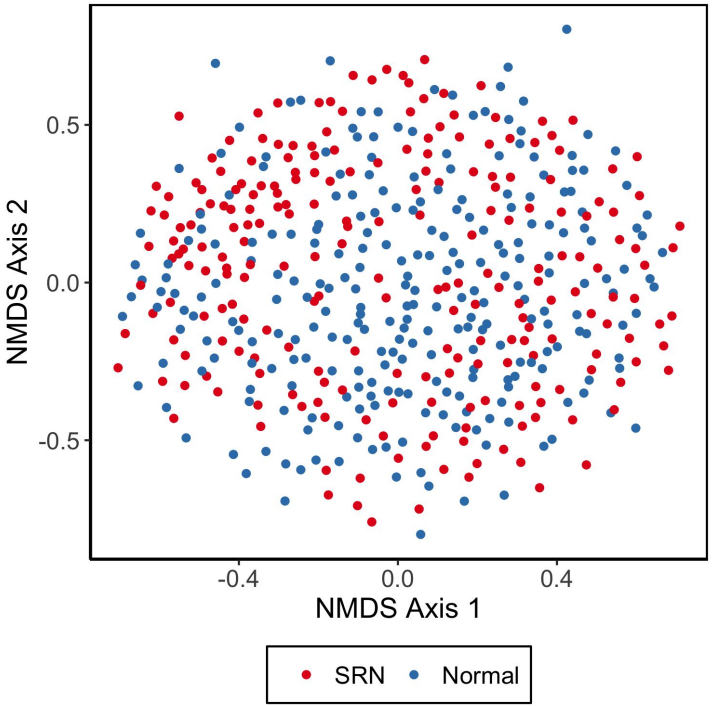
Range of MCC:

- 1 - perfect prediction
- 0 - random
- -1 - completely wrong

NMDS

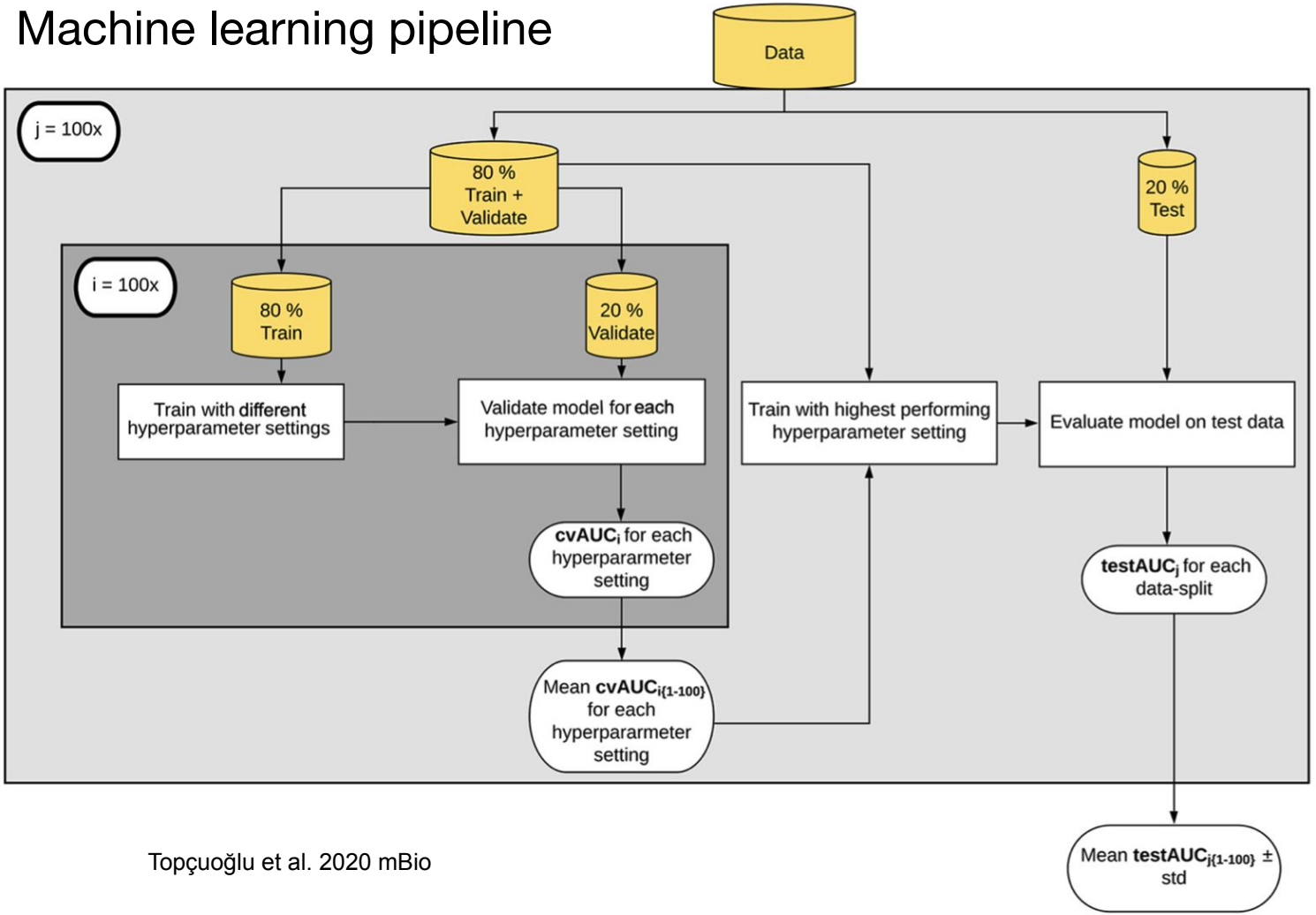


Zackular et al 2013 mBio



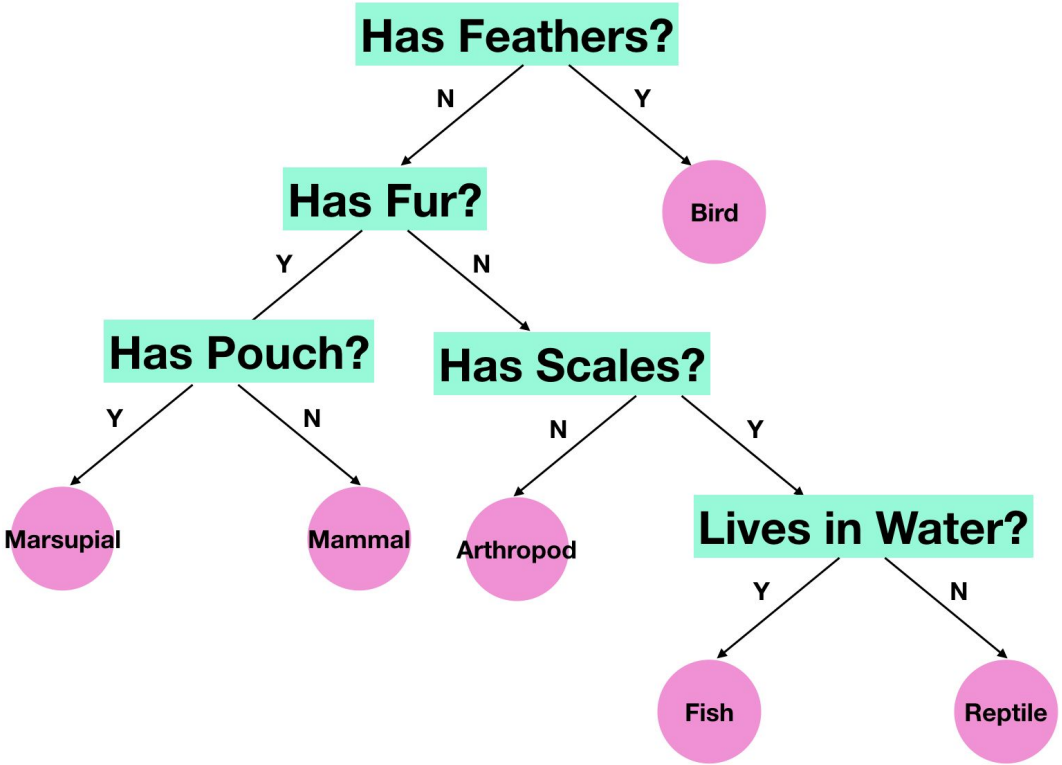
Topçuoğlu et al. 2020 mBio

Machine learning pipeline

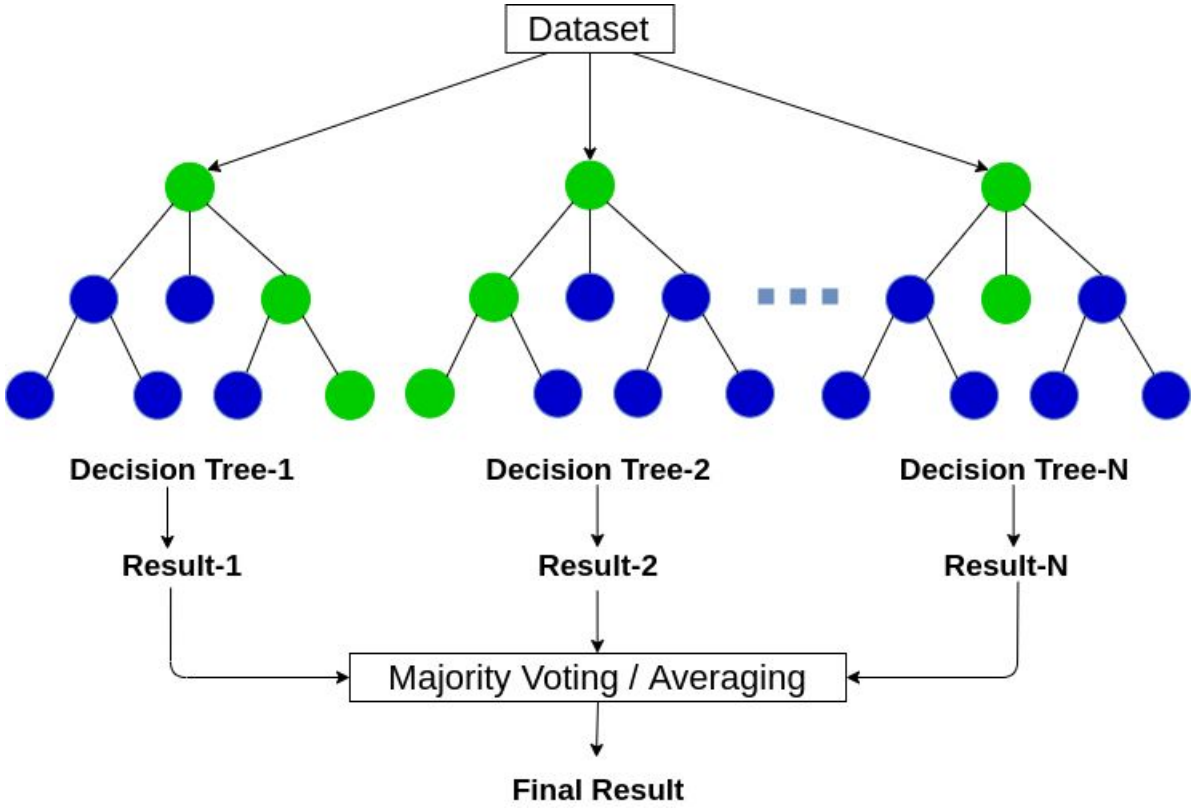


Topçuoğlu et al. 2020 mBio

Decision trees



Random forest



Wilcoxon test

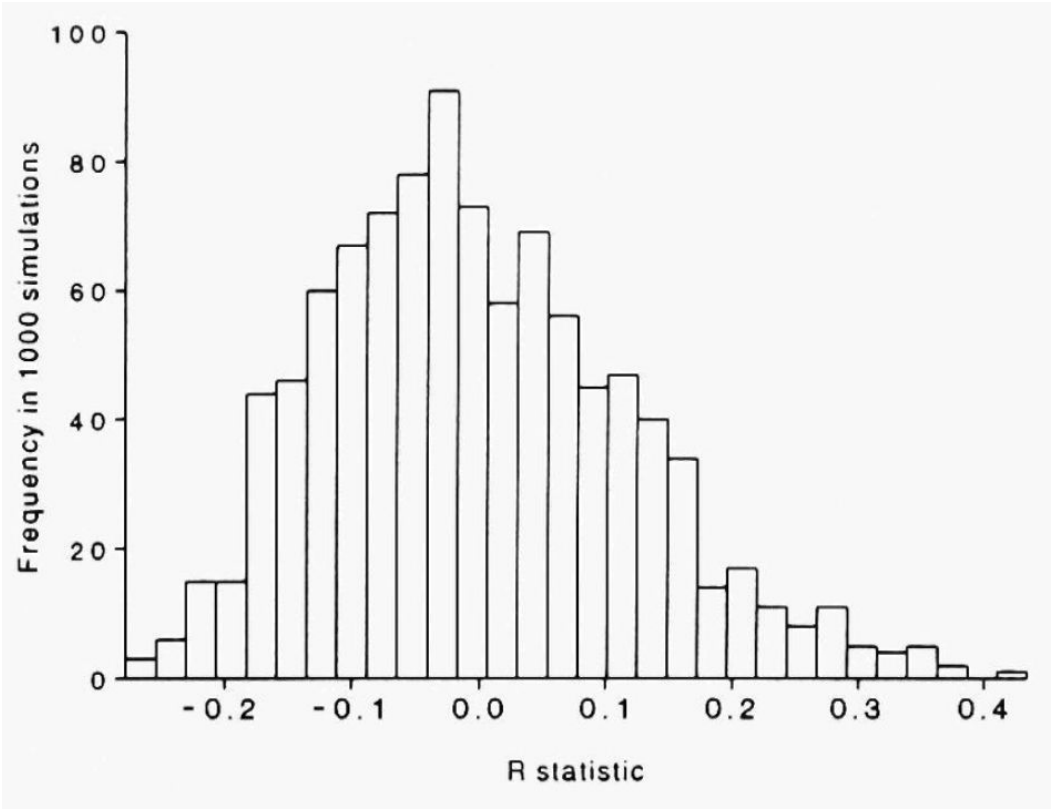
- Rank AUROCs.
- Calculate average rank for each group (model).
- Calculate U statistic for each:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

- U corresponds to the number of “wins” out of all pairwise comparisons.
- U is ~ normally distributed for large sample sizes, *P* value from normal table.

ANOSIM permutation test

Actual R was 0.45, which is greater than all sampled permutations.



Clarke 1994 AJoE

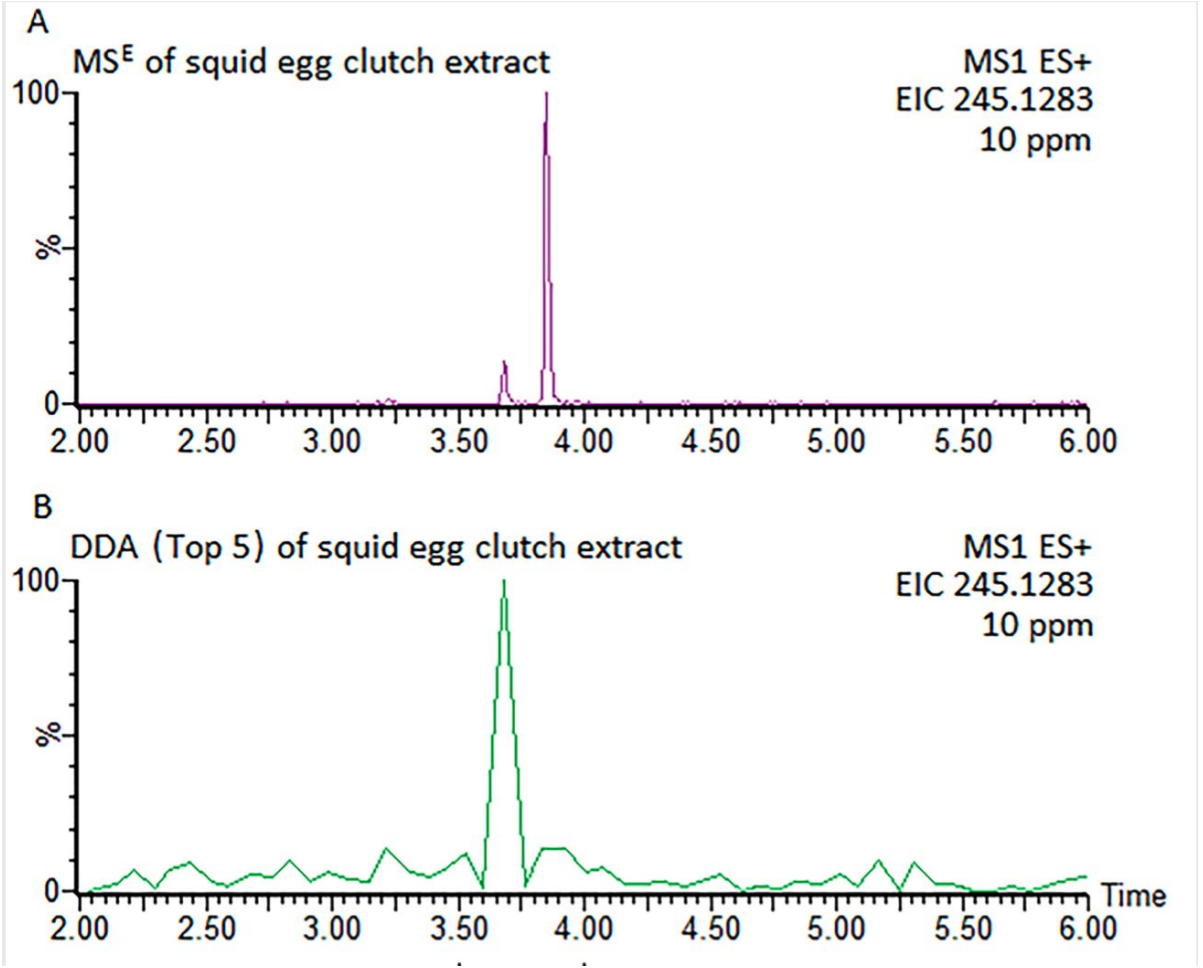
Cosine (dis)similarity

- After matching features by parent mass and retention time, consider MS2 fragments with:

$$D_{cosine} = 1 - \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2} \sqrt{\sum_{i=1}^{N} B_i^2}}$$

- A_i and B_i are the relative intensities of fragment i in features A and B
- Range: 0 to 1

DIA vs DDA



Puckett et al. 2020 Metabolomics