Marie F. Jensen

DirectPred Status Update

From detecting *Listeria* to understanding the risk

- *Listeria Monocytogenes* is routinely monitored in the food industry because it can cause severe disease.

- Routine testing answers "**Is *Listeria* present**"

- **DNA sequencing** can provide more detailed information
  - What type of *Listeria* is it?
  - How concerning is it?
  - What traits could affect control and cleaning?

- BUT this requires **bioinformatics expertise and time**

DirectPred project

Goal

Develop an open-access tool for the food industry to characterise *Listeria monocytogenes* using long-reads and machine learning



DirectPred

1. Detection (established)

Test sample

Enrichment
ISO 11290-1:2017
(48 hours)

qPCR-positive

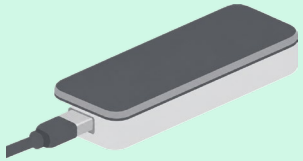
Current Workflow

> 3 weeks

Purification and colony picking → Transport to external facility → Single isolate sequencing → Expert based analysis and interpretation

DirectPred Workflow

1 workday



Direct ONT sequencing → Genome assembly → Characterisation and AI predictions

- Strain type
- Virulence potential
- Disinfectant tolerance

Faster information

From multiple weeks to a single workday

Automated and user friendly

Reduced need for bioinformatic expertise and easily interpretable results

Actionable results

Strain type, virulence and disinfectant tolerance

Where we are

1. Development

Stage 1 – controlled DNA mixtures

How accurately can the genomes be assembled?

Which bioinformatic tools should be used?

Under which conditions does the pipeline fail?

Stage 2 – Inoculated food samples

How does the food microbiomes affect the enrichment?

Do we get enough *Listeria* DNA for reliable analysis?

Future work

2. Peracetic acid ML

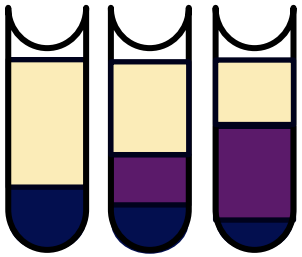
3. Validation

4. Implementation

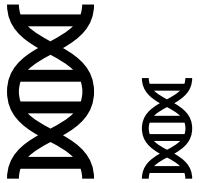
Development - Stage 1 composition

Prepare 51 samples with DNA mixtures from the **5** selected test strains – No enrichment

- 3 different *L. monocytogenes* (pathogenic), 1 *L. innocua* and 1 *L. welshimeri*
- One target *L. monocytogenes* strain (LM1)



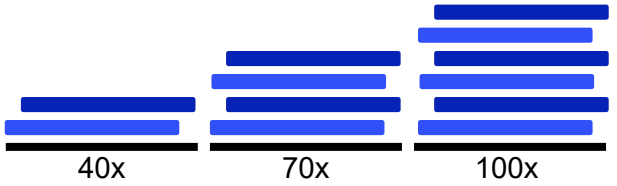
Different DNA proportions



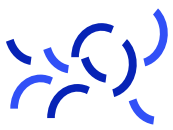
Different ng of DNA



ONT sequenced

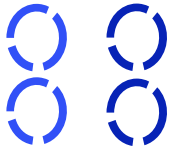


Adjusted to different depths



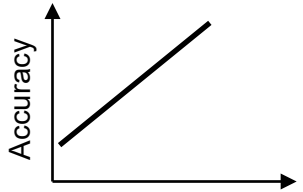
Assembly

- **metaFlye**
- **metaMDBG**
- **myloasm**



Binning

- **metabat2**
- **metacoag**
- **vamb**



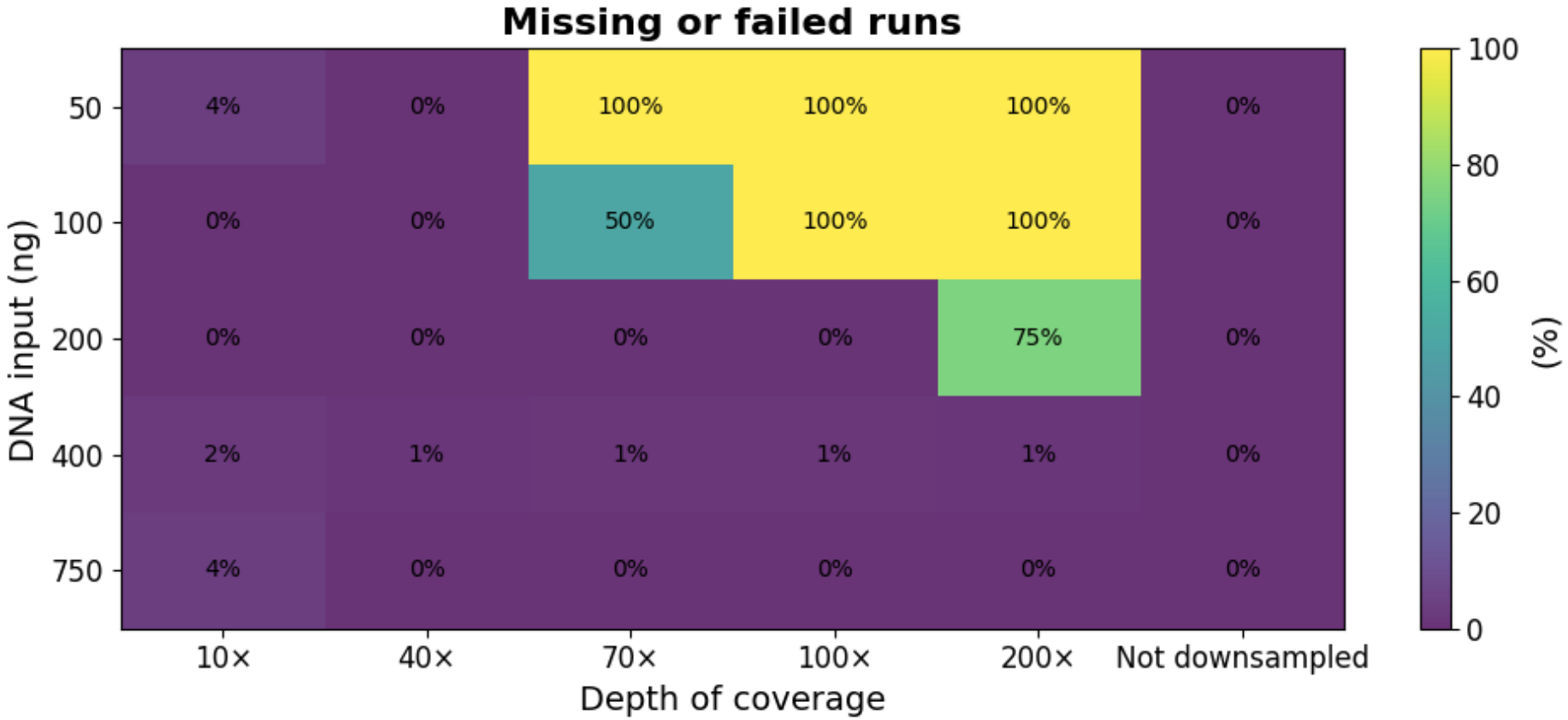
Scoring

- **LM1 focused**

5 things are varied:

- DNA proportions
- ngDNA in the sample
- Target depth of coverage
- Assembler
- Binner

Overall recovery



- ~9% of all possible runs produced no output
- There were not always enough DNA to downsample to higher depths

How are recovered genomes scored?

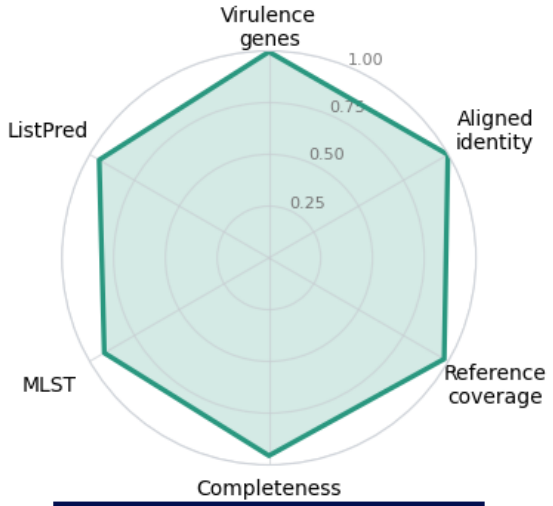
Do we find the same virulence genes?		Do we get the correct genome size?		How similar is the sequence to the reference?		Do we get the same MLST?		How much of the reference genome was reconstructed?		Did we get the same predicitons?
Virulence genes	+	Completeness	+	Alignment identity	+	MLST	+	Coverage	+	ListPred
Ratio		Ratio		Score		Binary		Score		Binary

$$\textbf{Total score} = \frac{\textbf{Virulence genes} + \textbf{Completeness} + \textbf{Alignment Identity} + \textbf{MLST} + \textbf{Coverage} + \textbf{ListPred}}{6}$$

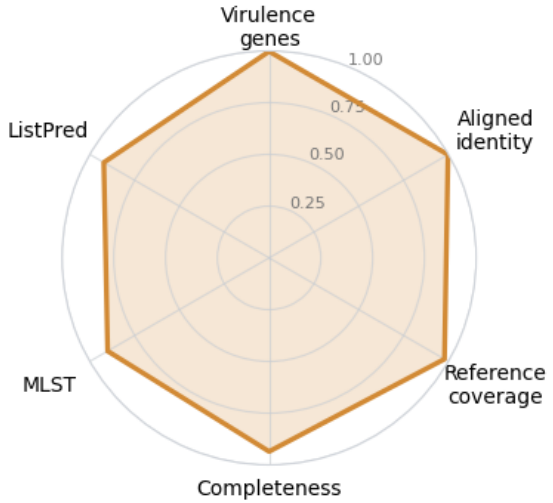
0 Poor → **1** Perfect

Score range 0 to 1

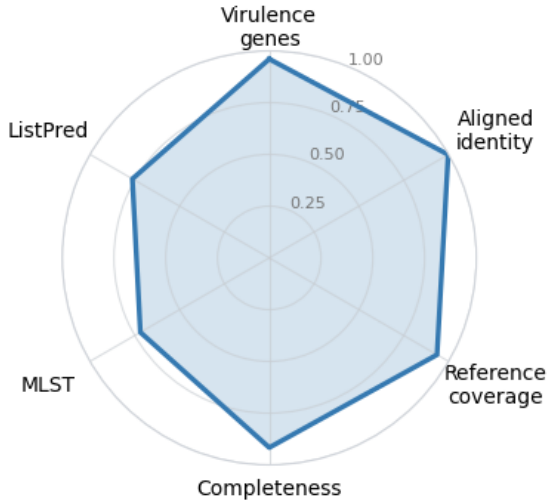
Best assembler-binner



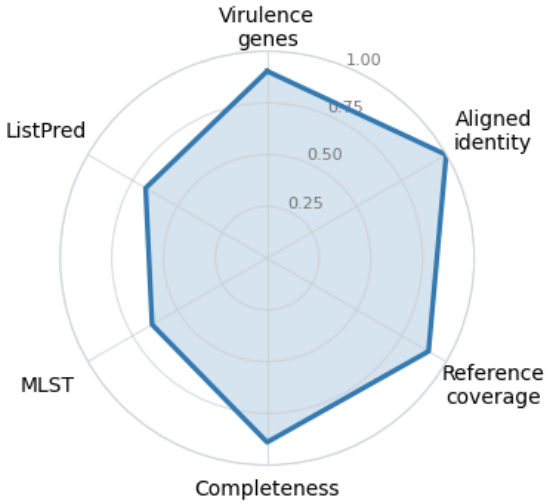
metaMDBG + MetaBAT2
Mean score: 0.965
Recovered: 281 / 281 (100.0%)



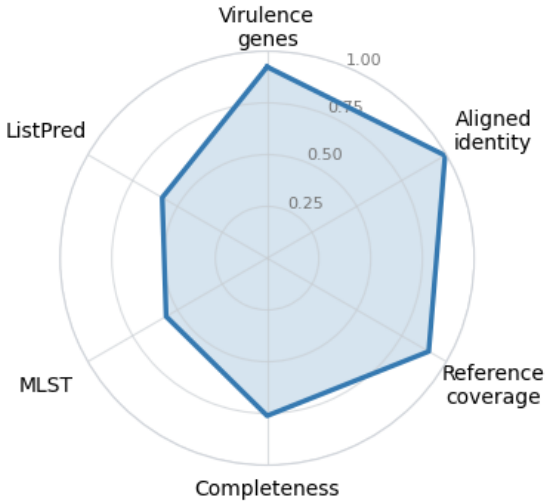
myloasm + MetaBAT2
Mean score: 0.955
Recovered: 281 / 281 (100.0%)



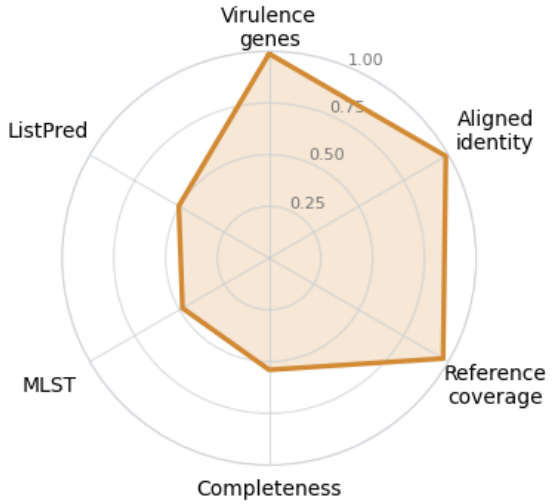
metaFlye + MetaBAT2
Mean score: 0.882
Recovered: 281 / 281 (100.0%)



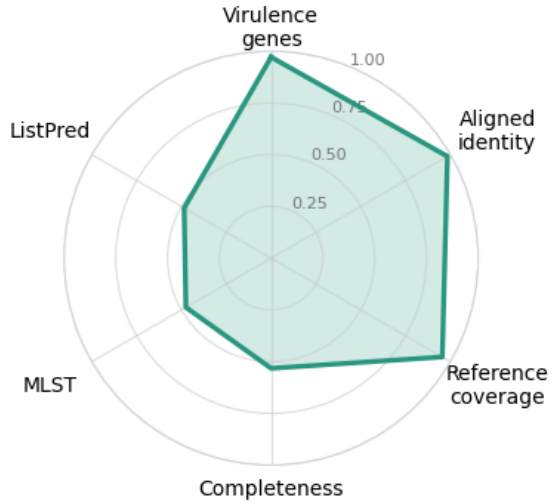
metaFlye + MetaCoAG
Mean score: 0.835
Recovered: 265 / 281 (94.3%)



metaFlye + VAMB
Mean score: 0.788
Recovered: 277 / 281 (98.6%)

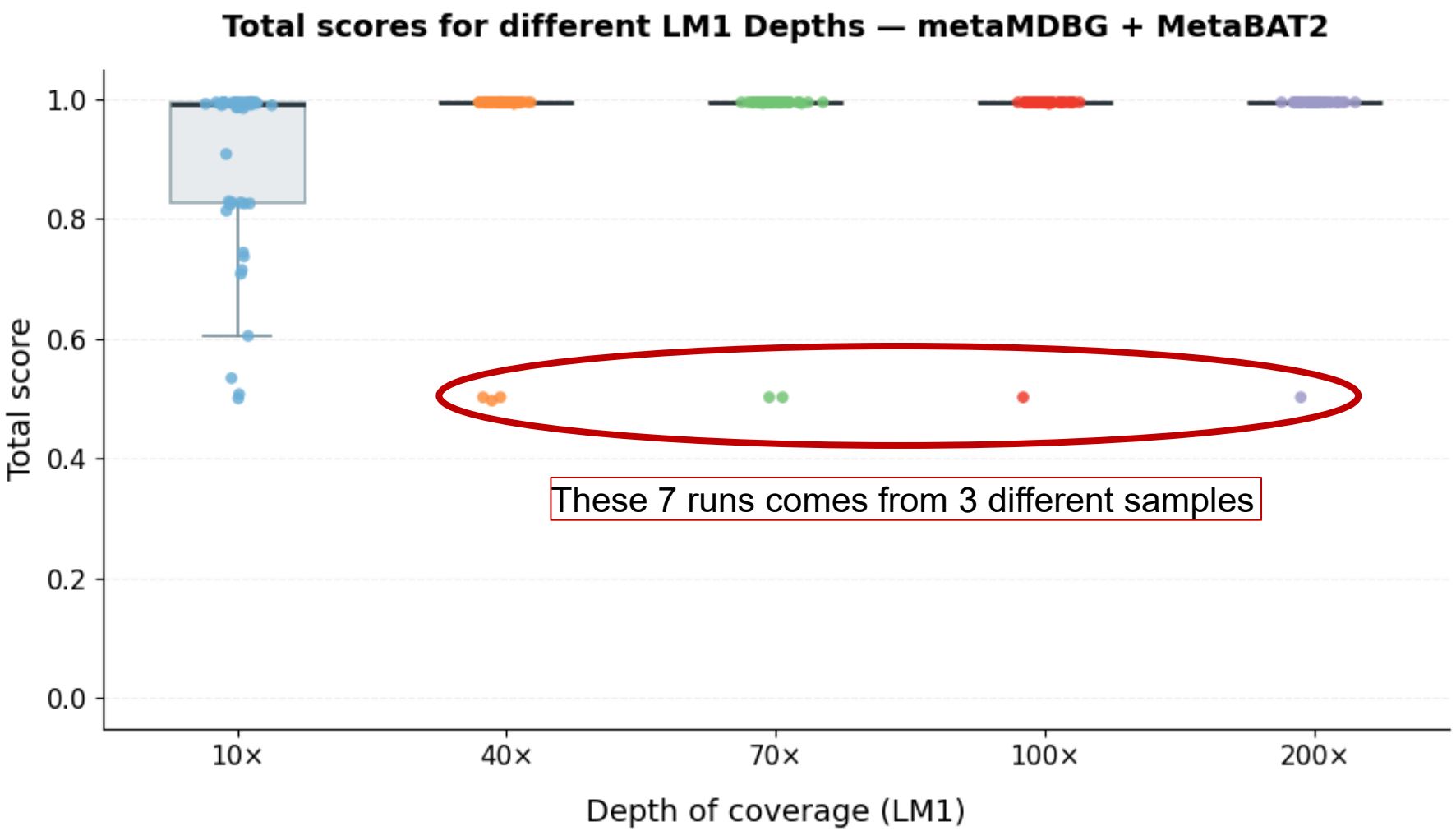


myloasm + VAMB
Mean score: 0.745
Recovered: 265 / 281 (94.3%)



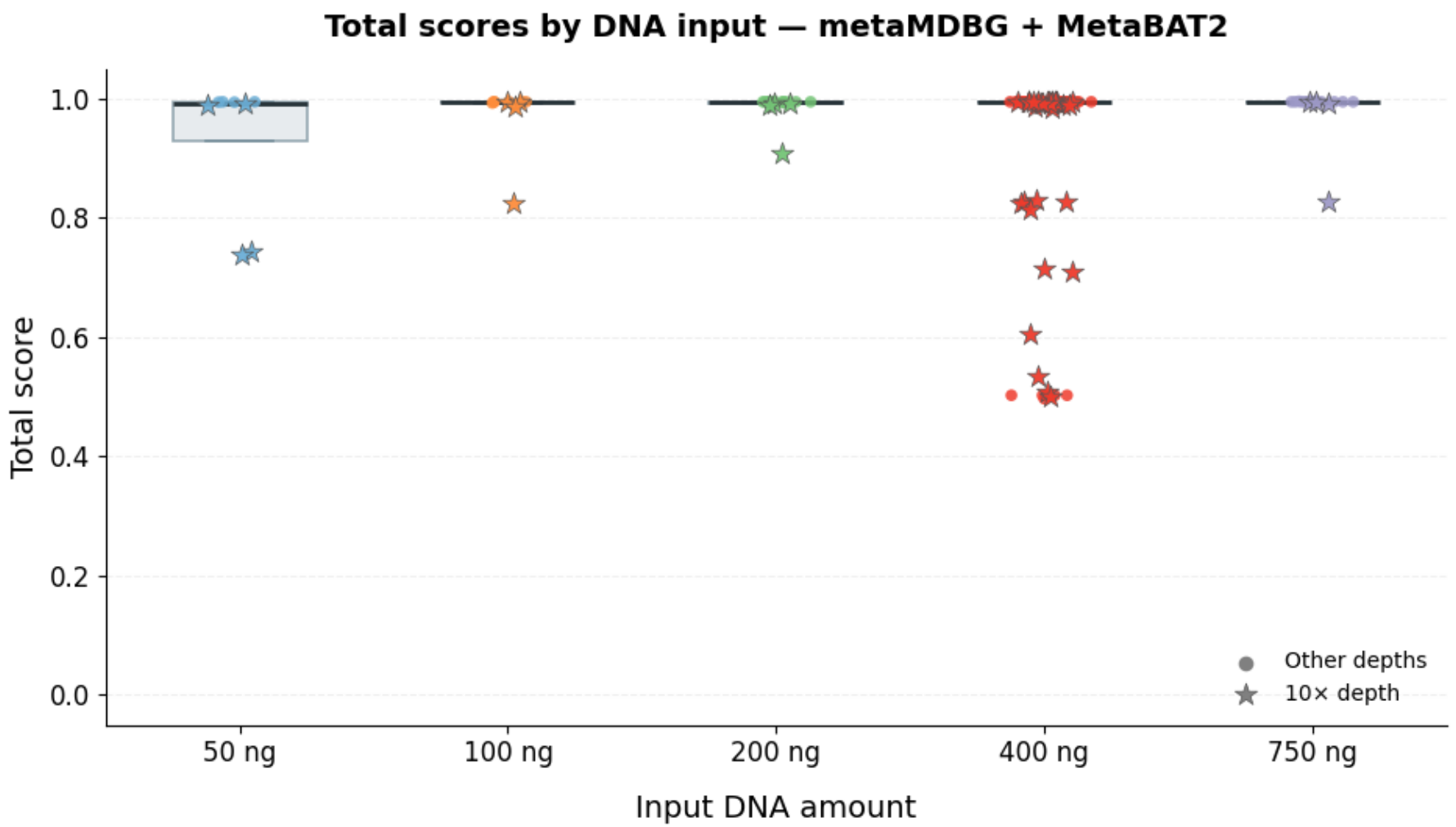
metaMDBG + VAMB
Mean score: 0.734
Recovered: 263 / 281 (93.6%)

Depth of Coverage



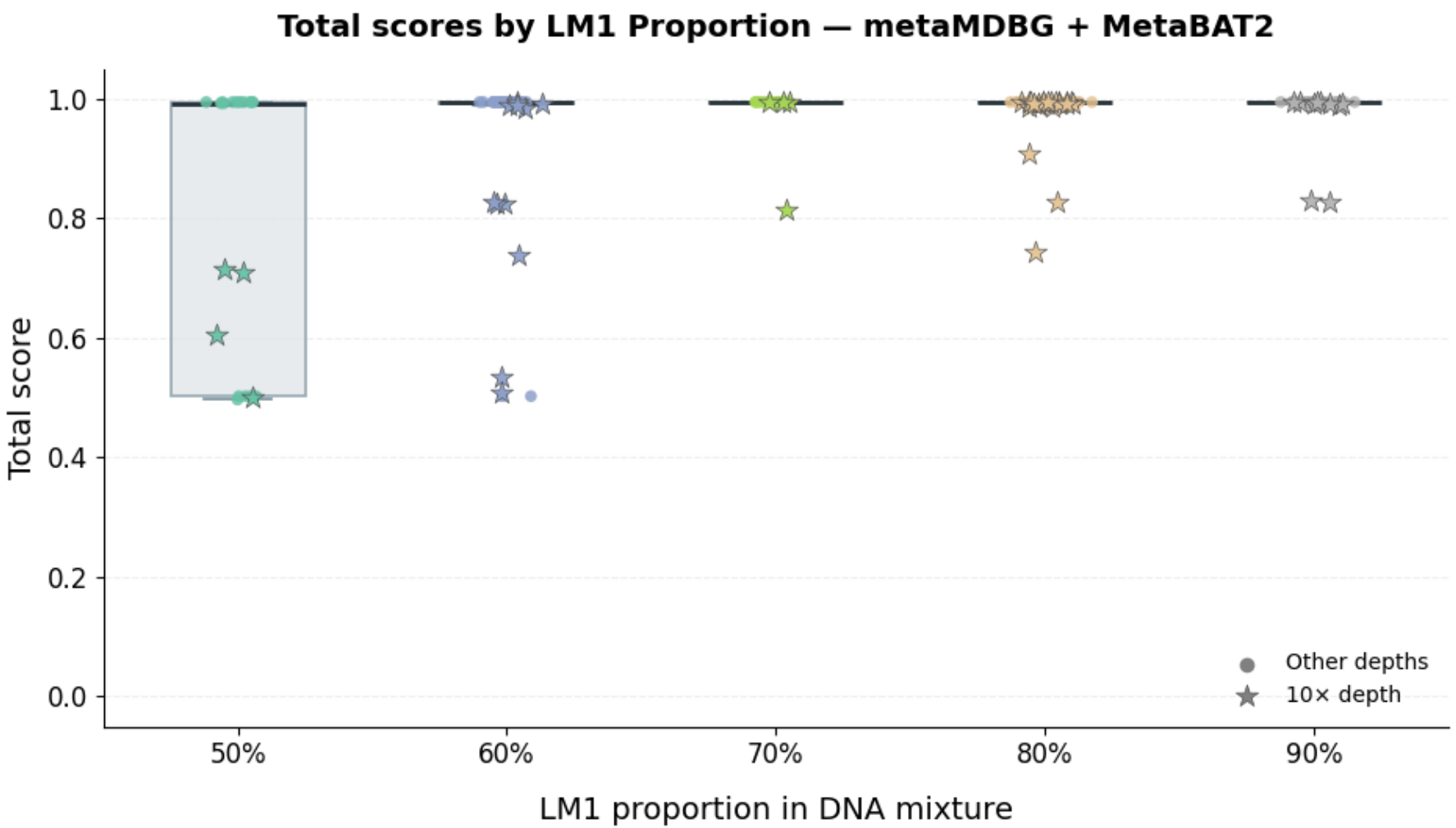
- 40x and above have little difference
- The lowest scoring for >10x results have 1 thing in common:
  - Genome size was around 5 to 6 Mbp compared to target 3

ngDNA scores



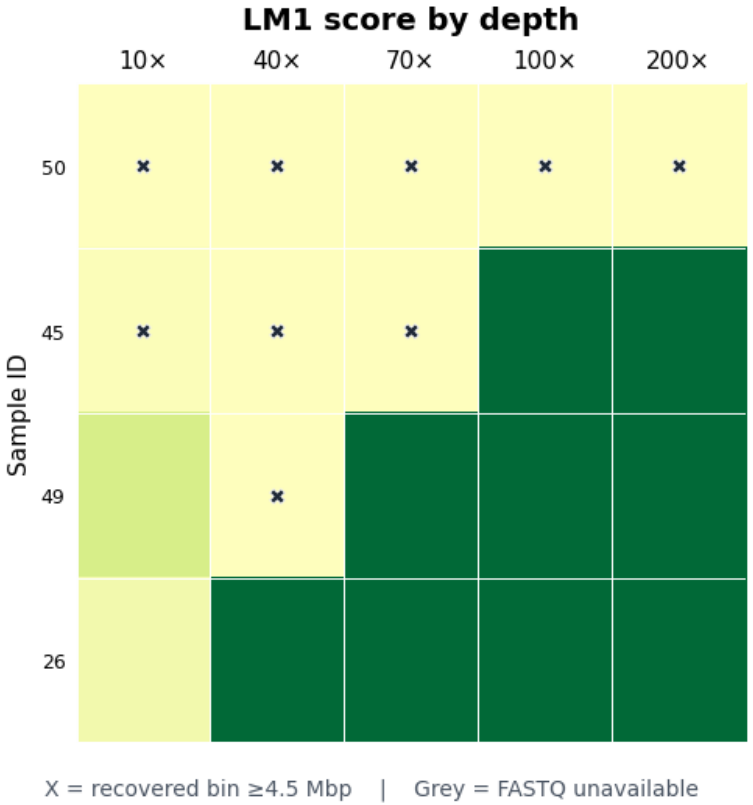
- Low DNA input means we don't get high depth
- Low scoring samples are mostly because of low depth

Target LM1 proportion



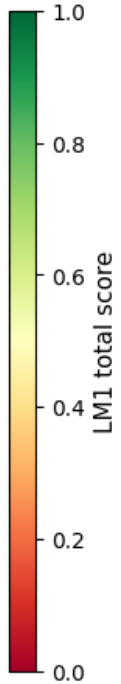
- With low LM1 % the scores decrease

Trouble samples



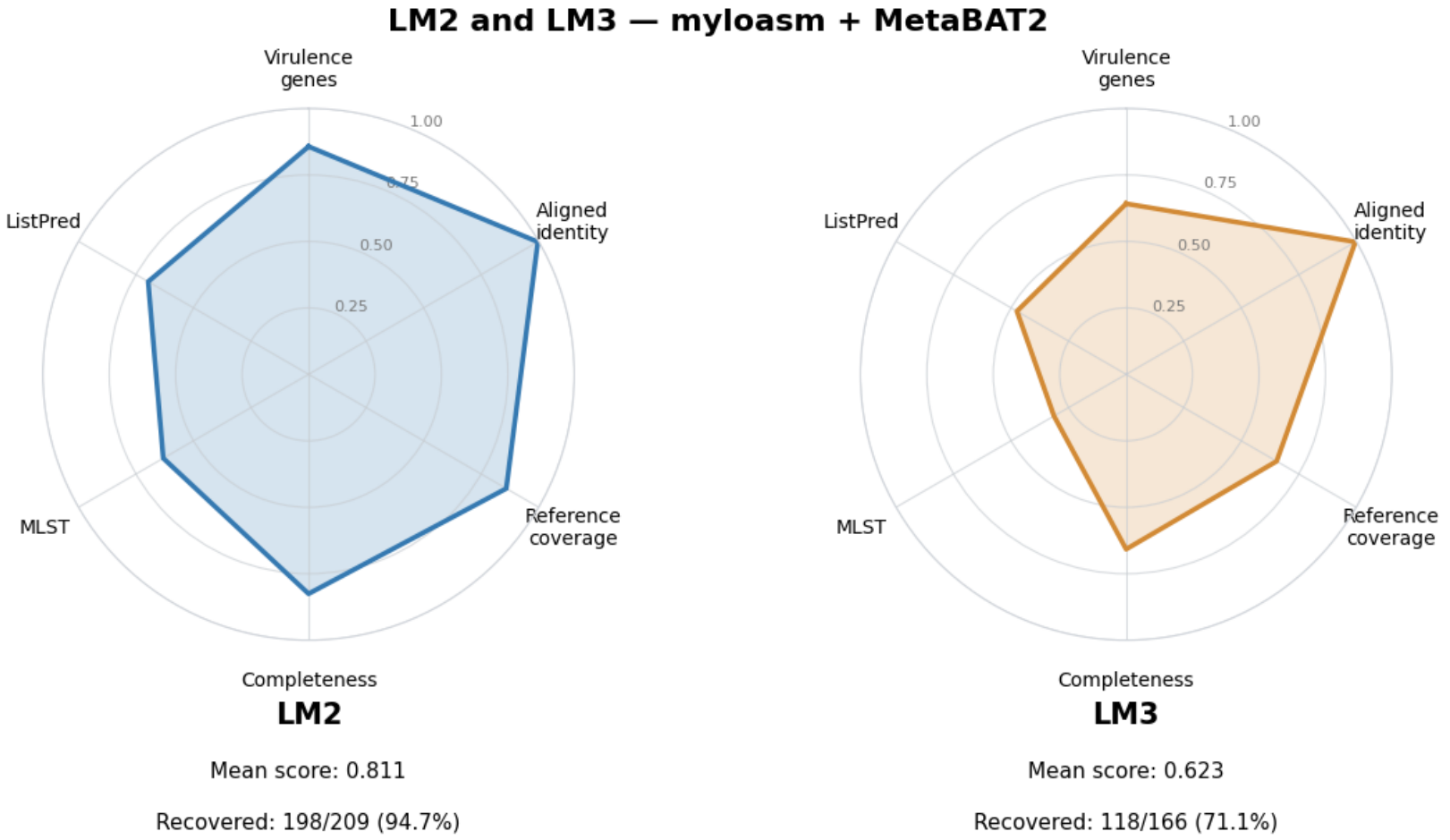
DNA mixture composition

	LM1 %	LM2 %	LM3 %	Lw %	Li %	ngDNA
50	50%	0%	10%	0%	**40%**	400
45	60%	0%	0%	10%	**30%**	400
49	50%	40%	0%	10%	0%	400
26	60%	30%	10%	0%	0%	400



- High % of Li (*L. Innocua*) causes the binner to mix Lm1 and Li
- This is particularly concerning because Li has been observed to co-enrich with Lm

What about lower abundance strains?



- LM2 results improve by increasing depth
- Results improve with a different assembler

Stage 1 main take aways

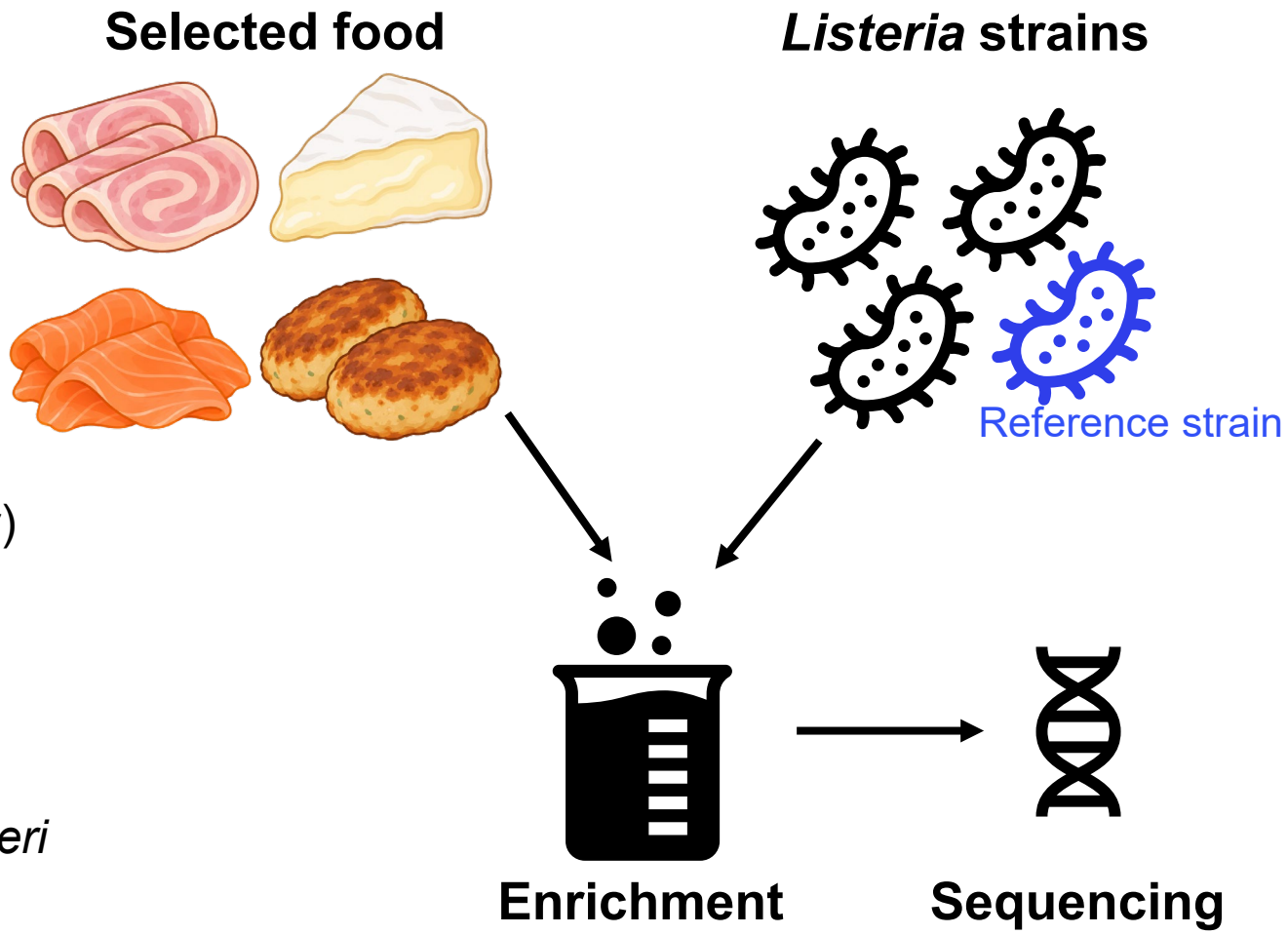
- Depth and proportion % matters
  - Depth 40x and above is robust for LM1
  - % LM1 matters only in the context of what strains are in the mix
- When depth is adequate, **poor LM1 scores are driven by oversized/mixed bins**
- *Listeria Innocua* causes trouble
- Not robust scores for low abundance strains but improves when changing the assembler

Stage 2

- 4 different risk foods
  - Soft cheese (from arla?)
  - Sliced meat (rullepølse)
  - Salmon (smoked or cured)
  - Fish cake/patties (fiskefrikadeller)

- Include the same LM strains but switch one out with the reference strain from ISI
- Include *L. innocua* but not *welshimeri*

- Environmental samples?



Questions?