LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK, LMU | HELMHOLTZ ZENTRUM FÜR INFEKTIONSFORSCHUNG, BRAUNSCHWEIG

Interpreting Sequence-based Deep Learning Models for Genomic Data

Yichen Han,
10.09.2024

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

1. Background

i. Research on model interpretability in Genomics focuses more on gene-based models, instead of sequence-based.

ii. Attempt to embed prior biological knowledge into the sequence-based model architecture (visible networks) is not applicable to DeepG.

iii. Papers published about feature selection in seq-based models, without discussing biological significance of selected features.

iv. Existing permutation-based feature importance methods are inefficient and expensive for long sequences.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

2. Project Overview

The DeepG package has already implemented Integrated Gradients (IG), a gradient-based feature attribution method with some desirable mathematical properties.

i. How reliable is IG? Does it produce biologically meaningful explanations?
   a) Expectation: task-specific or species-specific
   b) If not, is it due to our model or due to IG?

ii. Can we improve the current IG so that it fits genomic data better?
   a) Affordable computational costs given very long sequences
   b) Explanations can be facilitated by sequence annotations

iii. Comparison with other interpretable deep learning methods.

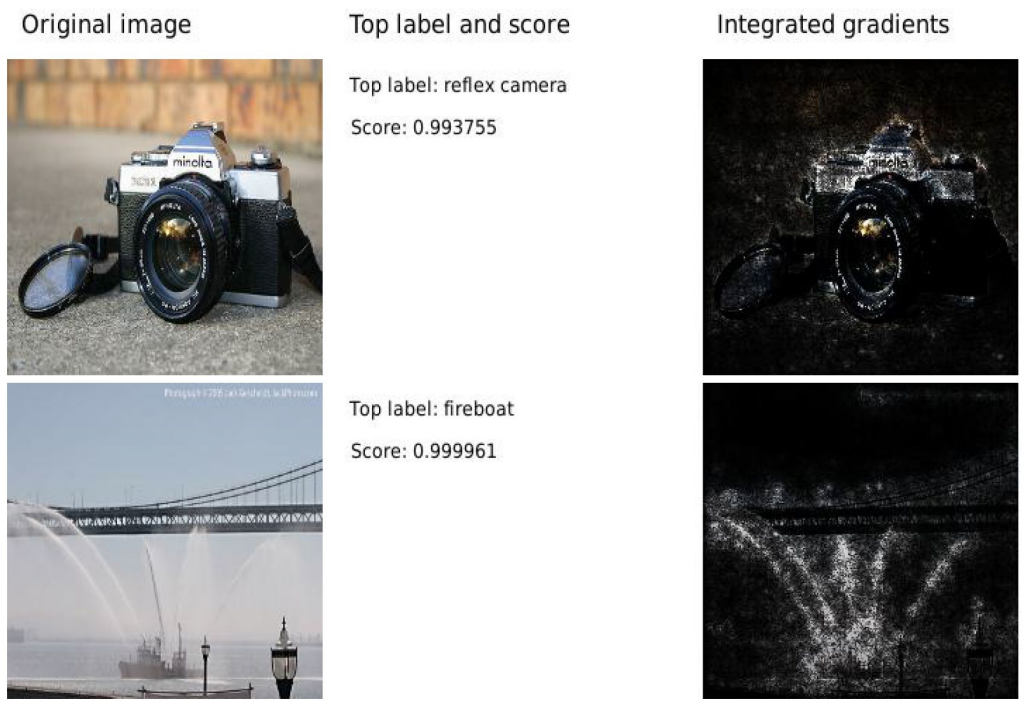
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Integrated Gradients

Empirical Approximation (Sundararajan et al. 2017):

$$IG_i^{emp}(x) := (x_i - x_i') \times \sum_{k=1}^{m} \frac{\partial F\left(x' + \frac{k}{m} \times (x - x')\right)}{\partial x_i} \times \frac{1}{m}$$

➔ Contribution of the i-th feature to model F prediction, given baseline x' and instance x. Iterated over m steps.



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Example Input

Our model takes a one-hot encoded (1, maxlen, 4) tensor / array as input.
The input has fixed length, so that the instance of IG needs to be a subsequence.

➔ Input of sequence –AG– [[[1, 0, 0, 0],
[0, 0, 1, 0]]]

➔ Corresponding least-informative baseline (“Baseline 0.25”, current default):
[[[0.25, 0.25, 0.25, 0.25],
[0.25, 0.25, 0.25, 0.25]]]

➔ Interpretation: each nucleotide base has equal probability to be present.

➔ Experimenting on different baselines is also part of the project.

➔ Result visualization: aggregate each locus by sum, then scatter plot.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Example Tasks

➔ Locus level: Motif detection with synthetic sequences

➔ Gene level: 16s rRNA gene vs. random bacteria sequence

➔ Genome level: Sporulation phenotype prediction

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Motif Detection

➔ Experiment 1: motifs with different information entropy

➔ Experiment 2: single vs. recurrent motifs

➔ Experiment 3: motifs with different lengths (TODO: model training)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Experiment 1: motifs with different information entropy

➔ 10000 sequences 600bp sampled from A,C,G,T (random sequences). 50% with motif.

One dataset for each of the 30bp motifs:

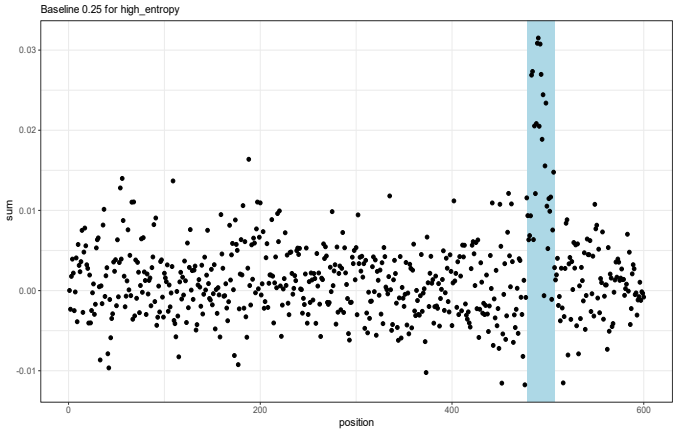
- High: GGTCGACACGAGTATAGCTAGTATACTCCG
- Medium: TATAGCGCAGCCCTTATGGAGAGTCCATGA
- …
- Low: TATAG * 6
- Lowest: A * 30

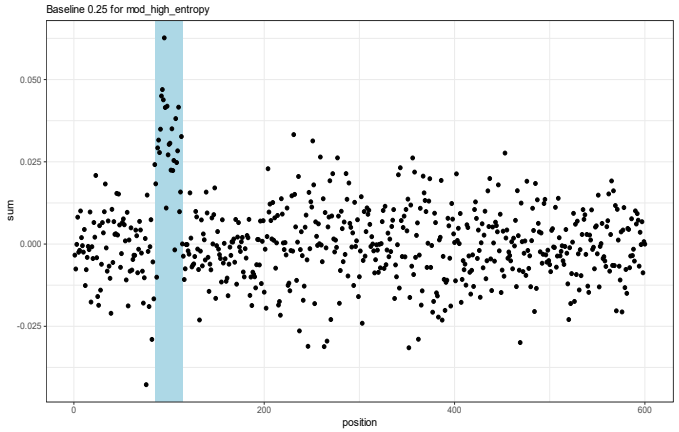
➔ Model: CNN. Same HPs for each dataset. Balanced Acc. > 97%.

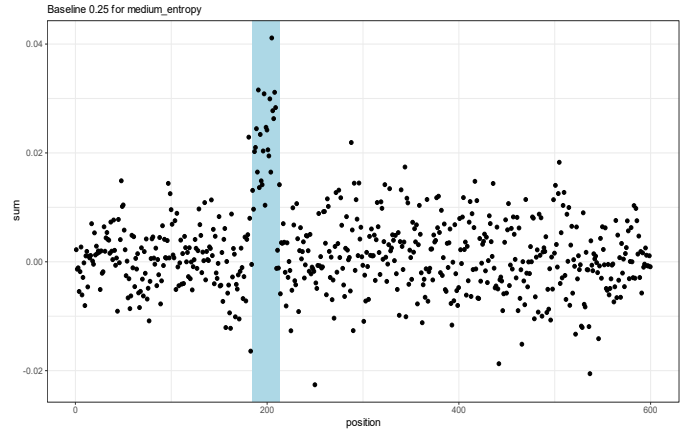
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

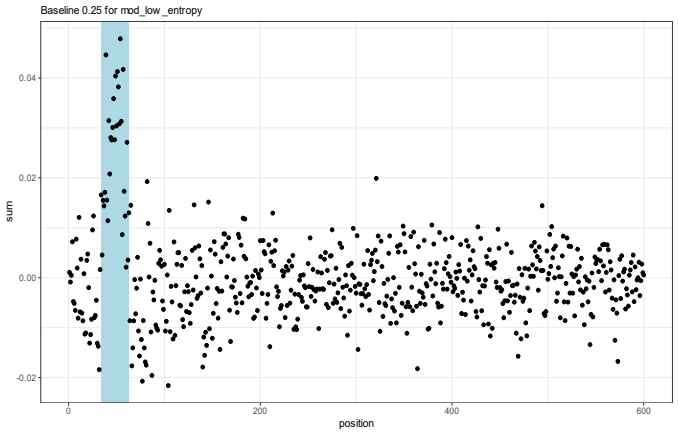
IG can detect a motif well regardless of its information entropy.

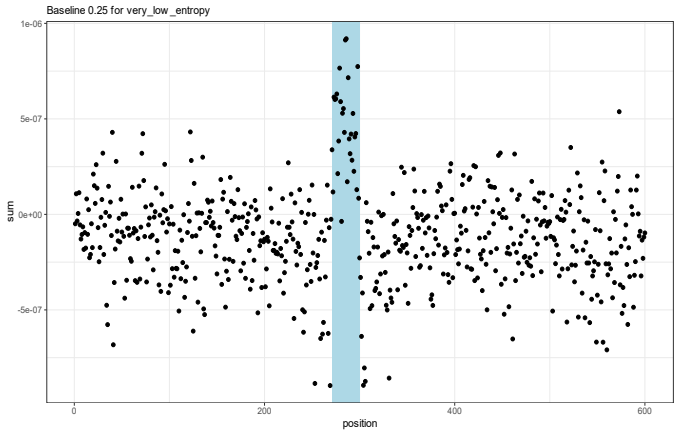
Blue area indicates motif presence.

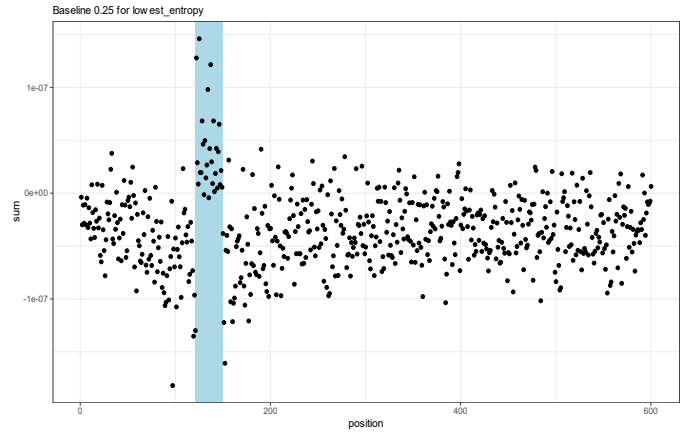






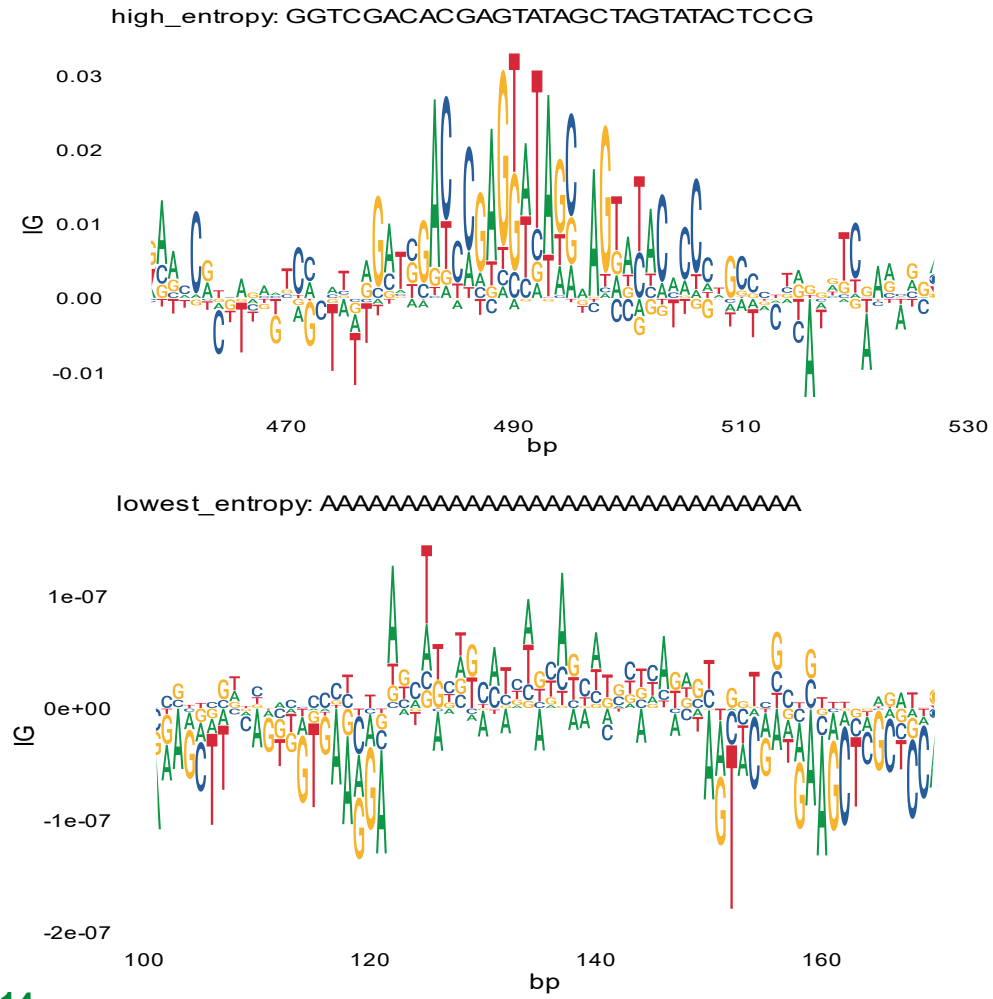


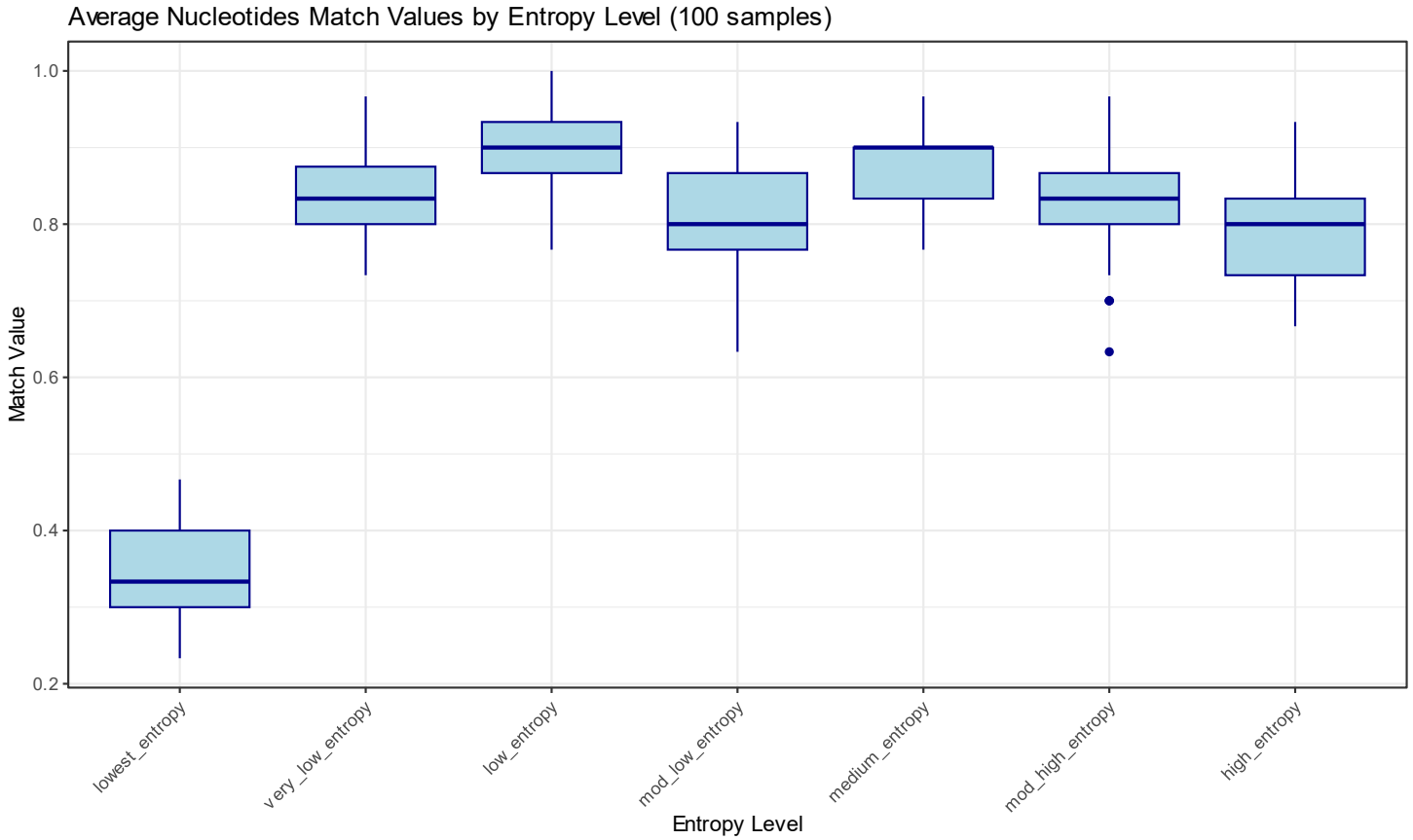




IG does not always assign highest importance to the correct base at each locus of the motif and is affected by information entropy.

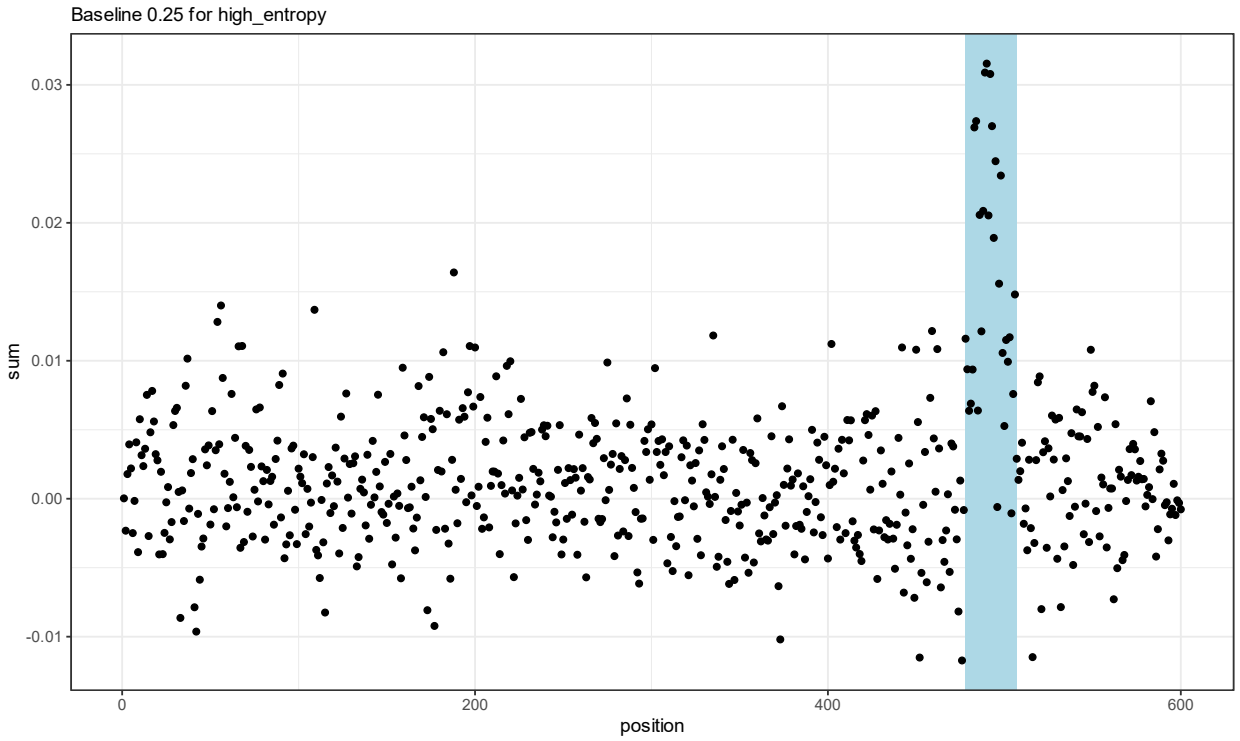
Seq logo plot. Avg.Match%: average percentage of *nt with max IG score == true nt in motif*

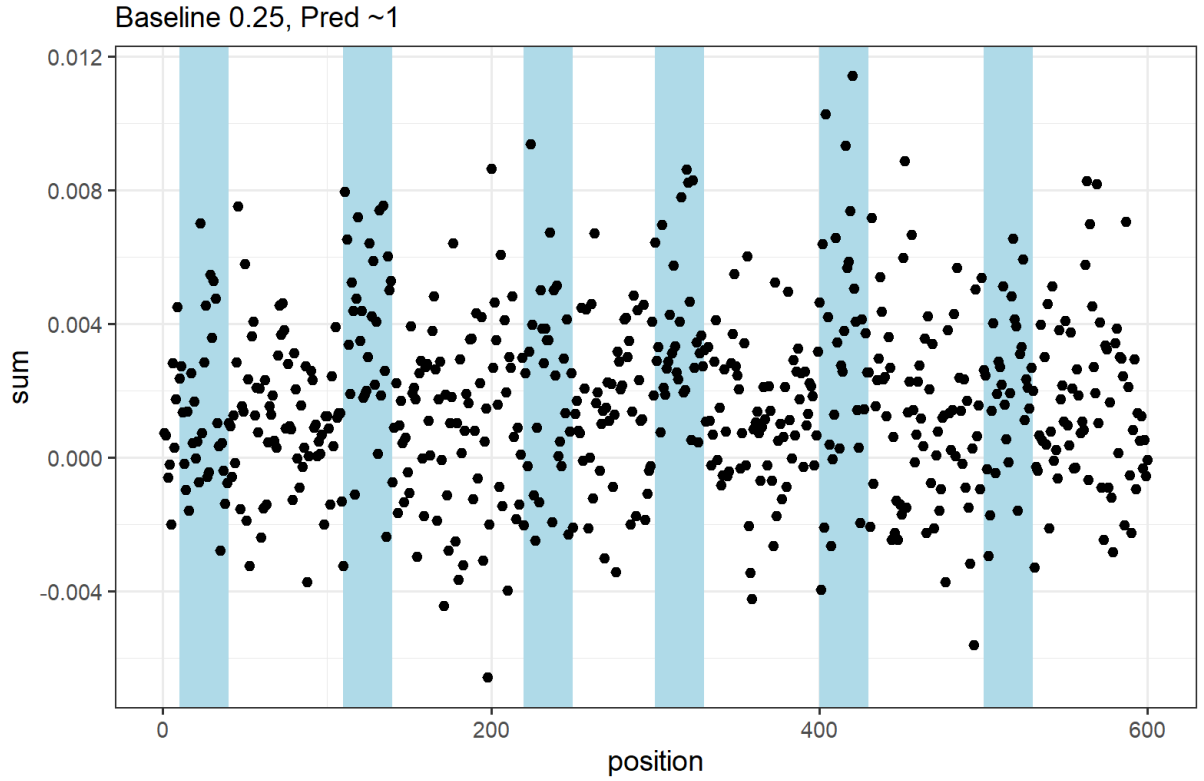




(Experiment 2: Recurrent Motifs) IG seems to lose precision in detecting motif if it is recurrent in the sequence.

Left: former example with high entropy motif.
Right: same motif, recurrent at about every 100 bp. Model: CNN, balanced accuracy ~ 97%





LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Sporulation

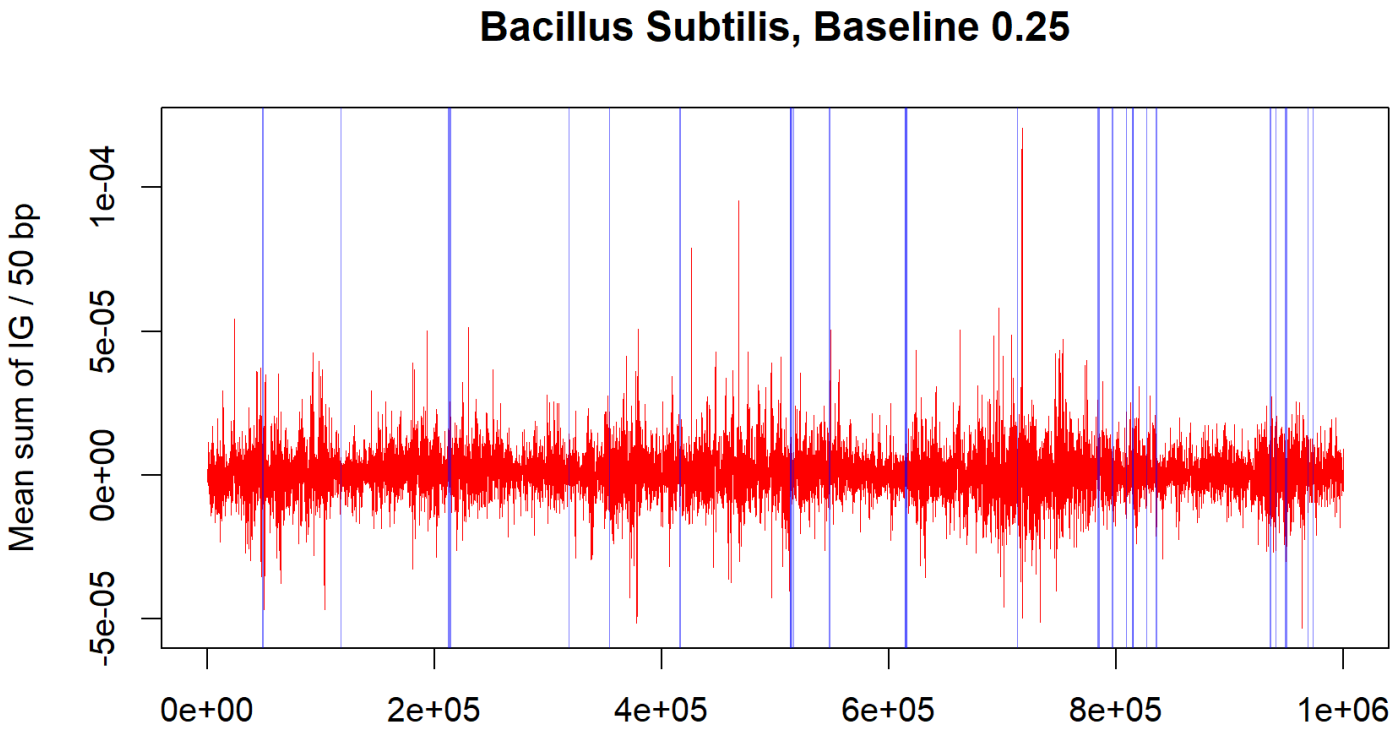
➔ Pretrained model with accuracy 96.25%

➔ Using the WA Subset. We also annotated each of them using the latest version of *Prokka*.

➔ Maxlen = 1 Mp. Usually computationally expensive operations are involved.

IG scores do not correlate with spore-related genes in general.

Instance: Bacillus subtilis, first 1 Mp subseq. Baseline: 0.25
Red curve is the zoo roll mean with k = 50.



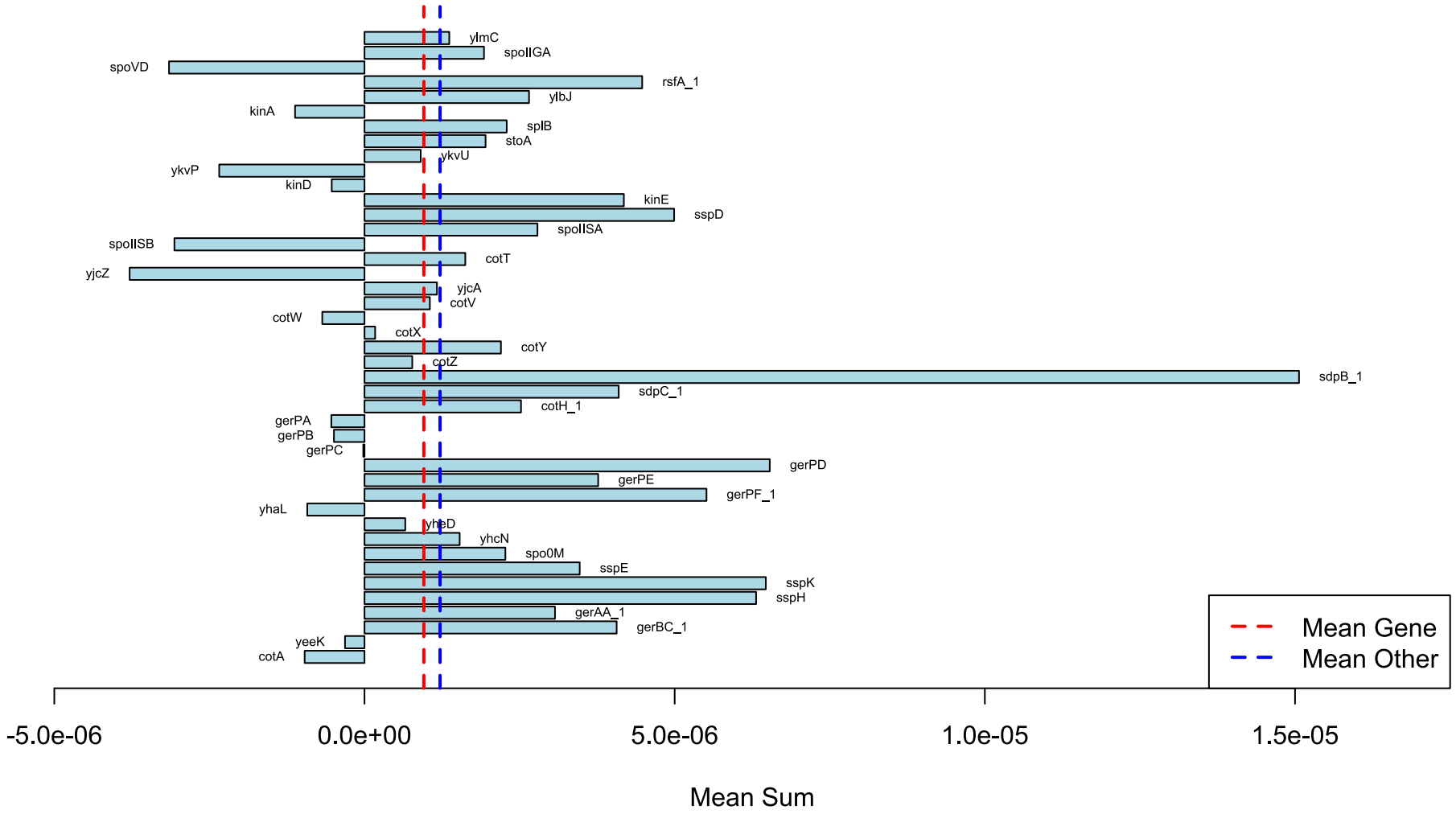
Annotated using downloaded GFF. Blue area indicates genes related to sporulation.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

IG scores do not correlate with spore-related genes in general.

Instance: Bacillus subtilis, first 1 Mp subseq. Baseline: 0.25

The spore-related genes can be positive or negative, at a not very impressive quantile.

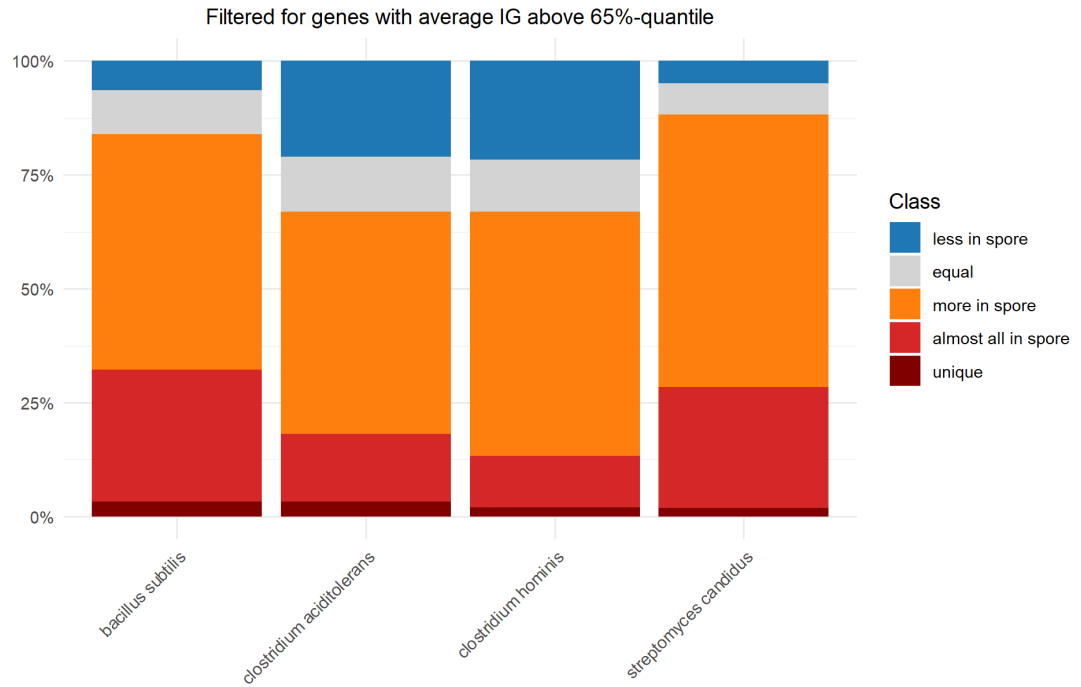


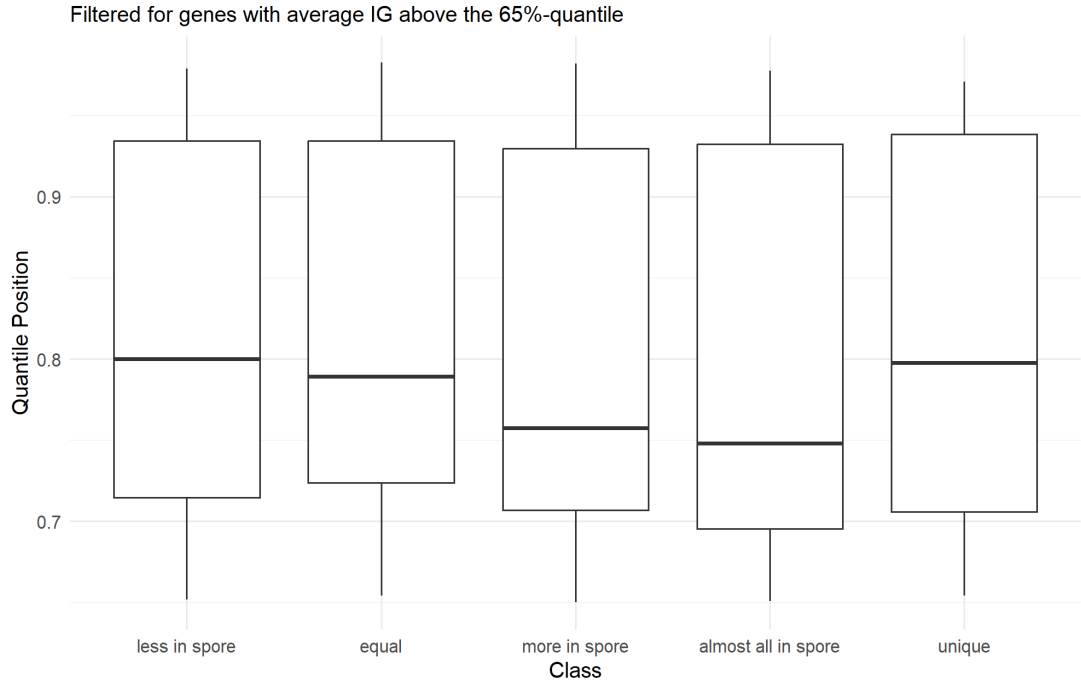
IG scores could be species-aware. Stress-related genes may be preferred.

Bacillus subtilis

Gene	Description	Quantile Position	Prevalence
steT	Serine/threonine exchanger SteT	0,927647	Equal
ohrR	Organic hydroperoxide resistance transcriptional regulator	0,833909	More in spore
sdpB_1	Sporulation-delaying protein SdpB	0,82038	Almost all in spore
opuE	Osmoregulated proline transporter OpuE	0,814637	Equal
yitU	5-amino-6-(5-phospho-D-ribitylamino)uracil phosphatase YitU	0,799016	More in spore
khtS	K(+)/H(+) antiporter modulator KhtS	0,772803	Unique to B. subtilis
nhaX	Stress response protein NhaX	0,753741	More in spore
yfiZ_1	putative siderophore transport system permease protein YfiZ	0,747927	More in spore
trpP	putative tryptophan transport protein	0,738799	Almost all in spore
yfkM	General stress protein 18	0,723087	More in spore
gutB_2	Sorbitol dehydrogenase	0,719846	More in spore
yodF_1	putative symporter YodF	0,713505	More in spore
gltT	Proton/sodium-glutamate symport protein	0,708389	More in spore
swrC	Swarming motility protein SwrC	0,702472	More in spore
fetB	putative iron export permease protein FetB	0,701218	More in spore

Consider investigating non-coding regions.





LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

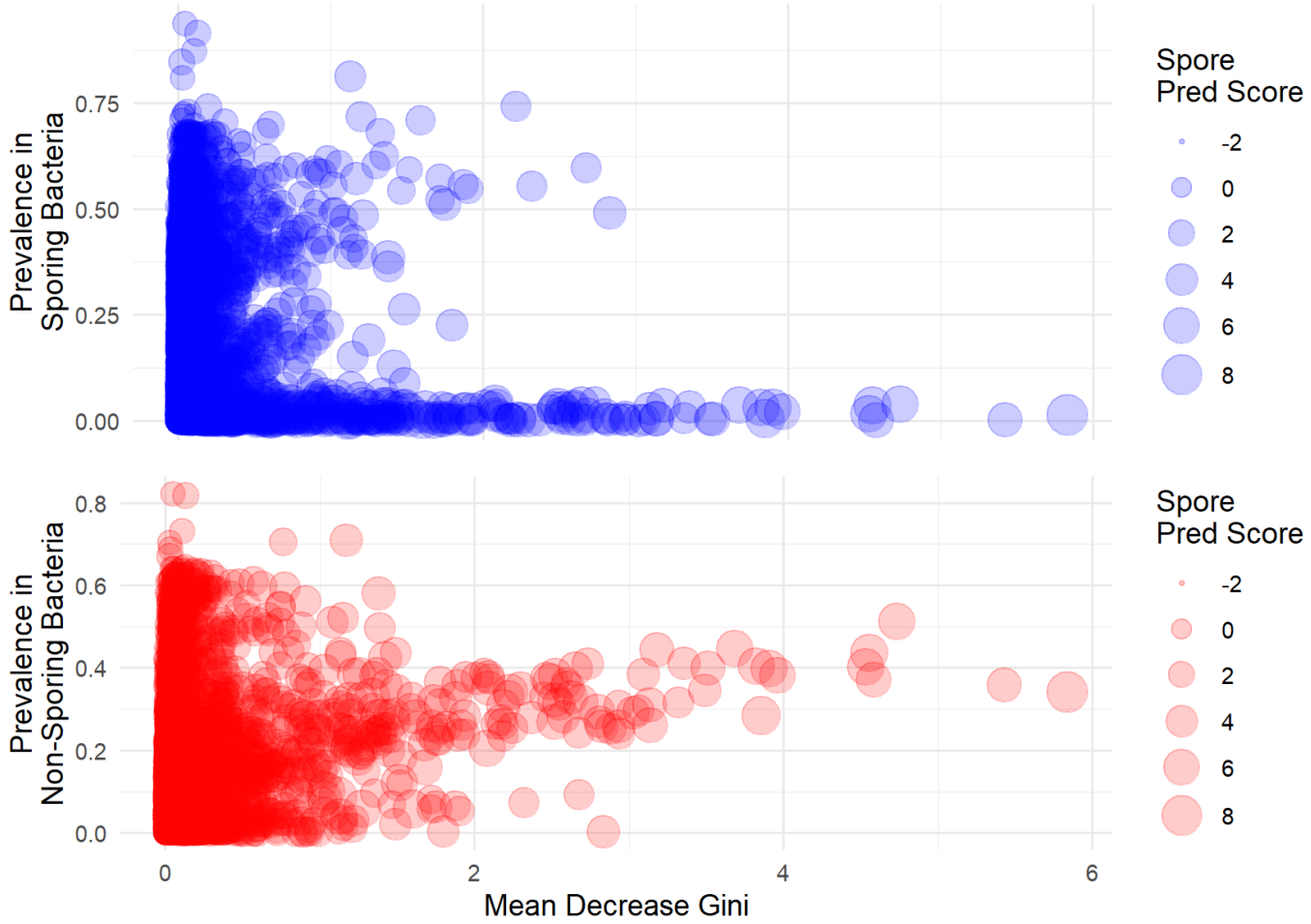
Random Forest taking binary gene matrix shows no preference for spore-related genes.

Random forest trained on gene names (0/1) for all named genes in all GFF files.
Tree = 500, OOB error rate = 3.14%. ~3k Obs vs ~40k features.
Low interpretability: genes rarely seen in spore-building bacteria have positive Spore scores?

Gene	Description	Mean Decrease Gini (RANKED)	non_spore	spore	Mean Decrease Accuracy	Prevalence in spore	Prevalence in non-spore	Class
pepN	Aminopeptidase N	5,84	-7,19	8,72	7,75	0,01	0,34	less in spore
ubiK	Ubiquinone biosynthesis accessory factor UbiK	5,43	-3,80	5,32	5,00	0,00	0,36	less in spore
dmlR_1	HTH-type transcriptional regulator DmlR	4,73	-3,72	6,42	6,06	0,04	0,51	less in spore
rlmJ	Ribosomal RNA large subunit methyltransferase J	4,58	-5,75	5,79	5,32	0,00	0,37	less in spore
btuB_1	Vitamin B12 transporter BtuB	4,56	-3,62	6,69	6,52	0,03	0,44	less in spore
ygfZ	tRNA-modifying protein YgfZ	4,53	-5,82	6,40	6,05	0,02	0,40	less in spore
gcvA_1	Glycine cleavage system transcriptional activator	3,96	-5,84	6,16	5,72	0,02	0,38	less in spore
hslU	ATP-dependent protease ATPase subunit HslU	3,90	-5,83	6,46	5,96	0,03	0,40	less in spore
oprM_1	Outer membrane protein OprM	3,85	-4,14	7,18	6,81	0,01	0,29	less in spore
bamD	Outer membrane protein assembly factor BamD	3,82	-5,27	6,38	5,62	0,03	0,41	less in spore
dmlR_2	HTH-type transcriptional regulator DmlR	3,68	-3,17	6,32	6,07	0,04	0,45	less in spore
secB	Protein-export protein SecB	3,51	-5,16	5,23	4,64	0,00	0,40	less in spore

Random Forest shows some preference for species-specific or rare genes.

Gene prevalence in spore building bacteria has negative corr with MDG.
Gene prevalence in non-spore-building bacteria has positive corr.




```
Call: lm(formula = MeanDecreaseGini ~
spore_prop + non_spore_prop, data =
combined_df2)

Coefficients:
(Intercept) 0.0063384 ***
spore_prop -0.2539945 ***
non_spore_prop 0.9482056 ***

Residual standard error: 0.1437 on 40637
degrees of freedom
Multiple R-squared: 0.1333,
Adjusted R-squared: 0.1332.
F-statistic: 3124 on 2 and 40637 DF,
p-value: < 2.2e-16
```

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

3. Ideas

i. How is a gene preferred? Examination on nucleotide level.

ii. How is a non-coding area preferred?

iii. Blacking a selected gene out, how would it affect prediction and interpretation?

iv. How can we improve the model?

v. Will different baselines provide different results?

vi. …

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN