A novel statistical approach for the identification of DNA sequences characterised by single-molecule fluorescence microscopy

Raffaele Vitale, Arno Bouwens, Jochem Deen, Kris Janssen, Cyril Ruckebusch, Johan Hofkens
Where everything started: the ADGut project (H2020)

These findings add Alzheimer’s disease to the growing list of diseases associated with gut microbial alterations, as well as suggest that gut bacterial communities may be a target for therapeutic intervention.

Vogt, N.M. et al., Scientific Reports, 2017 (7), 13537

Determining the composition of the gut microbiome in real samples is **fundamental** to develop new diagnostic tools and maybe even come up with a way to delay the onset of this neurodegenerative disorder.

How can this be done?
DNA mapping by single-molecule fluorescence microscopy

Which species does this subsequence come from?
What if our target species is not present in the reference database?

How reliable is an assignment based on such a global maximum?
Assessing the statistical significance of the matching

Statistically significant matching ($p$-value $<< 0.05$)
A first *fast* attempt: Lambda vs. T7

![Bar chart showing the fraction of assigned fragments (p-value < 0.05) for Bacteriophage T7 and Lambda. The bar for T7 is significantly higher than that for Lambda.](image-url)
A more comprehensive test: a simulated dataset

- T7 - V01146
- Stx2 Phage I - AP004402
- Stx1 Phage I - NC_004913
- Stx2 Phage II - AP005154
- Entero. HK629 - NC_019711
- Lambda – J02459
- Entero. HK630 – NC_019723
- Entero. EP390 – NC_019723
- HK97 – NC_002167
- HK022 – NC_002166

Pair-wise Jukes-Cantor distance

Assignation Matrix - $L_{eff}$ 70% - Widefield - $\alpha = 0.05$
Super-resolution fluorescence microscopy

Cell lysis → DNA extraction → DNA labelling → DNA stretching → Microscopy imaging

Pattern extraction

Super-resolution Optical Fluctuation Image (SOFI image)
Reducing ambiguity by increasing resolution

**Average ROC curve - $L_{\text{eff}} 70\%$**

- Sensitivity, True positive rate
- 1-Specificity, False positive rate

**Average area under ROC curve - $L_{\text{eff}} 70\%$**

- Area under ROC

Legend:
- Wide-field
- SOFI 2
- SOFI 3
Reducing ambiguity by statistical bootstrapping
Conclusions

- A novel permutation-based computational procedure was developed to assess the statistical reliability of the identification of DNA sequences characterised by single-molecule fluorescence microscopy.

- Such a procedure exhibited a very promising performance in a simulated and a real case-study involving viral species.

- Both super-resolution and resampling techniques like bootstrapping improved the matching quality reducing the ambiguity given by multiple assignations.

- The potential of the implemented methodology is currently being assessed in more complex real scenarios also involving bacterial species.
Many thanks for your kind attention