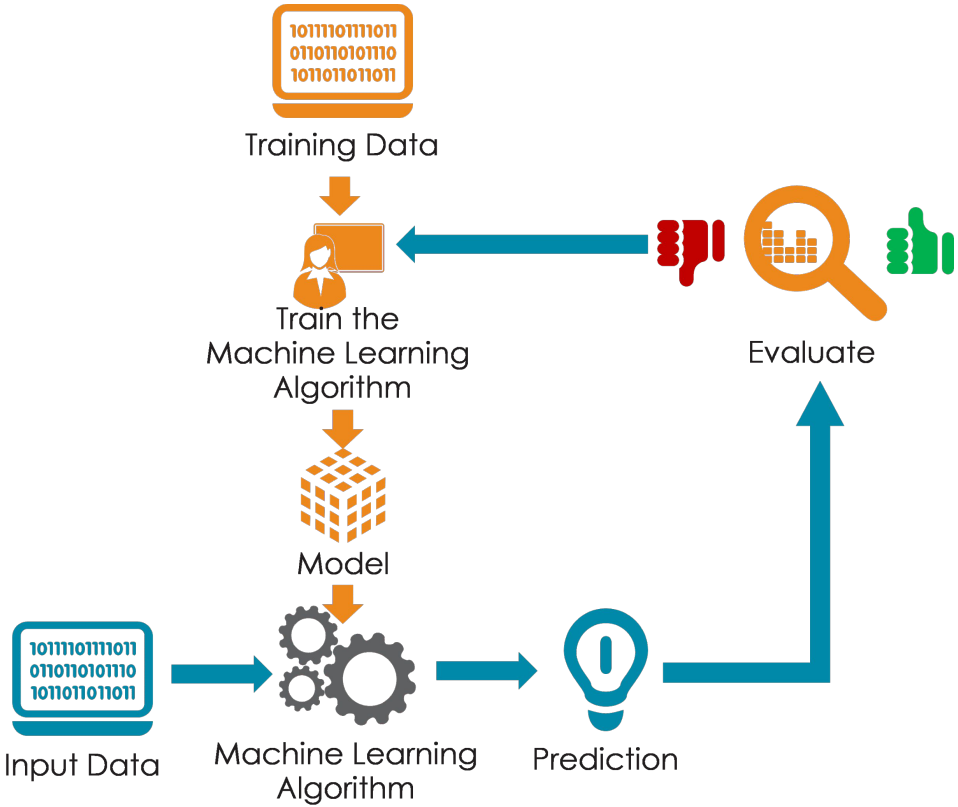
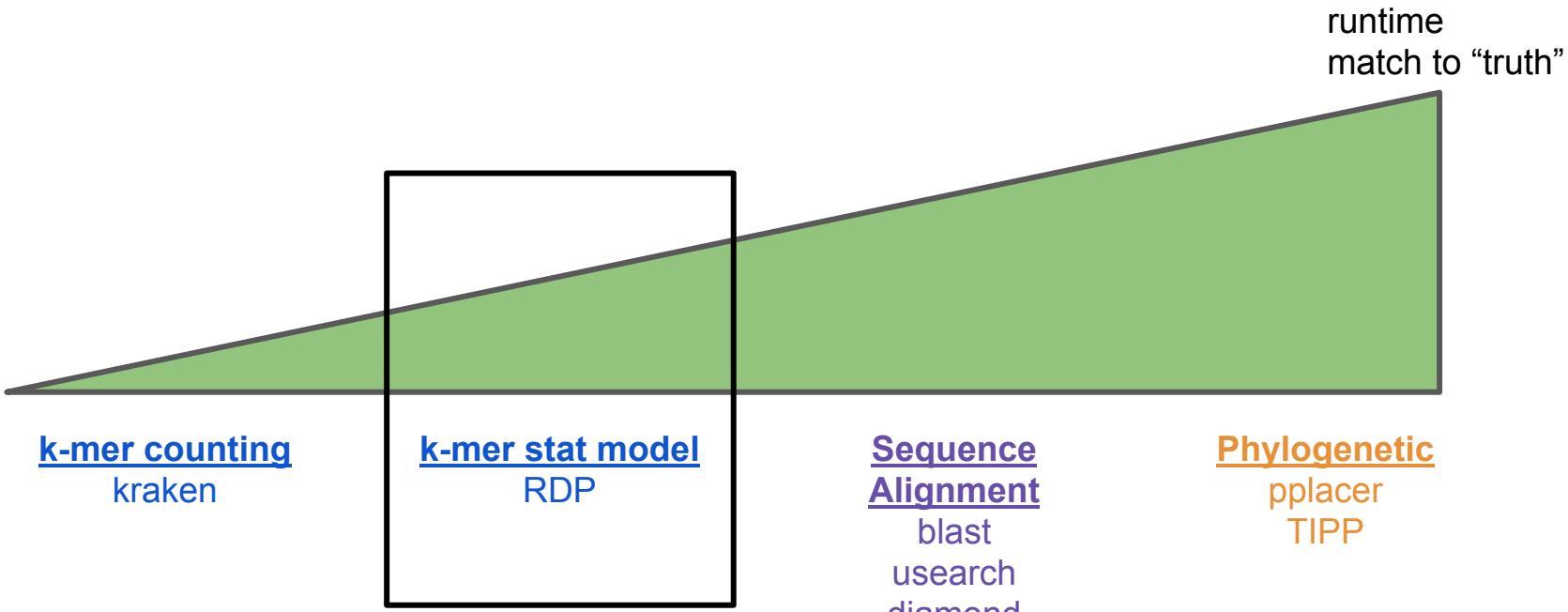
Machine Learning Approaches: Ribosomal Database Classifier

M^3 Workshop
January 10, 2019

Using machine learning to predict taxonomic labels



Landscape of Taxonomic Approaches



Ribosomal Database Project Classifier

- https://rdp.cme.msu.edu/
- Github: https://github.com/rdpstaff/classifier
- Publication: Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Wang et al. Appl Environ Microbiol 2007.
- Default for mothur's classify.seqs (method=wang) and option for QIIME's assign_taxonomy.py

Building the RDP Classifier

- Uses **k-mers**-- all possible substrings of length k that are contained in a string

Sequence: **ATGGAAGTCGCGGAA**

8-mers
ATGGAAGT
TGGAAGTC
GGAAGTCG
GAAGTCGC
AAGTCGCG
AGTCGCGG
GTCGCGGA
TCGCGGAA

Building the RDP Classifier

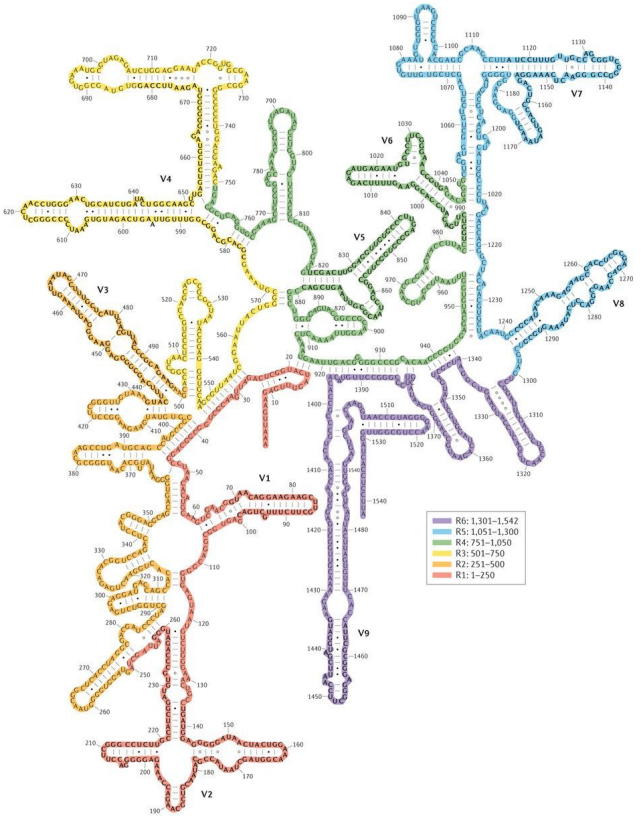
- Uses **k-mers**-- all possible substrings of length k that are contained in a string

Sequence: **ATGGAAGTCGCGGAA**
8-mers ATGGAAGT
TGGAAGTC
GGAAGTCG
GAAGTCGC
AAGTCGCG
AGTCGCGG
GTCGCGGA
TCGCGGAA

- Identifies all possible 8-mers in a taxon database and builds a table of how many times each 8-mer appears in each taxon

RDP taxonomy database is based on marker genes

- Why marker genes?
  - Compare apples to apples
  - Avoid lateral gene transfer
  - Can (in theory) extrapolate based on evolutionary principles
  - Can target experimentally ($'s matter)
- RDP has classifiers for:
  - 16S ribosomal RNA
  - Fungal LSU
  - Fungal ITS
- You can also train your own classifier with updated taxonomy or different marker genes!



Using the RDP Classifier to assign taxonomy

- Calculates all 8-mers in a query sequence and looks them up in the reference table to calculate the probability that the query sequence is a member of a genus

By Bayes' theorem, the probability that an unknown query sequence, S, is a member of genus G is

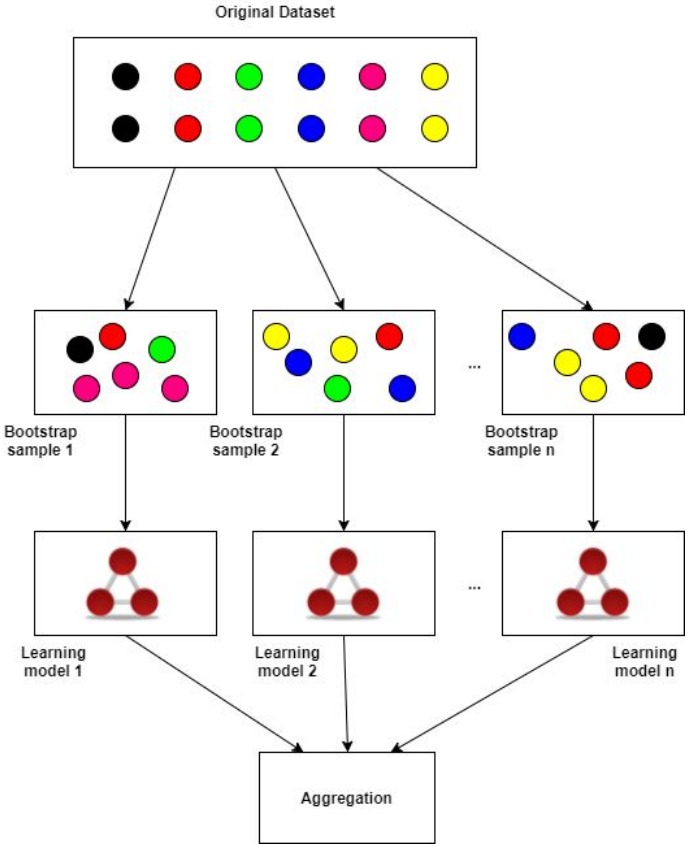
$$P(G|S) = P(S|G) \times P(G)/P(S)$$

P(G) = the prior probability of a sequence being a member of G

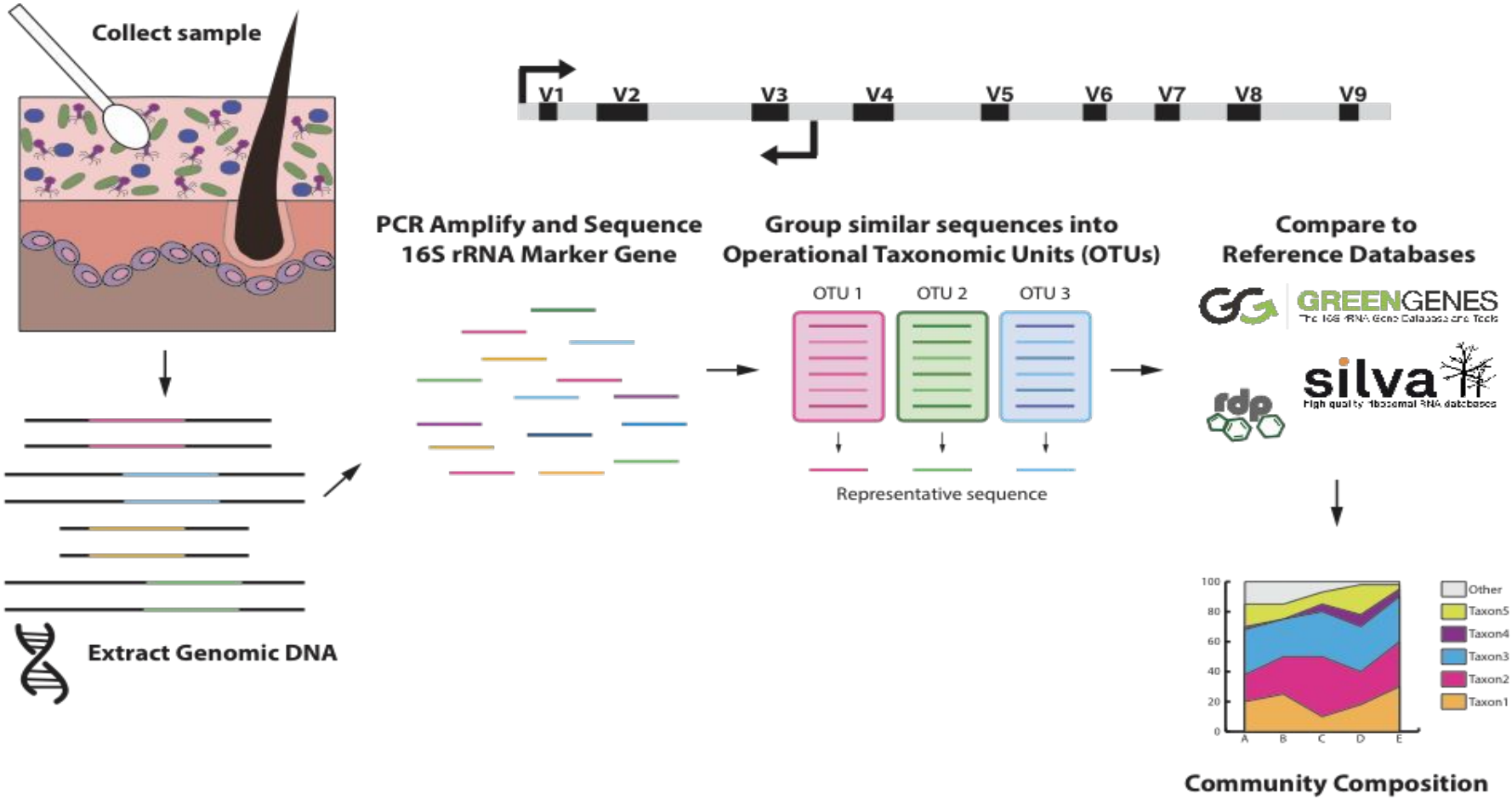
P(S) = the overall probability of observing sequence S (from any genus)

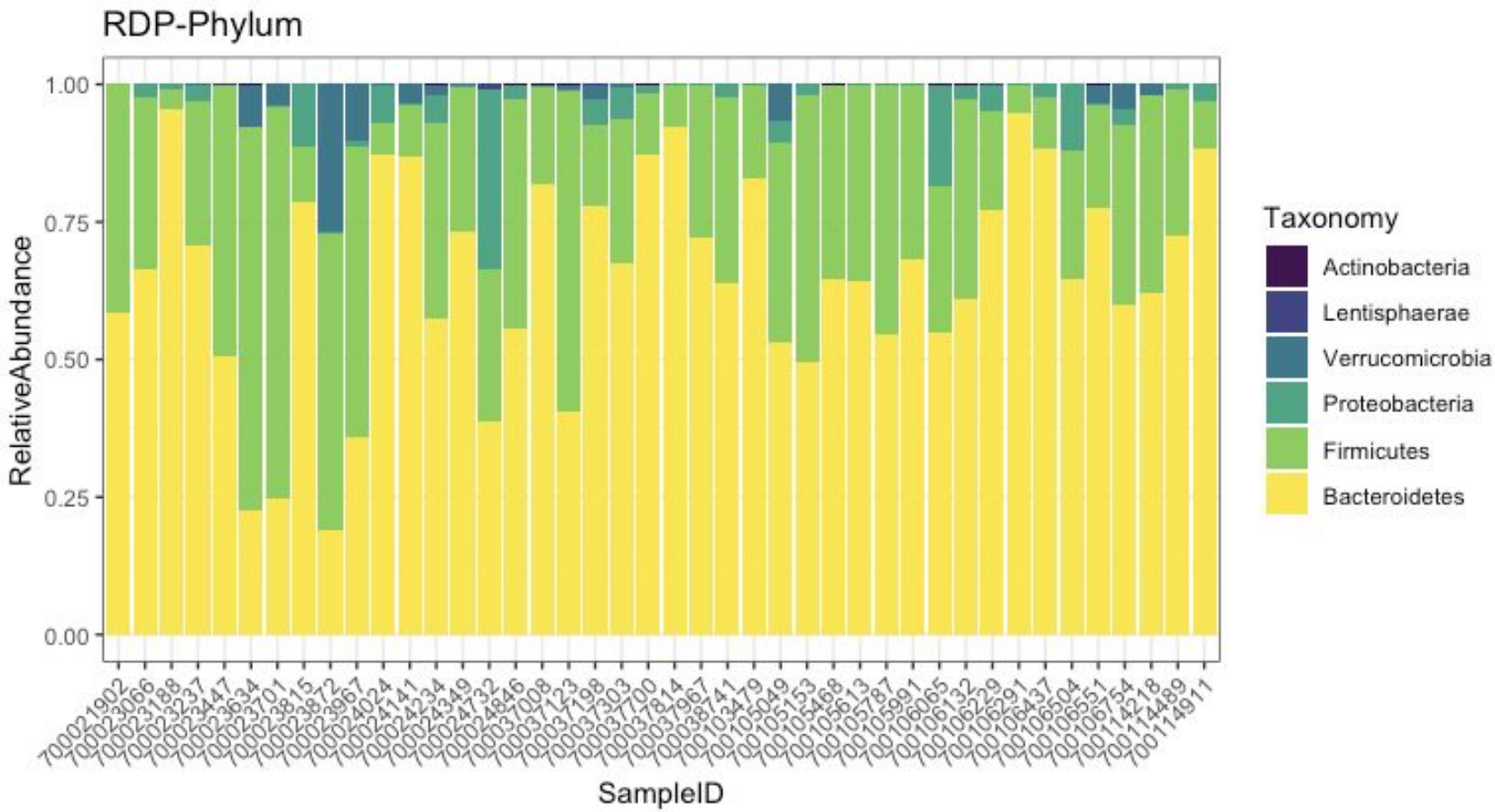
P(S|G) = the joint probability of observing from genus G a (partial) sequence, S, containing a set of words, V = {v1, v2, …, vf} = Π P(vi|G)

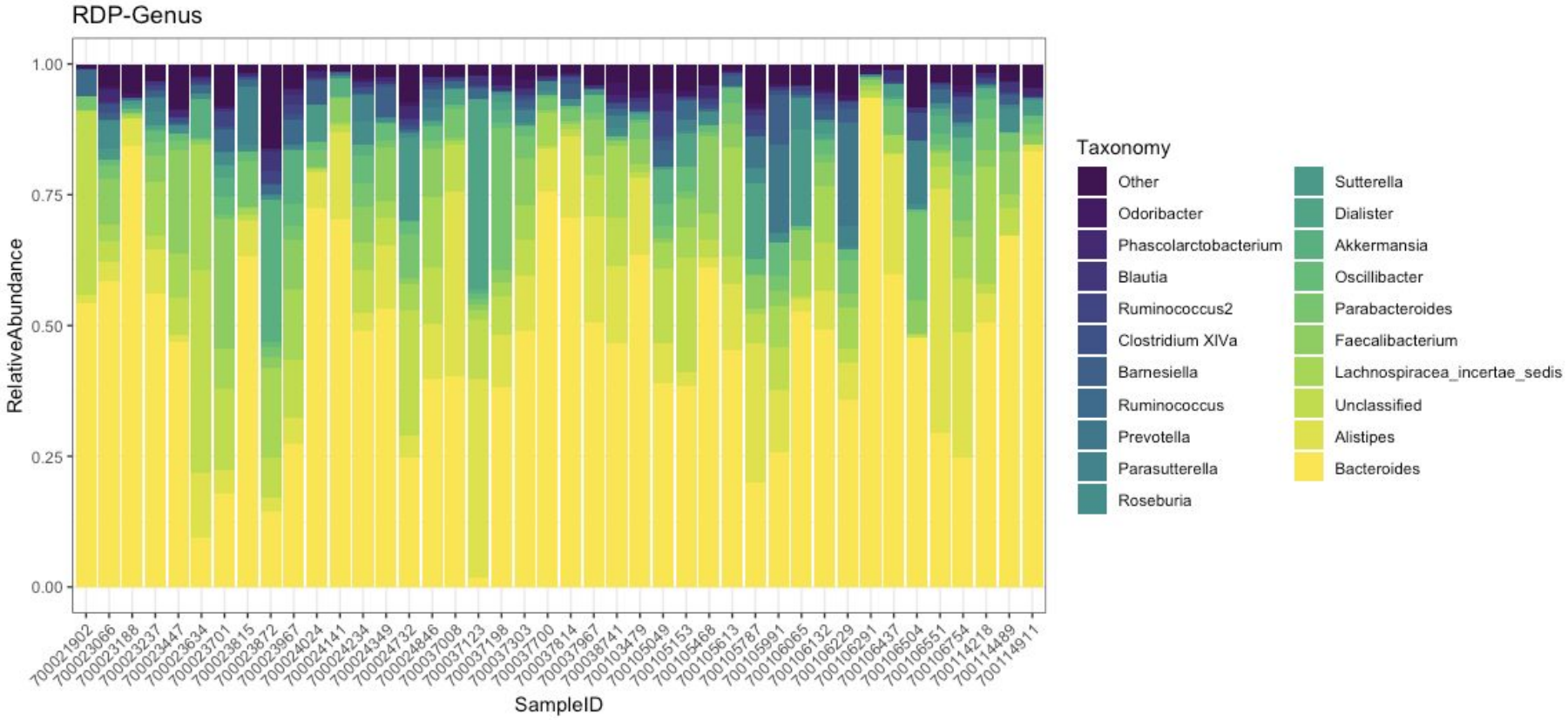
Assigning confidence to taxonomy predictions



- Use random-resampling (bootstrapping) to assign confidence
- Runs 100 times; each time it chooses a subset of ⅛ of all possible 8-mers from the query sequence
- For a particular taxonomic assignment, confidence is assessed by the number of times a particular genus is assigned out of 100 trials
- Confidence score ranges from 0 (not confident) - 1 (very confident)
- User can set a minimum confidence score [default: 0.8]


Collect sample
V1
V2
V3
V4
V5
V6
V7
V8
V9
PCR Amplify and Sequence
16S rRNA Marker Gene
Group similar sequences into
Operational Taxonomic Units (OTUs)
OTU 1
OTU 2
OTU 3
Representative sequence
Compare to
Reference Databases
GREENGENES
silva
rdp
Extract Genomic DNA
100
80
60
40
20
0
A
B
C
D
E
Other
Taxon5
Taxon4
Taxon3
Taxon2
Taxon1
Community Composition


RDP-Phylum
RelativeAbundance
1.00
0.75
0.50
0.25
0.00
SampleID
700021902
700023066
700023188
700023237
700023447
700023634
700023701
700023815
700023872
700023967
700024024
700024141
700024234
700024349
700024732
700024846
700037008
700037123
700037198
700037303
700037700
700037814
700037967
700038741
700103479
700105049
700105153
700105468
700105613
700105787
700105991
700106065
700106132
700106229
700106291
700106437
700106504
700106551
700106754
700114218
700114489
700114911
Taxonomy
Actinobacteria
Lentisphaerae
Verrucomicrobia
Proteobacteria
Firmicutes
Bacteroidetes


RDP-Genus
RelativeAbundance
1.00
0.75
0.50
0.25
0.00
SampleID
700021902
700023066
700023188
700023237
700023447
700023634
700023701
700023815
700023872
700023967
700024024
700024141
700024234
700024349
700024732
700024846
700037008
700037123
700037198
700037303
700037700
700037814
700037967
700038741
700103479
700105049
700105153
700105468
700105613
700105787
700105991
700106065
700106132
700106229
700106291
700106437
700106504
700106551
700106754
700114218
700114489
700114911
Taxonomy
Other
Odoribacter
Phascolarctobacterium
Blautia
Ruminococcus2
Clostridium XlVa
Barnesiella
Ruminococcus
Prevotella
Parasutterella
Roseburia
Sutterella
Dialister
Akkermansia
Oscillibacter
Parabacteroides
Faecalibacterium
Lachnospiracea_incertae_sedis
Unclassified
Alistipes
Bacteroides

If you were to train a classifier, what features would you use?